GPT as a measurement tool

Hemanth Asirvatham
OpenAI

Elliott Mokski*

Andrei Shleifer
Harvard University

Abstract

We present the **GABRIEL** software package, which uses GPT to quantify attributes in qualitative data (e.g. how "pro innovation" a speech is). We assess GPT performance on classification and regression against 1000+ human annotated tasks across a range of topics and data. We find that GPT as a measurement tool is accurate across domains and indistinguishable from human evaluators in many applications. Our evidence indicates that GPT is not relying on memorized training data or inferring attributes from other attributes. We show that labeling results do not depend on the exact prompting strategy used. We showcase the possibilities of GABRIEL by quantifying novel and granular trends in Congressional remarks, social media toxicity, and county-level school curricula. We then apply GABRIEL to study the history of tech adoption, using it to assemble a novel dataset of 37,000 technologies. We document a tenfold decline of time lags from invention to adoption over the industrial age, from ~50 years to ~5 years today. We quantify the increasing dominance of companies and the U.S. in innovation, alongside characteristics that explain whether a technology will be adopted slowly or speedily.

*Work conducted while at OpenAI.

Acknowledgements: We thank Julien Berman, Ronnie Chatterji, Hemanth Bharatha Chakravarthy, Kevin Chen, Tess Cotter, Karina Dotson, Rob Friedlander, Zoe Hitzig, Nathan Lane, Jimmy Lin, Aakash Rao, Cristopher Rosas, Suproteem Sarkar, Carl Shan, Jesse Shapiro, Cassandra Duchan Solis, Larry Summers, Adi Sunderam, Keyon Vafa, Ben Workman, Gawesha Weeratunga, and David Yang for helpful comments. Any remaining errors are the authors' alone.

The views expressed in this paper are those of the authors and do not necessarily reflect the views of Harvard University, OpenAI, or any other organization with which the authors are affiliated. GABRIEL — the Generalized Attribute-Based Ratings Information Extraction Library — is an open source Python library developed by the first two authors for the use of LLMs in social science research tasks.

1. Introduction

The vast majority of data generated by human activity is qualitative: social media posts, teaching curricula, policy documents, books, diaries, interviews, court opinions, historical newspapers, photographs, websites, advertisements, audio recordings, and more. This data is rich but unstructured. It cannot easily be used to test hypotheses in a statistically sound way. The alternative is quantitative data, which is testable but scarce and abstracted away from human nuance. This forces a dilemma: researchers must choose either quantitative rigor or qualitative texture.

The advent of large language models (LLMs) offers a solution. LLMs like GPT are powerful comprehension machines. Like a human, they can richly understand qualitative data, and then label and measure quantitative attributes on that data. Relative to human labeling, they can be cheaper, faster, more accessible, and easier to scale for many measurement tasks.

This paper introduces GABRIEL, a Python library designed to leverage LLMs to label qualitative data. GABRIEL (the **G**eneralized **A**ttribute-**B**ased **R**atings **I**nformation **E**xtraction **L**ibrary) is a prompt-based wrapper around OpenAI's GPT API designed to facilitate the use of LLMs for measurement on qualitative data like text, images, and audio recordings.

The basic intuition is simple. If a researcher can describe an attribute in clear natural language — how "anti establishment," "pro innovation," or "conciliatory" a passage sounds — then GABRIEL can ask that question consistently, across thousands of observations, and return numeric or textual measurements for every observation. The package constructs standardized prompts, executes them through the GPT API, and converts the results into structured spreadsheets ready for analysis.

GABRIEL provides methods to automate a variety of social science procedures. Researchers can rate speeches on populism, classify court cases by issue area, extract named entities from historical texts, deidentify interview transcripts, or even apply the same logic to multimodal data such as campaign posters or radio segments. There are also helper tools for passage coding, creating crosswalks to merge datasets, deduplicating a large dataset, and more (see Table 2 for a full list of methods).[0]

A growing literature uses LLMs as high-throughput measurement instruments: converting unstructured text (and other qualitative inputs) into variables that can enter standard empirical designs. That shift is promising but it raises a core question: when we treat a model's label as data, what exactly have we measured? The applied econometrics framework of Ludwig et al. (2025) emphasizes that LLM outputs are best viewed as *measurements* of latent constructs, and therefore inherit risks from measurement error and construct validity. In this setting, three concerns recur. First, *contamination and look-ahead bias*: pretrained models may draw on memorized or post-period knowledge, which is especially problematic for forecasting or any design requiring a real-time information set (Sarkar and Vafa, 2024; Lopez-Lira and coauthors, 2025; He et al., 2025; Wongchamcharoen and Glasserman, 2025). Second, the output is a *noisy proxy*: substituting LLM labels for a gold standard can attenuate coefficients or bias estimands unless the error is characterized and corrected with validation data and

[0]The tutorial to use GABRIEL can be accessed here.

appropriate estimators (Ludwig et al., 2025; Egami et al., 2023). Third, *shortcut inference and uneven generalization*: models may infer the target attribute from correlated cues rather than directly reading the relevant signal in the content, and their behavior may shift across domains in ways that surprise researchers (Vafa et al., 2024a). These concerns do not imply that LLM measurement is unusable; they imply that more work is needed to establish how broadly usable LLM measurement is across different data and methods. Here we do some of this work.

This paper has three aims. First, we validate LLMs as measurement tools on qualitative data, with an emphasis on two properties that matter for empirical work: **accuracy** and **directness**. By **accurate** we mean that, given an input, the model's measurement agrees with accepted benchmarks (human labels where the construct is inherently subjective, and objective external outcomes where available). By **direct** we mean that the measurement is driven by the signal *in the content itself*, rather than by leakage, memorized facts, or correlated cues that allow the model to guess the label without substantively reading what the researcher intends it to read. Our validation tests against a large sample of actual human labeled datasets from real empirical research, unlike prior work, which often uses cherrypicked positive or negative examples. Second, we showcase what this new measurement paradigm enables by walking through applied examples that mirror research in empirical economics and adjacent social sciences, culminating with a substantive application to the history of technology adoption. Third, we provide a transparent, standardized, and easy-to-use implementation in the form of GABRIEL, so that researchers (including non-technical ones) can deploy LLM measurement at scale without reinventing the workflow each time.

Why use GABRIEL instead of ChatGPT or the GPT API? Since GABRIEL is simply taking qualitative data and running it through GPT prompts, the natural question is why a researcher shouldn't just directly use GPT themselves. What does GABRIEL do that is harder or less convenient than direct ChatGPT usage?

GABRIEL does *nothing that couldn't be done through GPT alone*. This paper's aim is to validate and illustrate LLMs / GPT generally in measuring attributes on qualitative data, not GABRIEL specifically. Indeed, in our prompt variation section, we find that the prompt does not appear to matter much: GABRIEL's standard prompts do not yield different results from seemingly simplistic prompts. A major difference between today's AI models — with their general comprehension skills — and prior machine learning approaches is they are more generalizable and can work across many contexts without having to train or deploy a use-case specific model. LLMs do not require much technical expertise to set up, they understand the core task and ignore irrelevant details, and they can work with most qualitative data without any changes to the underlying model. We show that equivalent definitions of the same core concept receive very similar output – but definitions for genuinely different concepts are correctly differentiated.

Instead, **the purpose of GABRIEL is akin to Stata**. Stata has no secret sauce in performing a regression; any researcher could code up the same underlying math and do it themselves. Stata's advantages are that it has been optimized to handle data more efficiently, it scales better, it is a well

validated reference point relative to DIY code, it contains a number of more complex methods which would be difficult for researchers to build, and most importantly, it is simply more convenient and accessible. Our intention with GABRIEL is the same: take what researchers could do with ChatGPT or the GPT API but make it far more scalable to larger datasets, easier to use, more accessible to less technical researchers, a validated point of reference for the nascent method of AI labeling, and offering a set of difficult to build functions (e.g. creating crosswalks, ranking texts through pairwise comparisons). Our code is freely available for researchers to use and modify. Our package is simply a tool of convenience, accessibility, and reference validity, like most software.

What do we need to validate about GPT labeling data? Because we use GPT outputs as data, validation is about measurement: are we capturing the construct we claim to capture, and are we capturing it *from the input itself*? Concretely, we focus on two requirements.

First, **accuracy**. Given an input text (or image/audio segment), GPT should assign measurements that align with accepted benchmarks. Where the construct is inherently judgmental (e.g., how pessimistic a passage sounds), the relevant benchmark is typically human coding. Where the construct has an external ground truth (e.g., a county's credit score, an election outcome, a firm's realized performance), the benchmark can be objective. We also test robustness to prompt wording: do measurements remain stable across semantically equivalent prompts?

Second, **directness**. Even if a measurement correlates with a benchmark, it may be "right for the wrong reasons." We therefore test whether GPT's measurement is driven by information *in the text* (or image / audio recording) rather than by contamination, memorized facts, or correlated shortcuts. In practice, we study both (i) *look-ahead and contamination* channels, where post-period knowledge can leak into labels, and (ii) *shortcut inference* channels, where correlated cues induce the model to guess the label without reading the intended signal.

If GPT measurements are both *accurate* and *direct* across diverse settings, then researchers can treat them as usable inputs to standard empirical designs, subject to the same discipline we apply to any measurement, suited to domain-specific best practices. In Appendix B, we provide our view of best practices for using GPT as a measurement tool.

Roadmap The remainder of the paper is structured as follows. Section 2 describes the design and functionality of the package. Section 3 illustrates GPT as a measurement tool through three examples: congressional floor speeches, internet toxicity, and grade school curricula at the county level. Section 4 evaluates validity and bias, including tests designed to detect contamination and shortcut inference. Section 5 applies GABRIEL to a substantive research problem — the history of technology adoption — to showcase how the package enables empirical analysis of an important question at new scales. Section 6 concludes.

2. The GABRIEL package

GABRIEL is not a new machine-learning model and performs no training or fine-tuning of its own. Instead, it is a prompt wrapper around any modern language model, such as OpenAI's GPT API. It is a straightforward set of code that makes ChatGPT's intelligence usable for research at scale. Each GABRIEL function — whether to rate, rank, classify, code passages, or extract information — corresponds to a prompt template that the package sends to the GPT model. The model itself does all the reasoning; GABRIEL organizes how the question is asked and how its answers are captured.

For example, if one were to open ChatGPT and paste a passage of text, one could type, "Rate how optimistic this speech is about technology on a 0–100 scale," and receive what would likely be a sensible reply. GABRIEL automates that same interaction. It faithfully reproduces that question across thousands of observations, recording each answer, and returning the results in a structured format. For simple applications, the researcher could come with the data — for instance a dataset of political speeches — and directly analyze or convert them to numerical datasets using GABRIEL. We also present more complex applications in which GABRIEL can help create the analysis data itself, for instance through systematic web scraping.

What distinguishes GABRIEL from ordinary prompting is validity, consistency, accessibility, and scalability.

Validity In later sections, we establish the validity of GPT measurements on qualitative data against hundreds of human labeled datasets. These tests are run using GABRIEL's prompts and pipelines, although we expect them to generalize to other pipelines. We provide evidence that, within the family of semantically equivalent well-posed prompts we test, results are highly stable.

Consistency Every observation is evaluated with the exact same wording, output schema, model parameters, and attribute definitions. This reduces the capriciousness of ad-hoc prompting. The goal is not perfectly reproducible measurements — no stochastic model can promise that, just like no human could — but rather consistency good enough for credible measurement.

Accessibility Building boutique tools around LLM APIs for specific research projects is time consuming and requires technical know-how. GABRIEL is designed so that non-programmers can leverage LLM measurements. The outputs are standard spreadsheets, so that results can feed directly into quantitative analysis.

Scalability Copying and pasting to ChatGPT may work as a method for analyzing 20 texts, not 2,000 or 200,000. By using and parallelizing the GPT API, hundreds of GPT calls are executed concurrently, allowing corpora with tens of thousands of items to finish in minutes.

GABRIEL turns GPT from a chatbot into a research instrument: it asks the same question, in the same way, of every datapoint. Where classical machine learning methods require bespoke training for each variable, GABRIEL achieves generality through GPT's broad prior on language itself. The researcher defines what to measure in words, and the system scales that judgment across an entire

corpus of entities, texts, images, or audio recordings.

The optimization within GABRIEL and the tiny cost of LLMs relative to human labelers allows for massive cost savings relative to crowdsourcing. This enables measuring attributes on text or images at much larger scales, with more granular detail. Based on gauges of human reading speed from the literature (Brysbaert, 2019), we estimate that GABRIEL could cut labeling run cost by 700× relative to a human crowdsourcer paid $15 / hour when using the most premium gpt-5 model, and by 3500× or even 17,500× when using cheaper but still high-performing distilled models — see Table 38 in Appendix A. This would turn a project possibly costing hundreds of thousands or millions of dollars for a single run into something feasible on a low budget and easily replicable for multiple labels and runs.

Table 1: Cost estimates of rating each text on ten attributes (e.g. "promoting family values", "isolationist", "focused on human suffering"). Human target wage $15 / hr. Full calculations in Appendix E.

	gpt-5-nano	gpt-5-mini	gpt-5	**human**
240 State of the Union speeches	**$0.14**	$0.69	$3.46	~**$2,600**
100k full-text church sermons	**$43**	$217	$1,083	~**$700,000**

GABRIEL is applied to data via simple, one line Python commands, like the following call to rate thousands of speeches on populism.[1]

```
gabriel.rate(df, attributes={"populism": "How populist is the rhetoric in this speech?"})
```

Each function in GABRIEL represents a structured way of asking GPT to perform a comprehension task on qualitative data. Most functions share a similar architecture: take a spreadsheet with qualitative data (texts from any language, images, audio, entity names) → run each datapoint through our prompt templates → LLM model call → spreadsheet output. But each function differs greatly in how it asks the LLM to process the data.

Table 2 overviews the various functions supported by the toolkit, which range from classification to pairwise ranking to feature discovery to de-identification utilities. More details can be found in our tutorial notebook and GitHub repository (Asirvatham and Mokski, 2026). Figure 36 in Appendix A displays the default GABRIEL prompt for rating attributes on text. Table 3 below shows a sample of GABRIEL ratings measured on short State of the Union snippets. See Appendix B for our view on best practices for using GPT as a measurement tool in research.

[1]Full details on using GABRIEL are in table 2 and in the tutorial.

Table 2: GABRIEL methods available to researchers. See tutorial for more details.

(a) Measuring Attributes on Qualitative Data

Function	Purpose & Output Scale	Example Use
`gabriel.rate`	Asks GPT to score each text / image / audio / item on natural language attributes. Output = 0–100 rating.	Measure "populist rhetoric" in a speech; "toxicity" of tweets; "luxury" in ad images.
`gabriel.rank`	Pairwise comparisons between texts yields ELO-like attribute ratings. Output = grounded, relative z scores for each text.	Rank technologies by "bulkiness" or artworks by "fine brushwork".
`gabriel.classify`	Classifies texts / images / audio / items on whether provided labels apply. Output = one or more classes per item.	Tag news articles, product photos, or interview clips into topical categories.
`gabriel.extract`	Structured fact extraction on each item. Output = string / numeric values.	For each product, provide the "company", "CEO", and "year of invention".
`gabriel.discover`	Discovers natural language features which discriminate two classes of data.	Identify what distinguishes 5 star vs. 1 star reviews or successful vs. failed startups.

(b) Data Cleaning

`gabriel.merge`	Creates crosswalks. Output = merged table with GPT-matched identifiers.	Match two distinct job title directories; link patent titles to product names.
`gabriel.deduplicate`	Detects conceptual duplicates. Maps all duplicates to one representative term.	Collapse "F-18", "Super Hornet Fighter Jet", "f-18 hornet" into "F-18".
`gabriel.filter`	High-throughput boolean screening. Outputs items which meet natural language condition.	Subset 18M Wikipedia titles to only technologies.
`gabriel.deidentify`	Replaces PII with realistic, consistent fake PII. Outputs anonymized text + mapping.	Replace names, employers, addresses before sharing interview corpora.

(c) Helper Tools

`gabriel.codify`	Passage coding: highlights snippets in text that match qualitative codes.	Flag sentences about "economic insecurity" in speeches; "stressors" mentioned in interview.
`gabriel.compare`	Identifies similarities / differences between paired items. Output = list of differences.	Contrast op-eds from different districts; compare two ad campaigns.
`gabriel.bucket`	Builds taxonomies from many terms. Output = bucket/cluster labels.	Group technologies, artworks, or HR complaints into emergent categories.
`gabriel.seed`	Enforces a representative distribution / diversity of seeds.	Initialize unique personas that match US population distribution.
`gabriel.ideate`	Generates many novel scientific theories and filters the cream of the crop.	Procure novel theories on inflation for potential research.
`gabriel.debias`	Post-process measurements to remove inference bias.	Ensure GPT isn't guessing climate opinions in speeches based on general political lean.
`gabriel.load`	Prepares a folder of text / image / audio files into a spreadsheet for use in GABRIEL.	Image directory converted into spreadsheet of file paths.
`gabriel.view`	UI to view sample texts with ratings / passage coding.	Spot-check classify / rating outputs; view coded passages.
`gabriel.paraphrase`	Rewrites texts consistently per instructions.	Summarize earnings call transcripts to remove company specifics.
`gabriel.whatever`	Run any GPT prompts, but leverage GABRIEL's parallelization / checkpointing.	Any set of prompts; slots into any pipeline.

Table 3: Illustrative `gabriel.rate` measurements on State of the Union snippets. **FP**: foreign policy.

President	State of the Union snippet	Patriotic	FP	Individualism	Populist	Tech optimism
George W. Bush	From expanding opportunity to protecting our country, we've made good progress. Yet we have unfinished business before us, and the American people expect us to get it done. In the work ahead, we must be guided by the philosophy that made our Nation great. As Americans, we believe in the power of individuals to determine their destiny and shape the course of history.	52	7	74	10	0
Ulysses S. Grant	On the south we have extended to the Gulf of Mexico, and in the west from the Mississippi to the Pacific. One hundred years ago the cotton gin, the steamship, the railroad, the telegraph, the reaping, sewing, and modern printing machines, and numerous other inventions of scarcely less value to our business and happiness were entirely unknown. In 1776 manufactories scarcely existed even in name in all this vast territory. In 1870 more than 2,000,000 persons were employed in manufactories, producing more than $2,100,000,000 of products in amount annually, nearly equal to our national debt.	33	3	4	1	86
Harry S. Truman	Remember their power has no basis in consent. Remember they are so afraid of the free world's ideas and ways of life, they do not dare to let their people know about them. Think of the massive effort they put forth to try to stop our Campaign of Truth from reaching their people with its message of freedom. The masters of the Kremlin live in fear their power and position would collapse were their own people to acquire knowledge, information, comprehension about our free society.	56	91	7	45	2
Barack Obama	Sixty years ago, when the Russians beat us into space, we didn't deny Sputnik was up there. We didn't argue about the science or shrink our research and development budget. We built a space program almost overnight. And 12 years later, we were walking on the Moon.	58	20	4	4	90
George W. Bush	And one of the reasons that there is so much support across this country for term limitations is that the American people are increasingly concerned about big-money influence in politics. So, we must look beyond the next election to the next generation. And the time has come to put the national interest above the special interest and to totally eliminate political action committees. And that would truly put more competition in elections and more power in the hands of individuals.	8	0	6	82	0

3. GABRIEL in practice: three examples

To illustrate how GABRIEL functions in real empirical settings, we present three examples that move from the straightforward to the complex. Each demonstrates a distinct use case — text, online discourse, and web-based data — and together they show how a single prompt-based framework can adapt across content types and research questions. In every case, we define an attribute in ordinary language, provide a corpus, and let the package perform measurement at scale. What changes is only the data, not the method.

3.1. Example 1 — political rhetoric

Our first example centers on measuring U.S. political rhetoric at scale. In the text-as-data tradition, congressional speech has long been used to construct quantitative measures of ideology, partisanship, framing, and tone at scale (Grimmer and Stewart, 2013; Slapin and Proksch, 2008; Lauderdale and Herzog, 2016; Gentzkow et al., 2019; Roberts et al., 2014; Card et al., 2022; Aroyehun et al., 2025). We ask whether an LLM can serve as the measurement instrument, in similar function but for more complicated and useful concepts. We define attributes *ex ante*, apply a consistent measurement with `gabriel.rate` on each text, and study historical patterns in congressional rhetoric. The aim here is not to advance a new theory of politics, but to provide a concrete example of the ways in which LLMs can measure and quantify complex concepts in political text.

The `gabriel.rate` function turns each input text into quantifications on a 0–100 scale, where 0 denotes absence and 100 denotes full expression, as Table 3 shows.

We apply this method throughout a massive corpus of Congressional remarks spoken by representatives and senators since 1880, sampling hundreds of congresspeople per year for 1880 to 2022, drawn from Aroyehun et al. (2025).[2] We rate each transcript using `gabriel.rate` on a multiple of political and rhetorical attributes, such as `state-controlled economy` (defined so that a high score means they promote this idea), `confrontational rhetoric`, `optimistic about technological progress`, etc. Verbatim definitions for selected attributes used in the body are provided in Table 4, and the full set of all attributes is in Table 17 in Appendix A. The attributes are defined simply, in natural language. We are not overly prescriptive or verbose and frame the measurements through concise conceptual explanations of each construct (much like the instructions which would be given to a human coder), as seen in Table 4.

The data labels produced by GABRIEL from congressional speeches could be used to assess partisanship and polarization over any issue.

[2]For computational tractability, we sample 250 random congresspersons per year, concatenating the full set of their remarks in that year into one datapoint. This leaves us with 250 concatenated speech transcripts ($\sim$ 5000 words each) for every year from 1880 to 2022. Each transcript is unique to a specific congressperson in the specified year. Even after sampling, this still provides a very substantial level of granularity into evolution over time.

Attribute	Definition
`international interventionism`	Supports active engagement in foreign affairs, including diplomatic, economic, or military actions beyond national borders.
`moral universalism`	Frames rights and moral claims as broadly applicable across groups and contexts, emphasizing general principles over group-specific status hierarchies or local custom.
`optimistic about technological progress`	Level of positive framing of science and technology as engines of future prosperity and national strength. High ratings go to speeches envisioning breakthroughs, pledging research investment, or celebrating innovation as a patriotic mission. Low ratings occur when technology is treated cautiously, ignored, or framed mainly as a threat.

Table 4: Selected attribute definitions for Congressional remarks analysis. See Appendix A for all measured attributes.

As we can see with the `international interventionism` attribute in Figure 1, attitudes towards foreign action are remarkably bipartisan. In most years, it is not possible to distinguish the average Republican from the average Democrat, in terms of how much they support foreign intervention. We observe trends we might expect, such as brief pro-intervention language being used on the Congressional floors around the McKinley presidency and WWI, as well as a large and secular increase in such language following America's entrance into WWII. These averages come entirely from individual instances of GPT reading individual transcripts — there is no awareness of the overall picture or trend, only the narrow task at hand. The ratings capture even recognizable smaller shifts, such as the Cuban missile crisis era, the Reagan presidency, and the wars in Afghanistan and Iraq.

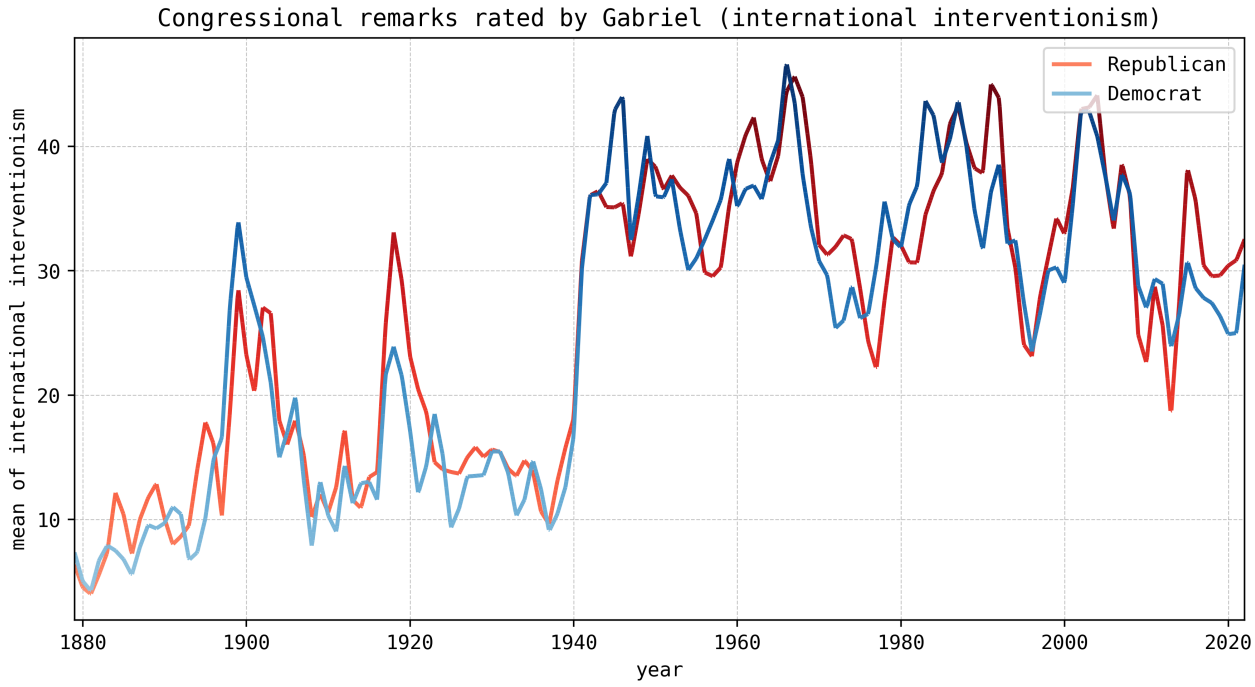


Figure 1: Support for international interventionism over from congressional speeches.

