Forecasting Research Institute

The Longitudinal Expert AI Panel

Understanding Expert Views on AI Capabilities, Adoption, and Impact

Authors: Connacher Murphy, Josh Rosenberg, Jordan Canedy, Zach Jacobs, Nadja Flechner, Rhiannon Britt, Alexa Pan, Charlie Rogers-Smith, Dan Mayland, Cathy Buffington, Simas Kučinskas, Amanda Coston, Hannah Kerner, Emma Pierson, Reihaneh Rabbany, Matthew Salganik, Robert Seamans, Yu Su, Florian Tramèr, Tatsunori Hashimoto, Arvind Narayanan, Philip E. Tetlock, Ezra Karger

First released on November 10, 2025
FRI Working Paper #5

The Longitudinal Expert AI Panel:

Understanding Expert Views on AI Capabilities, Adoption, and Impact

Authors: Connacher Murphy, Josh Rosenberg, Jordan Canedy, Zach Jacobs, Nadja Flechner, Rhiannon Britt, Alexa Pan, Charlie Rogers-Smith, Dan Mayland, Cathy Buffington, Simas Kučinskas, Amanda Coston, Hannah Kerner, Emma Pierson, Reihaneh Rabbany, Matthew Salganik, Robert Seamans, Yu Su, Florian Tramèr, Tatsunori Hashimoto, Arvind Narayanan, Philip E. Tetlock, Ezra Karger[1]

Abstract

Public debates about AI revolve around bold claims and counterclaims, but rarely culminate in precise, falsifiable forecasts of AI capabilities, adoption, and impact. The Longitudinal Expert AI Panel (LEAP) improves this signal-to-noise ratio by gathering such forecasts, monthly, from a carefully chosen panel of 339 experts spanning industry, academia, and policy. The median expert foresees that by 2030 AI will be responsible for 7% of U.S. electricity usage, assist in 18% of work hours in the U.S., and provide daily companionship for 15% of adults—roughly 7x, 4x, and 2.5x current levels, respectively. The median expert also gives a 60% chance that AI systems solve or substantially assist in solving a Millennium Prize Problem by 2040, which would be a major achievement in mathematics. There is substantial within-individual uncertainty and between-individual disagreement among experts, each accounting for roughly half of the total variation in expert forecasts across all questions. Nevertheless, the vast majority of LEAP forecasts fall far short of the warnings from AI lab leaders about imminent artificial superintelligence. We analyze 1.7 million words of participant rationales to provide a complementary qualitative overview of the key mechanisms underpinning fast and slow forecasts of AI progress.

---

[1] Corresponding author: Ezra Karger, ezra.karger@chi.frb.org
Affiliations:

1. **Forecasting Research Institute**: Connacher Murphy, Josh Rosenberg, Jordan Canedy, Zach Jacobs, Nadja Flechner, Rhiannon Britt, Alexa Pan, Charlie Rogers-Smith, Dan Mayland, Cathy Buffington, Simas Kučinskas
2. **University of California, Berkeley**: Amanda Coston, Emma Pierson
3. **Arizona State University:** Hannah Kerner
4. **McGill University:** Reihaneh Rabbany
5. **Princeton University**: Matthew Salganik, Arvind Narayanan
6. **New York University**: Robert Seamans
7. **The Ohio State University**: Yu Su
8. **ETH Zürich**: Florian Tramèr
9. **Stanford University**: Tatsunori Hashimoto
10. **University of Pennsylvania**: Philip E. Tetlock
11. **Federal Reserve Bank of Chicago**: Ezra Karger

The views expressed in this paper do not necessarily represent the views of the Federal Reserve Bank of Chicago or the Federal Reserve System. This research was funded with support from Open Philanthropy and Craig Falls. We thank Victoria Schmidt, Morgane Bascle, and Jonah Black for research assistance.

Table of Contents

[Abstract](#)
[Executive Summary](#)
[Introduction](#)
[Connection to Prior Work](#)
[Panel Construction](#)
[Monthly Surveys and Forecasting Questions](#)
[Results](#)
[Next Steps](#)
[Conclusion](#)
[References](#)
[Appendix A. Panel Construction and Sampling](#)
[Appendix B. Monthly Surveys and Forecasting Questions](#)
[Appendix C. Pooled Distribution Estimation](#)
[Appendix D. Public Accuracy Stratification](#)
[Appendix E. Survey Questions](#)
[Appendix F. Question-by-Question Results](#)

Executive Summary

Despite the clashing narratives around AI, there is little work systematically mapping the full spectrum of views among experts (computer scientists, economists, technologists) and the general public. What do these groups believe about AI's future capabilities, adoption, and effects? And why do they believe what they do? The Longitudinal Expert AI Panel (LEAP) fills this gap with monthly surveys tracking the quantified beliefs of experts, historically accurate forecasters ("superforecasters"),[2] and the general public.

Policymakers and stakeholders routinely consult experts to ground their decision-making in coherent, informed beliefs, especially when faced with new technologies and high levels of uncertainty. LEAP facilitates this process by cutting through the anecdote and speculation that currently dominates AI discourse. We specifically target prominent experts whom policymakers would be most inclined to consult on the progression of AI capabilities and their impact. Expert participants include top-cited AI and ML scientists, prominent economists, key technical staff at frontier AI companies, and influential policy experts from a broad range of NGOs. Our goal is to provide a clear summary of expert views, analyzed within and across these key groups.

[2] Forecasters are denoted "superforecasters" if they (1) were in the top 2% of the accuracy distribution in a given year of the Intelligence Advanced Research Projects Activity (IARPA) Aggregative Contingent Estimation (ACE) tournament (IARPA ACE Program n.d.; Mellers et al. 2014) or (2) they were a highly accurate performer on Good Judgment Open, an online continuous geopolitical forecasting tournament. Good Judgment Inc., which runs Good Judgment Open, then adds these top forecasters to the "superforecaster" pool. Most superforecasters come from the first selection criteria. Mellers et al. (2015) finds persistent performance of these superforecasters across several years of geopolitical forecasting.

Since we launched LEAP in June 2025, we have completed three survey waves focused on (1) high-level predictions about AI progress; (2) the application of AI to scientific discovery; and (3) the adoption and social impact of AI. Experts provided thoughtful engagement, spending a median of 44 minutes per survey and writing over 460,000 words of rationales explaining their beliefs, across all surveys and questions.[3] In this paper, we share results from these first three waves along with a detailed methodological description of the project as a whole.

Across the first three waves of LEAP, five patterns stand out:

1. Experts expect sizable near-term societal effects from AI by 2040.
2. Substantial disagreement and uncertainty underlie expert forecasts.
3. The median expert expects much slower progress than prominent leaders of frontier AI labs.
4. Experts anticipate faster progress than the public on most outcomes.
5. Experts and superforecasters mostly agree. Where they disagree, experts tend to expect more AI progress. Also, there are no systematic differences between the predictions of different types of experts: computer scientists, economists, industry professionals, and policy professionals.

First, **experts expect sizable near-term societal effects**: by 2030, the median expert predicts that 18%[4] of all U.S. work hours will be assisted by generative AI, up from 4.1% in November 2024 (Bick et al. 2025);[5] AI training and deployment will consume 7% of U.S. electricity;[6,7] autonomous vehicles will provide 20% of U.S. ride-hailing trips, and annual global private investment (as reported by Our World in Data)[8] will reach $260 billion, up from the $130 billion reported total for this series in 2024. The median expert predicts that by 2030, 15% of adults will self-report using AI for companionship, emotional support, social interaction, or simulated relationships at least once daily, up from 6% today. By 2040, that number doubles to 30% of adults. Experts expect substantial improvements in the ability of AI systems to complete difficult

[3] Expert rationales averaged 92 words, with 25% of expert rationales exceeding 100 words and 8% exceeding 200 words. Among superforecasters, 47% of rationales exceeded 100 words and 23% exceeded 200 words.

[4] If not otherwise stated, we report values from the 50th percentile forecasts given by each expert. We elaborate on the use of quantile forecasting in Monthly Surveys and Forecasting Questions.

[5] Respondents were shown a historical baseline value of 2%, based on an earlier version of the cited paper. A new version of the working paper estimates a range of 1.6% to 6.6%. We select the midpoint, 4.1%, as the historical baseline value.

[6] The median expert expects electricity consumption used for AI to rise to 12% by 2040.

[7] See Appendix E.II., 4. Electricity Consumption for information on the baseline estimate of 1.0%.

[8] Regarding their private AI investment indicator, Our World in Data (2025) notes: 1. “The data likely underestimates total global AI investment, as it only captures certain types of private equity transactions, excluding other significant channels and categories of AI-related spending;” 2. “The source does not fully disclose its methodology and what's included or excluded. This means it may not fully capture important areas of AI investment, such as those from publicly traded companies, corporate internal R&D, government funding, public sector initiatives, data center infrastructure, hardware production, semiconductor manufacturing, and expenses for research and talent.” More details on what is likely excluded can be found at Our World in Data (2025).

math questions: the median expert believes that AI systems will achieve performance of 75% on the FrontierMath benchmark[9] by 2030; and 23% of experts expect saturation of the benchmark.[10] The median expert also believes it is more likely than not (a 60% chance) that AI solves or substantially assists in solving a Millennium Prize Problem by 2040.[11] We summarize these expert forecasts in the figure below.

[9] The FrontierMath benchmark consists of math problems that resemble those a math PhD student might spend several days solving.

[10] This estimate of 23% reflects the fraction of experts whose median forecast is that AI systems will achieve performance of at least 90% (which we call saturation) on Tiers 1–3 of FrontierMath. We take the average of the proportions calculated under weak and strict inequality.

[11] The Millennium Prize Problems are seven mathematical problems identified by the Clay Mathematics Institute in 2000 as the most important unsolved questions in mathematics (Clay Mathematics Institute n.d.). Only one has been solved to date.

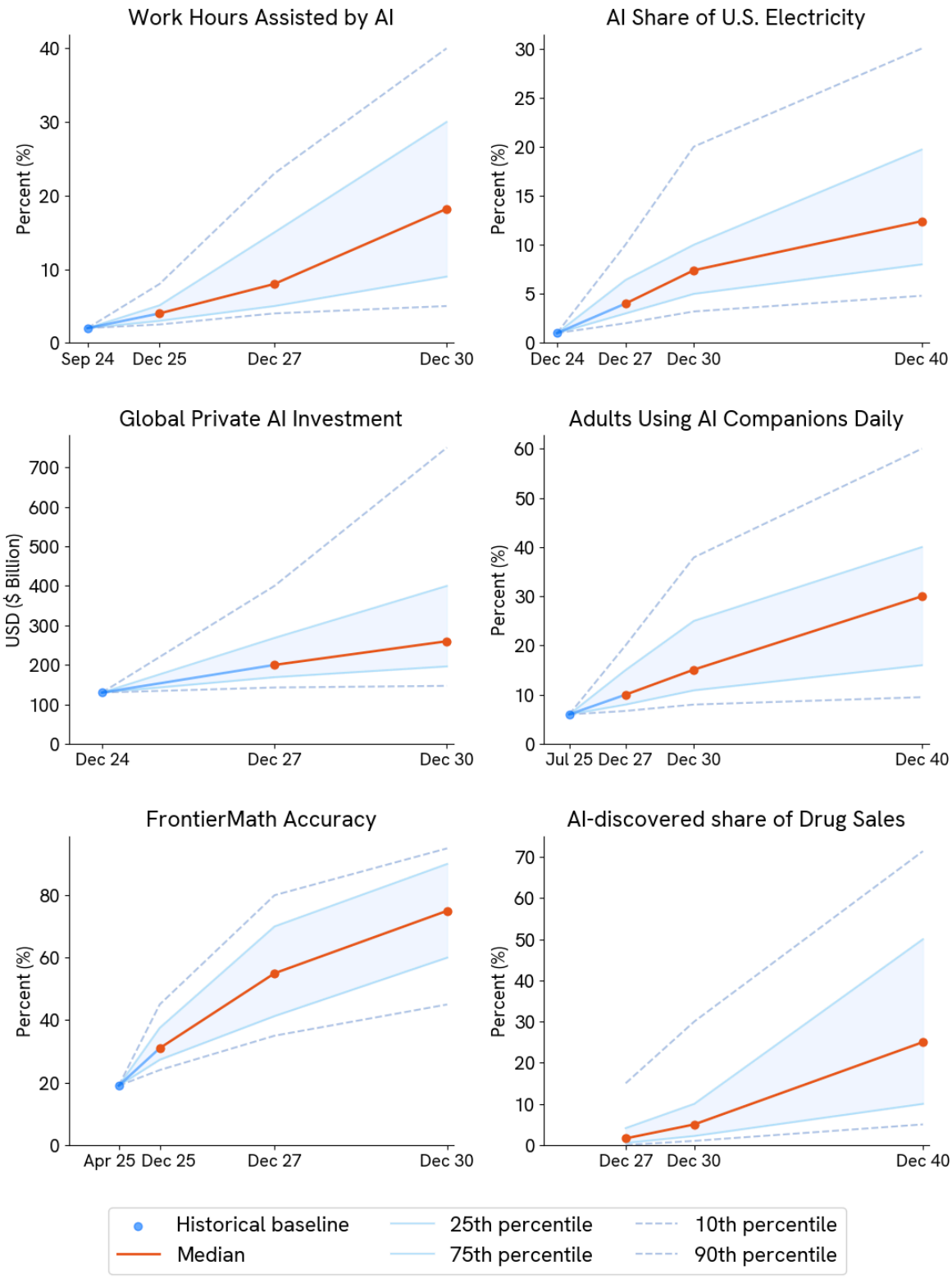


Figure: Median expert forecasts for various questions. We display the 10th, 25th, 50th, 75th, and 90th percentiles of the median forecasts given by experts at each date. For example, if 25% of experts give a median forecast of $10 billion or less, the 25th percentile series in the graph will lie at $10 billion; these series are *not* confidence intervals. Where available, we include a historical baseline in light blue.

Even experts with lower forecasts of AI progress, capabilities, and adoption still expect substantial impacts relative to current levels. Recall that, by 2030, the median expert predicts that 18% of all U.S. work hours will be assisted by generative AI, up from 4.1% in November 2024 (Bick et al. 2025);[12] however, respondents were shown a historical baseline value of 2%, based on an earlier version of the cited paper. We also asked forecasters for their 25th percentile forecast,[13] and panelists on this question gave a forecast of 9%, still more than a four-fold increase from the historical baseline level provided to forecasters at the time of the survey. In other words, the median expert gives a 75% chance that at least 9% of work hours will be assisted by generative AI in 2030.

Second, **substantial disagreement and uncertainty underlie expert forecasts**. The top quartile of experts estimate that the majority of revenue from newly approved U.S. drugs in 2040 will be from AI-discovered drugs, but the bottom quartile of experts predicts that less than 10% of new drug sales will be from AI-discovered drugs.[14] On the question of whether AI will independently solve or substantially assist in solving a Millennium Prize Problem by 2040, a quarter of experts think it is quite likely (>81% chance) while another quarter of experts believe it is unlikely (<30% chance). On many questions, individual experts also express substantial uncertainty in their own forecasts. We construct a composite measure of the total variation in expert beliefs in the figure below (for two questions), taking into account both the within-forecaster uncertainty and the between-forecaster disagreement. We discuss the methodology in greater detail in Uncertainty and Disagreement, where we describe the underlying assumptions that we rely on to construct these figures. These “pooled distributions” represent the full distribution of outcomes predicted by the average expert. We find that, across all forecasting questions where we allow forecasters to express their uncertainty, within-forecaster uncertainty explains 49% of the total variation in forecasts, while 51% of variation is explained by between-forecaster disagreement.

[12] Respondents were shown a historical baseline value of 2%, based on an earlier version of the cited paper. A new version of the working paper estimates a range of 1.6% to 6.6%. We select the midpoint, 4.1%, as the historical baseline value.
[13] A forecaster expects the realized outcome to be below their 25th percentile forecast in just 25% of cases, compared to 50% of cases for a median forecast. We additionally ask for a 75th percentile forecast whenever we collect a 25th percentile forecast.
[14] Forecasters are asked what percentage of sales revenue from recently FDA-approved drugs will come from those discovered using AI methods available after 2022. See Appendix E.II. 3. Drug Discovery.

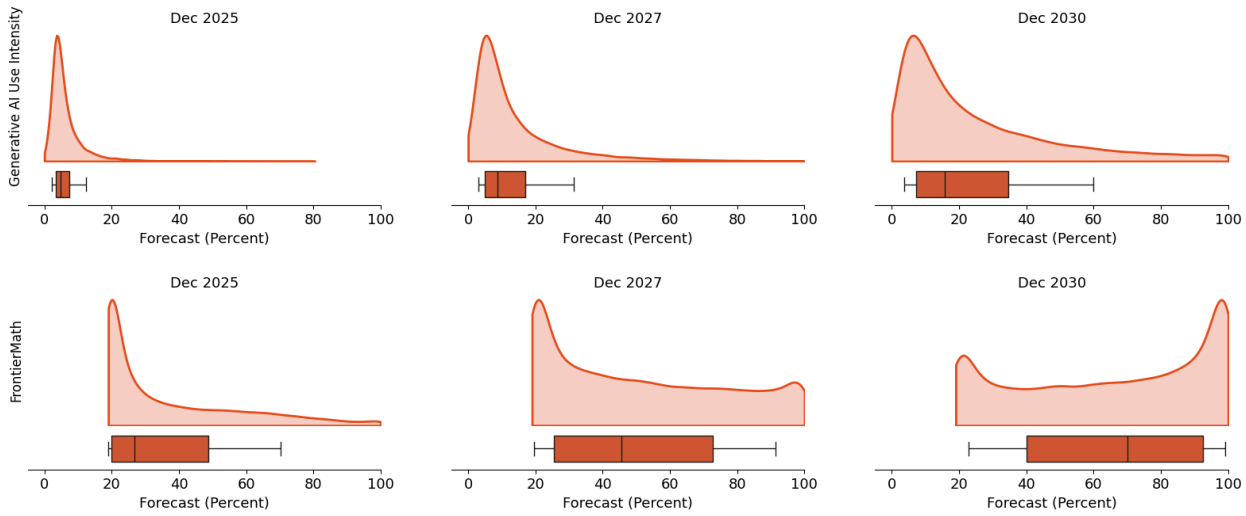


Figure: Pooled distributions for expert forecasts on *Work Hours Assisted by Generative AI* (top panels) and *FrontierMath scores* (bottom panels). These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. See Uncertainty and Disagreement for details.

Third, **the median expert expects much slower progress than prominent leaders of frontier AI labs**. These lab leaders predict human-level or superhuman AI by 2026–2029, while most of our expert panel rejects these shorter timelines. We ask respondents to forecast whether the average LEAP panelist will say, in 2030, that we are closest to a "slow-," "moderate-," or "rapid-" AI progress scenario. We define these scenarios in detail in the appendix (see Appendix E.I. 4. General AI Progress). The average expert thinks that 23% of LEAP panelists in 2030 will say the world most closely mirrors a "rapid" AI progress scenario that, of our three scenarios, most closely matches some of the lab leaders' claims. On the other hand, the average expert believes that 28% of panelists will indicate that progress plateaued at close-to-current levels, with fewer improvements in capabilities relative to today (a "slow" AI progress scenario).[15]

Fourth, **experts anticipate faster progress than the public**[16] on most outcomes (e.g., the accuracy AI systems will achieve on the FrontierMath benchmark by 2030: median expert 75% vs. median member of the public 50%, $p < 0.001$; generative-AI work assistance by 2030: median expert 18% vs. median member of the public 10%, $p < 0.001$). Experts predict a 32% chance that AI will be at least as impactful as a "technology of the millennium"—like the printing press or the Industrial Revolution—whereas the public gives this only a 22% chance. Across forecasts that exhibit a clear relationship with AI capabilities and impacts, a randomly selected

[15] Note that for two categorical questions about overall AI scenarios, we report averages instead of medians. Since respondents assign probabilities that sum to 100%, we use average aggregation to maintain this property.

[16] To assess the extent to which low-effort or relatively lower comprehension from the public could drive these results, we compare members of our public with high levels of forecasting accuracy in other studies to those with low accuracy. We do not find that one group systematically expects more or less progress. Public Accuracy Stratification details this analysis.

expert is 16% more likely than a randomly selected member of the public to predict faster progress than would be expected by random chance.

We summarize some of the differences in aggregate forecasts in the figure below.

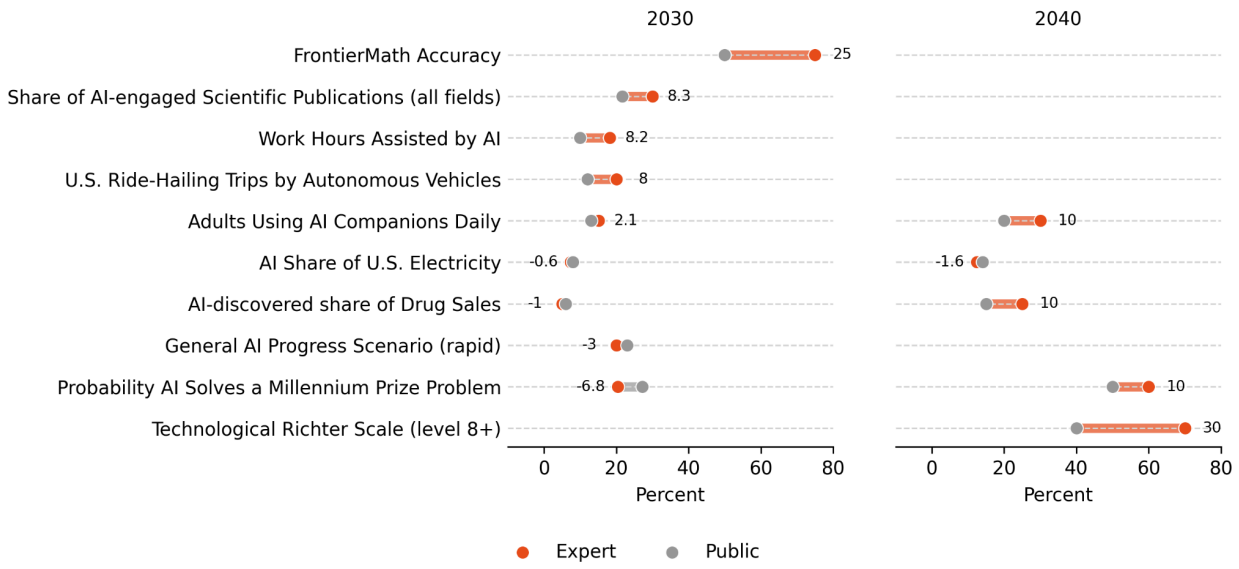


Figure: Differences between the expert and public median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of each group's 50th percentile forecasts. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

Fifth, **highly accurate forecasters (superforecasters) and experts are largely aligned, with superforecasters expecting slightly less progress overall**. Differences across expert subgroups—those specializing in computer science or economics, those working in industry, and those working in the policy and think tank space—are small and rarely statistically significant. We summarize some of the differences in aggregate forecasts in the two figures below. The first compares experts and superforecasters, while the second groups by category of expertise.

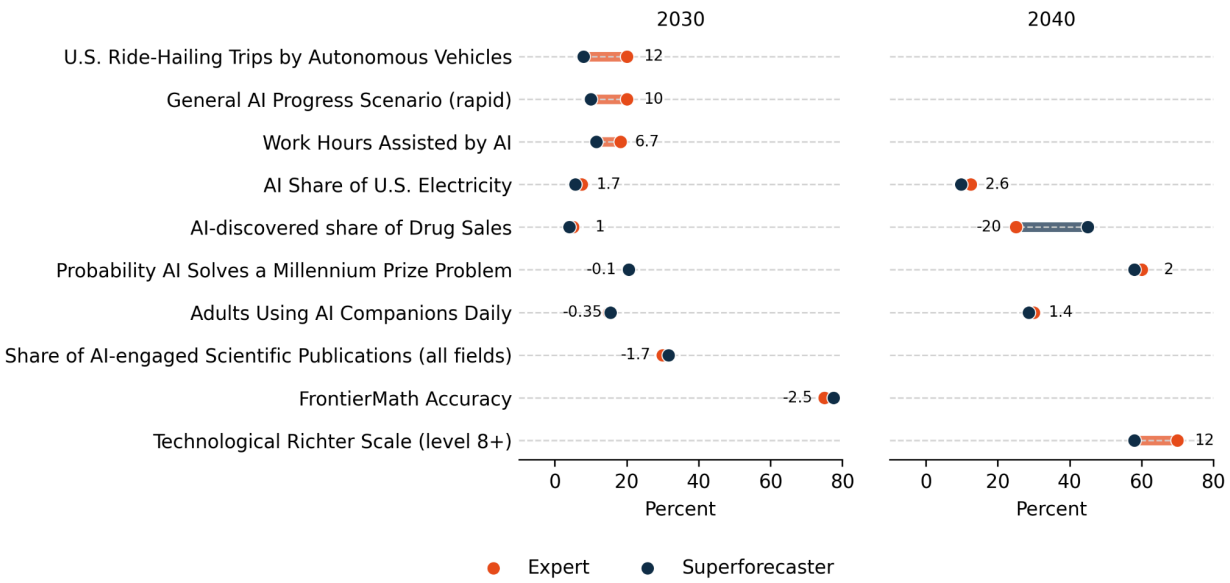


Figure: Differences between the expert and superforecaster median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of each group's 50th percentile forecasts. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

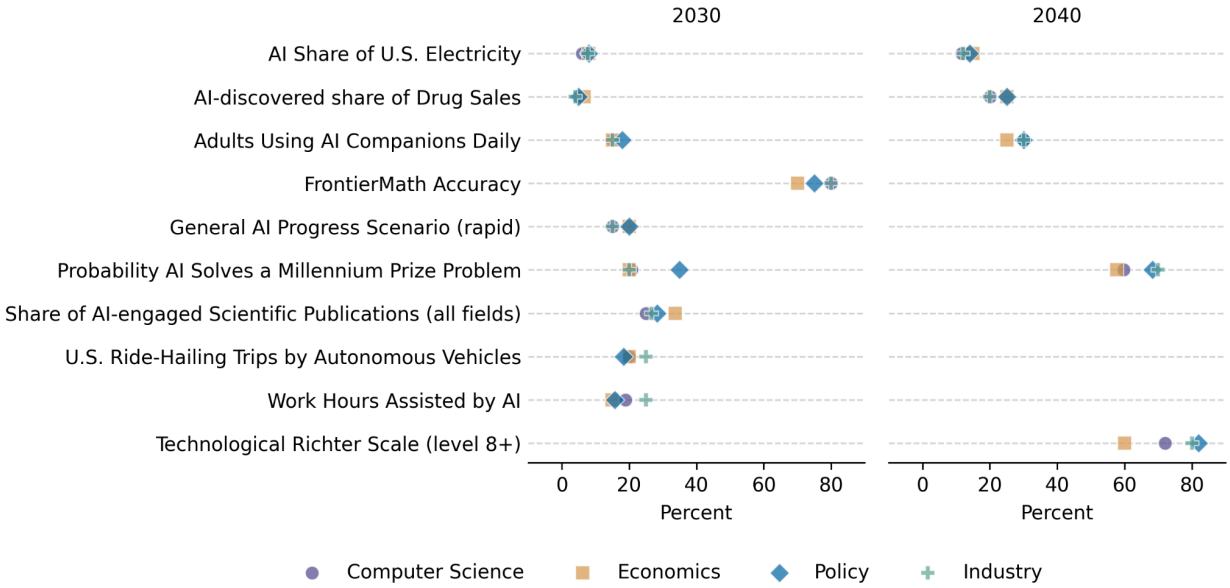


Figure: Expert category median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of 50th percentile forecasts for each category. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

Introduction

As artificial Intelligence (AI) reshapes culture,[17] science,[18] labor markets,[19] and the aggregate economy,[20] experts debate its value, risks, and how quickly it will integrate into everyday life. Leaders of AI companies forecast transformative AI systems that cure all diseases,[21] replace whole classes of jobs,[22] and supercharge GDP growth by the 2030s.[23] Skeptics see small gains at best, with AI's impact amounting to little more than a modest boost in productivity—if anything at all (Acemoglu 2024).

Despite these clashing narratives, there is little work systematically mapping the full spectrum of views among computer scientists, economists, technologists in the private sector, and the public. What do these groups believe about AI's future capabilities, adoption, and effects? Why do they believe what they do, and what mechanisms support those beliefs? Prior surveys offer opinions, but rarely comprehensively quantify those opinions, hampering policy guidance and evaluation. We fill this gap with the Longitudinal Expert AI Panel (LEAP), a monthly survey tracking the probabilistic forecasts of experts, historically accurate forecasters ("superforecasters"),[24] and the public.

---

[17] While surveys show that fewer than half of Americans report using AI products (NORC 2025), nearly all (99%) actually use AI-enabled tools like navigation apps, streaming services, and social media weekly (Maese 2025); this gap reveals that AI has become ubiquitous in some applications, but is sometimes invisible to users.

[18] AI is already accelerating scientific discoveries across a wide range of disciplines such as medicine and materials science (Dai et al. 2025; Kay 2025; Russell et al. 2023; Stanford Medicine News Center 2025; Sundermier 2024). While some of these claims may be overblown or overstated to encourage media attention, AI is certainly affecting many aspects of science and research.

[19] Despite public figures predicting extensive job destruction—for example Jamie Dimon speaking at Fortune's Most Powerful Women summit: "[AI] will eliminate jobs. People should stop sticking their head in the sand" (Gerut 2025)—overall employment effects of AI remain small (Chandar 2025; Gimbel et al. 2025; Eckhardt and Goldschlag 2025), the impact of rapid technological change on jobs in the short run is ambiguous, and in the longer run is neutral under standard economic models (e.g., Aghion, Jones, and Jones 2018; Agrawal, Gans, and Goldfarb 2019). Some recent evidence suggests that AI may be contributing to significant declines in entry-level hiring for workers in especially AI-exposed occupations (Brynjolfsson, Chandar, and Chen 2025), with those workers experiencing increased wages. This could suggest a negative supply shock, rather than a negative demand shock from substitution to AI. The ambiguity of technological change arises because automating some human processes often augments others (Agrawal, Gans, and Goldfarb 2023).

[20] Jason Furman, in an interview with Ross Douthat, estimates that 92% of the increase in demand in the U.S. in the first half of 2025 comes from categories related to AI investment and services (information processing equipment and software). Accounting for equilibrium effects, Furman posits that roughly half of GDP growth is from the AI boom (Douthat 2025). This view is echoed by Karen Dynan, an economics professor at Harvard University, who argues, "in a mechanical sense it's fair to say that AI has been the main driver of U.S. GDP growth this year" (Curran and Niquette 2025).

[21] Demis Hassabis: "I think one day maybe we can cure all disease with the help of AI." (Hassabis 2025).

[22] "There Will Be Very Hard Parts like Whole Classes of Jobs Going Away" (Altman 2025).

[23] "[A] dream scenario—perhaps a goal to aim for—would be 20% annual GDP growth rate in the developing world" (Amodei 2024).

[24] Forecasters are denoted "superforecasters" if they (1) were in the top 2% of the accuracy distribution in a given year of the IARPA ACE tournament (IARPA ACE Program n.d.; Mellers et al. 2014) or (2) they were a highly accurate performer on Good Judgment Open, an online continuous geopolitical forecasting tournament. Good Judgment Inc., which runs Good Judgment Open, then adds these top forecasters to

Policymakers and stakeholders routinely consult experts to ground their decision-making in coherent, informed beliefs, especially when faced with new technologies and high levels of uncertainty. LEAP is designed to facilitate this process by cutting through the anecdote and speculation that currently dominates AI discourse. We specifically target prominent experts whom policymakers would be most inclined to consult regarding the progression of AI capabilities and its technological impact. LEAP expert invitees (and participants) include top-cited AI and ML scientists, prominent economists, key technical staff at frontier AI companies, and influential policy experts from a broad range of NGOs. Our goal is to provide a clear summary of expert views, analyzed both across and within these key groups. To address concerns that our respondents might be a biased or selective sample, we reweight our data to account for different response propensities within our expert sampling frame, providing a representative summary of our expert frame's beliefs.[25] The Panel Construction section contains more details. This reweighting leaves our headline conclusions unchanged (see Sensitivity of Results to Reweighting).

We evaluate LEAP respondents' forecasts against detailed, pre-specified resolution criteria—avoiding the problem of ambiguous resolution criteria that introduce noise in forecasts by permitting both different interpretations of a question by forecasters, and disagreement about how a question is later resolved. In contrast, clear resolution criteria put the debate in common terms and enable policymakers to understand the range of opinion. LEAP's resolution criteria are also used to measure accuracy—and tie respondents' compensation to this accuracy—to encourage participants to provide high-quality forecasts. Lastly, we can use this accuracy measure to identify the most accurate forecasters and explore how their views compare to other participants.[26]

LEAP captures not just forecasts, but participants' rationales: 1.7 million words of detailed explanations across the first three survey waves of detailed explanations. Over 600,000 of these words come from our expert and superforecaster samples, with the remainder coming from public respondents. We use this data to identify key sources of disagreement and to analyze why participants express significant uncertainty about the future effects of AI.

Since we launched LEAP in June 2025, we have completed three survey waves focused on (1) high-level predictions about AI progress; (2) the application of AI to scientific discovery; and (3)

---

the "superforecaster" pool. Most superforecasters come from the first selection criteria. Mellers et al. (2015) finds persistent performance of these superforecasters across several years of geopolitical forecasting.

[25] "Representative" is a contested, and sometimes fraught concept in social science (Chasalow and Levy 2021). Here, we adopt a narrow definition for our expert sample. We specify an expert sampling frame that we believe closely tracks groups of experts that policymakers are most inclined to consult about AI. We then apply standard survey reweighting methods to combat nonresponse bias, reweighting our respondent sample to match the observable characteristics in our initial sampling frame. For our public sample, we use known characteristics of the U.S. population as our reweighting targets. See Reweighting for more details and a discussion of limitations of this approach.

[26] The Monthly Surveys and Forecasting Questions section contain further detail on resolution criteria.

widespread adoption and societal impact. In this paper, we share results from these first three waves along with a detailed methodological description of the project as a whole.

In what follows, we first present aggregate forecasts from experts. Aggregation exploits the "wisdom of the crowd" phenomenon, in which the accuracy of aggregated predictions exceeds the accuracy of a large majority of their constituent parts. This practice is supported by work in numerous fields such as prediction markets (Bassamboo, Cui, and Moreno 2015), political forecasting (Sjöberg 2009, Murr 2011), and more (Hueffer et al. 2023, Adjodah et al. 2021). Aggregated predictions of many individuals tend to be remarkably accurate (Da and Huang 2020; Lichtendahl Jr, Grushka-Cockayne, and Pfeifer 2013; Surowiecki 2004).[27] While aggregate forecasts are helpful, it is important to understand the extent of disagreement between experts and where individuals have more and less uncertainty in their forecasts. We discuss our approach to answering these questions below in the Uncertainty and Disagreement section.

We will continue to field new LEAP surveys every month for at least three years. We will revisit questions from previous waves beginning in year two, in order to assess how expert views are changing, both in aggregate and within individuals. First, when we report results, participants can compare their own forecasts to other participants. As forecasts resolve, participants will be able to see how the accuracy of their past forecasts compare to other participants and adjust their future forecasts accordingly. We will also explore the relationship between short-term accuracy and long-run forecasts about AI capabilities, effects, and diffusion.

In this report, we discuss LEAP in the context of prior AI forecasting surveys (Connection to Prior Work), detail the procedures we use to build our panel (Panel Construction), describe the surveys and forecasting questions (Monthly Surveys and Forecasting Questions), share preliminary results (Results), outline plans for future work (Next Steps), and conclude (Conclusion).

Connection to Prior Work

While prior work on AI progress has taken a variety of forms, we focus here on studies that, like LEAP, use surveys of the AI expert community and the general public to measure the range of opinions and expectations on the progress of AI, its adoption, and the associated societal impacts. This type of work faces five challenges. First, it is difficult to specify a comprehensive sampling frame, and nonresponse bias is challenging to avoid. Both of these hurdles can distort aggregate estimates. Second, views on AI could change over time. Third, ambiguous resolution criteria introduce noise in forecasts and limit our ability to identify the most accurate forecasters, whom policymakers may want to rely on in the future. Fourth, while we might care about

[27] As typically specified, the "wisdom of the crowd" phenomenon appeals to independent and unbiased judgments, which lead the aggregate to outperform randomly selected components of the aggregate (Davis-Stober et al. 2014). An extensive literature develops improved aggregation approaches (Baron et al. 2014; Himmelstein, Budescu, and Han 2023; Himmelstein, Budescu, and Ho 2023), but a key theme of that work is that a simple median performs quite well as an aggregation mechanism across contexts.

long-term outcomes, it is challenging to assess the credibility of long-term forecasts on short-run decision timelines. Fifth, participants often devote limited bandwidth to surveys, and, even when thoughtfully engaged, might not report their true beliefs on forecasting questions, instead using the survey to express their preferences. While no survey can fully overcome these limitations, we designed LEAP to thoughtfully engage with and mitigate each of these five challenges.

Challenge 1. Narrow frames and selective response can distort estimates of AI progress

A survey can fail to represent aggregate views in two ways. First, the sampling frame might fail to capture the population of interest.[28] Second, survey respondents might systematically differ from nonrespondents. Both problems affect AI forecasting surveys.

Most surveys of AI experts that predate LEAP target participants in top AI/ML or computer science conference proceedings (Muller and Bostrom 2014; Walsh 2017; Grace et al. 2018; Zhang et al. 2021; Stein-Perlman 2022; Zhang et al. 2022; Grace et al. 2024).[29] LEAP seeks to broaden the range of opinions as well as the expertise of participants and thus defines *four* target populations of experts: computer scientists researching topics in AI, economists studying the economic impacts of AI, AI industry professionals, and AI policy professionals. LEAP is the first survey to comprehensively measure the beliefs of each of these groups on AI-related questions so that we can compare each group's forecasts and accuracy in the future. The Sampling section details our sampling procedures.

LEAP additionally includes members of the general public[30] and forecasters with a demonstrated track record of uniquely high accuracy on forecasting tasks.[31]

Publicly available demographic data on AI experts is often collected and used to stratify results in these studies, but have not historically been used to reweight results to match a target expert population (Grace et al. 2018; Zhang et al. 2022). Many studies find associations between

[28] LEAP identifies distinct target populations (experts, superforecasters, and the general public) whose views we wish to study. We carefully construct sampling frames, or the part of the target populations that have a chance of being sampled, using multiple data sources to maximize the coverage of these populations. See Sampling for more information on how sampling frames are constructed.
[29] Top citations in AI publications have also been used to target experts (Muller and Bostrom 2014), as well as a broader search of the literature using targeted publication classification codes (O'Donovan et al. 2025).
[30] Participation from the general public in these types of surveys relies on well-established panels (McClain et al. 2025), online opinion polling platforms (Zhang and Defoe 2019), or by identifying non-experts with a demonstrated interest in AI (Walsh 2017).
[31] These individuals, labeled "superforecasters," are distinguished by their ability to sustain high-accuracy forecasts and "avoid regression to the mean" across multiple prediction instances (Mellers et al. 2015). This was demonstrated by findings from three consecutive years of geopolitical forecasting tournaments conducted by the Good Judgment Project under the U.S. Intelligence Advanced Research Projects Activity (IARPA).

collected demographics and outlook,[32] suggesting that results may be sensitive to nonresponse bias. LEAP directly addresses issues of nonresponse bias by using common survey weighting methods, discussed in greater detail in Reweighting.

Challenge 2. Views on AI can change over time

One-time surveys reflect respondent views at a given time, so they cannot track the evolution of individual or group opinions over time. As a panel survey that will continue for many years,[33] LEAP enables this type of tracking over time.

Challenge 3. Ambiguous forecasting questions complicate *ex post* evaluation and comparison between forecasters

When forecasting questions lack detailed resolution criteria, assessing accuracy and comparing forecasts is fraught. If a question is ambiguous, forecasters may have different interpretations of how the question will be resolved, so those with disparate views can each claim to have superior accuracy. In contrast, forecasting questions with prespecified resolution criteria permit evaluation of forecasting accuracy after events resolve, which proves crucial for the subsequent two challenges.

Early surveys of AI experts focused on measuring the range of opinions on AI (Muller and Bostrom 2014) or the relationship between those opinions and other forecasts (Walsh 2017) rather than defining resolvable forecasting questions. Research that followed set a goal of generating useful, more resolvable, forecasts. Grace et al. (2018) ask experts to forecast on AI progress milestones. Stein-Perlman and Grace (2022) and Grace et al. (2024) use slightly modified versions of the earlier questions. Zhang et al. (2022) modify many of the AI milestone forecasts in Grace et al. (2018). LEAP builds on these recent efforts and attempts to create clear resolution criteria for all questions, establishing a longitudinal panel of beliefs from the same set of experts to understand how opinions change over time and to better quantify uncertainty and disagreement across a wider set of experts. O'Donovan et al. (2025) and McClain et al. (2025) present forward-looking questions but do not ask respondents for forecasts; Pew asks respondents their opinions on the impact of AI over the next twenty years but does not generate measurable outcomes (McClain et al. 2025). LEAP forecasting questions use precise and specific language and are designed with measurable outcomes; survey

[32] Muller and Bostrom (2014) find experts working in theoretical AI were more likely to respond and more likely to be concerned about the negative effects of AI. Grace et al. (2018) find respondents to have less time in the field and lower citation indexes which in turn was associated with more optimistic views on the timing of human-level machine intelligence (HLMI). O'Donovan et al. (2025) find an association between views on AI safety and governance and respondent views on the inevitability of HLMI, as well as an association between their categorization of respondents as AI optimists and the inevitability of AI.

[33] A panel survey is a survey that repeatedly collects data from the same group of respondents over time, and LEAP is the first continuous panel survey of AI experts. There is one other repeated expert sample of AI experts that we have identified in the literature: Zhang et al. (2022) and Stein-Perlman and Grace (2022) do "matched panel" analysis by matching their respondents to earlier responses in Grace et al. (2018).

respondents are provided with the resolution criteria at the time of the forecast, a feature that enables us to address the remaining two challenges. Monthly Surveys and Forecasting Questions discusses this feature in greater detail.[34]

Challenge 4. Long-term forecasts are difficult to evaluate in the near-term

When forecasters disagree about the long-run, how can policymakers evaluate those forecasts on short-run timelines? One potential path forward is to evaluate the short-run accuracy of forecasters, relying on the most accurate forecasters over the short-run to better understand long-run outcomes. Resolving this challenge first requires clear resolution criteria, discussed in the preceding challenge. However, it also requires forecasts on both short- and long-run questions.

When using fixed-year format questions, earlier work used 10-, 20-, and 50-year time frames for predictions (Muller and Bostrom 2014; Grace et al. 2018; Zhang et al. 2019; Zhang et al. 2022; Stein-Perlman and Grace 2022; Grace et al. 2024). LEAP includes questions with a wide array of resolution dates, including both near- (as early as the end of 2025) and long-term resolution, allowing us to explore whether short-term accuracy correlates with longer-run beliefs.[35]

Challenge 5. Forecasting is time-intensive and cognitively taxing, and truthfulness is rarely incentivized

Expert and public participants have limited time to complete surveys, and long surveys risk dissuading potential participants from completion. Three features of LEAP counteract this challenge. First, we provide historical baselines and relevant background information for each question (see Appendix E.I. Survey Questions: Wave 1 for representative examples). Second, the survey instrument contains interactive interfaces—which integrate historical baseline data where available—to facilitate the forecasting process (see Appendix B.V. Survey instrument). Third, much like other surveys, LEAP compensates forecasters for their participation. In the first three waves, the median expert respondent spent 44 minutes on each survey, while the median member of the public and superforecaster spent 29 and 90 minutes, respectively. In contrast, the American Trends Panel (ATP) from Pew Research, a popular public opinion survey, targets ten to fifteen minutes per survey (Pew Research Center 2024).

Thoughtful engagement need not translate into forecasters reporting their true beliefs. If forecasters' rewards are not positively related to the quality of their forecasts, they lack the incentive to provide accurate forecasts. Much past work relies on traditional participation

[34] Nevertheless, the creation of unambiguous forecasting questions remains difficult. We dropped one question on the use of AI-use during K-12 instructional hours from this analysis, due to substantial misinterpretations.
[35] Some of the cited studies extrapolate forecasts backwards to obtain estimates for earlier dates, but these methods require parametric assumptions and generally do not allow for discontinuities in the time paths of forecasts.

payments that are independent of forecasting accuracy (Grace et al. 2018; Grace et al. 2024). Surveys of the general public (Zhang 2019; McClain et al. 2025) tend to provide similar recruitment incentives. In contrast, LEAP's detailed resolution criteria on timely forecasting questions allow us to incentivize participants to provide accurate forecasts by tying rewards to proper scoring rules (Brier 1950; Jose and Winkler 2009; see Appendix B.III. Scoring). Past work demonstrates that proper scoring rules yield more accurate forecasts than unincentivized forecasts (Karger et al. 2021).

Panel Construction

We discuss in this section our procedures for sampling and reweighting, as well as summarize the characteristics of our panel.

Sampling

We target prominent experts whom policymakers, business and nonprofit leaders, and other stakeholders would be most inclined to consult regarding the progression of AI capabilities and its technological impact. Our complete panel consists of experts, forecasters with a demonstrated track record of high, differential accuracy on forecasting tasks ("superforecasters"), and the members of the general public.

Expert Sampling

We target four expert communities. First, we include computer scientists researching topics in AI by including top-cited authors, stratified by age, and the authors of the top-rated papers at leading AI and ML conferences. Second, we identify leading economists, both across fields and within the subfield of economics focused on the economic effects of AI and new technology. We include top-cited authors of papers on AI and technology, members of the U.S. Economic Experts Panel (Clark Center 2025), and attendees of economics conferences on AI. Third, we include industry professionals, identified via their contributions to frontier models or employment at AI-related companies with extensive fundraising. Fourth, we identify institutions leading the discussion on AI development, policy, and impacts and invite research staff.

We sample from two other sources and sort them into one of the four communities above. First, we invited the honorees from TIME's 100 Most Influential People in AI in 2023 and 2024 (Barker Bonomo and Javed 2024). Second, we allowed invited respondents to recommend other qualified candidates for the survey, yielding 172 additional invitees. In order to filter this group, we require that an individual:

- meets the requirements of another sampling category;
- has over 1,000 academic citations; or
- has over 300 academic citations if a PhD student or postgraduate researcher.

These requirements excluded only 7 of the recommended candidates that ultimately enrolled. After exclusion, the referred group has 11.6 years of experience on average and 75% have a postgraduate degree. Like other expert sample expansions, referred contacts are not included

in frame targets, but their responses do receive positive weights through the reweighting process.

Within each community, we build our initial frame by identifying potential respondents from a number of sources, described in greater detail in Appendix A. Panel Construction and Sampling. We largely create non-probability samples composed of all respondents who meet criteria for inclusion in the frame, but we randomly sample from some sources that yielded a large frame (e.g., industry and policy professionals).

To reach sufficient respondent counts, we expand beyond our initial frame in a number of categories below–this results in our "full frame," as shown below. To correct for the consequent change in sample composition, we use our initial frame to define reweighting targets (see Reweighting for further details on reweighting and Appendix A. Panel Construction and Sampling for a detailed specification of our initial frame). While individuals identified in these late expansions do not contribute to our targets for weighting, they do receive positive weights in our results.

Expert Type	Initial Frame Count	Full Frame Count	Participant Count
Computer Science	454	719	61
Economics	391	773	66
Industry	561	1,640	57
Policy	367	881	96

Table 1: Counts by expert category in our initial frame (target population), full frame (the initial frame plus expanded sampling), and participant pool (individuals who have completed at least 1 survey)

When an individual enrolls in LEAP, we collect information on their *primary* affiliation, which we use to reassign their category. For example, a respondent might have been sampled through a well-cited computer science publication, but they currently work at a leading AI lab. In such cases, we use their current affiliation, as self-reported in the enrollment survey, as their assigned category. However, we use the initial, rather than updated, categorization to define our weighting targets, as we intend to measure the characteristics of the broader population reflected by this sampling group, rather than the identity of particular individuals. Additionally, we only have this updated affiliation information for enrollees and not experts in the broader sampling frame who do not enroll. This section will accordingly use these initial classifications. All sample statistics outside of this subsection instead use this final classification, rather than a classification based on how we sampled an individual.

Superforecaster Sampling

We include a sample of superforecasters, sourced through Good Judgment Inc., a company that maintains and adds to a list of highly accurate forecasters, many of whom were among the top-2% of forecasters by accuracy in IARPA's ACE tournament (Good Judgment Inc. n.d.).

These superforecasters have a demonstrated track record of providing the most accurate forecasts across a wide array of topics.[36] We invited 67 superforecasters through this search.

Public Sampling

We include in our sample highly-engaged participants[37] from past FRI research projects on the CloudResearch Connect platform. We initially invited approximately 2,600 individuals. To ensure representativeness across underrepresented populations in our initial sample, we also reach out to additional respondents who are either (1) over the age of 50 and identify as Republican; or (2) have a high school degree (or equivalent) as their highest level of completed education.

Reweighting

Individuals with certain viewpoints might be disproportionately likely to respond to our survey conditional on receiving an invite, skewing our results towards these viewpoints associated with a high propensity to respond to the survey. To address concerns about nonresponse bias, we use a standard approach in the public polling field, raking, to adjust aggregate statistics to be representative of the sampling frames.[38]

These adjustments do not substantially change any of our results. The section Sensitivity of Results to Reweighting shows the impact of weighting on our aggregate results. We default to reporting weighted summary statistics for any results in this paper, but any tests or statistics related to differences in distributions (Mann-Whitney U tests and Cliff's δ) are currently unweighted.

For the expert sample, we use our *initial* invite pool to generate benchmarks for reweighting. Appendix A. Panel Construction and Sampling provides more information on the composition and selection of this initial invite pool. We reweight on years of relevant experience, age,

[36] Forecasters are denoted "superforecasters" if they (1) were in the top 2% of the accuracy distribution in a given year of the IARPA ACE tournament (IARPA ACE Program n.d.; Mellers et al. 2014) or (2) they were a highly accurate performer on Good Judgment Open, an online continuous geopolitical forecasting tournament. Good Judgment Inc., which runs Good Judgment Open, then adds these top forecasters to the "superforecaster" pool. Most superforecasters come from the first selection criteria. Mellers et al. (2015) finds persistent performance of these superforecasters across several years of geopolitical forecasting.
[37] Highly-engaged participants were recruited from previous high effort projects spanning multiple weeks.
[38] Pew describes "raking," also known as iterative proportional fitting, as the most common approach to reweighting public opinion surveys, "For public opinion surveys, the most prevalent method for weighting is iterative proportional fitting, more commonly referred to as raking" (Mercer et al. 2018).

affiliation with Effective Altruism,[39] gender, continent, education, and affiliation with top AI labs.[40] We also equally weight participants based on their category of expertise.[41] Respondents from expanded sampling and referred contacts do not contribute to our frame targets for weighting, but they do receive positive weights in our results. The target populations for each reweighting category are displayed in Appendix A.VIII. Reweighting.

We do not reweight our superforecaster sample.

For the public sample, we reweight on age, gender, race/ethnicity, household income, educational attainment, and political party identification. We target U.S. population demographics.[42] After our initial public invites, we conducted two targeted recruitment waves in cells with low response counts in our initial sample: first, we recruited individuals who were over age 50 and identified as Republicans. Second, we recruited individuals with a high school degree (or equivalent) as their highest level of completed education.

We assessed our sample for differences in experience, prominence (measured by citations), ideology (measured by affiliation with Effective Altruism), and top-lab affiliation. First, there are similar levels of experience in the sampling frame and actual respondents; 50% of respondents have more than 10 years of experience in their field, contrasted with 56% of invitees. Second, there is no clear difference in academic or research prominence for the invitees versus respondents; for example, top-cited computer scientist respondents averaged 120,000 citations, similar to the 148,000 in our invited pool. Third, however, there is a clear difference in ideological affiliation; we found that 28% of respondents had ties to Effective Altruism, in contrast with 14% of invitees. But, when we downweighted this group of respondents to match the invitee proportion during reweighting, we saw no meaningful change in aggregate results. Lastly, there is also a difference in leading lab membership. 18% of invitees were employed at one of the top 20 AI labs (as defined above), but only 8% of our respondents belong to this group.

---

[39] Effective Altruism (EA) is a philosophical and social movement focused on directing resources towards improving the world. It is not a monolithic group. Many effective altruists strongly disagree about how best to direct resources towards improving the world, the philosophical framework that determines what an 'improvement' means, the risks one should take to improve the world, and the focus one should place on the short- and long-run in pursuit of improving the world. We consider someone EA-affiliated if they or their employer have or have had funding ties to EA, they publicly endorse EA, or they self-identify as EA-affiliated or EA-adjacent. Further, if their work focuses on EA, AI safety, global catastrophic risks, or existential risks, we consider them to be EA-affiliated. These criteria are permissive and favor overinclusion, as we chose to measure an upper bound of EA affiliation. Nevertheless, the proportion of our frame tagged as having EA affiliation remains relatively small, at 14.3%.

[40] We define "top AI labs" as the 10 unique labs which provide the 20 most computationally intensive models (in terms of training FLOP) on Epoch AI's Data on AI Models table (Epoch AI 2024b).

[41] As seen in the table below, we split our expert population into four categories of expertise: Computer Science, Economics, Industry and Policy. For the purposes of reweighting, we equally weight these categories. Our initial frame slightly overrepresented Computer Science and Industry professionals and underrepresented Economics and Policy professionals.

[42] We use the IPUMS USA combined Census and American Community Survey (ACS) data to derive population targets for all variables except party identification (Ruggles et al. 2025). For party identification, we use data from Pew Research Center (Nadeem 2024).

Table 2 presents a comparison of several key demographic and professional features between the target sampling frame and the Wave 1 respondent population before and after weighting. The table illustrates how our reweighting process adjusted the respondent sample to more closely reflect the target population's composition.

Characteristic	Target	Responded (unweighted)	Responded (weighted)
Average age	40 years	37 years	40 years
Affiliated with Effective Altruism	14%	28%	12%
Men	77%	78%	78%
Lives in North America	60%	67%	62%
Average years of experience	14 years	12 years	13 years
Has a postgraduate degree	75%	77%	79%
Affiliated with "Industry" category	25%[43]	23%	25%
Affiliated with top AI lab	18%	8%	16%

Table 2: Comparisons between characteristics of the sampling frame population and the population of unweighted/weighted Wave 1 respondents.

Respondent Characteristics

339 experts provided complete forecasts for at least one survey. We report respondent counts by domain of expertise in Table 3.

Expert type	Number of respondents
Computer scientists	76
AI Industry employees	76
Economists	68
AI policy experts	119

[43] Note that the actual proportion of participants invited from Industry was 36%. For the purposes of reweighting, each respondent category of expertise was weighted equally as discussed above.

Total	**339**

Table 3: Number of respondents per expert category. Respondents completed at least one of the first three survey waves.

To better characterize these experts, our respondent sample includes:

- *Top computer scientists:* 41 of our 76 computer science experts (54%) are professors, and 30 of these 41 (73%) are from top-20 institutions (Berger 2025). Twenty-three (30%) had top-rated (top-40 or better) papers at NeurIPS or ICLR in recent years, and eight others (11%) are PhD students or postdocs who are highly cited according to our criteria.[44] Ten (13%) are among the 200 top-cited authors in AI (OpenAlex n.d.). This category also includes researchers at academic and non-academic research institutions. Our CS respondents have a median of 7,100 citations (for the 95% of panelists for whom data is available).
- *AI industry experts:* 20 of our 76 industry respondents (26%) work for one of five leading AI companies: OpenAI, Anthropic, Google DeepMind, Meta, and Nvidia. Twenty-one of the remaining industry respondents (28% of the total) work for either a top AI company (top-20 model providers, by training compute, as measured by EpochAI 2024b), were identified as contributors to top-15 LLMs according to training compute or performance on Chatbot Arena in our sampling procedure (Epoch AI 2024b; LMArena 2024), or work for one of the top 30 AI-related companies, as measured by total funds raised (Crunchbase 2025). The remaining respondents were recategorized from our CS literature sampling pools, referral sampling, or other categories. Our industry respondents have a median of 9,100 citations (for the 59% of panelists for whom data is available).
- *Top AI economists:* 54 of our 68 economist respondents (79%) are professors, and 30 (44%) are from top-50 economics institutions (RePEc 2025).[45] Our economist respondents have a median of 2,200 citations (for the 96% of panelists for whom data is available).
- *Policy and think tank group:* Of our 119 AI policy respondents, 75 (60%) work for the following most-represented organizations (unordered): Brookings, RAND, Epoch AI, Federation of American Scientists, Center for Security and Emerging Technology, AI Now, Carnegie Endowment, Foundation for American Innovation, Stanford's Institute for Human-Centered Artificial Intelligence and related groups, GovAI, Institute for AI Policy and Strategy, Future of Life Institute, Institute for Law & AI, Center for a New American Security, Data & Society Research Institute, Abundance Institute, and the Centre for International Governance Innovation.
- *TIME 100:* Our panel includes 12 honorees from TIME's 100 Most Influential People in AI in 2023 and 2024 (Bajekal 2023; Barker Bonomo and Javed 2024). TIME 100 honorees are categorized by their expertise and distributed among the above categories.

