Revealing Life Preferences Through LLMs

Omar Abdel Haq Amitabh Chandra Tomáš Jagelka
Erzo F.P. Luttmer Joshua Schwartzstein *

April 30, 2026

Abstract

Large Language Models (LLMs) are trained on a prodigious corpus of human writing and may reveal human preferences over characteristics of life courses, such as income, longevity, and working conditions. We present OpenAI's *GPT-5.4* and a broadly representative sample of Americans with pairs of life stories and ask them to choose the life they would prefer for themselves. A person's choice is better predicted by the LLM's choice than by another person's choice over the same stories, and LLM valuations of several life attributes are similar to those derived from human responses. Our results suggest that LLM responses offer a scalable and cost-effective complement to existing methods for studying human preferences.

*Abdel Haq: Harvard Business School (email: oabdelhaq@hbs.edu); Chandra: Harvard Business School and Harvard Kennedy School (amitabh_chandra@harvard.edu); Jagelka: University of Bonn, Dartmouth College, and CREST-Ensae (tjagelka@uni-bonn.de); Luttmer: Dartmouth College (email: erzo.fp.luttmer@dartmouth.edu); Schwartzstein: Harvard Business School (email: jschwartzstein@hbs.edu).
For helpful comments, we thank Ben Bushong, Katherine Coffman, Sendhil Mullainathan, Paul Novosad, Ziad Obermeyer, Suproteem Sarkar, Andrei Shleifer, Adi Sunderam, Dmitry Taubinsky, and Hans-Joachim Voth, and participants at various seminars and conferences. We thank Isabel Galea, Alex Philip, and Julia Schwed for outstanding research assistance. This work is supported by Harvard Business School; the European Research Council (ERC) under the FELICITAS grant (No. 101165518); and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/1 – 390838866. Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

1 Introduction

Understanding how people value the fundamental attributes of their lives, such as longevity, health, and working conditions, has long posed significant challenges across economics, psychology, and public policy. These *life preferences* shape consequential decisions about health, relationships, and careers. They determine whether rising health insurance premiums are worth their cost, how to conduct the welfare analysis of speed limits and environmental standards, and how to translate differences in life expectancy across income groups into equivalent differences in living standards. Despite their importance, life preferences are difficult to estimate. First, stated and revealed preferences often diverge: people may report that relationships matter more than income but nevertheless choose higher-paying jobs over more fulfilling ones, suggesting that the reliability of either method cannot be taken for granted when recovering underlying preferences. Second, we generally lack exogenous, well-measured variation in the prices of life attributes, making it difficult to recover life preferences from observed behavior. Hazardous jobs, for example, offer higher pay, but other unobserved differences between jobs make it hard to attribute wage premia specifically to mortality risk. Third, intention-action gaps caused by limited self-control or inattention, such as failing to wear a seatbelt while believing that life is priceless, create further wedges between observed choices and underlying preferences. Together, these challenges mean that revealed-preference approaches are not automatically superior to stated-preference ones: inferring longevity valuations from wage differentials in dangerous occupations, for instance, relies on populations with non-representative preferences and requires people to evaluate small changes in mortality risk, a task that behavioral evidence suggests humans do poorly. More broadly, they indicate the need for new methods to elicit life preferences.

This paper asks whether a large language model (LLM) can recover previously unknown human life preferences in a way that complements existing approaches to preference elicitation. We gave OpenAI's *GPT-5.4* the identical choice task we gave a broadly representative sample of Americans: choosing between pairs of completed life stories that varied in income, longevity, and other life attributes. These life stories were designed to measure life preferences directly, eliciting preferences over entire life courses without requiring probability judgments. The human responses yield a previously undocumented distribution of income-longevity tradeoffs across the U.S. population (Jagelka ⓡ al., 2026). We conducted the LLM elicitation before those results were published: an LLM trained on existing research might reproduce known facts about human preferences by pattern-matching against published estimates, but it cannot recover estimates that did not yet exist.

We therefore ask whether LLMs encode something about what humans value that goes beyond what has been explicitly documented and published.

There are reasons to be optimistic. LLMs are trained on an extensive corpus of human writing, including literature, journalism, personal essays, and online discussion, and fine-tuned with human feedback. This corpus is rich with implicit information about what people value: which lives sound appealing, which tradeoffs feel fair, which circumstances seem enviable or pitiable. Insofar as human writing encodes these judgments, LLM training may allow models to extract preference-relevant structure even in domains where no explicit estimates exist (Horton et al., 2026; Asirvatham et al., 2026).

The competing view is that LLMs are "stochastic parrots," models that mimic the statistical regularities of text without encoding anything about preferences (Bender et al., 2021). On this view, an LLM asked to choose between two life stories is doing something like identifying which story sounds more like a positive description of a life, without any representation of the tradeoffs humans make when they think carefully about what they value. The concern is not that LLMs will fail trivially. Any model that understands language will agree that winning the lottery beats being assaulted. The concern is that LLM responses may not reflect the quantitative structure of preferences: how much income people trade for longevity, how much they discount hardship at work, or how they value the attribute bundles of relationships, career, and community embedded in completed life stories. These are the objects our survey was designed to measure, and they are precisely the objects we ask LLMs to recover.

We propose a *revelation conjecture*: that an LLM, when appropriately prompted, can reveal human preferences over consequential life choices in a statistically informative manner. This conjecture makes no claim about whether LLMs understand human preferences in any deep sense. It asks only whether LLM choices, when prompted to act as humans, yield preference estimates that are statistically related to those derived from human responses. The conjecture generates hypotheses of varying stringency: weak tests ask whether LLMs predict the direction of majority preferences, while strong tests ask whether they recover dollar-valued tradeoffs. The revelation conjecture may hold for some populations, some preference domains, and some attributes but not others, so each hypothesis in this family is specific to a population, a set of scenarios, and an estimation method. Consequently, the strongest possible test, whether LLMs can reveal the preferences of specific individuals rather than population-level averages, lies beyond what we can establish.