We observe a very different story when it comes to universalist moral framing (`moral universalism`)

in Figure 2 (Enke, 2020; Enke et al., 2023; Graham et al., 2009). We observe a general positive trend over time, with Democrats generally scoring higher than Republicans. But the overall level of progressive social discourse is fairly low, and remarkably bipartisan, until 2008 and the election of Barack Obama (alongside the Great Recession). 2008 marks a clear discontinuity, with a sudden and striking divergence between the parties' congresspersons. Many other socially coded attributes and interparty conflict attributes exhibit similar discontinuities in 2008; a few like opposing immigration and focusing on racial issues diverge only in the 2010s.

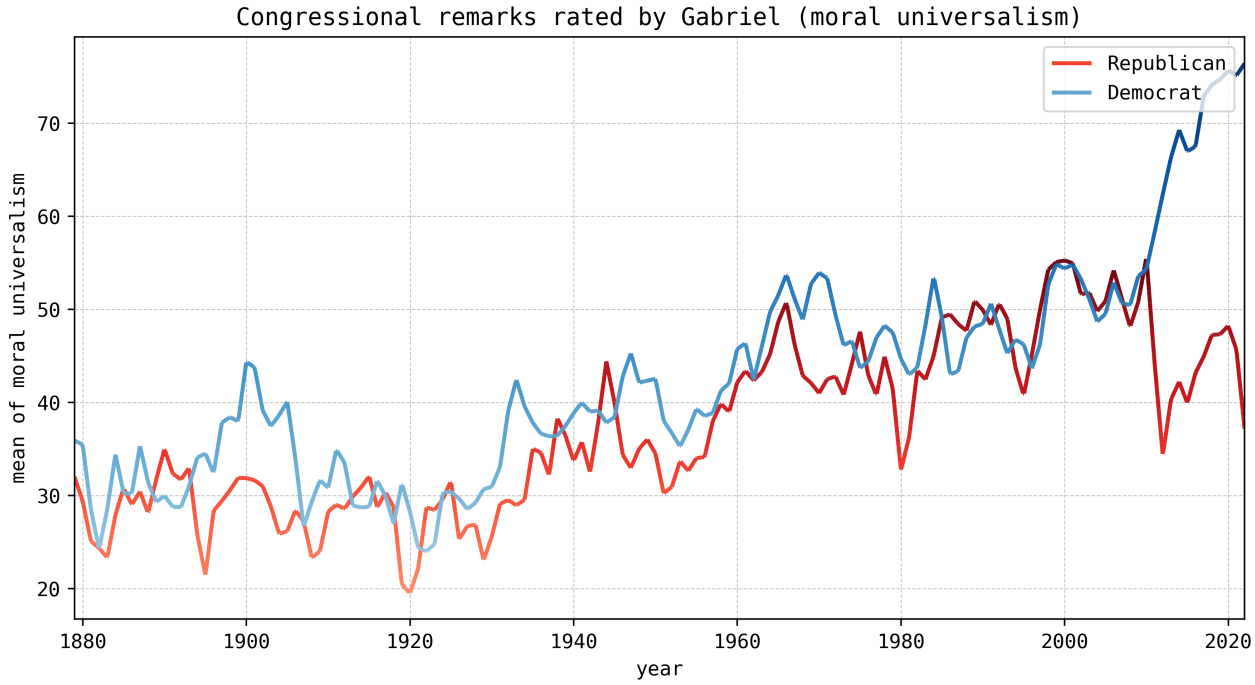


Figure 2: Universalist moral values exhibited in congressional speeches.

Scholarly accounts differ on when the modern era of U.S. polarization begins. One view locates its roots in the mid–1970s, when ideological distances in congressional roll-call votes began widening sharply (McCarty et al., 2006; Poole and Rosenthal, 1984; Shor and McCarty, 2011). Others emphasize a rhetorical realignment in the 1990s, when partisan identity became markedly easier to infer from speech and media exposure expanded through cable news (Gentzkow et al., 2019; Martin and Yurukoglu, 2017). A third line of work highlights the Obama era (2008–2010) as a turning point in the *content* of conflict, documenting a racialization of policy attitudes and the mobilization of new conservative movements such as the Tea Party (Tesler, 2012; Madestam et al., 2013). Finally, many studies interpret the time around the 2016 presidential election as an intensification of long-running trends toward affective and "sectarian" polarization rather than their origin (Iyengar et al., 2019; Finkel et al., 2020).

Our evidence here most closely aligns with the work that points to a fracture around 2008, at least in Congressional speech. It is notable how Democrats and Republicans are essentially indistinguishable on progressive social attributes from the mid-80s until 2008 — see Figure 2. While a few attributes

signal pre-2008 strife in Congress (e.g. Figure 38 in Appendix A shows a spike in `confrontational rhetoric` during the Newt Gingrich era), most attributes reinforce the 90s and early 2000s as the most bipartisan era of the past century. The sudden and complete bifurcation around 2008 is striking, also observed in Figure 37 in Appendix A. This may be explained by collegiality and Congressional norms leading to a more rapid rupture than broader society, but it challenges narratives of a gradual increase in polarization over decades.

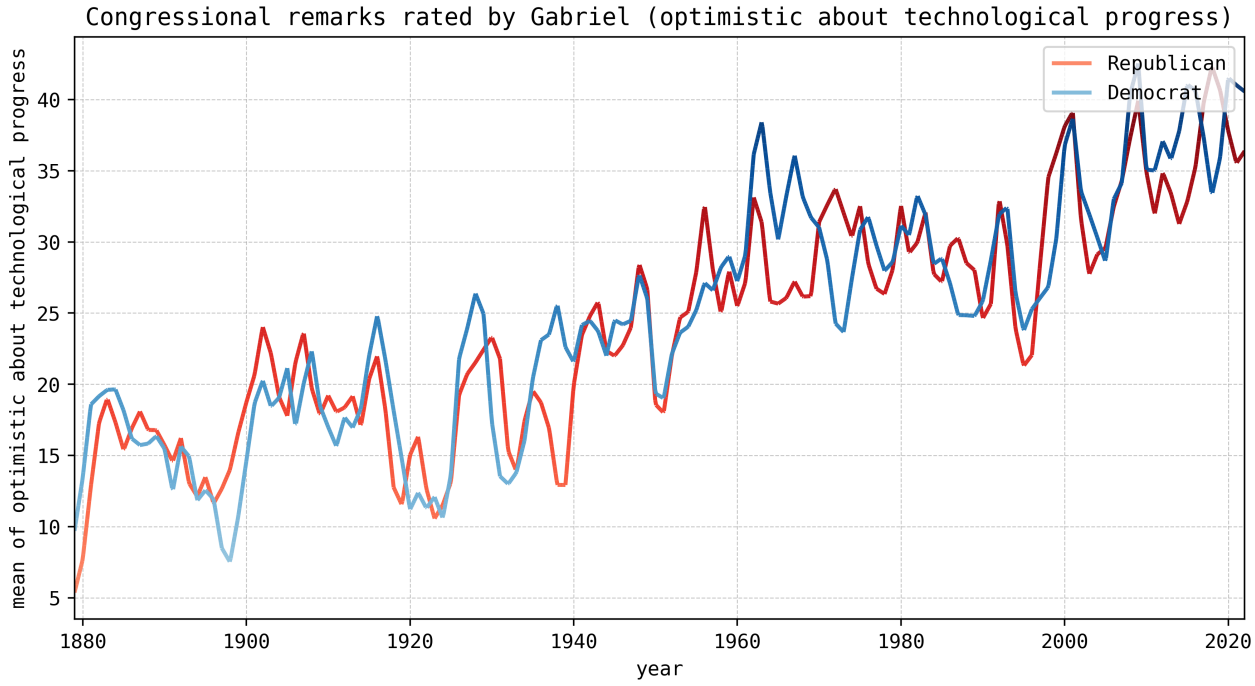


Figure 3: Optimism about new technologies and their role in society, from congressional speeches.

Another surprising result is on how Congress speaks of technology and the future. Narratives often argue that new technology was most beloved in early industrialization and has gained a negative reputation today. Our analysis in Figure 3 shows the opposite is true, at least in Congress. Tech optimism is near historical highs even in the social media era when looking at broad historical trends. Belief in technology within Congress is also bipartisan and has not diverged between the parties in recent years, unlike many social attributes. It tracks with bipartisan support for business and economic growth still enduring in Congress even as some social attitudes diverge.

3.2. Example 2 — social media toxicity

The second example applies the same logic to a less formal setting: the internet. How toxic is the conversation online? Once again, we use `gabriel.rate`, but this time on conversation threads on Reddit.[3] Our data is obtained from ConvoKit Developers (2025).

[3]Our sample consists of ~100 Reddit conversation threads for each of 100 subreddits (community groups on Reddit, such as "politics" or "Minecraft"). Each attribute — a list of samples is presented in Table 20 and the remainder are in Appendix A — is scored on a 0–100 scale for every thread, and aggregated by subreddit community.

Attribute	Definition
`toxic towards other users`	High when, across the conversation, participants direct rude, abusive, insulting, hateful, or profane language at other users (e.g., using "you" + insults, tagging users to attack them). Score higher if toxicity is frequent, sustained, escalatory, or involves multiple turns/users. Do not count playful banter clearly taken in good faith or general world-negativity not aimed at users.
`expressing fandom`	High when enthusiasm for a topic/person/work is a recurring theme across the thread (sharing favorites, lore, recommendations). Score higher if multiple participants join in or the excitement drives the conversation.
`content uninteresting to kids or teens`	High when topic/content is adult-oriented with minimal youth appeal, even to teenagers. Low if the content is particularly interesting to kids or teenagers.

Table 5: Selected attributes analyzed on Reddit threads.

We find substantial heterogeneity in toxicity across topic matter areas. Some subreddit communities, like `POLITIC` and `conspiracy` are overwhelmed by toxicity. Others, like `pokemontrades` and `askscience` are largely devoid of it. This suggests that *online content as a whole* is not toxic — pockets of it are. Figure 41 in Appendix A showcases this heterogeneity.

Variations in target audience explain the variations in heterogeneity. In Figure 4, we present evidence that the subreddits that are most relevant to adults are also the most toxic. In general, content areas likely to be consumed by children exhibit much lower levels of toxicity — suggesting that exposure to toxicity by children online may be less than one would instinctively believe.

3.3. Example 3 — county-level high school curricula

In our final example, we push LLMs beyond the analysis of existing data and into the consolidation of genuinely new data which would be prohibitively costly and slow to assemble via manual collection. We introduce a new web analysis method for understanding granular geographic variation in variables of our choosing.

We focus on US history curricula by county. We want to know what rendition of American history is taught in each county. This information does not exist in any clean or easily usable format. Curricula are hosted on individual school websites or on a teacher's webpage; they are, perhaps, described in local news or discussed in parent forums. Bits and pieces of information are sprinkled throughout the web, without aggregation.

We first use web-enabled GPT models to "scrape" the dump of unstructured information about each county and turn it into a cleanly organized report on what each county's curriculum emphasizes. In essence, we use a web-enabled model to conduct numerous web searches for each county to retrieve and synthesize county-relevant curriculum sources into a structured report. From the vast pile of disparate information online, GPT consolidates a raft of relevant primary sources into these county specific reports. We present samples of these for three counties in Table 16 in Appendix A.

We now have a clean database: for each county, a text representing an amalgam of the US history

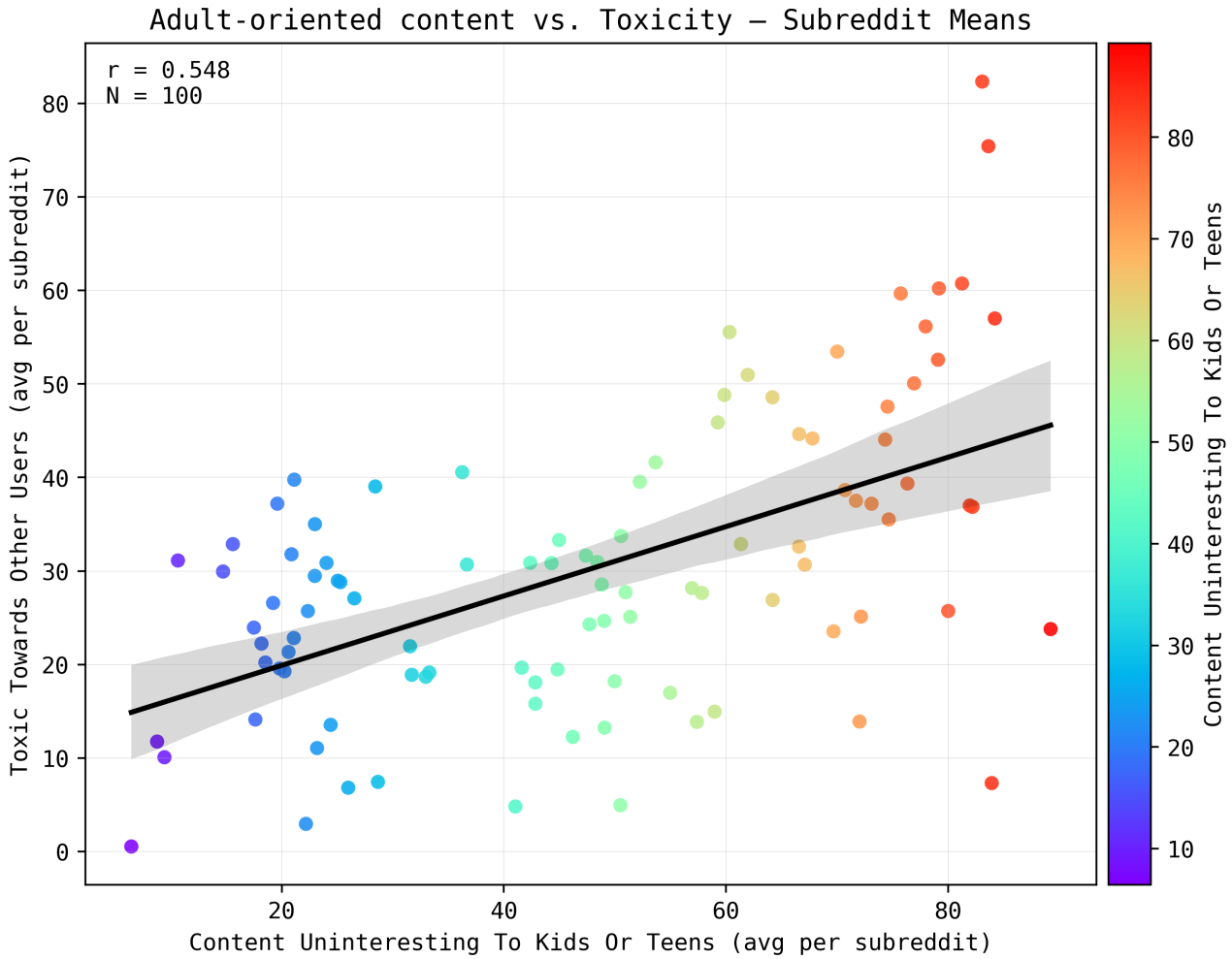


Figure 4: The subreddits most likely frequented by kids or teens are the least toxic.

curricula taught in that county's schools.[4] We use the `gabriel.rank` method to rank these county reports on various attributes, listed below in Table 6 and presented fully in Table 18 and Table 19 in Appendix A. These attributes include how positively technology is portrayed, how much certain phases of history are emphasized in the county, and so forth.[5] Below, we include full map results for a sample of attributes, as well a correlation table against demographic variables for the entirety of the attribute set in Figure 7. Across all maps in this paper, blue means that the relevant attribute is being manifested more (e.g. for `focus on technology` blue areas have a higher focus on technology while red areas have a lesser focus).

[4]We can directly apply a method like `gabriel.rate` to the internet. The intermediate step of generating reports is optional but boosts consistency.

[5]The rankings are created through a large number of pairwise comparisons between a random pair of county reports. Similar to the chess ELO system, GPT decides which county "wins" on manifesting each attribute more. That county's score increases on an attribute for every win. The magnitude of increase depends on the prior rating of its competition — if it beats a county which already had a low score, its increase is small since the win was expected. These rankings give us continuous numerical variables which represent the essence of core features about what is taught in each county, allowing us to understand the spatial variation in education with an extreme level of granularity.

Attribute	Definition
`positive portrayal of rugged individualism`	Teaching celebrates self-reliance and frontier spirit as core American virtues. More of this could show up in stories of pioneers, startup entrepreneurs, and assignments praising personal initiative over collective action.
`focus on technology and historical innovation`	Highlights how inventions and scientific breakthroughs shaped U.S. history. More of this might involve case studies of the cotton gin, railroads, wartime tech, or Silicon Valley with timelines of major milestones.
`positive portrayal of the new deal`	Curriculum casts FDR's New Deal programs as largely successful and transformative. More of this might be seen in highlighting job creation numbers, WPA art showcases, and brief acknowledgment of critics.

Table 6: Selected U.S. History curriculum attribute definitions.

Many patterns are as expected: we see more socially liberal curricula in politically liberal counties, and more religious content in conservative counties, as measured by `net democrat vote share`. Even so, in Figure 7, the correlation strength between many curriculum attributes and political lean is striking. A county's political lean explains much of its degree of focus on race. The county's percent Black population does not, after controlling for political lean — see Table 22 in Appendix A.

Other attributes follow non-political patterns. We see a greater focus on historical technology and innovation in the South (definition in Table 18 in Appendix A). We also see a more positive portrayal of the New Deal in the South. This may be explained by New Deal projects in the South, particularly in the Tennessee Valley, which exhibits the most positive portrayal on the map. We also find rugged individualism and America's natural beauty most celebrated in the rural West (attributes from Table 19 in Appenfix A).[6]

Using GPT to measure concepts on the web — or within its extensive internal knowledge — unlocks a great deal more possibility than just text analysis itself. People have written trillions of words and filmed billions of hours of content online. If captured correctly, GABRIEL could analyze the reviews of every local park in Europe, or quantify attributes of every church in America. Text analysis alone affords a manyfold increase in usable data for social scientists; unstructured text analysis offers even more.

These examples highlight the potential of GABRIEL as a measurement tool. But concerns remain. How provably accurate is GPT at these measurements, when compared to human labeled data? Is it adept at only a narrow range of tasks, or does its measurement acumen generalize broadly? Is GPT actually measuring an attribute, or has it memorized its label from its training data? Is it directly measuring an attribute like `pro environment` from a speech transcript, or is it merely guessing based on general political lean? We seek to address these concerns in our next section and return to the example of school curricula later in the section as a validation case.

[6]For illustrative purposes, we apply the same methods to capture broader cultural attributes at the county level. We allow GPT to scour social media and other local channels for each county and compile a compendium of how locals from each county talk, and what they value. The map in Figure 42 in Appendix A shows where hard work is most valued by local netizens, in what they post online.

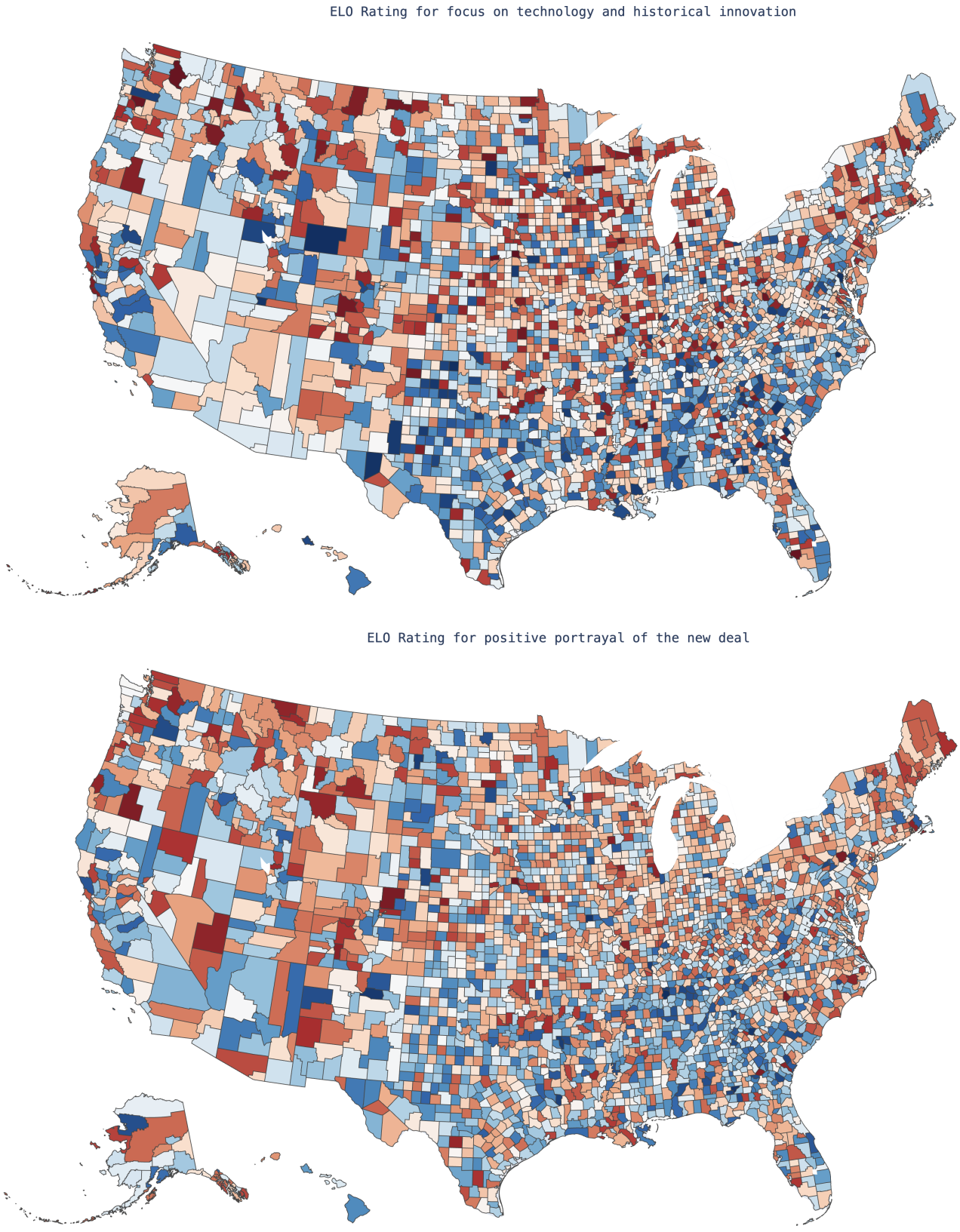


Figure 5: GPT with web scraping allows us to examine extremely granular novel data. In this case, GPT examines US History curricula at the county level and quantifies where different ideas are emphasized.