[44] These respondents are either 200 top-cited according to OpenAlex, or part of our age-stratified CS author list.

[45] Research rankings for economics schools and journals, based on publication records. Authors at an institution hosting one of the top-50 economics departments according to RePEc are considered.

In addition to domain experts, respondents included:

- 60 highly accurate forecasters ("superforecasters"), based on performance in prior geopolitical forecasting tournaments.
- 1,400 members of the public, largely consisting of especially engaged participants in previous research, reweighted to be nationally representative of the U.S.

We report the count of respondents in each wave in Figure 1 below. The drop in expert completions from Wave 1 to Wave 2 was much larger (23%) than from Wave 2 to 3 (4%).[46]

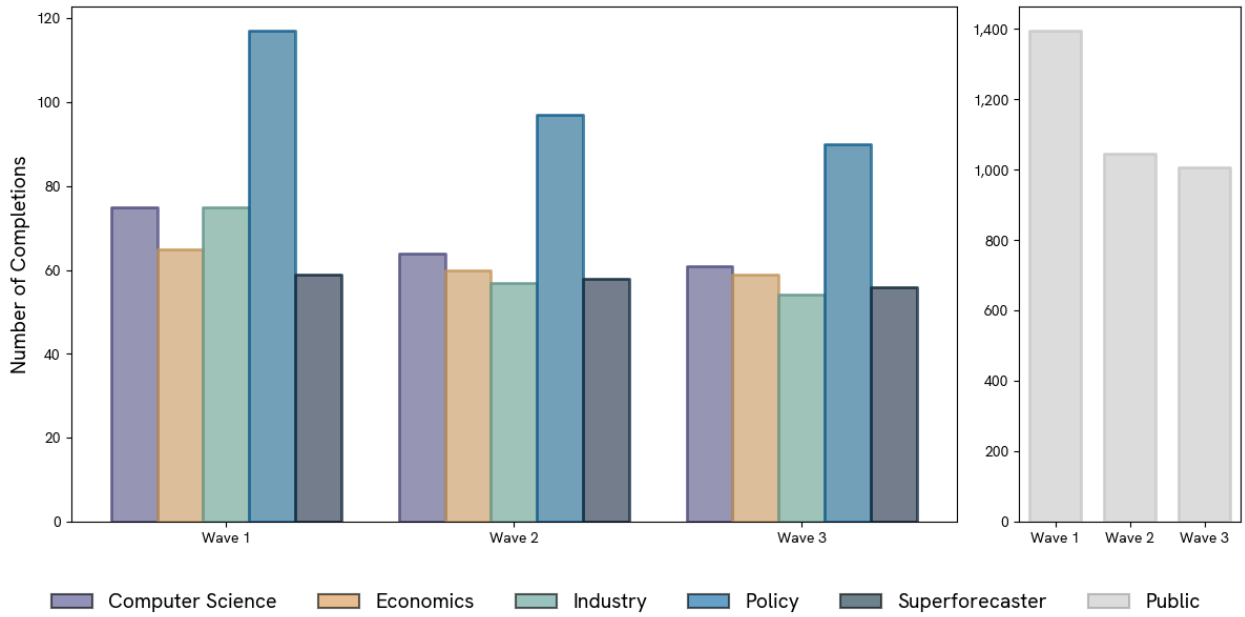


Figure 1. The number of survey completions per participant group, across the first three survey waves of LEAP.

In the first three waves, the median expert respondent spent 44 minutes on each survey,[47] while the median member of the public and superforecaster spent 29 and 90 minutes, respectively. Respondents provided 1.7 million words of rationales across three waves. Over 600,000 of these words come from our superforecaster and expert samples, with the remainder from public respondents.

Monthly Surveys and Forecasting Questions

We conduct surveys approximately every month, each consisting of 5-6 forecasting questions. We expect each survey to take experts approximately 30-40 minutes to complete, and

[46] The percentage decrease in completions from Wave 2 to 3 per category were as follows: Economics: 1.7%, Superforecasters: 3.5%, Public: 3.8%, Computer Science: 4.7%, Industry: 5.3%, Policy: 7.2%.
[47] The survey platform records the active time spent on the survey window on respondents' devices. If they leave the page (even to do additional research relevant to the survey), the timer is paused. Hence, this measure likely serves as a lower bound for time spent on the survey.

respondents are informed of this time estimate. Respondents receive a standardized payment for each survey they complete.[48] In addition to quantitative forecasts, we collect rationales for each forecasting question, in the form of plain text. We use these rationale data to understand the underlying reasons for forecaster responses.

LEAP includes forecasting questions across five categories:

1. *AI inputs:* drivers of AI progress like investment, electricity consumption, and other AI R&D inputs, such as talent.
2. *AI capabilities:* measures of AI progress such as benchmark performance on difficult and consequential tasks.
3. *AI adoption:* drivers of AI impact such as the prevalence of AI applications and the intensity of AI use in economically meaningful and high-stakes contexts.
4. *AI impacts:* downstream societal effects, such as AI incidents and labor market disruption.
5. *AI scenarios:* bundles of outcomes that represent different trajectories for the technology. See an example from Wave 1 here (General AI Progress).

We source forecasting questions and topics from academic papers, technical and policy reports, prediction platforms, public writings by leading AI figures, our past research output, our academic advisory board, and suggestions from survey respondents. We then create resolvable forecasting questions from these various input sources. You can view example questions with associated resolution criteria in Appendix E. Survey Questions. The resolution criteria are intended to reduce noise in forecast collection by minimizing the space for different question interpretations among forecasters, as well as permit accuracy assessment.

Second, we ask various types of forecasting questions:

1. *Probabilistic:* We ask participants to assign a probability to a binary or discrete event.
   For example, “Will AI solve or substantially assist in solving a Millennium Prize Problem in mathematics by the following resolution dates?” (Wave 2)[49]
2. *Quantile:* We ask participants to forecast quantiles of a continuous outcome (typically the 25th, 50th, and 75th percentiles).
   For example, “What will be the highest percentage accuracy achieved by an AI model on FrontierMath, by the following resolution dates?” (Wave 1)[50]
3. *Point Estimate:* We ask participants for a point estimate of a continuous outcome.
   For example, “By the end of 2030, what percent of LEAP expert panelists will agree that each of the following is a serious cognitive limitation of state-of-the-art AI systems?” (Wave 2; list omitted for brevity but available in Appendix E.II. Survey Questions: Wave 2).

[48] We provide expert participants with $2,000 per each year of full participation (prorated for the number of surveys they complete, with an expectation that we will complete 12 surveys in a typical year). In other words, experts receive $166.67 per survey completed. We provide superforecasters with $1,000, prorated similarly, or $83.33 per survey completed. We pay public participants in line with CloudResearch platform norms, or $8 per survey completed ($13.71 per hour).
[49] Respondents are prompted to give their forecasts in the form of probabilities.
[50] Respondents are prompted to give their forecasts in the form of 25th, 50th, and 75th percentiles.

Scoring first requires us to resolve the values being forecasted. Then, these scoring rules are used to provide accuracy prizes to incentivize truthful reporting. Respondents are provided with detailed resolution criteria and relevant historical baseline data in order to inform their forecasts. We resolve questions using either ground truth data or data generated from LEAP itself. We discuss these methods in greater detail in Appendix B. Monthly Surveys and Forecasting Questions.

We score rationales with a combination of human and LLM judges, and provide prizes to the highest quality rationales in each survey wave. We do not publicly share our scoring criteria to prevent gaming by participants.

Results

Uncertainty and Disagreement

On many questions, we ask respondents to express their uncertainty in the form of quantile forecasts. We describe quantile forecasts in greater detail in Monthly Surveys and Forecasting Questions. This approach allows us to quantify two sources of variation: *between-expert disagreement* (i.e., how much experts disagree with each other) and *within-expert uncertainty* (i.e., how wide each expert's predicted range is). To measure total variation, we generate a "pooled" distribution of respondent beliefs, representing the full variation in expert views.[51] We discuss this procedure in greater detail in Appendix C. Pooled Distribution Estimation. We plot these pooled distributions for several questions in Figure 4, and we then decompose the variation in expert views into the proportion due to within-forecaster uncertainty and the component due to between-forecaster disagreement. For example, if we look at the share of work hours participants forecast to be assisted by generative AI, the variance of the distribution grows over time: the standard deviation is 5.7% in 2025, 13% in 2027, and 22% in 2030, and 65% to 53% of the variation is explained by within-forecaster uncertainty across the three time horizons.

While this approach to understanding uncertainty and disagreement is used across questions and groups of forecasters through the rest of the paper, we outline it with expert responses to one example question here. By 2030, the median expert predicts that 18% of all U.S. work hours will be assisted by generative AI. We will discuss this finding in more detail below. The 18% median forecast should be interpreted alongside information about variation in beliefs. To capture the full extent of expert uncertainty and disagreement, we construct an aggregate distribution of expert beliefs by combining each expert's distributional forecasts on the question into a pooled (mixture) distribution. The 25th–75th percentile range of this pooled distribution is (7.3%, 34.6%). The ranges represent the total variation of beliefs in the expert pool. We use the

[51] The pooled distribution is not necessarily the optimal way to aggregate forecasts in terms of forecasting accuracy. For example, Ranjan and Gneiting (2010) show that combining well-calibrated forecasts in this fashion yields forecasts that are miscalibrated. Similar results are reported by Lichtendahl et al (2013).

law of total variance to decompose the variance of this pooled distribution into between-forecaster disagreement and within-forecaster uncertainty. For example, 53% of the variation in forecasts on this question are due to within-forecaster uncertainty and 47% are due to between-forecaster disagreement. Appendix C. Pooled Distribution Estimation contains more details on the method we use for this analysis.

Because fitting a distribution to a respondent's forecasts relies on a parametric assumption, especially about the tails of individual distributions, we also present two other measures of variation. First, we present a basic measure of the amount of between-forecaster disagreement in the forecast: the interquartile range of experts' 50th percentile forecasts was 9%–30%, indicating substantial variation among experts in their predictions of the median outcome. In other words, a quarter of all experts believe that 9% of work hours (or less) will be assisted by generative AI, and a quarter believe it is likely to be above 30%.

Second, we present a measure of individuals' uncertainty based on the quantile forecasts we elicited. In addition to the median, we elicited 25th and 75th percentile quantiles from each forecaster on this question. The median 25th percentile forecast across all experts was 9% and the median 75th percentile forecast was 28%. This 9%–28% range cannot be interpreted as a typical confidence interval, but it can indicate the degree of uncertainty forecasters had about the outcome of interest.[52]

In the footnotes for each point estimate, we report three measures of uncertainty as follows: we first summarize the total variation, pooling all expert beliefs and discussing the variation in that distribution. We then report the variation in median forecasts. If we only collected a central estimate, we do not report additional statistics. If we did collect other quantile forecasts, we report the median forecasts across experts of those quantiles. For example, on this example question, we would report the median 25th (and 75th) percentile forecasts for each quantity of interest.

Key Insights

We draw five insights from the results for Waves 1, 2, and 3. First, most experts expect sizable near-term societal effects from AI. Second, substantial disagreement underlies these forecasts. Third, the median expert expects much less progress than prominent leaders of frontier AI labs. Fourth, experts anticipate faster progress than the public on most outcomes. Fifth, highly accurate forecasters (superforecasters) and experts are largely aligned, with superforecasters expecting slightly less progress overall. Differences across expert subgroups (CS, economics, industry, policy) are small and rarely statistically significant, and reweighting our expert sample to match a carefully constructed expert sampling frame leaves our headline conclusions unchanged.

[52] In future work, we will summarize forecaster-level gaps between 25th and 75th percentile forecasts.

1. Experts expect sizable societal effects from AI by 2040.

In particular, the median expert expects substantial impacts on the ability of AI systems to solve difficult math problems, the use of AI for companionship and work, electricity usage from AI, and investment in AI. Even the lower end of the expert belief distribution still implies substantial impacts of AI:

- *Work:* The median expert forecast is that 18% of work hours will be assisted by generative AI in 2030,[53] up from approximately 4.1% in November 2024 (Bick et al. 2025), over a 4x increase.[54] The bottom quartile of experts give a forecast of 9%, while the top quartile gives a forecast of 30%. The median expert gives a 25% chance the value is 9% or lower (still over a 2x increase), and a 25% chance it exceeds 28%.[55]
- *Private AI investment:* The median expert predicts that Our World in Data will report $260 billion of global private AI investment by 2030, up from the $130 billion baseline for the series in 2024.[56] The median expert gives a 25% chance that investment will be at or below $175 billion, over a third higher than the baseline value, and another 25% chance that investment matches or exceeds $400 billion, just over 3x larger than the baseline level.
- *Electricity usage:* The median expert predicts that 7% of U.S. electricity consumption will be used for training and deploying AI systems in 2030, and close to double that (12%) in 2040. For context, 7% is 1.5x today's entire data-center load, 13% is all of Texas' electricity use, 23% is almost all of the industrial sector's electricity use, and 40% accounts for all residential electricity use. Even experts expecting less electricity consumption give substantial median forecasts: the bottom quartile of experts still predict values of 5% in 2030 and 8% in 2040.
- *Math research:* 23% of experts predict that the FrontierMath benchmark will be saturated by the end of 2030,[57] meaning that AI can autonomously solve a set of math problems that resemble those a math PhD student might spend several days completing. The

[53] *Pooled distribution*: IQR (7.3%–34.6%); variance decomposition: 47% between–forecaster disagreement, 53% within–forecaster uncertainty. *Raw data*: IQR on the 50th percentile was (9%–30%); median 25th and 75th percentile forecasts were 9% and 28% respectively. Uncertainty and disagreement metrics on other claims made in this list can be found in the Monthly Reports at https://leap.forecastingresearch.org/reports/.

[54] Respondents were shown a historical baseline value of 2%, based on an earlier version of the cited paper. The most recent draft gives a range of 1.6% to 6.6%. We select the midpoint, 4.1%, as the historical baseline value.

[55] Sentences of the form, "The median expert gives an X% chance," report the median of experts' Xth percentile forecasts.

[56] Regarding their private AI investment indicator, Our World in Data (2025) notes: 1. "The data likely underestimates total global AI investment, as it only captures certain types of private equity transactions, excluding other significant channels and categories of AI-related spending;" 2. "The source does not fully disclose its methodology and what's included or excluded. This means it may not fully capture important areas of AI investment, such as those from publicly traded companies, corporate internal R&D, government funding, public sector initiatives, data center infrastructure, hardware production, semiconductor manufacturing, and expenses for research and talent." More details on what is likely excluded can be found at Our World in Data (2025).

[57] This estimate of 23% reflects the fraction of experts whose median forecast is that AI systems will achieve performance of at least 90% (which we call saturation) on Tiers 1-3 of FrontierMath. We take the average of the proportions calculated under weak and strict inequality.

bottom quarter of experts expect 60% or less of these problems to be solved in the same timeframe, substantially more than the 19% baseline at the time of the survey. By 2040, experts predict it is more likely than not (60%) that AI will substantially assist in solving a Millennium Prize Problem, a set of problems comprising some of the most difficult unsolved mathematical problems.

- *Companionship:* The median expert predicts that by 2030, 15% of adults will self-report using AI for companionship, emotional support, social interaction, or simulated relationships at least once daily, up from 6% today. By 2040, that number doubles to 30% of adults.

To assess the broader scope of AI's impacts, we asked experts to assess "slow" versus "moderate" or "fast" scenarios for AI progress, and how AI will compare to other historically significant developments such as the internet, electricity, and the Industrial Revolution. We found:

- *Speed of AI progress:* By 2030, the average expert thinks that 23%[58] of LEAP panelists will say the state of AI most closely mirrors a "rapid" progress scenario, which we described as: AI writes Pulitzer Prize-worthy novels, collapses years-long research into days and weeks, outcompetes any human software engineer, and independently develops new cures for cancer.[59] Conversely, the average expert believes that 28% of panelists will indicate that reality is closest to a slow-progress scenario, in which AI is a widely useful assisting technology but falls short of transformative impact.
- *Societal impact:* By 2040, the median expert predicts that the impact of AI will be comparable to a "technology of the century," akin to electricity or automobiles. Experts also give a 32% chance that AI will be at least as impactful as a "technology of the millennium," such as the printing press or the Industrial Revolution and a 16% chance the AI is equally or less impactful than a "technology of the year" like the VCR.[60]

The median expert predicts 2%[61] growth in white-collar jobs between January 2025 and December 2030. This is significantly slower than a recent linear trend, which would predict 6.8% growth. However, we did not collect forecasts on the causal effect of AI on white collar employment.[62] While some experts expect AI to cause white collar job loss (See Occupational Employment Index), this question does not allow for a clear understanding of that causal relationship.

We summarize the expert forecasts for these various indicators in Figures 2 and 3.

[58] *Raw data*: IQR on the 50th percentile was (10%–30%).
[59] See Appendix E.I. Survey Questions: Wave 1 for background information. We ask participants what proportion of LEAP panelists will choose "slow progress," "moderate progress," or "rapid progress" as best matching the general level of AI progress.
[60] See Appendix E.I. Survey Questions: Wave 1 for background information
[61] *Raw data*: IQR on the 50th percentile was (-4%–5%)
[62] We have a complementary survey in the field exploring these topics which we plan to release results from in early 2026.

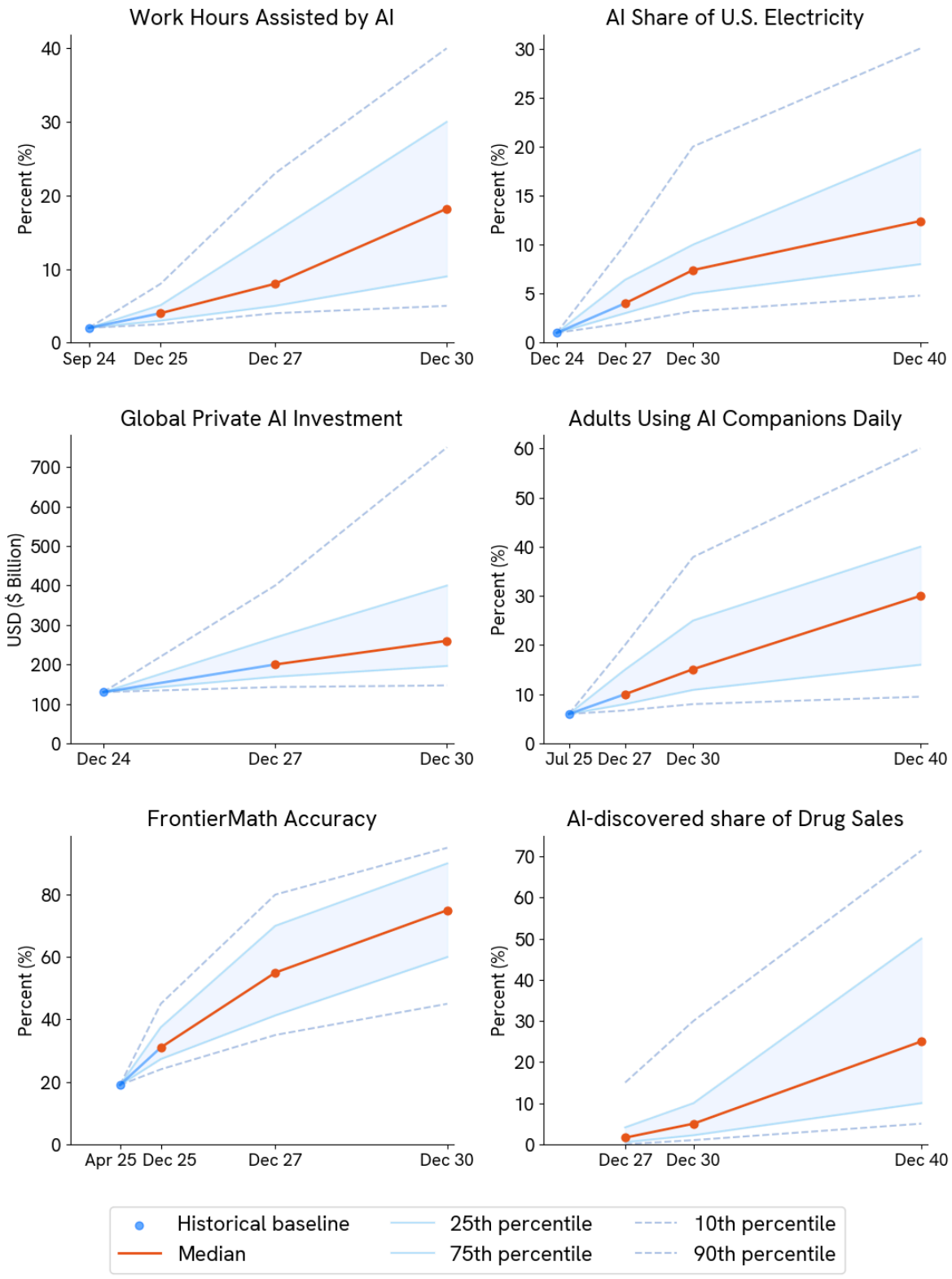


Figure 2: Median expert forecasts for various questions. We display the 10th, 25th, 50th, 75th, and 90th percentiles of the median forecasts given by experts at each date. For example, if 25% of experts give a median forecast of 5% or less, the 25th percentile series in the graph will lie at 5%; these series are *not* confidence intervals. Where available, we include a historical baseline.

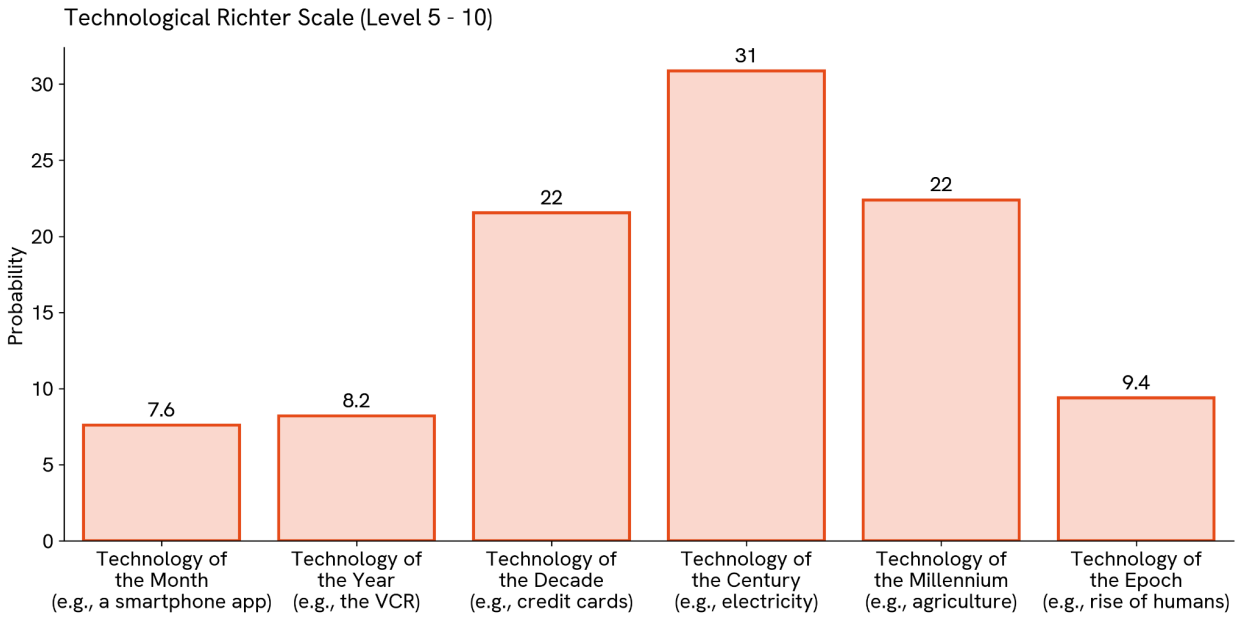


Figure 3: Average expert forecasts on the Technological Richter Scale question.

Experts lend insight into the aggregate results above with their rationales. For example, when considering the societal impact of AI (as defined by a technological Richter scale),[63] one expert is well aligned with the median in that they assign the greatest probability to AI being the Technology of the Century, but they also argue:

> ...if progress can continue at its present rate, Technology of the Millennium is a possibility. Given the level of investment and the scramble for talent, Technology of the Decade is assured. We are on a long runway stretching back 50 years and have finally achieved liftoff… AI could conceivably rival the printing press at giving [the] everyman a level of intelligence where it once provided [the] everyman with information. The industrial revolution greatly increased the material productivity of society; AI could provide the same boost for both material and service products by trading electrical energy for intellect.

Echoing that sentiment, an expert who forecasted a slightly above-the-median impact writes, “While there are parallel examples in the rise of agriculture and industrial production, particularly in terms of general-purpose innovation (steam, fossil fuels, electricity, etc.), AI is unique because it both augments human intelligence and will eventually surpass it.”

2. Experts disagree and express substantial uncertainty about the trajectory of AI.

While the median expert predicts substantial AI progress, and a sizable fraction of experts predict fast progress, experts disagree widely. Notably, the top quartile of experts give a median forecast that 50% of newly approved drug sales in the U.S. in 2040 will be from AI-discovered

[63] See Appendix E.I. Survey Questions: Wave 1, Question 5, Technological Richter Scale for details.

drugs, compared to a median forecast of just 10% for the bottom quartile of experts.[64] Further, the top quartile of experts gives a forecast of at least 81% that AI will solve or substantially assist in solving a solution to a Millennium Prize Problem by 2040, compared to a forecast of just 30% from the bottom quartile of experts.[65] We use our pooled distributions to express the relative importance of within-forecaster uncertainty and between-forecaster disagreement. We find that, across all forecasting questions that allow forecasters to express their uncertainty, within-forecaster uncertainty explains 49% of the total variation in forecasts, compared to the 51% explained by between-forecaster disagreement.

In Figure 4 below, we plot the pooled distributions for expert forecasts on the share of work hours assisted by generative AI and FrontierMath scores by the ends of 2025, 2027, and 2030.

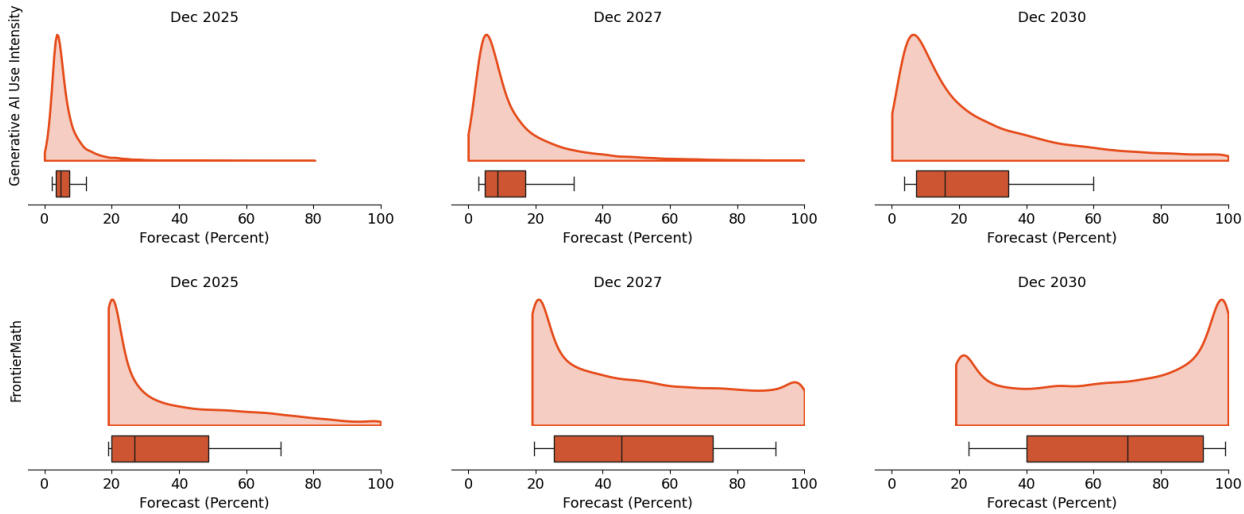


Figure 4: Pooled distributions for expert forecasts on *Work Hours Assisted by Generative AI* (top panels) and *FrontierMath* scores (bottom panels). These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. See Uncertainty and Disagreement for details.

We report summary statistics for the expert pooled distributions for select questions in Table 4 below. You can find tables for all other relevant questions on the LEAP website.[66]

[64] The median forecast for this question was 25%. *Pooled distribution*: IQR (8.4%–53.8%); variance decomposition: 52% between–forecaster disagreement, 48% within–forecaster uncertainty. *Raw data*: IQR on the 50th percentile was (10%–50%); median 25th and 75th percentile forecasts were 10% and 43% respectively.
[65] The median forecast for this question was 60%.
[66] See https://leap.forecastingresearch.org/reports/ to access these tables.

Question	Resolution Year	25th Pctl.	50th Pctl.	75th Pctl.	Within Share	Between Share
AV Trips	2027	2.7	7.1	20.1	0.53	0.47
	2030	6.9	19.2	46.2	0.53	0.47
Drug Discovery	2027	1.1	2.5	6.8	0.41	0.59
	2030	2.4	6	15.7	0.41	0.59
	2040	8.4	22.8	53.8	0.48	0.52

Table 4: Summary of expert pooled distributions for U.S. Ride-Hailing Trips by Autonomous Vehicles (AV Trips) and AI-discovered share of Drug Sales (Drug Discovery). "Within Share" and "Between Share" denote the proportion of variance from within-forecaster uncertainty and between-forecaster disagreement, respectively.

The degree of within-forecaster uncertainty revealed above is also well-represented in the rationales. As one expert writes, "Predicting when such breakthroughs come is notoriously difficult, hence broad confidence intervals," reflecting a common sentiment. Another expresses a more expansive view:

> My forecasts reflect the ambivalence I feel between competing narratives—one being that AI's potential to change the course of human history has been overhyped, and that we'll run out of decent training data, and so on; the other being that we're in a delusional, pre-Copernican state where we humans still cling to a 'we're the center of the universe' notion that intelligence is something unique to our species, or at least has to be rooted in biological entities, even as it becomes blindingly self-evident that this is not true.

The sharp between-forecaster disagreement documented by the forecasts is also revealed in the rationales, particularly through their juxtaposition. For example, when considering pace-of-AI-progress scenarios, one expert writes, "ChatGPT was first publicly released in late 2022. I don't believe what we witnessed over these past 2.5 years would justify expecting rapid progress over the next 5," while another offers, "I think the last three years of progress have been qualitatively immense, so the next five years seem like they could lead to highly autonomous systems capable of very impressive things." In other instances, the contrast is starker: "I believe that there is wide consensus on the rapid evolution of AI," writes one expert. "Rapid progress scenario is unhinged," writes another.

3. The median expert expects significantly less AI progress than leaders of frontier AI companies.

Leaders of frontier AI companies have made aggressive predictions about AI progress. Dario Amodei, co-founder and CEO of Anthropic, predicts:

- January 2025: "By 2026 or 2027, we will have AI systems that are broadly better than almost all humans at almost all things." (World Economic Forum 2025)

- March 2025: Anthropic also claimed in a response to the U.S. Office of Science and Technology Policy that it anticipates that by 2027 AI systems will exist that equal the intellectual capabilities of "Nobel Prize winners across most disciplines—including biology, computer science, mathematics, and engineering." (D'Souza 2025)
- May 2025: Amodei has stated that AI could increase overall unemployment to 10-20% in the next one to five years, a prediction highlighted by Barack Obama. (Allen and VandeHei 2025; Obama 2025)

Sam Altman of OpenAI states that:

- January 2025: "I think AGI will probably get developed during [Donald Trump's second presidential] term, and getting that right seems really important." (Tyrangiel 2025)

Elon Musk, leader of xAI and Tesla, writes:

- December 2024: "It is increasingly likely that AI will superset [sic] the intelligence of any single human by the end of 2025 and maybe all humans by 2027/2028. Probability that AI exceeds the intelligence of all humans combined by 2030 is ~100%." (Musk 2024)
- August 2025: When a user posted "By 2030, all jobs will be replaced by AI and robots," Musk responded: "Your estimates are about right." (Musk 2025)

Demis Hassabis, CEO and co-founder of Google DeepMind predicts:

- August 2025: "We'll have something that we could sort of reasonably call AGI, that exhibits all the cognitive capabilities humans have, maybe in the next five to 10 years, possibly the lower end of that." (Rose 2025)
- August 2025: "It's going to be 10 times bigger than the Industrial Revolution, and maybe 10 times faster." (Rose 2025)

While we cannot directly compare these claims to LEAP questions, we offer clear evidence based on LEAP forecasts that the median expert expects substantially smaller effects of AI than is expected by frontier AI company leaders:

- *General capabilities:* Lab leaders predict human-level or superhuman AI by 2026-2029, while our expert panel indicates longer timelines for superhuman capabilities. By 2030, the average expert thinks that 23% of LEAP panelists will say the state of AI most closely mirrors an ("rapid") AI progress scenario that matches some of these claims.[67,68]
- *White-collar jobs:* The median expert predicts 2% growth in white-collar employment by 2030 (compared to a 6.8% trend extrapolation).[69] This contrasts with Elon Musk's suggestion that all jobs might be replaced by 2030.[70] Relatedly, Dario Amodei predicts 10-20% overall unemployment within the next five years.

[67] *Raw data*: The median forecast for this question was 20%. IQR on the 50th percentile was (10%–30%).
[68] See Appendix E.I. Survey Questions: Wave 1 for background information. We ask participants what proportion of LEAP panelists will choose "slow progress," "moderate progress," or "rapid progress" as best matching the general level of AI progress.
[69] *Raw data*: IQR on the 50th percentile was (-4%–5%).
[70] In a future survey wave, we plan to collect forecasts of the predicted relationship between AI capabilities and employment growth in each sector by asking respondents to forecast employment growth conditional on low-, moderate-, and rapid-progress scenarios.

- *Millennium Prize Problems:* The median expert gives a 60% chance that AI will substantially assist in solving a Millennium Prize Problem by 2040[71] (and 20% by 2030).[72] Amodei's prediction of general “Nobel Prize winner” level capabilities by 2026-2027 could imply a much more aggressive timeline, but the implication of Amodei’s predictions are somewhat unclear.[73]

The rationales help explain why the median expert forecasted slower progress than AI company leaders predict. Some reference specific claims made by company leaders: “Anthropic CEO's forecast of 90% of coding in the USA done by AI ‘within six months’ has been a fantastic dud,” wrote one expert, and another, “Musk has been talking about autonomous driving for ages, and it´s always been worse in the end than he said.”

Others offer broader arguments as to why they believe the timelines predicted by the leaders of frontier AI companies are unlikely to materialize. Below are two examples of these arguments:

> Radical change in major systems just takes longer than 4-5 years. I also think that [in] many of these domains, even unexpectedly fast advancement in AIs will not easily translate to improvements for quite some time because of unexpected barriers. That is, at least until we have strong artificial general intelligence (AGI), which we will not by 2030. To paraphrase an old saying, every job looks easy for those not actually doing it.

> The force of its impact will likely be slowed by bottlenecks in areas AI hasn’t yet conquered. An important concept is that an economic bottleneck grows in significance when productivity elsewhere increases. For instance, if global shipping dramatically increases, a bottleneck in the Suez or Panama Canal becomes much more costly. There likely exist thousands (millions?) of potential bottlenecks in the economy which will only become legible as other processes are sped up by orders of magnitude.

4. Experts predict much faster AI progress than the general public.

Of 68 total forecasts[74] (across 14 questions with multiple time horizons and quantiles) with a clear valence of AI capabilities,[75] the general public holds views about AI progress, capabilities,

[71] *Raw data*: IQR on the 50th percentile was (30%–81%).
[72] *Raw data*: IQR on the 50th percentile was (10%–50%).
[73] The degree to which progress on Millennium Prize Problems is serial or parallel, as well as the general difficulty of the Problems, complicates this comparison. Eliciting forecasts from multiple experts on consistent forecasting questions with clear resolution criteria helps us bring clarity to debates often plagued by ambiguous definitions.
[74] Questions include FrontierMath, Autonomous Vehicle Trips, Millennium Prize, Diffusion of AI Across Sciences, Drug Discovery, Electricity Consumption, Cognitive Limitations, AI Investment, Generative AI Use Intensity, Open vs Proprietary Polarity, AI Companions, Barriers to Adoption, General AI progress, and Technological Richter Scale.
[75] Forecasting questions with clear valence have an unambiguous directional association with progress. For some questions, like employment by sector, it is unclear whether higher or lower levels of unemployment would be associated with more advanced or less advanced AI progress, so we exclude those questions from this analysis. We also transform some forecasts to establish the progress valence. First, we take the average across all fields in Diffusion of AI Across Sciences. Second, for Cognitive Limitations and Barriers to Adoption, we average across all categories and use the complementary

and diffusion that are statistically indistinguishable[76] from experts in 9% of cases, predict less progress[77] than experts in a large majority (71%) of all cases, and predict more progress in 21% of forecasts. Where experts and the public disagree, the public predicts less progress over three times as often as more progress. Across these forecasts that exhibit a clear valence of AI capabilities, a randomly selected expert is 16% more likely than a randomly selected member of the public to predict faster progress than would be expected by random chance. Note, all Mann-Whitney U-tests and Cliff's δ[78] values are calculated on unweighted data, but we plan to integrate weights into all such analyses in future work. We summarize some of the differences in aggregate forecasts in Figure 5, and Figure 6 plots pooled distributions for experts against those for the public. The questions selected in Figure 5 reflect the progress-valenced questions that easily map onto a percentage scale.

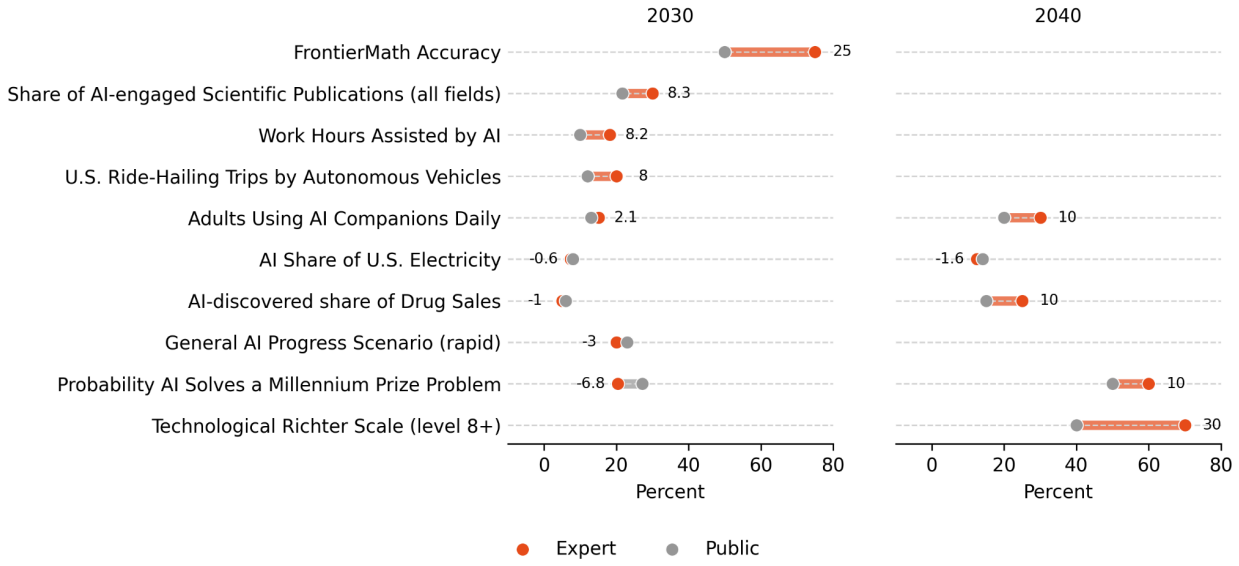


Figure 5: Differences between the expert and public median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of each groups' 50th percentile forecasts. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

probability. Third, we take the average of closed- and open-weight performance for Open vs Proprietary Polarity. Lastly, we take the values assigned to the "Rapid" scenario and TRS levels 8 and above for the General AI Progress and Technological Richter Scale questions, respectively.

[76] We use Mann-Whitney U-tests for equality in distribution unless otherwise stated, with a 5% significance threshold. All Mann-Whitney U-tests and Cliff's δ values are currently unweighted.

[77] We claim a group predicts statistically significantly less progress according to a Mann-Whitney U-test *and* a negative Cliff's δ.

[78] Cliff's δ performs pairwise comparisons between all values of two empirical distributions. It takes the number of comparisons where the value from the first distribution exceeds the second and subtracts the number of comparisons where the value from the second distribution exceeds the first and reports this difference as a proportion of the count of comparisons. In other words, it reports the probability that a randomly drawn value from the first distribution exceeds a randomly drawn value from the second distribution, over and above what would be expected by pure random chance if the two distributions were identical.

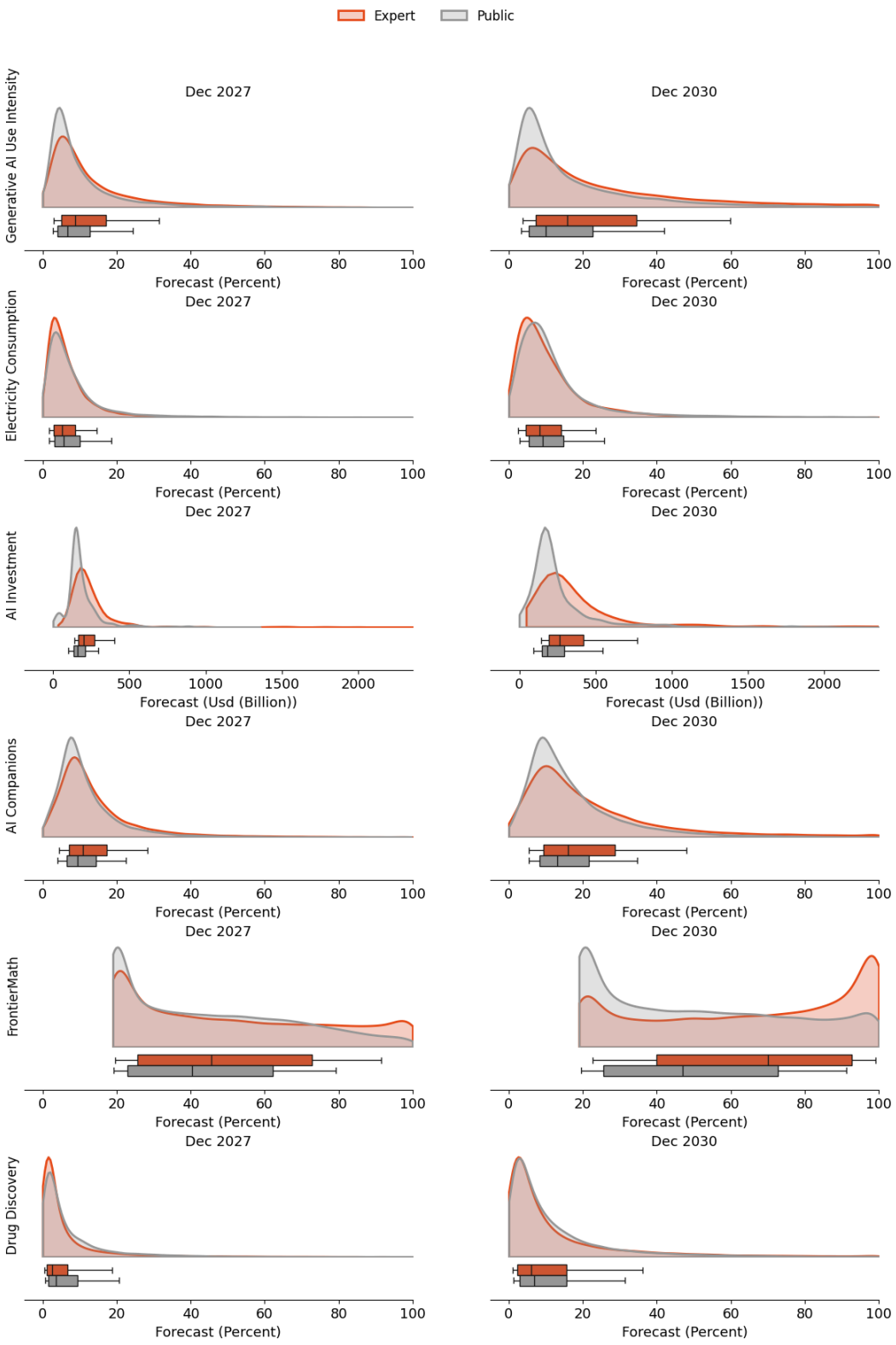


Figure 6: Pooled distributions for expert and public forecasts on various questions across years. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. See Uncertainty and Disagreement for details.

We report statistical comparisons between experts and the public for select questions in Table 5 below. You can find tables for all other relevant questions on the LEAP website.[79]

Question	Resolution Year	Percentile	p-value	Cliff's δ
AV Trips	2027	25	0.017	0.085
	2027	50	<0.001	0.16
	2027	75	<0.001	0.23
	2030	25	<0.001	0.12
	2030	50	<0.001	0.26
	2030	75	<0.001	0.35
Drug Discovery	2027	25	<0.001	-0.26
	2027	50	<0.001	-0.19
	2027	75	0.018	-0.093
	2030	25	<0.001	-0.14
	2030	50	0.28	-0.043
	2030	75	0.15	0.056
	2040	25	<0.01	0.12
	2040	50	<0.001	0.28
	2040	75	<0.001	0.37

Table 5: Statistical comparisons for experts and the public for U.S. Ride-Hailing Trips by Autonomous Vehicles (AV Trips) and AI-discovered share of Drug Sales (Drug Discovery). We report p-values from Mann-Whitney U tests and Cliff's δ values. A positive value indicates that expert forecasts tend to be larger than public forecasts.

We summarize some of the major differences below:

- *Societal impact:* On average, experts give a 63% chance that AI will be at least as impactful as a "technology of the century"—like electricity or automobiles—whereas the public gives this a 43% chance. Further, experts give a 32% chance that it will be at least as impactful as a "technology of the millennium" (akin to the printing press or the Industrial Revolution), while the public gives this a 22% chance.

[79] See https://leap.forecastingresearch.org/reports/ to access these tables.

- *Autonomous vehicles:* The public predicts about half as much autonomous vehicle progress as experts by 2030, as suggested by each group's 50th percentile forecasts. The median expert in our sample predicts that usage of autonomous vehicles will grow dramatically—from a baseline of 0.27% of all U.S. rideshare trips in Q4 2024 to 20% by the end of 2030.[80] In comparison, the general public predicts 12%[81] (*$p < 0.001$, Cliffs $\delta = 0.26$*).
- *Generative AI use:* The public predicts about half as much generative AI use in 2030. Experts predict that 18% of U.S. work hours will be assisted by generative AI in 2030, whereas the general public predicts 10% (*$p < 0.001$, Cliffs $\delta = 0.29$*).
- *Mathematics:* 23% of experts predict that FrontierMath[82] will be saturated by the end of 2030 in the median scenario,[83] meaning that AI can autonomously solve a typical math problem that a math PhD student might spend multiple days completing. Only 6% of the public predict the same, about 3x less.
- *Diffusion into science:* Experts predict a roughly 10x increase (from 3% to roughly 30%) in AI-engaged papers across Physics, Materials Science, and Medicine between 2022 and 2030.[84] The general public predicts two thirds as much diffusion into science: that roughly 20% of papers in these fields will be AI-engaged.[85]
- *Drug discovery:* By 2040, the median expert predicts that 25% of sales from newly approved U.S. drugs will be from AI-discovered drugs, compared to 15% for the public (*$p < 0.001$*, *Cliff's $\delta = 0.28$*). The median expert also thinks there's a 25% chance that AI-discovered drugs will account for more than 43% of recent drug sales, whereas the general public predicts there's a 25% chance of a share greater than 23%—about half the expert forecast. In contrast, the public expects a larger share from AI-discovered drugs in the short-run, predicting that AI-discovered drugs will account for 2.5% of recent drug sales in 2027; experts predict 1.6% of 2027 sales will be attributed to AI-discovered drugs (*$p < 0.001$, Cliff's $\delta = 0.19$)*.[86]

---

[80] *Pooled distribution*: IQR (6.9%–46.1%); variance decomposition: 47% between–forecaster disagreement, 53% within–forecaster uncertainty. *Raw data*: IQR on the 50th percentile was (10%–40%); median 25th and 75th percentile forecasts were 8% and 35% respectively.

[81] *Pooled distribution*: IQR (4%–31%); variance decomposition: 66% between–forecaster disagreement, 34% within–forecaster uncertainty. *Raw data*: IQR on the 50th percentile was (5%–29%); median 25th and 75th percentile forecasts were 5% and 20% respectively. Uncertainty and disagreement metrics on other claims made in this list can be found in the Monthly Reports at https://leap.forecastingresearch.org/reports/.

[82] "To gauge the difficulty of FrontierMath problems, we organized a competition at MIT involving around 40 exceptional math undergraduates and subject-matter experts. Participants formed eight teams of four or five members, each with internet access, and had four and a half hours to solve 23 problems. On a subset of 23 tier 1-3 problems, the average team scored 19%, while 35% of the problems were solved collectively across all teams." (Epoch AI 2025).

[83] This estimate of 23% reflects the fraction of experts whose median forecast is that AI systems will achieve performance of at least 90% (which we call saturation) on Tiers 1–3 of FrontierMath. We take the average of the proportions calculated under weak and strict inequality.

[84] Physics: 32%; Materials Science: 37%; Medicine: 37%.

[85] Physics: 27%; Materials Science: 30%; Medicine: 30%.

[86] For this comparison, we switch to reporting Cliff's δ values calculated with the public as the first distribution.

Contrary to this result, the public assigns more weight to the “Rapid Progress” scenario in the General AI Progress question: the average member of the public expects 26% of LEAP panelists in 2030 will select the rapid scenario (95% confidence interval [25.5%, 26.4%]),[87] compared to 23% for experts (95% confidence interval [22.1%, 23.7%]).[88]

To assess the extent to which low-effort or relatively lower comprehension from the public could drive these results, we compare members of our public sample with high forecasting accuracy in other studies to those with low accuracy. We do not find that one group systematically expects more or less progress. Public Accuracy Stratification details this analysis. Additionally, while within-forecaster uncertainty explains 49% of the total variation in expert forecasts, this component explains just 37% of the variation in public forecasts. As forecasting questions resolve, we will compare the calibration (and accuracy) of expert and public forecasts.

5. There are few differences in prediction between superforecasters and experts, but, where there is disagreement, experts tend to expect more AI progress. We don’t see systematic differences between the beliefs of computer scientists, economists, industry professionals, and policy professionals.

There are no discernible differences between forecasts from different groups of experts. Across all pairwise comparisons of expert categories for each of the questions with a clear AI progress valence, only 32 out of 408 combinations (7.8%) show statistically significant differences (at a 5% threshold), similar to what you would expect from chance. This means that computer scientists, economists, industry professionals, and policy professionals largely predict similar futures *as groups*, despite there being significant disagreement about AI among experts. This raises questions about popular narratives that economists tend to be skeptical of AI progress and that employees of AI companies tend to be more optimistic about fast gains in AI capabilities. In other words, while we do see widespread disagreement among experts about the future of AI systems, capabilities, and diffusion, we fail to find evidence that this disagreement is explained by the domain in which experts work. As LEAP continues, we plan to study what factors most drive expert disagreement. However, the groups used for these comparisons are subsets of our expert sample, so these comparisons are necessarily less powered.

Superforecasters and expert groups predict similar futures. Superforecasters are statistically indistinguishable from experts in 69% of valenced forecasts, predict less progress than experts in 26% of forecasts, and more progress in 4% of forecasts. A randomly selected expert is 9.8% more likely than a randomly selected superforecaster to predict faster progress than would be expected by random chance.

Where superforecasters and experts disagree, superforecasters usually (86% of such cases) predict less progress. Further, some of these disagreements are quite large**.** For example, the median expert predicts that use of autonomous vehicles will grow dramatically—from 0.27% of

[87] The median forecast was 23%. *Raw data*: IQR on the 50th percentile was (12%–35%)
[88] The median forecast was 20%. *Raw data*: IQR on the 50th percentile was (10%–30%)

all U.S. rideshare trips in 2024 to 20% by the end of 2030,[89] whereas the median superforecaster predicts less than half that, 8%[90] (*p < 0.001)*. A randomly selected superforecaster predicts, in the median scenario for 2030, less AV penetration than a randomly selected expert 37% more often than would be expected by random chance. Superforecasters also predict less societal impact from AI and less AI-driven electricity use. Drug discovery is the only setting where superforecasters are more optimistic than experts: By 2040, experts predict that 25% of sales from recently approved U.S. drugs will be from AI-discovered drugs. Superforecasters predict 45%, almost double (*p < 0.01*). Here, a randomly selected superforecaster predicts a higher share, in the median scenario, than a randomly selected expert 23% more often than would be expected by chance.

This pattern is consistent with the follow-up to our Existential Risk Persuasion Tournament, where a small sample of experts were more optimistic about AI progress than around 80 superforecasters about AI progress and capabilities in a 2022 (pre-ChatGPT) survey, although both experts and superforecasters significantly underestimated AI progress by 2025 (Kučinskas et al. 2025; Karger et al. 2025). We summarize some of the differences in aggregate forecasts in Figure 7, and Figure 8 plots pooled distributions for experts against superforecasters. In Figure 9, we compare aggregate forecasts of the various expert groups.