We present three lines of evidence, ordered from least to most demanding. First, a person's choice between two life stories is better predicted by the LLM's choice than by

the choice of a different human respondent over the same two stories. This is a striking finding on its face, but it could partly reflect the LLM having lower noise than individual human respondents. We use a decomposition model to show that the result is more fundamental: after separately identifying human noise, preference heterogeneity, and LLM accuracy from four empirical moments, we estimate that the LLM correctly infers the majority human preference in roughly 86 to 92 percent of story pairs. Second, we estimate how much each of our 28 template life stories is worth as a bundle, expressed in income units, and find that human and LLM valuations of these attribute bundles are highly correlated, with slopes close to one. This extends the result from individual choices to quantitative valuations of rich story content: relationships, career trajectories, community involvement, and circumstances of death. Third, for the specific life attributes that we separately randomize within template stories, including income, longevity, hardship at work, good health, and era of birth, LLMs not only match the direction of human responses but also provide estimates of dollar-denominated tradeoffs that are close to those derived from human responses. For longevity and hardship at work, the two attributes estimated with sufficient precision to support meaningful comparisons, *GPT-5.4* with direct choice produces valuations within 50 percent of the human estimates, and we cannot reject equality. Taken together, the evidence supports the revelation conjecture: LLM output, when appropriately prompted, reflects not only which lives sound appealing but also the rates at which people trade one life attribute against another.

Understanding the quantitative nature of this result requires context. Recovering preferences over life attributes is not merely a matter of knowing that people prefer more income to less or longer lives to shorter ones. It requires knowing the marginal rate of substitution between them: how much income compensates for a year of life lost, which in turn requires forming implicit estimates of the relative marginal utilities of the two quantities. Economists identify this tradeoff from experimental variation in prices. The LLM was given no such variation during training, yet when exposed to it in our choice task, it responds in a way that implies a marginal rate of substitution close to the one humans reveal.

Our contribution connects several streams of research. A growing literature examines whether LLM responses can substitute for or complement human survey data (Argyle et al., 2023; Dominguez-Olmedo et al., 2024; Santurkar et al., 2023; Brand et al., 2025; Horton et al., 2026), with some work focused on whether LLMs replicate demographic differences in responses (Aher et al., 2023; Cheng et al., 2023). Ludwig et al. (forthcoming) caution that LLM-generated data should not substitute for human data without validation against

actual human responses, a requirement our design is built to satisfy. Our work differs from this literature in two ways. First, we focus on preferences over life attributes, including income, longevity, health, and working conditions, rather than political opinions or factual beliefs, where LLMs may have absorbed published survey results directly. Second, we validate against a novel survey whose results were not available online at the time of the LLM elicitation, making contamination from training data implausible. Some economists now use rich text data to "reveal the invisible" (Stantcheva, 2023); our work suggests that LLMs may accelerate this program, not as a substitute for human data, but as a complement that can extend preference elicitation to populations and domains where direct surveys are costly or infeasible.

The remainder of the paper proceeds as follows. Section 2 formalizes the revelation conjecture and describes the experimental design, including how we constructed the life stories, recruited and surveyed human respondents, and elicited choices from LLMs. Section 3 presents results in two parts: story-level correlations between human and LLM choices together with a decomposition model that separates noise, preference heterogeneity, and LLM accuracy; and valuations of specific life attributes that allow direct comparison of human and LLM preferences in dollar terms. Section 4 discusses implications and Section 5 concludes.

2 Methods

We elicit preferences by offering human survey respondents and LLMs a binary choice between two hypothetical lives, written as completed life stories. These stories were written by a research assistant by early November 2022 (before LLMs were widely available) and subsequently edited by us. Stories are typically three to five paragraphs long and give a bird's eye view of someone's life, conveying hobbies, work, friendships, family, and volunteer work. All stories include their protagonist's name, gender, job, household income, marital and parental status, age at death, cause of death, timing of any terminal health condition, and information from which the year of birth can be inferred. An example story is shown in Appendix Figure A1, and a sample choice is shown in Appendix Figure A2.

2.1 The Revelation Conjecture

What do we mean by the conjecture that LLM choices, when appropriately prompted, usefully reveal human preferences? This section formalizes the idea within an inference framework. The key question is whether preferences estimated from LLM responses provide valuable information about human preferences in direction and magnitude—not whether LLM responses mimic the distribution of human responses, nor whether LLMs possess any underlying "understanding" of human preferences.

Each person i is endowed with preferences P_i^* over life courses, which are vectors of life attributes. For simplicity, we assume these preferences satisfy standard assumptions, including menu-independence (abstracting from issues raised by, e.g., Kőszegi and Rabin (2008) and Bernheim et al. (forthcoming)). For any scenario S_v consisting of two life courses L_{v1} and L_{v2}, P_i^* determines whether the person prefers L_{v1} over L_{v2}.

When a person is asked to make choices between richly described life courses in scenario S_v, her stated choice R_{iv} is a stochastic function of the scenario given her true preferences. Consider a set of scenarios $\mathcal{V}$ that may vary some life attributes (e.g., longevity) while fixing others (e.g., country of birth), as well as a human (sub)population $\mathcal{X}$. Let R_{Human} denote the vector of choices made by representative members of $\mathcal{X}$ on randomly selected scenarios from $\mathcal{V}$. Using estimation method $g(\cdot)$, the analyst estimates preferences $\hat{P}_{Human} = g(R_{Human})$ for a representative member of $\mathcal{X}$.

Suppose the analyst then prompts an LLM to act as a member of population $\mathcal{X}$ and choose between life courses across the same scenarios, yielding choices R_{LLM}. Applying the same estimation method gives $\hat{P}_{LLM} = g(R_{LLM})$, the "preference" estimate that rationalizes the LLM's choices.

The *revelation conjecture* for population $\mathcal{X}$ on scenarios $\mathcal{V}$ under estimation method $g(\cdot)$ is that knowledge of the LLM estimate provides statistical information about the human estimate: The revelation conjecture is that, at a minimum, the two estimates are not statistically independent; if the LLM perfectly reveals the life preferences of population $\mathcal{X}$, the two coincide in a statistical sense, i.e., $\hat{P}_{LLM} = \hat{P}_{Human}$, where equality is understood to hold up to sampling variation.

We next describe the experimental design used to test hypotheses derived from this conjecture for life preferences in a representative population of adult Americans.