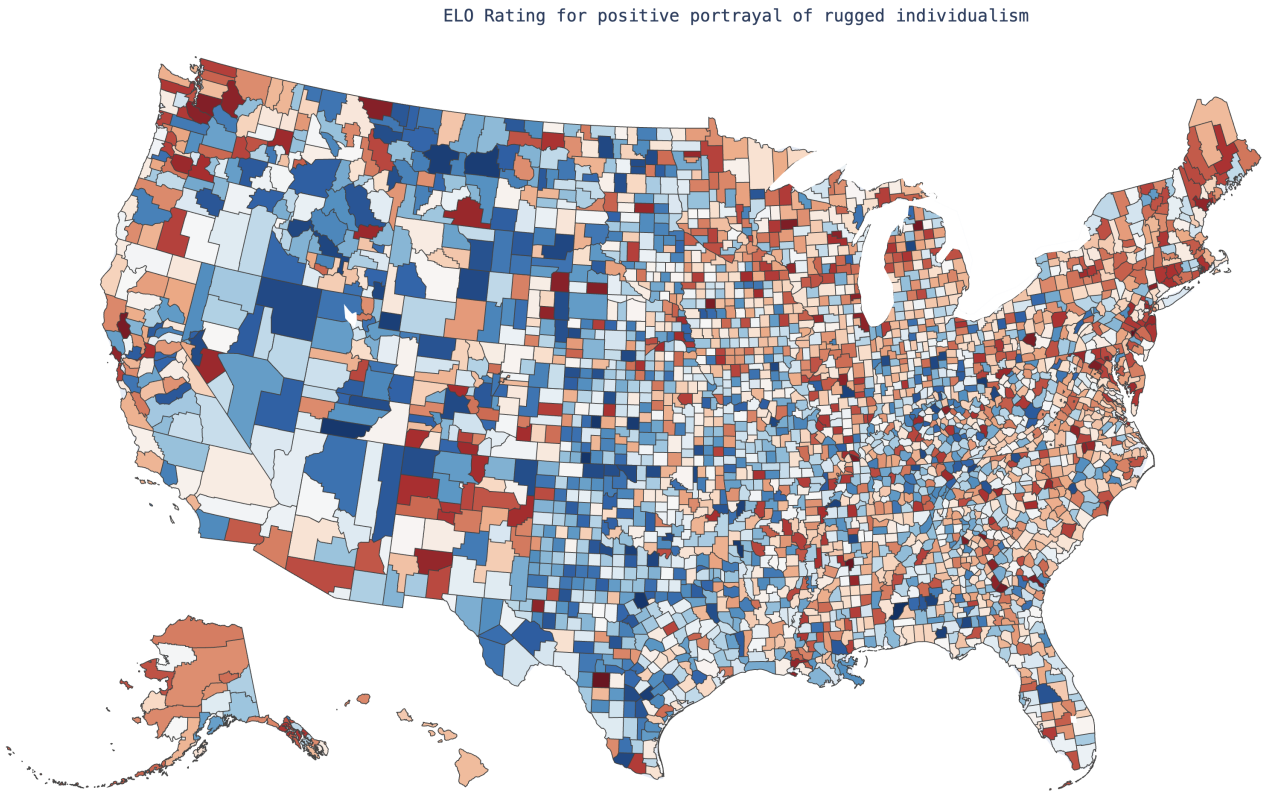


Figure 6: Rugged individualism in school curricula follows a different pattern from race or technology.

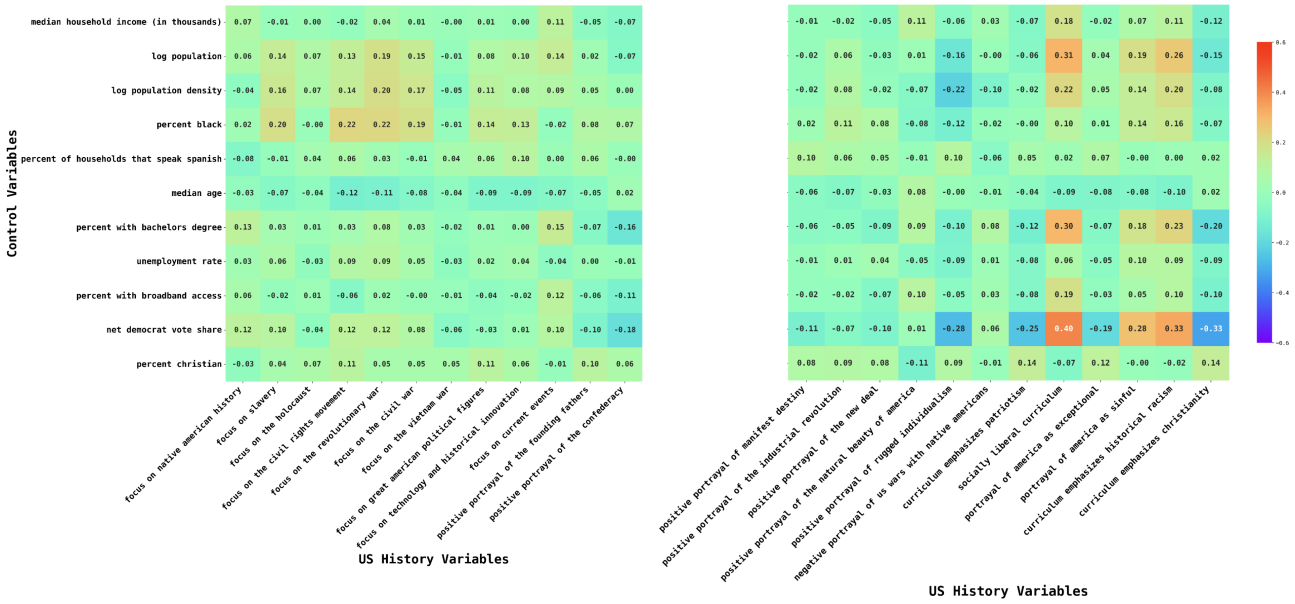


Figure 7: County US history curriculum attributes correlate with political lean (`net democrat vote share`), but also many other demographic characteristics.

4. Concerns

The preceding section illustrated what GABRIEL can do: it can quantify abstract qualities in text, assess tone across online forums, and generate complex, spatially detailed maps of cultural or institutional attributes. These examples demonstrate the flexibility of the system. But they also raise a deeper methodological question: are these measurements valid?

The capacity to produce a map or rating at scale is one thing; the harder task is to ensure that these numbers actually represent the constructs we claim to measure. This section turns to that question. We formalize the main sources of potential error, design tests to evaluate each, and interpret their results within the framework of empirical measurement. The goal is not to assert that GABRIEL or LLM-based evaluation is flawless. Instead we provide empirical backing to clarify when and why it can be trusted, compare GPT performance to commonly used human labeling baselines, and establish transparent procedures for detecting and mitigating bias.

4.1. What can go wrong? (threats to validity)

We focus on two core features of measurement, each with multiple specific manifestations. First, we want to ascertain that the LLM's measurement is **accurate**, that is that it correctly reflects some true version of the concept we want to measure. This can be either a human label of the concept, or some measurement generated by a process independent from human labels. Second, we want to verify that the LLM's measurement is **direct**, namely that the LLM measures the concept at hand and does not primarily rely on indirect inference (like shortcut inference or lookahead bias).

For *accuracy*, we focus on three distinct concerns:

1. **Agreement with human judgments**. Do GABRIEL's ratings track what trained coders would record on the same items? Across hundreds of human labeled datasets, we compare GABRIEL's ratings to the human annotations. In datasets with multiple human labelers, we check whether GPT measurements are different from the natural variability amongst a group of humans.

2. **External validity and prediction of ground truth**. Can measurements predict outcomes beyond human labels (e.g., stock returns or credit scores)? Performing as well as humans is already very useful, but human labels are not necessarily the gold standard. "Performing well" should include evidence that the variables capture real explanatory signal: the human defined concept, not just the human evaluation. Similarly, disagreement with human annotations does not necessarily mean GPT error — it could indeed be *outperforming* the humans against ground truth.

3. **Prompt brittleness and noise**. Do different prompts or attribute definitions yield different measurements? For GPT measurements to be reliable, we need the *noise* (randomness in exact wording) of the prompt to not affect the output. GPT should produce similar ratings if the prompt is brief or discursive. If the intended task is the same, we wish for GPT to understand the core intention and concept irrespective of stochastic noise in phrasing. If the intended attribute definition has a subtle but salient difference, GPT should respect that difference where it applies.

For *directness*, we further address two possible sources of bias:

1. **Contamination risk**. Are performance estimates inflated because the model previously "saw" the dataset and its labels? This is not about whether memorization is occurring, but whether GPT is *using* any such knowledge in its measurements rather than *comprehending* the text at hand. If GPT is indeed relying on memorized labels, we should observe poor performance against human labeled datasets released only after the model's training cutoff. We should also see knowledge-poor tiny models (such as gpt-5-mini) perform much worse, given these models are too small to memorize random texts and datasets. This addresses past evidence from papers including Ludwig et al. (2025).

2. **Shortcut inference (faking the measurement by measuring something easier)**. Are ratings contaminated by correlated cues (e.g., inferring "pro-environment" from partisan identity rather than the text's environmental content)? If we remove the direct signal from the text (all environmental content) but preserve everything else: the "pro-environment" should attenuate. If it does not, GPT is inferring it from other attributes, like political lean. We can use these "signal stripped" ratings econometrically to create debiased ratings as well.

The subsections that follow address each of these concerns sequentially, beginning with an overview of the accuracy of LLM measurement.

4.2. Accuracy of LLM measurement

4.2.1. Evaluation #1: performance against labeled data

Our first step in validating the accuracy of LLM measurements is to compare the ratings directly to established human labels. If the model's assessments track human judgments on the same items, we gain initial evidence that it measures the intended construct rather than noise. This is imperfect, since human evaluations are not ground truth; but it provides a first pass understanding of model performance. Human labels are also widely used in research, and just achieving performance parity with humans at computational scale would be very useful.

We begin with four demonstrative datasets, comparing GABRIEL ratings to the human labels. We assess whether GABRIEL ratings are indistinguishable from human labeling. We then apply GABRIEL at scale across hundreds of labeled datasets, showing high accuracy across a broad range of text topics and attribute types. Unlike most prior work, our assessment does not cherrypick single positive or negative example datasets but tests GPT labeling against a large sample of actual datasets.

GABRIEL performance against illustrative datasets We start with three datasets to measure performance against human labels on a battery of different numerical features. The datasets were chosen because each has continuous numerical outcomes *rated by multiple humans per observation*. This allows us measure the natural inter-rater disagreement for humans, meaning we can not only test accuracy, but whether GABRIEL is *indistinguishable* from any given human labeler. The datasets are:

- Varieties of Democracy (vDem). This data includes expert-coded assessments of political and

institutional characteristics across countries and years. Each country-year observation is coded by several independent experts (typically 5–10), who evaluate roughly 500 political indicators on ordered categorical scales. The resulting dataset spans over 200 countries and territories from 1789 to the present, yielding more than 30,000 country-year observations.[7]

- Formality scores (Pavlick & Tetreault, 2016). This dataset contains 11,274 English sentences drawn from diverse online sources including news, blogs, emails, and question / answer forums. Each sentence was annotated by multiple human raters who evaluated the perceived formality of language on a continuous scale ranging from -3 (very informal) to +3 (very formal).

- Stanford Politeness Corpus (Danescu-Niculescu-Mizil et al., 2013). The Stanford Politeness dataset consists of 10,956 utterances collected from Wikipedia Talk pages and Stack Exchange forums. Each utterance was annotated by several crowdworkers who rated its politeness on a 1–25 scale.

In each instance, we first compare the human mean to a single run of GABRIEL. Then we test whether the GABRIEL rating can be told apart from a human rating, using inter-rater variability.

We also replicate a simpler textbook task from traditional machine learning. We download thousands of Metacritic reviews — user reviews of films — and estimate the score attached with the review (out of 100) based purely on the text. This is our fourth curated example. Together, these four examples span a wide conceptual range and collectively provide a test of whether GPT-based measurement can reproduce human labels. Definitions are included in Appendix A in Table 26 for the 21 vDem variables. We attempt to match the original definitions given to human labelers wherever possible. When these are not provided, we still measure the same attributes as the humans but we define the attributes as closely as we can match the available information.

On this initial set of tests, we find that GABRIEL measurements are very similar to human labels. We find high correlations with consensus human labels across the board: lowest for politeness (around 0.55) and in the range of 0.7–0.8+ for vDemocracy (average of the 21 outcomes — see Figure 8 for the overall score), formality, and Metacritic (Figure 9). We see a clear pattern of improvement across models. While the oldest generations of models are passable but mediocre at the tasks, frontier models like gpt-5 achieve the highest performance. Our Metacritic example in Figure 9 also provides initial evidence that contamination bias is not a major driver of performance — there is no dropoff in accuracy after the model training cutoff. We return to contamination bias in more detail later.

In Figure 10, we see the average performance of three leading LLMs across our sample texts. The most advanced model, gpt-5, obtains top performance across the three task sets, at 0.78 correlation to humans for formality, around 0.71 across the vDem tasks, and 0.55 for politeness. gpt-5-mini performs nearly as well across the tasks despite being a far smaller model, with gpt-5-nano lagging further behind.

[7]We select 25 of the measured variables with consistent availability and reflecting a range of outcomes and randomly subset to 10,000 country year pairs. For each pair (e.g. US 1823) we ask the model to generate a numerical rating of the feature at hand — for instance, "freedom of the press".

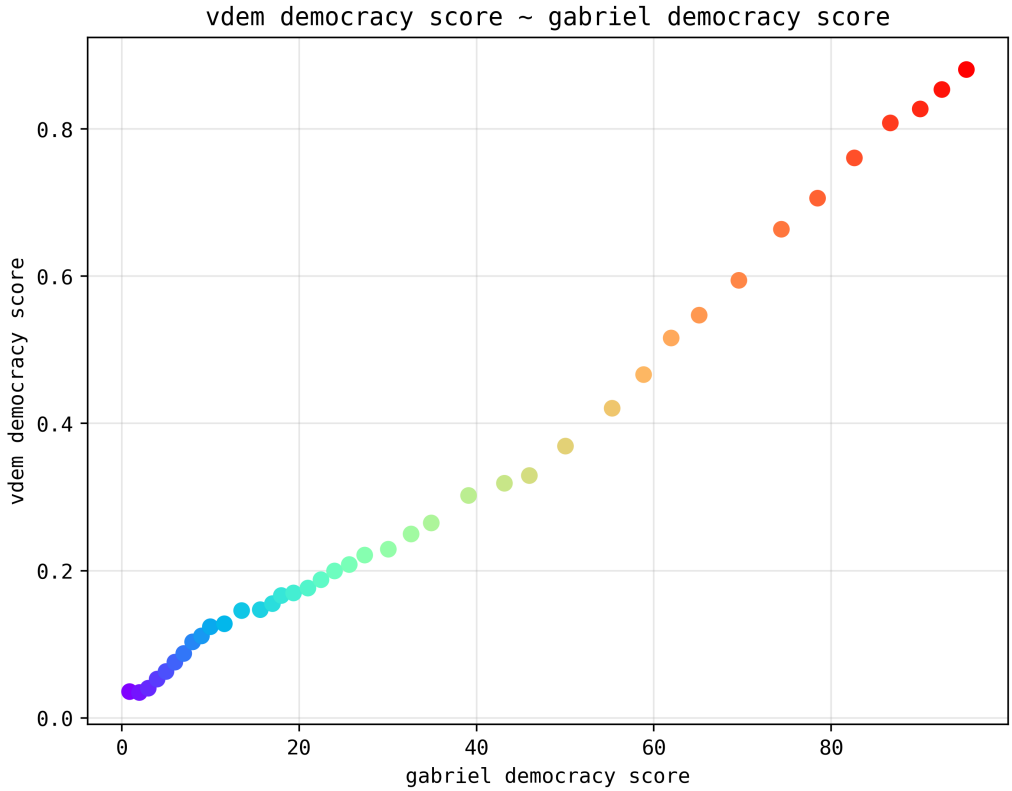


Figure 8: In replicating overall human expert vDemocracy ratings, GABRIEL on gpt-5 is a nearly exact match (so is the tiny gpt-5-mini model, where label contamination is unlikely). GPT comprehends each country's history and is well-calibrated on rating the provided democracy definition.

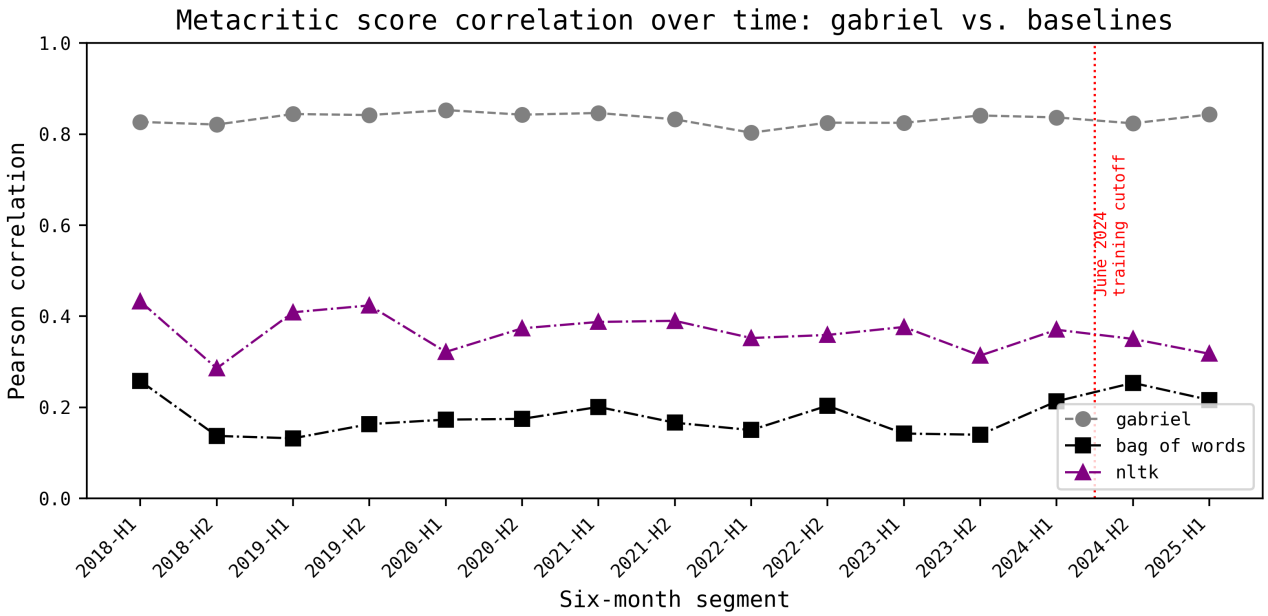


Figure 9: Predicting Metacritic rating given review. GABRIEL dramatically outperforms older zero shot techniques. Identical performance before / after training cutoff indicates no contamination bias.

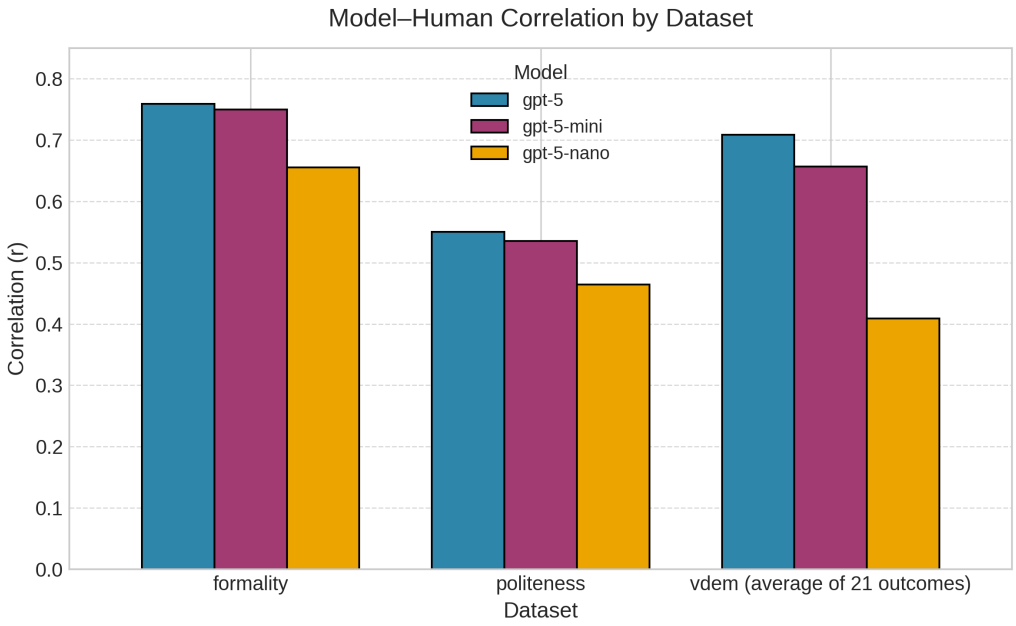


Figure 10: Average performance for frontier models on three illustrative datasets.

Is GABRIEL indistinguishable from interhuman variance? These correlations, particularly for gpt-5 and gpt-5-mini, are strong and appear close to the ceiling of what r is plausibly achievable. We test this below. The attributes we are evaluating are subjective — they depend on a qualitative understanding and *interpretation* of a text and data. They require understanding of a broad and abstract construct such as what it means for something to be "polite". To achieve correlations of 0.7 or above means that there is in practice strong agreement with human consensus. It would be very difficult to achieve correlations of the order of 1 or even 0.9 on a task that allows such a significant degree of discretion to the labeler.

We can test whether models are as good as human labelers by leveraging the fact that each of these texts was rated by multiple humans. If GPT's agreement with the human average is indistinguishable from each individual human labeler's agreement with the human average, then GPT measurements are *as good as* a human's. Indeed, humans display substantial levels of disagreement among themselves — see Table 24 in Appendix A. We compare the model correlation to human average (MH) to the human correlation to human average (HH). For model correlation MH, we simply correlate the GABRIEL rating for each text against the human annotator average for that same text. For human correlation HH, we take the correlation of each human i against the consensus judgment of all other humans and take the average of these one-hold-out correlations across the set of all humans. This one-versus-rest measurement is an easily interpretable test for correlation with humans, and it has been used widely, including in Buse and Weimer (2008); Lau et al. (2014); Koto et al. (2021); Bavaresco et al. (2025).

For each attribute we measure (`formality`, `politeness`, and the 21 vDem outcomes), we test whether HH differs from MH and whether humans match consensus better than the model. We find that HH does not exceed MH and in many cases MH exceeds HH: the model matches consensus better than

individual humans. More detail on these tests is provided in Appendix D.

While the top models slightly underperform humans on the formality task, they outperform humans on both the politeness task and on the vast majority of the 21 vDem variables. Our tests verify that these differences are statistically significant. For gpt-5, the frontier model, across all 23 tested variables we find that GABRIEL outperforms humans on 13 (12 vDem outcomes and politeness), underperforms humans on 3 (2 vDem outcomes and formality), and is statistically indistinguishable from humans on the remaining 7 outcomes. Small models including gpt-5-mini perform similarly well.

These results show that on this set of tasks, the correlations we observe to human consensus should be interpreted as near the ceiling of potential r. The capriciousness of these tasks means even individual humans correlate with the human consensus about as well as GABRIEL. GABRIEL is generally not distinguishably worse in performance than a typical human rater and frequently *exceeds* humans in terms of ability to match the consensus judgment.

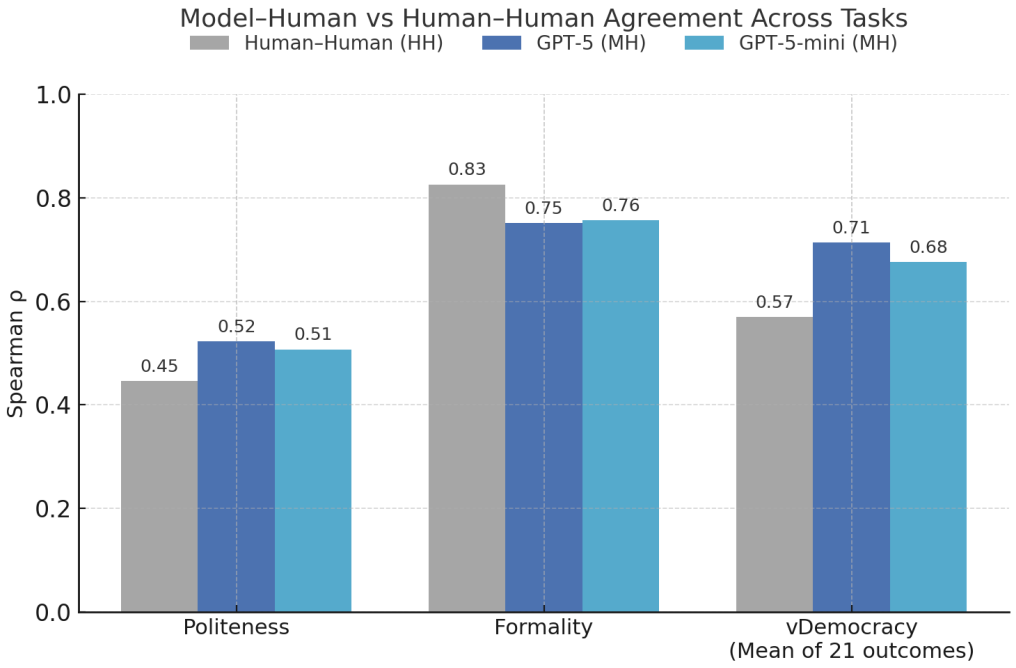


Figure 11: Human-Human (HH) and Model-Human (MH) consistency. GPT agrees with the human consensus as well as humans agree with each other; this indicates that GABRIEL performance is as high as possible against human labels.