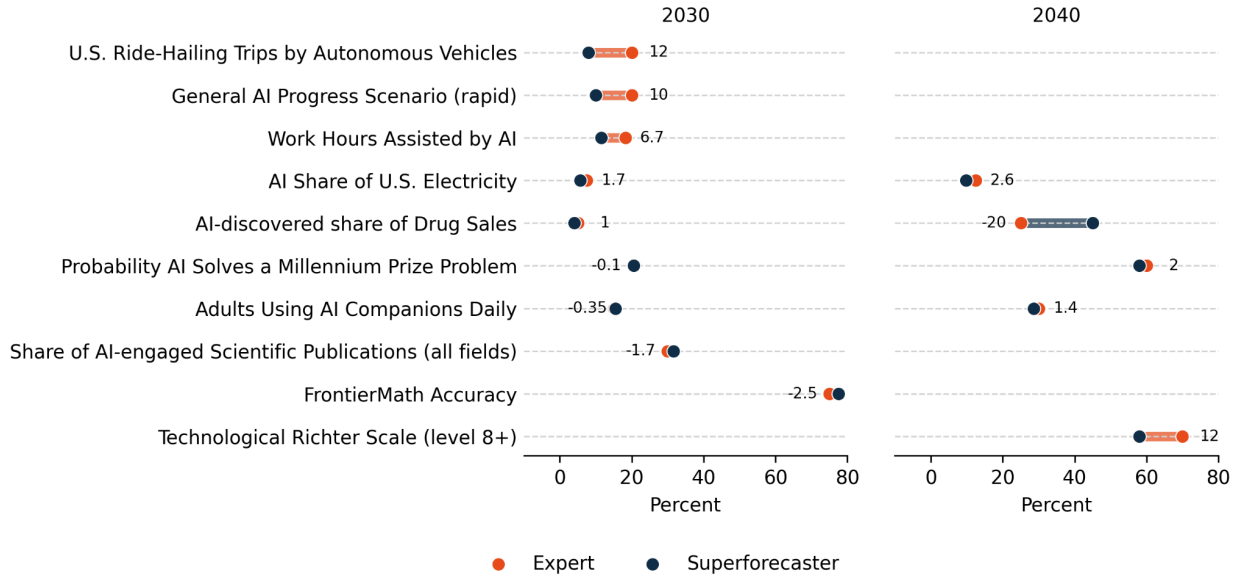


Figure 7: Differences between the expert and superforecaster median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of each groups' 50th percentile

[89] *Pooled distribution*: IQR (6.9%–46.2%); variance decomposition: 47% between–forecaster disagreement, 53% within–forecaster uncertainty. *Raw data*: IQR on the 50th percentile was (10%–40%); median 25th and 75th percentile forecasts were 8% and 35% respectively.
[90] *Pooled distribution*: IQR (3%–25%); variance decomposition: 48% between–forecaster disagreement, 52% within–forecaster uncertainty. _Raw data_: IQR on the 50th percentile was (3%–25%); median 25th and 75th percentile forecasts were 4% and 20% respectively. Uncertainty and disagreement metrics on other claims made in this paragraph can be found in the Monthly Reports at https://leap.forecastingresearch.org/reports/.

forecasts. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

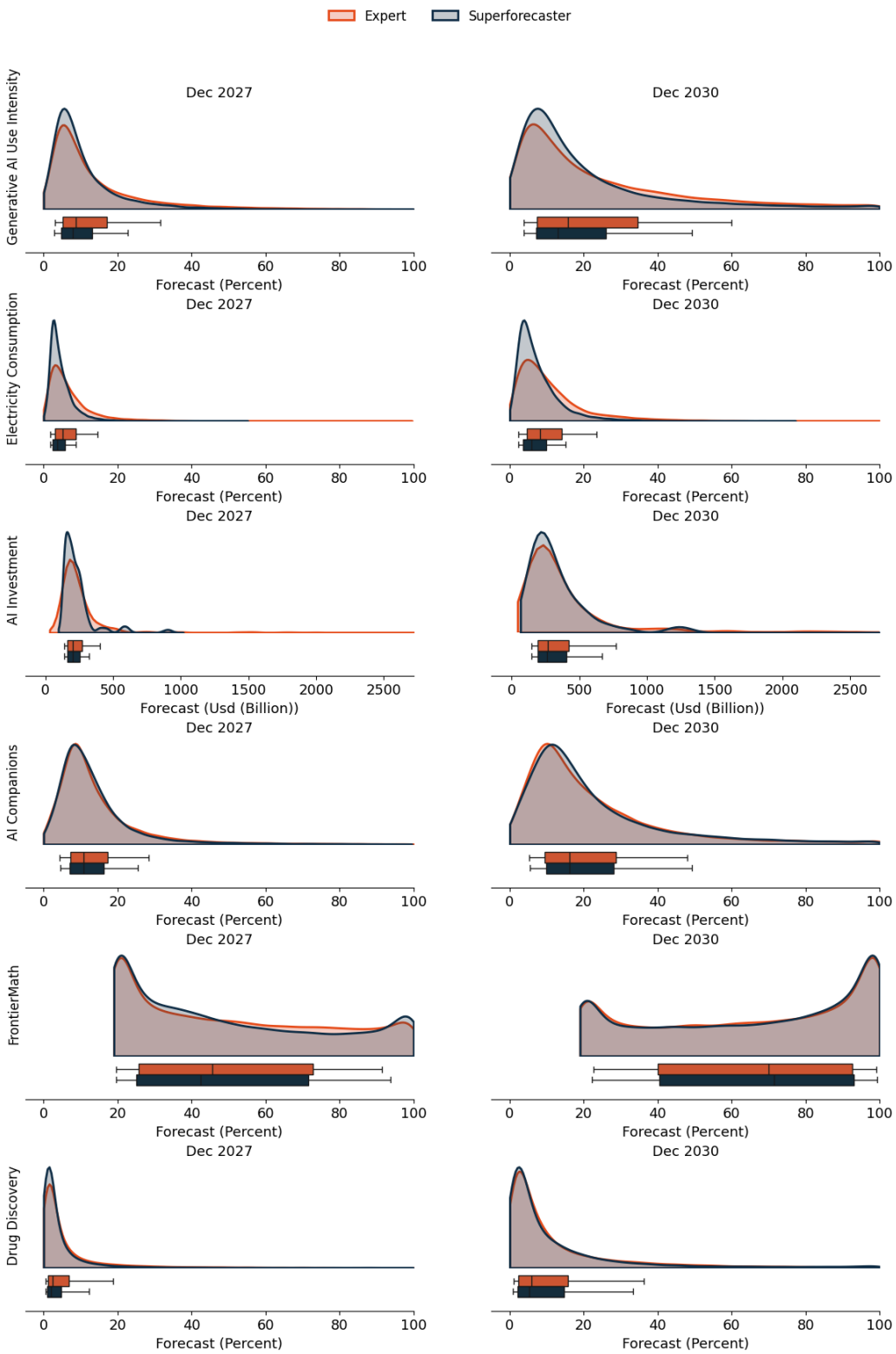


Figure 8: Pooled distributions for expert and superforecaster forecasts on various questions across years. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. See Uncertainty and Disagreement for details.

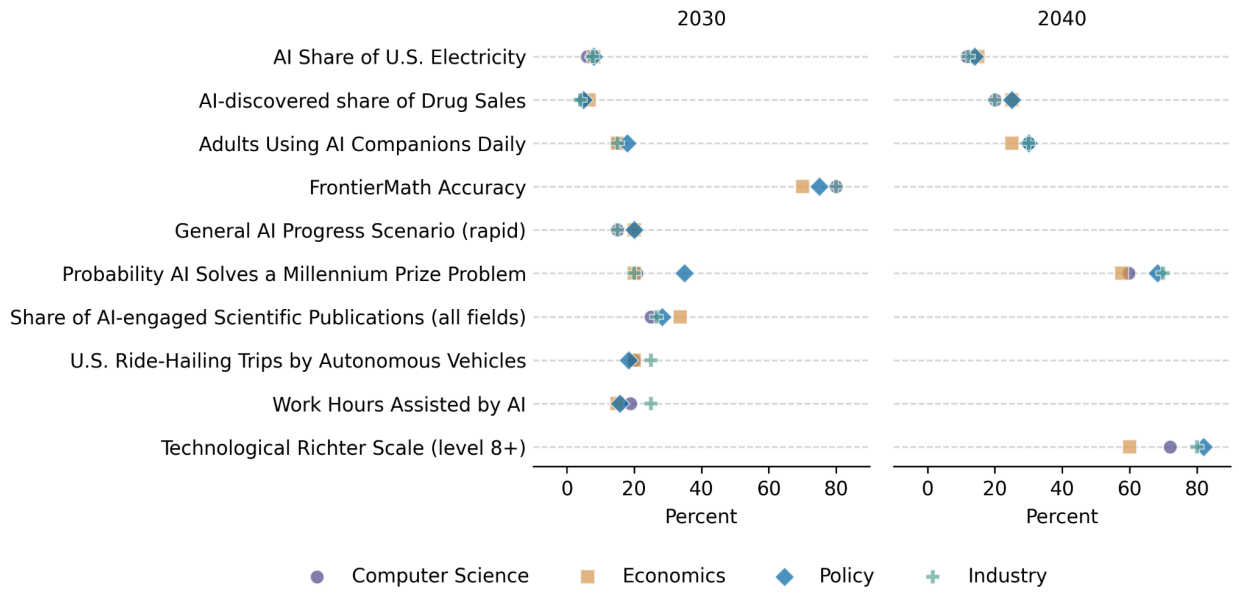


Figure 9: Expert category median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of 50th percentile forecasts for each category. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

We report statistical comparisons between experts and superforecasters for select questions in Table 6 below. You can find tables for all other relevant questions on the LEAP website.[91]

[91] See https://leap.forecastingresearch.org/reports/ to access these tables.

Question	Resolution Year	Percentile	p-value	Cliff's Delta
AV Trips	2027	25	<0.001	0.34
	2027	50	<0.001	0.45
	2027	75	<0.001	0.44
	2030	25	<0.001	0.31
	2030	50	<0.001	0.37
	2030	75	<0.001	0.32
Drug Discovery	2027	25	0.58	0.042
	2027	50	0.084	0.14
	2027	75	0.09	0.14
	2030	25	0.44	0.064
	2030	50	0.66	0.037
	2030	75	0.81	0.021
	2040	25	0.014	-0.2
	2040	50	<0.01	-0.23
	2040	75	<0.01	-0.22

Table 6: Statistical comparisons for experts and superforecasters for U.S. Ride-Hailing Trips by Autonomous Vehicles (AV Trips) and AI-discovered share of Drug Sales (Drug Discovery). We report p-values from Mann-Whitney U tests and Cliff's δ values. A positive value indicates that expert forecasts tend to be larger than superforecaster forecasts.

Highlights from Question-by-Question Analyses

To provide a window into some of the details available in the question-level analysis, below we highlight some key reasoning offered by experts in their written rationales. We present these highlights for a handful of questions, but more detail is available in Appendix F. Question-by-Question Results.

FrontierMath:[92] Experts who forecast a high degree of progress on this benchmark often point to recent trends. One argues, “We've seen jumps of around 5 points[93] on this benchmark every couple of months so far and these jumps will only accelerate as scores approach 50% (benchmark scores tend to be roughly sigmoid-shaped over time).” Many also emphasize inference scaling potential, with one observing that “the current top scorer is a reasoning model, and the reasoning model paradigm is relatively new; this suggests that rapid improvements are likely as the paradigm evolves,” and another opining that “with very large amounts of inference compute, it's possible that o3 or o4-mini could already get well over 30% today.” Beyond advances in technical capabilities, a significant minority of high-forecast respondents point to the likelihood of sustained investment in math-related AI capabilities due to its prestige and R&D value. Writes one: “Math is highly relevant to many R&D domains, so progress in math has been, and is highly likely to continue to be, a focus for leading AI companies.” Among low-forecast respondents, a common sentiment is that “the fastest . . . progress is behind us and we are now approaching the flat/end point portion of the S-Curve of advancement.” Many also express skepticism that inference scaling will be enough to overcome fundamental architectural limitations.

Autonomous vehicles:[94] A common consideration among experts who forecast that a high percentage of U.S. ride-hailing trips will eventually be provided by Level 4 autonomous vehicles was that Level 4 technology has a proven track record. One writes, “Historically, when a technology finally gets to be used in the wild, it improves very rapidly.” Another highlights Waymo's exponential expansion, noting that “Waymo is currently more-than-doubling every year,” and that this could result in a data flywheel where “broader deployment will generate more data, which in turn enhances safety—creating a positive feedback loop.” Conversely, low-forecast respondents often emphasize historical overpromising and argue the current technology is difficult to scale “because it requires lots of case-by-case optimization for a particular region (down to individual intersections),” and "progress in Phoenix or Miami does not generalize easily to New York, Boston, or Chicago.”

White-collar jobs:[95] Experts who forecast that the percentage of white-collar jobs will be higher in the future relative to 2025 often cite historical precedent when arguing that white-collar employment is more likely to adapt than collapse: “Historically, human labor patterns have experienced quite radical transformations over time, even in established sectors. Emerging technologies, rather than sucking people out of the labor market of white-collar work, are more likely to make them work differently and lead to new white-collar roles that can capitalize on this transition.”Low-forecast respondents tend to view AI's speed and cognitive focus as a key differentiator from prior periods of transition: “White collar workers...work on symbolic tasks,

[92] Experts were asked, “What will be the highest percentage accuracy achieved by an AI model on FrontierMath, by Jan 1 of 2026, 2028, and 2031?”
[93] Actual progress was marginally slower. The top Tier 1-3 accuracy rate rose from 1.03% in June of 2024 to 29% in August of 2025, where it remained as of the publication of this paper.
[94] Experts were asked, “What percentage of U.S. ride-hailing trips will be provided by autonomous vehicles that are classified SAE Level 4 or above in the years 2027 and 2030?”
[95] Experts were asked, “What will the percent change in the number of jobs (compared to Jan 1, 2025) in the U.S. be for white-collar, blue-collar, and service occupations, by Jan 1 of 2028 and 2031?”

generate language, and make decisions based on analysis of data, all tasks to which LLMs are well suited," writes one. Another points to evidence of layoffs already occurring, particularly in the tech sector: "From Intel to Microsoft, many top executives and management staff were laid off to make room for other investments at the organization. Google laid off 10% of its managerial staff last December." Still another adds, "software engineering is already being hugely impacted and this is only accelerating."

Speed of AI progress:[96] Many experts who forecast a high likelihood that AI progress will be rapid mention the consistent pace of capability improvements, with one noting, "historically AI system development has followed a steep scaling curve and increases in model size, data and compute have led to rapid capability gains." Another points out that, "METR [Model Evaluation and Threat Research] results imply a roughly 4 to 10x improvement in time horizon every year, which means that we'll have systems capable of doing weeks or months of work by the late 2020s."[97] Many also note the potential for a recursive self-improvement feedback loop, in which AI improves itself, to significantly accelerate growth. Slow-progress forecasters often highlight physical constraints, pointing to the need for better training data and the cost of compute—especially as it relates to energy needs: "I expect energy to be the chief bottleneck to AI progress such that it will be a rate-limiter for progress in general." Another common sentiment shared by slow-progress forecasters is that "for any of the moderate and rapid progress criteria to be met there would need to be a massive paradigm shift in AI technology," and that such a shift in the underlying LLM architecture is unlikely to materialize in the near term.

Technological Richter Scale:[98] Intended to be analogous to the measurement of earthquake magnitudes, the technological Richter scale instead attempts to measure the impact of technologies. Experts who predict AI will have an exceptionally high impact tend to see it as uniquely positioned to fundamentally restructure society, with one going so far as to suggest it could "challenge capitalism" and "has the potential to replace human labor in most fields [and] might force our societies to shift to a new economic model." This transformative potential is believed by many to stem from AI's dual nature in that "it both augments human intelligence and will eventually surpass it." High-impact respondents also point to the quick pace of AI deployment and emphasize the likelihood of rapid AI progress: "People fundamentally don't think in exponentials. 2040 is a LONG time away, technologically. And AI will modify AI, at which point its improvement will go even more second-order." Low-impact respondents often question the sustainability of recent growth patterns, with one noting that "upper levels, and in particular 10, require sustained exponential growth; this is unlikely to materialize given that natural growth

[96] Experts were presented with three scenarios that detailed the development of AI capabilities and asked, "At the end of 2030, what percent of LEAP panelists will choose 'slow progress,' 'moderate progress,' or 'rapid progress' as best matching the general level of AI progress?"

[97] Actual rate of improvement likely falls within this range, especially given recent acceleration trends, but there is considerable uncertainty and domain variability. See: https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/ for more information.

[98] Nate Silver's book "On the Edge" proposes the technological Richter scale (TRS) which, analogous with earthquake magnitudes, rates the impact of technologies on a roughly logarithmic scale with10 representing that greatest impact (Silver 2024). Experts were asked, "At the end of 2040, what is the probability for AI achieving the following levels of net impact [ranging from 5-10] on human society as compared to the impact of past technological events?"

(e.g., of bacteria) follow a sigmoid shape." Another points to regulations and physical bottlenecks as constraints and challenges whether AI will deliver tangible benefits sufficient to drive transformation: "The average citizen will not have much benefit to buy from AI. Improved games or art? Cheaper manufactured goods? A robot to clean your house? How does AI deliver things that humans want, like better, cheaper healthcare?"

Millennium Prize:[99] The most common reason given by experts who think the likelihood is high that AI will solve (or substantially assist in solving) one of the notoriously difficult Millennium Prize math problems is that in January of 2025 the CEO of Google's DeepMind claimed that, in partnership with a team of mathematicians, they're "close to solving" one of the problems (later identified as Navier-Stokes) within "a year or year and a half" (Ansede 2025). Many also point to AI's gold medals in the International Mathematical Olympiad (Luong and Lockhart 2025) and progress on FrontierMath as evidence of rapid capability growth in mathematical reasoning. Low-forecast respondents generally don't mention the DeepMind claim; a few do, but dismiss corporate pronouncements. One instead cites the president of the Clay Mathematics Institute (the organization responsible for awarding the Millennium Prize) who in June of 2025 claimed, "We're very far away from AI being able to say anything serious about any of those problems" (Heaven 2025). Regarding purported progress, one mathematician notes: "The Math Olympiad is targeted toward gifted high school students spending an afternoon on a problem solvable with known techniques…[whereas] Millennium Prize problems can consume entire careers without a solution." Multiple low-forecast respondents also note FrontierMath Tier 4 (which poses much harder problems than Tiers 1-3) has <10% solve rates.

Diffusion into science:[100] When considering the percentage of publications in the fields of physics, materials science, and medicine that will be 'AI-engaged', experts predicting high engagement commonly extrapolate recent exponential growth, noting the 2018-2022 data shows engagement roughly tripling across all three fields, and emphasize this baseline predates ChatGPT: "What we see in the baseline data is just the beginning, resulting from the application of foundational AI models but largely without the generative AI models exploding on the scene." Many cite domain-specific breakthroughs (AlphaFold in protein folding, GNoME and MatterGen in materials discovery, AI-driven imaging and diagnostics in medicine) as evidence AI can deliver transformative results that will drive rapid diffusion. Low-forecast respondents instead tend to focus on interpretability and reliability issues that could slow diffusion: "AI is a black box that hallucinates," writes one. Others add that diffusion may be slow due to, "natural resistance to change from the existing body of researchers in these fields," or because, in the case of medical studies, AI isn't useful: "A significant portion of papers are observational, often reporting causal effects. There isn't much room for AI in these sorts of papers, as current statistical methods are more reliable and bias-free."

[99] Experts were asked, "What is the probability that AI will solve or substantially assist in solving a Millennium Prize Problem in mathematics by Dec 31 of 2027, 2030, and 2040?"
[100] Experts were asked, "What percent of publications in the fields of Physics, Materials Science, and Medicine in 2030 will be 'AI-engaged' as measured in a replication of this study?" (Duede et al. 2024)

Drug discovery:[101] Several experts who forecast that AI will accelerate drug discovery-to-market timelines point to faster design-make-test-analyze loops and potentially AI-enabled pharmacodynamic simulations that could streamline clinical trials. Others note AI-discovered drugs already demonstrate significantly higher Phase I success rates, and that extrapolating from current growth trends suggests "these will constitute a majority of new clinical trial submissions." Some also point out that discovery-to-market timelines can be shortened significantly during times of crisis via EUAs (emergency use authorizations). Low-forecast respondents commonly emphasize regulatory realities that may limit AI's impact on approval timelines. One writes: "Given that the median time it takes to get through the FDA approval process is over 10 years, and no AI-discovered drugs appear to have started Phase III trials yet,[102] 2027 is likely too soon for many, if any, new AI drugs to be approved." Although some low-forecast respondents acknowledge improved Phase I results, several make the point that "the turnaround time between Phase I and approval will not speed up substantially for AI-invented drugs," because "early entrants sped through Phase I but then quickly reverted to the mean in Phase II."

Electricity:[103] When assessing the percent of U.S. electricity consumption that will be used for training and deploying AI systems, high-forecast respondents often emphasize that the massive capital expenditure plans already announced by competing AI companies signal the type of unprecedented infrastructure investment that could result in "an explosion of energy usage." Others note that the geopolitical competition for AI supremacy may trigger an energy arms race, with one forecaster warning, "China is also investing massive amounts in datacenters," leading to the possibility that "we enter an arms race that is mostly determined by who can pump the most electricity into AI." Low-forecast respondents tend to focus more on potentially formidable constraints, particularly when considering "the material and political investments necessary to get significant growth—physical data centers, chips, permitting, water for cooling, transmission lines," and suggest these constraints could push infrastructure development offshore: "Major developers will possibly respond by increasingly outsourcing the physical infrastructure of data processing to locales outside of the US—there's no particular reason why models need to be trained inside of U.S. borders where the economic and political expenses are potentially much higher."

Private investment in AI:[104] Among experts who predict high future levels of global private investment in AI, most view the current level of investment as fundamentally justified and advance arguments along the lines of, "AI adoption is still in its early stages across many industries," and "the strong rebound to ~$130 billion in 2024 is critical. It occurred despite higher interest rates, signaling powerful, non-speculative belief in the transformative potential of generative AI." In contrast, frequently expressed sentiments among low-forecast respondents

[101] Experts were asked, "What percent of sales of recently approved U.S. drugs will be from AI-discovered drugs and products derived from AI-discovered drugs in the years 2027, 2030 and 2040?"
[102] This was true at the time this expert completed the survey.
[103] Experts were asked, "What percent of U.S. electricity consumption will be used for training and deploying AI systems in the years 2027, 2030 and 2040?"
[104] Experts were asked, "What will be the global private investment (in billion USD) in AI in the years 2027 and 2030?"

include doubts that productivity gains "will materialize quickly enough to justify high levels of investment," and that this could lead to the bursting of an AI bubble, with one expert noting, "both Deutsche Bank and Bain & Co. have just warned that the current AI boom is not sustainable" (Edwards 2025) and another likening the current situation to, "the dot com bubble in 2000."

AI companions:[105] Many experts who believe that a high proportion of U.S. adults will eventually use AI for companionship cite loneliness as a key driver. One points out that "the U.S. Surgeon General declar[ed] loneliness an epidemic in 2023, with about half of U.S. adults experiencing measurable levels of loneliness" (Office of the Surgeon General 2023). Others predict that as AI capabilities advance, AI companions will become more "sophisticated, emotionally intelligent, and capable of forming deeper connections with users," and that this will facilitate their integration into existing platforms and devices, driving use and normalization to the point where "ambient access through devices turns companionship into a series of micro-interactions throughout the day." Low-forecast respondents tend to believe humans are likely to have a strong preference for human companionship, with one arguing that "most people would find [AI] companionship unfulfilling, perhaps even viewing reliance on it as a kind of failure." Others point to lower saturation limits than the number of people who experience frequent loneliness, with one stating: "About a quarter of U.S. adults go to therapy.[106] If that's the market size, then I expect AI to eventually saturate [at] that."

Example of a Question-Level Analysis: Millennium Prize Problem

For each question, we conduct an analysis like the Millennium Prize example below, and present them in Appendix F. Question-by-Question Results. Each analysis summarizes the question and background information, summarizes the results, and analyzes rationales to uncover the core differences in view between low and high forecasts. In the first wave alone, experts and superforecasts wrote over 600,000 words supporting their beliefs. Analyzing these rationales alongside predictions provides significantly more context on why experts believe what they believe, and the drivers of disagreement, than the forecast alone.

Question. *Will AI solve or substantially assist in solving a Millennium Prize Problem in mathematics by 2027, 2030, and 2040?*

Background. The seven Millennium Prize Problems[107] were chosen by the founding Scientific Advisory Board of the Clay Mathematics Institute (CMI) of Cambridge, Massachusetts to be the most significant and difficult mathematics problems unsolved by 2000.

[105] Experts were asked, "What proportion of U.S. adults will self-report using AI for companionship at least once daily by Dec 31 of 2027, 2030, and 2040?"
[106] This claim may refer to the ~23% of U.S. adults who, according to a 2024 KFF (formerly Kaiser Family Foundation) study, "say they received mental health counseling and/or prescription medication for mental health concerns in the last year." See Panchal and Lo (2024).
[107] *Link provided to participants*: https://www.claymath.org/millennium-problems/

Historical Baseline. As of July 2025, only one of the seven problems has been solved (Clay Mathematics Institute, 2025).

For full question background and resolution details, see Appendix E.II. 1. Millenium Prize.

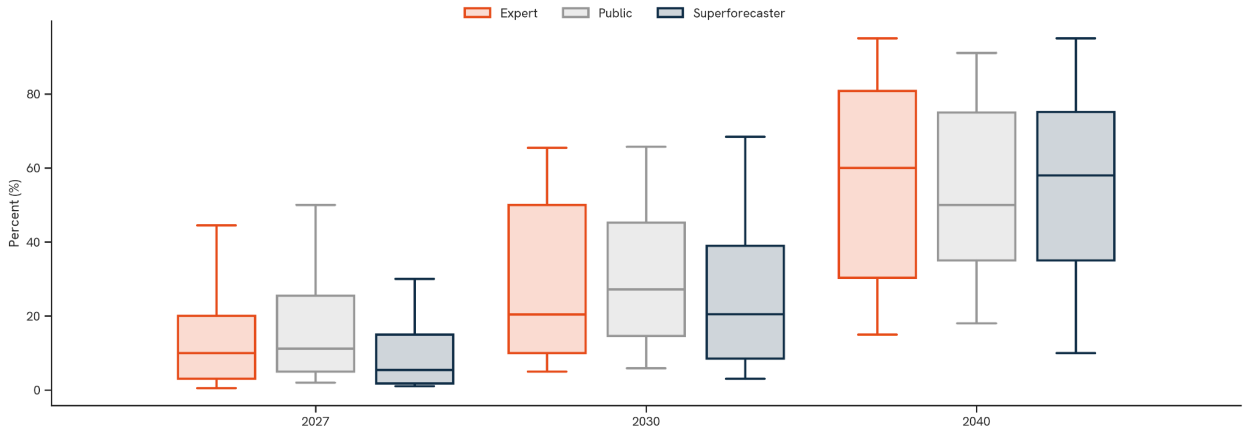


Figure 10: In this question, participants make 50th percentile forecasts for various resolution dates. This figure shows the 10th, 25th, 50th, 75th, and 90th percentiles of these 50th percentile forecasts, split by participant group. The 25th expert percentile for Dec 2027 represents the number that 25% of experts' median forecasts are lower than.

Results. Experts estimate a 10% chance that AI will solve or substantially assist in solving a Millennium Prize Problem by 2027,[108] up to 20% by 2030,[109] and **60% by 2040**.[110] All categories of experts, superforecasters, and the public largely predict similarly across timescales. However, there is wide disagreement *between* experts: the top quartile of experts think there's at least an even (50%) chance of AI assistance **by 2030**, whereas the bottom quartile of experts think there's only a 10% chance. The disagreement by 2040 is even larger: the interquartile range for expert medians is 30%–81%, while the top decile of experts think there's a 95% chance and the bottom decile of experts think there's only a 15% chance.

For full results tables, see here.

The rationales experts wrote to explain their forecasts lend considerable insight into their core areas of disagreement, in particular:

- **DeepMind/Navier-Stokes**: High-forecast respondents frequently cite DeepMind CEO's January 2025 statement that, in partnership with a team of mathematicians, they're "close to solving" one of the problems (later identified as Navier-Stokes) within "a year or year and a half" (Ansede 2025). This is treated as strong concrete evidence. Low-forecast respondents generally don't mention this or dismiss corporate pronouncements. One expert cites the Clay Institute president's June 2025 claim: "We're

[108] *Raw data*: IQR on the 50th percentile was (3.0%–20.0%). 90th percentile of median forecast: 44.5.
[109] *Raw data*: IQR on the 50th percentile was (10.0%–50.0%). 90th percentile of median forecast: 65.4.
[110] *Raw data*: IQR on the 50th percentile was (30.3%–80.8%). 90th percentile of median forecast: 95.0.

very far away from AI being able to say anything serious about any of those problems" (Heaven 2025).

- **Benchmarks**: Many high-forecast respondents point to International Mathematical Olympiad gold medals and FrontierMath progress as evidence of rapid capability growth in mathematical reasoning that will likely continue (Luong and Lockhart 2025). Low-forecast respondents tend to argue these are fundamentally different challenges. One mathematician notes: "The Math Olympiad is targeted toward gifted high school students spending an afternoon on a problem solvable with known techniques…[whereas] Millennium Prize problems can consume entire careers without a solution." Multiple forecasters note FrontierMath Tier 4 (which poses much harder problems than Tiers 1-3) has <10% solve rates.
- **The nature of Millennium problems**: High-forecast respondents commonly emphasize that math is verifiable, has clear structure, and that some problems (Navier-Stokes, Birch–Swinnerton-Dyer) may be suited to AI-assisted numerical exploration or pattern recognition. Low-forecast respondents often express doubts that Millennium Problems are solvable with the current AI paradigm, emphasizing doing so requires "deep conceptual breakthroughs," "developing new concepts and mathematical rules," and "truly out of the box thinking." One domain expert writes: "The current generation of AI does not seem to be able to do this sort of creative mathematical work at all. It can apply known techniques and get novel results, but these results would be very easy for top working mathematicians."
- **Base rates and timelines**: High-forecast respondents mostly don't engage with base rates, or they argue that AI changes the game fundamentally. By contrast, many low-forecast respondents emphasize that only one out of seven problems have been solved in the 25 years since the prize was announced, meaning some have remained unsolved for more than a century. They also highlight Millennium Prize rules: upon the publication of a solution, a minimum of two years must pass before a prize can be awarded, to allow time for adequate verification. (In the case of the one prize that was awarded, the gap between the publication of the solution and the awarding of the prize was over seven years.) This, many low-forecast respondents point out, renders the 2027/2030 dates almost impossible regardless of technical progress.
- **"Substantially assist" interpretation**: High-forecast respondents tend toward a broad interpretation—any meaningful acceleration of human-AI hybrid research counts, whereas low-forecast respondents tend toward restrictive interpretation. One notes the resolution criteria require contribution "likely not producible without AI," which is a higher bar.
- **Architecture sufficiency**: Most high-forecast respondents believe incremental improvements over current LLM capabilities will be sufficient, especially when paired with specialized tools (Lean, AlphaProof) and human collaboration. Low-forecast respondents frequently argue the current LLM paradigm fundamentally cannot do this. Multiple forecasters say we need "entirely new architectures" (neurosymbolic systems were mentioned several times) or that a "pattern matching paradigm doesn't extend to the deep creativity required."

- **Difficulty of achieving *superhuman* performance**: Although rarely discussed by high-forecast respondents, a few low-forecast respondents expressed doubts that this could be achieved, with one writing, "Training a model to do math at the level of human experts might be a qualitatively different ML problem from training a model to do math *surpassing* expert capabilities. RL training requires creating problems with reward functions...We haven't achieved that with reasoning post-training yet."

High-forecast rationale examples:

"I guess the elephant in the room is that DeepMind says they are close: The so-called Navier-Stokes Operation, underway for three years with a team of 20 people, has so far been carried out with complete discretion, although the chief of Google DeepMind, Demis Hassabis let slip in a January interview that they are 'close to solving a Millennium Prize Problem' without mentioning which one. 'We'll see that in the next year or year and a half."[111]

"Some of the problems, like the Riemann Hypothesis or the Birch and Swinnerton-Dyer Conjecture, are especially well-suited to AI-supported exploration. They bear a kind of family resemblance to the Four-Color Theorem in their relationship to computer-assisted mathematics. The Four-Color Theorem was famously solved through a hybrid of human conceptual framing and extensive computer verification. As Donald MacKenzie details in his socio-history of that episode [reference below[112]], much of the intellectual labor wasn't in the computation itself but in formalizing the problem in a way that machines could meaningfully engage with it and in managing the institutional consequences of proof-by-machine."

"My optimism that AI could achieve high level original mathematics is revised upward significantly since the Bubeck announcement about GPT-5 a few weeks ago regarding the first confirmed example of novel mathematical reasoning generated by a LLM."[113]

Low-forecast rationale examples:

"I have domain expertise here as a mathematician. The current generation of AI does not seem to be able to do this sort of creative mathematical work at all. It can apply known techniques, and get novel results, but these results would be very easy for top working mathematicians. The kind of pattern matching paradigm we have seen so far apparently doesn't extend at all to deep creativity required."

"If the millennium problems all require new insights absent from the training data, then current LLM technology is simply not up to the task: we will need instead new AI paradigms that are better at creating non-combinatorial insights (i.e., insights that do not originate from the recombination of patterns already learned by the AI). This will take

[111] The expert is referring to and quoting from Ansede (2025).
[112] The expert is referring to MacKenzie (1999).
[113] The expert appears to be referring to an August 2025 post by OpenAI researcher Sebastien Bubeck (Bubeck 2025).

time: it is not only the time to develop these new AI techniques, but also the time for the humans now riding the wave of machine learning and LLMs to accept that it might be worth their time to look into alternative approaches (more so after extensive efforts to trivialize these approaches as a loss of time). It is the second factor which I think will be the true time bottleneck and could push the resolution of this question further in time."

"Given the progress on Tier 3 FrontierMath problems, a Millennium Problem seems well away, notwithstanding bullish predictions from corporate spokespeople with vested interests."

"I would put 60% as some hard limit on whether any of the conjectures can be solved at all."

"I don't think there are many economic incentives to develop those kinds of systems. Millennium problems are very, very hard - much harder than most directly economically useful tasks. They require developing new mathematical theories and techniques to even approach them. As far as I know, current top AI models lack this ability, and I don't see an easy way for them to obtain such an ability (nor are there many economic incentives for building such abilities into them)."

Sensitivity of Results to Reweighting

As described in the Reweighting section, we use a standard approach in the public polling field, raking, to weight aggregate statistics to be representative of the sampling frames. We perform a sensitivity analysis to understand the impact of weighting on aggregate forecasts. We compare the median aggregated forecast to the weighted median aggregated forecast, where positive values of differences indicate that weighting participant responses increases the value of a forecast.

We begin by expressing the difference between the weighted median and unweighted median at the question level as a proportion of the forecast dispersion. This difference is divided by the standard deviation of the unweighted forecasts, within each question-participant group, to standardize the impact across questions with different units. To examine the full distribution of reweighting effects, we calculate this standardized difference for multiple aggregate statistics—including the 25th percentile, median, and 75th percentile of the standardized differences themselves. For example, a 25th percentile value of -0.07 indicates that across all questions, 25% of the standardized reweighting effects fall below -0.07. Table 7 below shows a summary of these differences by survey for experts and the public.

Survey	Participant	Min	p25	Median	Mean	p75	Max
Wave 1: Headliners	Expert	-0.24	-0.07	0	-0.03	0.01	0.2
Wave 1: Headliners	Public	-0.2	0	0	0	0	0.12
Wave 2: AI for science	Expert	-0.38	0	0	-0.02	0	0.19
Wave 2: AI for science	Public	-0.18	0	0.01	0.03	0.06	0.15
Wave 3: Broad Adoption of AI	Expert	-0.23	-0.09	0	-0.03	0	0.28
Wave 3: Broad Adoption of AI	Public	-0.21	0	0	-0.01	0	0.08

Table 7: Summary standardized reweighting effects. The difference between weighted and unweighted medians, expressed in standard deviation units, is summarized across surveys and participant groups.

These values show how much reweighting shifts the forecast relative to the dispersion of forecasts. For example, a value of 0.1 means the weighted forecast is 0.1 standard deviations higher than the unweighted forecast. This table shows that the median effect on both participant types and across all surveys is no change in the aggregate result. Reweighting has a marginally larger effect on expert participants than members of the public. This histogram below shows the effect of reweighting on all the forecasts in more detail.

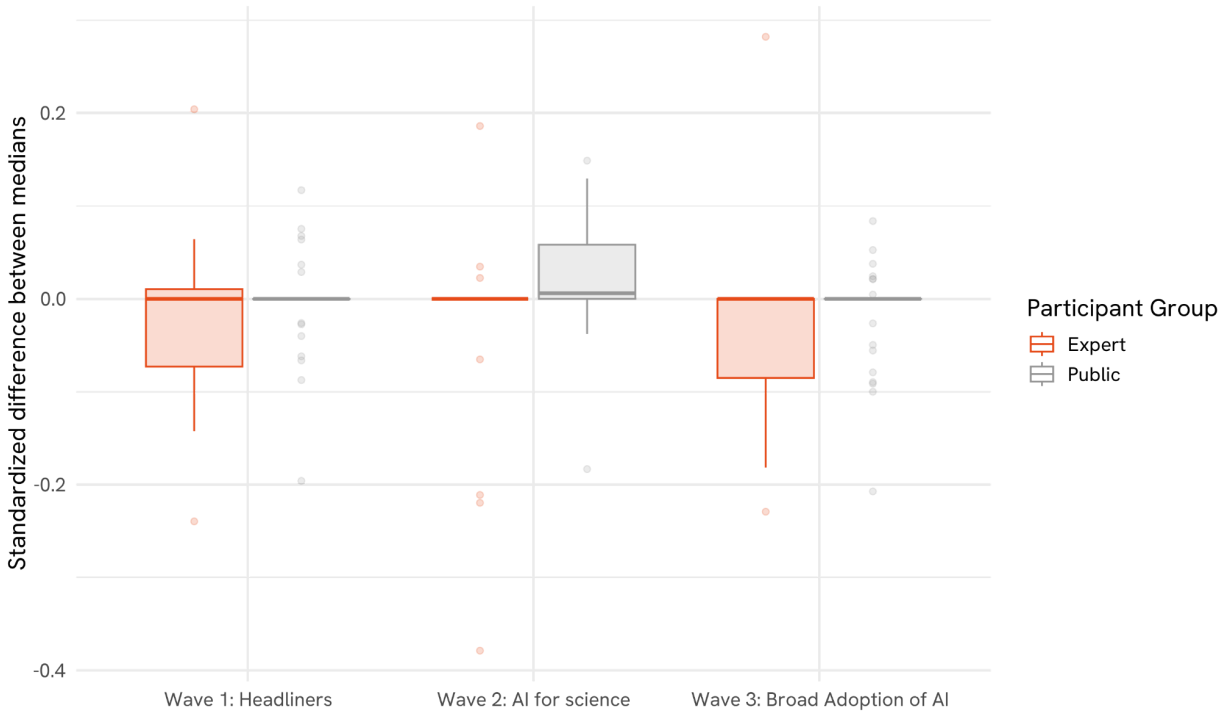


Figure 11: Histogram showing standardised difference between weighted and unweighted aggregate median

Public Accuracy Stratification

Given that we select for experience in AI in our expert sample but not in our public sample, public forecasts could be distorted by either a lack of comprehension or effort. To investigate this concern, we partition our sample by forecasting accuracy on out-of-sample, prior questions. For the 832 participants we can match to a prior forecasting record,[114] we calculate accuracy scores based on performance on 24 forecasting questions asking about near term (<6 months) events in an earlier research project (Barker et al. 2025).[115] We score these forecasts using an S-score, with a lower score indicating better performance.

Public participants are split into two accuracy groups: “High-accuracy,” representing the most accurate 50% of public participants, and “low-accuracy” representing the least accurate 50% of public participants. We do not adjust weights after partitioning the public sample.

Of 68 total forecasts with a clear valence of AI capabilities, the results are mixed. The high-accuracy group holds views about AI progress, capabilities, and diffusion that are

[114] Recall from Public Sampling that our public sample consists largely of highly engaged, past participants from FRI work.
[115] Examples of questions include: (1) What will be the National Average Temperature Rank for May 2025 in the contiguous United States, according to NOAA's Climate Data Center, where 1 is the coolest and 130 is the highest rank representing the warmest on record?; and (2) What will be the closing stock price of Meta on 30 May 2025?

statistically indistinguishable from the low-accuracy group in half of all cases. They predict less progress in 28% of cases and more in the remaining 22% of forecasts. We summarize some of the differences in aggregate forecasts in Figure 12.

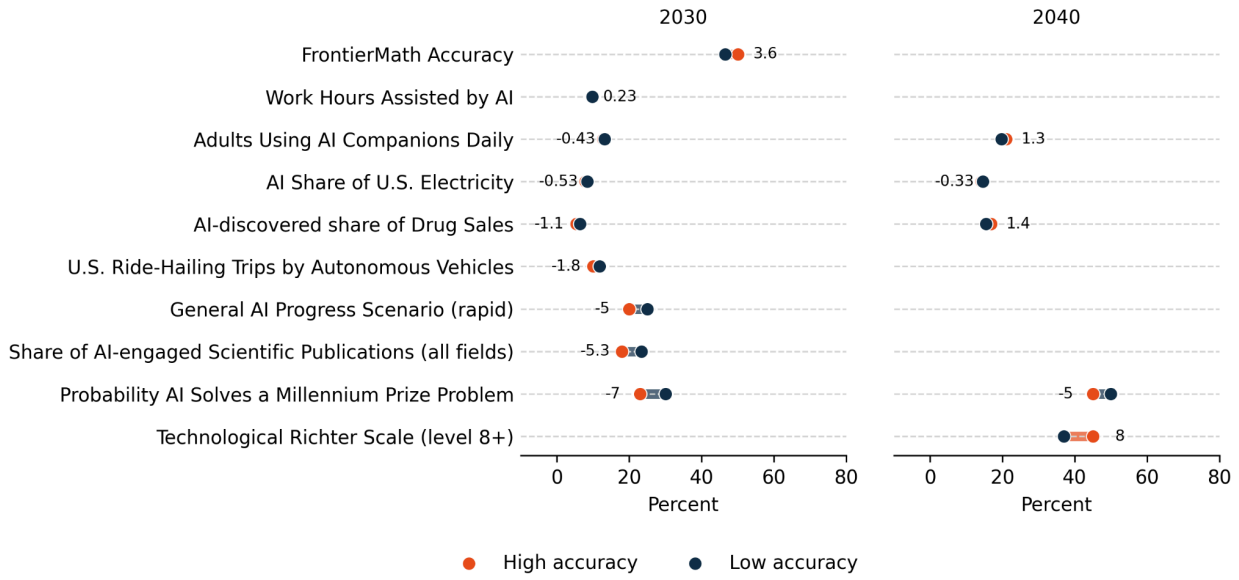


Figure 12: Differences between the high- and low-accuracy public median 50th percentile forecasts for several questions where the unit is a percentage. Points indicate the median of each groups' 50th percentile forecasts. We apply transformations to create valenced forecasts, where values closer to the left indicate slower progress and values to the right indicate faster progress.

We present some additional results in Appendix D. Public Accuracy Stratification. We plan to explore the drivers of these disagreements in future work.

Next Steps

This paper reports results from the first three monthly waves of LEAP and describes the project methodology in detail. We will elicit forecasts roughly each month for the next three years on timely topics related to the development, capabilities, adoption, and impact of AI. Both the set of forecasting questions and the space of possible analyses will expand over time: additional questions grow the number of datapoints we can test and learn from, and progressive question resolution enables a continued sharpening of accuracy measurement. We plan to release reports on each new wave of LEAP soon after we complete data collection. Additionally, we will periodically release more detailed reports and papers conducting more extensive and cross-wave analysis. We will discuss some of those analyses below.

Future waves of LEAP will be focused on forecasts related to topics including security and geopolitics, robotics, labor and automation, incidents and harms, and AI safety. For example, we tentatively plan to ask forecasters to predict how much AI will improve the productivity of software engineers, how AI will affect the productivity of workers, and how the use of AI may

cause harm. We welcome reader suggestions for LEAP questions or wave themes via outreach to our project team.[116] We now describe planned follow-up work for LEAP.

Accuracy Assessment

As questions resolve, we will be assessing the accuracy of forecasts to identify particularly accurate individual forecasters within the expert, superforecaster, and public groups, and to assess the relative accuracy of the different expert subgroups within our sample. We will also present forecasters with information on their own past performance, assessing how this feedback translates into accuracy on new forecasting questions. Kučinskas et al. (2025) perform a similar accuracy analysis, retrospectively analyzing a forecasting study focused on multi-year forecasts about AI-, nuclear-, climate-, and biotechnology-related progress and risks, finding no correlation between short-term accuracy and long-term beliefs in their study context, first introduced in Karger et al. (2023). LEAP is, in many ways, a follow-up to that project, improving on several choices the authors made: LEAP requires that forecasters answer all (or almost all) questions, elicits one-time forecasts, and does not provide a team structure or room for deliberation, since debate in that survey did not generally prove to resolve disagreements (Karger et al. 2025). We will evaluate these findings with significantly more precision in LEAP as questions resolve in 2027 and 2030.

Forecast Updating

We plan to re-survey the LEAP sample on many of our questions to track how respondents' views evolve over time. For example, one year from now, how will respondents update their forecasts of whether there will be an AI-reliant solution to a Millennium Prize Problem? How will progress (or a lack of progress) for AI systems on key benchmarks change respondents' views of longer-run effects of AI?

Schools of Thought Analysis and Crux-Finding

A "school of thought" is a cluster of similar responses to a set of forecasting questions. In future work, we will use standard clustering techniques to search for similar groups of forecasters across questions, and we will complement this work with analysis of qualitative information (rationales) for these various schools of thought. Are we able to distinguish consistent differences in sets of forecasts among subsets of our sample—for example, fast versus slow AI progress groups? The search for schools of thought in our forecasting data takes a related but opposite approach to our prior work using adversarial collaborations to identify cruxes for differences of opinion about the likelihood of harms from AI, in which we search for individuals from distinct schools-of-thought and then ask them to forecast on questions about AI on which they disagreed (Rosenberg et al. 2024). Both approaches can help us map beliefs about AI.

[116] Survey: https://airtable.com/appGCchUyUTPvT90e/pagPpnUX2SiiTUNpp/form. Project Team Contact: leap@forecastingresearch.org.

Relatedly, are we able to identify "cruxes"—i.e., strongly differential forecasts between schools of thought on near-term questions that enable faster assessment of which school is more likely to be accurate in the long term? This crux-finding effort builds on our earlier work (McCaslin et al. 2024; Rosenberg et al. 2024; Rosenberg et al. 2025), which finds that disagreement on the likelihood of extreme outcomes from AI is not easily resolved by debate, but we can identify nearer-term cruxes that could increase consensus.

Elicitation Experiments

We have already tested the impact of providing various defaults in the interactive forecasting interfaces (see Appendix B.V. Survey instrument) and the ordering of options within questions; we plan to present these results in a future report. Additionally, we plan to experiment with question wording, question ordering, and more.

Expanded Use of Rationale Data

We are exploring scalable and privacy-preserving methods for directly displaying a subset of rationales for high and low forecasts on particular questions, as part of our forecast explorer.[117] We will provide detailed rationale analyses (similar to the examples above) for all LEAP questions, available in our monthly reports. We are also awarding prizes to respondents for rationale quality, and we will highlight publicly some particularly high-quality rationales.

Public Engagement

We may experiment with broader public engagement on AI forecasting. For example, we may enable anyone to make their own forecasts on LEAP questions, and provide them with a report on where their forecasts fit among different schools of thought.

Conclusion

Policymakers, nonprofit and business leaders, and other stakeholders routinely consult experts to base their decisions on the perspective of experts, especially when faced with new technologies and high levels of uncertainty. While public discussion of AI and expert anecdotes are widespread, structured quantitative evidence on expert beliefs is lacking, impeding effective decision making. With the launch of LEAP, we fill an important gap by both measuring the full range of expert opinions on AI capability developments and their impact, and by capturing the underlying reasoning and evidence that supports these beliefs.

We have completed three survey waves focused on (1) high-level predictions about AI progress; (2) the application of AI to scientific discovery; and (3) widespread adoption and social impact. The first three rounds of LEAP reveal five key findings.

[117] See https://leap.forecastingresearch.org/forecasts.

First, collectively experts expect sizable societal effects from AI by 2040, even if effects materialize more slowly than expected.

Second, and in contrast with the first takeaway, considerable disagreement across experts, and uncertainty within individual experts, underlies these predictions of progress. This dynamic likely arises from multiple sources: the inherent difficulty of forecasting emerging technologies, sharp disagreements between competing schools of thought, and the fundamental uncertainty surrounding AI development and its impact. Forecasting emerging technologies is inherently difficult. For example, historical predictions about fusion power have consistently proven overoptimistic (Takeda et al. 2023). In our own prior work, both domain experts and superforecasters substantially underestimated AI progress (Kučinskas et al. 2025). Nevertheless, aggregate forecasts remain informative and offer the potential to cut through the noise of disagreement, as wisdom-of-the-crowd effects have proven robust across domains. As the project progresses and forecasting questions resolve, LEAP will evaluate the performance of aggregate and individual forecasts.

Third, expert predictions diverge substantially from the timelines articulated by frontier lab executives, with our median expert anticipating considerably slower progress. LEAP provides an expert view free from the potential distorting effects of financial incentives that may influence public statements from industry leaders. While the historical record will be the ultimate yardstick for these predictions, LEAP helps us understand what a broader swath of experts expect from AI.

Fourth, experts generally forecast faster AI progress than the public across most outcomes, and LEAP will continue to track the evolution of expert and public opinion, especially as the technology begins to be more front-of-mind for the public.

Lastly, we observe consistency in predictions between superforecasters and experts; in the instances where their views diverge, experts tend to predict somewhat faster AI progress. Importantly, we also find substantial consistency across our four categories of experts: computer scientists, economists, industry professionals, and policy professionals. The forthcoming resolution of near-term predictions will reveal whether specialized domain knowledge or general forecasting skill proves more valuable for predicting AI trajectories—a question with significant implications for weighing different sources of expertise in technology policy decisions.

Although we designed LEAP to overcome the major challenges that confront AI forecasting efforts, there are still some clear limitations of this work.

First, it is difficult to generate a sample frame that is representative of any key group of experts, and nonresponse bias is difficult to avoid, potentially biasing results. We construct comprehensive sampling frames of experts to minimize coverage bias; however, it is possible that there are experts who hold views different from those found in our frames. We reweight our data based on frame demographics to reduce nonresponse bias, but our set of target variables

might not capture all the variation in opinion. These sources of bias may affect the representativeness of our results.

In addition, while we have taken great care in constructing clear, specific, and resolvable questions, some questions contain inherent ambiguity and for others, the discontinuation or change of a data source may preclude resolution. It is also possible that, going forward, attrition will affect our ability to measure how views on AI progress and diffusion change over time; indeed, leading research organizations often experience high attrition and low response rates (NORC AmeriSpeak 2024). We will monitor attrition from LEAP to ensure sufficient sample sizes in future waves as well as to understand if attrition may be biasing our results.

Finally, survey participants typically have limited time for surveys. LEAP addresses this through three strategies: providing historical context and background for each question, offering interactive interfaces with baseline data to streamline forecasting, and providing significant compensation to participants for their time. These measures contribute to considerable effort by participants—the median expert took 44 minutes per survey, the median member of the public 29 minutes, and the median superforecaster 90 minutes. But, sharing background and baseline information among all participants reduces their independence and may dilute the wisdom-of-the-crowd effect, creating correlated forecasts or echo chambers. Additionally, it remains possible that some participants will not put effort into reporting their true beliefs on each question, speeding through the survey because of time constraints or disinterest. Unlike previous studies that used fixed payments regardless of accuracy, LEAP employs proper scoring rules that link compensation to forecast quality based on clear resolution criteria, in an effort to reduce this risk.

LEAP will continue to explore important questions regarding the future of AI. Public, high-profile proclamations about the technology are not necessarily representative of expert opinion, and we will search for agreement and disagreement among experts, the general public, and professional forecasters. As LEAP forecasting questions begin to resolve as early as the end of 2025, we will assess how short-run accuracy on AI-related questions correlates with long-run AI-related beliefs as we try to bring clarity to the many current high-stakes debates about AI.

References

Acemoglu, Daron. “Don’t Believe the AI Hype.” Project Syndicate, May 21, 2024. https://www.project-syndicate.org/commentary/ai-productivity-boom-forecasts-countered-by-theory-and-data-by-daron-acemoglu-2024-05.

Adjodah, Dhaval, Yan Leng, Shi Kai Chong, P. M. Krafft, Esteban Moro, and Alex Pentland. “Accuracy-Risk Trade-Off Due to Social Learning in Crowd-Sourced Financial Predictions.” *Entropy* 23, no. 7 (2021): 801. https://doi.org/10.3390/e23070801.

Aghion, Philippe, Benjamin F. Jones, and Charles I. Jones. “Artificial Intelligence and Economic Growth,” In *The Economics of Artificial Intelligence: An Agenda*, ed. by Ajay K. Agrawal, Joshua S. Gans, and Avi Goldfarb (University of Chicago Press, 2019). https://www.nber.org/books-and-chapters/economics-artificial-intelligence-agenda/artificial-intelligence-and-economic-growth.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. “Artificial Intelligence: The Ambiguous Labor Market Impact of Automating Prediction.” *The Journal of Economic Perspectives* 33, no. 2 (2019): 31–50. https://www.jstor.org/stable/26621238.

Agrawal, Ajay, Joshua S. Gans, and Avi Goldfarb. “Do We Want Less Automation?” *Science* 381, no. 6654 (2023): 155–58. https://doi.org/10.1126/science.adh9429.

Allen, Mike, and Jim VandeHei. “Behind the Curtain: Top AI CEO Foresees White-Collar Bloodbath.” Axios, May 28, 2025. https://www.axios.com/2025/05/28/ai-jobs-white-collar-unemployment-anthropic.

Altman, Sam. “The Gentle Singularity.” June 10, 2025. https://blog.samaltman.com/the-gentle-singularity.

Amodei, Dario. *Machines of Loving Grace*. October 2024. https://www.darioamodei.com/essay/machines-of-loving-grace#3-economic-development-and-poverty.

Ansede, Manuel. “Spanish Mathematician Javier Gómez Serrano and Google DeepMind Team up to Solve the Navier-Stokes Million-Dollar Problem.” EL PAÍS English, June 24, 2025. https://english.elpais.com/science-tech/2025-06-24/spanish-mathematician-javier-gomez-serrano-and-google-deepmind-team-up-to-solve-the-navier-stokes-million-dollar-problem.html.

Bajekal, Naina. “The 100 Most Influential People in AI 2023.” Time, September 7, 2023. https://time.com/collection/time100-ai/.

Barker, Jessie, Benjamin, Rhiannon Britt, Mark Himmelstein, David Budescu, Ezra Karger. “Team Dynamics in Forecasting.” Forecasting Research Institute Working Paper, 2025.

Barker Bonomo, Emma, and Ayesha Javed. “The 100 Most Influential People in AI 2024.” TIME, September 5, 2024. https://time.com/collection/time100-ai-2024/.

Baron, Jonathan, Barbara A. Mellers, Philip E. Tetlock, Eric Stone, and Lyle H. Ungar. “Two Reasons to Make Aggregated Probability Forecasts More Extreme.” *Decision Analysis* 11, no. 2 (2014): 133–45. https://doi.org/10.1287/deca.2014.0293.

Bassamboo, Achal, Ruomeng Cui, and Antonio Moreno. “The Wisdom of Crowds in Operations: Forecasting Using Prediction Markets.” *SSRN Electronic Journal*, ahead of print, October 27, 2015. https://doi.org/10.2139/ssrn.2679663.

Berger, Emery D. *CSRankings*. Accessed June 4, 2025. https://csrankings.org.

Bick, Alexander, Adam Blandin, and David Deming. “The Rapid Adoption of Generative AI.” Nos. 2024–027. Federal Reserve Bank of St. Louis, 2025. https://doi.org/10.20955/wp.2024.027.

Brier, Glenn W. “Verification of Forecasts Expressed in Terms of Probability.” *Monthly Weather Review* 78, no. 1 (1950): 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Brookings Institution. “Artificial Intelligence and Emerging Technology Initiative.” Accessed October 20, 2025. https://www.brookings.edu/projects/artificial-intelligence-and-emerging-technology-initiative/.

Brynjolfsson,Erik, Bharat Chandar, and Ruyu Chen. “Canaries in the Coal Mine? Six Facts about the Recent Employment Effects of Artificial Intelligence.” 2025. https://digitaleconomy.stanford.edu/publications/canaries-in-the-coal-mine/.

Bubeck, Sebastien (@SebastienBubeck). “Claim: gpt-5-pro can prove new interesting mathematics. Proof: I took a convex optimization paper with a clean open problem in it and asked gpt-5-pro to work on it. It proved a better bound than what is in the paper, and I checked the proof it's correct.” X, August 20, 2025. https://x.com/SebastienBubeck/status/1958198661139009862

Chandar, Bharat. “Tracking Employment Changes in AI-Exposed Jobs.” SSRN Scholarly Paper No. 5384519. Social Science Research Network, June 3, 2025. https://doi.org/10.2139/ssrn.5384519.

Chasalow, Kyla, and Karen Levy. “Representativeness in Statistics, Politics, and Machine Learning.” arXiv:2101.03827. Preprint, arXiv, February 10, 2021. https://doi.org/10.48550/arXiv.2101.03827.

Clark Center. “US Economic Experts Panel.” *The Clark Center for Global Markets.* Accessed October 20, 2025. https://kentclarkcenter.org/us-economic-experts-panel/.

Clay Mathematics Institute. “The Millennium Prize Problems.” Accessed November 6, 2025. https://www.claymath.org/millennium-problems/.

Council on Foreign Relations. “CEO Speaker Series with Dario Amodei of Anthropic.” Council on Foreign Relations, March 10, 2025. https://www.cfr.org/event/ceo-speaker-series-dario-amodei-anthropic.

"Crunchbase." Accessed April 8, 2025. https://www.crunchbase.com/.

Curran, Edna and Mark Niquette. "AI-Led Investments Are Driving US Economic Growth." *Bloomberg.* October 31, 2025. https://www.bloomberg.com/news/articles/2025-10-31/ai-boom-drives-us-gdp-growth-higher-stock-prices.

Cvitanić, Jakša, Dražen Prelec, Blake Riley, and Benjamin Tereick. "Honesty via Choice-Matching." *American Economic Review: Insights* 1, no. 2 (2019): 179–92. https://doi.org/10.1257/aeri.20180227.

Da, Zhi, and Xing Huang. "Harnessing the Wisdom of Crowds." *Management Science* 66, no. 5 (2020): 1847–67. https://doi.org/10.1287/mnsc.2019.3294.

Dai, Ji, Huiyu Xu, Tao Chen, et al. "Artificial Intelligence for Medicine 2025: Navigating the Endless Frontier." *The Innovation Medicine* Vol 3 (2025). Accessed October 31, 2025. https://the-innovation.org/article/doi/10.59717/j.xinn-med.2025.100120.