2.2 Life Story Variation

We construct choice scenarios using a total of 28 template life stories, 14 with a female protagonist and 14 with a male protagonist. Stories are paired subject to three restrictions: both protagonists must be of the same sex, baseline income levels cannot differ by more than a factor of five, and no respondent sees the same story twice within a survey.[1]

We independently randomize five attributes within the template stories of a scenario. The protagonist's age at death is drawn uniformly between 60 and 90 years. Annual household income (rounded to the nearest thousand) is drawn uniformly between two-thirds and four-thirds of an occupation-specific baseline. The year of death is drawn uniformly from 2000 to 2019, with the year of birth determined jointly by longevity and the year of death.[2] Two story-specific elements — one describing workplace hardship such as stress, irregular hours, or career setbacks, and one stating that the protagonist was in good health prior to the onset of any terminal condition — are each included in a scenario with 25% probability, independently of one another; when included, each is assigned to one of the two paired stories at random.

Additional sentence-level randomizations at the story-pair level, including material and nonmaterial utility sentences and a filler sentence, are randomized orthogonally to the five attributes above and are not directly relevant to this study. Full details on all randomizations, examples of each sentence type, and an overview of the story construction process are provided in Appendix A.2. Ultimately, after applying sentence randomizations and numeric randomizations to the 28 template stories and randomly pairing two stories into a choice scenario, we end up using $21,270$ distinct scenarios in the human choice experiment. We use the same scenarios for the LLM choice experiment.

2.3 Human Experiment Overview

Human choices were collected via a *Qualtrics* survey on the *Prolific* platform between March and April 2025. Respondents are U.S. citizens and residents with at least 100 completed surveys and an approval rating of $\geq$ 95%, with quotas ensuring equal proportions of female and male respondents.

[1] Each respondent who completed the two survey waves (see Subsection 2.3) is presented exactly one repeated scenario from the first wave, providing a simple measure of the noisiness of human responses.

[2] We chose this range to ensure protagonists lived in the not-too-distant past while avoiding the most recent years, which risk evoking Covid-era associations or triggering memories of recently deceased loved ones.

Each respondent was presented with six choice scenarios. For each scenario, they chose between two life stories after being asked "Which life would you prefer for yourself?" (see Appendix Figure A3 for full instructions). We collected 3,746 completed surveys, of which 3,050 passed our inclusion criteria. Exclusions were based on two criteria: respondents who spent less than 30 seconds on any choice scenario (85% of exclusions) and respondents who failed checks designed to detect AI use (15% of exclusions).[3] The median included respondent spent 2.2 minutes per choice scenario.

Respondents who completed the Wave 1 survey and met the inclusion criteria were invited to participate in Wave 2. The second survey closely followed the structure of the first, also eliciting six scenario choices from each respondent. We received responses from 2,441 respondents in Wave 2, of which 2,162 passed the inclusion criteria. In total, this yielded 31,272 scenario choices: 18,300 from Wave 1 and 12,972 from Wave 2.

To improve the representativeness of our results with respect to the broader U.S. adult population, we reweight respondents using inverse probability weights constructed by matching the Prolific sample to the March 2025 Current Population Survey on age, gender, race, education, and income (Flood et al., 2025). Respondents missing one or more of these variables or whose responses have no direct CPS analogue are excluded, yielding a weighted analysis sample of 2,907 respondents and 29,748 scenario choices. All results reported in the paper use this weighted sample. Details on the reweighting procedure and sample balance are provided in Appendix A.4.

2.4 LLM Elicitation Procedure

Our analysis focuses on `gpt-5.4-2026-03-05`, a snapshot of OpenAI's *GPT-5.4*, but our methods are designed to be portable to any instruction fine-tuned model that can engage with a binary choice task. We focus on the *GPT-5.4* family for three reasons. First, it was OpenAI's frontier model at the time of writing, and frontier models are the most informative test of the revelation conjecture: if any LLM encodes quantitative preference structure, it is most likely to be a model trained on the largest corpus with the most sophisticated fine-tuning. Second, OpenAI provides dated model snapshots via its API, which allows us to fix the model version and ensure that our results are reproducible; models accessed

[3]The AI-detection checks include: a two-part attention check in which respondents are instructed early on in the survey how to answer a question that appears much later (an LLM agent consistently failed this check during validation), an HTML question invisible to human respondents but visible to bots parsing the page directly, and a minimum threshold on within-respondent variation in response times. Details are provided in Appendix A.3.

without a dated snapshot may silently change between queries, introducing instability that would be difficult to diagnose. Third, we assess robustness across additional response sets from six predecessor and smaller OpenAI models, spanning earlier generations, smaller variants, and multiple reasoning thresholds of the same model family. This within-family variation allows us to ask whether the results depend on model scale, vintage, or reasoning depth, without conflating differences in model family with differences in training data, fine-tuning procedure, or output format. We restrict attention to instruction fine-tuned models throughout because fine-tuning is a methodological prerequisite: a base model may fail to engage with a binary choice format in a consistent and interpretable way.

Unlike human respondents, who saw six consecutive choice scenarios and selected stories by clicking, LLMs evaluate each scenario independently and choose between labeled options ('A' or 'B'). We collect LLM choices in two ways: directly, constraining output to a single token, and allowing the model to reason before making a final choice. We validate the elicitation procedure by testing sensibility and sensitivity during prompt development; details are provided in Appendix A.5.

3 Results

We now test hypotheses derived from the revelation conjecture. First, we compare respondents' choices to (i) choices made by other respondents who saw the same choice scenario, i.e., the same pair of life stories with the identical randomizations, and (ii) choices made by LLMs that received the same choice scenario. Then, we use a decomposition model to determine the degree to which the observed correlations reflect LLM "knowledge" of human preferences.

3.1 Correlations in Story-Level Choices

The first row of Table 1 shows that the correlation between the choices of different human respondents across choice scenarios is 0.22. Because choices are binary and, on average, each story is roughly equally likely to be chosen, this correlation has a simple interpretation: it equals the fraction of respondents who agree on a given scenario (61%) minus the fraction who disagree (39%).