Testing GABRIEL on hundreds of human labeled datasets Does this performance hold up in general? It is a common fear that machine learning performance may be highly uneven across subject matter. Vafa et al. (2024b) argue that LLM capabilities do not generalize across domains.

Evidence on frontier LLMs mostly supports the view that they are general purpose comprehension machines, and understand a broad range of text, just like humans. Frontier model cards from OpenAI, Google DeepMind, and Anthropic show that GPT-5, Gemini 2.5 Deep Think, and Claude Sonnet 4.5

attain frontier-level performance across an unusually wide range of tasks — competition math, broad reasoning, open-ended factual QA, high-stakes health advice, safety-sensitive dialogue, multimodal understanding, long-context retrieval, and software engineering (OpenAI, 2025b; Google DeepMind, 2025; Anthropic, 2025).

While those evaluations prove general text / image reasoning and comprehension on standardized tests, we provide empirical evidence on how model performance actually holds up across a large number of *labeling* and *measurement* tasks, consistent with our results from the four datasets above. We want large-scale empirical grounding that LLMs do understand concepts across the board with relatively small variation in competence between tasks, like humans and unlike prior machine learning approaches.

To this end, we aggregate a new set of hundreds of human labeled datasets available on the HuggingFace data repository. Many of these human labeled datasets are affiliated with real academic work that commissioned the labeling. Our methodology aims to test generalized performance by capturing as many topics and tasks as possible without any systematic source of bias. Pipeline details on how the datasets were obtained appears in Appendix C. Figure 12 shows a stylized representation of our process to create the new benchmark with around 300 full-text annotated datasets. For each dataset, we replicate the original human labels using GABRIEL and measure how well the model matches human evaluators.

Detailed results for the systematic replication task are in Appendix A, e.g. Table 29. Here, we study one subset of the tests that we run: a series of over 1000 distinct binary classification tasks from the data, where the goal is to determine whether a text is an instance of some class. Each task contains up to 200 distinct text observations. For instance, in a simple example we might ask whether a newspaper article is a `sports article`, or we might ask whether a part of a legal filing is a `pleading` or a `judgment`, or whether a given political speech is `anti immigration`. Each of these distinct tasks was sourced from the corpus of human labeled datasets from HuggingFace and contains both the texts and the human consensus label for each. We measure both raw accuracy and F1 score. F1 score balances precision and recall in cases where the observations are imbalanced across classes through $F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$ where 0 means total failure and 1 means perfect performance.

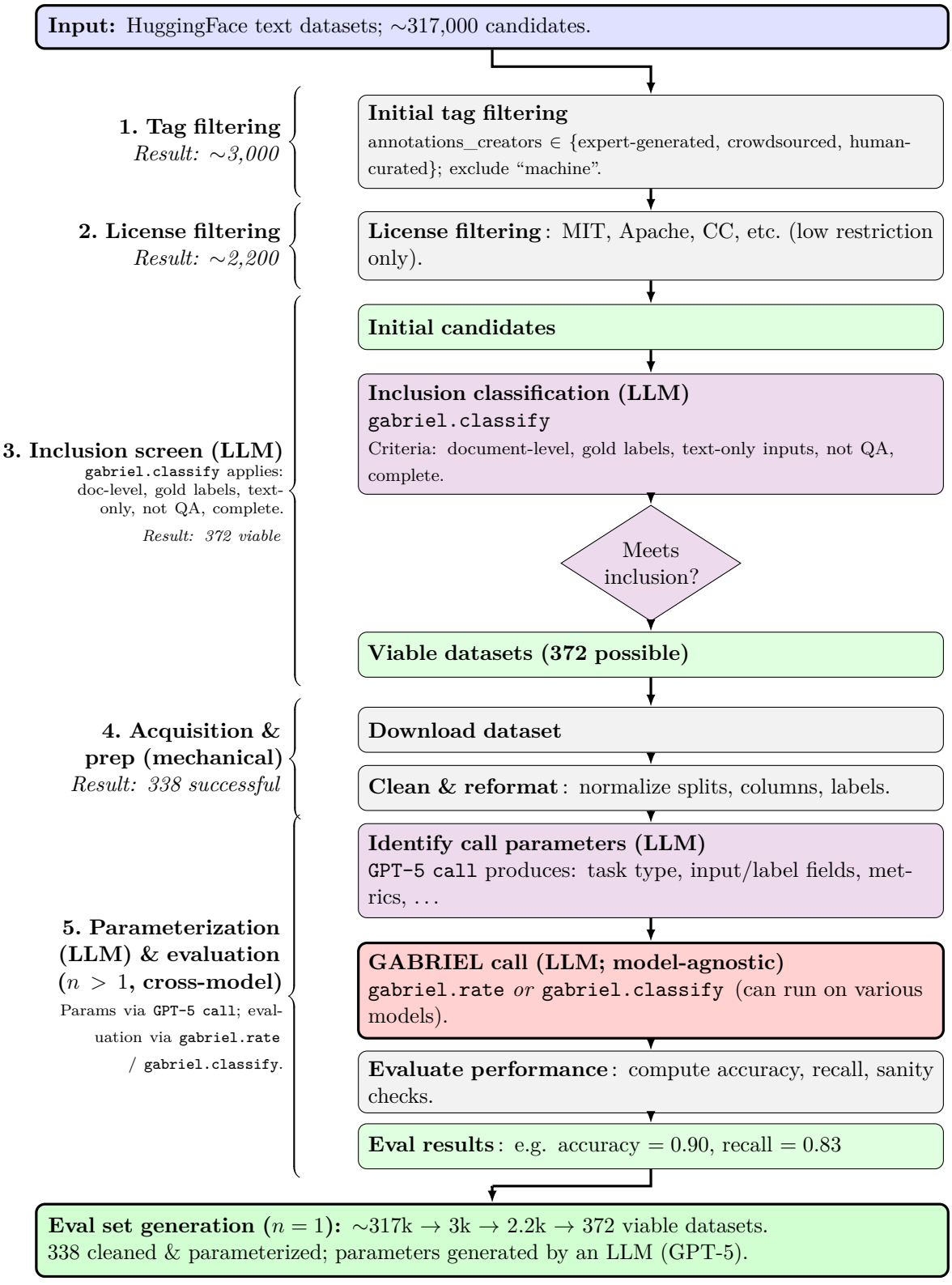


Figure 12: Pipeline schematic. Purple cells: *LLM-based* steps; gray: *mechanical filtering*; green: *dataset states*; red: main *GABRIEL* execution.

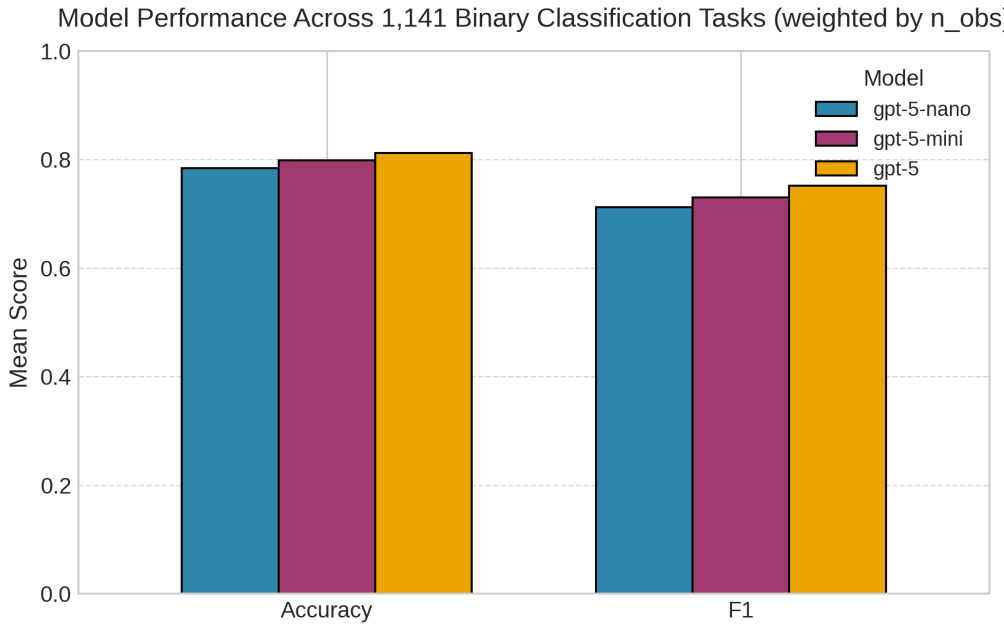


Figure 13: Classification accuracy (percentage agreement with humans) across over >1000 different classification tasks in >300 different text datasets, GPT labeling closely matches the human baseline. Inexpensive distilled models do well also.

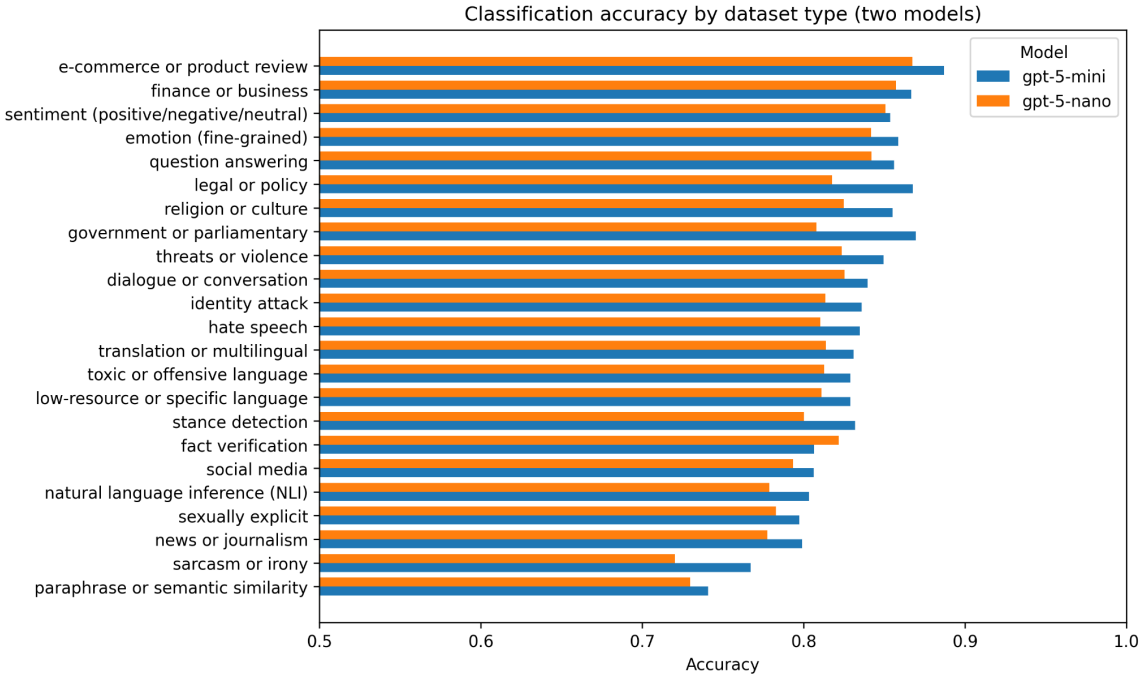


Figure 14: GABRIEL labels are similarly accurate on human labels irrespective of text topic / classification task. This indicates general usability, unlike prior machine learning approaches.

In the aggregate, as depicted in Figure 13, we find that the models perform well across the board on this large-scale evaluation. We observe high average performance in classification accuracy and F1

score for these binary classification tasks (of the form "is this text an instance of X") — in Figure 43 in Appendix A — and high correlation across a much smaller set of continuous numerical variables. While there is variation between tasks, the IQR range is on balance relatively narrow ($\sim$ 0.7–0.9+ for classification with leading models) and even accuracy on the 10th percentile task remains at around or over 0.5 for frontier models. This does not guarantee that the LLM will perform well on every single task. Instead, the evidence shows that LLMs in frameworks like GABRIEL can generally perform well across a scope of tasks reflecting real research and its use of human labeling.

This high accuracy is consistent across very different types of text and very different conceptual attributes being measured. This evidence indicates that LLMs are *comprehending in a general fashion*. A purpose-built approach using prior ML techniques would not have this flexibility to drop into any measurement task and perform accurately.

We again observe intermodel variation in performance, with the most recent frontier models in the gpt-5 series achieving leading results, outperforming legacy models like gpt-3.5-turbo. The best models achieve median accuracies of 0.85, with an inter-quartile range of more than 0.7 to near 0.95. This tightness suggests both that the model is performing well on average, and that it is performing well on the vast majority of tasks.

Two additional observations from this test suggest that performance comes from reasoning, not parroting of training data. First, more advanced and expensive models which are simply better on independent evaluations like reasoning ability (like gpt-5 versus gpt-4o) perform better here too. Simply put, more intelligent models perform better. There is a clear return to intelligence in improving output, which suggests that output quality is indeed a product of intelligence or reasoning.

Second, model *intelligence* appears more important in output accuracy than model *size*. Small distilled models like gpt-5-mini only slightly underperform the base model like gpt-5, and they outperform large models like gpt-4o which lack reasoning capabilities and perform worse on reasoning-intensive assessments. This indicates reasoning is the primary feature that explains performance. Further, small models do not have the ability to retain large amounts of facts / keep much data in memory. The fact that they only slightly underperform much larger base models suggests that contamination is not driving the results. If pure regurgitation were the driver instead of reasoning, we would expect slightly older large models like gpt-4o to outperform tiny distilled models like gpt-5-nano. The opposite is true.

Takeaways The tests in this section validate the use of LLMs as accurate measurement tools in three key ways:

1. **Broadly as valid as humans.** On our tests against human labeled data, GPT measured similar labels and ratings as did the human annotators.

2. **A general purpose measurement tool.** Unlike prior literature (and unlike the capabilities of prior machine learning approaches), GPT measurements are accurate across a range of labeling exercises, without any conditioning or manual validation / tweaking for the specific labeling task.

3. **Broadly indistinguishable from individual human labelers.** We not only observe a high correlation between GPT labels / ratings and the human consensus but also that this correlation is about as high as is possible, given that individual human raters disagree with the human consensus generally as much or more than GPT does.

4.2.2. Evaluation #2: predicting ground truth

Human labels are useful, but they are not always ground truth or the gold standard. To further assess the accuracy of LLM measurement, we shift to evaluating attributes which are directly observable from some external process or data which does not rely on subjective human annotation. Ideally, LLMs can exceed the capabilities of human labeling, not simply match it.

Estimating county level credit scores We begin by replicating recent estimates of creditworthiness across space from the Opportunity Insights Credit Atlas (Bakker et al. (2025); Opportunity Insights and U.S. Census Bureau (2025)). This example is notable in part because the data was released entirely after the training cutoff for the models we use in the analysis, meaning that the data is fully unseen by the model (and no data of this form or granularity had been published in the past).

Using the same methodology as our prior examples with school curricula, we generate county level reports of consumer purchasing trends and use these to estimate `willingness to borrow money` at the county level. The definition is provided in Table 7, and other attributes of interest we measured are in Table 21 in Appendix A. Our goal here is simply to measure one attribute about cultural attitudes which could plausibly be measured from online text and compare it to an empirical measure of a manifestation of that cultural attribute. We are not assessing causality: our interest is to measure a manifestation of the ground truth attribute purely from text.

We compare our measurements with the ground truth estimates of credit score and find a clear, strong monotonic relationship between our web-based estimates and the ground truth data in Figure 16. Because the text-based measurement never "sees" the credit data (and the data was released after the model's training cutoff, more on this in the next section), this co-movement is strong evidence that GABRIEL's variables capture real structure in the world.

Attribute	Definition
`willingness to borrow money`	Residents express comfort with debt as a normal tool for everyday life and big purchases. More of this might appear as upbeat comments about financing cars, payday loans, or juggling multiple credit cards.

Table 7: Attribute definition for credit score analysis.

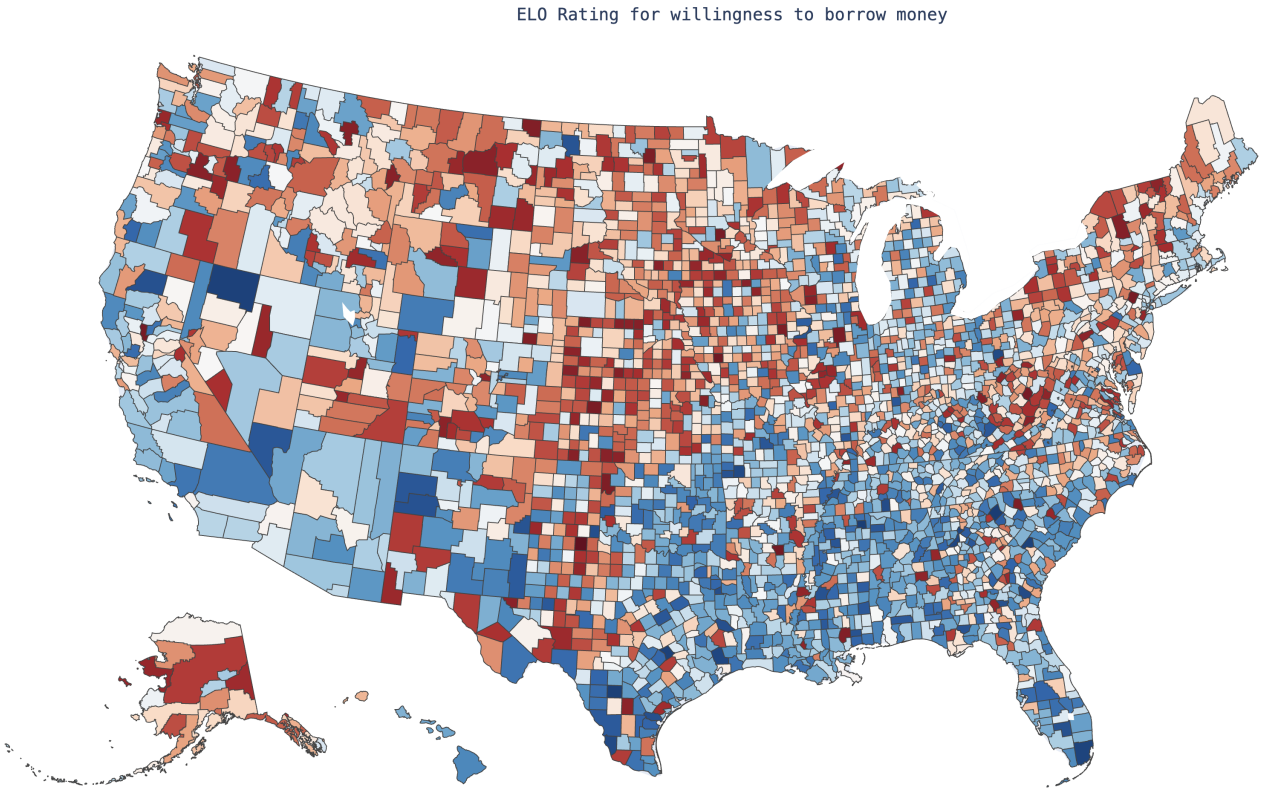


Figure 15: County level analysis of local internet content on people's willingness to borrow money.

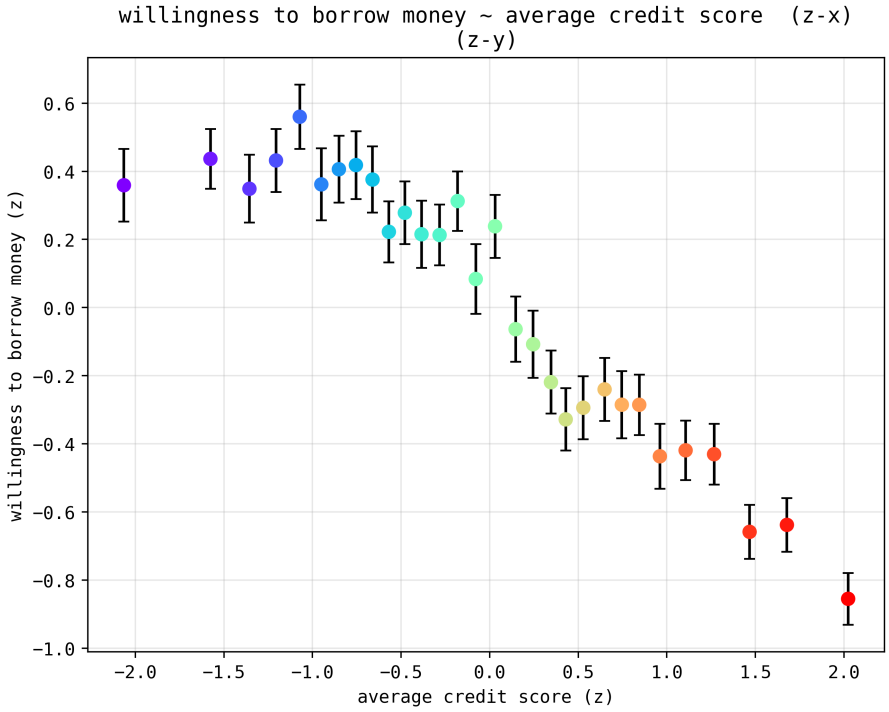


Figure 16: Our willingness to borrow ratings correlate well with county level credit scores. Our measurement was conducted prior to the credit score data release.

Replicating World Values Survey results with GPT's internal knowledge We also test GPT's measure of country-level opinion on certain value-based questions, a ground truth measurement collected by the World Values Survey (WVS) (Survey, 2022). We find that even one shot ratings purely based on GPT's internal knowledge (no additional web scraping like above) can replicate survey series. We don't expect pure reliance on internal knowledge to perform well at the county level like above, but GPT has expert level knowledge on each country's culture and manifests that here.

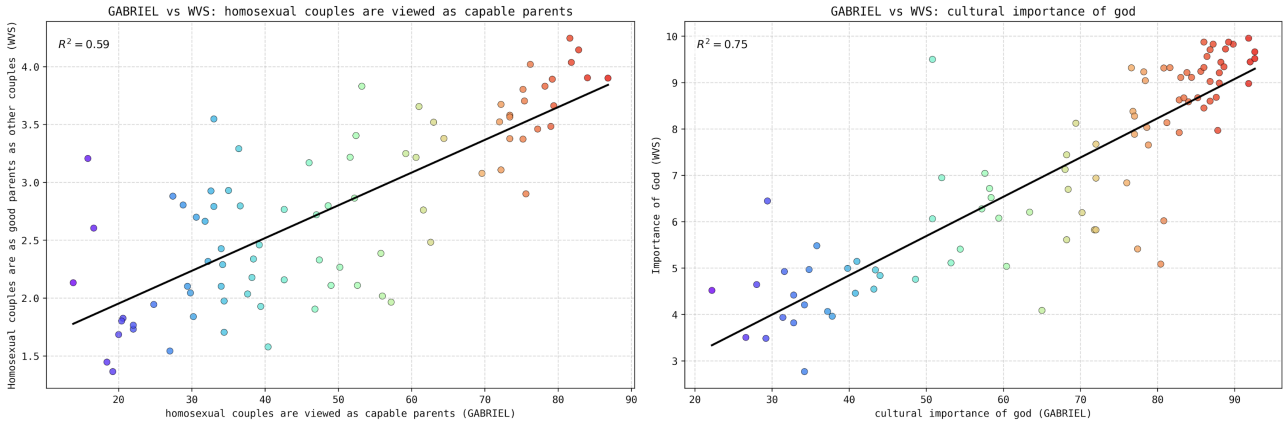


Figure 17: GPT's understanding of country level opinion / values correlates well with WVS results.

Taken together, these tests — alongside a third ground truth example around persuasiveness of arguments in Figure 46 in Appendix A, the Metacritic analysis, as well as other applied results throughout the paper (history classes, political event studies, tech history) — support the claim that GPT is measuring signal anchored in ground truth, not merely agreement with human labels.

4.2.3. Evaluation #3: prompt brittleness and noise

We next turn to how much the prompt matters in getting an accurate measurement. Prompts — like instructions to human labelers — can be very different from one another. In theory, this could lead to different measurements. Similarly, if expertise is needed in writing a prompt, this could limit possible usage by researchers. We desire the *noise* of prompt or attribute definition wording not to matter, while a *substantively different* attribute is measured differently.