Davis-Stober, Clintin P., David V. Budescu, Jason Dana, and Stephen B. Broomell. "When is a Crowd Wise?" *Decision* (US) 1, no. 2 (2014): 79–101. https://doi.org/10.1037/dec0000004.

Douthat, Ross. "The Next Economic Bubble Is Here." *The New York Times*, October 23, 2025. https://www.nytimes.com/2025/10/23/opinion/ai-bubble-economy-bust.html.

D'Souza, Faisal. "Re: Request for Information (RFI) on the Development of an Artificial Intelligence (AI) Action Plan ('Plan')." Anthropic (public comment letter), March 6, 2025. https://assets.anthropic.com/m/4e20a4ab6512e217/original/Anthropic-Response-to-OSTP-RFI-March-2025-Final-Submission-v3.pdf.

Duede, Eamon, William Dolan, André Bauer, Ian Foster, and Karim Lakhani. "Oil & Water? Diffusion of AI Within and Across Scientific Fields." Preprint, arXiv, 2024. https://doi.org/10.48550/ARXIV.2405.15828.

Eckhardt, Sarah, and Nathan Goldschlag. "AI and Jobs: The Final Word (Until the Next One)." *Economic Innovation Group*, August 10, 2025. https://eig.org/ai-and-jobs-the-final-word/.

EconBiz. "Journals, Working Papers & Conferences in Business Studies and Economics." Accessed May 20, 2025. https://www.econbiz.de/

Edwards, Jim. "The AI boom is unsustainable unless tech spending goes 'parabolic,' Deutsche Bank warns: 'This is highly unlikely.'" Fortune, September 23, 2025. https://fortune.com/2025/09/23/ai-boom-unsustainable-tech-spending-parabolic-deutsche-bank/

Epoch AI. "About FrontierMath." 2025. https://epoch.ai/frontiermath/about.

Epoch AI. "AI Benchmarking." Accessed November 22, 2024. https://epoch.ai/benchmarks.

Epoch AI. "Data on AI Models." Accessed November 22, 2024. https://epoch.ai/data/ai-models.

Federal Reserve Bank of St. Louis. “NBER based Recession Indicators for the United States from the Period following the Peak through the Trough [USREC].” FRED, Federal Reserve Bank of St. Louis. Accessed November 6, 2025. https://fred.stlouisfed.org/series/USREC.

Gerut, Amanda. “JPMorgan CEO Jamie Dimon Says People Who Don’t Think Job Losses Due to AI Are Inevitable, ‘Should Stop Sticking Their Head in the Sand.’” Fortune, October 15, 2025. https://fortune.com/2025/10/15/jpmorgan-ceo-jamie-dimon-ai-job-losses/.

Gimbel, Martha, Molly Kinder, Joshua Kendall, and Maddie Lee. “Evaluating the Impact of AI on the Labor Market: Current State of Affairs.” The Budget Lab at Yale, October 1, 2025. https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs.

Good Judgment Inc. “The First Championship Season.” Accessed October 30, 2025. https://goodjudgment.com/resources/the-superforecasters-track-record/the-first-championship-season/.

Grace, Katja, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. “Viewpoint: When Will AI Exceed Human Performance? Evidence from AI Experts.” *Journal of Artificial Intelligence Research* 62 (July 2018): 729–54. https://doi.org/10.1613/jair.1.11222.

Grace, Katja, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner. “Thousands of AI Authors on the Future of AI.” *Journal of Artificial Intelligence Research*, January 2024. https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf.

Hassabis, Demis. “What’s next for AI at DeepMind, Google’s Artificial Intelligence Lab.” Interviewed by Scott Pelley. *60 Minutes*, CBS. April 20, 2025. https://www.youtube.com/watch?v=1XF-NG_35NE

Heaven, Will Douglas. “What’s next for AI and math.” MIT Technology Review, June 4, 2025. https://www.technologyreview.com/2025/06/04/1117753/whats-next-for-ai-and-math/

Himmelstein, Mark, David Budescu, and Ying Han. “The Wisdom of Timely Crowds.” In *Judgment in Predictive Analytics*, ed. M. Seifert. International Series in Operations Research & Management Science, vol 343. Springer, Cham. https://doi.org/10.1007/978-3-031-30085-1_8.

Himmelstein, Mark, David V. Budescu, and Emily H. Ho. “The Wisdom of Many in Few: Finding Individuals Who Are as Wise as the Crowd.” *Journal of Experimental Psychology: General* (US) 152, no. 5 (2023): 1223–44. https://doi.org/10.1037/xge0001340.

Horizon Institute for Public Service.“Think Tanks.” Emerging Tech Policy Careers, accessed May 22, 2025. https://emergingtechpolicy.org/institutions/think-tanks/.

Hueffer, Karsten, Miguel A. Fonseca, Anthony Leiserowitz, and Karen M. Taylor. “The Wisdom of Crowds: Predicting a Weather and Climate-Related Event.” *Judgment and Decision Making* 8, no. 2 (2023): 91–105. https://doi.org/10.1017/S1930297500005039.

Intelligence Advanced Research Projects Activity - ODNI. “IARPA ACE Program.” https://www.iarpa.gov/research-programs/ace.

Jose, Victor Richmond R., and Robert L. Winkler. “Evaluating Quantile Assessments.” *Operations Research* 57, no. 5 (2009): 1287–97. https://doi.org/10.1287/opre.1080.0665.

Karger, Ezra, Joshua Monrad, Barbara Mellers, and Philip Tetlock. “Reciprocal Scoring: A Method for Forecasting Unanswerable Questions.” *SSRN Electronic Journal*, ahead of print, 2021. https://doi.org/10.2139/ssrn.3954498.

Karger, Ezra, Josh Rosenberg, Zachary Jacobs, Molly Hickman, and Philip E. Tetlock. “Subjective-Probability Forecasts of Existential Risk: Initial Results from a Hybrid Persuasion-Forecasting Tournament.” *International Journal of Forecasting* 41, no. 2 (2025): 499–516. https://doi.org/10.1016/j.ijforecast.2024.11.008.

Karger, Ezra, Josh Rosenberg, Zachary Jacobs, et al. *Forecasting Existential Risks: Evidence from a Long-run Forecasting Tournament*. Forecasting Research Institute, 2023. https://forecastingresearch.org/s/XPT.pdf.

Kay, Carly. “AI-Powered CRISPR Could Lead to Faster Gene Therapies, Stanford Medicine Study Finds.” Stanford Medicine Magazine, September 16, 2025. https://med.stanford.edu/news/all-news/2025/09/ai-crispr-gene-therapy.html.

Kučinskas, Simas, Josh Rosenberg, Rebecca Ceppas de Castro, et al. *Assessing Near-Term Accuracy in the Existential Risk Persuasion Tournament*. Forecasting Research Institute, September 2025. https://forecastingresearch.org/near-term-xpt-accuracy.

Kwa, Thomas, Ben West, Joel Becker, et al. “Measuring AI Capability to Complete Long Tasks.” March 30, 2025. https://doi.org/10.48550/arXiv.2503.14499

Lichtendahl Jr., Kenneth C., and Robert L. Winkler. “Probability Elicitation, Scoring Rules, and Competition among Forecasters.” *Management Science* (US) 53, no. 11 (2007): 1745–55. https://doi.org/10.1287/mnsc.1070.0729.

Lichtendahl, Kenneth C., Yael Grushka-Cockayne, and Phillip E. Pfeifer. “The Wisdom of Competitive Crowds.” *Operations Research* 61, no. 6 (2013): 1383–98. https://doi.org/10.1287/opre.2013.1213

Lichtendahl, Kenneth C., Yael Grushka-Cockayne, and Robert L. Winkler. “Is It Better to Average Probabilities or Quantiles?” *Management Science* 59, no. 7 (2013): 1594–611. https://doi.org/10.1287/mnsc.1120.1667.

LMArena. “Overview Leaderboard | LMArena.” Accessed November 21, 2024. https://lmarena.ai/leaderboard.

Luong, Thang and Edward Lockhart. “Advanced version of Gemini with Deep Think officially achieves gold-medal standard at the International Mathematical Olympiad.” Google DeepMind, July 21, 2025. https://deepmind.google/blog/advanced-version-of-gemini-with-deep-think-officially-achieves-gold-medal-standard-at-the-international-mathematical-olympiad/

Maese, Ellen. “Americans Use AI in Everyday Products Without Realizing It.” Gallup, January 14, 2025. https://news.gallup.com/poll/654905/americans-everyday-products-without-realizing.aspx.

McCaslin, Tegan, Josh Rosenberg, Ezra Karger, et al. *Conditional Trees: A Method for Generating Informative Questions about Complex Topics*. Forecasting Research Institute, August 12, 2024. https://forecastingresearch.org/ai-conditional-trees

McClain, Colleen, Brian Kennedy, Jeffrey Gottfried, Monica Anderson, and Giancarlo Pasquini. “How the U.S. Public and AI Experts View Artificial Intelligence.” *Pew Research Center*, April 3, 2025. https://www.pewresearch.org/internet/2025/04/03/how-the-us-public-and-ai-experts-view-artificial-intelligence/.

McGann, James G. *2020 Global Go To Think Tank Index Report*. University of Pennsylvania, Think Tanks and Civil Societies Program, 2021. https://guides.library.upenn.edu/publicpolicyresearchthinktanks/publicpolicyresearchthinktanks

MacKenzie, Donald. “Slaying the Kraken: The Sociohistory of a Mathematical Proof.” *Social Studies of Science* 29, no. 1 (1999): 7–60. http://www.jstor.org/stable/285445.

Mellers, Barbara, Eric Stone, Terry Murray, et al. “Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions.” *Perspectives on Psychological Science* 10, no. 3 (2015): 267–81. https://doi.org/10.1177/1745691615577794.

Mellers, Barbara, Lyle Ungar, Jonathan Baron, et al. “Psychological Strategies for Winning a Geopolitical Forecasting Tournament.” *Psychological Science* 25, no. 5 (2014): 1106–15. https://doi.org/10.1177/0956797614524255.

Mercer, Andrew, Arnold Lau, and Courtney Kennedy. “How Different Weighting Methods Work.” *Pew Research Center*, January 26, 2018. https://www.pewresearch.org/methods/2018/01/26/how-different-weighting-methods-work/.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. “Eliciting Informative Feedback: The Peer-Prediction Method.” *Management Science* 51, no. 9 (2005): 1359–73. https://doi.org/10.1287/mnsc.1050.0379.

Müller, Vincent C., and Nick Bostrom. “Future Progress in Artificial Intelligence: A Poll among Experts.” *AI Matters* 1, no. 1 (2014): 9–11. https://doi.org/10.1145/2639475.2639478.

Murr, Andreas Erwin. "'Wisdom of Crowds'? A Decentralised Election Forecasting Model That Uses Citizens' Local Expectations." *Electoral Studies* 30, no. 4 (2011): 771–83. https://doi.org/10.1016/j.electstud.2011.07.005.

Musk, Elon (@elonmusk). "@davidpattersonx Your estimates are about right. However, intelligent robots in humanoid form will far exceed the population of humans, as every person will want their own personal R2-D2 and C-3PO. And then there will be many robots in industry for every human to provide products & services." X, August 24, 2025. https: //x.com/elonmusk/status/1959667978158395737.

Musk, Elon (@elonmusk). " It is increasingly likely that AI will superset the intelligence of any single human by the end of 2025 and maybe all humans by 2027/2028. Probability that AI exceeds the intelligence of all humans combined by 2030 is ˜100%." X, December 23, 2024. https://x.com/elonmusk/status/1871083864111919134.

Nadeem, Reem. "The Partisanship and Ideology of American Voters." *Pew Research Center*, April 9, 2024. https://www.pewresearch.org/politics/2024/04/09/the-partisanship-and-ideology-of-american-voters/.

NORC Amerispeak. *Response Rate Transparency in U.S. Probability Household Panels*. October 2024. https://amerispeak.norc.org/content/dam/amerispeak/supporting-documents/response-rate-transparency-in-panels.pdf.

Obama, Barack (@BarackObama). "At a time when people are understandably focused on the daily chaos in Washington, these articles describe the rapidly accelerating impact that AI is going to have on jobs, the economy, and how we live. https://t.co/RSbMkhz3Xm." X, May 30, 2025. https://x.com/BarackObama/status/1928568801232138423.

O'Donovan, Cian, Sarp Gurakan, Xiaomeng Wu, et al. *Visions, Values, Voices: A Survey of Artificial Intelligence Researchers*. Zenodo, 2025. https://doi.org/10.5281/ZENODO.15080287.

O'Donovan, Peter, Kevin Leahy, Ken Bruton, and Dominic T. J. O'Sullivan. "Big Data in Manufacturing: A Systematic Mapping Study." *Journal of Big Data* 2, no. 1 (2015): 20. https://doi.org/10.1186/s40537-015-0028-x.

Office of the Surgeon General. *Our Epidemic of Loneliness and Isolation: The U.S. Surgeon General's Advisory on the Healing Effects of Social Connection and Community*. Washington (DC): US Department of Health and Human Services; 2023. PMID: 37792968. https://pubmed.ncbi.nlm.nih.gov/37792968/

OpenAlex. "OpenAlex." Accessed October 20, 2025. https://openalex.org/.

Our World in Data. "Annual Private Investment in Artificial Intelligence." Accessed November 3, 2025. https://ourworldindata.org/grapher/private-investment-in-artificial-intelligence.

Panchal, Nirmita, and Justin Lo. "Exploring the Rise in Mental Health Care Use by

Demographics and Insurance Status." KFF, August 1, 2024. https://www.kff.org/mental-health/exploring-the-rise-in-mental-health-care-use-by-demographics-and-insurance-status/.

Paper Copilot. *Papercopilot/Paperlists*. Accessed October 17, 2025. https://github.com/papercopilot/paperlists.

Pew Research Center. "The American Trends Panel." *Pew Research Center*, 2024. https://www.pewresearch.org/the-american-trends-panel/.

Prelec, Dražen. "A Bayesian Truth Serum for Subjective Data." *Science* 306, no. 5695 (2004): 462–66. https://doi.org/10.1126/science.1102081.

Priem, Jason, Heather Piwowar, and Richard Orr. "OpenAlex: A Fully-Open Index of Scholarly Works, Authors, Venues, Institutions, and Concepts." arXiv:2205.01833. Preprint, arXiv, June 17, 2022. https://doi.org/10.48550/arXiv.2205.01833.

Quorum Research. "Quorum Research." Accessed October 31, 2025. https://quorumresearch.com.

Ranjan, Roopesh, and Tilmann Gneiting. "Combining Probability Forecasts." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 72, no. 1 (2010): 71–91. https://doi.org/10.1111/j.1467-9868.2009.00726.x.

RePEc. "IDEAS." Accessed March 7, 2025. https://ideas.repec.org/.

Rose, Steve. "Demis Hassabis on Our AI Future: 'It'll be 10 Times Bigger than the Industrial Revolution – and Maybe 10 Times Faster.'" Technology. *The Guardian*, August 4, 2025. https://www.theguardian.com/technology/2025/aug/04/demis-hassabis-ai-future-10-times-bigger-than-industrial-revolution-and-10-times-faster.

Rosenberg, J., Karger, E., Jacobs, Z., et al. "Belief updating in AI-risk debates: Exploring the limits of adversarial collaboration." *Risk Analysis*, 1–17 (2025). https://doi.org/10.1111/risa.70023

Rosenberg, Josh, Ezra Karger, Avital Morris, et al. *Roots of Disagreement on AI Risk: Exploring the Potential and Pitfalls of Adversarial Collaboration*. Forecasting Research Institute, March 11 2024. https://forecastingresearch.org/ai-adversarial-collaboration.

Ruggles, Steven, Sarah Flood, Matthew Sobek, et al. "IPUMS USA: Version 16.0." With United States Census Bureau. Minneapolis, MN: IPUMS, 2025. https://doi.org/10.18128/D010.V16.0.

Russell, Regina G., Laurie Lovett Novak, Mehool Patel, et al. "Competencies for the Use of Artificial Intelligence-Based Tools by Health Care Professionals." *Academic Medicine: Journal of the Association of American Medical Colleges* 98, no. 3 (2023): 348–56. https://doi.org/10.1097/ACM.0000000000004963.

S&P Global. "Layoffs surge in US white collar jobs as rates, AI alter office work." October 31, 2024.

https://www.spglobal.com/market-intelligence/en/news-insights/articles/2024/11/layoffs-surge-in-us-white-collar-jobs-as-rates-ai-alter-office-work-85986794

Silver, Nate. *On the Edge: The Art of Risking Everything*. New York: Penguin Press, 2024.

Sjöberg, Lennart. “Are All Crowds Equally Wise? A Comparison of Political Election Forecasts by Experts and the Public.” *Journal of Forecasting* 28, no. 1 (2009): 1–18. https://doi.org/10.1002/for.1083.

Stanford Medicine News Center. “Stanford Medicine - Artificial Intelligence.” 2025. https://med.stanford.edu/news/topics/artificial-intelligence.html.

Stanford University Human Centered Artificial Intelligence. *The 2025 AI Index Report.* Edited by Nester Maslej. April 2025. https://hai.stanford.edu/ai-index/2025-ai-index-report.

Stein-Perlman, Zach, and Katja Grace. “2022 Expert Survey on Progress in AI.” AI Timeline Surveys. *AI Impacts*, August 4, 2022. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai/.

Steinwart, Ingo, and Andreas Christmann. “Estimating Conditional Quantiles with the Help of the Pinball Loss.” *Bernoulli* 17, no. 1 (2011): 211–25. https://doi.org/10.3150/10-BEJ267.

Sundermier, Ali. “New AI Approach Accelerates Targeted Materials Discovery and Sets the Stage for Self-Driving Experiments.” SLAC National Accelerator Laboratory, August 18, 2024. https://www6.slac.stanford.edu/news/2024-07-18-new-ai-approach-accelerates-targeted-materials-discovery-and-sets-stage-self.

Surowiecki, James. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*. Doubleday, 2004.

Takeda, S., Keeley, A.R. & Managi, S. “How Many Years Away is Fusion Energy? A Review.” *Journal of Fusion Energy* 42, 16 (2023). https://doi.org/10.1007/s10894-023-00361-z

Tyrangiel, Josh. “Sam Altman on ChatGPT’s First Two Years, Elon Musk and AI Under Trump.” *Bloomberg.Com*, January 5, 2025. https://www.bloomberg.com/features/2025-sam-altman-interview/.

Walsh, Toby. “Expert and Non-Expert Opinion about Technological Unemployment.” Preprint, arXiv, 2017. https://doi.org/10.48550/ARXIV.1706.06906.

World Economic Forum. *How Close Are We to Very Powerful AI? Experts Weigh In*. Centre for Frontier Technologies and Innovation. Davos 2025, 2025. https://www.weforum.org/videos/davos-day-3-am25/.

Yang, Jing. “Paper Copilot - Paper Copilot.” 2024. https://papercopilot.com/.

Zhang, Baobao, Markus Anderljung, Lauren Kahn, Noemi Dreksler, Michael C. Horowitz, and Allan Dafoe. “Ethics and Governance of Artificial Intelligence: Evidence from a Survey of

Machine Learning Researchers." Preprint, arXiv, 2021. https://doi.org/10.48550/ARXIV.2105.02117.

Zhang, Baobao, and Allan Dafoe. "Artificial Intelligence: American Attitudes and Trends." SSRN Scholarly Paper No. 3312874. Social Science Research Network, January 9, 2019. https://doi.org/10.2139/ssrn.3312874.

Zhang, Baobao, Noemi Dreksler, Markus Anderljung, et al. "Forecasting AI Progress: Evidence from a Survey of Machine Learning Researchers." Preprint, arXiv, 2022. https://doi.org/10.48550/ARXIV.2206.04132.

Appendix A. Panel Construction and Sampling

We outline the construction of our panel in this section.

Appendix A.I. Computer Scientists

Computer Science Papers (OpenAlex)

We targeted the authors of top-cited publications in the AI and machine learning (ML) literature, according to the citations database OpenAlex (OpenAlex 2025, Priem et al. 2022). The median member of this group (data was available for 68% of the group) has 137,732 citations and an h-index of 55.[118] The institutions representing the most experts in our sampling frame include Stanford, MIT, and CMU. In this group, we initially included 300 computer scientists.

We use the OpenAlex API to query a large corpus of scholarly works.

1. We select papers in the “Artificial Intelligence” and “Computer Vision and Pattern Recognition” with topics listed in Table A1.[119]
2. We restrict our search to papers with a publication date in 2012 or later.
3. We drop all papers with less than 30 total citations.
4. We divide a work’s citations by the publication age (in years)[120] to obtain a publication-age-adjusted citation count.
5. We total publication-age-adjusted citation counts by *author*.

Topic Name
Statistical Machine Translation and Natural Language Processing
Neural Network Fundamentals and Applications
Artificial Intelligence and Expert Systems
Machine Learning for Mineral Prospectivity Mapping
Natural Language Processing
Speech Recognition Technology
Swarm Intelligence Optimization Algorithms
Anomaly Detection in High-Dimensional Data
Deep Learning in Medical Image Analysis
Machine Learning for Earthquake Early Warning Systems
Learning and Inference in Bayesian Networks

[118] The h-index is the number of publications a researcher has that have been cited at least that many times. For example, a researcher with an h-index of 20 has published 20 papers that have each been cited at least 20 times.
[119] OpenAlex assigns 3 topics to each paper, and we search for matches on *any* of the three topics. If desired, we can select papers based only on the primary topic.
[120] Paper age is calculated as the difference in the reference date and the publication date, expressed in unrounded years. We set a floor at 0.5 to prevent outsize adjusted citation counts for new papers with high pre-publication citation counts.

Application of Genetic Programming in Machine Learning
Sentiment Analysis and Opinion Mining
Digital Image Processing and Artificial Neural Networks
Machine Learning Methods for Solar Radiation Forecasting
Artificial Intelligence in Education and Technology
Artificial Intelligence Planning and Reasoning
Reinforcement Learning Algorithms
Adversarial Robustness in Deep Learning Models
Advances in Transfer Learning and Domain Adaptation
Machine Learning for Internet Traffic Classification
Game Artificial Intelligence Research
Active Learning in Machine Learning Research
Multi-label Text Classification in Machine Learning
Learning with Noisy Labels in Machine Learning
Graph Neural Network Models and Applications
Machine Learning in Smart Healthcare
Handling Imbalanced Data in Classification Problems
Deep Learning Applications in Healthcare
Artificial Intelligence in Service Industry
Photonic Reservoir Computing for Neural Computation
Explainable Artificial Intelligence
Gaussian Processes in Machine Learning
Theory and Applications of Extreme Learning Machines
Automatic Text Simplification and Readability Assessment
Theoretical Framework of Cognitive Informatics and Computational Intelligence
Optimization Methods in Machine Learning
Deep Learning for Wireless Signal Classification
Artificial Intelligence and Technology Innovation
Deep Learning in Computer Vision and Image Recognition
Handwriting Recognition and Text Detection
Generative Adversarial Networks in Image Processing

Table A1: Topics included in OpenAlex literature search.

We use these data to create 2 sampling pools:

1. *Top-Cited CS Authors*: We select the top-200 cited authors, using our *unadjusted* citations measure.
2. *Age-Stratified CS Authors*: We use the date of an author's first work as a proxy for their age. We define "early"-career authors as those with their first publication within 0-6 years from present; "mid"-career authors are defined as 7-12 years, and "senior" authors are defined as 13+ years. We drop any authors already included through the Top-Cited CS Authors. We pool the early- and mid-career authors and include the top-100 cited authors in this group, using our publication-age-adjusted citation measure.

The *Top-Cited CS Authors* group is intended to capture leading voices in AI research, while the *Age-Stratified CS Authors* pools allow us to find researchers at all career stages.

Target for reweighting

We use the first rounds of sampling from the *Top-Cited CS Authors* and *Early- and Mid-Career CS Authors* to create our reweighting targets.

Additional sampling

We later sampled the top 75 authors in the *Senior CS Authors* category and expanded the *Top-Cited CS Authors* and *Early- and Mid-Career CS Authors* pools to 250 and 200 individuals, respectively.

Computer Science Conference Papers (Paper Copilot)

To identify younger authors in the AI and ML literature, we also included in our sampling frame the lead authors of top-reviewed papers at leading AI and ML conferences. We use Paper Copilot, which includes OpenReview ratings for NeurIPS (2021-2024) and ICLR (2017-2025), to find top-rated conference papers (Yang 2024; Paper Copilot 2025). We use submissions from 2021 to present.

1. We drop papers in the 'Datasets & Benchmarks' and 'Journal' tracks for NeurIPS. We drop all rejected papers from NeurIPS and all rejected, desk rejected, and withdrawn papers from ICLR.
2. For each paper, we multiply each reviewer's rating and confidence score. We then sum these scores across reviewers and divide by N^{γ}, where $\gamma = 0.9$ to place greater weight on papers with more reviews.

We use these data to create one sampling pool:

1. We select the top papers for each conference-year, skewed towards more recent years.
2. We select the first author for each paper, and we move to the next author if we are unable to find contact information.

We take 40 authors from ICLR 2025 and NeurIPS 2024. We next select 20 authors from the preceding year, followed by 10 authors from each conference-year, stopping with the 2021 instance of each conference.

Target for reweighting

We use the first waves of sampling from both conferences (NeurIPS 2021-2024 and ICLR 2021-2025) to create our reweighting targets.

Additional sampling

We do not conduct any additional sampling from this source.

Computer Science Professors (CSRankings)

We use CSRankings, a ranking of university computer science departments, to expand our sample of computer science academics (Berger 2025).

1. Select the “AI” category.
2. Select the top-ranked professors within each institution.

Target for reweighting

We did not include this source in our first wave of sampling, so we do not include it in our reweighting targets.

Additional sampling

In a round of additional sampling, we selected the top 8 professors within each of the top 10 institutions.

Appendix A.II. Economists

Economics Papers (RePEc)

We included top-cited economists in the field of growth and technology, as well as the top-cited authors studying the economics of artificial intelligence. We use RePEc IDEAS, a publications database, to find top-cited economics publications (RePEc 2025).

We conduct two searches of the economics literature. Our first search is narrower:

1. We search within JEL Code O33 (Technological Change: Choices and Consequences; Diffusion Processes).
2. We limit our search to articles and papers.
3. We select papers that meet at least one of the following criteria:
    a. Abstract contains "AI" or "artificial intelligence"
    b. Keywords contain "AI" or "artificial intelligence"
    c. Title contains "AI" or "artificial intelligence"
4. We calculate paper-age-adjusted citation counts (citations per year), and then sum these adjusted citation counts by author.
5. We invite all top-300 individuals on this list with publicly available contact information.

Our second search is broader:

1. We search within JEL Codes O32 (Management of Technological Innovation and R&D), O33 (Technological Change: Choices and Consequences; Diffusion Processes), and O4 (Economic Growth and Aggregate Productivity).
2. We limit our search to articles and papers.
3. We calculate paper-age-adjusted citation counts (citations per year), and then sum these adjusted citation counts by author.

4. We invite the top 57[121] individuals on this list.

Since this pool of around 335 invites is less selective than in other categories, we only included individuals that had at least 400 citations or were a member of a top-50 institution according to RePEc. This filtering left us with 323 potential invitees.

The median member of this group (the statistics were available for 66% of the group) has 2,441 citations and an h-index of 20. The top-represented institutions include the OECD, the U.S. Census Bureau, and the Bank for International Settlements (BIS).

Target for reweighting

We use the first waves in the *O33* and *O3 and O4* pools to create our reweighting targets.

Additional sampling

We do not conduct any additional sampling from this source.

Expert panel

We include prominent economists across all fields by targeting the U.S. Economic Experts Panel, which yielded 60 invited individuals (Clark Center 2025). We invite the publicly listed alumni and active participants in the Panel.

Target for reweighting

We use all publicly listed alumni and active participants to create our reweighting targets.

Additional sampling

We do not conduct any additional sampling from this source.

Economics events sampling

We later expanded the economics group by inviting publicly listed attendees of conferences on the economics of artificial intelligence, including events organized by NBER, Brookings, OpenAI, and various Regional Federal Reserve Banks. We apply the same filtering to this expansion group.

We invited publicly listed presenters and attendees for various conferences on the economics of artificial intelligence. The basic requirements for the events are as follows:

- Events must have taken place in the U.S. (or online) between 2022 and the present.
- The event name contains at least one primary keyword and one secondary keyword.
    - Primary keywords: AI, artificial intelligence, automation

[121] We initially intended to invite the top 50 individuals, but a clerical error led to inviting the top 57.

    - Secondary keywords: economics (of), economic implications (of), policy implications (of), labor market, work, productivity, growth, technological change, inequality

The events are manually checked to ensure relevance, and the implied meaning of the event name is more important than strict adherence to the above lists of keywords.

We focused on the following four sub-pools of events:

- Major academic conferences
    - We use an EconBiz (2025) search for general economics conferences (JEL code A1) and check the sessions of the resulting 39 conferences.
- Events organized by policy/research institutions
    - Events organized by the policy/research institutions listed in the 2020 Global Go To Think Tank Index Report (McGann 2021). We include all speakers and **invited** participants (where available). We take the union of the top-10 think tanks on the “2020 Top International Economics Policy Think Tanks” and “2020 Top Domestic Economic Policy Think Tanks," filtering down to U.S.-based institutions. This results in 7 institutions:
        - Brookings
        - National Bureau of Economic Research (NBER)
        - Peterson Institute for International Economics (PIIE)
        - Heritage Foundation
        - Center for American Progress (CAP)
        - Cato Institute
        - RAND Corporation
- Events organized by public institutions
    - Events organized by individual Federal Reserve banks. We include all speakers and **invited** participants (where available).
- Events organized by major AI companies
    - Events organized by the top-5 AI companies by Epoch Training Compute ranking (Epoch AI 2025c). The five companies include:
        - xAI
        - Google DeepMind
        - OpenAI
        - Meta AI
        - Anthropic

Organizer	**Conference**
OpenAI	Navigating the Future of Work in the Age of AI
OpenAI	Making AI Work for Everyone
OpenAI	Thinking Machines & AI Economics
OpenAI	AI Economics in the OpenAI Forum

OpenAI	Preparing the Workforce for Generative AI: Insights and Implications
OpenAI	Expertise, Artificial Intelligence, and the Work of the Future Presented
OpenAI	Digital transformation and artificial intelligence: Implications for inequality and global economic convergence
OpenAI	Work in the age of artificial intelligence
OpenAI	AI, innovation, and welfare
OpenAI	Economics of Artificial Intelligence, Fall 2024
NBER	Economics of Artificial Intelligence, Fall 2024
NBER	Economics of Artificial Intelligence, Fall 2023
NBER	Economics of Artificial Intelligence, Fall 2022
Federal Reserve Bank of Boston	Generative AI at Work
Federal Reserve Bank of Philadelphia	Frontiers in Machine Learning and Economics: Methods and Applications – 2025
Federal Reserve Bank of Philadelphia	Frontiers in Machine Learning and Economics: Methods and Applications – 2024
Federal Reserve Bank of Philadelphia	Frontiers in Machine Learning and Economics: Methods and Applications – 2022
Federal Reserve Bank of Philadelphia	AI in Finance: Evidence from Job Listings
Federal Reserve Banks of Boston, Atlanta, Richmond	Technology-Enabled Disruption: Implications of AI, Big Data, and Remote Work
Federal Reserve Bank of Chicago	The Automation of Jobs: Impacts on Workers and Inequality
Federal Reserve Bank of Chicago; Federal Reserve Bank of San Francisco	The Automation of Jobs: Impacts on Workers and Inequality; David Autor | Expertise, AI, and the Work of the Future
Federal Reserve Bank of San Francisco	Panel Discussion on AI and the Labor Market
Federal Reserve Bank of San Francisco	The Growth and Employment Effects of AI
Federal Reserve Bank of San Francisco	AI-nomics: The Nexus of GenAI + the Economy
Federal Reserve Bank of San Francisco	The Transformative Power of AI: How is Technology Changing Our Lives?
Federal Reserve Bank of San Francisco	Technology and Inequality in the Age of AI
Federal Reserve Bank of San Francisco	GenAI for Economic Opportunity: Early Learnings from Philanthropy and Nonprofits
Federal Reserve Bank of San Francisco	GenAI and Real-World Productivity
Federal Reserve Bank of San Francisco	Job Matching in the Age of AI
Federal Reserve Bank of San Francisco	The Economics of Transformative AI
Federal Reserve Bank of San Francisco	The Disruptive Economics of AI

ASSA 2025 Annual Meeting of the AEA	Economic Implications of AI
ASSA 2025 Annual Meeting of the AEA	AI and the Future of Work
ASSA 2025 Annual Meeting of the AEA	Policy Implications of Transformative AI
ASSA 2025 Annual Meeting of the AEA	Productivity and Competition Effects of AI
ASSA 2025 Annual Meeting of the AEA	Artificial Intelligence and the Future of Work
ASSA 2025 Annual Meeting of the AEA	The Local Dynamic Effects of AI and Technological Change
Southern Economic Association 2024	Automation, Growth, and the Labor Market
Southern Economic Association 2024	Artificial Intelligence, Labor Markets, and Human Welfare
Computational Social Science 2024	AI and the Future of Work
ASSA 2024 American Economic Association	Effects of Artificial Intelligence
ASSA 2023 American Economic Association	Economic Uses and Applications of AI and Big Data
2022 Society for Economic Dynamics	Data, AI, and Automation II
2022 Society for Economic Dynamics	Data, AI, and Automation I

Table A2: Economics events-based sampling.

Target for reweighting

We did not include this source in our first wave of sampling, so we do not include it in our reweighting targets.

Additional sampling

In a round of additional sampling, we invited the publicly listed presenters and attendees of the above conferences.

Appendix A.III. Industry Professionals

We identified the 20 frontier AI labs according to performance on the Chatbot Arena LLM Leaderboard and investment in training compute according to Epoch (LMArena 2024; Epoch AI 2024b); the final list of 20 labs came from the union of the two underlying sources. We identified core contributors to the most recent models from these labs. This list is led by OpenAI, xAI, Google DeepMind, Meta AI, Anthropic, Alibaba, and DeepSeek. We initially targeted 220 individuals from the labs, but ended up with an initial sampling frame of 209, due to limits in publicly available information on lab membership. We discuss the sources used to determine lab membership below.

Language Models, Training Compute (Epoch)

We use the data on notable AI models from Epoch. (Epoch AI 2024b)

1. We sort models by training compute (floating point operations [FLOP]).

2. We record all leading, non-academic institutions and, where available, authors. If a paper distinguishes between core- and non-core contributors, we select core contributors.

We disproportionately sample individuals from institutions associated with higher-ranked models, selecting 10 individuals from the top 5 institutions and 6 individuals from institutions ranked 6 to 15. However, not all companies had enough publicly listed contributors to meet these targets for each lab.

Target for reweighting

We use the first wave of recruitment from the top 15 labs to create our reweighting targets.

Additional sampling

In a round of additional sampling, we invited an additional 25 and 15 individuals from the top 5 institutions and institutions ranked 6 to 15, respectively.

Language Models, Performance (Chatbot Arena)

We use the Chatbot Arena LLM Leaderboard data (LMArena 2024).

1. We filter to the “Hard Prompts (Overall)” category.
2. We sort models by their leaderboard rating.
3. We record all leading, non-academic institutions and, where available, authors. If a paper distinguishes between core- and non-core contributors, we select core contributors.

We disproportionately sample individuals from institutions associated with higher-ranked models, selecting 10 individuals from the top-5 institutions and 6 individuals from institutions ranked 6 to 15. However, not all companies had enough publicly listed contributors to meet these targets for each lab.

Target for reweighting

We use the first wave of recruitment from the top-15 labs to create our reweighting targets.

Additional sampling

In a round of additional sampling, we invited an additional 25 and 15 individuals from the top-5 institutions and institutions ranked 6 to 15, respectively.

Industry Players (Crunchbase)

We also invited technical staff from active AI companies with $500 million or more in total venture capital (and other) funding. This list includes groups like Databricks, Waymo, and Anduril Industries, which are involved in the development and deployment of AI, despite not producing frontier LLMs in-house. Though we initially targeted 550 individuals from this industry

group to include in our frame, we could only identify 370 individuals, due to limits in publicly available information on employees from Crunchbase, a financial database (Crunchbase 2025).

1. We select all private AI-related companies that have raised $500 million or more in *total* funding, with the most recent fundraising date within 2 years.

We identify engineering and research and development staff with contact information from Crunchbase.[122] We disproportionately sample individuals from companies with more fundraising, targeting 25 individuals from the top-10 institutions and 15 individuals from institutions ranked 11 to 30. We select staff members at random.

Target for reweighting

We use the first wave of recruitment from the top-30 companies to create our reweighting targets.

Additional sampling

In a round of additional sampling, we invited an additional 50 and 25 individuals from the top-10 institutions and institutions ranked 11 to 30, respectively.

Appendix A.IV. Policy Professionals

Think Tanks (Horizon and Global Go To Think Tank Index)

We identified institutions leading the discussion on AI development, policy, and impacts. We use the list of “AI policy think tanks” from the Horizon Institute of Public Service and the “2020 Best AI Policy and Strategy Think Tanks” from the 2020 Global Go To Think Tank Index Report (Table 56) to find these think tanks working on AI policy and impacts (Horizon Institute of Public Service 2025; McGann 2021). These institutions include Brookings, RAND, AEI, AI Now, and the Stanford Institute for Human-Centered Artificial Intelligence.

1. We take the union of the 2 lists.
2. For general-purpose think tanks, we will limit our search to AI-specific initiatives, e.g., the Artificial Intelligence and Emerging Technology Initiative at Brookings (Brookings Institution 2025). Or, we will filter based on reported staff expertise.

We generated a list of 62 (initiatives within) think tanks. If a think tank has multiple AI-related initiatives or groups, we treat the groups as separate think tanks. We randomly select 5 research staff from each entity. We identify research staff from staff pages. In some cases, we exhausted contacts before selecting 5 individuals.

[122] Crunchbase did not have contact information for 3 of the top 10 companies and 7 of the companies ranked from 11 to 30, so we did not include these companies *in this sampling pool*.

Target for reweighting

We use the first wave of recruitment from the sampled institutions to create our reweighting targets.

Additional sampling

In a round of additional sampling, we invited contacts from institutions with low response rates. We sampled an additional 20 individuals at institutions with a response rate of 10% or lower and 10 individuals from institutions with a response rate between 20% (inclusive) and 10% (exclusive).

Other Policy Professionals

We also invited research staff at Epoch AI, a leading nonprofit research center that collects data on and studies AI-related trends. However, we exclude these individuals when calculating our population weighting targets.

Appendix A.V. Other Invitees

TIME 100

We supplemented the categories above with the honorees from TIME's 100 Most Influential People in AI in 2023 and 2024 (Bajekal 2023; Barker Bonomo and Javed 2024). We identified 173 unique individuals across these two lists. Some individuals in this category were previously included in other categories. Individuals not previously included were sorted into one of the four primary expert categories, both for the purposes of respondent classification and defining our reweighting targets. Individuals that did not fit any of our frame groups were excluded from the panel.

Target for reweighting

We use the unique individuals across the two lists, sorted into appropriate categories, to create our reweighting targets. We were not able to assign a small subset of this group to another category, and these individuals do not contribute to our reweighting targets.

Additional sampling

We did not conduct any additional sampling in this group.

Snowball Sampling

Finally, we allowed invited respondents to recommend other qualified candidates for the survey, yielding 172 additional invitees. We prompt respondents to list two individuals they expect to largely agree with and two they expect to largely disagree with, as well as provide room for additional suggestions. In order to filter this group, we require that an individual:

- meets the requirements of another sampling category;

- has over 1,000 academic citations; or
- has over 300 academic citations if a PhD student or postgraduate researcher.

These requirements excluded only 7 of the recommended candidates that ultimately enrolled. After exclusion, the referred group has 11.6 years of experience on average and 75% have a postgraduate degree. Like other expert sample expansions, referred contacts are not included in frame targets, but responses do receive positive weights through the reweighting process.

We assigned respondents from this group into the appropriate sampling category, if one exists. We were not able to assign a small subset (less than 10 individuals) of this group to another category.

Target for reweighting

We do not use snowball-sampled contacts in creating our weighting targets.

Appendix A.VI. Superforecasters

We include a sample of Superforecasters, sourced through FRI and Good Judgment Inc.[123] These superforecasters have a demonstrated track record of providing the most accurate forecasts.

Target for reweighting

We do not reweight the superforecaster sample.

Additional sampling

We do not conduct any additional sampling.

Appendix A.VII. Public

We include in our sample highly-engaged participants from past FRI research projects on the CloudResearch Connect platform. We initially invited approximately 2,600 individuals.

Target for reweighting

We use the U.S. population to generate our reweighting targets. We discuss our targets for reweighting in Reweighting.

[123] Forecasters are denoted “superforecasters” if they (1) were in the top 2% of the accuracy distribution in a given year of the IARPA ACE tournament (IARPA ACE Program n.d.; Mellers et al. 2014) or (2) they were a highly accurate performer on Good Judgment Open, an online continuous geopolitical forecasting tournament. Good Judgment Inc., which runs Good Judgment Open, then adds these top forecasters to the “superforecaster” pool. Most superforecasters come from the first selection criteria. Mellers et al. (2015) finds persistent performance of these superforecasters across several years of geopolitical forecasting.

Additional sampling

We later expanded our sample among individuals over the age of 50 and identifying as Republican, followed by a targeted sampling of individuals with a high school degree or equivalent as their highest level of completed education.[124]

Appendix A.VIII. Reweighting

We summarize the expert reweighting targets in Table A3 below.

Category	Target
Years of relevant experience	
0-10	43%
11-15	24%
16-20	13%
21+	20%
Age	
18-29	16%
30-44	56%
45-64	23%
65+	5%
Affiliation with Effective Altruism	
Non-Affiliated	85%
Affiliated	15%
Gender	
Male	78%
Female	20%
Non-binary/prefer to self-describe	2%
Continent	

[124] We include individuals in the following categories: no formal education; less than a high school diploma; high school graduate - high school diploma or the equivalent (for example: GED); and some college, but no degree.

Asia	14%
Europe	23%
North America	60%
Other regions	3%
Education	
Associate's/Bachelor's or less	25%
Post-graduate degree	75%
Proportion of Participants in Each Expert Category[125]	
Computer Science	25%
Economics	25%
Industry	25%
Policy	25%
Affiliation with Top AI Labs	
Non-Affiliated	82%
Affiliated	18%

Table A3: Population targets for the expert reweighting process.

[125] As noted above, participants were weighted equally on their category of expertise for the purposes of reweighting, so that each of the four expert subpopulations gets 25% weight in the overall expert numbers we present. The actual (unweighted) fraction of invitees in each category of our sampling frame was as follows: Computer Science: 27%; Economics: 21%; Industry: 36%; Policy: 16%. The fraction of respondents in each expert subcategory, after reassignment, was as follows: Computer Science: 22%; Economics: 24%; Industry: 20%; Policy: 34%.

Appendix B. Monthly Surveys and Forecasting Questions

Appendix B.I. Monthly Surveys

We conduct surveys approximately every month, each consisting of 5-6 forecasting questions. We expect each survey to take experts approximately 30-40 minutes to complete, and respondents are informed of this time estimate. Respondents receive a standardized payment for each survey they complete.[126] In addition to quantitative forecasts, we collect rationales for each forecasting question, in the form of plain text. We use these rationale data to understand the underlying reasons for forecaster responses.

Participants are invited to complete all surveys through the Quorum survey platform. Examples of this survey instrument are shown in Appendix B.V. Survey Instrument (Quorum Research 2025).

While respondents are able to edit their forecasts on each question during a short window of time following each survey wave's release, we use the respondents' final forecast for a given question in our analysis.

We further validate respondent forecasts with three steps:

1. We ensure that forecast values fall within an appropriate range for the unit of the corresponding question (e.g., proportions should fall between 0% and 100%). In cases where forecasts fall above or below the appropriate range, we set the forecast to either the minimum limit or the maximum limit respectively.[127]
2. We check if quantile forecasts are coherent, meaning that forecast values for quantile forecasts should be weakly monotonically increasing. For example, when we elicit quantile forecasts, the 25th percentile forecasted value should be less than or equal to the 50th percentile value, which should be less than or equal to the 75th percentile value. If a forecaster is incoherent on a given question, we remove their forecast for that question.[128]
3. We adjust forecasts that are not within reasonable bounds given the context of a question. Specifically, for FrontierMath forecasts we adjust forecasts below the 19% historical baseline upward to the baseline because it is impossible for the "best" performance by a model as of 2030 to be worse than the best performance of a model

[126] We provide expert participants with $2,000 per each year of full participation (prorated for the number of surveys they complete, with an expectation that we will complete 12 surveys in a typical year). In other words, experts receive $166.67 per survey completed. We provide superforecasters with $1,000, prorated similarly, or $83.33 per survey completed. We pay public participants in line with CloudResearch platform norms, or $8 per survey completed ($13.71 per hour).
[127] While the survey instrument is designed to reject these forecasts, respondents can sometimes edit forecasts after validation.
[128] While the survey instrument is designed to reject these forecasts, respondents can sometimes edit forecasts after validation.

now. Similarly, for Occupational Employment Index forecasts we winsorize forecasted values down to a +4,000 percentage point increase if forecasted values exceed that level.

In practice, these validations adjust or exclude less than 1% of respondent forecasts.

To account for low-effort participants, we excluded participants based on their responses to the first three waves. Specifically, participants were excluded from the sample if they were unusually fast, and had low effort rationales, as defined below:

- Unusually fast participants are those that had an average completion time (across the surveys that they completed) of surveys of under 15 minutes.
- Low effort rationales are defined as less than 10 word rationales on average (for all rationales that they submitted).

There are cases in which a participant is unusually fast but we do not want to exclude them. Several panel members waive questions due to insider information. Hence, their survey completion will be fast. Therefore, we use the combination of the two criteria above to determine whether a participant should be excluded for low effort. Using these criteria, we exclude 2 expert participants (0.5%) and 35 public participants (2.5%) based on the results from Wave 1 through Wave 3.

Appendix B.II. Forecasting Questions

LEAP forecasting questions fall into five categories: inputs, capabilities, adoption, impacts, and scenarios. This framework tracks AI development from the drivers of progress and technical progress on highly consequential tasks through real-world deployment, the resulting consequences, and possible future trajectories.

We ask various types of forecasting questions:

1. *Probabilistic:* We ask participants to assign a probability to a binary or discrete event.
2. *Quantile:* We ask participants to forecast quantiles of a continuous outcome (typically the 25^{th}, 50^{th}, and 75^{th} percentiles).
3. *Point Estimate:* We ask participants for a point estimate of a continuous outcome.

In addition to quantitative forecasting questions, we also include open-ended questions. For example, in Wave 1, participants gave text-based responses to the following two questions:

- What do you see as the main cognitive limitations of AI systems in 2025?
- What will AI experts identify as the main cognitive limitations of state-of-the-art AI systems by 2030?

We used the cognitive limitations that experts identified in these two questions to develop a question for the following wave. Specifically, in Wave 2, we asked participants, “By the end of 2030, what percent of LEAP expert panelists will agree that each of the following is a serious cognitive limitation of state-of-the-art AI systems?” The list of options was derived from the responses in the previous wave. Options included:

- *Hallucination / Inaccuracy:* they give plausible-sounding but incorrect or fabricated information.
- *Shallow reasoning:* their reasoning capabilities are shallow beyond math and coding and lack “genuine” causal and logical reasoning.
- *Lack of long-term memory:* they cannot retain and utilize information across sessions or long interactions.

In summary, we used unstructured responses from Wave 1 to generate a resolvable forecasting question in Wave 2.

We include both recurring and one-shot forecasting questions. We intend to repeat recurring questions once each year. Our recurring questions allow us to track how expert views evolve over time. On the other hand, one-shot questions are domain- and timeline-flexible, allowing us to quickly add questions addressing unforeseen events that generate significant interest (e.g., new model/benchmark releases or contemporary policy debates). We source forecasting questions from academic papers, technical and policy reports, prediction platforms, public writings by leading AI figures, our past research output, our academic advisory board, and suggestions from survey respondents. We then create resolvable forecasting questions from these various input sources.

Appendix B.III. Scoring

For ground truth and dynamically resolved questions, the resolution method depends on the question type.

1. *Probabilistic:* We use the Brier score (Brier 1950) to score probabilistic predictions.
2. *Quantile:* We use the S-score as described in Jose and Winkler (2009) to score quantile forecasts. The S-score is identical to the tilted absolute value loss function, the loss function used for quantile regression.[129]
3. *Point Estimate:* We use mean squared error to score point estimates.

All three of these scoring rules are *proper*: respondents yield the best score (in expectation) if they report their true belief.[130,131]

The Brier score’s loss function is the squared difference between the forecast probability of an outcome and an indicator function representing the realized outcome. Forecasting accuracy is incentivized as any deviation between the forecaster’s true subjective probability and their reported probability is penalized (in expectation). Similarly, scoring using mean squared error incentivizes reporting true forecasted values.

The S-score uses a titled absolute value loss function that penalizes errors asymmetrically

[129] This loss function is sometimes called “check loss” or “pinball loss” (Steinwart and Christmann 2011).
[130] For point predictions scored using MSE, reporting one’s average belief minimizes the score.
[131] This incentive for truthfulness might not hold if accuracy payments are not linear in accuracy. For example, large prizes for top-performers can introduce incentives for reporting more extreme forecasts (Lichtendahl and Winkler 2007).

based on whether the forecast value is greater or less than the realized value. At the 25% quantile, overestimation is more penalized than underestimation, while at the 75% quantile underestimation is more penalized than overestimation. At the median, errors are symmetrically penalized. Forecasters minimize their expected loss by reporting their subjective quantiles truthfully.

For questions with intersubjective resolution, we use FRI's preferred intersubjective metric. FRI is concurrently conducting experiments to inform which intersubjective metric best incentivizes accuracy while remaining transparent to forecasters, which will inform our final choice of metric and resolution method for these types of questions.

When calculating prizes, we normalize[132] scores across questions, adopting approaches similar to other FRI work. These normalizations adjust for differences in question difficulty and, in the case of non-probabilistic forecasts, differences in outcome scale. We leave a complete specification of these methods to future project updates.

Some respondents might join the panel late, and their exposure to prizes will be proportional to the share of questions they answer: answering a question will never reduce a respondent's chance of winning a prize.

We score rationales with a combination of human and LLM judges, and provide prizes to the highest quality rationales in each survey wave. We do not publicly share our scoring criteria to prevent gaming by participants.

Appendix B.IV. Question Resolution

We use three methods to resolve forecasting questions and determine forecasting accuracy:

1. *Ground truth:* We prioritize resolving our forecasting questions with reputable data sources or independent research with commissioned expert input.
2. *Dynamic resolution:* Where ground truth data are unavailable, we use LEAP panel 'nowcasts' on the resolution date as data source.[133]
    - If the question is not resolvable using ground truth data, LEAP panelists will be asked to 'nowcast' the question (i.e., give their best estimate of the true value) on the resolution date.
    - Nowcasts on the resolution date will be elicited using a truth-telling metric, likely an intersubjective metric (discussed in greater detail below).
3. *Intersubjective metrics:* We aggregate LEAP panel forecasts in order to create resolved values, using an intersubjective metric to incentivize truth-telling.[134]

[132] As in other work, we often trim or winsorize data before calculating the values used for recentering (mean) and scaling (standard deviation).
[133] Karger et al. (2021) find that aggregated expert forecasts of a similar form produce accurate estimates evaluated in hindsight.
[134] No questions from the first three waves use intersubjective resolution, but we plan to explore intersubjective resolution in future waves.

- The forecasting literature now offers an array of intersubjective metrics that allow us to reward and score forecasters when we do not have access to ground truth, by evaluating how a respondent's forecasts compare to those of other participants (e.g., Prelec 2004, Miller et al. 2005, Cvitanić et al. 2019, Karger et al. 2021). In general, these metrics do not rely on access to ground truth. FRI is concurrently conducting experiments to inform which intersubjective metric best incentivizes accuracy while remaining transparent to forecasters, which will inform our final choice of metric and resolution method for these types of questions.
- Possible metrics include reciprocal scoring (Karger et al. 2021) and Choice Matching (Cvitanić et al. 2019).

However, it remains challenging to foresee possible changes to data sources or the relevance of certain indicators. Additionally, some forecasting questions may lead to participant confusion, even after extensive vetting.[135] A benefit of LEAP's longitudinal nature is that we can continue to improve and adjust questions that we re-ask in the coming years.

Appendix B.V. Survey Instrument

Participants completed the LEAP survey on the Quorum platform.[136] For each question, they were presented with relevant background information, historical data, and resolution criteria, along with a quantile forecasting explanation. Participants could submit their forecasts either by dragging values on the time series chart or by entering them directly into the provided textboxes. Screenshots of the survey are shown below.

[135] For example, we included a forecasting question on K-12 classroom use of AI in Wave 3 of LEAP, but we excluded it from analysis due to forecasters adopting opposing interpretations of the question.
[136] https://quorumresearch.com/



Figure B1: Screenshot of the survey instrument showing the AI Investment question from Wave 3

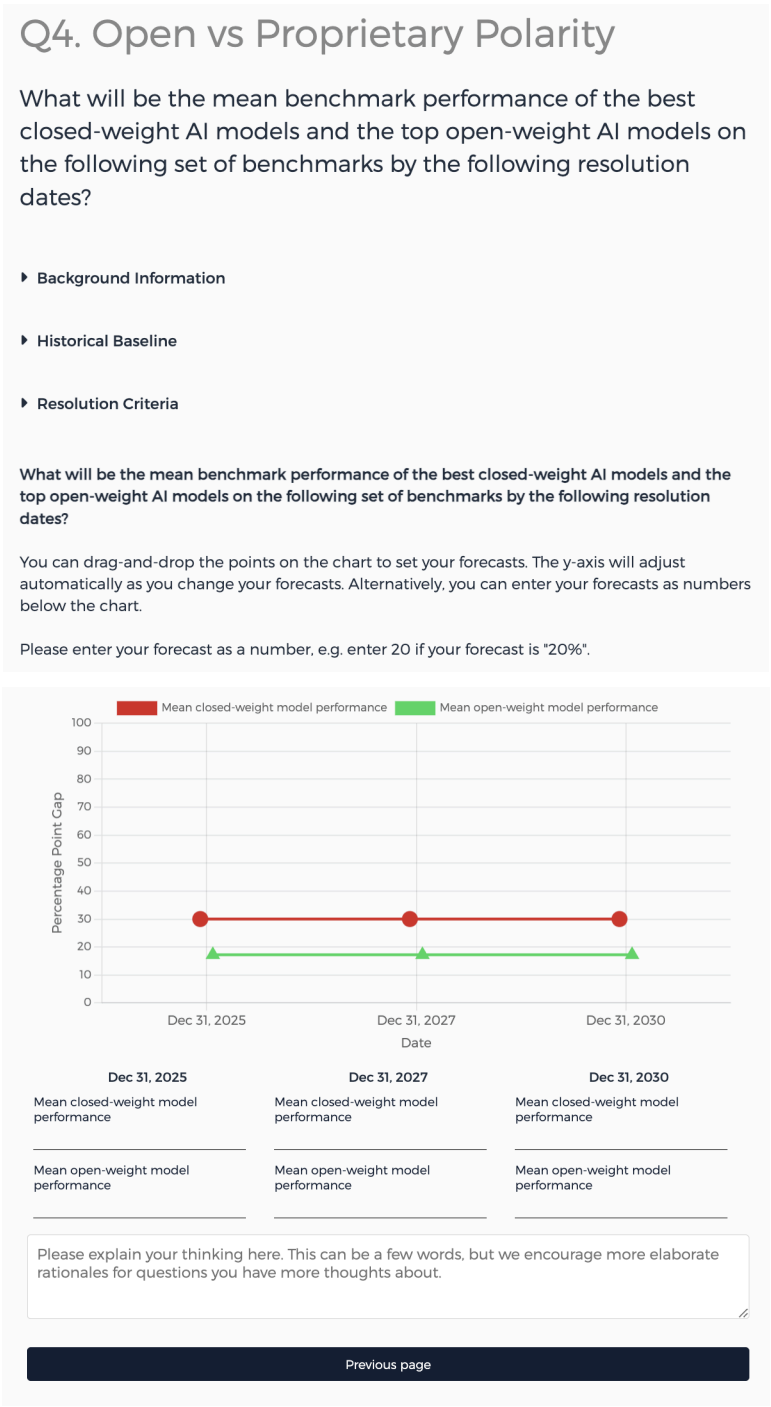


Figure B2: Screenshot of the survey instrument showing the Open vs Proprietary Polarity question from Wave 3

Appendix C. Pooled Distribution Estimation

When a respondent provides a set of quantile forecasts, we estimate a distribution over the entire outcome space by fitting a beta or gamma distribution via nonlinear least squares on the quantile function. For variables with support bounded above and below, we fit a Beta distribution. For support that is bounded only above or only below, we fit a Gamma distribution. The pooled distribution is then the mixture distribution over respondents, with mixture weights proportional to our analysis weights. Notably, this procedure requires extrapolation from below the 25^{th} and above the 75^{th} percentiles, and results might be less reliable in the tails.