The remaining columns show that the human-LLM correlation (0.33 for *GPT-5.4* with direct choice and 0.36 for *GPT-5.4* with a medium reasoning setting) exceeds the between-

person correlation. This is a striking finding. If we want to predict which life story a random respondent will choose, we do better by asking an LLM than by using the choice of another respondent.

3.2 Decomposing the Correlations

It would be wrong to conclude that LLMs know our preferences better than we do ourselves. First, the higher correlation for LLMs could reflect that LLM responses have less noise. Second, there is heterogeneity in human preferences whereas LLM choices may track the preferences of the typical American.

To disentangle noise from preference heterogeneity, we examine test-retest correlations (second row of Table 1): the correlation when the same respondent sees the same scenario again (four weeks later for humans; a new API call for LLMs). The human test-retest correlation is 0.49, implying that on average people choose the same story 74% of the time in repeated choices. LLM test-retest correlations are higher: 0.83 for *GPT-5.4* with direct choice and 0.88 for *GPT-5.4* with a medium reasoning setting. As such, less noise in LLM responses is part of the explanation for the high correlation between LLM and human choices relative to the correlation in choices between different people.

We use a simple decomposition model to quantify three sources of divergence between observed choices and the choices of a representative person in a deterministic setting: response noise, preference heterogeneity, and LLM knowledge of human preferences.[4]

The first step quantifies the role of noise in responses. We assume that respondents choose according to their true preferences a fraction of the time and randomly otherwise — a pattern that Belzil and Jagelka (2025) show can arise from an endogenous effort model. We assume a parallel structure for LLM noise, which arises from built-in randomness.[5] The parameter "Random Response Probability" is a monotonically decreasing function of the test-retest correlation. Respondents behave as if they were answering randomly 30.1% of the time; this fraction is 8.8% for *GPT-5.4* with direct choice and 6.4% for *GPT-5.4* with a medium reasoning setting (Panel B of Table 1).

[4]We solve the model analytically given the four aforementioned correlations in the data. See Appendix Section A.6 for details.

[5]LLM noise results from the model's "temperature," kept at the default value of 1. Some residual noise would remain even at minimum temperature due to parallel processing and floating-point precision (see, e.g., Yuan et al., 2025).

Table 1: Noise, Heterogeneity, and LLM Accuracy

	Human	OpenAI *GPT-5.4*	
	Stated Choice	Direct Choice	Reasoned Choice
	(1)	(2)	(3)
Panel A: Correlations			
Same-Scenario Agreement			
Correlation with Random Person	0.217	0.326	0.356
(Standard Error)	(0.043)	(0.007)	(0.007)
[N]	[7,126]	[24,612]	[24,612]
Repeat-Scenario Consistency			
Test-Retest Correlation	0.489	0.832	0.876
(Standard Error)	(0.019)	(0.012)	(0.011)
[N]	[2,051]	[2,051]	[2,051]
Panel B: Implied Structural Parameters			
Noise Rate			
Random Response Probability	0.301	0.088	0.064
(Standard Error)	(0.014)	(0.007)	(0.006)
Preference Heterogeneity			
Average Minority Share	0.204	—	—
(Standard Error)	(0.036)		
LLM Accuracy			
Majority Following Rate	—	0.862	0.918
(Standard Error)		(0.118)	(0.125)

Notes: This table reports empirical correlations and the structural parameters they imply under the decomposition model described in Appendix A.6. Column (1) uses choices from human respondents on the *Prolific* platform, reweighted using inverse probability weights to match the March 2025 Current Population Survey; see Appendix A.4 for details. Columns (2) and (3) use choices elicited from the LLM listed in the column header.

The test-retest correlation is computed over the choice scenarios for which a respondent saw the same scenario in both waves, with the LLM queried independently each time. The correlation with a random person is computed over observations from respondents who completed both survey waves: for humans, it is the correlation between two distinct respondents on the same scenario; for LLMs, it is the correlation between the LLM's choice and the corresponding human's choice on the same scenario.

The model assumes latent preferences follow a symmetric Beta distribution across scenarios — with estimated heterogeneity parameter 0.62 — and characterizes each respondent and LLM by a noise rate and the LLM additionally by an accuracy rate reflecting how often its non-random response matches the latent majority preference of human respondents. Standard errors are computed via a respondent-level cluster bootstrap with 100,000 replications and 2,051 respondent clusters; see Appendix A.7.1 for details.

The second step quantifies preference heterogeneity, which must be present because the human test-retest correlation (0.49) exceeds the across-person correlation (0.22). We express the amount of heterogeneity by the minority-preference share: the average fraction of respondents whose true preference for a given life story in a scenario differs from that

of the majority, after filtering out noise.[6] We find an average minority share of 0.20 — that is, on average 80% of respondents truly prefer one life story in a given scenario while 20% prefer the other. Due to noise, observed shares are closer to chance: 71% and 29%.

In the final step, we ask: if LLMs could perfectly infer the majority's true preference, what correlation with human choices would we expect, given our estimates of noise and heterogeneity? Based on our model, this hypothetical correlation is 0.38 for *GPT-5.4* with direct choice and 0.39 for *GPT-5.4* with a medium reasoning setting. Actual correlations are slightly lower, indicating imperfect inference. We estimate the probability that the LLM correctly infers the majority human preference at 86% for *GPT-5.4* with direct choice and 92% for *GPT-5.4* with a medium reasoning setting. However, we cannot reject that the majority following rate is 100% (p-values are 0.241 and 0.512, respectively). We next ask whether LLM revelation extends to people's valuations of bundles of life attributes, captured by our 28 template life stories, as well as to specific life attributes that we randomize independently.

3.3 Valuations of Life Attributes

The previous section treated each life story — including its specific randomizations such as income and age of death — as an indivisible whole. It described the similarity of *choices* and interpreted the correlations using a decomposition model, recovering preferences over specific choice scenarios rather than over attributes in the stories featured in the scenarios. We now open the black box by distinguishing the attribute bundles implicit in each of the 28 template stories from the attributes that we explicitly randomized within each template story. This distinction allows us to estimate a choice model over both attributes and attribute bundles, and to compare the implied preferences for each.