Testing the brittleness of a measurement prompt We test whether prompts matter by performing the same attribute rating task on 300 State of the Union snippets, but with 100 different variants of the standard GABRIEL prompts. All the prompts aim to do the same core rating task, but the variants are dramatically altered in language and length from the base GABRIEL prompt. We achieve this diversity by using `gabriel.seed` to create 100 mutually exclusive style guides, then `gabriel.whatever` to vary the base `gabriel.rate` prompt in accordance with each style guide. Then we rate the SotU snippets with each prompt, and correlate their ratings to the baseline GABRIEL run. Examples of the prompt variants are seen in Table 8.

Example prompt variants
Prompt #1. Read `{{ text }}` front to back, then for every trait in `{{ attributes | shuffled_dict }}` hit one crisp 0–100 JSON stat you independently assign on that attribute's own direct track. **(32 words)**
Prompt #2. Peruse thou the whole of the foregoing text with utmost diligence—from the first syllable unto the last. Skim not, nor hurry thy gaze; rather, apprehend the entire discourse in depth, heeding even the most obscure and well-hidden subtleties that may lurk in its latter parts. Thy charge: For each several quality or property hereafter named, thou shalt render a judgment. . . **(563 words)**
Prompt #3. OK LISTEN 4 INSTRUCTIONS: U gotta read all this text thing, like from top to bottom, no skimming, no "meh good enough." Try to actually get what it means, even the sneaky subtle stuff hidden in the middle or wherever. . . **(423 words)**

Table 8: Example prompt variants used to elicit attribute ratings. See Figure 36 in Appendix A for the default GABRIEL prompt.

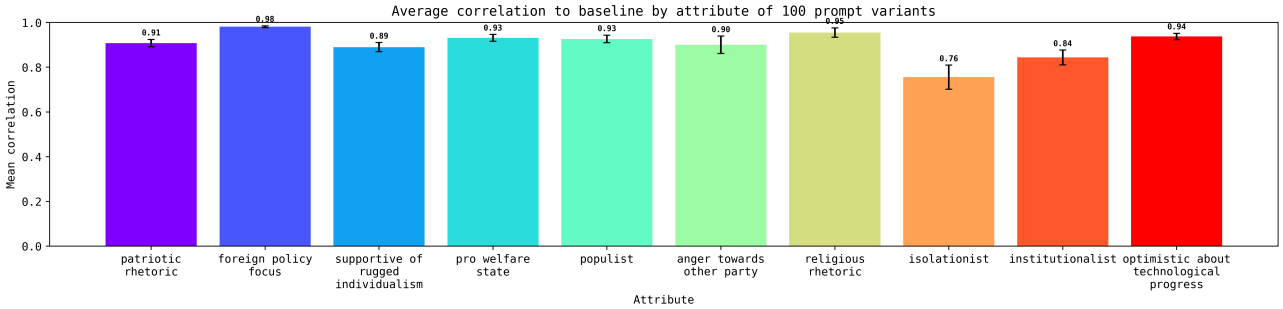


Figure 18: Significant variation of the measurement prompt has little effect on the measurement. Prompts need not be perfect.

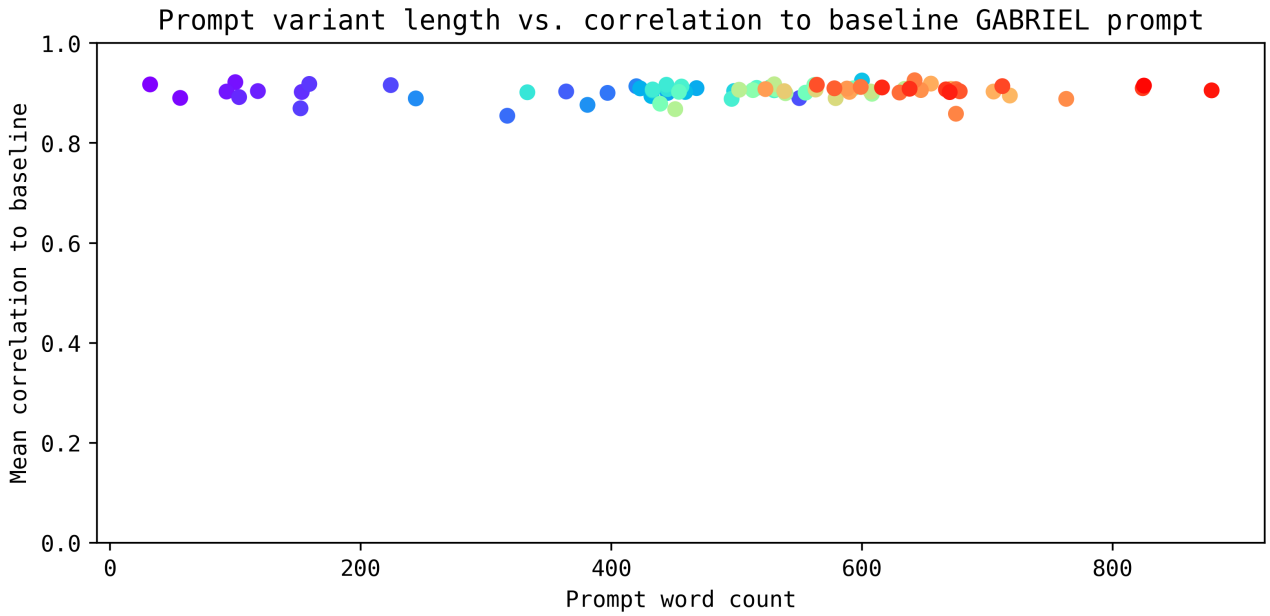


Figure 19: Very short prompts perform similarly to the GABRIEL default prompt.

We find that different prompts get very similar results. The accuracy of GABRIEL measurements as shown previously is not due to our specific prompt, but a general capability of GPT which can be utilized with many prompting approaches. More importantly, long and sophisticated prompts do not appear to add much value. More broadly, prompt engineering can matter in other task families (e.g., multi-step reasoning or tool use) (Chen et al., 2025), but in our rubric-based measurement setting we do not see systematic gains from longer or more elaborate prompts. In Figure 19, we see that our 100 prompt variants are diverse in word count, with some under 100 words. But even **very short and unsophisticated prompts result in essentially the same ratings as the baseline GABRIEL prompt**, or as other prompt variants.

Testing the brittleness of attribute definitions We also test attribute noise: when the same concept is intended, how much the specific wording of an attribute definition matters. We use the same approach as above to create 100 different definitions for each of the 10 attributes. Like before, each variant aims to reflect the same core concept to measure, but with wildly different style and wording. Each set of attribute variants are passed through the default GABRIEL prompt on the same 300 SotU snippets. Unlike before, the baseline here is the attributes with no definition provided at all, meaning GPT has to simply rely on its interpretation of concepts like `patriotic rhetoric`.

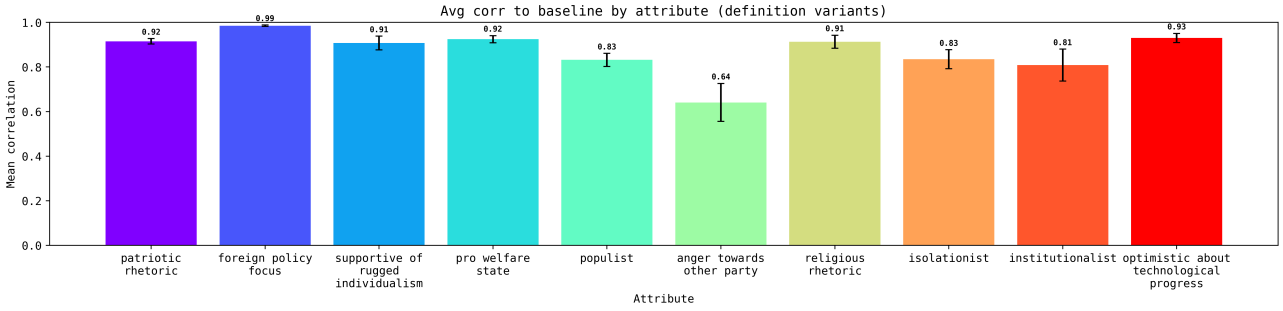


Figure 20: When the same concept is intended, noise in attribute definition wording has little effect. Anger is less consistent, but it has a low base rate.

Like prompt noise, attribute noise appears to have a minimal effect on ratings. The results indicate GPT attends to the core task at hand and core concepts to measure, and is is not meaningfully affected by format and styling. It appears that **defining attributes is not always necessary** because GPT understands the underlying concept of an attribute like `institutionalist` well. The exception here is the `anger towards other party` attribute. This is perhaps due to it having near the lowest average rating, though it may indicate that more ambiguous concepts like anger are more subject to change based on a definition. See Figure 19 in Appendix A for the baseline ratings distributions across snippets.

Are measurements overanchored to a generic concept? The prior results are a positive sign that GPT has a strong and consistent understanding of core concepts behind attributes, and is attending to concept and not the noise. But this can be problematic if these understandings are *too* anchored. It

is desirable that researchers can consistently measure the concept they intend, regardless of phrasing. It is undesirable if GPT is not measuring the concept they intend because it is anchored in a different more generic definition, where there is a subtle but important distinction.

We test this by synthetically generating 3000 tweets (using `gabriel.whatever`) that invoke pro-environment views. One third of these focus only on clean energy; another third on plants and landscapes; the final third on animals and wildlife. Using `gabriel.rate`, we measure each of four facially similar attributes on each of the three datasets. The attributes, defined in Table 9, assess `pro environment` and `pro nature` with no definitions. Separately, the same attributes are assessed but with leading definitions (`pro environment` only through the lens of clean energy, `pro nature` only through the lens of plants / landscapes).[8]

Attribute	Definition
`pro environment`	Supports the environment specifically by promoting **clean/renewable energy** or reducing fossil fuels. *Only consider this specific lens.*
`pro environment`	**No definition provided.**
`pro nature`	Values nature specifically through the lens of **plant life and landscapes** (trees, forests, flowers, wild places) and protecting them. *Only consider this specific lens.*
`pro nature`	**No definition provided.**

Table 9: Attributes measured on synthetic tweets.

Table 10: Illustrative GABRIEL ratings on synthetic short-form posts across three treatments.

Treatment	Text	pro-env (def)	pro-env (no def)	pro-nature (def)	pro-nature (no def)
Clean energy	Clean energy is the vibe—more wind, solar, and electrified rides so we can cut fossil fuels fast. Let's keep swapping gas for renewables everywhere, because cleaner power means a cleaner future.	92	93	12	71
Plants / landscapes	Give me a trail with tall pines, wildflowers, and a river cutting through the valley—instant reset. Mountains on the horizon and green everywhere just hits different, like the whole landscape is breathing.	3	42	86	90
Animals / wildlife	Just saw a hawk circling and a line of deer slipping through the trees—wildlife is pure magic. Let's give animals the space, quiet, and protection they need to thrive out there.	6	74	22	91

[8] Each of the four attributes is measured alone in its own distinct run.

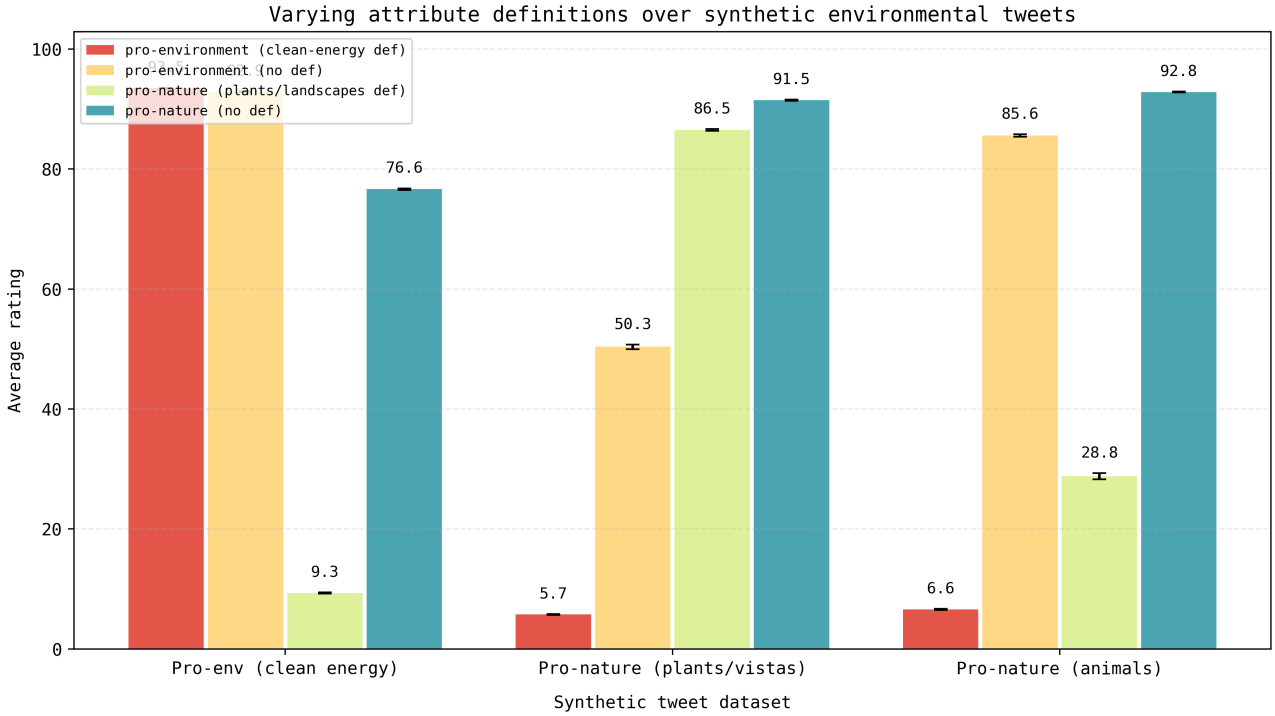


Figure 21: The intended concepts are being measured. Ratings match the specific definition.

The test shows that **when definitional differences matter, GPT correctly measures what was specifically defined rather than the generic concept**. When `pro environment` is defined to only consider clean energy, it scores highly on the synthetic clean energy tweets and near zero on the natural beauty tweets. The `pro environment` attribute with no definition does not behave this way. Even subtle but salient differences between the definitionless `pro environment` and `pro nature` attributes show up with `pro nature` not scoring as highly on the clean energy tweets (a common strife amongst environmental activists). There will always be a tradeoff between too little and too much conceptual anchoring. Too little means the model is brittle to noisy changes in wording. Too much means the researcher cannot measure the subtleties they intend. This applies to both human and GPT labelers. Our initial findings here indicate GPT is at a good balance point between these two concerns.

These results do not mean there is no potential for prompt bias. If a prompt or attribute is written in leading fashion to rate one group of data higher than another, there will be bias, just like with human labelers. However, these results show that as long as the researcher is writing a prompt with scientifically sound intention — even if short and unsophisticated — the results will likely not be worse than a "more engineered" prompt. Nevertheless, with GABRIEL we offer a standard set of validated prompts to be used to avoid any concern, and we provide guidance in Appendix B on general best practices for model usage.

4.3. The directness of LLM measurements

The previous section addressed the accuracy of measurements from an LLM. Through a battery of tests and across a range of applications, we find that LLM measurements to be highly accurate. We next address whether the measurements are direct. An LLM's output could be accurate but indirect, based on inference and not direct measurement of the concept of interest. We address two possible manifestations of this: **contamination bias** (relying on memory instead of measurement) and **shortcut inference** (relying on an alternate construct as a proxy for the one of interest).

4.3.1. Evaluation #4: contamination bias

We assess the degree to which GABRIEL's outputs are conditioned by knowing its training data. A frequent objection to LLM evaluations is the claim that the LLM has "seen and remembers" the data. For instance, if we take a dataset of Metacritic reviews off the internet and predict the score from the rating, some would claim that the reviews were in the training data and thus the model has already seen those reviews and their corresponding scores.[9]

There are two main ways of thinking about indirect measurements. The first theory is brute *label contamination*: when we ask the model to measure something about a previously labeled text, those label-text pairs are in its internal knowledge *and* are used when we ask it to label the text, as opposed to freshly comprehending it. Both conditions must be met for this to be a worry, which may mean our prior tests are not indicative of out of sample performance. This is what Sarkar and Vafa (2024) and Ludwig et al. (2025) observe in narrowly scoped situations, and it would constitute a form of indirect measurement because the model relies on contaminated internal knowledge instead of reasoning for measurement. The second theory concerns *look-ahead bias* or *shortcut inference*, where the model knows something related and gives an answer based on inference. For instance, if the model measures optimism in a 2008 earnings report it could "infer" that the company did poorly that year purely because of the year 2008, or because it knows how the stock ultimately performed in subsequent years. Either would be an indirect measurement. Here, we focus on the label contamination risk, like that elicited in Ludwig et al. (2025). The following section tackles shortcut inference. We test look-ahead bias, a form of shortcut inference, in Appendix A.4.1.

To provide empirical evidence on whether contamination bias is significant in practice, we develop a set of statistical tests to measure bleedthrough. Our method relies on staggered training cutoffs: the various models that OpenAI has released were trained with a series of different cutoffs for the latest information the model had "seen" in training. For the older models like gpt-3.5-turbo, these may be in 2021. For the latest models, the training run may have included data running through 2024.

We exploit this exogenous variation in cutoff dates to identify possible contamination bias. We can simply compare pre- versus post-cutoff performance across many models, with the knowledge that the

[9]There are structural reasons we believe this to be an unlikely source of bias, namely that contamination bias is most plausible for observations frequently repeated in the training data. We see it as unlikely that a single observation on a single spreadsheet located somewhere on the internet is likely to bleed through meaningfully to a model with limited space for internal knowledge. We see look-ahead bias as a far more salient concern.

cutoff date is arbitrary between models. If there is contamination bias, we would expect performance to decay after the training cutoff, because the models can no longer rely on the in-training knowledge they had obtained previously.[10][11]

Our analysis in Figure 22 (for top models) and Table 30 in Appendix A fails to find significant evidence of contamination bias in our use case, based on the hundreds of HuggingFace datasets. None of the models show a statistically significant difference in mean accuracy before versus after their respective cutoff. The same holds for F1 (overall retrieval performance) scores.

On balance, performance is essentially identical pre- versus post-training cutoff. This suggests that the magnitude of contamination bias is likely to be relatively small. Indeed, the table clearly shows that the pre-cutoff and post-cutoff difference in average performance is nowhere close to statistical significance across every model and every staggered cutoff; and the practical difference in pre versus post average performance is very small regardless (in the range of 0.01-0.02 across all of the models).

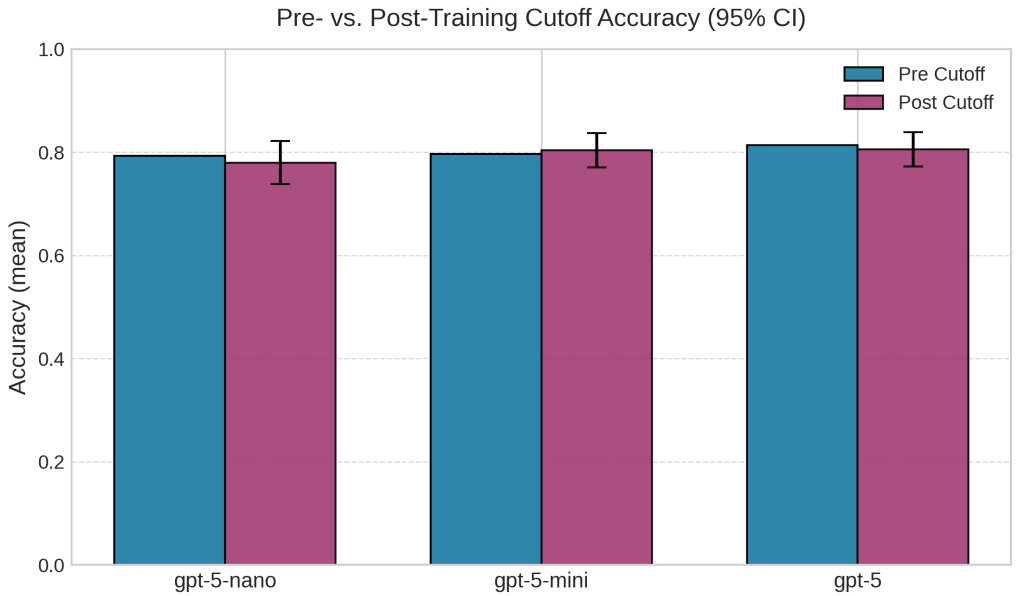


Figure 22: Models perform identically at labeling HuggingFace datasets released before vs.after the training cutoffs in mid-2024. Tiny models perform as well as big models. This evidence is inconsistent with contamination being the primary driver of accuracy in these tasks.

This finding is particularly notable insofar as some of the models entail more risk of contamination than others. Compared to small distilled models like gpt-5-mini, the far larger base models like gpt-4.1

[10] For our evaluation, we obtain the first date of posting of each HuggingFace dataset in our large-scale compiled set from the previous section, and use this as the relevant date to evaluate. We compare the dataset posting date to the training cutoff date for each model to assess whether the data was in-training. For instance, a dataset posted on June 10th, 2023 would be in-training for a model cutoff of December 31st, 2023 but out-of-sample for a model cutoff of December 31st, 2022.

[11] Strictly speaking, it is possible that the dataset was posted elsewhere prior to the cutoff date and therefore that the HuggingFace date is a later bound on inclusion on the Internet. While possible, we believe it is unlikely this would materially affect the results of the test based on manual inspection of sample datasets in which we did not find widespread evidence of easily accessible pre-posted versions of the datasets.

or gpt-5 have much more information stored inside the model, which could in theory be used for regurgitation. The fact that neither these large models nor the relatively low-information small models exhibit contamination trends suggests that the labels are indeed a product of reasoning and not of regurgitation.

This is further supported by the evidence from Metacritic reviews seen previously in Figure 9. There is no evidence of any performance dropoff following the training cutoff for gpt-5, consistent with our more systematic analysis in this section. Our analysis of the out-of-sample credit score work also substantiates this.

The results provide large-scale quantitative evidence that contamination bias is a low risk. It could potentially be triggered with highly engineered prompting, but reasonably standard use shows no evidence of contamination bias. Our results suggest that the kinds of regurgitation documented elsewhere like in Sarkar and Vafa (2024) and Ludwig et al. (2025) may be more likely under elicitation settings designed to trigger recall than under standard measurement prompting.

4.3.2. Evaluation #5: shortcut inference

The final concern is shortcut inference. By this, we mean the potential validity concern of GPT not actually measuring the attribute of interest, but rather inferring it from other attributes. For example, if GABRIEL is given a speech by a Labour MP in British Parliament and asked to rate how pro-environment the speech is, we do not want GABRIEL to give a higher rating simply because the speaker is liberal, or because other liberal opinions are mentioned in the speech. This would be shortcut inference. We want GABRIEL to directly measure environmentalism in the speech; we do not want this rating interpolated or affected by other attributes, like general political lean.

Here, we test whether shortcut inference is a legitimate concern in practice. We also propose an econometric approach to debiasing GPT measurements, which should work with any text data.

Testing shortcut inference on synthetic speeches Here we look at political content. We want to confirm that if GPT reads a Labour MP campaign speech and is asked to measure how pro-environment the speech is, we are measuring the true pro-environment signal, not an inferred signal from general left wing content. We create an empirical test for this by using GPT to generate 1000 synthetic (i.e. fake, LLM-generated) campaign speeches for Labour MPs. These base speeches are standard, except GPT is told not to include any environment related content in them. Separately we generate 1000 pro-environment paragraphs to append to the base speeches. The prompts used to make the speeches are in Table 31, in Appendix A.

The key idea is to test shortcut inference using *signal stripping*. A speech may have environmental content, and other political content from which environmental content could be inferred. Rather than trying to mask any number of sources of shortcut inference, we instead *remove the signal* (all environmental content) while holding the rest fixed.

We then use GABRIEL to rate the `pro environment` and `left wing` attributes on the base, environ-

ment stripped speeches. We separately rate the composite speeches which include the pro-environment paragraphs. If there is shortcut inference, we should see non-zero `pro environment` ratings even on the base, environment stripped speeches. This would be because GPT is inferring pro-environmental positions from context. The definitions for the attributes are in Table 11. Figure 45 in Appendix A shows a detailed flowchart for this methodology.

Attribute	Definition
`pro environment`	Politician expresses a clearly pro-environmental stance within the speech.
`left wing`	Politician expresses left wing ideas and policies (by UK standards) within the speech.

Table 11: Attributes measured by GABRIEL in synthetic speeches.