We draw bootstrap samples from this mixture distribution to find the quantiles of and plot the pooled distributions.

We additionally use the law of total variance to decompose the total variance of the mixture distribution into within- and between-forecaster components.

Appendix D. Public Accuracy Stratification

For our public sample with accuracy scores, we correlate forecasts for each question with out-of-sample forecasting accuracy. Figure D1 shows these results.

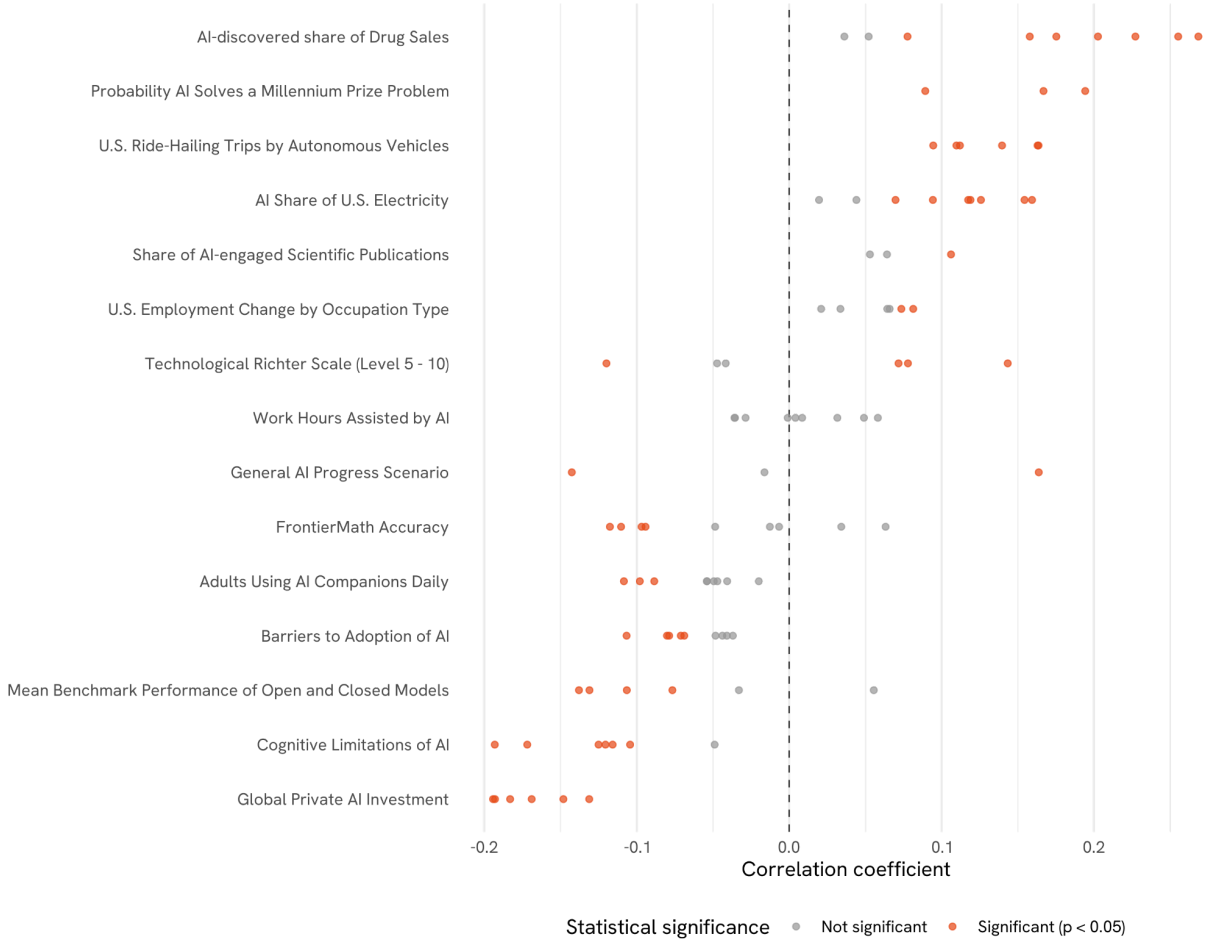


Figure D1: Correlation between forecasting accuracy and predicted AI progress. Spearman correlation coefficients between participants' prior accuracy scores and their forecasts on LEAP questions. Each point represents an individual question, with orange colors indicating statistical significance ($p < 0.05$). Lower S-scores indicate higher forecasting accuracy; therefore, positive correlations (top right) suggest that more accurate forecasters predicted smaller values, while negative correlations (bottom left) indicate that more accurate forecasters predicted higher values.

Correlation analyses revealed statistically significant relationships between forecasting accuracy and predicted outcomes on several questions. However, all Spearman correlation coefficients fell within the range of -0.2 to 0.27, indicating weak associations that explain minimal variance in forecast values. Despite the weak magnitudes, correlations showed some consistent directional patterns within question types. Questions related to specific AI applications (Drug Discovery, Millennium Prize, Electricity Consumption, Autonomous Vehicle Trips) exhibited positive

correlations, where lower accuracy forecasters (higher S-Scores) predicted faster progress. Conversely, questions about AI ecosystem factors (AI Investment and AI companions) showed negative correlations, with higher accuracy participants forecasting greater advancement.

Appendix E. Survey Questions

We first list all questions for the first three waves, and the following subsections include the full text of each question.

Appendix E.I. Survey Questions: Wave 1

1. ***FrontierMath***: What will be the highest percentage accuracy achieved by an AI model on FrontierMath, by the following resolution dates?

   FrontierMath is a benchmark of hundreds of original, expert-crafted mathematics problems, which typically require hours or days for expert mathematicians to solve. The problems are unpublished and designed to be "guessproof," with less than a 1% chance of guessing correctly without the mathematical work. For more information on the benchmark and further details, see Appendix E.I. 1 FrontierMath.

2. ***Autonomous Vehicle Trips***: What percentage of U.S. ride-hailing trips will be provided by autonomous vehicles that are classified SAE Level 4 or above in the years 2027 and 2030?

   SAE Level 4 "High Automation" vehicles do not require any human intervention when automated driving features are engaged, but may only operate in limited conditions or environments. For further details, see Appendix E.I. 2 Autonomous Vehicle Trips.

3. ***Occupational Employment Index***: What will the percent change in the number of jobs (compared to Jan 1, 2025) in the U.S. be for white-collar, blue-collar, and service-sector occupations, by the following resolution dates?

   For this question, each group of occupations is substituted with a representative sample of professions, e.g., childcare workers and nurses (amongst others) for service-sector occupations. For the full list and further details, see Appendix E.I. 3 Occupational Employment Index.

4. ***General AI Progress***: At the end of 2030, what percent of LEAP panelists will choose "slow progress," "moderate progress," or "rapid progress" as best matching the general level of AI progress?

   In the "slow progress" scenario, AI is a capable assistant that augments human work but doesn't replace it; "moderate progress" refers to a scenario where AI functions as an effective collaborator: and the "rapid progress" scenario describes a future in which AI systems match or surpass the best human capabilities across most domains. For further details on the different scenarios, see Appendix E.I. 4 General AI Progress.

5. ***Technological Richter Scale***: At the end of 2040, what is the probability for AI achieving the following levels of net impact on human society as compared to the impact of past technological events?

   Nate Silver's book "On the Edge" proposes the technological Richter scale (TRS) which, analogous with earthquake magnitudes, rates the impact of technologies on a roughly logarithmic scale. We ask for levels between 5 (equal to a commercially successful invention important in its category, e.g., a leading brand

of windshield wipers) to 10 (epoch-defining event that alters the fate of the planet, e.g., the rise of humans). For full question background and resolution details, see Appendix E.I.5.

6. ***Cognitive Limitations of AI***

   (1) What do you see as the main cognitive limitations of AI systems in 2025?
   (2) What will AI experts identify as the main cognitive limitations of state-of-the-art AI systems by 2030?

   These are both open response questions whose results were used in Wave 2. For further details on these questions see Appendix E.I. 6 Cognitive Limitations of AI, for the follow-up question in Wave 2 see Appendix E.II. 5 Cognitive Limitations, Part II.

Appendix E.II. Survey Questions: Wave 2

1. ***Millennium Prize***: Will AI solve or substantially assist in solving a Millennium Prize Problem in mathematics by the following resolution dates?

   The seven Millennium Prize Problems were chosen by the founding Scientific Advisory Board of the Clay Mathematics Institute (CMI) of Cambridge, Massachusetts to be the most significant and difficult mathematics problems unsolved by 2000. As of July 2025, only one of the seven problems has been solved. For further details, see Appendix E.II. 1 Millenium Prize.

2. ***Diffusion of AI Across Sciences***: What percent of publications in the fields of Physics, Materials Science, and Medicine in 2030 will be 'AI-engaged' as measured in a replication of this study?

   A 2024 study by researchers at Harvard University and the University of Chicago tracked scholarly engagement with AI across 20 scientific fields by measuring the change in percentage of "AI-engaged publications" within each field. "AI-engaged publications" are defined as papers with abstracts that contain at least one keyword related to contemporary approaches to AI. For further details, see Appendix E.II. 2 Diffusion of AI Across Sciences.

3. ***Drug Discovery***: What percent of sales of recently approved U.S. drugs will be from AI-discovered drugs and products derived from AI-discovered drugs in the years 2027, 2030 and 2040?

   We define "recently approved U.S. drugs" as drugs which received approval for sale by the U.S. Food and Drug Administration (FDA) in the one year preceding the resolution year. "AI-discovered drugs" are drugs developed through significant use of AI techniques in processes that would likely not have occurred without post-2022 AI capabilities. For further details, see Appendix E.II. 3 Drug Discovery.

4. ***Electricity Consumption***: What percent of U.S. electricity consumption will be used for training and deploying AI systems in the years 2027, 2030 and 2040?

   AI model training and deployment occur mainly through the use of data centers. Servers in data centers are split into two main categories: conventional and AI specialized servers. This question asks about the electricity consumption of AI

specialized servers. For further details, see Appendix E.II. 4 Electricity Consumption.

5. ***Cognitive Limitations, Part II***: By the end of 2030, what percent of LEAP expert panelists will agree that each of the following is a serious cognitive limitation of state-of-the-art AI systems?

   We identified the limitations via an open response question in Wave 1. The limitations are: hallucination / inaccuracy; shallow reasoning; lack of long-term memory; limited ability to generalize; limited metacognition and continual learning; limited embodiment / robotics; limited inter-system collaboration. For more information on these limitations and further question details, see Appendix E.II. 5 Cognitive Limitations, Part II.

6. ***Barriers to Adoption, Part I***: What do you see as the main barriers to adopting current or future state-of-the-art AI systems for broader use in society?

   This is an open response question whose results were used in Wave 3. For further details on these questions see Appendix E.II. 6 Barriers to Adoption, Part I, for the follow-up question in Wave 2 see Appendix E.III. 6 Barriers to Adoption, Part II.

Appendix E.III. Survey Questions: Wave 3

1. ***AI Investment***: What will be the global private investment (in billion USD) in AI in the following calendar years? (2027, 2030)

   According to the AI Index Report (2025), private investment in AI includes investment in AI startups that have received over $1.5 million in investment since 2013. For further details, see Appendix E.III. 1 AI Investment.

2. ***Generative AI Use Intensity***: What percent of work hours in the U.S. at the following dates will be estimated as assisted by generative AI, according to a future iteration of the St. Louis Fed study or a similar study selected by an FRI-appointed expert panel?

   A June 2025 study by the Federal Reserve Bank of St. Louis, based on the Real-Time Population Survey (N=3,216), estimated that 1.3%–5.4% of all U.S. work hours were assisted by generative AI based on self-reports via a nationally representative survey. This question tracks how much self-reported generative AI adoption intensifies in work settings, as measured by updated or similar surveys. For further details, see Appendix E.III. 2 Generative AI Use Intensity.

3. ***Personalized Education***: What percentage of weekly instructional hours on average will K-12 students in the United States spend using AI-powered tutoring or teaching tools during instructional hours?

   Note: this question was excluded from our analysis due to substantial misinterpretation.

4. ***Open vs Proprietary Polarity***: What will be the mean benchmark performance of the best closed-weight AI models and the top open-weight AI models on the following set of benchmarks by the following resolution dates?

   The set of benchmarks is:

   - Mathematical reasoning: FrontierMath (Tier 1-3 and Tier 4 separately)
   - Software engineering: SWE-Bench Verified

- General reasoning and abstraction: ARC-AGI-2

For further details, see Appendix E.III. 4 Open vs Proprietary Polarity.

5. ***AI Companions***: What proportion of U.S. adults will self-report using AI for companionship at least once daily by the following resolution dates?

   A July 2025 poll from the AP-NORC Center for Public Affairs Research found that 6% of U.S. adults reported using AI for companionship at least once a day. For further details, see Appendix E.III. 5 AI Companions.

6. ***Barriers to Adoption, II***: By the end of 2030, what percent of LEAP expert panelists will say that each of the following factors has significantly slowed AI adoption relative to popular expectations around AI adoption progress in 2025…

   We identified the barriers via an open response question in Wave 2. The barriers are: lack of reliability; restrictive regulations; cost issues; data quality issues; integration challenges; not enough use cases; lack of AI literacy; social-cultural anomie. For more information on these barriers and further question details, see Appendix E.II. 6 Barriers to Adoption, Part II.

Appendix E.I. Survey Questions: Wave 1

1. FrontierMath

What will be the highest percentage accuracy achieved by an AI model on FrontierMath,[137] by the following resolution dates?

- **Jan 1, 2026**
- **Jan 1, 2028**
- **Jan 1, 2031**

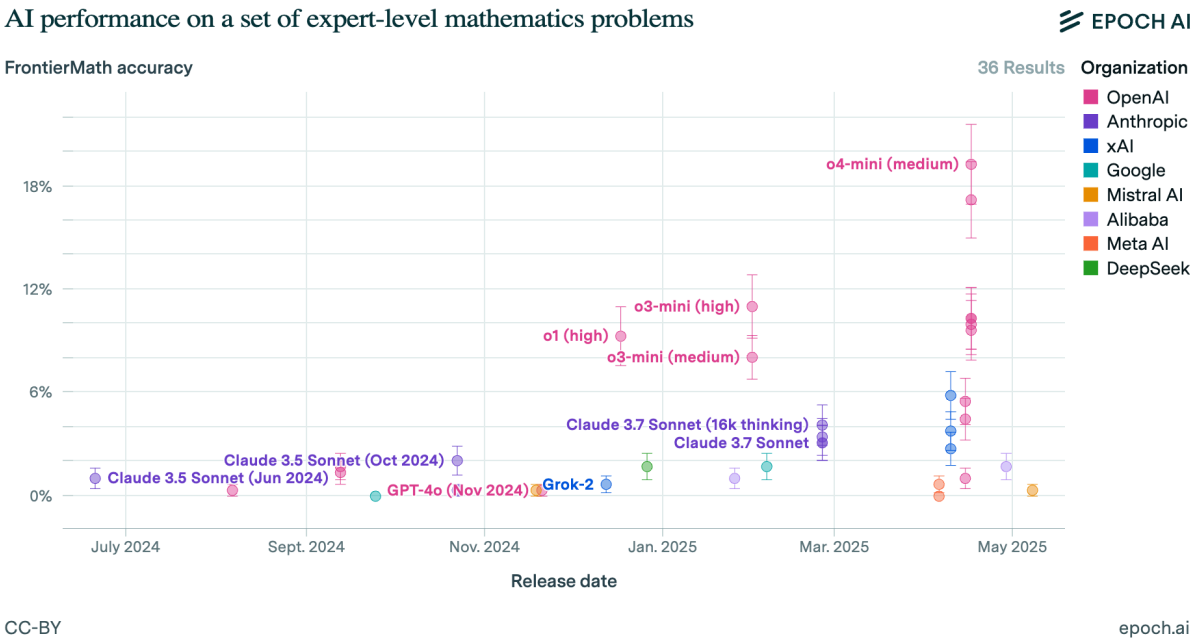
Background information

FrontierMath is a benchmark of hundreds of original,[138] expert-crafted mathematics problems, which span major branches of modern mathematics and typically require hours or days for expert mathematicians to solve. The problems are unpublished and designed to be "guessproof," with less than a 1% chance of guessing correctly without the mathematical work.

Historical baseline

As of April 16, 2025, the best AI performance according to Epoch AI, the benchmark developer, is achieved by o4-mini solving **19%** of problems.[139]

[137] *Link provided to participants:* https://epoch.ai/data/ai-benchmarking-dashboard?cameraX0=2024-8-16&cameraY0=0.009294645423959632&cameraX1=2025-4-27&cameraY1=0.18079213872900524#explore-the-data
[138] *Link provided to participants:* https://epochai.org/frontiermath/the-benchmark
[139] *Link provided to participants:* https://epoch.ai/data/ai-benchmarking-dashboard



The problems in Epoch's evaluation so far fall into three tiers of difficulty:[140]

- Tier 1: Advanced, near top-tier undergrad/olympiad, 25% of problems.
- Tier 2: Needs serious grad-level background, 50% of problems.
- Tier 3: Needs an expert mathematician with relevant specialization, 25% of problems.

As of March 2025, the percentage of problems solved by o3-mini in each difficulty tier (out of 100% for each tier)[141]:

- Tier 1: 15.6%
- Tier 2: 6.8%
- Tier 3: 5.8%

Epoch AI is currently adding Tier 4 problems to the FrontierMath benchmark,[142] with difficulty level requiring weeks or months of expert effort. In this question, **we ask about the highest percentage accuracy achieved on the current FrontierMath dataset (Tiers 1-3),** not including Tier 4 problems even if they have been used in evaluations by the resolution date.

Resolution criteria

Resolution body: Epoch AI; fallback to FRI and FRI-appointed panel of experts if unavailable.

Resolution criteria: Ground truth

[140] *The following was not shown to participants*. This forecasting question was developed before the release of Tier 4 questions, hence the discussion starting with, “Epoch AI is currently adding.”
[141] *Link provided to participants:* https://x.com/tamaybes/status/1898131897458606506
[142] *Link provided to participants:* https://epoch.ai/frontiermath/tier-4

The question will be resolved by

- FrontierMath performance reported by Epoch AI, confirmed by an FRI staff.
- If credible reports by EpochAI are not available, FRI will resolve the question with input from at least 3 experts not in the LEAP panel selected for expertise in AI model evaluations.
- Cases in which we'd consider credible reports to be unavailable:
    - If there is evidence that tasks in the FrontierMath benchmark have been updated beyond adding Tier 4 problems: FRI will commission either Epoch AI or a third-party contractor to evaluate frontier models on the original FrontierMath benchmark (as of April 2025). If this is not possible, we resolve the question independently as above.
    - If there are no reports of FrontierMath performance in the year leading up to resolution date, for reasons including saturation (>90% performance): we resolve the question independently as above.
    - If there is evidence of training contamination leading to substantially increased performance, the resolution value may be adjusted or disqualified.

2. Autonomous Vehicle Trips

What percentage of U.S. ride-hailing trips will be provided by autonomous vehicles that are classified SAE Level 4 or above in the years 2027 and 2030?[143]

Background information

The SAE Levels of Driving Automation defines the level of automation involved in motor vehicles operating on roadways.[144] SAE Level 4 "High Automation" vehicles do not require any human intervention when automated driving features are engaged, but may only operate in limited conditions or environments.

Ride-hailing companies match passengers with vehicles for hire via electronic means (e.g., websites and mobile apps). As of May 2025, the only ride-hailing company providing U.S. based trips by autonomous vehicles at SAE Level 4 or above is Waymo. Waymo currently offers SAE Level 4 services in Phoenix and San Francisco, and plans to expand services to Atlanta,[145] Miami, and Washington D.C. in 2026.

[143] *Link provided to participants:* https://www.sae.org/blog/sae-j3016-update
[144] *Link provided to participants:* https://www.sae.org/blog/sae-j3016-update
[145] *Link provided to participants:* https://www.reuters.com/technology/alphabets-waymo-aims-2026-self-driving-ride-hailing-launch-washington-dc-2025-03-25/#:~:text=Waymo%20One%2C%20the%20company's%20fully,million%20paid%20trips%20in%202024

Several other companies have announced plans to develop autonomous vehicles for ride-hailing, such as Motional,[146] Uber,[147] and Lyft.[148]

Historical baseline
There is currently no official reporting on the percentage of U.S. ride-hailing trips provided by autonomous vehicles at SAE Level 4 or above. We estimate that this is **0.27%** as of Q4 2024.

More information about this estimate
- Waymo vehicles currently operate at SAE Level 4. In April 2025, Waymo reported delivering more than 250,000 paid rides per week in the U.S.,[149] approximately 36,000 rides a day on average.
- Uber and Lyft hold over 99% of the U.S. ride-hailing market share.[150] Using this and other publicly available data, we estimate approximately 13.6 million ride-hailing rides per day in the U.S. in 2024. You can read more about our reasoning here.[151]
- The percentage of U.S. ride-hailing trips provided by autonomous vehicles at SAE Level 4 or above (Waymo) is estimated to be 36000/13,600,000 = **0.265%**.

Resolution criteria

Resolution body. Financial reports by major ride-hailing companies; fallback to FRI and FRI-appointed panel of experts if unavailable.

Resolution criteria: Ground truth
This question will be resolved by:
- Credible reports about SAE level 4 and total ride-hailing trips, e.g., financial reports by major ride-hailing companies (companies with at least 1% market share of the U.S. ride-hailing market, or at least 5% of the U.S. autonomous vehicle ride-hailing market). We will resolve this question using data with a cutoff date of Jan 1, 2028 and Jan 1, 2031.
- If credible reports are not available, FRI will resolve the question with input from at least 3 experts not in the LEAP panel selected for knowledge of the autonomous vehicle industry.

[146] *Link provided to participants:* https://motional.com/
[147] *Link provided to participants:* https://investor.uber.com/news-events/news/press-release-details/2024/Wayve-and-Uber-Partner-to-Accelerate-the-Future-of-Automated-Driving/default.aspx
[148] *Link provided to participants:* https://www.lyft.com/blog/posts/autonomous-revolution-drivers
[149] *Link provided to participants:* https://x.com/Waymo/status/1915507165378257100
[150] *Link provided to participants:* https://secondmeasure.com/datapoints/rideshare-industry-overview/
[151] *Link provided to participants:* https://docs.google.com/document/d/1B7uykf5IdEMClcDwSt8yfAyU7Xpb4M9ItcnMVWizgmg/edit?tab=t.0

3. Occupational Employment Index

What will the percent change in the number of jobs (compared to Jan 1, 2025) in the U.S. be for white-collar, blue-collar, and service-sector occupations, by the following resolution dates?

- **Jan 1, 2028**
- **Jan 1, 2031**

Background information

White-collar occupations, as represented by
- Financial Managers (O*NET code 11-3031)[152]
- Human Resources Specialists (13-1071)[153]
- Architects (17-1011)[154]
- Lawyers (23-1011)[155]
- Public Relations Specialists (27-3031)[156]
- Software developers (15-1252)[157]

Blue-collar occupations, as represented by
- Construction Laborers (47-2061)[158]
- Electricians (47-2111)[159]
- Industrial Machinery Mechanics (49-9041)[160]
- Welders, Cutters, Solderers, and Brazers (51-4121)[161]
- Maintenance and Repair Workers, General (49-9071)[162]

Service-sector occupations, as represented by
- Registered Nurses (29-1141)[163]
- Elementary School Teachers (25-2021)[164]
- Customer Service Representatives (43-4051)[165]
- Waiters and Waitresses (35-3031)[166]
- Childcare Workers (39-9011)[167]

---

[152] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/11-3031.00
[153] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/13-1071.00
[154] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/17-1011.00
[155] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/23-1011.00
[156] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/27-3031.00
[157] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/15-1252.00
[158] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/47-2061.00
[159] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/47-2111.00
[160] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/49-9041.00
[161] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/51-4121.00
[162] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/49-9071.00
[163] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/29-1141.00
[164] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/25-2021.00
[165] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/43-4051.00
[166] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/35-3031.00
[167] *Link provided to participants:* https://www.onetcodeconnector.org/ccreport/39-9011.00

You can find detailed descriptions of these occupations and related occupations by looking up O*NET codes (also linked in O*NET code above): https://www.onetcodeconnector.org/.

Historical baseline

We estimate that annual percent changes in each group of occupations in May 2024 as compared to May 2023 are:

- White-collar: **+1.61%**
- Blue-collar: +**2.67%**
- Service-sector: **+0.44%**

White-collar occupations

Occupation	O*Net Code	Employment (May 2023)	Employment (May 2024)
Financial Managers	11-3031	787,340	818,620
Human Resources Specialists	13-1071	895,970	917,460
Architects	17-1011	111,170	111,140
Lawyers	23-1011	731,340	747,750
Public Relations Specialists	27-3031	275,550	280,590
Software Developers	15-1252	1,656,880	1,654,440
Total	-	4,458,250	4,530,000[168]

Percent change in 2024, compared to 2023: +1.61%

[168] The correct total is 4,178,000, but we preserve the total (mistakenly) reported to participants here.

Blue-collar occupations

Occupation	O*Net Code	Employment (May 2023)	Employment (May 2024)
Construction Laborers	47-2061	1,019,090	1,057,660
Electricians	47-2111	712,580	742,580
Industrial Machinery Mechanics	49-9041	412,650	421,940
Welders, Cutters, Solderers, and Brazers	51-4121	421,730	424,040
Maintenance and Repair Workers, General	49-9071	1,503,150	1,531,700
Total		4,069,200	4,530,000[169]

Percent change in 2024, compared to 2023: +2.67%

Service-sector occupations

Occupation	O*Net Code	Employment (May 2023)	Employment (May 2024)
Registered Nurses	29-1141	3,175,390	3,282,010
Elementary School Teachers	25-2021	1,410,070	1,393,310
Customer Service Representatives	43-4051	2,858,710	2,725,930
Waiters and Waitresses	35-3031	2,237,850	2,302,690
Childcare Workers	39-9011	497,450	520,180
Total	-	10,179,470	10,224,120

Percent change in 2024, compared to 2023: +0.44%

[169] The correct total is 4,178,000, but we preserve the total (mistakenly) reported to participants here.

The above historical baselines are computed from May 2024 and May 2023 occupational employment data by the U.S. Bureau of Labor Statistics. [170,171]

You can also view occupational employment data in 2023 and official employment for 2033.[172]

Resolution criteria

Resolution body: U.S. Bureau of Labor Statistics; fallback to FRI and FRI-appointed panel of experts if unavailable. [173]

Resolution criteria: Ground Truth
The question will be resolved by

- Credible reports of U.S. occupational employment data, as confirmed by FRI staff.
- Where employment data for a specific occupation is not available, we will calculate percentage change using the data for the closest parent category.
- If credible reports are not available, FRI will resolve the question with input from at least 3 experts not in the LEAP panel selected for expertise in U.S. labor statistics.

4. General AI Progress

At the end of 2030, what percent of LEAP panelists will choose "slow progress," "moderate progress," or "rapid progress" as best matching the general level of AI progress?

In the following scenarios, we consider the development of AI capabilities, not adoption. Regulation, social norms, or extended integration processes could all prevent the application of AI to all tasks of which it is capable.

Reasonable people may disagree with our characterization of what constitutes slow, moderate, or rapid AI progress. Or they may expect to see slow progress observed with some AI capabilities and moderate or fast progress in others. Nevertheless, we ask you to select which scenario, in sum, you feel best represents your views.

We consider a capability to have been achieved if there exists an AI system that can do it:

- ***Inexpensively**: with a computational cost not exceeding the salary of an appropriate 2025 human professional using the same amount of time to attempt the task.*
- ***Reliably**: what this means is context-dependent, but typically we mean as reliably as, or more reliably than, a human or humans who do the same tasks professionally in 2025.*

---

[170] *Link provided to participants:* https://www.bls.gov/news.release/ocwage.t01.htm
[171] *Link provided to participants:* https://www.bls.gov/oes/2023/may/oes_nat.htm
[172] *Link provided to participants:* https://www.bls.gov/emp/tables/occupational-projections-and-characteristics.htm
[173] *Link provided to participants:* https://data.bls.gov/oes/#/industry/000000

Slow Progress

By the end of 2030 in this slower-progress future, AI is a capable assisting technology for humans; it can automate basic research tasks, generate mediocre creative content, assist in vacation planning, and conduct relatively standard tasks that are currently (2025) performed by humans in homes and factories.

Researchers can benefit from literature reviews on almost any topic, written at the level of a capable PhD student, yet AI systems rarely produce novel and feasible solutions to difficult problems. As a result, genuine scientific breakthroughs remain almost entirely the result of human-run labs and grant cycles. Nevertheless, AI tools can support other research tasks (e.g., copy editing and data cleaning and analysis), freeing up time for researchers to focus on higher-impact tasks. AI can handle roughly half of all freelance software-engineering jobs that would take an experienced human approximately 8 hours to complete in 2025, and if a company augments its customer service team with AI, it can expect the model to be able to resolve most complaints.

Writers enjoy a small productivity boost; models can turn out respectable short stories, but full-length novels still need heavy human rewriting to avoid plot holes or stylistic drift. AI can make a 3-minute song that humans would blindly judge to be of equal quality to a song released by a current (2025) major record label. At home, an AI system can draft emails, top up your online grocery cart, or collate news articles, and—so long as the task would take a human an hour or less and is well-scoped—it performs on par with a competent human assistant. With a few prompts, AI can create an itinerary and make bookings for a weeklong family vacation that feels curated by a discerning travel agent.

Self-driving car capabilities have advanced, but none have achieved true level-5[174] autonomy. Meanwhile, household robots can make a cup of coffee and unload and load a dishwasher in some modern homes—but they can't do it as fast as most humans and they require a consistent environment and occasional human guidance. In advanced factories, autonomous systems can perform specific, repetitive tasks that require precision but little adaptability (e.g., wafer handling in semiconductor fabrication facilities).

Moderate Progress

[174] Level 5 autonomy is defined as "can operate the vehicle on-road anywhere within its region of the world and under all road conditions in which a conventional vehicle can be reasonably operated by a typically skilled human driver. This means, for example, that there are no design-based weather, time-of-day, or geographical restrictions on where and when the ADS can operate the vehicle. However, there may be conditions not manageable by a driver in which the ADS would also be unable to complete a given trip (e.g., white-out snow storm, flooded roads, glare ice, etc.) until or unless the adverse conditions clear." *Link provided to participants:* https://saemobilus.sae.org/standards/j3016_202104-taxonomy-definitions-terms-related-driving-automation-systems-road-motor-vehicles#view

By the end of 2030, in this middle-of-the-road future, AI is capable of being an effective collaborator across creative, corporate, and technical arenas.

R&D teams can lean on autonomous lab systems with human research assistants to make rapid advances in solar-cell chemistries or fusion-reactor components, although the inventions are improvements rather than revolutions. Almost all freelance software-engineering jobs requiring 5 days or fewer of effort from experienced humans can be tackled by AI, and if a company wants to replace nearly all of its human customer service team with AI, it can probably get away with it, provided customers are not opposed in principle to AI customer service.

AI can draft 100-thousand-word novels solid enough for current (2025) mainstream publishers to release with typical levels of editing by a human publisher and can create a 5-minute song that humans would blindly judge to have breakout potential, exceptional emotional depth, and be of equal quality to the best songs released by current (2025) record labels. An AI agent can not only craft a well-curated weeklong vacation itinerary, but it can also make last-minute adjustments to bookings in response to unexpected weather, travel delays, or individual requests. Boardrooms are still run by people, but executive teams can delegate multiday operational projects to agentic software that reliably juggles vendors, budgets, and contingencies.

Level-5 robo-taxis that can go anywhere a human driver can go finally exist, as do robots that can navigate any home in the US, make a cup of coffee and unload and load a dishwasher as fast and reliably as most humans and without any human guidance. Robots in advanced factories can adapt to different tasks without extensive reprogramming or equipment changes, making real-time decisions, and have increased physical mobility.

Rapid Progress

By the end of 2030, in the rapid-progress world, AI systems are capable of competing with the best human minds and workers, and can surpass them.

Human creativity and leadership remain valued, but mostly for setting high-level vision—day-to-day execution can be left to silicon-based systems. Autonomous researchers can collapse years-long research timelines into days, weeks, or months, creating game-changing technologies, such as materials that revolutionize energy storage, or bespoke cancer cures. No human freelance software engineer can outperform AI. The same goes for customer service (e.g., call center and support chat), paralegal, and administrative workers (e.g., bookkeepers or scheduling assistants).

Indeed, models have become so capable that AI can create an album of the same caliber as the Grammy Album of the Year. Additionally, a single AI agent can generate a Pulitzer- (or Booker Prize-) caliber novel according to current (2025) standards, adapt the book into an engaging two-hour movie, negotiate the resulting book and movie contracts, and launch the marketing

campaigns for both while its sibling agents manage the book publishing company and movie studio at the level of highly competent CEOs.

Not only do Level-5 robo-taxis exist, but they are, on average, 99.9% safer than human-piloted cars and can venture anywhere off-road that a competent human driver can. Meanwhile, robots can navigate an arbitrary home anywhere in the world, make a cup of the most popular local hot beverage, clean and put away the dishes according to the local custom, fix any plumbing issues that arise while they're doing the dishes--and they can do it all faster and more reliably than most humans and without human guidance. Robots in advanced factories can autonomously perform the full range of tasks requiring the highest levels of dexterity, coordination, and adaptive decision-making.

Resolution criteria

Resolution body: FRI; LEAP Panel

Resolution criteria: Ground truth
This question will be resolved by FRI surveying the LEAP panel, or another expert panel with similar representation to LEAP, on the following question in December 2030: **"Which scenario best matches the level of AI capabilities progress to date?"**

As a reminder, the current LEAP panel consists of experts in:
- Computer Science: 40%
- Economics: 10%
- Industry: 15%
- Policy: 35%

Future LEAP panel composition may change depending on enrollment. If this changes significantly, or if the LEAP panel is not available by the resolution date, FRI will recruit an expert panel with the above composition to resolve the question.

The forecasts will be elicited using an intersubjective metric which incentivizes respondents to be truthful. You can read more about the intersubjective metrics we may use in the "Instructions" tab.

5. Technological Richter Scale

At the end of 2040, what is the probability for AI achieving the following levels of net impact on human society as compared to the impact of past technological events?

Background information
Nate Silver's book "On the Edge" proposes the technological Richter scale (TRS) which, analogous with earthquake magnitudes, rates the impact of technologies on a roughly logarithmic scale. Some levels are associated with a frequency inversely related to the magnitude of impact.

Resolution criteria

Resolution body: FRI; LEAP Panel

Resolution criteria:
This question will not be resolved as it is inherently subjective.

Please choose the scenario which is closest to your true beliefs for this question. We encourage you to explain your thinking. We will give out ten $200 prizes for rationales that our team votes as the most thoughtful and informative for our analysis.

What is the probability that AI will fall in each of the following TRS levels by the end of 2040?

- Level 5: Technology of the month, a commercially successful invention important in its category, e.g., a leading brand of windshield wipers.
- Level 6: Technology of the year, an invention that is disruptive in its field and has some ripple effects beyond it, e.g., VCR.
- Level 7: Technology of the decade, measurably impacts daily lives, e.g., credit cards, social media.
- Level 8: Technology of the century, broadly disruptive in society, e.g., electricity, automobiles.
- Level 9: Technology of the millennium, altering the course of human history, e.g., agriculture, the wheel, Industrial Revolution.
- Level 10: Epoch-defining event that alters the fate of the planet, e.g., the rise of humans.

6. Cognitive Limitations of AI

(1) What do you see as the main cognitive limitations of AI systems in 2025?

Background information
"Cognitive limitations" are defined as deficiencies in AI capabilities that constrain their economic value or real-world impact. As of May 2025, examples include hallucination, short task horizons, and limited metacognition. This question elicits open-ended forecasts about the types of cognitive limitations likely to remain prominent in future leading AI systems.

For both (1) and (2), please feel free to answer in prose or a bulleted list.

Resolution criteria

Resolution criteria:
This pair of open-ended questions will not be resolved.

We will:

- Collect free-text responses about expected cognitive limitations.

- Group responses into a structured list of key themes or categories.
- Release a forecasting question about cognitive limitations of AI in a future survey, scored using an intersubjective metric.

Rationale prizes:
We will give out ten $200 prizes for rationales across all questions in this survey that our team votes as the most thoughtful and informative for our analysis.

For example, you are likely to win a prize if you propose an important cognitive limitation that is rarely mentioned by other panelists but upon reflection many panelists would agree with, or give the strongest, most convincing argument for or against a cognitive limitation.

(2) What will AI experts identify as the main cognitive limitations of state-of-the-art AI systems by 2030?

Background information
By "AI experts," we mean people with the most extensive experience designing, developing, using, and analyzing AI systems. Please assume that these experts are reporting their honest beliefs about AI systems.

Resolution criteria
Same as in (1).

Appendix E.II. Survey Questions: Wave 2

1. Millennium Prize

Will AI solve or substantially assist in solving a Millennium Prize Problem in mathematics by the following resolution dates?

- **Dec 31, 2027**
- **Dec 31, 2030**
- **Dec 31, 2040**

Background Information
In 2000, the founding Scientific Advisory Board of the Clay Mathematics Institute (CMI) of Cambridge, Massachusetts, chose seven of the most difficult and deepest mathematical problems unsolved at that time - the Millennium Prize Problems.[175]

For each of the problems, $1 million is designated to its solution.

[175] *Link provided to participants:* https://www.claymath.org/millennium-problems/

As of July 2025, only one of the seven problems has been solved (the Poincaré Conjecture, solved by Perelman in 2002).[176]
The yet unsolved problems are:

- Birch and Swinnerton-Dyer Conjecture
- Hodge Conjecture
- Navier-Stokes Equation
- P vs NP
- Riemann Hypothesis
- Yang-Mills & the Mass Gap

Resolution Criteria

Resolution Body: Clay Mathematics Institute; FRI

Resolution Criteria: Ground truth
The question will be resolved by

- Credible reports by CMI that a solution has been found by AI or humans who testify to substantial AI assistance.
    - We will determine that there is "substantial assistance" if a majority of a panel of 5 expert mathematicians specializing in the field of the solved problem agree with the statement that "The AI-assisted part of the solution is original, important, and likely could not have been produced without the AI."
- If no credible reports are available, the question resolves to "No."

2. Diffusion of AI Across Sciences

What percent of publications in the fields of Physics, Materials Science, and Medicine in 2030 will be 'AI-engaged' as measured in a replication of this study?[177]

Background Information
A 2024 study by researchers at Harvard University and the University of Chicago tracked scholarly engagement with AI across 20 scientific fields by measuring the change in percentage of "AI-engaged publications"[178] within each field.

"AI-engaged publications" are papers with abstracts that contain at least one keyword related to contemporary approaches to AI. The list of keywords is generated by the authors using text-processing methods and expert review (e.g., 'artificial intelligence' and 'deep neural network'). AI-engaged papers are then identified by binary classification of their abstracts.

[176] *Link provided to participants:* https://arxiv.org/abs/math/0211159
[177] *Link provided to participants:* https://arxiv.org/pdf/2405.15828
[178] *Link provided to participants:* https://arxiv.org/pdf/2405.15828

According to the study, modes of engagement identified in this way include “the development of novel AI theory and approaches, technologies, or applications; the general use of AI models for domain-specific tasks; and critical engagement with AI.”

The study considers all English language papers in the Semantic Scholar Academic Graph dataset from 1985 to 2022. Papers are identified as belonging to a distinct scientific field by the Semantic Scholar paper metadata dataset.

Methodology Details
AI-engaged keyword list: The keyword list is not publicly available. A summary of the generation procedure is as follows:.

- Start with a small “seed list” of obvious AI terms (e.g., “artificial intelligence,” “deep neural networks,” “machine learning”)
- Iteratively expanded this list through five main steps:
    - Download papers from relevant arXiv categories (cs.AI, cs.LG, cs.NE, stat.ML).
    - Used word2vec to create vector representations of all words in the corpus.
    - For each seed term, find the 10 most similar terms using cosine similarity.
    - Ask active AI experts from the University of Chicago, Harvard, and the Argonne National Laboratory to review the expanded list to remove false positives and add missing terms.
- Repeated the expansion and evaluation process twice more to create the final keyword list.

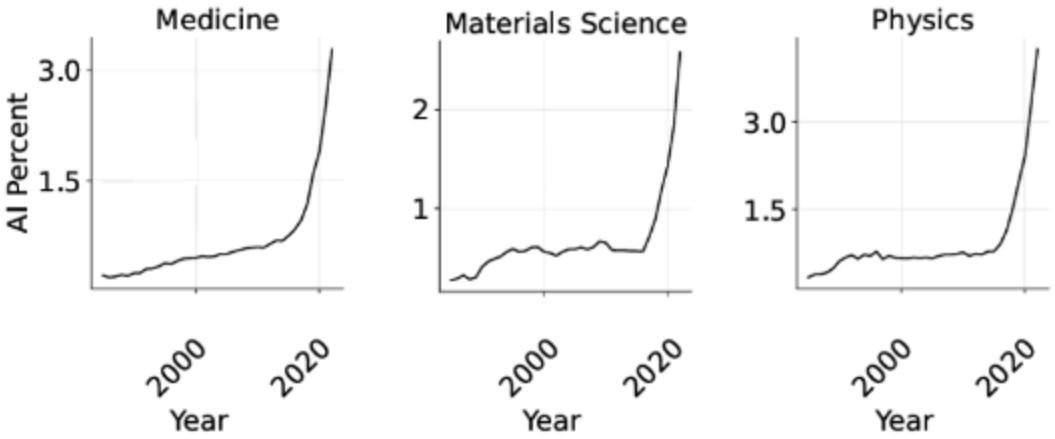
Papers considered: You can find details on the scope of publications considered on pages 3-4 of the paper.[179]

Historical Baseline
The study finds that, in 2022, the percentage of AI-engaged publications in scientific fields of interest as follows (rounded to the nearest percent):

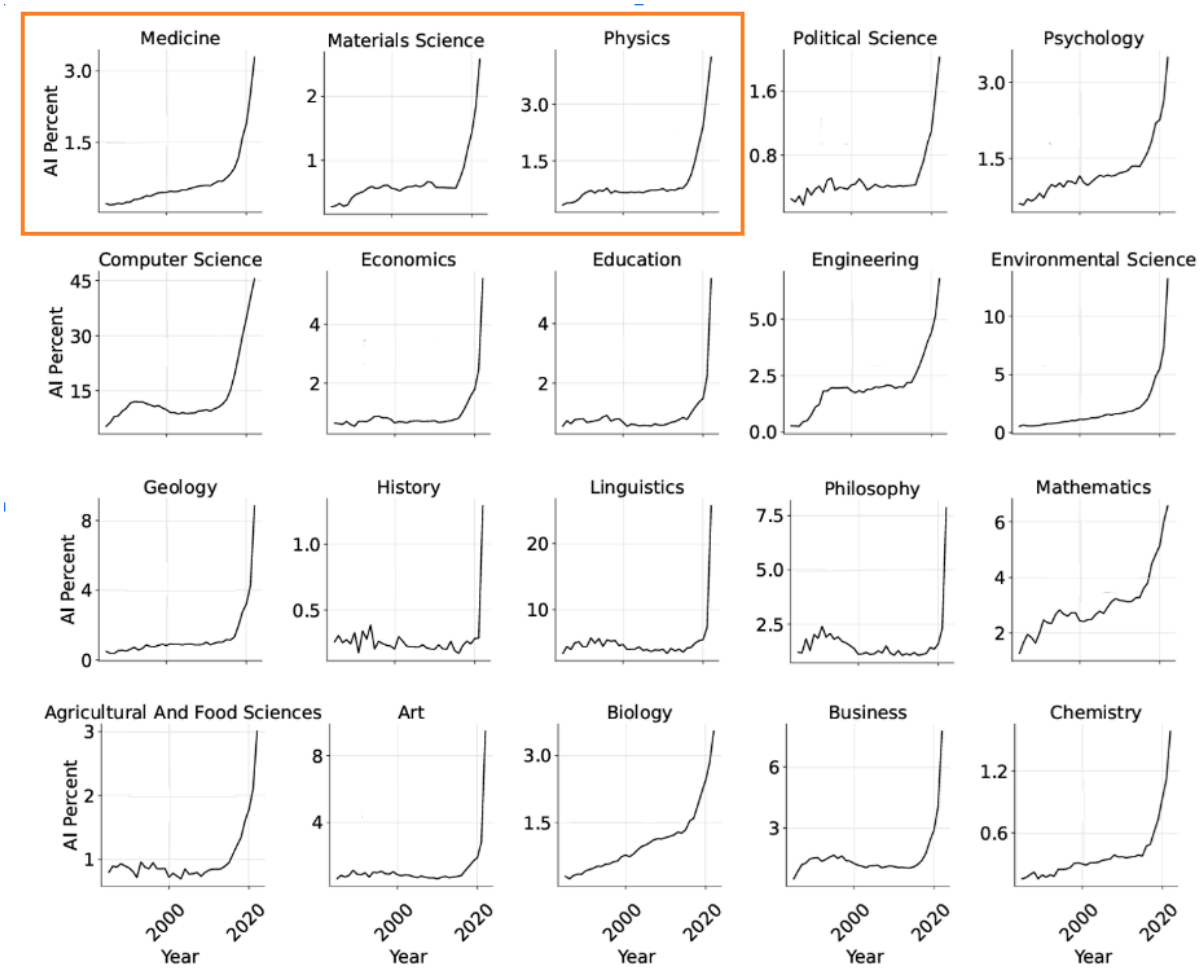
- Physics: 3%
- Materials Science: 2%
- Medicine: 3%

You can view detailed results in the below figure, which is a subset of results from the original paper. The y-axis of each plot in the figure shows the percentage of AI-engaged publications in each field from 1985 to 2022. Note that the scale of the y-axis is different for each field.

[179] *Link provided to participants:* https://arxiv.org/pdf/2405.15828



You can view results for additional fields in the below figure.



Resolution Criteria

Resolution Body: FRI

Resolution Criteria: Ground truth

- If available, use a future iteration of this study measuring the percentage of AI-engaged publications in scientific fields of interest
- If no future iteration is published until January 31, 2031, FRI researchers will replicate or commission a replication of this paper.

3. Drug Discovery

What percent of sales of recently approved U.S. drugs will be from AI-discovered drugs and products derived from AI-discovered drugs in the years 2027, 2030 and 2040?

Background Information

Recently approved U.S. drugs:

- We define "recently approved U.S. drugs" as drugs which received approval for sale by the U.S. Food and Drug Administration (FDA) no earlier than in the year preceding the resolution year.
- For example, for forecasts resolving in 2027, we would consider all drugs approved on or after Jan 1, 2026.

FDA Drug Approval:

- *In **2023** and **2024**,[180,181] the FDA approved **55** and **50** new drugs for sale respectively.*

Revenues within 3 Years:

- A report by L.E.K. Consulting finds that the annual revenue of a new drug in the first three years after launch as a *percentage of its peak annual revenue* is 20% in Year 1,[182] 41% in Year 2, and 52% in Year 3 (Fig. 1 in report).
- A second report by L.E.K consulting finds that the median revenue of a drug in the third year after approval is $260 million, while the average revenue is $600 million.[183]

Market Size:

- According to ASPE,[184] total sales revenues of prescription drugs amounted to $716 billion in 2022.

---

[180] *Link provided to participants:* https://pmc.ncbi.nlm.nih.gov/articles/PMC10856271/
[181] *Link provided to participants:* https://www.fda.gov/drugs/novel-drug-approvals-fda/novel-drug-approvals-2024
[182] *Link provided to participants:* https://www.lek.com/sites/default/files/PDFs/Launch-Monitor-Lessons-Learned_v2.pdf
[183] *Link provided to participants:* https://www.lek.com/insights/hea/us/ei/variability-large-pharma-launch-performance
[184] *Link provided to participants:* https://aspe.hhs.gov/sites/default/files/documents/4326cc7fe43bc11770598cf2a13f478c/international-market-size-prices.pdf

Average Trial Phase Duration:

- BIO analyzed 6,151 successful phase transitions from 2011-2020.[185] On average, a drug took 10.5 years to get from Phase I to approval. Average durations for each phase were:
    - 2.3 years for Phase I
    - 3.6 years for Phase II
    - 3.3 years for Phase III
    - 1.3 years between Phase III and regulatory approval

More Information About Derived Products

The FDA defines a biosimilar as a biologic that is highly similar, but not structurally identical, to an existing FDA-approved biologic with no meaningful differences in efficacy, safety, or purity. See an example of biosimilars here.[186]

Drug review process:

- In the United States, drugs must obtain the Investigational New Drug (IND) approval in order to enter clinical trials, then undergo Phase I, II, and III trials before receiving approval for sale by the FDA.
- A study in 2022 reported that there was publicly available information on about 160 AI drug discovery programs,[187] of which 15 products are reportedly in clinical trials, representing a rate of **9.3%** of AI-discovered drugs entering clinical trials.
- A study by Boston Consulting Group in June 2024 reported that 67 AI-discovered molecules were in ongoing trials as of December 2023.[188] Among these, 24 AI-discovered molecules had completed **Phase I** trials, at a success rate of **80–90%**, which is substantially higher than historical industry averages that range from ~40% to ~55–65%. In **Phase II** the success rate is **~40%**, comparable to historic industry averages. Historically, drugs that enter phase III have a success rate of ~58%.

Examples of AI systems for biomedical discovery:

- AlphaFold 3 is an AI system developed by Google DeepMind that predicts the structures and interactions of proteins, DNA, RNA, and ligands.[189,190] In predicting the interactions of proteins with other molecules, it is said to achieve “50% improvement [in accuracy] compared with existing prediction methods.”

---

[185] *Link provided to participants:* https://go.bio.org/rs/490-EHZ-999/images/ClinicalDevelopmentSuccessRates2011_2020.pdf
[186] *Link provided to participants:* https://www.gabionline.net/biosimilars/general/biosimilars-of-adalimumab
[187] *Link provided to participants:* https://www.cas.org/resources/cas-insights/ai-drug-discovery-assessing-the-first-ai-designed-drug-candidates-to-go-into-human-clinical-trials
[188] *Link provided to participants:* https://www.sciencedirect.com/science/article/pii/S135964462400134X?via%3Dihub
[189] *Link provided to participants:* https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/
[190] *Link provided to participants:* https://deepmind.google/

- Isomorphic Labs,[191] a drug discovery company, is reportedly collaborating with pharmaceutical companies to apply AlphaFold 3 and other proprietary AI models to real-world drug design challenges. Evo is a publicly available AI model that can "deduce how bacterial and viral genomes operate and use that information to design new proteins and even whole microbial genomes."[192]

Historical Baseline

Since no AI-discovered drug has been approved for U.S. sale, the current percentage of sales of recently approved U.S. drugs generated by AI-discovered drugs and derived products is **0%**. As pharmaceutical companies do not fully disclose their research processes, the real percentage could be higher. Future sales will be measured in USD.

Resolution Criteria

Resolution Body: FDA database of annual Novel Drug Approvals; Financial reports by U.S. pharmaceutical companies; fallback to FRI and FRI-appointed panel of experts if unavailable.

Resolution Criteria: Ground truth

The question will be resolved by

- Sales revenue of specific drugs in FDA's database of Novel Drug Approvals,[193] and the use of AI/ML in drug discovery by sources such as the FDA and U.S. pharmaceutical companies, as confirmed by an FRI staff.
    - We will identify AI-discovered drugs by iterating through the list of FDA-approved drugs in the resolution year and the year prior, and checking that reports on the drug discovery process given by the company which proposed the drug mentions at least one significant use of AI techniques as we defined above.
- If the FRI staff is not able to confidently confirm resolution, FRI will resolve the question with input from at least 5 experts not in the LEAP panel, selected for knowledge of AI drug discovery and the U.S. pharmaceutical industry.
    - We will consider a drug AI-discovered if a majority of this expert panel specializing in biopharmaceutical science agree with the statement that "The AI-assisted part of the drug discovery process constitutes 'significant use of AI techniques' as defined in the question."

4. Electricity Consumption

What percent of U.S. electricity consumption will be used for training and deploying AI systems in the years 2027, 2030 and 2040?

---

[191] *Link provided to participants:* https://www.isomorphiclabs.com/articles/rational-drug-design-with-alphafold-3
[192] *Link provided to participants:* https://www.science.org/content/article/meet-evo-dna-trained-ai-creates-genomes-scratch
[193] *Link provided to participants:* https://www.fda.gov/drugs/novel-drug-approvals-fda/novel-drug-approvals-2024

Background Information

AI model training and deployment occur mainly through the use of data centers,[194] which are giant computing centers which host computers ("servers").

Servers in data centers are split into two main categories: conventional and AI specialized servers,[195] the latter of which are better designed for and dedicated to AI workloads. This question asks about the electricity consumption of AI specialized servers, regardless of the end application which is in practice difficult to determine.

Historical Baseline

There is currently no official reporting on the percentage of U.S. electricity consumption used for training and deploying AI systems. We estimate that this is **1.0%** in 2024. You can read more about our reasoning here.[196]

Potentially relevant sources which estimate future U.S. electricity consumption by data centers:

- The 2024 United States Data Center Energy Usage Report estimates that **data centers will consume approximately 6.7 to 12% of total U.S. electricity by 2028.**[197]
- A 2025 report by the National Academy of Sciences estimates U.S. **data center demand as a percentage of total U.S. electricity demand under four different growth scenarios as between 4.6-9.1% in 2030.** [198]

Potentially relevant sources which estimate electricity consumption due to AI:

- In February 2025, Epoch AI estimates that "AI could reach fairly eye-popping levels of energy usage by 2030, on the order of **10% of U.S. electricity."**[199]
- A 2024 white paper estimates **40%** of AI energy usage by model training and development, and 60% for deployment, based on analysis of existing models.[200]

Resolution Criteria

---

[194] *Link provided to participants:* https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
[195] *Link provided to participants:* https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf
[196] *Link provided to participants:* https://docs.google.com/document/d/1S1h7sH8KnP42NepoSRRoVTu5n_uBe2SxB6gewVjqdWk/edit?usp=sharing
[197] *Link provided to participants:* https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf
[198] *Link provided to participants:* https://nap.nationalacademies.org/read/29101/chapter/4#11
[199] *Link provided to participants:* https://epoch.ai/gradient-updates/how-much-energy-does-chatgpt-use#:~:text=With%20reasonable%20and%20somewhat%20pessimistic,queries%20with%20very%20long%20inputs.
[200] *Link provided to participants:* https://www.epri.com/research/products/3002028905

Resolution Body: U.S. Energy Information Administration; fallback to FRI and FRI-appointed panel of experts if unavailable.

Resolution Criteria: Ground truth
The question will be resolved by

- Credible reports of total U.S. electricity consumption and fraction of U.S. used for AI training and deployment, by sources such as the U.S. Energy Information Administration, the International Energy Agency, and peer-reviewed publications.
- If credible reports are not available, FRI will resolve the question with input from at least 3 AI experts selected for expertise in estimating AI's energy consumption.

5. Cognitive Limitations, Part II

By the end of 2030, what percent of LEAP expert panelists will agree that each of the following is a serious cognitive limitation of state-of-the-art AI systems?

Note: This question builds on input from LEAP participants in the previous wave. For that reason, we are including it in this wave, although the theme of this survey is AI for science.

Background Information
A cognitive limitation is considered serious if it continues to constrain the economic value or real-world impact of state-of-the-art AI systems.

Given the overlapping and causally interacting nature of cognitive limitations, reasonable people may disagree with our sorting and labeling process. Nevertheless, we ask you to consider the limitations below, as they have been presented, as best you can.

In Wave 1 of LEAP, panelists proposed that current AI systems face the following cognitive limitations (in no particular order):

1. **Hallucination / Inaccuracy**: they give plausible-sounding but incorrect or fabricated information.
2. **Shallow reasoning**: their reasoning capabilities are shallow beyond math and coding and lack "genuine" causal and logical reasoning.
3. **Lack of long-term memory**: they cannot retain and utilize information across sessions or long interactions.
4. **Limited ability to generalize**: they perform badly in tasks they have not been trained on and are limited by the quality of human-generated training data, limiting their creativity.
5. **Limited metacognition and continual learning**: they cannot reliably assess and regulate their own cognitive processes, e.g., notice and correct their own errors, question previous assumptions, know when to ask for help or to defer, know how to improve a capability they don't have, continue to acquire new capabilities after training, etc.
6. **Limited Embodiment / Robotics**: they perform badly in processing multimodal inputs beyond text and vision and at navigating the physical world.