These attribute bundles include relationships, family, hobbies, passions, career, location, community involvement, and cause of death. With only 28 templates we cannot credibly decompose the bundles, but we can test whether humans and LLMs value them similarly. We also examine whether LLMs and humans respond in the same direction and at similar magnitudes to the specific life attributes that we explicitly randomized, such as income, longevity, hardship at work, good health, and era of birth. Because we experimentally vary these attributes within stories, we can credibly estimate how much weight each attribute receives in choices and express preferences as dollar-denominated tradeoffs that

[6] Because the minority share varies across scenarios, we model it as a draw from a symmetric Beta distribution.

are directly relevant for economic analysis.

Table 2 reports seemingly unrelated regressions of how respondents and LLMs respond to the five life attributes randomized within each template story: annual household income, longevity, hardship at work, good health, and era of birth. Without loss of generality, one life story is labeled the "reference story"; the dependent variable is an indicator for whether it was chosen. The explanatory variables are the differences between the reference story and the alternative in log income, log longevity, a hardship-at-work indicator, a good-health indicator, and year of birth. We also include an indicator for whether the reference story was shown first, to capture order effects. Story fixed effects ensure that coefficients on specific life attributes are identified from within-story variation, not from features of the template story (names, career path, family relationships, etc.). All coefficients and standard errors are normalized by the human coefficient, so the human column reports ones and the LLM columns show responsiveness relative to human respondents.

Table 2 yields three findings. First, all LLM coefficients on life attributes are positive, meaning LLMs match the direction of human responses for the five randomized life attributes. However, for biases in choice, such as whether to choose the first-shown story in a scenario, LLM responses do not necessarily go in the same direction as human responses. Second, the standard errors on all LLM responses to life attributes are smaller than the standard errors on the corresponding human estimates. This helps explain why the LLM estimates are significant for all three life attributes with a significant human coefficient. Moreover, the LLM estimates are even significant in three out of four cases for the two life attributes with an insignificant human coefficient. Third, the adjusted R^2 of both LLM equations substantially exceeds that of humans (0.52 and 0.59 versus 0.20), consistent with less noise and no preference heterogeneity across LLM responses. Next, we compare the magnitude of preferences implied by the coefficient estimates of Table 2.

To examine *preferences* rather than choice responsiveness, we need to calculate tradeoffs between life attributes. Specifically, we divide the coefficient on each non-income attribute by the income coefficient, yielding money-metric valuations.[7] Similarly, we turn fixed-effect estimates into preferences by dividing by the coefficient on log income, which yields "fixed-effect valuations." For example, a fixed-effect valuation of 0.5 means the attribute bundle embedded in the text of a template story is valued as much as 50 log points of income relative to the attribute bundle in the average story (given that we normalized the average fixed effect to zero).

[7] If, say, 20% of respondents randomly chose stories, coefficients would be attenuated by 20%. Ratios of coefficients are immune to such attenuation.

Table 2: Predictors of Story Choice: Human Respondents versus LLMs

	Human	OpenAI *GPT-5.4*	
	Stated Choice	Direct Choice	Reasoned Choice
	(1)	(2)	(3)
Δ Log Income	1.00***	1.27***	0.89***
(normalized by human coefficient)	(0.14)	(0.10)	(0.11)
Δ Log Longevity	1.00***	1.84***	2.02***
(normalized by human coefficient)	(0.06)	(0.04)	(0.04)
Δ Hardship-at-Work Sentence	1.00***	0.87***	1.56***
(normalized by human coefficient)	(0.20)	(0.13)	(0.14)
Δ Good-Health Sentence	1.00	4.69***	3.59***
(normalized by human coefficient)	(1.37)	(1.21)	(1.13)
Δ Year of Birth	1.00	2.37***	0.66
(normalized by human coefficient)	(0.69)	(0.51)	(0.45)
Reference Story Shown First	1.00***	-5.15***	1.68***
(normalized by human coefficient)	(0.23)	(0.18)	(0.16)
Story Fixed Effects	Yes	Yes	Yes
N	29,748	29,748	29,748
Adjusted R^2	0.20	0.52	0.59

Notes: Without loss of generality, we refer to one of two stories as the reference story and the other as the alternative story. This table presents regression coefficients where the dependent variable is a binary indicator equal to one if the reference story is selected as the preferred life. Column (1) uses choices from human respondents on the *Prolific* platform, reweighted using inverse probability weights to match the March 2025 Current Population Survey; see Appendix A.4 for details. Columns (2) and (3) use binary choices elicited from the LLM listed in the column header, presented with the same choice scenarios as shown to human respondents. All equations are estimated jointly via Seemingly Unrelated Regression using the Moore–Penrose pseudoinverse with sampling-weighted clustered sandwich standard errors (2,907 respondent clusters); see Appendix A.7.2 for details.