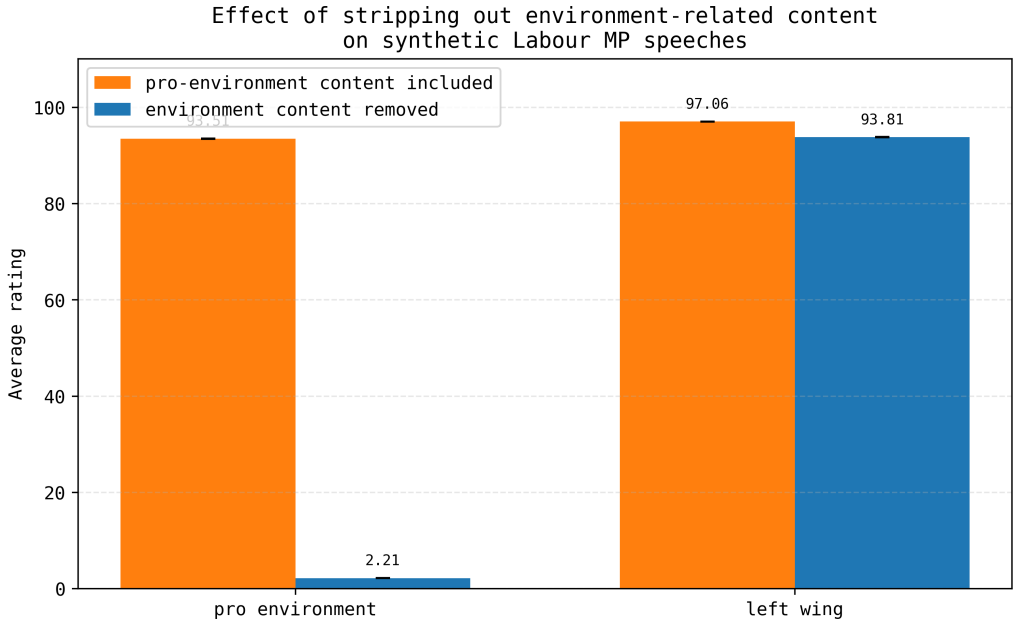


Figure 23: We find little evidence of shortcut inference. When the signal is removed (environmental content), GABRIEL ratings of `pro environment` almost fully attenuate (98%).

In Figure 23, we observe no apparent shortcut inference. When environmental content is stripped / not included in the speech, the `pro environment` attribute rating almost completely attenuates (**a 98% reduction**). Importantly, the `left wing` attribute barely shifts, indicating the modification did not alter the context where inference could be drawn from.

From this experiment, shortcut inference appears minimal. Using synthetic data allows the cleanest proof of this. The results from Figure 21 support this finding when even easier shortcuts are available. Next, we test real data that would be very high risk (in theory) for shortcut inference.

Testing shortcut inference on real data As detailed in Section 3.3, we employ GPT with web searching to create reports on specific county level characteristics. Here we deploy this method to

tabulate local regulations of business in each county. One instance of GPT is deployed to one county, and sets about scouring the internet from local government websites to Chambers of Commerce and beyond. Then it pens a report on every piece of locally enforced business regulation it finds. We take consolidated versions of all these county reports and use `gabriel.rate` on many attributes, including the `restrictive environmental regulation` attribute.

In theory, this test case should be among the most likely candidates for shortcut inference. There are many smaller counties where detailed regulations might be harder to find. At the same time, GPT knows much about each of these counties, and could easily infer an attribute like `restrictive environmental regulation` from its knowledge of the county's general political lean and demographic character (this attribute is strongly correlated with political lean).

We approach our shortcut inference test like before by removing the signal. We use GPT to find and excise all environmental content (with `gabriel.codify`), leaving everything else untouched. We are essentially making all reports have absolutely no signal on how that locality regulates environmental issues, while leaving in powerful political identifiers like the county name and how everything else is regulated. We can remeasure the same attributes on these signal stripped reports. If GABRIEL is acting as we wish, we should observe the `restrictive environmental regulation` attribute rating attenuate to zero across the country, as there is no environment related signal remaining in the reports to directly measure.

Attribute	**Definition**
`restrictive environmental regulation`	Local rules impose substantial environmental constraints on business activity (beyond typical state or federal baseline), such as strict waste disposal requirements, stormwater controls, limits on certain materials (like single use plastics), recycling or compost mandates, idling limits, or green building requirements for commercial properties. More of this shows up as detailed ordinances, extra permits, and strong enforcement language aimed at business compliance.
`high local business taxes`	Local government imposes substantial business tax burdens (business license taxes, gross receipts taxes, local sales tax add ons, special assessment districts, high permit and impact fees, or recurring annual fees). More of this shows up as fee schedules with multiple recurring charges and revenue programs explicitly targeted at business activity.

Table 12: Attributes measured by GABRIEL in local business regulation reports.

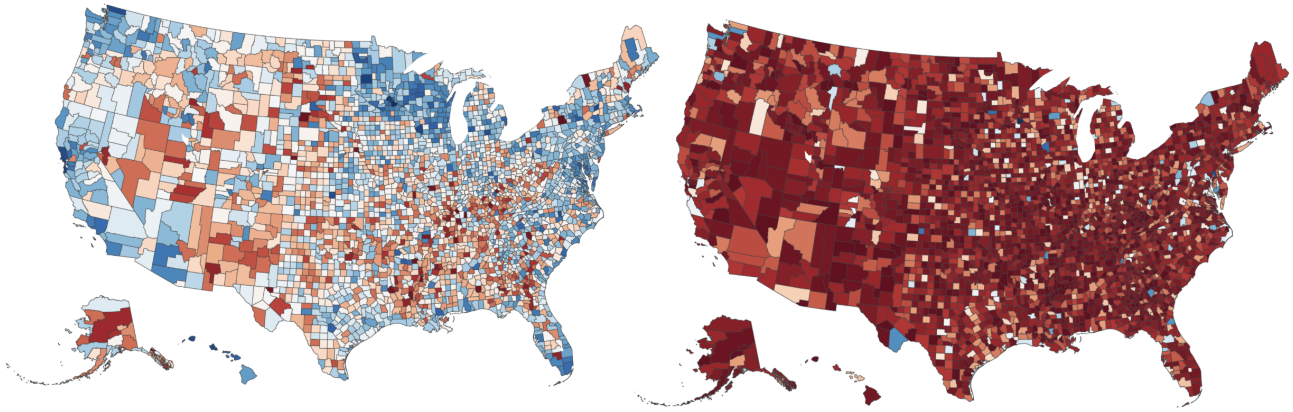


Figure 24: Maps of `restrictive environmental regulation`. Left is original; right is after signal stripping. Removing environmental content attenuates the ratings by 81%.

We observe substantial attenuation in Figure 24. We see ratings drop toward zero even in areas that previously had high signal. This evidence supports that GPT is doing what it is asked to do: it is directly measuring the attribute in front of it. If that attribute is removed but could still easily be inferred, it is not inferring it because we have asked it to measure the text only. Importantly, we can see in Figure 25 that while signal stripping for environment attenuates the `restrictive environmental regulation` attribute, it has little effect on other politically correlated attributes, like `high local business taxes`. This indicates our environment stripping is precise, and only removing text related to environment and nothing else (also confirmed with manual inspection).

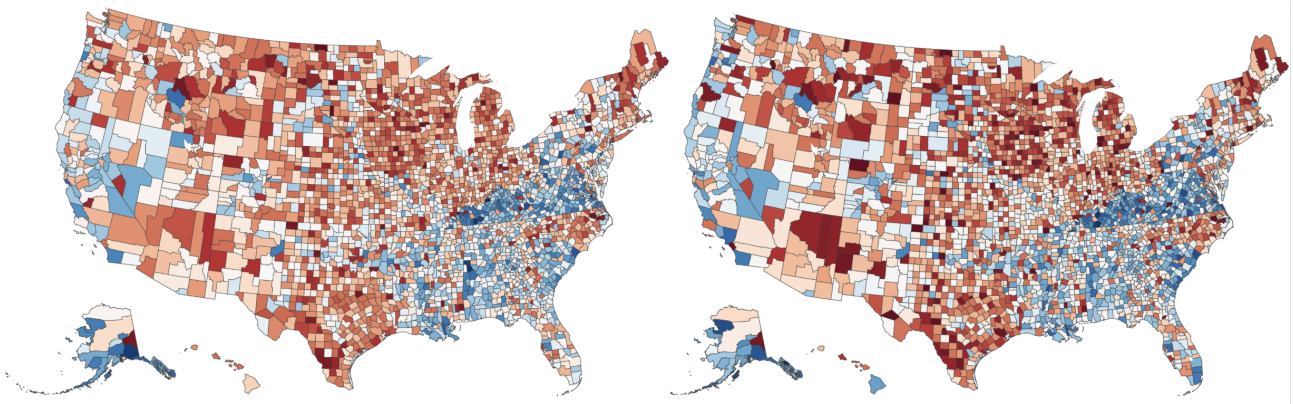


Figure 25: Maps of `high local business taxes`. Left is original; right is after signal stripping. Removing environmental content does not affect non-environmental attributes (<10% change).

Similarly, if we simply anonymize the county name and any descriptors in the reports that might give away the county, we see no effect on the ratings. This is consistent with the ratings being driven by text content rather than by county identity.

We prefer the signal stripping approach to this anonymization method. There could be any number of sources for shortcut inference — it could be the county, it could also be other political aspects of the

report. By stripping the signal itself (environmental content) rather than the potential sources of bias, we can address the full scope of shortcut inference rather than any one specific possible source.

Why don't the ratings completely attenuate? Looking closely at the environment stripped ratings map from earlier, attenuation is substantial but not complete (81% reduction), unlike the synthetic experiment (98% reduction). Ratings do not fully reach zero, and there is still a correlation (albeit weak) between the original ratings and the environment stripped ratings. This could be due to shortcut inference, but it could also simply mean the signal stripping process was not complete. Examining the data, we see that the environment stripping process was imperfect. This would explain the difference from the synthetic experiments, where the synthetic data generation process allows more explicit control to ensure all environment content was removed. For that reason, we view the synthetic results as a cleaner signal on the existence of shortcut inference.

But we can still account for this imprecision in real data signal stripping with a separate GABRIEL run to measure a much simpler attribute — `prevalence of environmental content`. This attribute should tell us whether environment related content persisted after signal stripping. We find that it does, and that much of the remaining `restrictive environmental regulation` non-zero ratings after environment stripping is explained by this `prevalence of environmental content` attribute. In other words, imperfect signal stripping explains much of why the ratings did not fully attenuate to the degree of the controlled synthetic experiment. We propose next an econometric approach to creating debiased ratings in real data, factoring in imperfect signal stripping.

An empirical approach to debiased GPT ratings For each document i, we measure an attribute twice: (i) on the original text, yielding an original rating y_i, and (ii) on a signal-stripped version of the same text, yielding a stripped rating s_i.[12]

We conceptualize the original rating as the sum of a direct-signal component and an inference component:

$$y_i = t_i + b_i + \varepsilon_i, \tag{1}$$

where t_i is the desired direct measurement of the signal present in the text, and b_i captures any inference-based component that the model might measure from contextual cues rather than direct signal.

If stripping were perfect, the stripped rating would isolate the inference component ($s_i \approx b_i$), and a natural debiased estimator is the difference:

$$\widehat{t}_i^{\text{diff}} = y_i - s_i. \tag{2}$$

[12]We confirm linearity applies to the signal stripping process in fig. 48. We do this by using the `gabriel.codify` function to perform the stripping, by first identifying all snippets of text from the underlying documents that are related to the signal. We can then remove a varying percentage of those snippets randomly. We see that as the percentage of snippets removed increases, measured ratings consistently decline. Establishing linearity is important because of the possibility that shortcut inference only occurs at higher levels of signal-related content.

In practice, stripping may be incomplete, so the stripped text may still contain residual signal. To quantify this, we additionally measure r_i, the prevalence of signal-related content remaining in the stripped text. We treat r_i as a proxy for the amount of leftover direct signal that survived the stripping step. We then use the relationship between s_i and r_i to estimate how much of the stripped rating is mechanically driven by remaining signal:

$$s_i = \alpha_s + \delta r_i + \eta_i, \tag{3}$$

where δ translates the remaining signal measure into units of the stripped rating, and η_i captures the part of the stripped rating not explained by remaining signal.[13]

We estimate δ by OLS in the equation above and define the inferred (non-signal) component for document i as the portion of the stripped rating not attributable to remaining signal:

$$\widehat{b}_i = s_i - \widehat{\delta}\, r_i. \tag{4}$$

Finally, we define the debiased estimate of the direct signal as:

$$\widehat{t}_i = y_i - \widehat{b}_i = y_i - s_i + \widehat{\delta}\, r_i. \tag{5}$$

When stripping is close to complete (so r_i is near zero), this reduces to the simple difference estimator $\widehat{t}_i = y_i - s_i$.

Debiasing results on our data We find that the debiased rating strongly correlates with the original rating — see Figure 26 (regression table is Table 32 in Appendix A). The shift is modest and the explanatory power is high. This is true whether or not we include the `remaining content` variable; excluding it leads to a slightly lower R^2 of 0.68. The R^2 here of 0.86 is as high as the R^2 between the signal stripped and original `high local business taxes` ratings (also 0.86). Recall that this was the control attribute meant to be unaffected by environment stripping, implying a ceiling on how consistent the ratings can be here. The debiased ratings show no practical difference from the original ratings and do not indicate a significant shortcut inference problem at hand.

Takeaways All our methods described in this shortcut inference section are accessible using the `gabriel.debias` function on any text data, to ensure debiased measurements. But our initial finding is that debiasing may be unnecessary: we observe little evidence of shortcut inference in our tests. Our main result is on synthetic data, which allows us to eliminate the complication of incomplete signal stripping. In these experiments, ratings almost entirely attenuate and are not inferred from context when the signal is removed. Our debiasing method is more compatible with real data and shows a similar lack of shortcut inference. We also test the related concern of look-ahead bias (GPT's foreknowledge of outcomes affecting measurements) in Appendix A.4.1 (Sarkar and Vafa, 2024). In

[13]This method is vulnerable if higher remaining signal correlates with with high true signal (which is likely) *and* high true signal is particularly vulnerable to shortcut inference (which is unlikely, given our linearity findings in fig. 48).

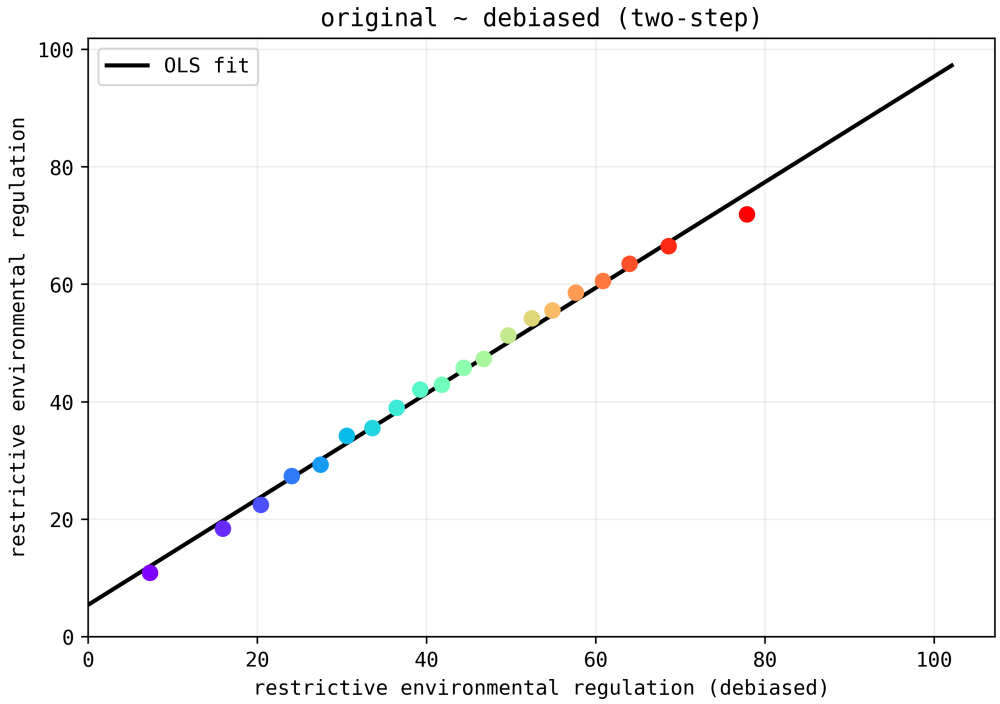


Figure 26: Regressing the debiased ratings against the original. There is little change; there appears to be little to no shortcut inference in practice. Correlation is important because most applications rely most on the relative ordering of entity ratings, not the raw value.

that test, we similarly do not observe shortcut inference.

We do not prove here that shortcut inference bias is zero, and low power statistical methods are sensitive to even small biases. Yet our evidence suggests that the theoretical problem of shortcut inference is not a significant problem in practice. GPT appears to act as intended: comprehending a text and measuring the signal it is tasked with measuring, not inferring a measurement from context.

5. Applying GABRIEL to study the history of technology adoption

Here we cover a full usage example to motivate what is possible for a social science researcher today by using in conjunction many LLM methods like those GABRIEL offers.

How much time does it take for technology to be adopted? This is a challenging question to answer empirically. There is no comprehensive dataset of all technologies with their lags from invention to adoption or their core characteristics. It is difficult to understand what makes a technology slow or fast to be adopted, or how the lags to adoption have changed over time *at scale*.

A large diffusion literature shows that adoption is typically gradual and S–shaped, reflecting complementary investments and organizational change rather than invention alone (Griliches, 1957; Mansfield, 1961; Rogers, 2003). Work on general-purpose technologies argues that systems like electrification and computing require substantial reorganization before productivity gains are visible, producing

long adjustment paths (Bresnahan and Trajtenberg, 1995; David, 1990; Gordon, 2000; Brynjolfsson et al., 2017). Cross-country studies measure adoption timing and document wide dispersion across places and periods, but available datasets cover a limited set of technologies and provide relatively few comparable attributes for explaining why some diffuse faster than others (Comin and Hobijn, 2010; Comin and Mestieri, 2018). This gap motivates our approach: a large-scale, attribute-rich new dataset that identifies invention to adoption time lags and technology characteristics for systematic analysis.

In this section, we leverage a sequencing of many GABRIEL tools to create this novel dataset with tens of thousands of industrial age technologies. For each one, we extract a series of data including invention and widespread adoption dates; inventor and location; technology class and various attributes of the product itself. With this, we show how GABRIEL can measure new data usable in a real research setting.

5.1. Methods

To create our dataset of historical technologies, we begin with the ~18 million set of all article titles in the English Wikipedia. Wikipedia is a good source for extracting the names of historical technologies. It is a far more massive encyclopedia than any other. This means it more thoroughly covers every time period and every category of technology. In addition, Wikipedia links to each technology are unique, helping to prevent duplicate technologies with slightly altered names from appearing in the dataset. The existence of a Wikipedia page on a technology is itself a starting indicator that the technology is not completely obscure. We discuss validity of Wikipedia as a data source in Appendix C.

Our method involves a succession of GABRIEL steps — a fully GPT based pipeline. We filter the initial articles, categorize them, identify important technologies, and extract information about them. Figure 27 shows the core steps, and the full method is described below. These steps narrow us from ~18 million initial candidate articles to ~25k industrial age technologies, and populate each with numerous extracted and measured characteristics. This is the dataset on which our results are based.

Initial filtering We first pass all 18M Wikipedia article titles through `gabriel.filter`, a high throughput classification filter for massive datasets. This method sends 500 random article titles to each individual GPT call (gpt-5-nano). GPT outputs the subset of those titles which are historically significant technologies (the specific "historically significant technology" filter condition is in Table 35 in Appendix A). The filter call is repeated three times, and any title which is classified as a technology at least once moves on to the next round of more rigorous screening. Approximately 1M article titles meet this criterion.

Next, we simply repeat the same exact filter step on these 1M article titles. The smaller population allows for a more detailed filtering, at a higher standard. This time, only 100 random titles go to each GPT call (since a higher fraction are likely to be technologies), and this filtering is rerun 11 times. Article titles classified as historically significant technologies more than 50% of the time (at least 6 out of 11) move on. After some programmatic string deduplication, this leaves us with ~200k candidate technologies.

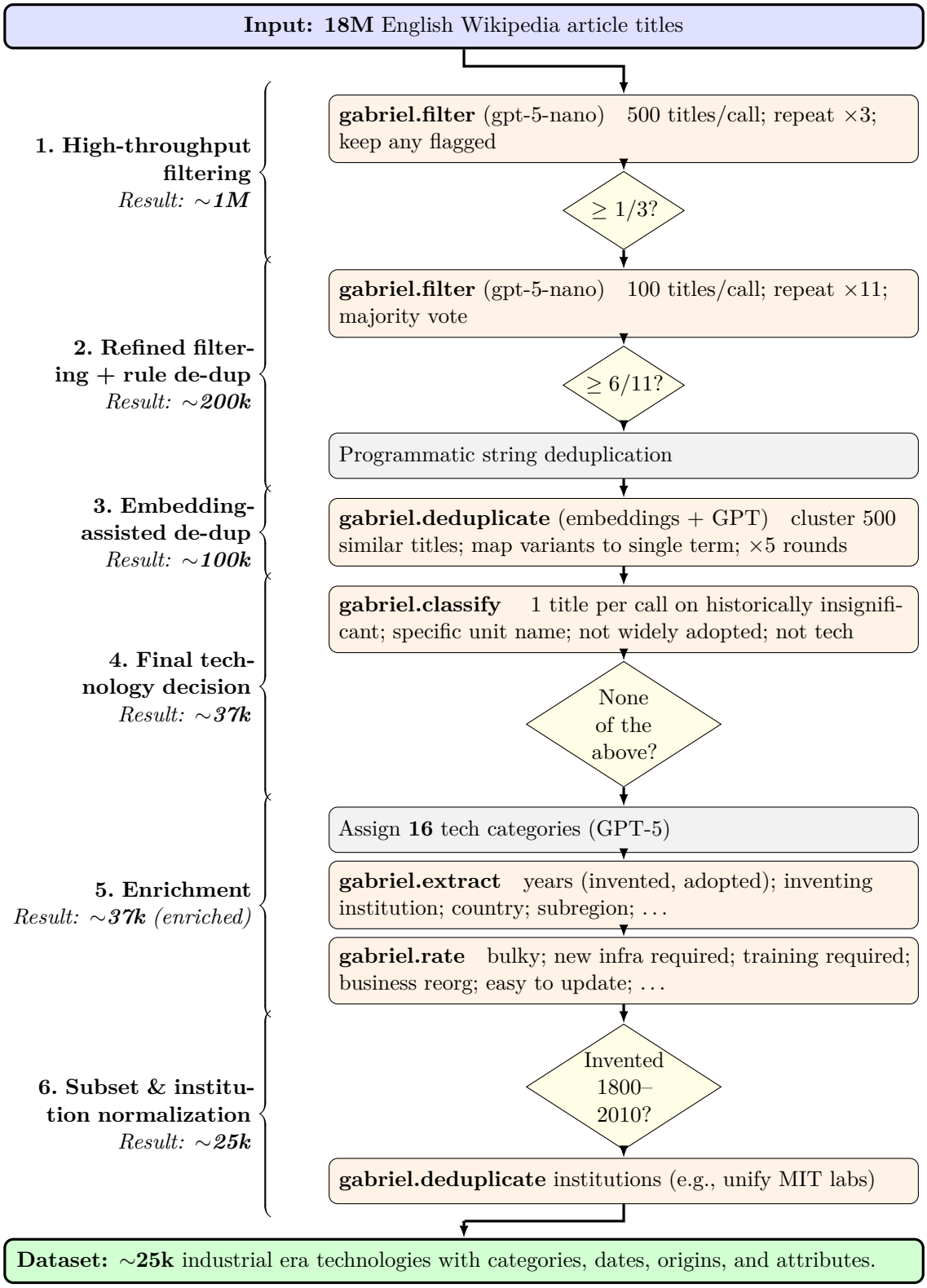


Figure 27: The GABRIEL pipeline used to assemble the historical technology dataset. Left braces denote stages and the number of candidates after each stage; right column shows the detailed methods.

Deduplication The next step is to perform a more intelligent and thorough deduplication using the `gabriel.deduplicate` method. While Wikipedia helps to avoid duplication (since riffs on the same concept map to a single article title), duplicates still sneak through. Examples include different variants of the F-16 fighter jet, or different iterations of the iPhone. Our method works by first using embeddings similarity (vector representations of the conceptual meaning behind sentences) to create clusters, each containing 500 conceptually similar technology candidates. Each cluster of 500 is passed to its own GPT call, which is instructed to map all slight variants of the same technology to their best representative term, also from the same list. We repeat this for five rounds of consecutive deduplication. This results in ~100k unique candidate technologies.