7. **Limited inter-system collaboration:** they cannot effectively coordinate with other AI systems to negotiate goals, allocate subtasks, and work in parallel under standardized protocol.

Resolution Criteria

Resolution Body: FRI; LEAP Panel

Resolution Criteria: Ground truth
This question will be resolved by FRI surveying the LEAP panel, or another expert panel with similar representation as LEAP, on the following question in January 2031: **"Do each of the following cognitive limitations of AI systems remain serious as of today? Choose 'Agree' or 'Disagree'."**

The resolution values will be generated by measuring the percentage of "agree" responses across all expert panelist responses for each cognitive limitation.

As a reminder, the current LEAP panel consists of experts in roughly the following proportions:
- Computer Science: 30%
- Economics: 20%
- Industry: 15%
- Policy: 35%

Future LEAP panel composition may change depending on enrollment. If this changes significantly, or if the LEAP panel is not available by the resolution date, FRI will reweight future panel responses to produce an aggregated hypothetical panel with similar representativeness to resolve the question. If this cannot be done for methodological reasons, FRI will recruit an expert panel with a composition similar to the above to resolve the question.

The forecasts about limitations by the end of 2030 will be elicited using an intersubjective metric that incentivizes respondents to be truthful. You can read more about the intersubjective metrics we may use in the "Instructions" tab.

6. Barriers to Adoption, Part I

What do you see as the main barriers to adopting current or future state-of-the-art AI systems for broader use in society?

Note: This question is not specifically about the theme for this survey, AI for science.

Background Information
Barriers to adoption are defined as factors that prevent or slow down AI systems being used more widely and more intensively. As of 2025, examples include unavailability of quality training data for use case, concerns about robustness to outliers, cost of compute, lack of AI-related

skills, and regulation or legal restrictions, etc. This question elicits open-ended forecasts about the types of adoption barriers likely to remain prominent for future AI systems.

"Broader use in society" refers to AI applications beyond using AI for frontier AI development. Examples include augmenting or automating routine work, scientific research and discovery, personal entertainment, uses for social good, etc.

Resolution Criteria

This pair of open-ended questions will not be resolved.

We will:

- Collect free-text responses about expected barriers to adoption.
- Group responses into a structured list of key themes or categories.
- Release a forecasting question about barriers to adopting AI systems in a future survey, scored using an intersubjective metric.

Rationale Prizes:
*We will give out **ten $200 prizes** for rationales across all questions in this survey that our team votes as the most thoughtful and informative for our analysis.*

For example, you are likely to win a prize if you propose an important barrier to adoption that is rarely mentioned by other panelists but upon reflection many panelists would agree with, or give the strongest, most convincing argument for or against a barrier to adoption**.**

Appendix E.III. Survey Questions: Wave 3

1. AI Investment

What will be the global private investment (in billion USD) in AI in the following calendar years?[201]

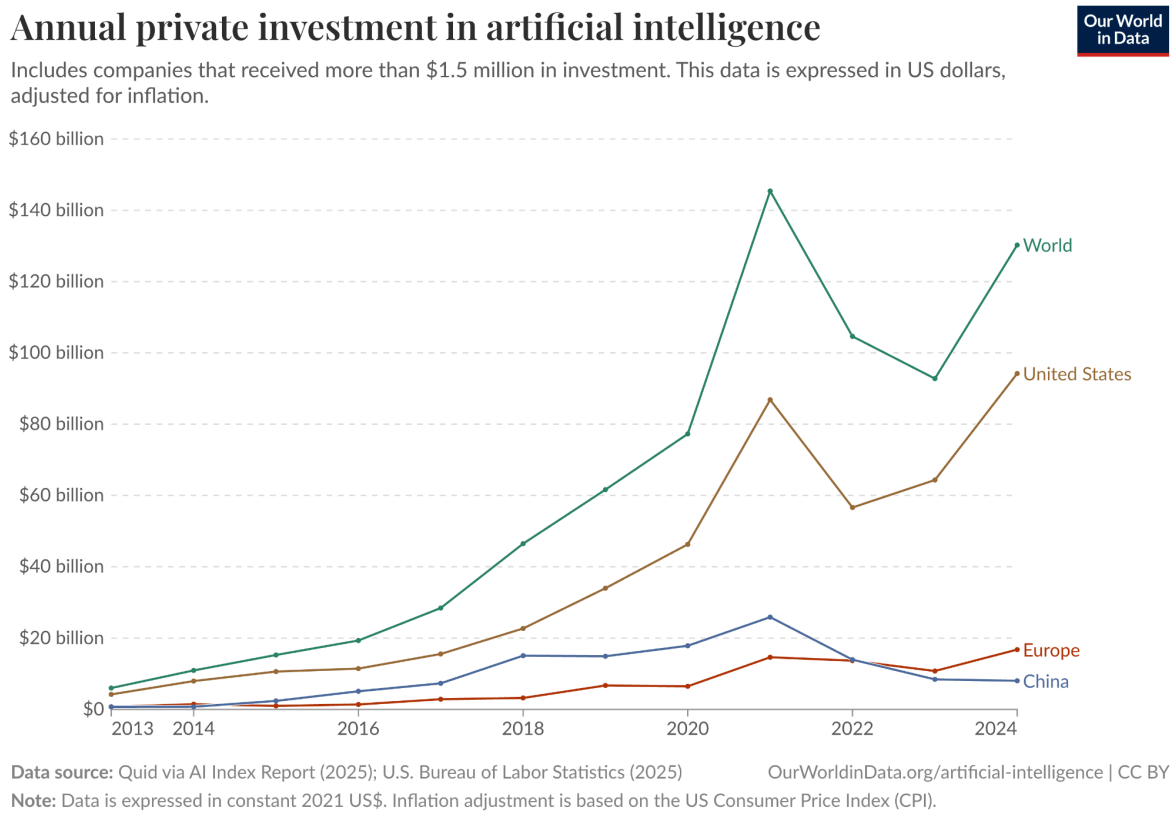
- **2027**
- **2030**

Background. According to the AI Index Report (2025),[202] private investment in AI includes investment in AI startups that have received over $1.5 million in investment since 2013 (pp. 248).

[201] *Link provided to participants:* https://ourworldindata.org/grapher/private-investment-in-artificial-intelligence
[202] *Link provided to participants:* https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

Our World in Data (OWID) notes that this indicator is likely to underestimate total global AI investment. More details on what is likely excluded are here.[203]

Historical baseline. According to OWID, global private AI investment was approximately **$130 billion** in 2024.[204]



Resolution Body: *Our World in Data*;[205] fallback to FRI or FRI-appointed panel of experts

Resolution Criteria: Ground Truth

This question will be resolved by:

- Total global private AI investment (in USD) for the resolution year, as reported by Our World in Data.
- If no credible report is available, FRI will produce an independent estimate with input from at least three AI experts selected for expertise in tracking AI investments.

[203] *Link provided to participants:*https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type#sources-and-processing
[204] *Link provided to participants:* https://ourworldindata.org/grapher/private-investment-in-artificial-intelligence#sources-and-processing
[205] *Link provided to participants:* https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type

2. Generative AI Use Intensity

What percent of work hours in the U.S. at the following dates will be estimated as assisted by generative AI, according to a future iteration of the St. Louis Fed study or a similar study selected by an FRI-appointed expert panel?

- **Dec 31, 2025**
- **Dec 31, 2027**
- **Dec 31, 2030**

Background: A June 2025 study by the Federal Reserve Bank of St. Louis,[206] based on the Real-Time Population Survey (N=3,216), estimated that 1.3–5.4% of all U.S. work hours were assisted by generative AI based on self-reports via a nationally representative survey. This question tracks how much self-reported generative AI adoption intensifies in work settings, as measured by updated or similar surveys.

Historical baseline. We estimate that the fraction of work hours assisted by generative AI in the U.S. as of September 2024 is **2.0%**.[207] This is the average of the estimated range in the above study.

Resolution Body: FRI staff; fallback to LEAP panel nowcast.

Resolution Criteria: Dynamic Resolution

- If available, use a future iteration of the St. Louis Fed study estimating U.S. work hours assisted by generative AI.
- If no future iteration is published, use a similar study selected by FRI based on comparability (e.g., national coverage, survey-based, measuring work-hour assistance by generative AI).
- If a range rather than a point estimate is reported, we will resolve the question using the average of the reported range.
- If no acceptable study is available by the resolution date, fallback to LEAP panel dynamic nowcast.

3. Personalized Education

Note: this question was excluded from our analysis due to substantial misinterpretation.

What percentage of weekly instructional hours on average will K-12 students in the United States spend using AI-powered tutoring or teaching tools during instructional hours?

- **Dec 31, 2027**

[206] *Link provided to participants:* https://fedinprint.org/item/fedlwp/98805/101172
[207] This estimate is based off of an earlier version of the paper, hence the discrepancy.

- **Dec 31, 2030**

Background: As of 2024, AI-powered education tools (such as research and writing tools and AI tutors) are increasingly used in schools around the world. A 2024 report by Cengage Group found that AI usage among global K-12 students rose sharply, with self-reported student usage at any point in the past increasing from 37% in 2023 to 75% in 2024.[208]

Historical baseline: There is currently no official reporting on the time spent by K-12 students in the U.S. using AI tools for educational purposes. Related studies report:

- Whether students have used AI tools (according to Cengage Group, 75% of K-12 students globally have)
- How frequently they use them (according to a May 2024 survey conducted by Impact Research commissioned by Walton Family Foundation, 49% of K-12 students self-reported using ChatGPT for school-related work at least once a week; this represented a 27 p.p. increase from February 2023)[209]
- What they use them for (e.g., searching for information, proofreading drafts)

We estimate that the average total weekly instructional hours across the United States is about **28 hours**. Our sources for this estimate can be found in the dropdown list below.

Baseline Source and Calculation (dropdown):

- Source: Pew Research analysis found U.S. instructional-time minimums are "943 hours for 1st-graders, 1,016 hours for 7th-graders and 1,025 hours for 11th-graders"[210]
  Calculation:
  - Average across grade levels: (943 + 1,016 + 1,025) ÷ 3 = 995 hours per year
  - Typical U.S. school year: ~36-38 weeks of instruction
  - Hours per week: 995 ÷ 36 weeks = 27.6 hours per week

Resolution Body: FRI staff; fallback to LEAP panel nowcast.

Resolution Criteria: Dynamic Resolution

- This question will be resolved by the latest official figures reported by U.S. school systems or Department of Education, if available.
- We will take an unweighted average of figures reported for U.S. states if data is split by state. If multiple sources exist, prioritize national-level reports; otherwise, use representative state-level data aggregated by reweighting by population size.

---

[208] *Link provided to participants:* https://www.cengagegroup.com/news/perspectives/2024/2024-in-review-ai--education
[209] *Link provided to participants:* https://static.waltonfamilyfoundation.org/bf/24/cd3646584af89e7c668c7705a006/deck-impact-analysis-national-schools-tech-tracker-may-2024-1.pdf?utm_source=chatgpt.com
[210] *Link provided to participants:* https://www.pewresearch.org/short-reads/2014/09/02/school-days-how-the-u-s-compares-with-other-countries/

- If insufficient data is available, resolve using LEAP panel dynamic nowcast as of the resolution date.

4. Open vs Proprietary Polarity

What will be the mean benchmark performance of the best closed-weight AI models and the top open-weight AI models on the following set of benchmarks by the following resolution dates?

- **Dec 31, 2025**
- **Dec 31, 2027**
- **Dec 31, 2030**

Background: The set of benchmarks is**:**

- **Mathematical reasoning:** FrontierMath (Tier 1-3 and Tier 4 separately)[211]
- **Software engineering:** SWE-Bench Verified.[212]
- **General reasoning and abstraction:** ARC-AGI-2.[213]

Open-weight models (publicly released model weights) and closed-weight models (proprietary access only) are competing for top performance. This question measures the mean performance of the best open-weight models and the best closed-weight models across a set of critical benchmarks. Note that this could result in different models of interest being chosen for each benchmark. Then, we'll take the mean of the percentage point gaps across the three benchmarks. Note also that "bespoke" models custom-made for the ARC Prize, as defined by the ARC-AGI-2 leaderboard, are not being considered for this question.

Historical baseline. As of September 2025, we find that the current mean performances of the top open-weight LLM and the top closed-weight LLM are **17.2%** and **30.0%** respectively**.** Note that some leaderboards may have changed since this baseline was last updated.

Benchmark	Best open-weight model	Best closed-weight model
FrontierMath (tier 1-3)	2.1% (Kimi K2)	24.8% (GPT-5 medium, high)
FrontierMath (tier 4)	0.0% (Kimi K2)	8.3% (GPT-5, high)
SWE-Bench Verified	65.4% (OpenHands + Kimi K2)	70.8% (Moatless Tools + Claude 4 Sonnet)

[211] *Link provided to participants:* https://epoch.ai/data/ai-benchmarking-dashboard
[212] *Link provided to participants:* https://www.swebench.com/
[213] *Link provided to participants:* https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025

ARC-AGI-2	1.3% (DeepSeek-R1)	16% (Grok- 4 (Thinking))

Resolution Body: Official leaderboards of included benchmarks; FRI staff

Resolution Criteria: Ground Truth

- For each benchmark, use the official leaderboard to identify the most capable (i.e., highest overall score) open-weight and closed-weight model, then calculate the mean performance across the open-weight and closed-weight models separately.
    - Open-weight models will be identified by FRI staff according to the Open Weight Definition by Open Source Alliance.[214]
    - If official leaderboard scores are not available, use scores reported by third-party evaluators such as Epoch AI.
- If multiple models are tied for the top open/closed weight scores due to overlapping confidence intervals, use the average score of the tied models.
- If credible reports on one or more selected benchmarks are unavailable, FRI will resolve the question by producing independent estimates with input from at least 3 experts not in the LEAP panel selected for expertise in model evaluations.

5. AI Companions

What proportion of U.S. adults will self-report using AI for companionship at least once daily by the following resolution dates?

- Jan 1, 2028
- Jan 1, 2031
- Jan 1, 2041

Background: Several prominent news outlets have recently noted a sharp increase in the use of LLMs to simulate companionship-style interactions.[215,216,217]

While the proportion of LLM conversations devoted to these types of interactions remains relatively small (Anthropic reported in June 2025 that just under 3% of conversations with Claude are affective, i.e., "motivated by emotional or psychological needs"),[218] services specifically aimed at providing AI companionship like Replika and Character.ai, the latter of which Wired reported as having 20 million monthly users,[219] have been growing in popularity.

[214] *Link provided to participants:* https://openweight.org/
[215] *Link provided to participants:* https://www.nbcnews.com/tech/ai-companions-friendship-rcna194735
[216] *Link provided to participants:* https://www.nytimes.com/2024/05/09/technology/meet-my-ai-friends.html
[217] *Link provided to participants:* https://www.newyorker.com/magazine/2025/09/15/playing-the-field-with-my-ai-boyfriends
[218] *Link provided to participants:* https://www.anthropic.com/news/how-people-use-claude-for-support-advice-and-companionship
[219] *Link provided to participants:* https://www.wired.com/story/character-ai-ceo-chatbots-entertainment/

Historical Baseline: A July 2025 poll from the AP-NORC Center for Public Affairs Research found that **6%** of U.S. adults reported using AI for companionship at least once a day; 16% reported having ever used AI for companionship.[220] 25% of young adults (18-29) also reported that they had used AI at least once for companionship. Note that the AP-NORC survey does not specifically define 'companionship', instead allowing respondents to determine for themselves whether they consider their particular use of AI to fit into this category.

Resolution Body: Representative polls of U.S. adults; FRI staff

Resolution Criteria: Ground Truth

- This question will be resolved by nationally representative public opinion polls from credible organizations (e.g., AP-NORC, Pew, Gallup, etc.) which ask respondents to self-report whether they use AI for companionship, emotional support, social interaction, or simulated relationships at least once per day.
    - Note that AP-NORC does not provide a definition for companionship, and instead allows respondents to determine for themselves whether their use capacity constitutes a companionship use case.
- Polls conducted up to 12 months before each date are acceptable to resolve the question. If multiple valid polls exist, we will default to using the most recent, reliable measure.
- If no valid poll data is available within the past 12 months at resolution, FRI will run a nationally representative survey of at least 1,000 U.S. adults, using appropriate quotas and weights (age, gender, race/ethnicity, region, and education), via a reputable online survey platform (e.g., CloudResearch, Prolific). The survey will replicate question wording from prior public polls as closely as possible, specifically asking about daily AI use for companionship.

6. Barriers to Adoption, II

By the end of 2030, what percent of LEAP expert panelists will say that each of the following factors has significantly slowed AI adoption relative to popular expectations around AI adoption progress in 2025…

- **Lack of reliability**
- **Cultural Resistance**:
- **Restrictive Regulations**
- **Cost Issues**
- **Data Quality Issues**
- **Integration Challenges**
- **Not Enough Use Cases**
- **Lack of AI Literacy**
- **Social-Cultural Anomie**

[220] *Link provided to participants:* https://apnorc.org/wp-content/uploads/2025/07/July-W1-Topline_AI.pdf

Background: Barriers to adoption are defined as factors that prevent or slow down AI systems being used more widely and more intensively. As of 2025, examples include unavailability of quality training data for use case, concerns about robustness to outliers, cost of compute, lack of AI-related skills, and regulation or legal restrictions, etc. This question includes the types of adoption barriers likely to remain prominent for future AI systems as elicited in the previous wave of this survey.

“Broader use in society” refers to AI applications beyond using AI for frontier AI development. Examples include augmenting or automating routine work, scientific research and discovery, personal entertainment, uses for social good, etc.

Given the overlapping and causally interacting nature of these factors, reasonable people may disagree with our sorting and labeling process. Nevertheless, we ask you to consider the barriers below, as they have been presented, as best you can.

In Wave 2 of LEAP, panelists proposed that current AI systems face the following barriers to adoption (in no particular order):

1. **Lack of reliability**: Hallucinations, unpredictable behavior, a lack of interpretability, results skewed by biased training data, and other reliability issues limit their usefulness.
2. **Cultural Resistance**: Fear of job losses and rising inequality, along with a common preference for humans over AI, will curtail use.
3. **Restrictive Regulations**: Ambiguous, fragmented, and evolving new regulations regarding AI, and the fact that the preponderance of existing regulations were written before the emergence of AI, will slow deployment.
4. **Cost Issues**: High infrastructure and energy costs will render some large AI projects impractical.
5. **Data Quality Issues**: High-quality, unbiased data that is useful to specialized industries is limited by privacy constraints, copyright issues, and the high cost of data labeling.
6. **Integration Challenges**: At a corporate level, AI adoption often requires a complex redesign of workflows, IT systems, and the physical environment—and many of the people tasked with executing the redesign may lack motivation due to fear of replacement.
7. **Not Enough Use Cases**: Until trust issues are resolved, and progress in robotics advances, the number of ways in which AI can create value in the real world will be limited and adoption will be delayed.
8. **Lack of AI Literacy**: There aren’t enough people with the necessary skills to develop, implement, and adopt new AI systems quickly, at scale, in diverse organizations around the world.
9. **Social-Cultural Anomie**: Societal intellectual atrophy and damage to the environment due to use of AI, paired with the mental health consequences of AI replacing much of human companionship, will lead to a backlash that slows adoption.

Resolution Body: FRI; LEAP Panel

Resolution criteria: Ground Truth

- At the resolution date (Dec 31, 2030), the LEAP expert panel will be asked to **respond to**: *In 2025, LEAP panelists highlighted possible obstacles to AI adoption. Indicate whether each of the following items turned out to be a significant barrier to adoption over the past five years, relative to popular expectations around AI adoption progress.*
- Panelists are instructed to choose only one option between ‘Agree’ or ‘Disagree’, choosing the option closest to their beliefs if they are uncertain.

Appendix F. Question-by-Question Results

Appendix F.I. Question-by-Question Results: Wave 1

FrontierMath

Question. *What will be the highest percentage accuracy achieved by an AI model on FrontierMath (Tiers-1-3), by 2025, 2027 and 2030?*

Background. FrontierMath[221] is a benchmark of hundreds of original, expert-crafted mathematics problems, which typically require hours or days for expert mathematicians to solve. The problems are unpublished and designed to be "guessproof," with less than a 1% chance of guessing correctly without the mathematical work.

- The problems in Epoch's evaluation so far fall into three tiers of difficulty:
    - Tier 1: Advanced, near top-tier undergrad/olympiad, 25% of problems.
    - Tier 2: Needs serious grad-level background, 50% of problems.
    - Tier 3: Needs an expert mathematician with relevant specialization, 25% of problems.
    - Epoch AI has since added Tier 4 problems to the FrontierMath benchmark (Epoch AI 2025a) which require weeks or months of expert effort. This question asks only about Tiers 1-3.

Historical baseline. As of April 16, 2025, the best AI performance according to Epoch AI, the benchmark developer, was achieved by o4-mini, solving **19%** of problems.[222] Since we elicited forecasts, in Aug 2025 Gemini 2.5 surpassed o4-mini, scoring **29**%.[223]

For full question background and resolution details, see Appendix E.I. 1. FrontierMath.

Figures

[221] *Link provided to participants:* https://epochai.org/frontiermath/the-benchmark
[222] *Link provided to participants:* https://epoch.ai/data/ai-benchmarking-dashboard
[223] See https://epoch.ai/frontiermath

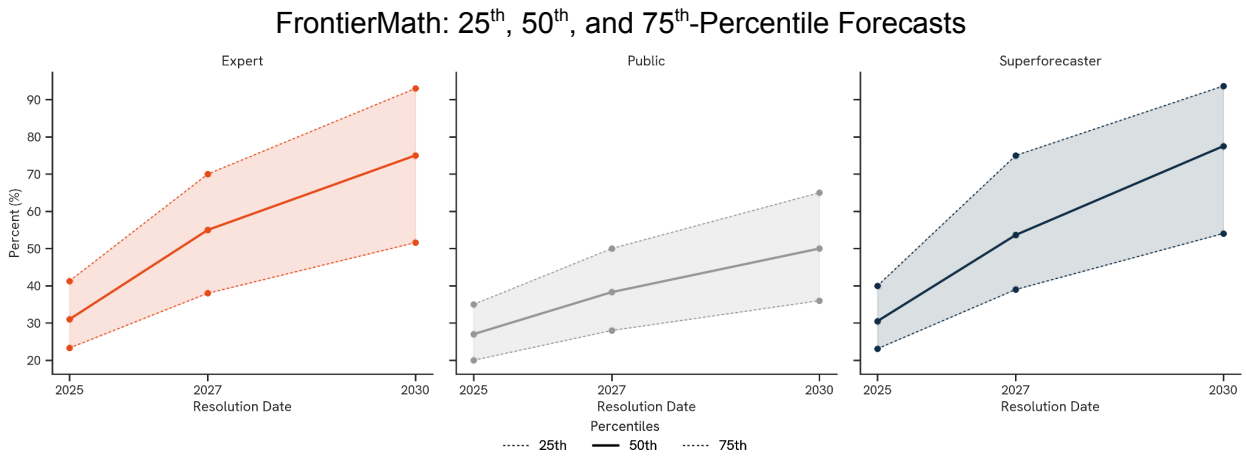


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.

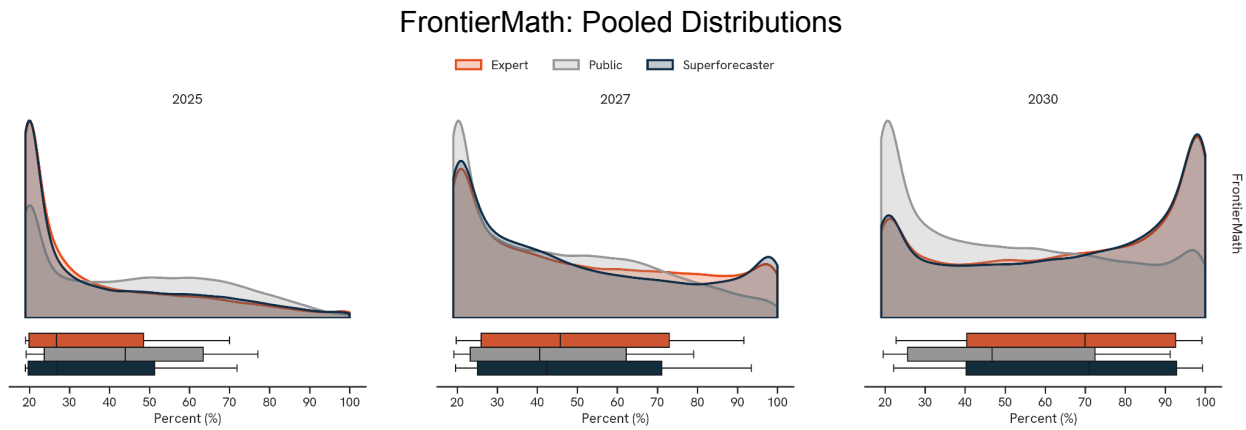


Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. We estimate each forecaster's full probability distribution from their 25th, 50th, and 75th percentile forecasts, by fitting the cumulative density function of an appropriate distribution (i.e., beta or gamma distribution) to the observed forecasts using nonlinear least squares. We then sample from these distributions and plot the aggregated distribution.

Results. Experts predict[224] state-of-the-art (SOTA) accuracy on FrontierMath of 31% by the end of 2025,[225] 55% by the end of 2027,[226] and 75% by the end of 2030.[227] 23% of experts predict that FrontierMath will be saturated by 2030 (>90%), meaning that AI can autonomously solve almost any math problem that resembles a problem a math PhD student might spend multiple

[224] Unless otherwise stated, when stating what a group "predicts," we are stating what the median member of that group predicts.
[225] *Raw data*: IQR on the 50th percentile was (27.3%–37.5%); median 25th and 75th percentile forecasts were 23.3% and 41.3% respectively.
[226] *Raw data*: IQR on the 50th percentile was (41.3%–70.0%); median 25th and 75th percentile forecasts were 38.0% and 70.0% respectively.
[227] *Raw data*: IQR on the 50th percentile was (60.0%–90.0%); median 25th and 75th percentile forecasts were 51.6% and 93.0% respectively.

days completing. While experts and superforecasters predict similarly, the public expects much less progress, predicting 27% by 2025,[228] 38% by 2027,[229] and 50% by 2030[230]—**25**p.p. less than experts. Since eliciting forecasts, in Aug 2025 Gemini 2.5 surpassed o4-mini,[231] scoring **29**+/-3%, 10p.p. higher than the previous SOTA of 19%.

For full results tables, see here.

Rationale analysis:

- **Recent progress:** High-forecast respondents often cite impressive recent trends: "Evaluated scores across benchmark types have tended to jump over 3 year periods. For example: MATH from 7% in 2021 to 72% in 2023 and 90% in 2024." Another observes, "The historical data show a very rapid increase in the accuracy [on FrontierMath] over a time period just short of a year. In that time the accuracy increased from 0% to 19%."[232]
- **Trajectory**: High-forecast respondents frequently emphasize likely continued progress: "We've seen jumps of around 5 points on this benchmark every couple of months[233] so far and these jumps will only accelerate as scores approach 50% (benchmark scores tend to be roughly sigmoid-shaped over time)." Among low-forecast respondents, a common sentiment is that "the fastest…progress is behind us and we are now approaching the flat/end point portion of the S-Curve of advancement."
- **Test-time compute**: High-forecast respondents often highlight potential gains from scaling inference compute: "The current top scorer is a reasoning model, and the reasoning model paradigm is relatively new; this suggests that rapid improvements are likely as the paradigm evolves." Another noted, "with very large amounts of inference compute, it's possible that o3 or o4-mini could already get well over 30% today." Many low-forecast respondents, however, are skeptical that inference scaling will be sufficient to overcome fundamental architectural limitations.
- **Data**: High-forecast respondents tend to believe that labs will have the training data they need to drive improvements. One notes that an "increased use of synthetic data…will likely lead to broader problem-solving capabilities," and another that, "by 2031, the frontier labs will have enormous amounts of math data if they believe it's valuable and the RL process will solve the problem." Low-forecast respondents are more skeptical that training data will be suitable for the task: "One of the key bottlenecks is the scarcity of high-quality, human-generated mathematical data."

[228] *Raw data*: IQR on the 50th percentile was (22.0%–35.0%); median 25th and 75th percentile forecasts were 20.0% and 35.0% respectively.
[229] *Raw data*: IQR on the 50th percentile was (28.0%–50.0%); median 25th and 75th percentile forecasts were 28.0% and 50.0% respectively.
[230] *Raw data*: IQR on the 50th percentile was (35.0%–70.0%); median 25th and 75th percentile forecasts were 36.0% and 65.0% respectively.
[231] https://epoch.ai/frontiermath
[232] This was true at the time the expert completed the survey.
[233] Actual progress was marginally slower. The top Tier 1-3 accuracy rate rose from 1.03% in June of 2024 to 29% in August of 2025, where it remained as of the publication of this paper.

- **Architecture suitability**: Many high-forecast respondents highlight math's suitability for current architectures. One states, “Existing LLMs are well-suited to make progress in math because math is about logical pattern matching, can be well-represented by text, and the answers are typically verifiable allowing for easy rewards.” Low-forecast respondents express considerable doubt that the current architecture will be sufficient. One mathematician states: “I am deeply skeptical of LLMs ever being able to excel at complex math. LLMs are linguistic pattern-replication machines, and advanced math requires a high degree of conceptual creativity and flexibility that does not derive from linguistic patterns.”
- **Incentives**: Some high-forecast respondents expect sustained investment due to prestige and R&D value. One notes, “Math is highly relevant to many R&D domains, so progress in math has been, and is highly likely to continue to be, a focus for leading AI companies.” But many low-forecast respondents question that assumption: “Performance from 2028 to 2031 entirely depends on whether or not the frontier AI labs believe it's economically useful to continue improving performance on this benchmark,” writes one, with another bluntly concluding, “I don’t think current companies are focused on solving math.”

High-forecast rationale examples:

> “We’ve seen jumps of around 5 points on this benchmark every couple of months so far and these jumps will only accelerate as scores approach 50% (benchmark scores tend to be roughly sigmoid-shaped over time). Reasoning models are quickly improving their math abilities, and math is a priority for labs like OpenAI and Google DeepMind. One important factor here is that these will be Epoch-run evaluations, with a baseline of 19% today, and I assume future evaluations will be run with similar approaches. With very large amounts of inference compute, it’s possible that o3 or o4-mini could already get well over 30% today. Most benchmarks are saturated within a few years. For example, GPQA was a very difficult benchmark when it was released in late 2023, but has now been saturated just under two years later…”

> “Evaluated scores across benchmark types have tended to jump over 3 year periods. For example: MATH from 7% in 2021 to 72% in 2023 and 90% in 2024. Similarly, GPQA jumped from mid 30s for GPT3, Grok 1 and Gemini to above 80% currently (across models). If the FrontierMath proves similarly susceptible to both scale and training focus a jump from 19% to the 70-80% plus range seems plausible by 2028.”

Low-forecast rationale examples:

> “FrontierMath … breaks slightly with previous benchmark creation ideals in that it is intended as a test that saturates human limits; harder math benchmarks would be difficult for humans to generate. Many older benchmarks are instead designed to be just within reach for current AI.”

> “One of the key bottlenecks is the scarcity of high-quality, human-generated mathematical data. These problems are original and solving them typically requires deep intuition and multi-step reasoning that’s hard to find in any existing dataset. Most available math training data is either synthetic, shallow, or not aligned with the kind of work required for those problems. Until models can train on richer and more realistic human proofs or at least be fine-tuned on reasoning traces that resemble the way mathematicians think, performance will improve slowly.”

Autonomous Vehicle Trips

Question. *What percentage of U.S. ride-hailing trips will be provided by autonomous vehicles that are classified SAE Level 4 or above in the years 2027 and 2030?*

Background. SAE Level 4 “High Automation” vehicles do not require any human intervention when automated driving features are engaged, but may only operate in limited conditions or environments.

Historical baseline. There is currently no official reporting, but we estimate a share of **0.27%** as of Q4 2024.

For full question background and resolution details, see Appendix E.I. 2. Autonomous Vehicle Trips.

Figures

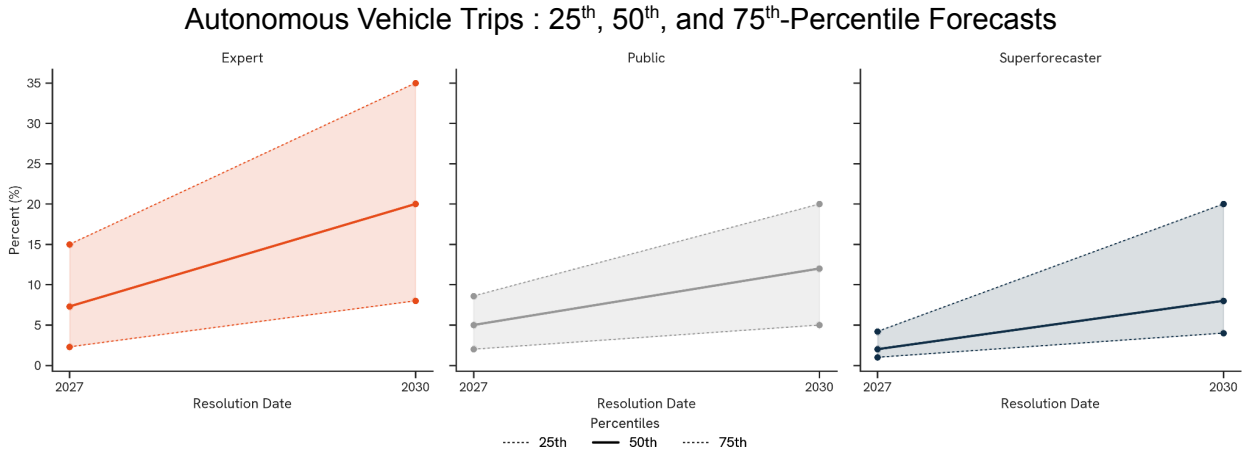


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.

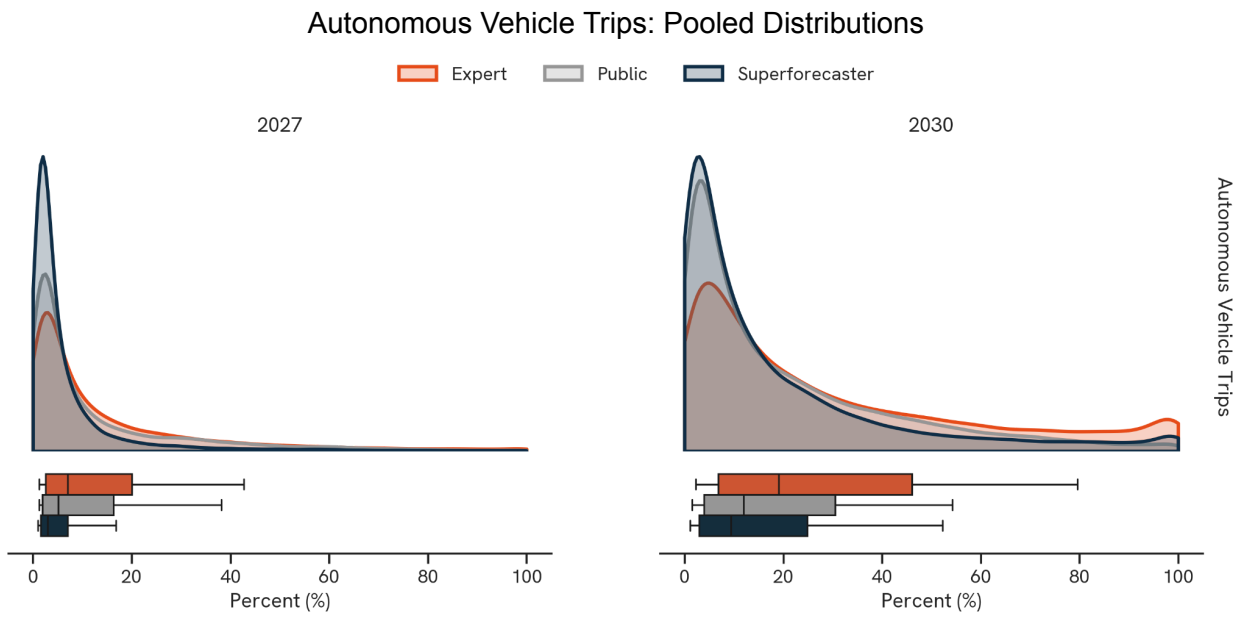


Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. For more details on how we calculate pooled distributions, see Appendix C. Pooled Distribution Estimation.

Results. Experts predict that 7% of U.S ride-hailing trips in 2027 will be provided by autonomous vehicles that are classified as SAE Level 4 or above,[234] and **20% in 2030**,[235] up from our estimate of 0.27% in Q4 2024. Experts have a long tail of prediction, with the median expert assigning a 25% chance to more than 35% AV use in 2030. Superforecasters and the public forecast a much lower autonomous vehicle share, predicting 2%[236] and 5%[237] respectively in 2027, and 8%[238] and 12%[239] respectively in 2030.

For full results tables, see here.

Rationale analysis:

- **Technology readiness:** High-forecast respondents commonly emphasize that Level 4 technology "has a proven, real-world track record and is already being commercially

[234] *Raw data*: IQR on the 50th percentile was (3.0%–15.0%); median 25th and 75th percentile forecasts were 2.3% and 15.0% respectively.
[235] *Raw data*: IQR on the 50th percentile was (10.0%–40.0%); median 25th and 75th percentile forecasts were 8.0% and 35.0% respectively.
[236] *Raw data*: IQR on the 50th percentile was (1.0%–5.0%); median 25th and 75th percentile forecasts were 1.0% and 4.2% respectively.
[237] *Raw data*: IQR on the 50th percentile was (1.6%–13.3%); median 25th and 75th percentile forecasts were 2.0% and 8.6% respectively.
[238] *Raw data*: IQR on the 50th percentile was (3.0%–25.0%); median 25th and 75th percentile forecasts were 4.0% and 20.0% respectively.
[239] *Raw data*: IQR on the 50th percentile was (5.0%–29.0%); median 25th and 75th percentile forecasts were 5.0% and 20.0% respectively.

deployed," suggesting that the fundamental technical challenges have mostly been solved. In particular, many high-forecast respondents point to Waymo's demonstrated safety record, noting, "Studies have demonstrated that its vehicles had far fewer airbag deployment crashes and injury-causing crashes as compared to human drivers." Conversely, a large number of low-forecast respondents emphasize that autonomous driving "is a tech that is absolutely notorious historically for overpromising and snail-like progress" and argue the current technology "is not scalable because it requires lots of case-by-case optimization for a particular region (down to individual intersections)."

- **Data flywheel:** High-forecast respondents frequently highlight Waymo's exponential expansion, observing, "Waymo is currently more-than-doubling every year," which they believe will result in a data flywheel: "Broader deployment will generate more data, which in turn enhances safety—creating a positive feedback loop." Another states: "Historically, when a technology finally gets to be used in the wild, it improves very rapidly." Low-forecast respondents often acknowledge progress but stress constraints, arguing that "progress in Phoenix or Miami does not generalize easily to New York, Boston, or Chicago."
- **Economic viability:** Many high-forecast respondents argue that cost advantages will drive adoption, stating that "once AV ride-hailing proves substantially safer and cheaper, adoption could accelerate sharply." Low-forecast respondents tend to focus on higher near-term costs, with one noting that "Waymo cars are still 30-40% more expensive compared to Uber and Lyft," and that "AVs still require high capital expenditure, vehicle downtime, and supervision costs."
- **Tesla:** Some high-forecast respondents see Tesla as a game-changer, arguing that "Tesla alone could drive explosive growth if their FSD [Full Self-Driving] vision model is successful in select markets, as their capex is just the mass-produced vehicle cost," and that Tesla's manufacturing advantage over Waymo could enable rapid scaling. Low-forecast respondents largely remain skeptical, with one predicting, "it is well under 25% likely that Tesla will achieve its dream of all of their 2m cars/year being capable of serving as robotaxis straight out of the factory."
- **Regulatory environment:** Some high-forecast respondents believe regulatory barriers will diminish, with a few noting that the Trump administration's anti-regulatory stance may help to speed things up. Low-forecast respondents emphasize fragmented regulation, arguing "the U.S. will continue to have a patchwork of rules on autonomous vehicles," and that "investors are less likely to invest in areas that present less regulatory certainty."
- **Geographic and weather constraints:** High-forecast respondents often downplay geographic limitations, with one noting that even cities like Boston have "more simple grid-like" streets than some might expect. By contrast, low-forecast respondents tend emphasize both weather and geography constraints, noting initial deployment has been "in cities with no or little ice and snow," and arguing this creates "a chicken and egg problem where they need to train in [icy and snowy] conditions to get better, but [doing so is] really risky and unpredictable."
- **Social and labor resistance:** Some high-forecast respondents believe "consumer acceptance is likely to rise with exposure," while some low-forecast respondents warn

that “the cities where taxis are most used are also likely to have strong unions and/or lobbying efforts to limit the damage of job losses.” They note growing concern about “the replacement of blue collar labor with more automation,” and predict that “ride share drivers aren't going to go down without a fight.”

High-forecast rationale examples:

> “Waymo is currently more-than-doubling every year, e.g., from March 2024 to March 2025 it grew by about 8.5x.”
>
> “Waymo currently has a fleet of 1.5k cars, with expansion plans to 3.5k by 2026. Extrapolating this growth rate of 2.3x per year, they should have 8.2k cars by EOY 2027 and 104k cars by EOY 2030. This would correspond to 196k and 2.5 million rides respectively. Assuming the 13.6 daily rides don’t grow significantly (for simplicity), Waymo would constitute 1.4% of all rides in Jan 2028 and 18% of all rides in 2031. I’m updating on this in the following ways: I expect that Tesla could deploy five times as many cars…given they already have factory capacity. I also expect there’s a chance most normal cars are equipped with self-driving features by 2030.”

Low-forecast rationale examples:

> “At every point in the past this has seemed much closer than in actuality.”
>
> “The current Waymo technology is not scalable because it requires lots of case-by-case optimization for a particular region (down to individual intersections). It is not a general-purpose technology that can be deployed to new cities with marginal extra costs. Thus, I expect the scaling to be more like linear instead of exponential.”
>
> “Not enough ride-hailers in big cities with no ice/snow to support massive near-term growth: the next challenge would be tackling different weather conditions. I see this as a big problem (kind of like a chicken and egg problem where they need to train in such conditions to get better, but the behavior stays really risky and unpredictable for a long time which makes it harder to scale fast).”
>
> “I also expect manufacturing capacity to be a bottleneck, as well as infrastructure/capacity in rural areas. Even if prices drop significantly, and everyone who normally uses public transport relies on ride-hailing services, our cities are not made for this.”

Occupational Employment Index

Question. *What will the percent change in the number of jobs (compared to Jan 1, 2025) in the U.S. be for white-collar, blue-collar, and service-sector occupations, by the end of 2027 and 2030?*

Historical baseline. We estimate that annual percent changes in each group of occupations in May 2024 as compared to May 2023 are:

- White-collar: **+1.61%**
- Blue-collar: +**2.67%**
- Service-sector: **+0.44%**

For full question background and resolution details, see Appendix E.II. 3. Occupational Employment.

Figures

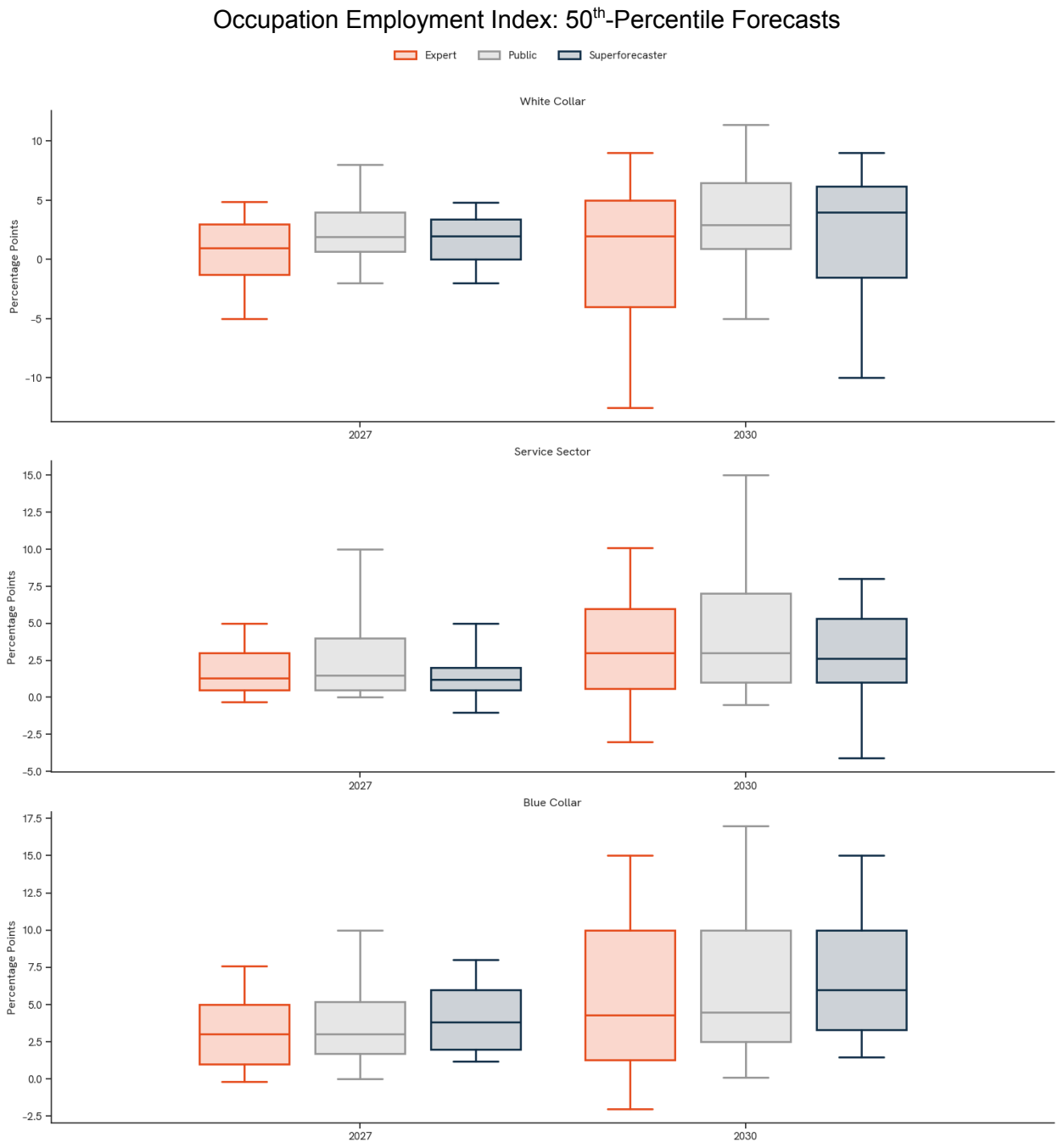


Figure: Median forecast for the percent change in employment (compared to Jan 1, 2025) in the U.S. for service-sector, white-collar, and blue-collar occupations by participant category.

Results. Experts, the public, and superforecasters generally do not expect significant job losses for blue-collar or service-sector workers by 2027 or 2030. However, the median expert predicts 2% growth in white-collar jobs between January 2025 and December 2030. This is significantly slower than the pre-existing trend, which predicts 6.8% growth. 25% of experts predict more than 5% white-collar job *gain* (close to baseline), meaning over 75% of experts predict slower

growth than recent trends. Moreover, 25% of experts predict more than 4% white-collar job *loss* by 2030.

For full results tables, see here.

Rationale analysis:

- **White-collar jobs**: Most high-forecast respondents (those predicting more growth) believe white-collar employment will adapt rather than collapse: “Revolutionary technologies rarely deliver the job losses that people fear,” writes one, and another, “AI adoption can engender task-level displacement, but the broader digital ecosystem…generates complementary roles that absorb displaced workers and create net employment gains.” Low-forecast respondents (those predicting more losses) frequently express the belief that AI's speed and cognitive focus make this disruption different: “I am most pessimistic about white-collar jobs due to two developments - the tech sector becoming an increasingly bigger proportion of the white collar sector and AI systems making sufficient progress to replace a small but significant proportion of jobs in the tech sector.”
- **Blue-collar jobs**: Forecasters broadly agree blue-collar jobs will be the most resilient to automation. High-forecast respondents particularly emphasize physical complexity barriers: “Blue collar jobs involving complex physical manipulation will likely be among the last job categories to be fully automated by AI. Humans excel at intricate physical tasks requiring dexterity, problem-solving in unpredictable environments, and adaptability.” Even among forecasters who foresee long-term losses, many acknowledge that “blue-collar occupations are likely resilient to AI trends short of advanced post-AGI robotics.” Infrastructure investment is commonly thought to be another factor bolstering blue-collar work: “There will be significant investment as part of the AI revolution (data centers, research, new types of companies being spun out), which will create new other jobs across the spectrum.”
- **Service-sector jobs**: Views diverge on service sector outcomes. Some emphasize that “people have a stronger intrinsic preference for many of these roles to be done by humans,” and that there are demographic tailwinds: “Pink collar will increase because of longevity and old-age care.” Others focus on automation vulnerability, particularly in customer service.
- **Demand elasticity**: Experts are split on whether AI productivity gains will increase or decrease employment. High-forecast respondents invoke economic theory: “A correct prediction depends upon predicting the effects of the Jevons Paradox (where price declines lead to increased purchases) as well as wealth effects (where efficiency boosts overall wealth increasing consumption).” Several low-forecast respondents argue elasticity won't offset displacement: “Even for software, elasticity is probably not high enough to increase the number of SWEs if the number needed to produce software falls by 1000x.”
- **Recent trends**: Low-forecast respondents often emphasize current evidence: “From Intel to Microsoft, many top executives and management staff were laid off to make room for other investments at the organization. Google laid off 10% of its managerial staff last

December." Another notes: "In October 2024, S&P Global claimed that one in every four American workers that lost their jobs last year had worked in professional and business services."[240]

High-forecast rationale examples:

"Rather than decreasing it's highly likely that this area of labor will simply adapt. Historically, human labor patterns have experienced quite radical transformations over time, even in established sectors. Emerging technologies, rather than sucking people out of the labor market of white-collar work, are more likely to make them work differently and lead to new white-collar roles that can capitalize on this transition."

"The easiest trap on forecasting is to overestimate the short-term impact and underestimate the long-term impact. While AI is going to revolutionize the world, and especially the workforce, it won't happen overnight. Forecasting in 1995 that by 2000 e-commerce would represent more than 5% of total sales would [have been] a big overestimate. But by 2010 it would be low balling. It is similar for impact on jobs."

Low-forecast rationale examples:

"Reports of white-collar jobs shrinking haven't just come to light this month but have been a topic of conversation for months now. In October 2024, S&P Global claimed that one in every four American workers that lost their jobs last year[241] had worked in professional and business services and that in September, 497,000 professional and business services jobs were eliminated. From Intel to Microsoft, many top executives and management staff were laid off to make room for other investments at the organization. Google laid off 10% of its managerial staff last December. More recently, companies like HP and Block are also looking at large-scale layoffs in their organization."

"Software engineering is already being hugely impacted and this is only accelerating…[although] it will take much longer for more experienced roles to be eliminated. The only way that we may see an increase in white collar jobs is through new roles and industries enabled by AI. We have seen this before through the industrial revolution and the invention of computers for example, but these occurred over decades, giving people time to learn new skills. AI is coming much much faster."

General AI Progress

Question. *At the end of 2030, what percent of LEAP panelists will choose "slow progress," "moderate progress," or "rapid progress" as best matching the general level of AI progress?*

Scenario summaries:

[240] The October 2024 S&P Global report referred to job losses over the course of the first nine months of 2024 (S&P Global 2024).
[241] As per the footnote above, the October 2024 S&P Global report referred to job losses over the course of the first nine months of 2024.

- **Slow progress**: AI is a capable assistant that augments human work but doesn't replace it. AI can produce PhD-level literature reviews (but rarely solutions to difficult problems), handle about half of 8-hour freelance coding jobs, resolve most customer service complaints with human teams, and generate respectable creative content (3-minute songs at major label quality, short stories that need editing). AI can manage simple, well-defined tasks that take a human under an hour—like drafting emails or planning vacations—on par with a competent human assistant. Self-driving cars haven't achieved full autonomy, and household robots can perform basic tasks like making coffee or loading dishwashers but require consistent environments and occasional human guidance. Scientific breakthroughs remain almost entirely human-driven.
- **Moderate progress**: AI is an effective collaborator. Autonomous lab systems with human assistants drive rapid (though incremental) advances in fields like solar cells and fusion. AI can handle nearly all software freelance projects taking 5 days or less, can manage complex multi-day business operations, and can fully replace customer service teams. AI can draft 100,000-word novels ready for mainstream publishing (with standard human editing) and can create 5-minute songs with breakout potential. Level-5 autonomous vehicles finally exist, and robots can navigate any U.S. home independently, performing basic tasks like making coffee or loading dishwashers as quickly and reliably as humans.
- **Rapid progress**: AI systems match or surpass the best human capabilities across most domains. Autonomous researchers can collapse years of R&D into weeks or months, creating revolutionary technologies like materials that revolutionize energy storage or bespoke cancer cures. No human freelance software engineer can outperform AI. AI can create Grammy-level albums, Pulitzer-level novels, and can run companies at the level of highly competent CEOs. Level-5 robo-taxis are 99.9% safer than human drivers and can go off-road anywhere a competent human driver can. Robots can navigate an arbitrary home, performing many household tasks faster and more reliably than most humans and without guidance. Robots in advanced factories can autonomously perform the full range of tasks requiring the highest levels of dexterity, coordination, and adaptive decision-making.

For full question background and resolution details, see Appendix E.I. 4. General AI Progress.

Figures

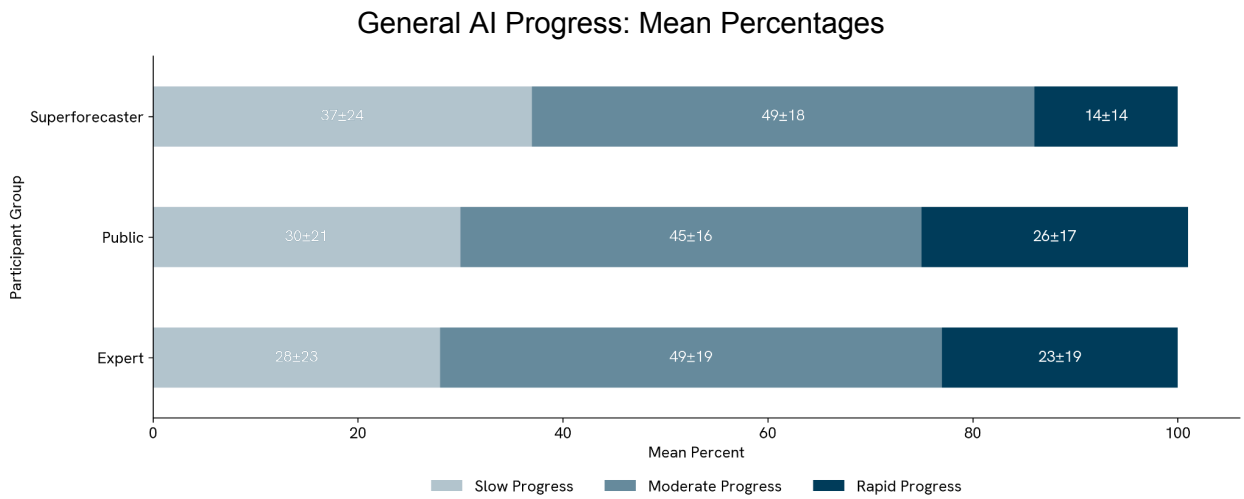


Figure: Participants estimated what proportion of LEAP panelists will choose "slow progress," "moderate progress" or "rapid progress" as best matching the general level of AI progress in 2030 (and the standard deviation of responses). This figure shows the mean percent on each scenario (split by color) and by participant group (y axis).

Results. Experts, superforecasters and the public, all modally expect[242] a "moderate" progress scenario. Most of the difference between these groups is in the weight they put on a "rapid" progress scenario, with superforecasters estimating 14%,[243] experts 23%,[244] and the public 26%.[245] This is one of the few cases where the public expects faster progress than experts.

For full results tables, see here.