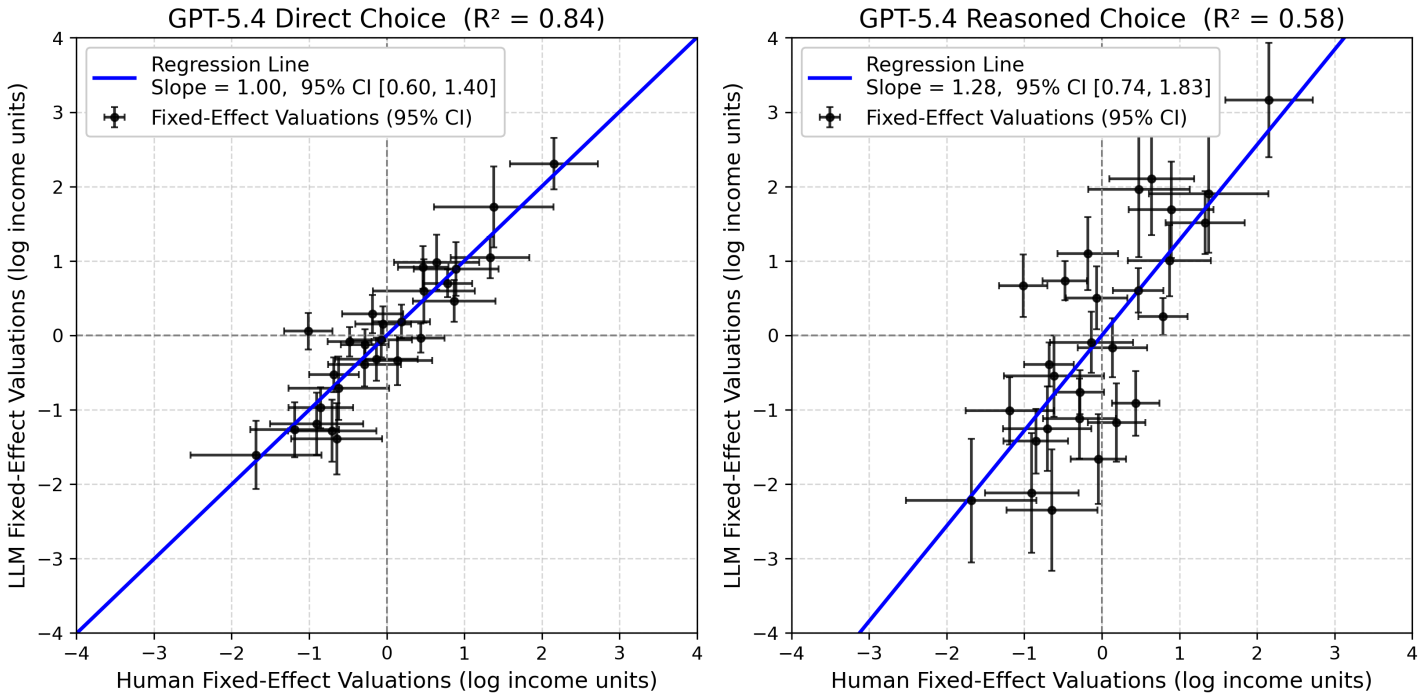
All reported coefficients and standard errors are normalized by the corresponding point estimate from column (1), so that every entry in column (1) equals one by construction. Statistical significance at the 10%, 5%, and 1% level is denoted by *, **, and ***, respectively.

We first examine preferences over the attribute bundles implicit in each story template, as measured by fixed-effect valuations. Because these bundles are arguably more representative of important life attributes than the specific life attributes that we will analyze below, they offer a more general test of the revelation conjecture. We leverage the fact that these template stories were written by a research assistant and designed to feel natural to readers and to encompass a wide variety of life attributes such as family relationships, community involvement, and career paths.

In Figure 1, we plot the estimated fixed-effect valuations for each template story from the regressions in Table 2. Story fixed-effect valuations from the LLM models are on

the y-axis, and those from the human model are on the x-axis. Since every fixed-effect valuation is estimated, there are standard errors for both (reflecting statistical uncertainty in both the fixed effect and the income coefficient). The left panel plots the fixed-effect valuations from the *GPT-5.4* with direct choice regression, and the right panel uses the *GPT-5.4* with reasoned choice regression.

Figure 1: Comparison of Human & LLM Story Fixed-Effect Valuations



Notes: This figure is derived from the same weighted SUR reported in Table 2. For each equation, each story's fixed effect is divided by that equation's log-income coefficient, yielding a valuation expressed in income units. Because the Moore–Penrose pseudoinverse imposes a sum-to-zero constraint on the fixed effect coefficients, the regression line passes through the origin by construction. Human story valuations are plotted on the x-axis and LLM story valuations on the y-axis. The left panel plots valuations from the human and OpenAI *GPT-5.4* with direct choice equations; the right panel plots valuations from the human and OpenAI *GPT-5.4* with reasoned choice equations. Error bars show 95% confidence intervals derived via the delta method. The regression slope and its standard error are likewise computed via the delta method, with uncertainty propagated through the full cross-equation SUR covariance matrix.

Figure 1 shows that story fixed-effect valuations are highly correlated between LLMs and human respondents, with R^2 statistics of 0.84 and 0.58. Hence, there is a high correlation between life stories that were more desirable to human respondents and those more likely to be chosen by LLMs, holding randomized within-story attributes such as income and longevity fixed. If LLMs and humans valued life attributes equally, the slope of the regression line would be 1. The point estimate of the slope for *GPT-5.4* with direct choice is indeed 1.00, though this precise alignment is likely coincidental given estimation noise. Indeed, for *GPT-5.4* with reasoned choice the slope is 1.28. The important lesson from

this figure is that both slopes are significantly different from 0 and neither is significantly different from 1 ($p = 0.99$ and $p = 0.31$, respectively). The visual and quantitative similarity of LLM and human valuations across the 28 template stories supports the revelation conjecture.

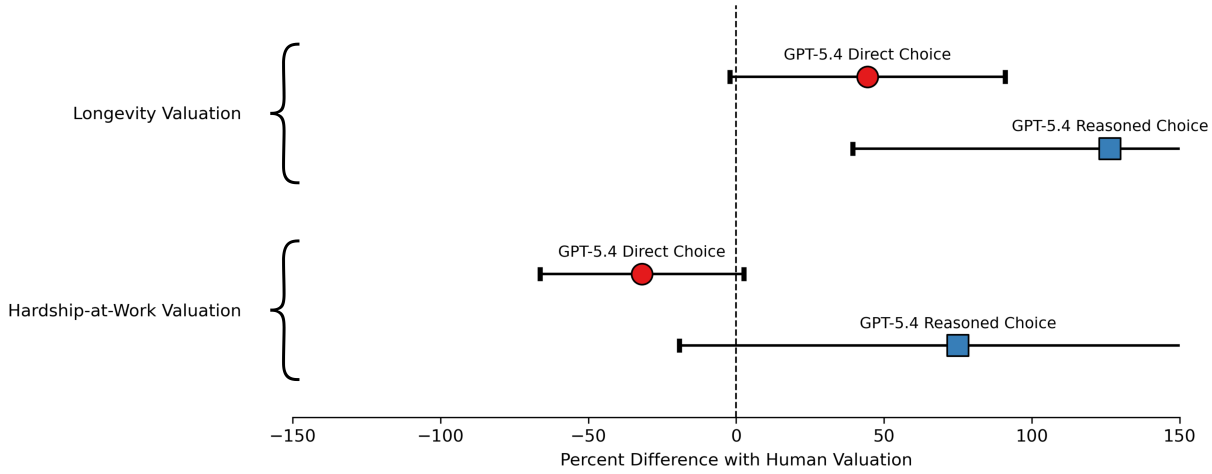
Finally, we examine the degree to which LLM valuations of specific life attributes correspond to human valuations. As shown in Table 2, human choices responded significantly to three life attributes: income, longevity, and hardship at work. Given that we use income to turn responsiveness into valuations, we can estimate human valuations with reasonable precision only for longevity and hardship at work. Panel A of Figure 2 reports the percent difference between LLM and human valuations with 95% confidence intervals for these two attributes, with the confidence intervals reflecting both the uncertainty in the estimated human valuations and the estimated LLM valuations. The valuations of *GPT-5.4* with direct choice lie within 50% of the human valuations. However, for these two life attributes, *GPT-5.4* with reasoned choice shows differences with human valuations that range between 50% and 150%, and with the confidence interval excluding zero for the longevity valuation. For the two attributes in Panel B (era of birth, good-health indicator), human responses are too noisy to yield informative comparisons (as shown in Table 2).