Precise filtering After this, we have a small enough dataset to cost-effectively use `gabriel.classify`. This method passes a single technology candidate to a GPT call to classify whether it meets certain conditions. It is thus relatively more expensive than `gabriel.filter`, but allows the model to devote much more attention to each individual candidate observation. The classification labels are "historically significant", "specific product edition", "specific unit" (e.g. a specific battleship), "widely adopted by target audience" (which can be quite small, in the case of something like EUV lithography), and "not a technology". The final technology list is composed of those candidates classified as `True` for "widely adopted by target audience" and `False` for the other labels. This results in ~37k technologies.

Extracting information and characterizing each technology With filtering concluded, we next characterize the technologies using gpt-5, a larger and more knowledge-rich model. Again with `gabriel.classify`, we classify each of the 37k technologies into 16 tech categories (e.g. "computer software", "medical", "communication", "military"). We use `gabriel.extract` to get specific and consistent information about each technology. Extracted information includes "year of invention", "year of wide adoption by target audience", "institution of invention", "institution type", "country of invention", "subregion of invention", etc. We also use `gabriel.rate` to quantify certain qualitative attributes on each technology, including "large and bulky", "productive usage requires new infrastructure", "requires highly specialized training", "easy to update", and so on. All attributes we measure are defined fully in Table 36 and Table 37 in Appendix A. No other features were measured beyond those listed. We now have a final dataset of ~37k technologies with information on the tech category, years of invention and adoption, region and institution of origin, and notable attributes quantified on each technology. For our purposes, we restricted the data to ~25k industrial era technologies, invented between 1800 and 2010 and widely adopted by the present day.

Specific definitions for extracted attributes, including the years of invention and adoption, are provided in Table 34 in Appendix A. The invention year was defined as the year that technology first had a working prototype / the year the tech is commonly agreed to have been invented. Adoption year is more complicated. Our definition uses the year that marks widespread adoption by the technology's target audience. This could mean wide adoption by Americans for the bicycle, or wide adoption by semiconductor foundries for EUV lithography machines. This way, we are capturing true adoption and regularized use of the technology, which in most cases would not be used by most of the population.

Human labeling to extract these datapoints for 37k technologies would be infeasible. Most people do not know almost any of these facts and would need to spend dozens of minutes of internet research to procure them. GPT holds much of the required information inside its internal knowledge, something impossible with humans. In our small scale manual validation tests, we found GPT extracted years to be accurate and general GPT knowledge about the technologies to be robust, without observed hallucinations. While there is inherent uncertainty in a `year of wide adoption by target audience` for a technology, the GPT measurements appear accurate based on manual verification on a sample. The years when many productivity enhancing technologies are recorded as becoming widely adopted correlates positively with GDP growth. There is no such correlation if the adoption years are lagged by 10 years or increased by 10 years. This provides some evidence that the adoption year measurements are relatively precise.

Attribute	Definition
`year of invention`	The singular, precise year that this technology was first invented or the first working prototype was built. The year of first prototype/invention is the year that the first working prototype of this technology was created (NOT some unrealistic concept work but a working, legitimate prototype); it will usually be the same year that this technology is commonly agreed to have been invented. The prototype may have been subpar in performance and capability, and it may not have been publicly revealed at the time. Careful consideration of the internal development and earliest versions of prototypes must be done to ensure you are isolating the true earliest working prototype and not just an early product. This is usually NOT the year of the first product release, but the year of the invention/first working prototype which comes before, sometimes well before. Report the precise year of invention alone (no rounding).
`year of wide adoption by target audience`	The single, specific year that marks widespread adoption and use of this technology by its target audience. By this year, the tech has clearly become widely adopted and accepted as normal by its target audience, and it has reached a level of polish sufficient for it to be genuinely useful and productively beneficial to users in its intended role. Importantly, the technology could still grow and improve after this year (e.g., the iPhone was widely adopted by 2012 even though it improved iteratively afterwards). This is the year it was complete and functional enough to have merited AND succeeded at acquiring mass adoption by its target audience.
`inventor`	The full name of the person most responsible for inventing this technology. Report the inventor name alone.
`institution of invention`	The institution where this technology was first invented. Report the institution name alone. Report "unknown" if you do not know, or if there was no specific institution.
`country of invention`	The country within which this technology was first invented (NOT the birthplace of the inventor, but where the invention was first made). Report the country name in standard modern English ("United States," "China," "United Kingdom," "Japan," etc.), even if the country went by a different name at the time of invention.

Table 13: Extracted attributes for each technology. All remaining attributes are in table 34.

5.2. Results

Adoption lags have shortened dramatically over the industrial age. Our key result, in Figure 28, is quantifying a secular and dramatic speed up in tech adoption lags over the industrial era. The 19th century was characterized by 40–60 year lags from when a technology was first invented/prototyped to when it was widely adopted by its target audience. By the 21st century, these

lags have reduced tenfold. Adoption lags today are on the order of five years, not fifty.[14] This is a defining shift of modernity, and it has potentially meaningful implications for how new technologies like AI might invade the economy much faster than older tech revolutions. The dynamo is not the computer, and the computer will not be AI.

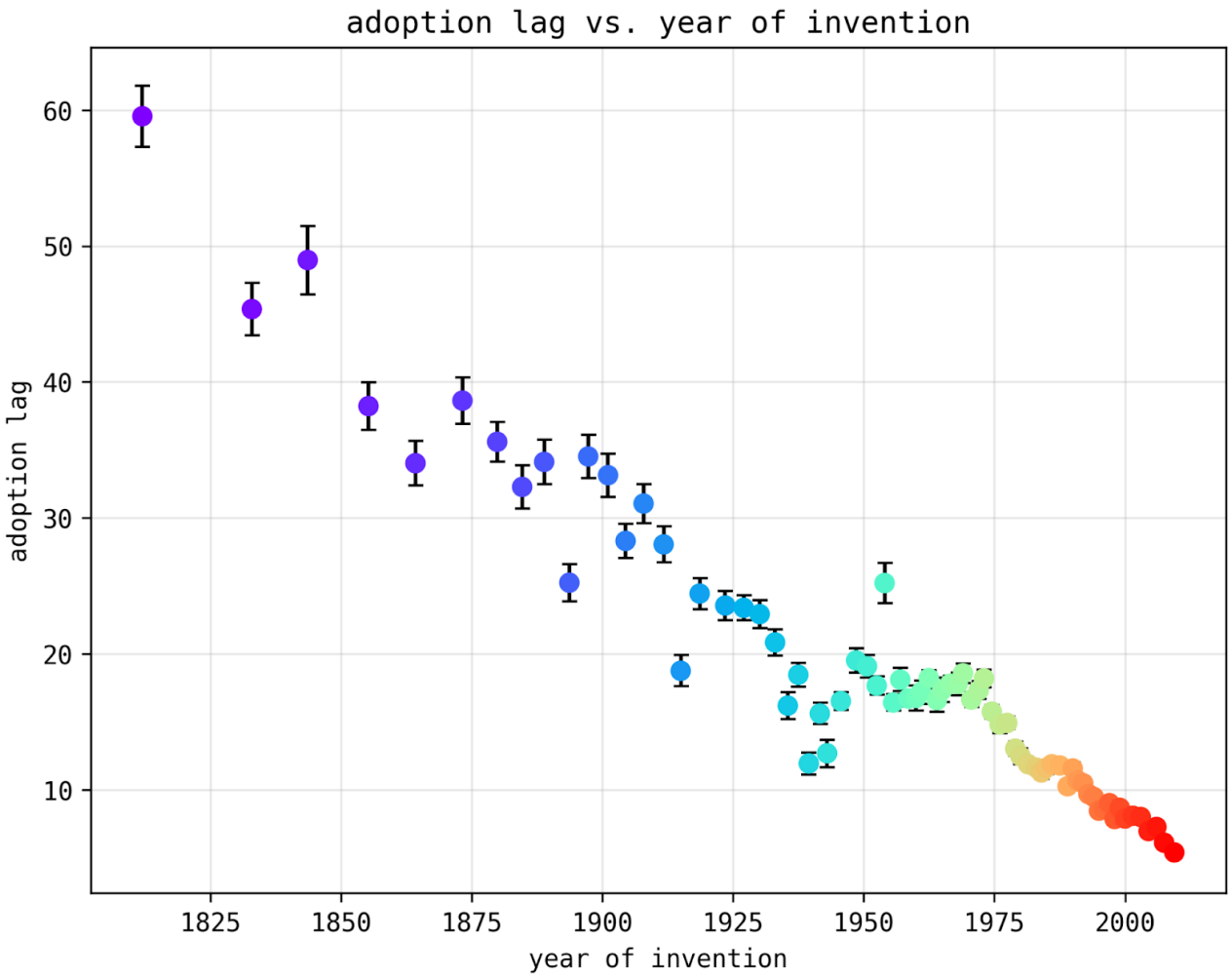


Figure 28: Time from prototyping to wide adoption of technologies has fallen tenfold over the industrial era.

Tech specific attributes explain speedier adoption. We wish to explain this secular trend, as well as the high variance between individual technologies: why does the adoption in some technologies lag while for others it does not? We can study this by examining excess lags, or how technologies behave relative to their contemporaries.[15] Table 15 shows the regression results for this analysis under various specifications.

We unexpectedly find that while computer software is a fast adoption tech category in absolute terms,

[14]We removed all technologies invented after the year 2010, to mitigate an artificial shortening of lags due to the 2025 terminus date (all techs have to have been widely adopted by the target audience at some point). This 15 year cushion

	Dependent variable: adoption lag
	(1)
year of invention	-0.224***
	(0.004)
Observations	24431
R^2	0.254
Note:	* p<0.1; ** p<0.05; *** p<0.01

Table 14: Tech adoption has sped up over time.

it is not particularly fast relative to its contemporaries in other categories. Our evidence suggests then military tech adopts the fastest, with 28% shorter adoption lags than its contemporaries — see Figure 29.

Examining the trends of tech categories over time, we find this result reflects the urgency of war: military kit devised at the beginning of a war must be made usable for that same war. We see a similar result looking at the type of institution each technology was invented at. Tech born out of the military sees the fastest adoption, and corporate tech is faster paced as well. On the other side, technology invented in academia is the slowest. The effects here are likely a combination of how much the institution cultivates speed and the selection of technologies each institution specializes in.

We now seek to understand how more nuanced, tech-specific attributes explain excess adoption lags. We wish to look beyond broad tech categories and examine the explanatory power of various attributes belonging to the technology and its development cycle. These include how large and bulky the technology is, its reliance on network effects for usefulness, how well financed its early development was, etc. Again, these attributes are specifically defined in Table 36 and Table 37.

Many of these attributes show patterns we might expect but previously have been unable to quantify. Complex supply chains, reliance on network effects, bottlenecks involving academic research, containing many parts, and so on all correlate with increased adoption lags. Competitive, geopolitically motivated, and iterative technologies move faster.

A few attributes show unexpected results. Requiring specialized training correlates with shorter lags while `easy to use` technologies correlate with longer lags. Large and bulky technologies also move somewhat faster. This may have to do with a business vs. consumer usage differential or multicollinearity; more analysis is required.

helps avoid this problem, alongside observing a long running trend with no apparent distortion at the tail.

[15] For each technology, we construct a window of 7 years around its year of invention (3 years before, 3 years after). We compute the average adoption lag across this window. We then take the ratio between the specific technology's actual lag and the surrounding window's average lag for its contemporaneous technologies. If our specific technology in question has a shorter lag than its contemporaries, we know that for some reason it was relatively fast in adoption. We can then seek to explain that relative speed or slowness. The "excess lag" approach allows us to control for the dominating secular trend. Computer software is adopted very quickly, but that could be because it is a recent technology, not because it has uniquely speedy properties.

	Dependent variable: adoption lag (ratio)		
	(1)	(2)	(3)
dramatically increased worker productivity (z)	-0.012***	-0.023***	-0.027***
	(0.005)	(0.005)	(0.005)
requires highly specialized training (z)	-0.032***	-0.025***	-0.022***
	(0.008)	(0.008)	(0.008)
large and bulky (z)	-0.047***	-0.050***	-0.054***
	(0.008)	(0.008)	(0.008)
contains many parts (z)	0.016**	0.027***	0.028***
	(0.007)	(0.007)	(0.007)
useless in early days (z)	0.195***	0.192***	0.186***
	(0.007)	(0.007)	(0.007)
intense competition to develop (z)	-0.023***	-0.016**	-0.019***
	(0.006)	(0.006)	(0.007)
inventor was highly eccentric (z)	-0.007	-0.008*	-0.013***
	(0.004)	(0.004)	(0.004)
long time to install (z)	0.049***	0.043***	0.040***
	(0.009)	(0.009)	(0.009)
expensive (z)	0.104***	0.114***	0.118***
	(0.009)	(0.009)	(0.009)
productive usage requires business reorganization (z)	0.073***	0.058***	0.055***
	(0.007)	(0.007)	(0.007)
productive usage requires new infrastructure (z)	-0.069***	-0.079***	-0.077***
	(0.009)	(0.009)	(0.009)
easy to update (z)	-0.014**	-0.011*	-0.013**
	(0.006)	(0.006)	(0.006)
high fixed costs in product development (z)	0.107***	0.095***	0.099***
	(0.010)	(0.010)	(0.010)
reliant on network effects (z)	0.009*	0.008	0.012**
	(0.005)	(0.005)	(0.005)
early mass manufacturing challenges (z)	0.032***	0.031***	0.021**
	(0.008)	(0.008)	(0.008)
well financed (z)	-0.099***	-0.096***	-0.074***
	(0.008)	(0.008)	(0.008)
strong geopolitical incentives (z)	-0.077***	-0.041***	-0.036***
	(0.006)	(0.008)	(0.008)
reliant on academic research (z)	0.163***	0.146***	0.143***
	(0.006)	(0.006)	(0.007)
easy to use (z)	0.077***	0.078***	0.081***
	(0.006)	(0.006)	(0.007)
widespread excitement by target audience (z)	-0.127***	-0.127***	-0.131***
	(0.005)	(0.005)	(0.005)
reliant on complex supply chains (z)	0.054***	0.073***	0.079***
	(0.009)	(0.010)	(0.010)
largely iterative (z)	-0.037***	-0.035***	-0.030***
	(0.005)	(0.005)	(0.005)
Observations	24404	24026	23492
R^2	0.203	0.220	0.228
primary category	-	✓	✓
institution type	-	-	✓
Note:			* p<0.1; ** p<0.05; *** p<0.01

Table 15: Tech specific qualitative attributes explain why technologies reach adoption slower or faster than their contemporaries. A lower ratio means speedier adoption than contemporary technologies.

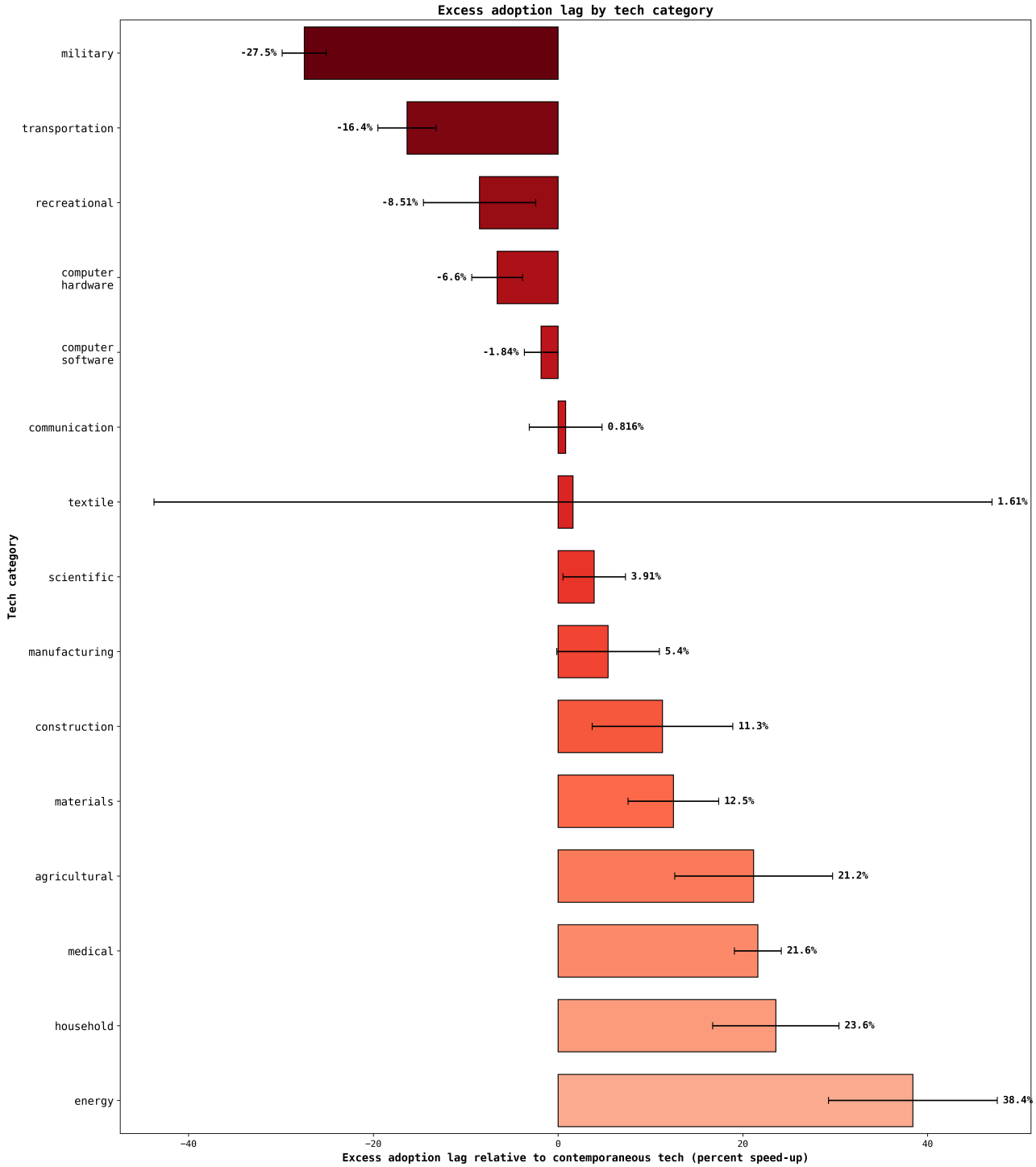


Figure 29: Though there is wide variance, some classes of tech (like military kit) are adopted much faster than others.

These results are not causal and indeed some outcomes we measure are plausibly endogenous. Here, we focus mainly on an initial descriptive attempt to describe these trends at large scale.

Put together, these attributes (which are the totality of those measured in this experiment) explain

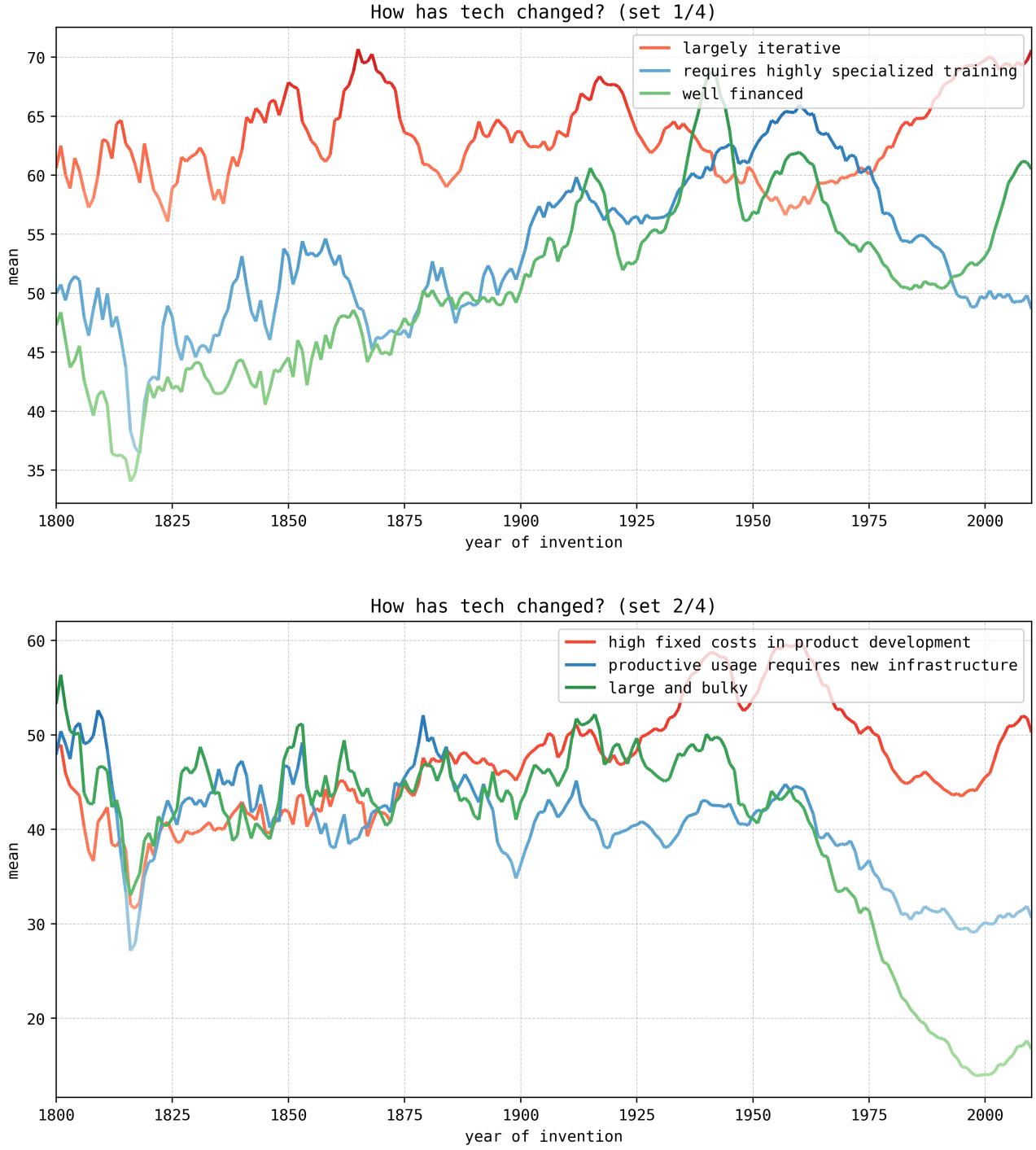
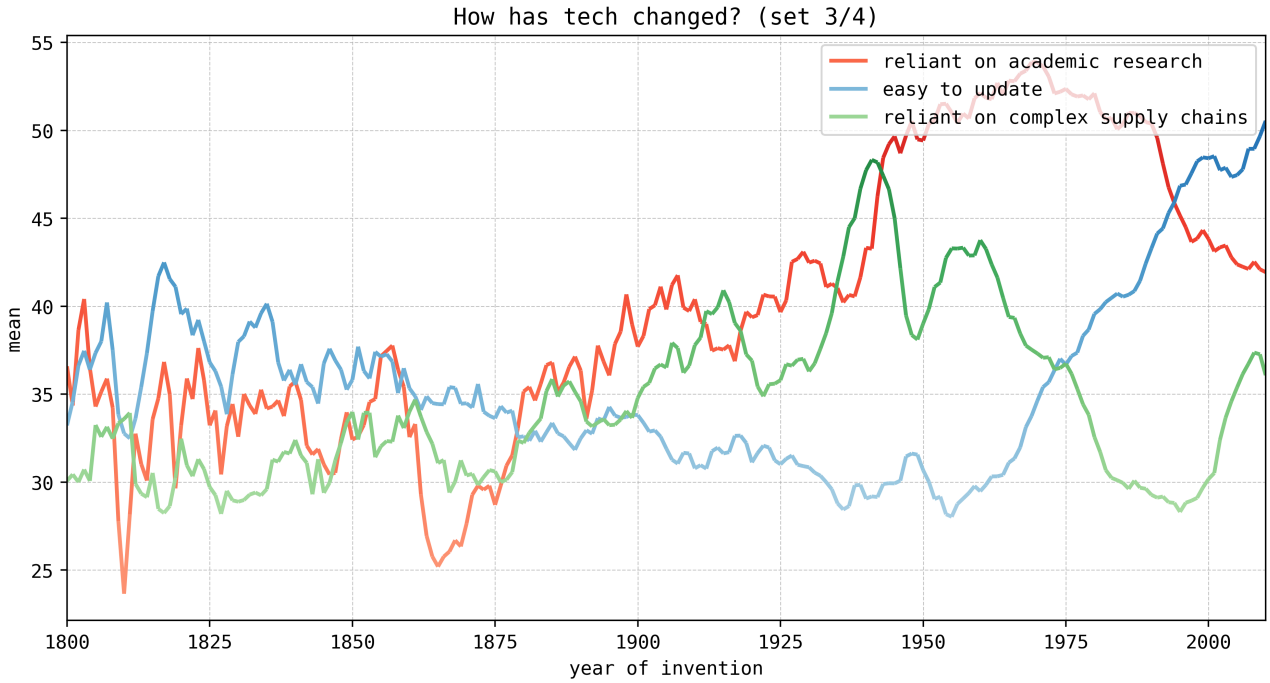


Figure 30: The nature of technologies has evolved greatly over time. The requirement of specialized training peaked in the 1960s and has fallen in the computer age. New tech has gotten smaller and less dependent on new infrastructure.