Rationale analysis:

- **Scaling:** Rapid-progress forecasters typically cite consistent capability improvements: "historically AI system development has followed a steep scaling curve and increases in model size-data and compute have led to rapid capability gains." One points out that "METR [Model Evaluation and Threat Research] results imply a roughly 4 to 10x improvement in time horizon every year, which means that we'll have systems capable of doing weeks or months of work by the late 2020s."[246] Many rapid-progress forecasters also argue that AI progress has been consistently underestimated, and that "rapid progress is quite likely due to the [...] possibility of AI improving themselves."
  Slow-progress forecasters often argue that current AI paradigms are likely insufficient: "For any of the moderate and rapid progress criteria to be met there would need to be a massive paradigm shift in AI technology. LLMs are unlikely to achieve these goals with the incremental progress made over the last 2-5 years."

[242] Expect LEAP panelists to choose as best-matching the general level of AI progress.
[243] *Raw data*: IQR on the 50th percentile was (5.0%–19.0%). 90th percentile of median forecast: 35.0
[244] *Raw data*: IQR on the 50th percentile was (10.0%–30.0%). 90th percentile of median forecast: 50.0
[245] *Raw data*: IQR on the 50th percentile was (12.0%–35.0%). 90th percentile of median forecast: 50.0
[246] Actual rate of improvement likely falls within this range, especially given recent acceleration trends, but there is considerable uncertainty and domain variability. See Kwa et al. (2025) for more information.

- **Physical world capabilities:** Robotics is the strongest consensus limitation: “As a roboticist I have serious doubts about how quickly the embodied aspect of AI will make progress due to real-world implementation issues (sensors, communication delays, etc.) regardless of how good the brain is.” Many slow-progress forecasters also believe that historical precedent suggests caution is merited: “Having end to end task completion ability is hard as can be told from the development of autonomous vehicles (15+ years of R&D and 100+ billions of dollars invested and we are still scaling L4 robotaxis today).” Most moderate-progress forecasters also acknowledge this challenge.
- **Input constraints:** Slow-progress forecasters often highlight the need for better training data, the cost of compute, and especially energy: “I expect energy to be the chief bottleneck to AI progress such that it will be a rate-limiter for progress in general.” Some forecasters expecting rapid progress, however, argue that such constraints can be eliminated by massive investment in AI, driven by corporate competition and national security incentives.
- **Reliability:** A common consideration among slow-progress forecasters is that “[reliability] will be a difficult bar to clear for advanced-capability systems given current cognitive limitations.” They tend to emphasize that barriers like robust generalization, complex symbolic reasoning, and end-to-end safe autonomy for complex tasks will likely hold “superhuman across the board” scenarios back.

Slow progress rationale examples:

> “Radical change in major systems just takes longer than 4-5 years. I also think that [in] many of these domains, even unexpectedly fast advancement in AIs will not easily translate to improvements for quite some time because of unexpected barriers. That is, at least until we have strong artificial general intelligence (AGI), which we will not by 2030. To paraphrase an old saying, every job looks easy for those not actually doing it.”

> “ChatGPT was first publicly released in late 2022. I don’t believe what we witnessed over these past 2.5 years would justify expecting rapid progress over the next 5.”

> “We are in 2025 [at] peak AI hype. I do not expect this to last for another 5 years. The expectations will evolve.”

Moderate progress rationale examples:

> “If the METR task horizon trends hold, then by 2030 AIs would be able to do software tasks that take humans months or weeks. This kind of time horizon mostly seems like the moderate progress world.”

> “Moderate sounds completely achievable even today with proper integration of the latest technologies (commercialized, announced, released & unreleased) of the SOTA labs into enterprise and consumer workflows. I’m less sure of the robot part.”

Rapid progress rationale examples:

> “The fast progress scenario essentially describes AGI, and my median timeline for that is 2030. This is mostly an intuitive guess. But I think the last three years of progress have been qualitatively immense, so the next five years seem like they could lead to highly autonomous systems capable of very impressive things. And the METR results imply a roughly 4 to 10x improvement in time horizon every year, which means that we'll have systems capable of doing weeks or months of work by the late 2020s.”
>
> “I expect a roughly ~35% chance of AGI by EOY 2030. AGI is defined here as capable of automating all remote worker tasks (as of 2024 EOY) at a cost comparable to hiring a human of equivalent skill or better.”
>
> “I predict AI development will follow the rapid progress scenario in most domains, with robotics being a notable exception where progress will likely align more closely with the moderate scenario. However, public perception will remain divided regardless of actual capabilities, because AI systems will continue to be imperfect—demonstrating remarkable proficiency in some tasks while exhibiting surprising blind spots in closely related areas. These inconsistencies will provide evidence to support conflicting viewpoints about AIs true capabilities.”

Technological Richter Scale

Question. *At the end of 2040, what is the probability that AI will achieve the following levels of net impact on human society as compared to the impact of past technological events?*

Background. Nate Silver’s book “On the Edge” (2024) proposes the technological Richter scale (TRS) which, analogous with earthquake magnitudes, rates the impact of technologies on a roughly logarithmic scale.

- **Level 5**: A commercially successful invention important in its category, e.g., a leading brand of windshield wipers.
- **Level 6**: An invention that is disruptive in its field and has some ripple effects beyond it, e.g., VCR.
- **Level 7**: Technology of the decade, measurably impacts daily lives, e.g., credit cards, social media.
- **Level 8**: Technology of the century, broadly disruptive in society, e.g., electricity, automobiles.
- **Level 9**: Technology of the millennium, altering the course of human history, e.g., agriculture, the wheel, Industrial Revolution.
- **Level 10**: Epoch-defining event that alters the fate of the planet, e.g., the rise of humans.

For full question background and resolution details, see Appendix E.I. 5. Technological Richter Scale.

Figures

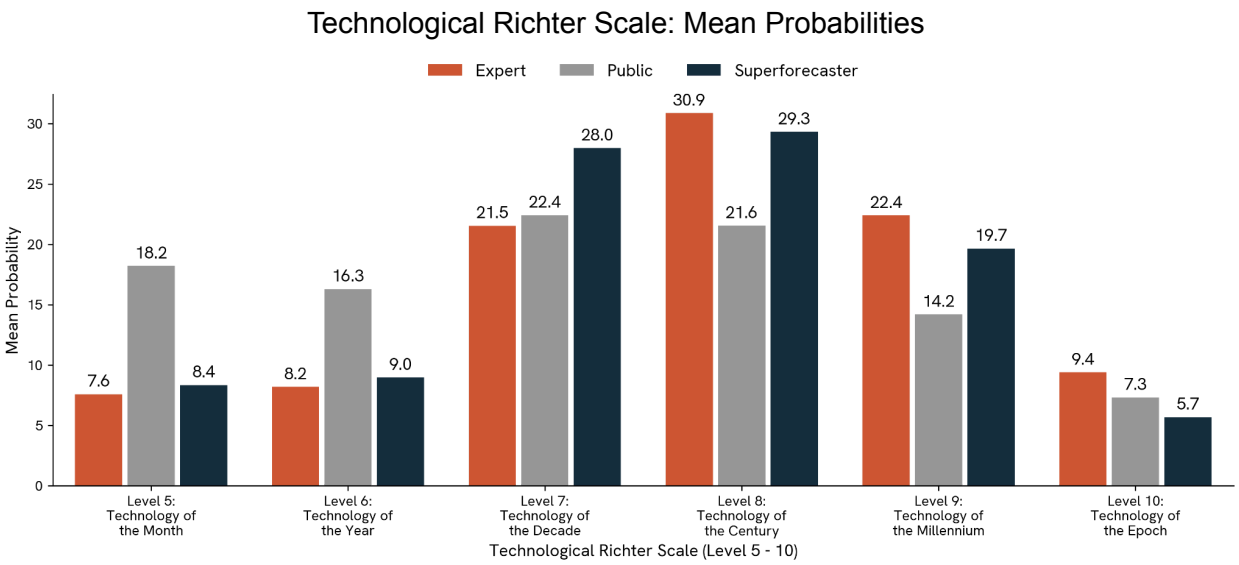


Figure: In this question, participants estimate the probability of AI achieving various levels of societal impact. This figure shows the mean probability assigned to each level.

Results. Experts modally expect[247] that the impact of AI by 2040 will be comparable to a "technology of the century" (akin to electricity or automobiles), while the public expects AI's impacts to be closer to the "technology of the decade" (more like social media). Experts give a 32% chance that AI will be at least as impactful as a "technology of the millennium"—like the printing press or the Industrial Revolution—whereas the public gives this a 22% chance. Superforecasters sit between experts and the public.

For full results tables, see here.

Rationale analysis:

- **Intelligence as a transformative force**: Many high TRS forecasters argue that, because AI will augment and eventually surpass human intelligence, its transformative potential is unique and eclipses all prior transformative technologies. "Intelligence is one of the most disruptive phenomena known to history," writes one. High TRS forecasts also tend to argue that progress in intelligence builds on itself: "People fundamentally don't think in exponentials. 2040 is a LONG time away, technologically. And AI will modify AI, at which point its improvement will go even more second-order."
- **Current societal impact**: Most high TRS forecasters believe AI has already surpassed levels 5 and 6. Many argue level 7 has already been met given its use in specialized fields (e.g., medical imaging, SWE, text editing, etc.) and its level of diffusion. Some even argue it can be credibly claimed that advanced capitalist societies have already reached level 8, given that "AI is strongly affecting geo-political considerations, changing energy demand and planning, disrupting education, [and] accelerating biomedical research." Many argue that even a slow, linear extrapolation of current trendlines

[247] Expect as best-matching the level of societal impact from AI.

suggests level 8 or higher by 2040, especially given that much of the existing potential of this new technology has yet to be adopted by society. Low TRS forecasters often view current AI as roughly level 5 or 6, essentially "an extension of the Internet."

- **Bottlenecks**: The main argument from low TRS forecasters is that progress will be substantially slowed by implementation constraints that are difficult to overcome with software improvements. One notes, "There likely exist thousands (millions?) of potential bottlenecks in the economy which will only become legible as other processes are sped up by orders of magnitude." Low TRS forecasters expect that many of these bottlenecks will be input bottlenecks, like energy, production capacity, and chip availability.
- **Diffusion speed**: High TRS forecasters frequently emphasize AI's rapid adoption rate compared to historical technologies, noting that it is "much faster than it was for prior technologies." Several low TRS forecasters, however, argue that historical precedent remains relevant: "The integrated circuit took about 15 years to change electronics. Computers took about 25 years. The Internet took 15 years to produce the WWW and another 10 or 15 years to change lives." Some low TRS forecasters also argue that, even if AI is proceeding on a faster diffusion timeline than other transformative technologies, the higher TRS levels require not only technical breakthroughs, but also changes to our political, economic, regulatory, and cultural structures.
- **Economic transformation**: Some high TRS forecasters see AI fundamentally restructuring society. One writes, "[Just as] the industrial revolution helped usher in capitalism, because it was an economic system that was compatible with that type of transformation, AI—being a technology that has the potential to replace human labor in most fields—might force societies to shift to a new economic model, which would place it at level 9." But most low TRS forecasters question whether AI will deliver enough tangible benefits to lead to such a transformation: "The average citizen will not have much benefit to buy from AI. Improved games or art? Cheaper manufactured goods? A robot to clean your house? How does AI deliver things that humans want, like better, cheaper healthcare?"

High-forecast rationale examples:

"The scale and scope of AI…[and] impact on human evolution is unique. While there are parallel examples in the rise of agriculture and industrial production—particularly in terms of general-purpose innovation (steam, fossil fuels, electricity, etc), AI is unique because it both augments human intelligence and will eventually surpass it. AI will ultimately displace capitalism as a form of civilization by eliminating scarcity while providing a basis for a shift into a new civilization rooted in consciousness evolution. AI is not a singular technology but rather is complemented by clean energy (at near -zero cost), bio-engineering (genetic design), quantum computing and distributed and/or networked culture. These technologies represent the end of the industrial era and a market-based society as human development shifts to culture and human evolution as an end in itself."

"It's already moved beyond the 5/6 level even today. Gemini does amazing videos, pictures and many things that are very comparable to human generated videos…Has it (AI) reached commercial success - of course…Has it disrupted? Of course."

> "If progress can continue at its present rate, Technology of the Millennium is a possibility. Given the level of investment and the scramble for talent, Technology of the Decade is assured. We are on a long runway stretching back 50 years and have finally achieved liftoff, more progress is required to assure a smooth flight and safe landing. AI could conceivably rival the printing press at giving everyman a level of intelligence where it once provided everyman with information. The industrial revolution greatly increased the material productivity of society, AI could provide the same boost for both material and service products by trading electrical energy for intellect."

Low-forecast rationale examples:

> "The force of its impact will likely be slowed by bottlenecks in areas AI hasn't yet conquered. An important concept is that an economic bottleneck grows in significance when productivity elsewhere increases. For instance, if global shipping dramatically increases, a bottleneck in the Suez or Panama Canal becomes much more costly. There likely exist thousands (millions?) of potential bottlenecks in the economy which will only become legible as other processes are sped up by orders of magnitude."

> "Our current AI can be seen as an extension of the Internet (which itself is an extension of the printing press, etc.). To me, current AI is like a super-Google search -- it's still all about the data that we give it (our cultural heritage), but now we are able to use/query/manipulate this in much more powerful ways than before."

Appendix F.II. Question-by-Question Results: Wave 2

Millennium Prize

Question. *Will AI solve or substantially assist in solving a Millennium Prize Problem in mathematics by 2027, 2030, and 2040?*

Background. The seven Millennium Prize Problems[248] were chosen by the founding Scientific Advisory Board of the Clay Mathematics Institute (CMI) of Cambridge, Massachusetts to be the most significant and difficult mathematics problems unsolved by 2000.

Historical Baseline. As of July 2025, only one of the seven problems has been solved (Clay Mathematics Institute, 2025).

For full question background and resolution details, see Appendix E.II. 1. Millenium Prize.

Figures

Millennium Prize: 50th-Percentile Forecasts

[248] *Link provided to participants*: https://www.claymath.org/millennium-problems/

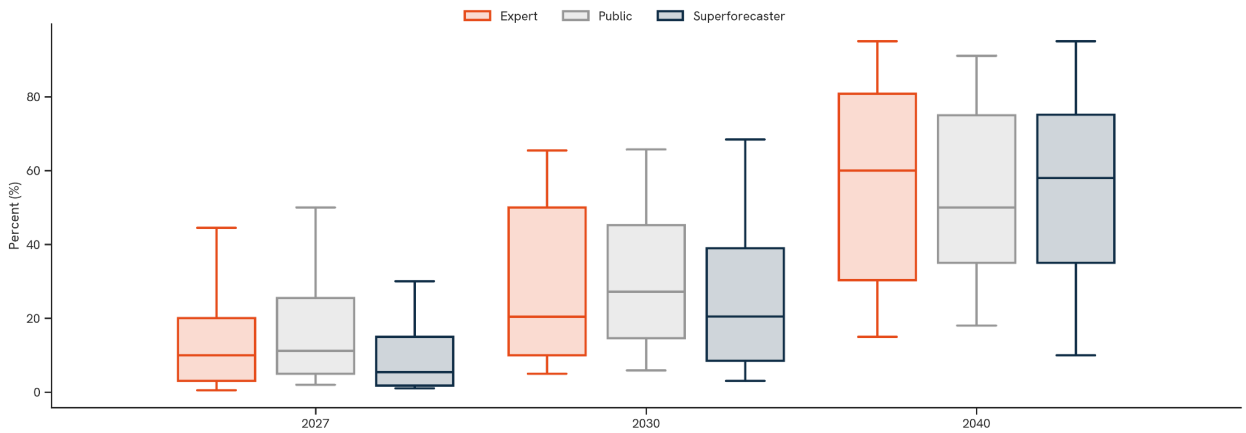


Figure: In this question, participants make 50th percentile forecasts for various resolution dates. This figure shows the 10th, 25th, 50th, 75th, and 90th percentiles of these 50th percentile forecasts, split by participant group. The 25th expert percentile for Dec 2027 represents the number that 25% of experts' median forecasts are lower than.

Results. Experts estimate a 10% chance that AI will solve or substantially assist in solving a Millennium Prize Problem by 2027,[249] up to 20% by 2030,[250] and **60% by 2040**.[251] All categories of experts, superforecasters, and the public largely predict similarly across timescales. However, there is wide disagreement *between* experts: the top quartile of experts think there's at least an even (50%) chance of AI assistance **by 2030**, whereas the bottom quartile of experts think there's only a 10% chance. The disagreement by 2040 is even larger: the interquartile range for expert medians is 30%–81%, while the top decile of experts believe there's a 95% chance and the bottom decile think there's only a 10% chance.

For full results tables, see here.

Rationale analysis:

- **DeepMind/Navier-Stokes**: High-forecast respondents frequently cite DeepMind CEO's January 2025 statement that, in partnership with a team of mathematicians, they're "close to solving" one of the problems (later identified as Navier-Stokes) within "a year or year and a half" (Ansede 2025). This is treated as strong, concrete evidence. Low-forecast respondents generally don't mention this or dismiss corporate pronouncements. One expert cites the Clay Institute president's June 2025 claim: "We're very far away from AI being able to say anything serious about any of those problems" (Heaven 2025).
- **Benchmarks**: Many high-forecast respondents point to International Mathematical Olympiad gold medals (Luong and Lockhart 2025) and FrontierMath progress as evidence of rapid capability growth in mathematical reasoning that will likely continue.

[249] *Raw data*: IQR on the 50th percentile was (3.0%–20.0%). 90th percentile of median forecast: 44.5.
[250] *Raw data*: IQR on the 50th percentile was (10.0%–50.0%). 90th percentile of median forecast: 65.4.
[251] *Raw data*: IQR on the 50th percentile was (30.3%–80.8%). 90th percentile of median forecast: 95.0.

Low-forecast respondents tend to argue that these are fundamentally different challenges. One mathematician notes: “The Math Olympiad is targeted toward gifted high school students spending an afternoon on a problem solvable with known techniques…[whereas] Millennium Prize problems can consume entire careers without a solution.” Multiple forecasters note FrontierMath Tier 4 (which poses much harder problems than Tiers 1-3) has <10% solve rates.

- **The nature of Millennium problems**: High-forecast respondents commonly emphasize that math is verifiable, has clear structure, and that some problems (Navier-Stokes, Birch–Swinnerton-Dyer) may be suited to AI-assisted numerical exploration or pattern recognition. Low-forecast respondents often express doubts that Millennium Problems are solvable with the current AI paradigm, emphasizing doing so requires “deep conceptual breakthroughs,” “developing new concepts and mathematical rules,” and “truly out of the box thinking.” One domain expert writes: “The current generation of AI does not seem to be able to do this sort of creative mathematical work at all. It can apply known techniques and get novel results, but these results would be very easy for top working mathematicians.”
- **Base rates and timelines**: High-forecast respondents mostly don't engage with base rates, or they argue that AI changes the game fundamentally. By contrast, many low-forecast respondents emphasize that only one out of seven problems have been solved in the 25 years since the prize was announced, meaning some have remained unsolved for more than a century. They also highlight Millennium Prize rules: upon the publication of a solution, a minimum of two years must pass before a prize can be awarded, to allow time for adequate verification. (In the case of the one prize that was awarded, the gap between the publication of the solution and the awarding of the prize was over seven years.) This, many low-forecast respondents point out, renders the 2027/2030 dates almost impossible regardless of technical progress.
- **"Substantially assist" interpretation**: High-forecast respondents tend toward a broad interpretation—any meaningful acceleration of human-AI hybrid research counts, whereas low-forecast respondents tend toward restrictive interpretation. One notes the resolution criteria require contribution “likely not producible without AI,” which is a higher bar.
- **Architecture sufficiency**: Most high-forecast respondents believe incremental improvements over current LLM capabilities will be sufficient, especially when paired with specialized tools (Lean, AlphaProof) and human collaboration. Low-forecast respondents frequently argue the current LLM paradigm fundamentally cannot do this. Multiple forecasters say we need “entirely new architectures” (neurosymbolic systems were mentioned several times) or that a “pattern matching paradigm doesn't extend to the deep creativity required.”
- **Difficulty of achieving *superhuman* performance**: Although rarely discussed by high-forecast respondents, a few low-forecast respondents expressed doubts that this could be achieved, with one writing, “Training a model to do math at the level of human experts might be a qualitatively different ML problem from training a model to do math *surpassing* expert capabilities. RL training requires creating problems with reward functions...We haven't achieved that with reasoning post-training yet.”

High-forecast rationale examples

> "I guess the elephant in the room is that DeepMind says they are close: The so-called Navier-Stokes Operation, underway for three years with a team of 20 people, has so far been carried out with complete discretion, although the chief of Google DeepMind, Demis Hassabis let slip in a January interview that they are 'close to solving a Millennium Prize Problem' without mentioning which one. 'We'll see that in the next year or year and a half."[252]

> "Some of the problems, like the Riemann Hypothesis or the Birch and Swinnerton-Dyer Conjecture, are especially well-suited to AI-supported exploration. They bear a kind of family resemblance to the Four-Color Theorem in their relationship to computer-assisted mathematics. The Four-Color Theorem was famously solved through a hybrid of human conceptual framing and extensive computer verification. As Donald MacKenzie details in his socio-history of that episode [reference below[253]], much of the intellectual labor wasn't in the computation itself but in formalizing the problem in a way that machines could meaningfully engage with it and in managing the institutional consequences of proof-by-machine."

> "My optimism that AI could achieve high level original mathematics is revised upward significantly since the Bubeck announcement about GPT-5 a few weeks ago regarding the first confirmed example of novel mathematical reasoning generated by a LLM."[254]

Low-forecast rationale examples

> "I have domain expertise here as a mathematician. The current generation of AI does not seem to be able to do this sort of creative mathematical work at all. It can apply known techniques, and get novel results, but these results would be very easy for top working mathematicians. The kind of pattern matching paradigm we have seen so far apparently doesn't extend at all to deep creativity required."

> "If the millennium problems all require new insights absent from the training data, then current LLM technology is simply not up to the task: we will need instead new AI paradigms that are better at creating non-combinatorial insights (i.e., insights that do not originate from the recombination of patterns already learned by the AI). This will take time: it is not only the time to develop these new AI techniques, but also the time for the humans now riding the wave of machine learning and LLMs to accept that it might be worth their time to look into alternative approaches (more so after extensive efforts to trivialize these approaches as a loss of time). It is the second factor which I think will be the true time bottleneck and could push the resolution of this question further in time."

[252] The expert is referring to and quoting from Ansede (2025).
[253] The expert is referring to MacKenzie (1999).
[254] The expert appears to be referring to an August 2025 post by OpenAI researcher Sebastien Bubeck (Bubeck 2025).

> “Given the progress on Tier 3 FrontierMath problems, a Millennium Problem seems well away, notwithstanding bullish predictions from corporate spokespeople with vested interests.”
>
> “I would put 60% as some hard limit on whether any of the conjectures can be solved at all.”
>
> “I don't think there are many economic incentives to develop those kinds of systems. Millennium problems are very, very hard - much harder than most directly economically useful tasks. They require developing new mathematical theories and techniques to even approach them. As far as I know, current top AI models lack this ability, and I don't see an easy way for them to obtain such an ability (nor are there many economic incentives for building such abilities into them).”

Diffusion of AI Across Sciences

Question. *What percent of publications in the fields of Physics, Materials Science, and Medicine in 2030 will be ‘AI-engaged’ as measured in a replication of the 2024 Harvard University and University of Chicago study?*

Background. A 2024 study (Duede et al. 2024) by researchers at Harvard University and the University of Chicago tracked scholarly engagement with AI across 20 scientific fields by measuring the change in percentage of “AI-engaged publications” within each field. “AI-engaged publications” are papers with abstracts that contain at least one keyword related to contemporary approaches to AI.

For full question background and resolution details, see Appendix E.II. 2. Diffusion of AI Across Sciences.

Figures

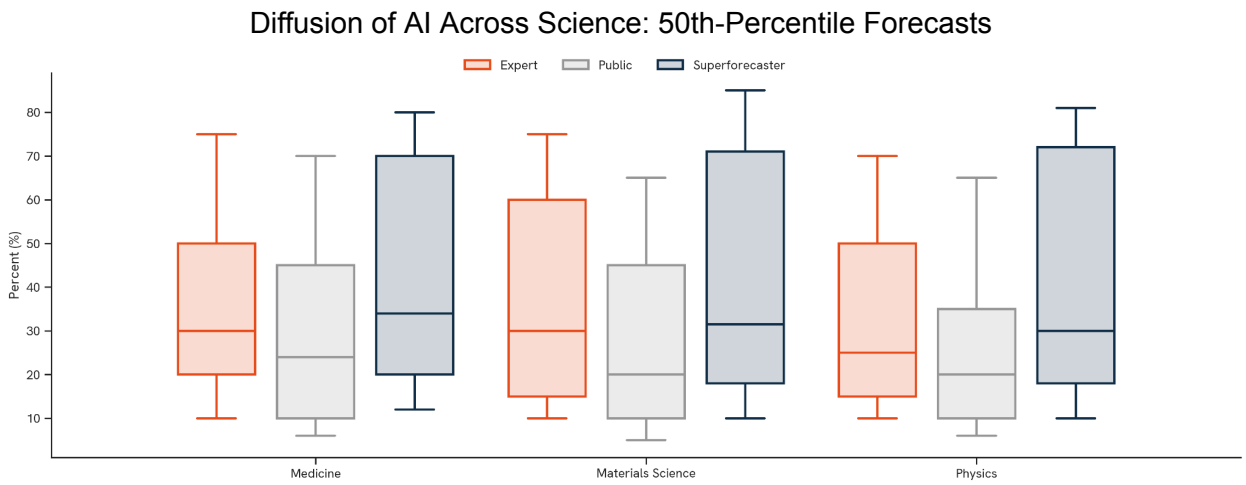


Figure: In this question, participants make 50th percentile forecasts for resolution in 2030. This figure shows the 10th, 25th, 50th, 75th, and 90th percentiles of these 50th percentile forecasts, split by participant group and by field. The 25th expert percentile represents the number that 25% of experts' median forecasts are lower than.

Results. Experts predict a 10x increase (from 3% to ~30%) in AI-engaged papers in Physics,[255] Materials Science,[256] and Medicine,[257] between 2022 and 2030. This means that experts predict roughly 5x faster year-by-year progress in the next 8 years than we saw from 1985 to 2022.[258] However, experts disagree: the interquartile range for expert medians is 15–50%, the bottom decile of experts believe *less than 10%* of papers will be AI-engaged, and the top decile of experts believe *more than 70%* of papers will be AI-engaged. The public expects meaningfully less diffusion than experts and superforecasters, predicting about 21%,[259] roughly two thirds as much as experts.

For full results tables, see here.

Rationale analysis:

- **Measurement concerns:** A significant proportion of forecasters predict low numbers because they believe it will be difficult to determine AI-engagement by abstracts alone: "AI will become so prevalent that its use will be assumed--like the use of a computer--and not mentioned in abstracts." But several high-forecast respondents point out that "major publishers (e.g., Nature, Elsevier) increasingly require AI/tool disclosure," even if "the details of AI tools are still discussed in the main body of the papers."
- **Trend extrapolation:** Many high-forecast respondents emphasize that the academic fields under study attained a combined increase of 1293% since 1985 (Duede et al.

[255] *Raw data*: Median 50th percentile forecast: 25%; IQR on the 50th percentile was (15.0%–50.0%)
[256] *Raw data*: Median 50th percentile forecast: 30%; IQR on the 50th percentile was (15.0%–60.0%)
[257] *Raw data*: Median 50th percentile forecast: 30%; IQR on the 50th percentile was (20.0%–50.0%)
[258] For comparison, from 1985 to 2022—37 years—these fields saw 11x, 8x, and 14x more AI-engaged papers.
[259] Medians for Physics: 20%; Materials Science: 20%; Medicine: 24%.

2024), that this “rapid increase is mostly associated with the seven years between 2015 and 2022,” and “large increases in the percentage of AI-engaged have surely already occurred by Q3 of 2025 compared with the results of the 2022 study.” Low-forecast respondents often question the wisdom of anchoring off these historical trends.

- **AI capabilities:** High-forecast respondents frequently emphasize that “AI will become a ubiquitous research tool” due to domain-specific models, agentic systems, and improved AI literacy among scientists. Low-forecast respondents tend to suspect that unreliability and a lack of interpretability will be a bottleneck. Some argue that to be truly useful in all aspects of research, a breakthrough in the underlying architecture is probably required.
- **Cultural resistance:** A few low-forecast respondents consider that older researchers may be reluctant to adopt AI, and the introduction of AI-literate researchers into these fields will likely be gradual: “Integrating AI into fields that have barely engaged with it in the past will be naturally slow, both because of the lack of interdisciplinary knowledge and natural resistance to change from the existing body of researchers in these fields.”
- **Physics:** Many high-forecast respondents highlight that extensive AI use is likely in data-rich, math-heavy fields of study where AI's mathematical and computational capabilities will be helpful, in particular: particle physics, astrophysics, high-energy physics, quantum systems, anomaly detection, and large-scale simulations. Some low-forecast respondents, however, express skepticism that theory-heavy subfields will see much engagement by 2030: “AI is way less useful for theoretical papers as there is black box problem of not being able to test the hypothesis against causal empirical findings.” Another writes that, in addition to lacking the requisite capabilities, “physics has a conservative publishing culture -- theorists in particular won't add AI to their abstracts unless the method is clearly central.”
- **Materials Science:** As with physics, high-forecast respondents often emphasize that AI is well-suited to assist with data-rich, math-heavy subfields like molecular design and discovery, computational chemistry, energy storage, inverse design, high-throughput screening, and simulation AI coupling. One notes: “Materials science is...the one with the highest potential to be AI-engaged in the next decade due to its large dependence on knowledge coming from sophisticated combinatorics from a fixed set of elements.” Some also stress that industrial demand for accelerated discovery, and the financial rewards that might follow, will provide a strong push. One low-engagement forecaster stressed that “materials science likely [won’t] benefit from vast troves of data crossing disciplines unless new methods of AI accessible data collection are developed.”
- **Medicine:** High-forecast respondents commonly note that AI is already used in imaging, diagnostics, genomics, drug discovery, medical natural language processing, and personalized medicine design: “Medicine shows the most rapid uptake, fueled by the pervasive use of AI in medical imaging diagnostic algorithms, and clinical decision support systems.” One emphasizes that “AI could enable the utilization of the enormous amount of medical data contained in electronic health records (EHRs).” Low-forecast respondents focus more on the possibility that regulatory, data privacy, validation, and ethical concerns could limit use, and that the number of observational clinical case studies and trials, which are unlikely to involve AI, will keep the percentage low: “A significant portion of papers are observational, often reporting causal effects. There isn't

much room for AI in these sorts of papers, as current statistical methods are more reliable and bias-free, compared to AI."

High-forecast rationale examples:

"Historical trending seems ineffective in projecting 2030 levels for these variables, due to the rapid increase in AI availability and the recent exponential growth in AI-engagement."

"Publication is a lagging indicator of the actual research work being done now, so the 3% numbers in 2022 was [a] significant understatement."

"Funding opportunities for AI-related research in these domains will grow over the next few years and yield a higher percentage of AI-engaged research papers. While the Trump administration has thrown the NSF research ecosystem into general chaos, they have heavily signaled that AI-related research funding will be prioritized and expanded. This is evidenced by NSFs July 2025 announcement of a $100 million investment in National Artificial Intelligence Research Institutes awards to secure American leadership in AI – this investment will be partially directed towards research efforts in medicine and materials science (among other focuses)."

Low-forecast rationale examples:

"Just to be clear, every researcher will be using AI by 2030, if they aren't already, to help with aspects of the research process. But will it show up in an abstract? If it's text analysis, NLP, or specialized AI tools, it will show up. But even substantial use in many fields won't make it to the abstract. No one says I used a computer and python to generate my empirical results, even if they did. Much AI use will be like that---a particular tool."

"AI is a black box that hallucinates and its ability to be truly creative / inventive (outside of hallucinating) is contested. Thus, I expect the physics community to only apply AI in an assistive manner for a narrow set of tasks / fields (e.g., cosmology, quantum error correction)."

Drug Discovery

Question. *What percent of sales of recently approved U.S. drugs will be from AI-discovered drugs and products derived from AI-discovered drugs in 2027, 2030, and 2040?*

Background. We define "recently approved U.S. drugs" as drugs which received approval for sale by the U.S. Food and Drug Administration (FDA) in the one year preceding the resolution year. "AI-discovered drugs" are drugs developed through significant use of AI techniques in processes that would likely not have occurred without post-2022 AI capabilities.

For full question background and resolution details, see Appendix E.II. 3. Drug Discovery.

Figures

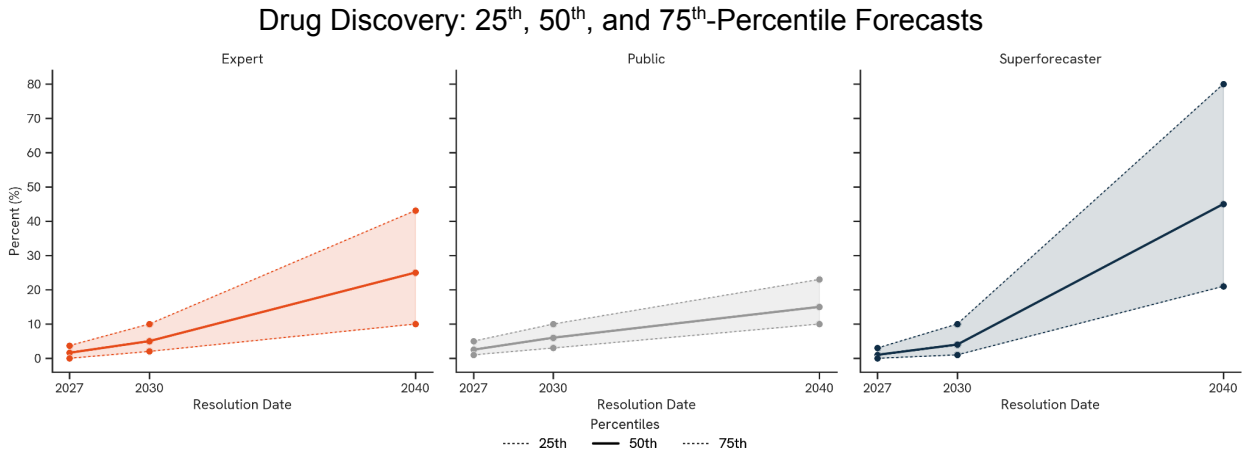
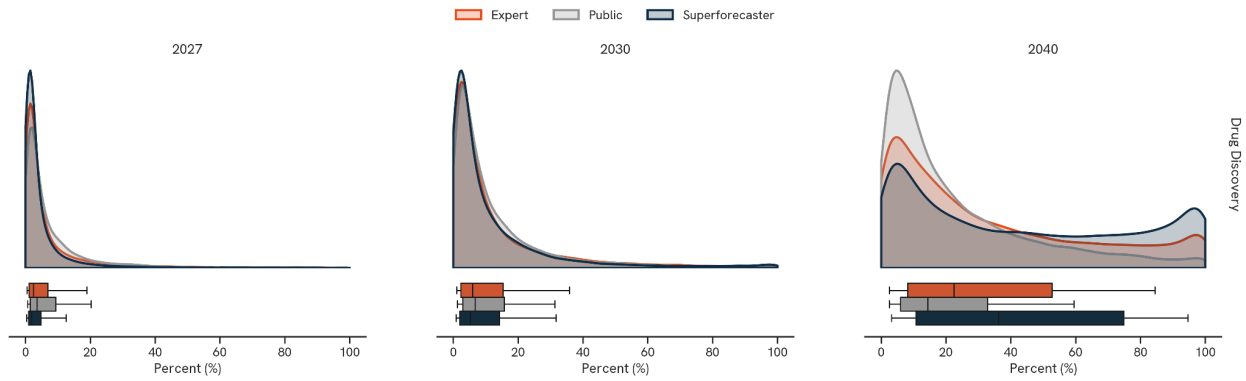


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.



Drug Discovery: Pooled Distributions

Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. For more details on how we calculate pooled distributions, see Appendix C. Pooled Distribution Estimation.

Results. Experts predict that 1.6% of recently approved drug sales in 2027 will come from AI-discovered drugs,[260] 5% by 2030,[261] and **25% by 2040**.[262] Experts, superforecasters and the public largely forecast similarly for 2027 and 2030, but sharply diverge in 2040. Experts predict almost double the sales relative to the public (25% vs 15%[263]), and superforecasters predict almost double relative to experts (45%[264] vs 25%), with superforecasters in aggregate giving 25% that more than 70% of sales will come from AI-discovered drugs. Drug discovery is one of the few settings where superforecasters are meaningfully more optimistic than experts.

For full results tables, see here.

Rationale analysis:

- **FDA approval timelines:** Several high-forecast respondents believe that AI will accelerate discovery-to-market timelines through faster design-make-test-analyze loops and potentially AI-enabled pharmacodynamic simulations that could streamline clinical trials. Others note that drug discovery-to-market timelines can be shortened significantly during times of crisis via EUAs (emergency use authorizations). Low-forecast respondents commonly emphasize regulatory realities that may limit AI's impact on approval timelines: “Given that the median time it takes to get through the FDA approval process is over 10 years, and no AI-discovered drugs appear to have started Phase III trials yet,[265] 2027 is likely too soon for many, if any, new AI drugs to be approved."L
- **Phase I success rates:** Many high-forecast respondents note that AI-discovered drugs already demonstrate significantly higher Phase I success rates, and that “extrapolating from current rates of increase in the number of proposed AI drugs, these will constitute a majority of new clinical trial submissions.” Several low-percentage forecasters, however, think that “the turnaround time between Phase I and approval will not speed up substantially for AI-invented drugs,” because “early entrants sped through Phase I but then quickly reverted to the mean in Phase II.”
- **AI ubiquity by 2040:** Many high-forecast respondents expect that by 2040, AI will become “a standard discovery tool.” One writes, “by then, I expect AI to be fully embedded in how drug discovery is done” and another that, “as some companies invest in AI adoption and see payoffs, more competitors and startups will do the same in order to compete.” But several low-percentage forecasters argue institutional inertia will slow adoption: “While there are some pharma companies that have widely embraced the usage of deep learning models...much of the industry is fairly slow moving to adopt new

[260] *Raw data*: IQR on the 50th percentile was (0.5%–4.1%); median 25th and 75th percentile forecasts were 0.0% and 3.7% respectively.
[261] *Raw data*: IQR on the 50th percentile was (2.2%–10.0%); median 25th and 75th percentile forecasts were 2.0% and 10.0% respectively.
[262] *Raw data*: IQR on the 50th percentile was (10.0%–50.0%); median 25th and 75th percentile forecasts were 10.0% and 43.1% respectively.
[263] *Raw data*: IQR on the 50th percentile was (6.9%–30.0%); median 25th and 75th percentile forecasts were 10.0% and 23.0% respectively.
[264] *Raw data*: IQR on the 50th percentile was (15.0%–70.0%); median 25th and 75th percentile forecasts were 21.0% and 80.0% respectively.
[265] This was true at the time this expert completed the survey.

techniques....[The] market shift will likely take a decade at least to fully change the nature of biomedical research."

- **What qualifies as "AI-discovered":** Most high-forecast respondents interpret the term broadly, with one expert noting, "It gets tricky to pin down what counts as an AI-discovered drug. Based on you using reports like the Boston Consulting Group… I'm assuming you go for a fairly broad sense, i.e., that earlier attempts at ML assistance in drug discovery count as AI." Low-percentage forecasters tend to question whether narrow tools will count: "Traditional methods (including Bayesian methods and random forest) with simple computational features can already perform well on drug property prediction...do Bayesian methods and random forest count as AI?"
- **Power laws:** Some high-forecast respondents hypothesize that AI-discovered blockbusters may dominate sales: "Sales will probably follow a power law of some sort (i.e., a small number of drugs will have a large number of sales). If AI invents one or more blockbusters, then sales might be highly skewed." Echoing that sentiment, another writes, "there's a fat tail from the possibility that one or more AI-discovered wonder drugs gets rapid approval and huge sales." Several low-forecast respondents point to the possibility that AI may instead "increase the availability of treatments for rare diseases, but these drugs would likely not make up a large part of total new drug sales."
- **Regulatory environment:** A few high-percentage forecasters cite the potential for deregulation under the current administration and emphasize what seems to be a "bipartisan recommendation to reduce [the] FDA drug timeline." Several low-percentage forecasters see the current political climate as constraining: "I kept my estimates low in the near term because of the ongoing disruption to FDA approval and drug discovery pipelines under the current U.S. Government and Trump Administration."

High-forecast rationale examples:

> "My 2040 distribution…presumes near-ubiquity of AI as the decisive engine for discovery, foundation models over multimodal omics, autonomous labs closing design–make–test–analyze loops. If those pieces land, the majority of recent approvals and their sales could plausibly trace to AI-originated, hence the very high median."
>
> "My mainline expectation is that by 2030 AI is very helpful here but isn't totally dominant, but that by 2040 we've had an intelligence explosion leading to artificial superintelligence, and AI systems have ~completely automated drug discovery."
>
> "The Trump administration's removal of CDC officials may trigger a move towards a faster approval process managed by third-party contractors or slow down. Given the bipartisan recommendation to reduce [the] FDA drug timeline, this will likely keep the projected trend."

Low-forecast rationale examples:

> "Base rate: 9% of drugs entering clinical trials are AI-discovered (and the proportion is increasing); on average, drugs take 10 years to get from trials to approval; AI-discovered

drugs are slightly more likely to pass trials; therefore, 10 years from now, slightly more than 9% of recently approved drugs should be AI-discovered."

"My first assumption is that the turnaround time between Phase 1 and approval will not speed up substantially for AI-invented drugs. Early entrants sped through Phase 1 but then quickly reverted to the mean in Phase 2. If this trend holds, then AI drugs might get through the process a little more quickly than the historical average…"

"While there are some pharma companies that have widely embraced the usage of deep learning models for exploration of molecular compounds (namely Regeneron) to accelerate drug discovery and protein synthesis, much of the industry is fairly slow moving to adopt new techniques….[The] market shift will likely take a decade at least to fully change the nature of biomedical research due to the duration of patents and the high levels of resistance to change in that industry."

Electricity Consumption

Question. *What percent of U.S. electricity consumption will be used for training and deploying AI systems in 2027, 2030, and 2040?*

Background. AI model training and deployment occur mainly through the use of data centers.[266] Servers in data centers are split into two main categories: conventional and AI specialized servers.[267] This question asks about the electricity consumption of AI specialized servers.

Historical baselines. There is currently no official reporting, but we estimate 1.0% in 2024.

For full question background and resolution details, see Appendix E.II. 4. Electricity Consumption.

Figures

Electricity Consumption: 25th, 50th, and 75th-Percentile Forecasts

[266] *Link provided to participants:* https://www.iea.org/reports/energy-and-ai/energy-demand-from-ai
[267] *Link provided to participants:* https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

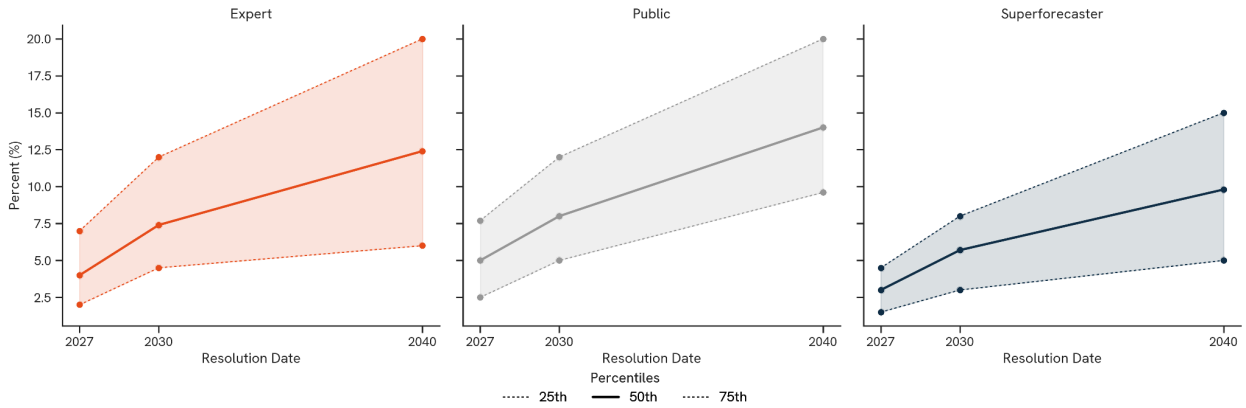


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.

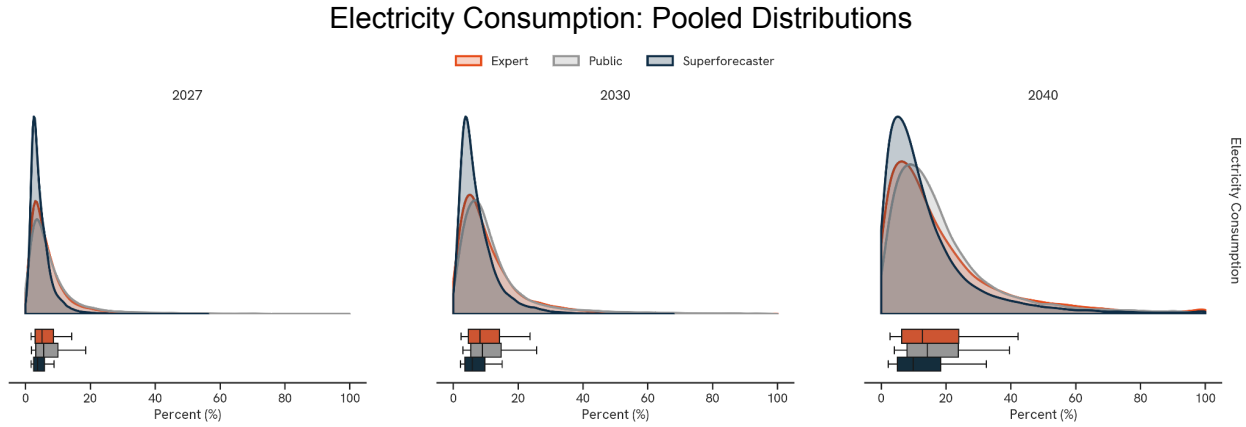


Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. For more details on how we calculate pooled distributions, see Appendix C. Pooled Distribution Estimation.

Results. The median expert predicts that 4% of U.S. electricity consumption will be used for training and deploying AI systems in 2027.[268] That rises to **7% of all electricity consumption in 2030**,[269] and close to double that (12%) in 2040.[270] For context: 7% is 1.5x today's entire data-center load, and 12% is close to all of Texas' electricity use. The top quartile of experts predict that AI training and deployment will account for more than 20% of total U.S. electricity consumption by 2040, and the top 10% believe it will account for more than 30%.[271] 20% is almost all of the industrial sector's electricity use. Experts and the public predict similarly across

[268] *Raw data*: IQR on the 50th percentile was (3.0%–6.4%); median 25th and 75th percentile forecasts were 2.0% and 7.0% respectively.
[269] *Raw data*: IQR on the 50th percentile was (5.0%–10.0%); median 25th and 75th percentile forecasts were 4.5% and 12.0% respectively.
[270] *Raw data*: IQR on the 50th percentile was (8.0%–19.7%); median 25th and 75th percentile forecasts were 6.0% and 20.0% respectively.
[271] *Pooled distribution*: IQR (6.31%–24.13%); variance decomposition: 46.37% between–forecaster disagreement, 53.63% within–forecaster uncertainty.

all dates, while superforecasters are slightly less optimistic—predicting 3%,[272] 6%,[273] and 10%.[274]

For full results tables, see here.

Rationale analysis:

- **Infrastructure**: Many high-forecast respondents emphasize that tech companies are making unprecedented capital commitments to AI infrastructure: "With the building of data centers and dramatic investments in large-scale infrastructure, such as Project Stargate, alongside policies and executive orders enabling the leasing of federal land for more AI data centers…I absolutely think electricity consumption for AI in the near term will skyrocket." Others point to the possibility that competition between companies and nations for supremacy in AI may lead to "an explosion in energy usage." Low-forecast respondents tend to focus more on potentially formidable constraints, particularly when considering "the material and political investments necessary to get significant growth—physical data centers, chips, permitting, water for cooling, transmission lines, etc."
- **Projections**: Many forecasters anchor on institutional projections: "The Boston Consulting Group says 7.5% in 2030 [for data centers]. Bloomberg says 8.6% [for data centers] by 2035…AI-specialized servers accounted for 15% of total global data center demand in 2024. Electricity usage by specialized AI hardware accounted for 11-20% in 2024. Demand implies that a share of AI-related U.S. data center electricity consumption of 50/180=27.8%."
- **Efficiency gains:** A common low-forecaster consideration was that "there is a huge amount of scope for efficiency gains [and] things will become more efficient as they scale up," and several noted that "DeepSeek appears to consume considerably less power" than frontier U.S. models. High-forecast respondents, however, largely don't believe efficiency improvements will meaningfully reduce overall consumption: "More efficient chips and algorithms will simply lead to more compute going into training and inference." One argues, "Jevons paradox will surprise on the upside - i.e., even more efficient chips will mean more usage with little savings in terms of energy efficiency."
- **AI bubble:** Many low-forecast respondents emphasize business model concerns: "The current economics of all of this are obviously not sustainable: we're spending billions of dollars on these data centers in support of AI companies, or AI segments of hyperscalers, that are losing huge amounts of money off their core AI businesses." Multiple forecasters note "AI will need to begin demonstrating economic and social utility

[272] *Raw data*: IQR on the 50th percentile was (2.0%–4.0%); median 25th and 75th percentile forecasts were 1.5% and 4.5% respectively.
[273] *Raw data*: IQR on the 50th percentile was (3.7%–8.0%); median 25th and 75th percentile forecasts were 3.0% and 8.0% respectively.
[274] *Raw data*: IQR on the 50th percentile was (6.0%–15.0%); median 25th and 75th percentile forecasts were 5.0% and 15.0% respectively.

rapidly to justify the investments in physical infrastructure necessary to sustain rapid growth in energy consumption."

- **Geopolitical competition:** Some high-forecast respondents emphasize that "China is also investing massive amounts in datacenters," with one writing "there's a chance that we enter an arms race that is mostly determined by who can pump the most electricity into AI." A low-forecast respondent suggests this dynamic, instead of resulting in U.S. growth, could drive infrastructure offshore: "Major developers will possibly respond by increasingly outsourcing the physical infrastructure of data processing to locales outside of the US—there's no particular reason why models need to be trained inside of U.S. borders where the economic and political expenses are potentially much higher."
- **Scaling plateau:** High-forecast respondents typically expect continued returns from scaling: "I think it's more likely than not that transformative superintelligence will have arrived by 2040, in which case almost all knowledge work …will be reliant on the use of AI models." Whereas low-growth forecasters commonly express that they anticipate diminishing returns: "Unless further scaling leads to qualitative leaps in AI capabilities, the growth in electricity consumption for AI training and deployment is unlikely to be exceptionally rapid;" "We have already seen indications that the limits of scaling may be soon reached."
- **Denominator effects:** Rarely emphasized by high-forecasters, low-forecast respondents frequently pointed out that "while growth in power production will likely be necessary for much growth in AI systems, overall power consumption is the denominator in this calculation, and increased demand from electric vehicles could play a role in growing this denominator" and therefore "data centers' energy consumption will grow in absolute terms, but not so much in relative terms."

High-forecast rationale examples:

> "... the exponential scaling of AI model training, the deployment of increasingly powerful systems across all industries, and the emergence of autonomous AI agents that run continuously. This explosive growth trajectory reflects AI becoming as fundamental to the digital economy as the internet itself."
>
> "Given the likely lack of regulation related to AI-energy consumption in the next several years…and the race to the bottom between AI companies and nation states to deploy these systems, I think we are likely to see an explosion in energy usage related to this technology."
>
> "The U.S. Department of Energy is seriously considering the possibility of allowing companies with large data centers dedicated to training AI models to install a small nuclear power plant nearby that would be able to meet their energy needs. This strategy would also reinforce the U.S. commitment to energy sources that do not emit greenhouse gases. The most solid option right now, both from a technical and economic perspective, are compact modular reactors known as SMRs (Small Modular Reactors)."

Low-forecast rationale examples:

"I'm still seeing this as a period of build-up: we'll see major investment in new data centers and specialized AI hardware, but constraints like power availability, construction timelines, and chip supply will keep growth modest."

"Major developers will possibly respond by increasingly outsourcing the physical infrastructure of data processing to locales outside of the US—there's no particular reason why models need to be trained inside of U.S. borders where the economic and political expenses are potentially much higher."

Cognitive Limitations, Part II

Question. *By the end of 2030, what percent of LEAP expert panelists will agree that each of the following is a serious cognitive limitation of state-of-the-art AI systems?*

- **Hallucination / Inaccuracy**: they give plausible-sounding but incorrect or fabricated information.
- **Shallow reasoning**: their reasoning capabilities are shallow beyond math and coding and lack "genuine" causal and logical reasoning.
- **Lack of long-term memory**: they cannot retain and utilize information across sessions or long interactions.
- **Limited ability to generalize**: they perform badly in tasks they have not been trained on and are limited by the quality of human-generated training data, limiting their creativity.
- **Limited metacognition and continual learning**: they cannot reliably assess and regulate their own cognitive processes, e.g., notice and correct their own errors, question previous assumptions, know when to ask for help or to defer, know how to improve a capability they don't have, continue to acquire new capabilities after training, etc.
- **Limited Embodiment / Robotics**: they perform badly in processing multimodal inputs beyond text and vision and at navigating the physical world.
- **Limited inter-system collaboration**: they cannot effectively coordinate with other AI systems to negotiate goals, allocate subtasks, and work in parallel under standardized protocol.

For full question background and resolution details, see Appendix E.II. 5. Cognitive Limitations, Part II.

Figures

Cognitive Limitation: 50th-Percentile Forecasts

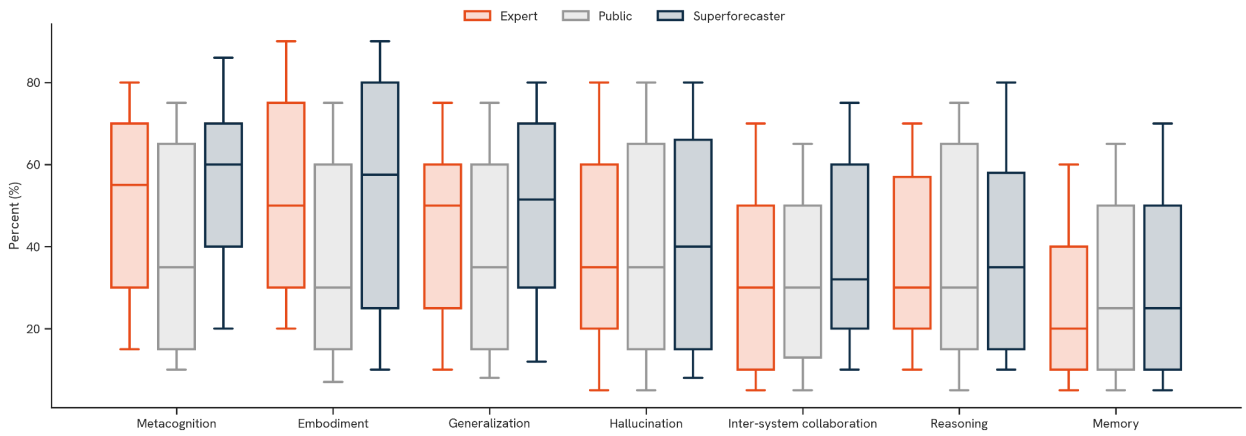


Figure: In this question, participants make 50th percentile forecasts for resolution in 2030. This figure shows the 10th, 25th, 50th, 75th, and 90th percentiles of these 50th percentile forecasts, split by participant group and by cognitive limitation. The 25th expert percentile represents the number that 25% of experts' median forecasts are lower than.

Results. Experts predict that memory, hallucination, reasoning, and inter-system collaboration are not likely to be judged as serious cognitive limitations for AI in 2030 (20%,[275] 35%,[276] 30%,[277] and 30% respectively[278]), whereas metacognition, embodiment, and generalization are predicted to be more serious (55%,[279] 50%,[280] 50%[281]). There's substantial disagreement between experts: for example, half of experts' best guesses for whether Embodiment will be rated a serious limitation span 30-75%, whereas the bottom decile of experts predicts less than 20%, and the top decile predicts more than 90%. There's no consensus that any particular limitation will turn out to be serious. The public predicts fairly uniformly across limitations,[282] which means they predict substantially lower than experts on Metacognition, Generalization and Embodiment.

For full results tables, see here.