Overall, we find that all estimates of *GPT-5.4* with direct choice are consistent with the strongest test of the revelation conjecture, establishing the existence of an LLM that makes choices that quantitatively replicate human preferences for a range of life attributes. The performance of the *GPT-5.4* with reasoned choice was more varied.[8] While it outperformed direct choice in predicting the choices of human respondents across life stories, it was not as accurate in estimating human valuations of life attributes, and we could reject equal valuations in one case. One interpretation is that reasoning helps the model focus on the most salient attributes in a story, improving its ability to predict choices. But for valuations — which require trading off attributes like income and longevity that are always present but vary in degree — reasoning may be less helpful. Such tradeoffs may rely more on intuition and feeling than on deliberation, which is why we used the life story framework to infer people's preferences in the first place, rather than presenting them with direct tradeoffs between life attributes.

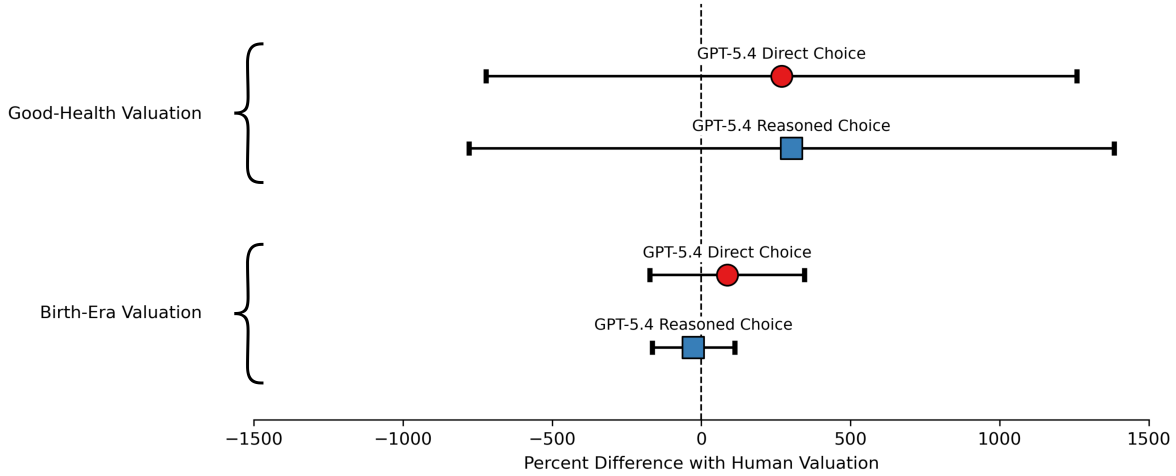
[8]Results for six predecessor and smaller OpenAI models are reported in Appendix A.8. Performance varies across models: some, such as *GPT-4.1* and *o3*, yield results broadly similar to *GPT-5.4*, while the smaller *GPT-4o mini* and *GPT-4.1 mini* models do not reveal human preferences as well.

Figure 2: Differences between Human and LLM Estimates of Life Preferences

Panel A: Differences with Precisely Estimated Human Valuations



Panel B: Differences with Imprecisely Estimated Human Valuations



Notes: Each marker shows the percentage by which the valuation implied by LLM choices differs from the corresponding human valuation, where each valuation is defined as the ratio of the attribute's regression coefficient to the log-income coefficient. All estimates are derived from the same weighted SUR reported in Table 2, with confidence intervals and standard errors computed via the delta method propagated through the full cross-equation SUR covariance matrix. Panel A reports comparisons for attributes for which the human valuation is precisely estimated; Panel B reports comparisons for attributes for which the human valuations are imprecisely estimated, resulting in percentage differences with wide confidence intervals.

If we provided LLMs with information about human respondents, would they be better able to recover their preferences? In principle, such information should be helpful. In practice, however, it is not clear that LLMs would learn the mapping between a person's characteristics and their preferences when many texts that LLMs were trained on provided

little or no information about the characteristics of the writer.

To answer this question, we re-ran the main analyses after providing the LLM with (i) the respondent's age, gender, marital status, and number of children, or (ii) those demographics plus the respondent's answers to an eight-item questionnaire on attitudes toward longevity, finances, work, hardship, risk, altruism, time preferences, and material-comfort tradeoffs. Neither modification meaningfully improved preference revelation. Details on this exercise can be found in Appendix A.9.

4 Discussion

We find that LLMs can *quantitatively* reveal human preferences over the quality-of-life attributes that we varied in our survey of adult Americans. This revelation conjecture is supported by three results: LLM choices predict individual human choices better than another respondent's choices do, LLM valuations of attribute bundles scale roughly one-to-one with human valuations, and LLM estimates of dollar-denominated tradeoffs between specific life attributes are generally within the confidence intervals of those derived from human responses. Crucially, the LLM produced these results before our survey instruments or data were published, ruling out the possibility that it simply absorbed findings from the literature.

LLMs may provide complementary signals even when human data are available, adding precision to estimates from human responses or yielding priors for Bayesian estimation. LLMs do not tire and can produce estimates at a granularity infeasible for human respondents. Appendix A.5.1 illustrates this by mapping indifference regions in income–longevity space through repeated LLM queries — a task prohibitively expensive with humans.