Rationale analysis:

- **General.** Forecasters disagree strongly on whether incremental improvements through scaling and post-training enhancements will suffice to overcome AI's cognitive limitations. Low-forecast respondents tend to believe that "reasoning, memory, and embodiment are all something that can be improved with more computing power, more parameters, and more data." High-forecast respondents often emphasize that limitations

[275] *Raw data*: IQR on the 50th percentile was (10.0%–40.0%)
[276] *Raw data*: IQR on the 50th percentile was (20.0%–60.0%)
[277] *Raw data*: IQR on the 50th percentile was (20.0%–57.0%)
[278] *Raw data*: IQR on the 50th percentile was (10.0%–40.0%)
[279] *Raw data*: IQR on the 50th percentile was (30.0%–70.0%)
[280] *Raw data*: IQR on the 50th percentile was (30.0%–75.0%)
[281] *Raw data*: IQR on the 50th percentile was (25.0%–60.0%)
[282] The median response for each limitation falls between 25-35%.

are “intrinsically interlinked,” and that to truly solve them, “we need entirely new architectures.”

- **Hallucination/inaccuracy.** Many low-forecast respondents emphasize that the frequency of hallucinations has already been substantially reduced with some predicting that “productized retrieval, tool use, and verifiers [will] cut hallucinations enough that only a minority will still call them serious.” Some also predict that users will become increasingly adept at prompting to avoid hallucinations. Several high-forecast respondents maintain that “even with retrieval-augmented generation and tool use, hallucinations will remain a widely recognized constraint especially in high-stakes settings like medicine and law.”
- **Shallow reasoning.** Low-forecast respondents often highlight recent progress, noting, “the advancements that have been made in just the last year to agentic systems and reasoning engines have [been] shocking levels of improvements.” They tend to think that test-time compute, tool-augmented reasoning, and other inference enhancements will lead to additional improvements. High-forecast respondents, however, tend to think that, “while AI can be expected to achieve superhuman performance in formal closed-system domains like math, code, and certain scientific domains, the ability to perform robust informal reasoning about the messy open world remains a frontier problem that will likely require new architectural breakthroughs.”
- **Long-term memory.** There is relative consensus that the memory issue is solvable, with disagreements mainly about timeline and completeness. Low-forecast respondents point to technical solutions: “Memory systems (vector databases, stateful agents, persistent contexts) are improving rapidly.” High-forecast respondents frequently acknowledge progress but emphasize limitations: “How to enable AI to work continuously over years like a human expert remains a significant and currently challenging problem.” They focus more on technical challenges, including “context rot which limits the use of explicit in-context prompting as a tool for memory,” and “the general limitations of RAG [retrieval augmented generation].”
- **Limited ability to generalize.** Low-forecast respondents often emphasize that transfer learning will help, along with "the explosion, and use, of synthetic data, greater progress in test-time learning, and programmatic reasoning.” Some high-forecast respondents acknowledge that “few-shot learning and transfer learning continue to improve,” but tend to think that an inability to generalize is inherent to the underlying architecture, and that “failures on out-of-distribution or truly novel tasks will still be seen as a serious barrier especially by economists and policymakers.”
- **Metacognition and Continual Learning.** Many forecasters identify this as the core unsolved limitation. Low-forecast respondents are scarce and many high-forecast respondents are emphatic: “I deeply believe metacognition is the number one limitation today and that it will remain the case by 2030.” They emphasize that models “don't know what they don't know” in that they are unable to reliably self-correct, defer when appropriate, or guide themselves through the acquisition of new skills and knowledge—and there is no clear solution to address these issues on the horizon. One forecaster bluntly states: “I don't know how to solve this with RL [reinforcement learning]. Serious fundamental advances are needed.”

- **Embodiment/Robotics.** Disagreements center on timelines. Low-forecast respondents commonly point to recent progress: “If you view a Waymo vehicle as basically a robot on wheels, it is apparent that the problem of navigating in physical space in the real world is manageable.” High-forecast respondents tend to emphasize fundamental constraints: “To generate training data we need to deploy robots in real-world settings to collect data. However, doing this on a large scale presents practical challenges.” A roboticist forecaster agrees: “Embodied AI is significantly more challenging than many in the AI community believe. It is severely data restricted.” Many high-forecast respondents think use cases in unstructured environments (agriculture, construction) are likely to be particularly constrained.
- **Inter-system collaboration.** Many low-forecast respondents emphasize that this “appears to be in large part a systems integration and standardization problem that will see major progress in light of strong commercial incentives.” They note rapid progress has already been made via the deployment of tools that “permit a supervised semi-autonomous optimization framework wherein instruction sets, for submission to subordinate models, are able to be optimized, evaluated, and aligned with human experts.” Some low-forecast respondents, however, point out that “AI systems represent various individuals, companies, and organizations, each with their own interests,” leading to a “lack of interoperable standards.” One notes: “While short-term collaboration between AI systems is feasible and not a major obstacle, sustained, coordinated collaboration over extended periods remains challenging.”

Rationale examples:

Hallucination: “Strategies such as tool-use, retrieval, constrained decoding, uncertainty-aware outputs, and model-graded self-checks (already somewhat in use) will be refined and come into wider use to help mitigate this limitation.”

Reasoning: “The advancements that have been made in just the last year to agentic systems and reasoning engines have already proven the shocking levels of improvements that can be made when an entire industry focuses on improving this skill in GenAI services. The focus on this capability isn't going away and there are a lot of highly capable people working on this.”

Memory: “Memory systems (vector databases, stateful agents, persistent contexts) are improving rapidly and likely to be well-integrated by 2030. Many limitations here may shift from technical (it can't be done) to engineering (it's expensive, brittle, or not universally applied).”

Metacognition: “This may remain the hardest frontier. By 2030, models will still struggle with self-evaluation, error detection, and self-directed learning. Without robust metacognition, systems can't fully know what they don't know. Even if continual learning improves, the deep challenge of autonomous capability acquisition will persist. I expect this to have the highest agreement among panelists.”

Intersystem collaboration: “This appears to be in large part a systems integration and standardization problem that will see major progress in light of strong commercial incentives. As a growing number of specialized AI agents are deployed, standardized communication protocols will be developed to allow them to negotiate and allocate tasks. The technical ability for systems to coordinate on well-defined tasks will be largely in place.”

Appendix F.III. Question-by-Question Results: Wave 3

AI Investment

Question. *What will be the global private investment (in billion USD) in AI in 2027 and 2030?*

Background. According to the *AI Index Report 2025 Annual Report*,[283] private investment in AI includes investment in AI startups that have received over $1.5 million in investment since 2013. Our World in Data (OWID) notes that this indicator is likely to underestimate total global AI investment.[284,285] We use this series for resolution.

Historical baseline. According to OWID, global private AI investment was approximately **$130 billion** in 2024.[286]

For full question background and resolution details, see Appendix E.III. 1. AI Investment.

Figures

AI Investment : 25th, 50th, and 75th-Percentile Forecasts

[283] *Link provided to participants:* https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf
[284] The survey instrument links to *Our World in Data (2025a), Annual Global Corporate Investment in Artificial Intelligence, by Type.* https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type.
[285] Regarding their private AI investment indicator, Our World in Data notes: 1. “The data likely underestimates total global AI investment, as it only captures certain types of private equity transactions, excluding other significant channels and categories of AI-related spending;” 2. “The source does not fully disclose its methodology and what's included or excluded. This means it may not fully capture important areas of AI investment, such as those from publicly traded companies, corporate internal R&D, government funding, public sector initiatives, data center infrastructure, hardware production, semiconductor manufacturing, and expenses for research and talent.” See Our World in Data (2025b) for more details on what is likely excluded.
[286] *Link provided to participants:* https://ourworldindata.org/grapher/private-investment-in-artificial-intelligence#sources-and-processing

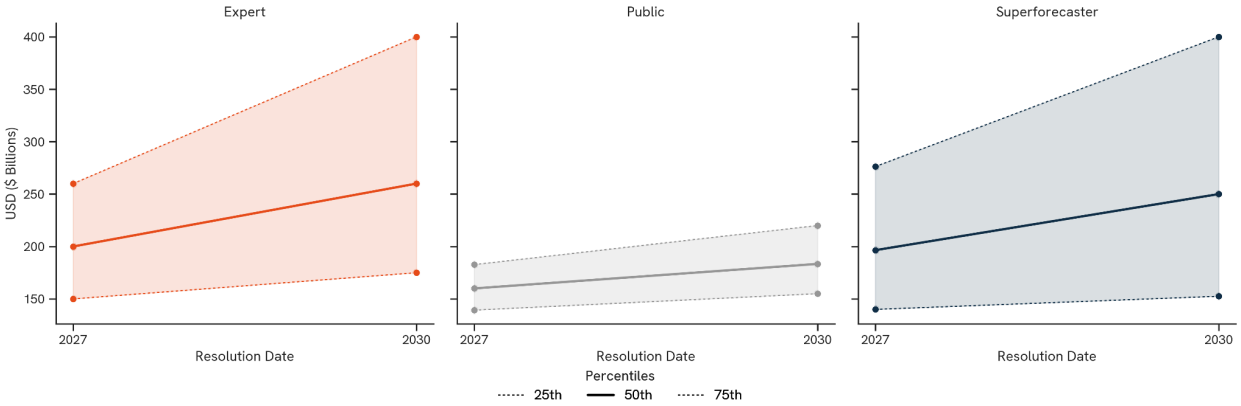


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.

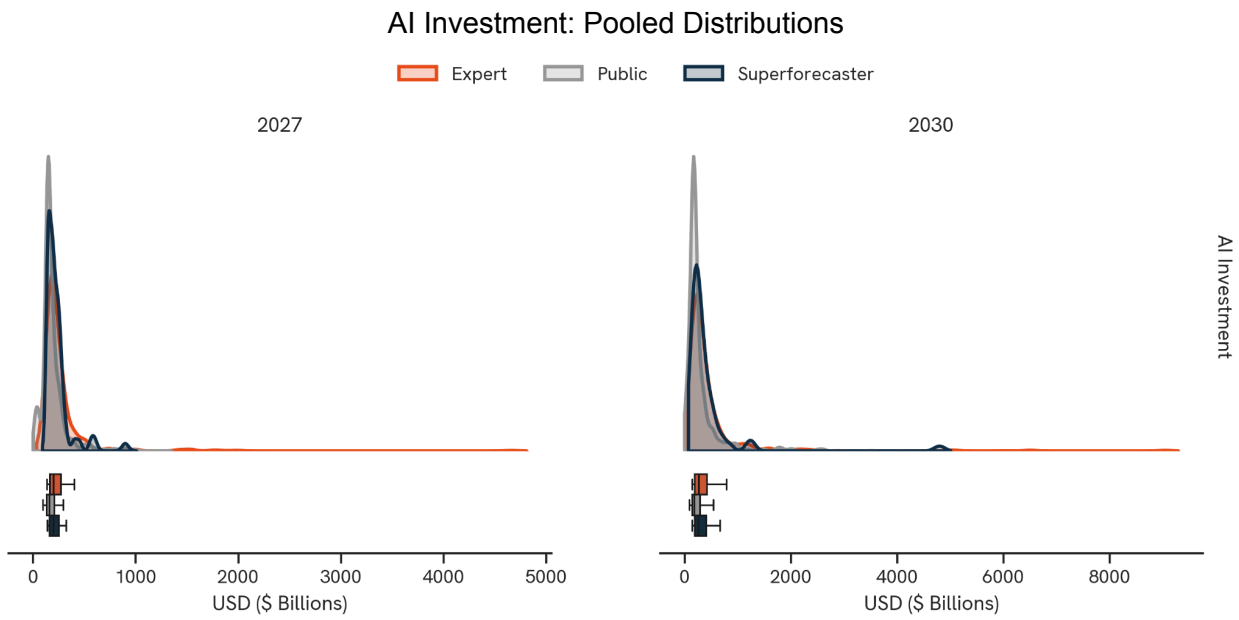


Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. For more details on how we calculate pooled distributions, see Appendix C. Pooled Distribution Estimation.

Results. The median expert predicts 200 billion USD in private investment in AI by 2027,[287] and 260 billion USD in 2030,[288] almost double the 130 billion today.[289] 50% of experts believe that investment in 2030 will be between 196–400 billion USD, with 25% each on either side of this interval. The top decile believes that investment will be greater than 750 billion USD in the median scenario. Experts and superforecasts largely predict similarly, whereas the general public believes there will be substantially less global private investment in AI: 160[290] vs 200 billion USD in 2027, and 183[291] vs 260 billion USD in 2030.

For full results tables, see here.

Rationale analysis:

[287] *Raw data*: IQR on the 50th percentile was (169.0–268.3); median 25th and 75th percentile forecasts were 150.0 and 260.0 respectively.
[288] *Raw data*: IQR on the 50th percentile was (196.3–400.0); median 25th and 75th percentile forecasts were 175.0 and 400.0 respectively.
[289] *Link provided to participants:* https://ourworldindata.org/grapher/corporate-investment-in-artificial-intelligence-by-type?country=~Private+investment#sources-and-processing
[290] *Raw data*: IQR on the 50th percentile was (138.0–210.0); median 25th and 75th percentile forecasts were 139.3 and 182.7 respectively.
[291] *Raw data*: IQR on the 50th percentile was (150.0–300.0); median 25th and 75th percentile forecasts were 155.0 and 220.0 respectively.

- **Trend extrapolation**: Most high-forecast respondents extrapolate from historical trends showing over 30% compound annual growth from 2013-2024 and believe this momentum can continue. One argues, “$1T [trillion] is not very much compared to the total pool of investable assets.” Another notes that “the strong rebound to ~$130 billion in 2024 is critical. It occurred despite higher interest rates, signaling powerful, non-speculative belief in the transformative potential of generative AI.” But many low-forecast respondents worry about the AI bubble bursting, with one forecaster noting “both Deutsche Bank and Bain & Co. have just warned that the current AI boom is not sustainable,” (Edwards 2025) and another likening the current situation to “the dot com bubble in 2000.”
- **Enterprise adoption timelines**: High-forecast respondents frequently note that “AI adoption is still in its early stages across many industries, suggesting there is substantial room for further expansion.” One forecaster quotes a 2025 McKinsey report: “Over the next three years, 92 percent of companies plan to increase their AI investments. But while nearly all companies are investing in AI, only 1 percent of leaders call their companies 'mature' on the deployment spectrum.” Some low-forecast respondents, however, argue that “the uptake for commercial purposes is still fairly slow,” and worry that productivity gains may not materialize quickly enough to justify high levels of investment. One noted, “Anthropic CEO's forecast of 90% of coding in the USA done by AI 'within six months' has been a fantastic dud” (Council on Foreign Relations 2025).
- **Economic downturn**: Although rarely emphasized by high-forecast respondents, pessimists often stress macroeconomic risks, with one noting “The NBER [National Bureau of Economic Research] lists 14 recessions since (but not including) the Big One in 1929,” (Federal Reserve Bank of St. Louis 2025) and from that calculates that “the chances of an economic contraction through 2027 and 2030 are 33% and 61% respectively, assuming a Poisson [i.e., random] process.”
- **Market evolution**: High-forecast respondents tend to expect continued startup ecosystem growth, believing, “AI start ups & companies are staying in the private market for longer as it's much easier to get funding.” One notes that: Open AI has raised billions this year in the private market and that “these companies can also remain more agile to compete against a very fast changing market.” But some low-forecast respondents anticipate consolidation, with one arguing that “some leading companies will raise less private capital as they mature, merge, or turn to other funding sources.”
- **Investment sources:** Some high-forecast respondents emphasize international expansion potential: “I suspect the number will be much higher than forecast as China's economy matures and begins to lift economies throughout the Asian region, including India. U.S. AI is just the tip of the iceberg.” Low-forecast respondents tend to express skepticism, with one noting that “investment is highly concentrated in the US,” and that “I don't expect the rest of the world to pick up the slack.”

High-forecast rationale examples:

> “If we assume constant linear growth, then investment should amount to $242 billion in 2027 and $355 billion in 2030. If growth is exponential, then we’re looking at quite a bit more: $360 billion in 2027 and $997 billion in 2030.”

> "Goldman Sachs 2023 forecast for 2024 [1] appears to be spot-on (132B vs actual 130B), although they seem to have overestimated the 2023 amount (110B vs actual 93B). Their projection for 2025 was at 158B; they seem to use largely the same data & methodology with the Stanford AI Index Report, including the amounts expressed in constant 2021 US$. Based on this forecast, moderate projections suggest that global private investment in AI could surpass $250-300 billion by 2027 and could even approach $500-600 billion by 2030."

Low-forecast rationale examples:

> "I'm not at all confident that current rates of investment in AI will stay the same, let alone continue to grow. Given that the U.S. is headed toward a major recession, and given the amount of money that companies like OpenAI are burning on rapidly expanding AI infrastructure without a clear sense of what future AI use cases will actually look like, I think we are in fact reaching the end of what is clearly a bubble."

> "My forecast reflects a shift in structure more than a shift in enthusiasm. Some leading companies will raise less private capital as they mature, merge, or turn to other funding sources. Others will drop out of the "AI" category altogether as their work gets folded into broader infrastructure or enterprise tooling. They will still be vital but no longer counted as "AI." Additionally, some of the work that currently spins out as startups, especially around compliance, safety-testing, audits, and assurance, will likely move inward, absorbed by incumbents or coordinated through standards efforts. Meanwhile, policy environments may tighten, and potential infrastructure limits around energy and chip access could slow the pace of new deployments."

Generative AI Use Intensity

Question. *What percent of work hours in the U.S. in 2025, 2027 and 2030 will be estimated as assisted by generative AI?*

Background. A June 2025 study by the Federal Reserve Bank of St. Louis,[292] based on the Real-Time Population Survey (N=3,216), estimated that 1.3–5.4% of all U.S. work hours were assisted by generative AI based on self-reports via a nationally representative survey.[293] This question tracks how much self-reported generative AI adoption intensifies in work settings, as measured by updated or similar surveys.

Historical baseline. We estimate that the fraction of work hours assisted by generative AI in the U.S. as of September 2024 is **2.0%**.[294] This is the midpoint of the estimated range in the above study.

[292] *Link provided to participants:* https://fedinprint.org/item/fedlwp/98805/101172
[293] These figures were later revised (Bick et al. 2025).
[294] This estimate is based off of an earlier version of the paper, hence the discrepancy.

For full question background and resolution details, see Appendix E.III. 2. Generative AI Use Intensity.

Figures

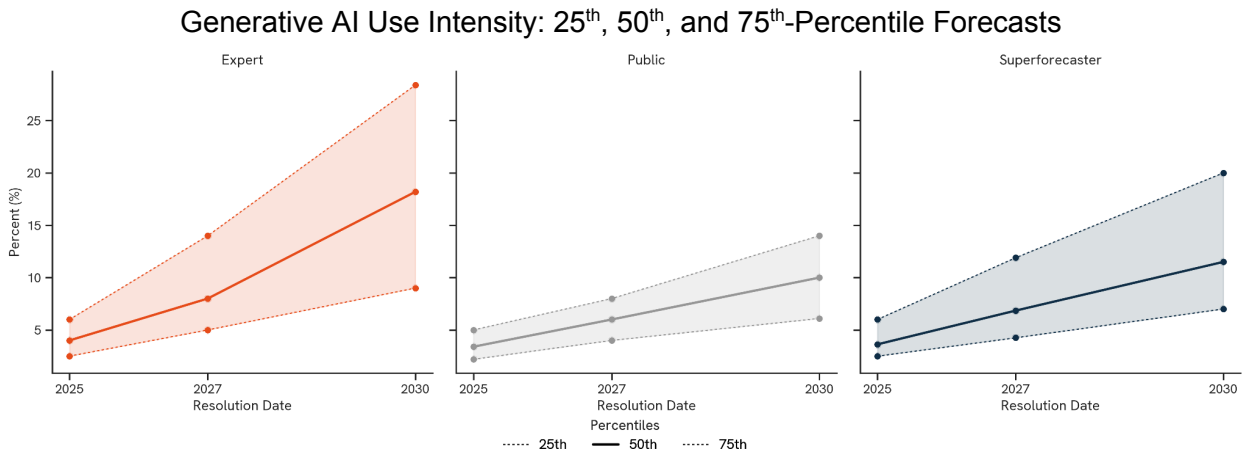


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.

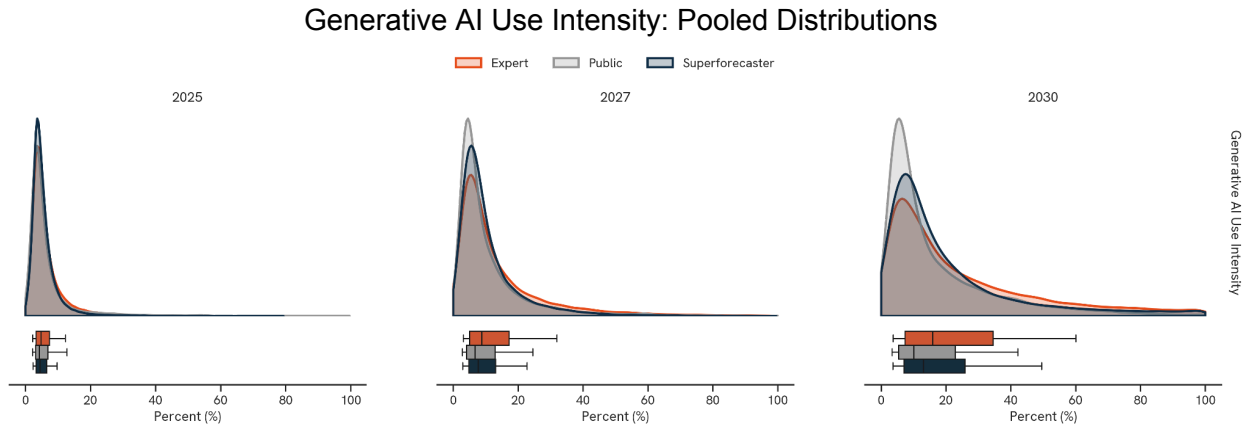


Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. For more details on how we calculate pooled distributions, see Appendix C. Pooled Distribution Estimation.

Results. Experts predict that 4% of 2025 work hours in the U.S. will be assisted by generative AI,[295] 8% in 2027,[296] and 18% in 2030.[297] That's 9 times our estimate for Sep 2024, 2%. The top 25% of experts believe that more than 30% of work hours will be assisted by generative AI in

[295] *Raw data*: IQR on the 50th percentile was (3.0%–5.1%); median 25th and 75th percentile forecasts were 2.5% and 6.0% respectively.
[296] *Raw data*: IQR on the 50th percentile was (5.0%–15.0%); median 25th and 75th percentile forecasts were 5.0% and 14.0% respectively.
[297] *Raw data*: IQR on the 50th percentile was (9.0%–30.0%); median 25th and 75th percentile forecasts were 9.0% and 28.4% respectively.

2030 and the top decile of experts believe that more than 40% of work hours will be assisted by generative AI. Experts and superforecasts predict similarly, whereas the public predicts much less progress in the medium-term: 10% by 2030,[298] almost half that of experts.

For full results tables, see here.

Rationale analysis:

- **Pace of enterprise adoption**: Many high-forecast respondents emphasize that fast distribution channels—through existing hardware like PCs and phones, and familiar software from Microsoft, Google, and Adobe—will accelerate enterprise adoption: "Tools are more and more integrated into everyday software products...it's shipping by default in editors, docs, email, calendars, crm [customer relationship management software], helpdesk [internal support software]... you don't 'go to AI', your software already has AI." Many high-forecast respondents also believe that competitive pressure will necessitate quick adoption: "Companies that don't adapt will go out of business." Low-forecast respondents frequently point to "institutional and bureaucratic barriers," and argue that "these changes are mediated by human organizations changing how they work. Most organizations are not that fast."
- **Use cases**: High-forecast respondents often highlight established use cases for routine writing, computational tasks, coding, report preparation, bookkeeping, graphic design, tax preparation, and legal briefing. They argue that "the primary long-term driver will be the shift from sporadic, task-specific use to continuous, deeply integrated AI assistance within core software platforms." Low-forecast respondents tend to acknowledge the potential but stress that there are "many jobs that simply do not need the use of GenAI, [for example] clerks, bartenders, plumbers, etc.," and that this creates a natural ceiling on adoption rates.
- **Impact on labor market**: Forecasters are divided on whether AI will predominantly assist or replace human workers. Most think an expansion of AI assistance is likely, but others argue that in many cases, "AI would eliminate rather than assist with jobs," and that when AI did assist, it would do so quickly, freeing up the human to do non-AI work—in which case the measured use of AI could flatten or decline, even as the impact of AI use was rising.
- **Capability requirements:** High-forecast respondents tend to believe current generative AI capabilities are sufficient for substantial workplace assistance: "I don't think AI tech improvement is needed here…This is just about how quickly tech can get integrated into organizations." Many pessimists disagree. One notes, "I'm bullish on AI as a whole but less persuaded by the current models."
- **Generational changes**: Several high-forecast respondents highlight that by the end of 2030, "more people who frequently used generative AI in their academic career will have entered the workforce and these younger people are often the ones spending a lot of

[298] *Raw data*: IQR on the 50th percentile was (5.0%–23.0%); median 25th and 75th percentile forecasts were 6.1% and 14.0% respectively.

time on repetitive deliverables." They see this demographic shift as a key driver, while pessimists tend to emphasize that the vast majority of laborers entered the workforce prior to the advent of generative AI and are still "used to doing their normal functions without AI systems."

- **Pace of blue-collar adoption**: A notable split exists on whether AI will penetrate manual labor roles. Low-forecast respondents frequently emphasize that manual labor is a key bottleneck given that substantial portions of the workforce perform tasks, "inherently unsuitable for generative AI assistance." But several high-forecast respondents push back on that notion, arguing that "AI can be integrated into practically any form of decision-making," for example, "construction workers can use AI assistants for safety checks, logistics planning, or instructional support. Retail and service workers may rely on AI-powered scheduling, training, or customer-facing chat systems. Thus, penetration will broaden across the economy, not remain confined to office work."

High-forecast rationale examples:

> "Generative AI assistance in work is likely to accelerate rapidly through 2030 as integration with productivity tools (Microsoft 365 Copilot, Google Workspace, etc.) becomes the default across knowledge industries."

> **"**Eloundou et al. (2023) found that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, with between 47 and 56% of all tasks potentially completable faster when incorporating LLM-powered software. This massive exposure gap between current usage (2% of work hours) and potential coverage (nearly half of all tasks) suggests explosive growth ahead as competitive pressures force adoption."

Low-forecast rationale examples:

> "While adoption rates among knowledge workers may continue to grow rapidly, a substantial portion of the U.S. workforce performs tasks that are inherently unsuitable for generative AI assistance in its current or foreseeable form. Manual labor sectors including construction, manufacturing assembly, and maintenance require physical presence and manipulation of objects. Direct service work such as food service, retail, healthcare aides, and childcare depends on human interaction and physical tasks. Transportation and logistics, skilled trades like plumbing and electrical work, agriculture, and direct patient care in healthcare all involve work that generative AI cannot meaningfully assist."

> "This is really hard because the number of work hours is itself dependent on generative AI. Suppose that instead of writing a 5-hour brief, AI can do it for me in 30 mins. Is that 30 mins of ai-assisted work or 5 hours? And what about entire jobs that are eliminated -- if we no longer have accountants running reports because AI automates it away, is that 0 work hours?"

Open vs Proprietary Polarity

Question. *What will be the mean benchmark performance of the best closed-weight AI models and the top open-weight AI models on the following benchmarks by 2025, 2027, and 2030?*

Background. The set of benchmarks is**:**

- **Mathematical reasoning:** FrontierMath (Tier 1-3 and Tier 4 separately).[299]
- **Software engineering:** SWE-Bench Verified.[300]
- **General reasoning and abstraction:** ARC-AGI-2.[301]

Historical baseline. As of September 2025, we find that the current mean performances of the top open-weight LLM and the top closed-weight LLM are **17.2%** and **30.0%** respectively**.**

For full question background and resolution details, see Appendix E.III. 4. Open vs Proprietary Polarity.

Figures

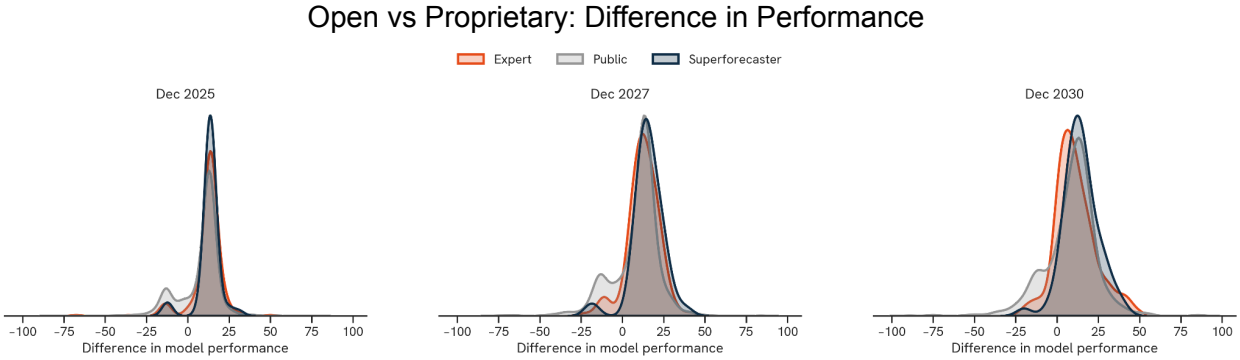


Figure: Each participant estimates the mean benchmark performance of the best closed-weight AI models and the top open-weight AI models by different dates. This figure shows the difference (closed performance minus open performance), with the y axis representing smoothed density of predicted difference across participant groups.

Results. Experts and superforecasts predict a ~15% performance gap between open weight and proprietary models across all time horizons.[302] Superforecasters predict similarly to experts, while the general public predict substantially less progress overall (48% vs 70% benchmark performance in 2030), as well as slightly smaller gaps in performance: ~10% in 2025 and 2027, and 8% by 2030—about half that of experts.[303]

---

[299] *Link provided to participants:* https://epoch.ai/data/ai-benchmarking-dashboard
[300] *Link provided to participants:* https://www.swebench.com/
[301] *Link provided to participants:* https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
[302] Expert medians: 35% vs 20% by 2025, 50% vs 37% by 2027, 70% vs 53% by 2030.
[303] Public medians: 31% vs 20% by 2025, 40% vs 31% by 2027, 48% vs 40% by 2030.

For full results tables, see here.

Rationale analysis:

- **Capital expenditure advantage:** Most high-polarity respondents emphasize that jumps in model capabilities increasingly require massive amounts of capital to pay for the high cost of compute, proprietary data, the best tool and retrieval stacks, and top-tier research teams. As a result, one forecaster concludes, “We should expect closed-weight models with heavy capital expenditure to perform particularly well at leading-edge tasks.” Low-polarity respondents often argue that capital-intensive techniques developed by closed labs will be adopted by open-source models. “We've seen this in the past,” notes one forecaster, adding “everything from telegraphy to computers (ENIAC onwards) to the internet, in which what starts as a more elite and constrained technology migrates widely into open knowledge.” Other low-polarity respondents believe that in lieu of massive capex, “aggressive community efforts” that result in rapid iteration, distributed innovation, and collaborative scaling will help open-source models close the gap, and that “open-weight progress will be [also] accelerated by expanded access to high-quality synthetic data.”
- **Trend extrapolation:** High-polarity respondents often point to historical trends that show a persistent gap between closed and open model performance: “Open-model performance has tended to lag closed-model performance by anywhere between about 6 and 22 months,” wrote one. Others estimated similar lags. Low-polarity respondents challenge this pattern, pointing to recent convergence: “According to Stanford's 2025 AI Index Report, the gap between closed and open models narrowed from 8.04% to 1.70% in just one year [on the Chatbot Arena Leaderboard]” (Stanford University Human Centered Artificial Intelligence 2025).
- **Benchmark saturation:** Many low-polarity respondents believe convergence could result from the four specific benchmarks becoming saturated prior to the end of 2030. SWE-bench was thought to be particularly susceptible: “I find it likely that the current version of SWE-Bench is fully saturated by 2027 by both closed and open models.” Many also thought the other benchmarks were also likely to saturate: “Eventually, benchmarks saturate (e.g., see what happened to GPQA),” wrote one forecaster, with another arguing “there is very little chance a benchmark with >1% score today is not completely saturated in 2030.” High-polarity respondents typically either do not view convergence due to benchmark saturation as likely, or focus less on the specific benchmarks than on polarity in general.
- **Architectural breakthroughs:** Most high-polarity respondents believe proprietary labs have advantages in breakthrough development. One notes that “a major architectural breakthrough (e.g., a successor to the Transformer), kept proprietary, could dramatically widen the gap.” As with the capex gap, low-polarity respondents tend to view rapid iteration, distributed innovation, and collaborative scaling as ways for open-source models to remain competitive. One points to the “proprietary and non-proprietary development of the human genome project” as an apt historical parallel.

- **Benchmark focus:** Some high-polarity respondents believe that FrontierMath and ARC-AGI II, lacking clear practical utility, will be more of a focus of closed model developers: “Math to impress the general public, ARC-AGI for the specialists and the VCs.” Several low-polarity respondents, however, come to the opposite conclusion: “FrontierMath and ARC-AGI II don't have obvious economic utility, so they may become more of a focus of academics, etc. using open source models.”

High-forecast rationale examples:

> “Closed models will benefit from continued massive compute, fresh proprietary data, and production-grade tool and retrieval stacks that boost performance on math, coding, and abstraction tasks.”
>
> “I also expect an increasing gap between closed-weight and open-weight models as compute expenditures heavily benefit the former. We should expect closed-weight models with heavy capital expenditure to perform particularly well at leading-edge tasks, while the performance gap on more mundane tasks will probably narrow.”
>
> “I am just going to do a (linear) projection of current trends for the closed-weight score on each benchmark, and then assume that the best Open Weight score remains a similar distance behind as in the baseline.”

Low-forecast rationale examples:

> **“**I'm assuming steady gains on SWE-bench verified and FrontierMath T1-3, slower progress on FrontierMath T4 and ARC-AGI-2, and a gradual narrowing of the open/closed gap as open weight ecosystems absorb techniques a bit later but at scale -- thinking strong distillation, community eval cycles, hardware access. Closed models would likely retain a modest edge from proprietary data and eval driven tuning and tooling.”
>
> “I expect convergence in performance of open and closed weight models on raw benchmarks like these. Closed weight models will likely outperform in many other ways not measured by the benchmarks though.”
>
> “Epoch's recent article: https://epoch.ai/blog/what-will-ai-look-like-in-2030 basically looks at this question. For SWE and math benchmarks it looks like it takes a bit over 2 years to go from ~10% to ~90%.”

AI Companions

Question. *What proportion of U.S. adults will self-report using AI for companionship at least once daily by 2027, 2030, and 2040?*

Historical Baseline. A July 2025 poll (AP-NORC Center for Public Affairs Research 2025) found that **6%** of U.S. adults reported using AI for companionship at least once a day.

For full question background and resolution details, see Appendix E.III. 5. AI Companions.

Figures

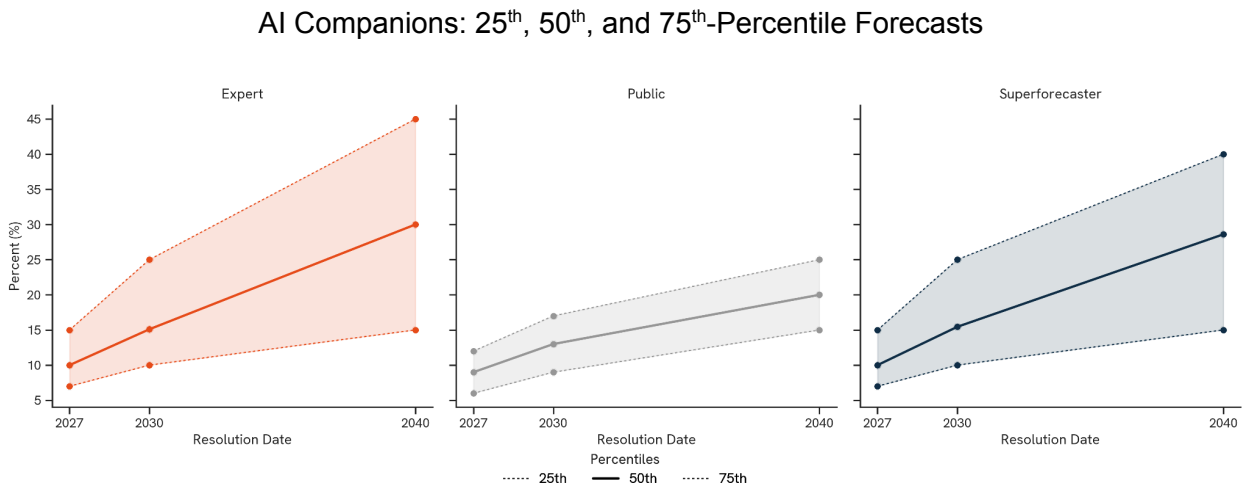


Figure: Participants gave 25th, 50th, and 75th percentile forecasts for each resolution date. This figure shows the median response for each of these forecasts, split by participant group.

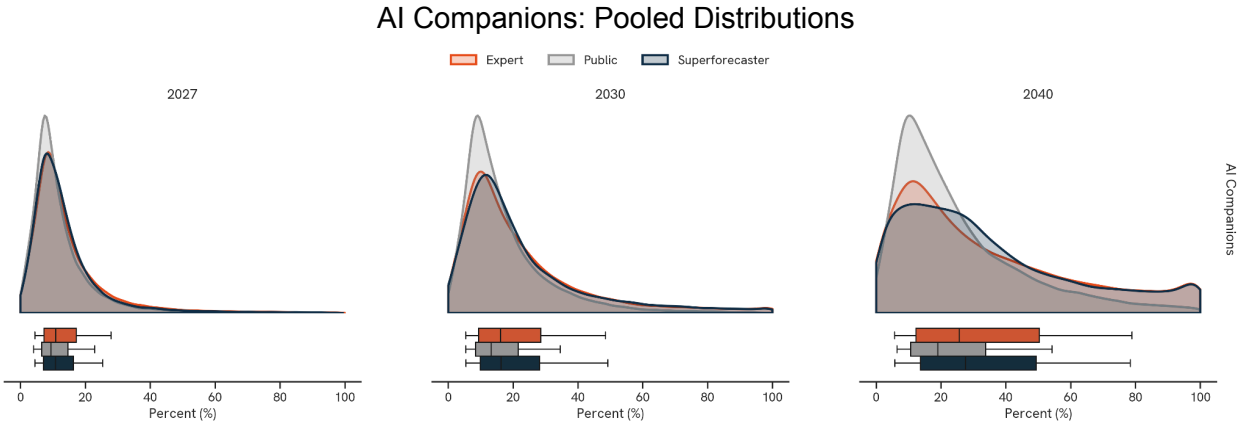


Figure: This figure shows pooled distributions split by participant category. These pooled distributions combine within-expert uncertainty and between-expert disagreement. Densities are normalized to the same peak for comparability. For more details on how we calculate pooled distributions, see Appendix C. Pooled Distribution Estimation.

Results. Experts predict that 10% of U.S. adults will self-report using AI for companionship at least once per day in 2027,[304] up from 6% in July 2025. That number increases to 15% of adults in 2030,[305] and 30% in 2040.[306] By 2040, the bottom quartile of experts believe that less than 16% of people will self-report using AI for companionship daily whereas the top quartile believe that more this figure will be more than 40%;[307] the top decile of experts believe that more than 60% of people will use AI for companionship daily. While experts and superforecasters predict similarly, the general public predicts substantially less adoption by 2040: 20% vs experts' 30%.[308]

For full results tables, see here.

Rationale analysis:

- **Loneliness**: Many high-forecast respondents project a substantial expansion of use, in large part due to increasing loneliness. One notes that "the U.S. Surgeon General declar[ed] loneliness an epidemic in 2023, with about half of U.S. adults experiencing measurable levels of loneliness." A few low-forecast respondents emphasize lower

[304] *Raw data*: IQR on the 50th percentile was (8.0%–15.0%); median 25th and 75th percentile forecasts were 7.0% and 15.0% respectively.
[305] *Raw data*: IQR on the 50th percentile was (10.9%–25.0%); median 25th and 75th percentile forecasts were 10.0% and 25.0% respectively.
[306] *Raw data*: IQR on the 50th percentile was (16.0%–40.0%); median 25th and 75th percentile forecasts were 15.0% and 45.0% respectively.
[307] *Raw data:* IQR on the 50th percentile was (16.0%–40.0%); median 25th and 75th percentile forecasts were 15.0% and 45.0% respectively.
[308] *Raw data*: IQR on the 50th percentile was (11.0%–32.3%); median 25th and 75th percentile forecasts were 15.0% and 25.0% respectively.

saturation limits. One writes, “About a quarter of U.S. adults go to therapy. If that's the market size, then I expect AI to eventually saturate [at] that.”[309]

- **Capabilities**: Many high-forecast respondents believe AI capabilities are likely to improve dramatically, making companions more “sophisticated, emotionally intelligent, and capable of forming deeper connections with users,” and that “personalized AI companions that learn and adapt to individual users' preferences and needs will become more common.” Low-forecast respondents tend to argue that AI is unlikely to be able to replicate genuine human connection: “Nothing can replace real human interactions, and the vast majority of the population will be reluctant to have emotional interactions with AI.” One forecaster argues that “most people would find such companionship unfulfilling, perhaps even viewing reliance on it as a kind of failure.”
- **Technological integration**: High-forecast respondents frequently cite integration with smartphone apps, wearable devices, and social media platforms as a likely driver of widespread use, even among those not searching for companionship. As one forecaster observed, “I have Grok in my car, Alexa in my kitchen, and Meta in my glasses and it's 2025. I do not use them for companionship, but I do miss them when not available.” Another emphasized that “ambient access through devices turns companionship into a series of micro-interactions throughout the day.” Low-forecast respondents often question whether the current technology can provide genuine companionship, with one arguing “it will not be a substitute for human interaction unless or until AIs achieve human levels of autonomy.”
- **Generational adoption patterns**: Both sides acknowledge strong generational effects, but interpret implications differently. High-forecast respondents tend to see use among young adults (25% of 18-29 year-olds have tried AI companionship) as indicative of future mainstream adoption. One notes, “Young teens have much higher usage patterns for companionship than adults, hence giving some early indication of usage by adults in future years.” Another thinks the proportion “will increase rapidly as children grow to adulthood accepting these conversations as natural, having known interactive systems like Siri and Alexa for all of their conscious lives.” Some low-forecast respondents acknowledge youth adoption but emphasize resistance from older demographics, anticipating “generational resistance from many older adults” and that the “novelty will wear off.”
- **Regulatory and social acceptance**: “Government regulation is unlikely,” is a common view among high-forecast respondents. Relatedly, they anticipate declining social stigma as AI companionship becomes normalized through integration with existing platforms. Several low-forecast respondents emphasize the potential for “persistent social/cultural resistance” that could lead to regulatory backlash. One notes, “Recent cases of suicides, possibly caused by AI companionship…may reduce trust and adoption. I assume that there will be more regulation limiting how people can/should use AI companion tools.”

High-forecast rationale examples:

[309] This claim may refer to the ~23% of U.S. adults who, according to a 2024 KFF (formerly Kaiser Family Foundation) study, “say they received mental health counseling and/or prescription medication for mental health concerns in the last year.” See Panchal and Lo (2024).

> "Facebook growth and penetration of the United States adult population may be the best analogy for the potential scale and trajectory of Companion-AI. About 50% of the adult population used Facebook daily as of 2021 per Pew Research. Adult Facebook general usage peaked at 72% in 2015 per Pew research via a Gemini query."
>
> "Per this set of statistics, over 60% of people experience loneliness as a chronic condition in the US. https://www.discoveryaba.com/statistics/loneliness. People spend huge amounts of money attempting to alleviate this."
>
> "AI companion experience will be fundamentally different: likely multi-sensory (holographic/VR), capable of long-term memory, and indistinguishable from human interaction for many. The proportion of daily users may approach the current daily usage rate of major social media platforms."

Low-forecast rationale examples:

> "Nothing can replace real human interactions, and the vast majority of the population will be reluctant to have emotional interactions with AI."
>
> "Alignment constraints, regulation, and social stigma will suppress widespread adoption in conservative or regulated contexts. Ethical debates over dependency, consent, and simulated intimacy are likely to trigger restrictions in schools, workplaces, and certain jurisdictions."

Barriers to Adoption, Part II

Question. *By the end of 2030, what percent of LEAP expert panelists will say that each of the following (9) factors has significantly slowed AI adoption relative to popular expectations?*

- **Lack of reliability**: Hallucinations, unpredictable behavior, a lack of interpretability, results skewed by biased training data, and other reliability issues limit their usefulness.
- **Cultural resistance**: Fear of job losses and rising inequality, along with a common preference for humans over AI, will curtail use.
- **Restrictive regulations**: Ambiguous, fragmented, and evolving new regulations regarding AI, and the fact that the preponderance of existing regulations were written before the emergence of AI, will slow deployment.
- **Cost issues**: High infrastructure and energy costs will render some large AI projects impractical.
- **Data quality issues**: High-quality, unbiased data that is useful to specialized industries is limited by privacy constraints, copyright issues, and the high cost of data labeling.
- **Integration challenges**: At a corporate level, AI adoption often requires a complex redesign of workflows, IT systems, and the physical environment—and many of the people tasked with executing the redesign may lack motivation due to fear of replacement.

- **Not enough use cases**: Until trust issues are resolved, and progress in robotics advances, the number of ways in which AI can create value in the real world will be limited and adoption will be delayed.
- **Lack of AI literacy**: There aren't enough people with the necessary skills to develop, implement, and adopt new AI systems quickly, at scale, in diverse organizations around the world.
- **Social-cultural anomie**: Societal intellectual atrophy and damage to the environment due to use of AI, paired with the mental health consequences of AI replacing much of human companionship, will lead to a backlash that slows adoption.

For full question background and resolution details, see Appendix E.III. 6. Barriers to Adoption, Part II.

Figures

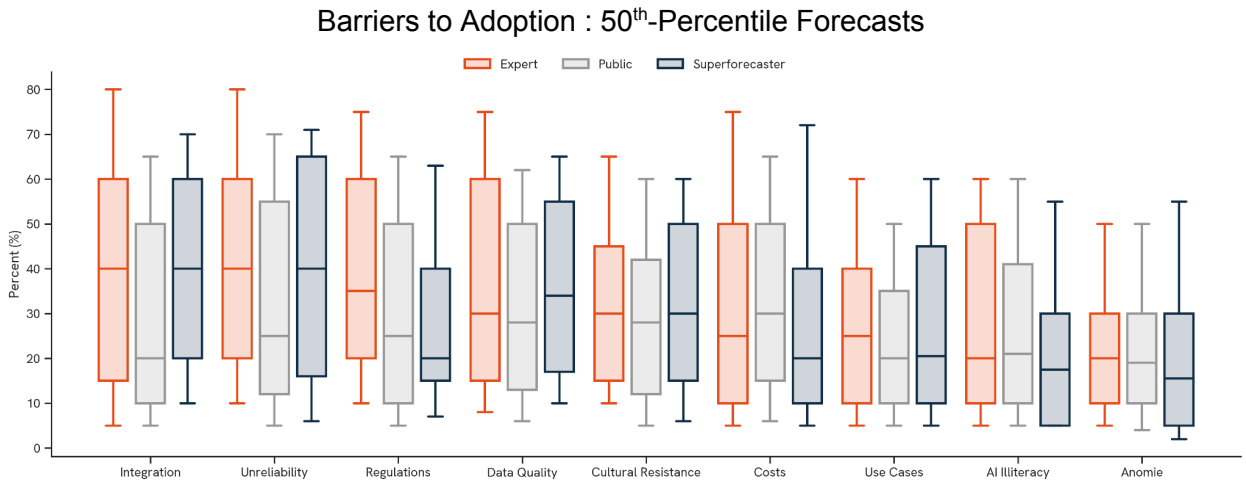


Figure: In this question, participants make 50th percentile forecasts for resolution in 2030. This figure shows the 10th, 25th, 50th, 75th, and 90th percentiles of these 50th percentile forecasts, split by participant group and by potential bottleneck. The 25th expert percentile represents the number that 25% of experts' median forecasts are lower than.

Results. Experts predict that AI literacy,[310] social-cultural anomie,[311] uses cases,[312] and costs[313] are not likely to be judged to significantly slow AI adoption in 2030 relative to popular expectations (medians 20%–25%). Data quality,[314] regulations,[315] and cultural resistance[316] are judged slightly more serious (medians 30%–35%), and integration[317] and unreliability[318] are

[310] *Raw data*: IQR on the 50th percentile was (10.0%–50.0%)
[311] *Raw data*: IQR on the 50th percentile was (10.0%–30.0%)
[312] *Raw data*: IQR on the 50th percentile was (10.0%–40.0%)
[313] *Raw data*: IQR on the 50th percentile was (10.0%–50.0%)
[314] *Raw data*: IQR on the 50th percentile was (15.0%–60.0%)
[315] *Raw data*: IQR on the 50th percentile was (20.0%–60.0%)
[316] *Raw data*: IQR on the 50th percentile was (15.0%–45.0%)
[317] *Raw data*: IQR on the 50th percentile was (15.0%–60.0%)
[318] *Raw data*: IQR on the 50th percentile was (20.0%–60.0%)

judged the most likely barriers to adoption, at 40%. There's no consensus that any particular barrier to adoption will turn out to be significant. Experts and superforecasters predict similarly. The general public also predicts similarly, except for integration (20%[319] vs experts' 40%) and unreliability (25%[320] vs experts' 40%).

For full results tables, see here.

Rationale analysis:

- **General:** Forecasters widely disagree about which AI adoption barriers will persist or diminish by 2030. Low-forecast respondents tend to believe barriers will fade as capabilities improve, with one forecaster pointing to "the rapid and continuing surge of AI adoption after ChatGPT [as] a sign that improving capabilities can wash away a lot of these barriers." High-forecast respondents largely emphasize structural challenges, arguing that "many of the issues "have no known perfect solutions" and that the pace of model deployment will likely "outstrip governance, integration capacity, and human skill adaptation." One concludes "The next 5 years are likely to bring a reality check to the current hyped expectations." A core divide is whether solutions exist: some view barriers as engineering problems while others see fundamental limitations requiring "entirely new architectures."
- **Lack of reliability:** Most low-forecast respondents believe reliability issues are rapidly improving through technical advances. They note that "hallucination has already dropped a lot with GPT 5 and it will keep going" and expect that "productized retrieval, tool use, and verifiers [will] cut hallucinations enough that only a minority will still call them serious." Some also believe that "by 2030, the AI community, including users, will have a better understanding of AI limitations, and AI applications will be designed to provide a known level of accuracy that is appropriate for that application." High-forecast respondents frequently emphasize persistent fundamental limitations: "Even with retrieval-augmented generation and tool use, hallucinations will remain a widely recognized constraint especially in high-stakes settings like medicine and law." One notes that "research shows these issues have a statistical lower bound making them an inherent limitation rather than an occasional glitch." Some high-forecast respondents also point to a creative-accuracy tradeoff: "For AI to be creative, it needs to hallucinate."
- **Cultural resistance:** Low-forecast respondents tend to think economic incentives will overcome resistance as "soft social factors do very little if there is a good economic case for use" and expect resistance to fade with demonstrated utility: "If the product is good/addictive enough, then cultural resistance will be muted." Some high-forecast respondents, however, focus on the potential for labor unrest to spark resistance: "I think [cultural resistance] will be very large as job losses start appearing" notes one, and another foresees "a societal push to reward human work and avoid AI produced content… like we have already seen in the world of publishing." Other forecasters

[319] *Raw data*: IQR on the 50th percentile was (10.0%–50.0%)
[320] *Raw data*: IQR on the 50th percentile was (12.0%–55.0%)

suspect that the “breakdown of trust and norms caused by rapid AI shifts will fuel confusion and resistance.”

- **Restrictive regulations:** Many low-forecast respondents believe competitive pressures will limit restrictive regulations. They note “regulation doesn't seem to be a priority now for the world, quite the opposite” and that “the competition is fierce…It seems both the U.S. and China are set to facilitate their companies to freely compete.” One argues, “Judging by how tech illiterate elected politicians are, they will be too late to put meaningful regulations in place that would hinder its adoption.” But many low-forecast respondents do expect significant regulatory barriers, especially internationally. One predicts “regulation might not be a huge problem in the U.S. but looking back in a few years, I believe most panelists would agree that the Europeans will have missed another technological change because of regulations.” Another argues that “heterogeneous, evolving rules across jurisdictions and sectoral compliance burdens will create real deployment frictions.”
- **Cost issues:** Low-forecast respondents tend to suspect “the costs of deploying AI are going to continue to fall, either from the investments in energy supply, more efficient training and inference, or the reorientation towards smaller models.” One points to the potential for small modular [nuclear] reactors to change the calculus. High-forecast respondents often stress “the current rate of capex is unsustainable” and emphasize that “spiraling infrastructure and energy demands make large-scale deployment economically inefficient outside of the biggest firms.”
- **Data quality issues:** Some low-forecast respondents think “data quality issues can be overcome with clever synthetic data generation” and note there is “ample…investment and [that] startups are working on improving data quality.” Several high-forecast respondents argue that “access to clean, unbiased, legally usable, and domain-specific datasets remains the single greatest obstacle to scalable adoption” and that this is especially the case in key fields like healthcare, finance and defense. They also worry about data degradation: “Without proper data management AI models will start cannibalising their own slop.”
- **Integration challenges:** Most low-forecast respondents believe competitive pressures will drive successful integration: “If future profits/relevance are at stake, corporations will make the investments to keep up--no one wants to be the next Kodak.” Many also expect “tooling, APIs, and education will improve enough that most organizations can technically adopt AI once legal and data issues are resolved” and that eventually AI may be able to assist in the integration process so that it occurs “almost seamlessly.” High-forecast respondents commonly emphasize the complexity of organizational change. One observes that “changing existing business processes is generally hard (ERP projects have a notoriously high failure rate, for example) and I struggle to think of an organization that has successfully moved to predominantly using AI.” Another that “large companies have enormous inertia, and integrating AI at a corporate level requires deep workflow and process redesign.”
- **Not enough use cases:** Many low-forecast respondents expect rapid expansion of viable applications and note that “the amount of use cases is already large and growing.” High-forecast respondents tend to point to ROI concerns and argue the issue is “a

shortage of economically viable and trusted applications that significantly outperform existing methods." Many also note constraints from robotics limitations: "With progress in robotics being slower than for intellectual work…I could see this [barrier] being somewhat significant."

- **Lack of AI literacy:** Low-forecast respondents typically expect to see rapid skill development materialize, driven by necessity and user-friendly, natural-language interfaces. They note "AI illiteracy [is] already fading (look at those adoption patterns)" and that "already it is easily accessed through search engines" and that "organisations are working hard to upskill staff." Some high-forecast respondents, however, stress that "workforce training takes time" and highlight the likelihood of persistent skill gaps. One argues that "skills shortages are going to become worse due to i) demographic change, ii) restrictive immigration policy, iii) cuts in research and science funding."
- **Social-cultural anomie:** Most low-forecast respondents believe economic benefits will ultimately override concerns. They argue that "anomie and cultural resistance are usually not long-lived during technological shifts." One notes, "I cannot think of any technological development that has been hindered by it," and another, "TV and cars and many automations made humans lazy but that didn't stop the industry." Some high-forecast respondents highlight the potential for growing environmental and social concerns to create resistance. One predicts: "Anomie will be a huge issue as people realize just how much electricity is needed for AI and the environmental cost of that."

Rationale examples:

General: "The rapid and continuing surge of AI adoption after ChatGPT is a sign that improving capabilities can wash away a lot of these barriers. If powerful general-purpose AI systems are achieved, they will not require too much hand-holding. Companies that decline to adopt AI will be outcompeted by those that do, and private equity-style "rollup" acquisitions that force companies to adopt AI may become common."

AI literacy: "This has always been the case with each new leap in information technology, yet it has always been overcome. The same thing will happen with AI use, which if anything has the potential of democratizing information technology much more than any other advancement in the past. I mean, we can interact with AI by means of natural language: what could be easier than this?"

Data quality: "Access to clean, unbiased, legally usable, and domain-specific datasets remains the single greatest obstacle to scalable adoption. Despite larger models and better architectures, progress will continue to be throttled by privacy constraints, copyright uncertainty, and the scarcity of trustworthy labeled data for specialized sectors such as healthcare, defense, and finance."

Regulations: "I don't think the regulatory regime has been all that hostile to AI, but some panelists would probably disagree if they were being surveyed today, and it is possible that AI regulations (or AI regulation-adjacent problems, like it becoming impossible to build new data centers or power generation in large parts of the country) will increase."