More broadly, the way people write about their experiences may sometimes be more informative about life preferences than observed behavior, which is subject to temptation, present bias, and inattention. Fiction in particular is a powerful form of social and emotional simulation: it lets people explore the complexities of life, witness diverse relationships, grapple with definitions of success, and confront mortality through the eyes of others. Whether it is Jane Austen's portraits of marriage and social ambition, Reddit threads about career regret, or obituaries that distill a life into its most valued elements, the training corpus contains an implicit preference-relevant structure that no single survey could capture.

LLM revelations derive their credibility from validation against human data. We cannot yet be certain whether LLMs recover life preferences or replicate systematic biases that masquerade as preferences (Mullainathan and Obermeyer, 2017). To shed light on this issue, researchers could examine settings where the degree of bias in human responses is understood, for example where optimal decision rules are known but heuristic use is frequent (Mu et al., 2025), to describe circumstances under which LLM responses reflect life preferences more accurately than human choices.

5 Conclusion

The estimation of life preferences has long been hindered by the noise and various biases contained in data from surveys, experiments, and observed behavior. We show that LLMs, trained on a vast corpus of human writing, can help. The three results laid out above — choice prediction, bundle valuation, and attribute-level tradeoffs — all support the revelation conjecture for life preferences held by adult Americans. LLM choices are also less noisy than human choices and more responsive to variation in life attributes, suggesting they may add precision to estimates from human responses even when such data are available.

Our findings have practical implications for researchers. First, LLMs can serve as cheap and scalable pilots for survey instruments. Before investing in a large representative survey, researchers can use LLMs to screen stimuli, estimate likely effect sizes, and identify which attributes matter enough to randomize, at negligible cost relative to a Prolific study. Second, LLMs may allow preference elicitation in domains where directly canvassing human subjects is costly, ethically challenging, or simply impossible. Populations that left written traces but cannot be surveyed, including historical cohorts, non-internet-connected communities, or people in settings where sensitive topics cannot be raised directly, may nonetheless be represented in the corpus on which LLMs were trained. Third, LLM reasoning traces offer a window into what drives choices that human response data alone cannot provide: which attributes capture attention, how tradeoffs are framed, and where deliberation departs from intuition. Together, this suggests that LLMs are a complement to existing methods of estimating life preferences, one that is faster, cheaper, and scalable to settings where direct elicitation is costly or infeasible.

Our results suggest that LLMs can be used to recover quantitative information on human preferences. Future research should explore which parts of the human-written

corpus on which LLMs are trained are especially useful for revealing preferences, and to what extent post-training facilitates this process. It is also important to understand whether and how LLMs can be prompted to reveal information on the preferences of particular groups of people.

References

Aher, Gati V., Rosa I. Arriaga, and Adam T. Kalai (2023) "Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies," in *Proceedings of the 40th International Conference on Machine Learning*, 202, 337–371, Proceedings of Machine Learning Research.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate (2023) "Out of One, Many: Using Language Models to Simulate Human Samples," *Political Analysis*, 31 (3), 337–351.

Asirvatham, Hemanth, Elliott Mokski, and Andrei Shleifer (2026) "GPT as a Measurement Tool," NBER Working Paper 34834, National Bureau of Economic Research.

Belzil, Christian and Tomáš Jagelka (2025) "Separating Preferences from Endogenous Effort and Cognitive Noise in Observed Decisions," IZA Discussion Paper 18315, Institute of Labor Economics (IZA).

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell (2021) "On the Dangers of Stochastic Parrots: Can Language Models be too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Bernheim, B. Douglas, Kristy Kim, and Dmitry Taubinsky (forthcoming) "Welfare and the Act of Choosing," *Journal of Political Economy*.

Brand, James, Ayelet Israeli, and Donald Ngwe (2025) "Using LLMs for Market Research," *Harvard Business School Marketing Unit Working Paper* (23-062).

Cheng, Myra, Tiziano Piccardi, and Diyi Yang (2023) "CoMPosT: Characterizing and Evaluating Caricature in LLM Simulations," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10853–10875: Association for Computational Linguistics.

Dominguez-Olmedo, Ricardo, Moritz Hardt, and Celestine Mendler-Dünner (2024) "Questioning the Survey Responses of Large Language Models," *Advances in Neural Information Processing Systems*, 37, 45850–45878.

Flood, Sarah, Miriam King, Renae Rodgers et al. (2025) "IPUMS CPS: Version 13.0 [dataset]," IPUMS, Minneapolis, MN.

Horton, John J., Apostolos Filippas, and Benjamin S. Manning (2026) "Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?" NBER Working Paper 31122, National Bureau of Economic Research.

Jagelka, Tomáš ⓡ Erzo F.P. Luttmer ⓡ Joshua Schwartzstein ⓡ Amitabh Chandra (2026) "Living Large or Long? Preference Estimates from Completed-Life Stories," Unpublished Manuscript.

Kőszegi, Botond and Matthew Rabin (2008) "Choices, Situations, and Happiness," *Journal of Public Economics*, 92 (8-9), 1821–1832.

Ludwig, Jens, Sendhil Mullainathan, and Ashesh Rambachan (forthcoming) "Large Language Models: An Applied Econometric Framework," *Annual Review of Economics*.

Mu, Tianshi, Pranjal Rawat, John Rust, Chengjun Zhang, and Qixuan Zhong (2025) "Who is More Bayesian: Humans or ChatGPT?," `https://arxiv.org/abs/2504.10636`.

Mullainathan, Sendhil and Ziad Obermeyer (2017) "Does Machine Learning Automate Moral Hazard and Error?," *American Economic Review*, 107 (5), 476–480.

Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto (2023) "Whose Opinions do Language Models Reflect?," in *Proceedings of the 40th International Conference on Machine Learning*, 202, 29971–30004, Proceedings of Machine Learning Research.

Stantcheva, Stefanie (2023) "How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible," *Annual Review of Economics*, 15 (1), 205–234.

Yuan, Jiayi, Hao Li, Xinheng Ding et al. (2025) "Understanding and Mitigating Numerical Sources of Nondeterminism in LLM Inference," `https://arxiv.org/abs/2506.09501`.