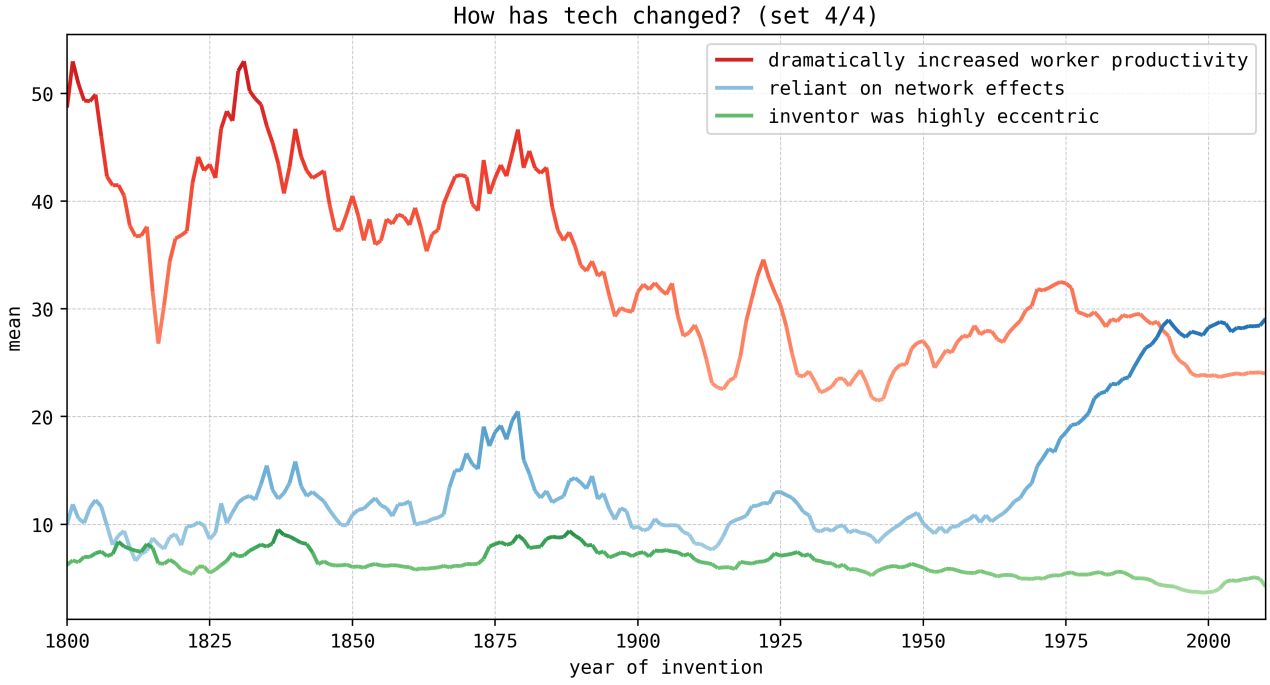


Figure 31: Tech is easier to update and iterate. Academic research maintains importance in new tech but peaked in the late 20th century. Network effects matter much more for recent tech.

23% of the variation in excess adoption lags. This is a powerful result towards understanding why some technologies move slower than others. Using GABRIEL allows us to test many hypotheses quickly.

The results provide evidence contrary to our original hypothesis that AI has properties friendly to particularly rapid adoption. From these results, it appears that computer software is not faster than its contemporaneous tech; easy to use, high fixed cost, reliant on research breakthrough technologies like AI appear to move meaningfully slower. At the same time, these effects appear to be dominated by the first order phenomenon we outlined earlier. There is a massive secular decline in adoption lags leading into the present, where today's tech is widely adopted on the order of 5 to 10 years.

These results lead us to the overall takeaway of expecting relatively fast adoption of major technology today. The implications for modern technologies like artificial intelligence would be significant.

The technologies and their origins have changed greatly over time. We next look at our measurements over the industrial era, to gain a descriptive understanding of how technology has evolved in its properties and origins.

We observe the rising dominance of American innovation in the late 19th century, coinciding with the US becoming the leading economic power (Figure 32). However, unlike the GDP statistics, while America reached its peak percentage of global GDP in the early 20th century, its total dominance in innovation is a more recent phenomenon.[16] In our dataset, the share of technologies attributed to U.S. origin rises to roughly 60–70% in recent periods, even though only 20 to 30% of global GDP has been American. This highlights how GDP today may be a poor proxy for who is driving GDP growth tomorrow: innovation appears much more concentrated in America than economic activity.

Looking at subregions / states within countries: the top five invention powerhouses over industrialization are Paris and four American states — see Figure 33. California's dominance today is particularly striking: this one state is responsible for over a quarter of new technologies in our dataset. No other subregion in the world has held such an expansive role in innovation over the past two centuries. Put together, in our sample just three states — California, New York, and Massachusetts — account for more inventions than the other 47 combined.

Beyond polity, the institutions where invention happens have changed dramatically (Figure 34). The pre-industrial and early industrial inventions were mostly the creation of independent builders. But over industrialization, the homemade invention has largely disappeared (with a small comeback in the software era) while corporate invention has become dominant. We might expect independent inventors to be more capricious and less resourced in turning prototypes into products. Indeed, our earlier result showed faster adoption of corporate innovation — but without strong explanatory power.

We run a final deduplication step, which allows us to group together similar institutions — like the various labs at MIT into the single "MIT" bucket. Then, we find the most prolific inventor in history to be AT&T — better known by its historical moniker of Bell Labs — see Figure 35. Bell alone was

[16]While US bias is worth keeping in mind, the prior work and replications using other data covered in the methods section indicate there should be no strong US bias.

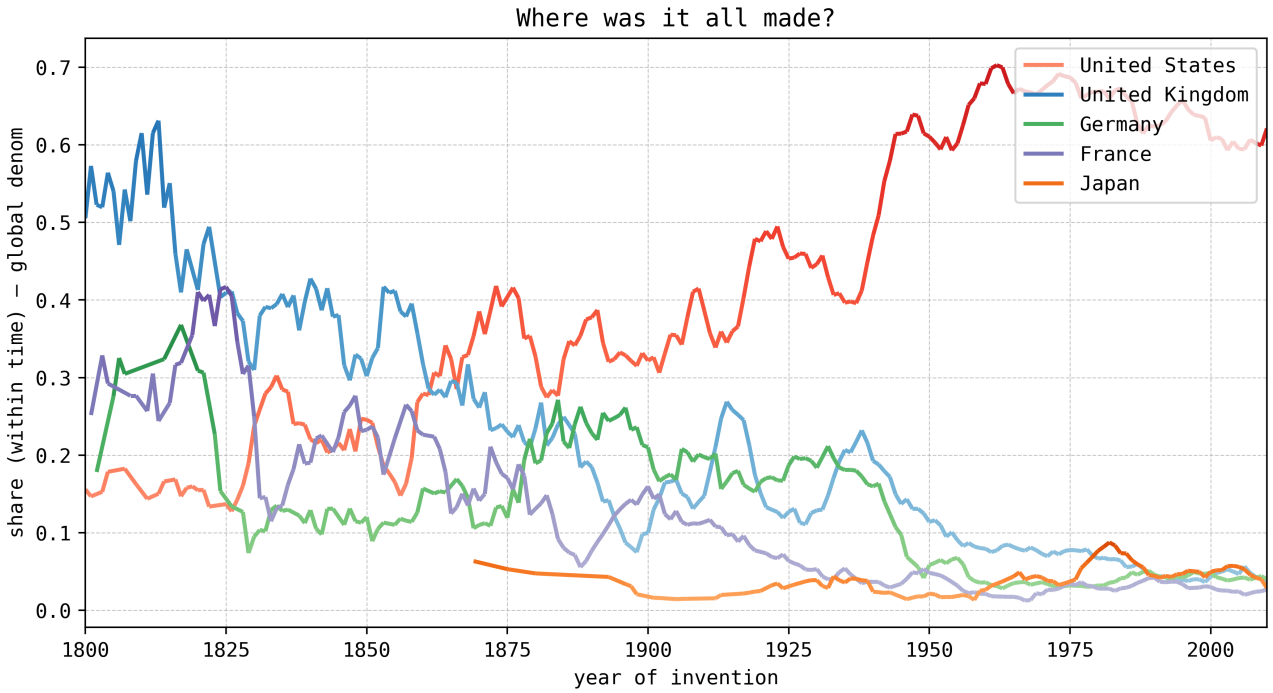


Figure 32: The countries seeding invention. The US slowly builds a dominant lead beginning in the late 19th century.

responsible for 3% of industrial era inventions that GABRIEL could attribute to an institution. IBM was a close second, and MIT led all universities. This data could offer new insight into how profitable (and potentially monopolistic) companies innovate in fields tangential to their core business.

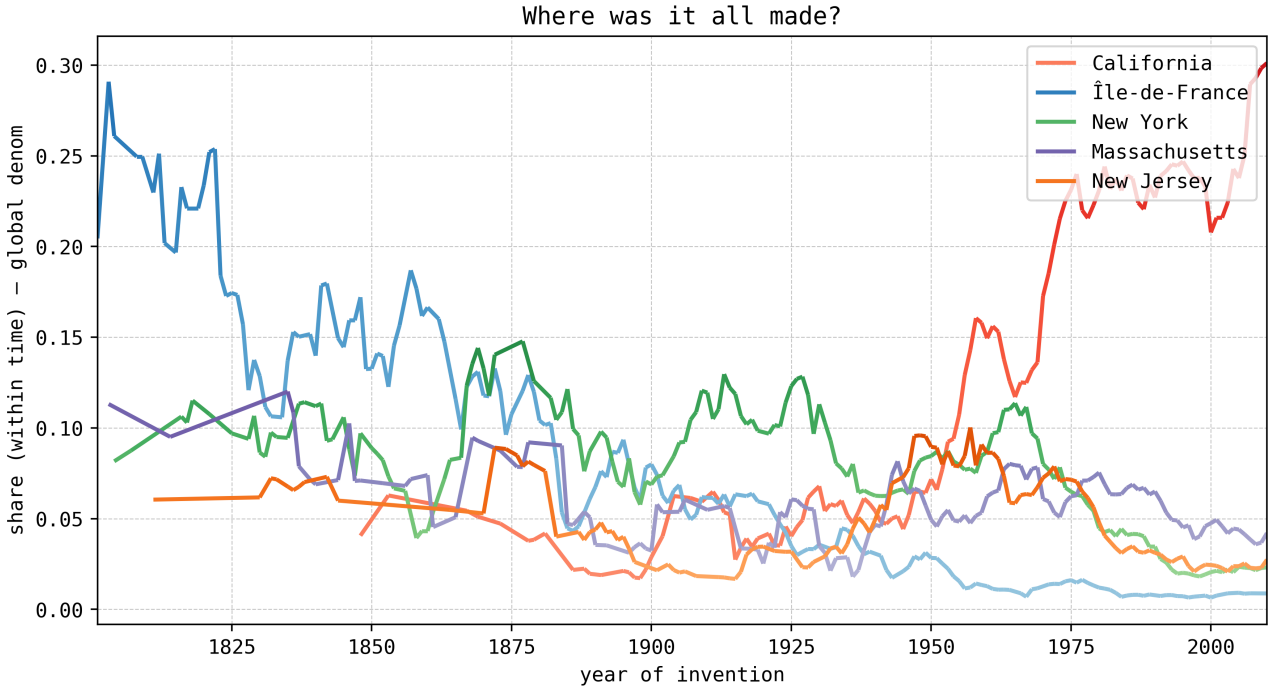


Figure 33: The states / subregions powering invention. California, New York, and Massachusetts account for more than half of all American invention over the industrial era.

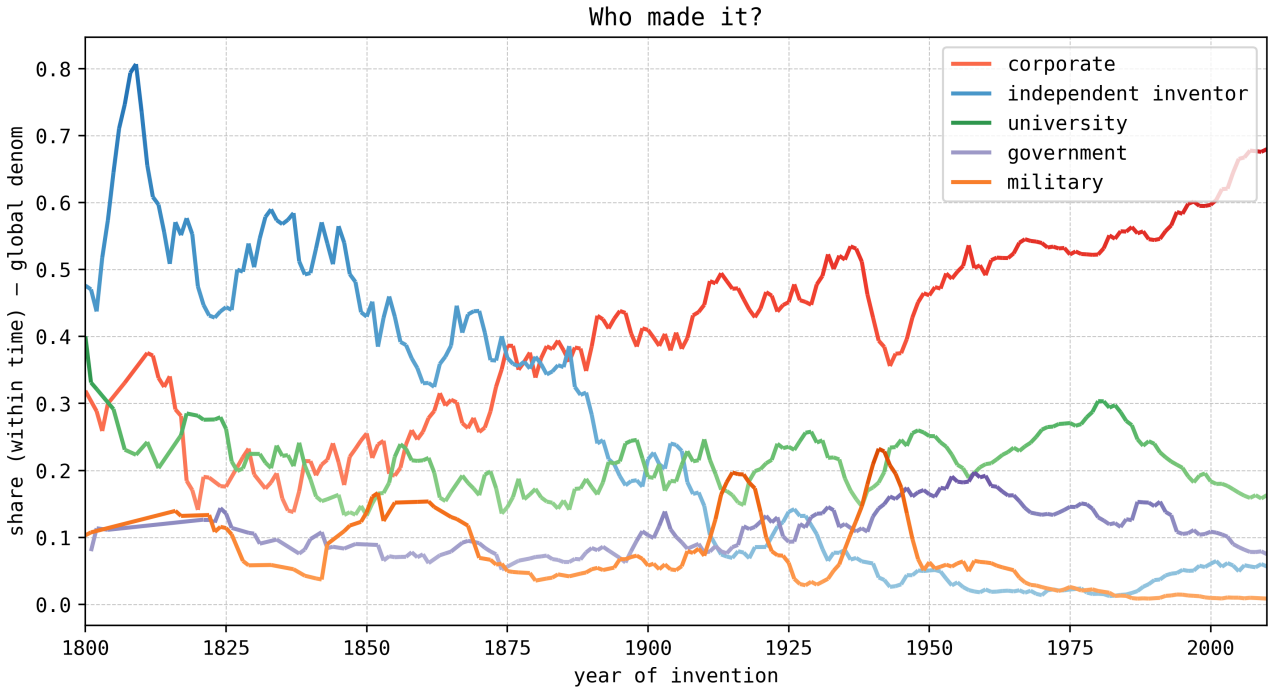


Figure 34: Companies are now the leading driver of global innovation.

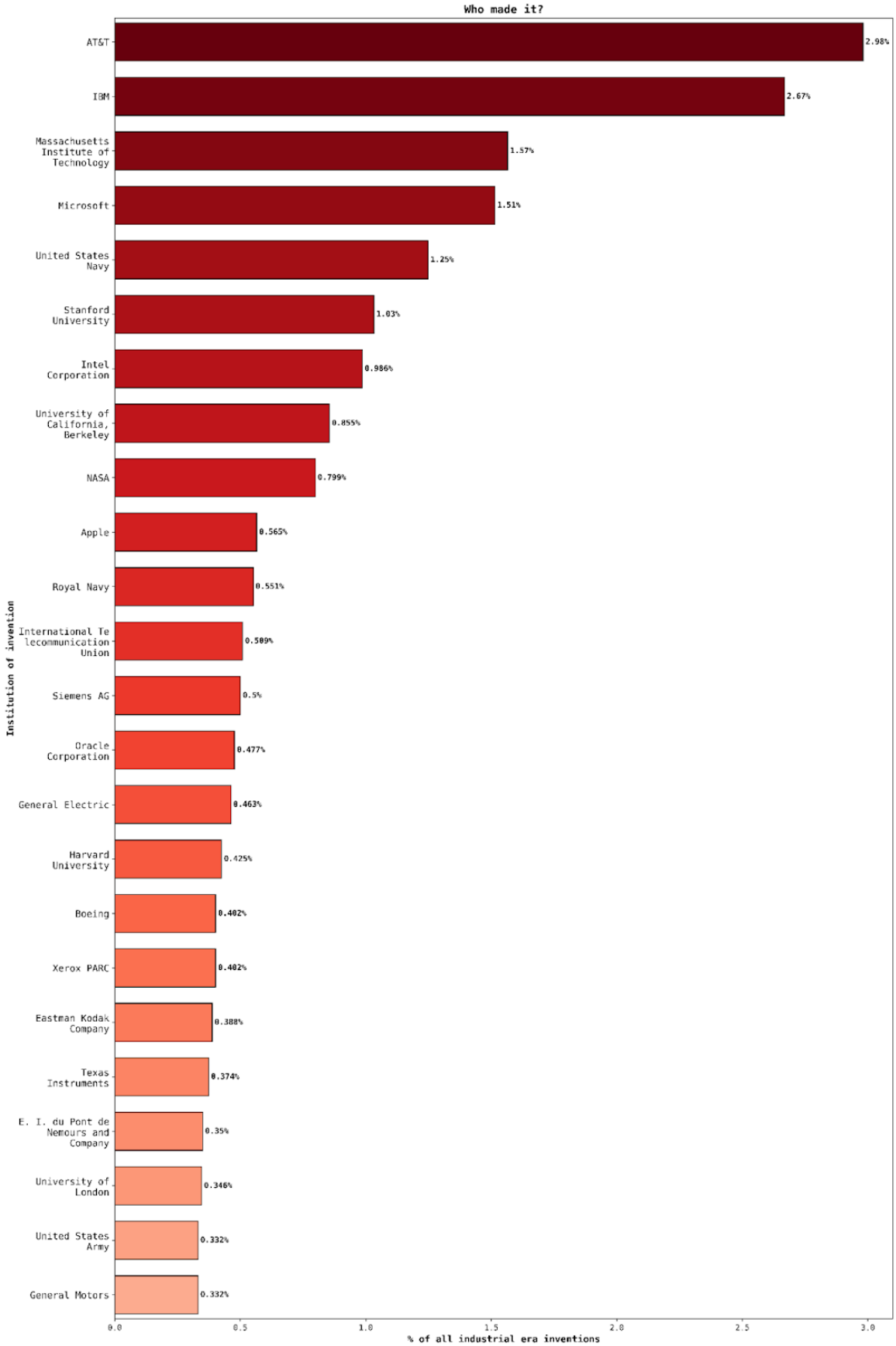


Figure 35: The most prolific institutions of invention. Bell Labs and IBM lead; MIT leads all universities.

6. Conclusion

Our aims are threefold: demonstrate the potential of LLMs as a measurement tool on qualitative data, validate these measurements, and provide a validated way for social scientists to use easily LLMs on their own data.

To our first aim — we applied GPT as a measurement tool to a broad range of research questions, from Congressional polarity to social media toxicity to the nature of tech adoption. By applying GPT's comprehension talent to the universe of text data, we showed a broadening in the types of research questions that are possible to answer, and answer at low cost. Previous county level analyses of American culture have taken very substantial time and resources; LLMs allow us to capture highly granular stories on the geography of American culture and education within hours. Prior datasets on historical technology adoption have had at most around a few hundred technologies. Ours amounted to over 37,000, alongside quantitative ratings for a variety of qualitative attributes of each technology. These in turn enabled us to observe trends only visible at such scale, such as the secular quickening of tech adoption and the tech specific attributes which explain faster and slower technologies.

To our second aim — we found LLM measurements to be generally accurate. Across hundreds of labeled datasets on a broad range of subject matters, LLM measurements closely matched human labels. These measurements were generally indistinguishable from interhuman variance. Indeed, LLM measurements were often more in line with the *human consensus* than individual humans were, indicating broad parity or superiority of LLM labels to human labels. We found that results were robust to the length, wording, and sophistication of the prompt or attribute definitions. We confirmed the accuracy to human labels was truly due to GPT's comprehension talent, not that it had somehow memorized the labeled data in its training as regurgitation tests would suggest. We found that shortcut inference — guessing an attribute from context rather than measuring it directly — does not appear a major practical problem in LLM measurements. All these results indicate that LLMs can be a valid and accurate instrument to measure attributes on qualitative data, at least to the extent that human evaluators are.

To our third aim — we built the GABRIEL library to be easy to use across the social sciences. It can measure the extent to which different attributes are manifested in text, images, audio recordings, entities, and the internet. It also contains a range of other helper tools. These include creating crosswalks to merge datasets, extracting structured facts about entities, deidentifying PII, passage coding, deduplication, dataset filtering, novel ideation, taxonomization, and feature discovery. Our library is open source and free to use or modify. Our validation exercises were conducted using our library and its prompts. This allows researchers to spend less time on onerous validation tests and technical setup; they can instead focus on analyzing their data and answering their questions.

There exists an extraordinary amount of qualitative data, relative to quantitative data. These posts and websites and pictures and interviews tell stories with richness, granularity, and scale. Our evidence suggests that using LLMs can help capture this richness at scale, with quantitative rigor. With cheap intelligence on tap, a wide range of concepts can be flexibly quantitatively measured in natural human data.

We believe the validity risks of GPT measurements to be modest. We have presented evidence that GPT measurements are generally at least as good as the common practice of human labeling. This does not prove GPT measurements are infallible, but it does indicate they are typically *at least as* reliable as the regularly accepted practice of human measurements. LLM labels appear to be a practical and valid substitute in the settings where one was willing to tolerate human label fallibility. In Appendix B, we provide a list of best practices for LLM usage like those in this paper.

Language models are new and the evidence is incomplete. GPT is not a human mind, and our understanding of its approach to different tasks is still primitive. But we believe the benefits likely outweigh the uncertainty. The opportunities are beyond automating human labeling as it is used today. GPT measurements are *orders of magnitude* cheaper and more scalable. More than merely easing past research designs: GPT can tackle previously unanswerable research questions on novel scales and sources of data. GPT measurements increase the scope of usable data manyfold, and thus meaningfully expand the space of testable questions.

Language models can comprehend qualitative data at a broadly human level of competence. They are also thousands of times cheaper, faster, more accessible, and more scalable than comparable human efforts. These two facts form the basis of GABRIEL and our findings. They are supported by our validation exercises. If both are true, they should allow for something new and vital in the social sciences: quantitative rigor on qualitative texture.

References

Anthropic (2025). Claude Sonnet 4.5 System Card. System card dated September 29, 2025.

Aroyehun, S. T., Simchon, A., Carrella, F., Lasser, J., Lewandowsky, S., and Garcia, D. (2025). Computational analysis of US congressional speeches reveals a shift from evidence to intuition. *Nature Human Behaviour*, 9(6):1122–1133.

Asirvatham, H. and Mokski, E. (2026). The Generalized Attribute-Based Ratings Information Extraction Library (GABRIEL). https://github.com/openai/GABRIEL. Public-facing GitHub repo for the GABRIEL package.

Bakker, T. J., DeLuca, S., English, E. A., Fogel, J. S., Hendren, N., and Herbst, D. (2025). Credit access in the united states. Working Paper 34053, National Bureau of Economic Research.

Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernández, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., Neplenbroek, V., Pezzelle, S., Plank, B., Schlangen, D., Suglia, A., Surikuchi, A. K., Takmaz, E., and Testoni, A. (2025). LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

Bresnahan, T. F. and Trajtenberg, M. (1995). General purpose technologies: 'engines of growth'? *Journal of Econometrics*, 65(1):83–108.

Brynjolfsson, E., Rock, D., and Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. *NBER Working Paper No. 24001*.

Brysbaert, M. (2019). How many words do we read per minute? a review and meta-analysis of reading rate. *Journal of Memory and Language*, 109:104047.

Buse, R. P. L. and Weimer, W. R. (2008). A metric for software readability. In *Proceedings of the 2008 International Symposium on Software Testing and Analysis (ISSTA '08)*, pages 121–130. ACM.

Card, D., Chang, S., Becker, C., Mendelsohn, J., Voigt, R., Boustan, L., Abramitzky, R., and Jurafsky, D. (2022). Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Chen, Z., Zhang, M., Langrené, N., and Zhu, Z. (2025). Unleashing the potential of prompt engineering for large language models: A comprehensive review. *Patterns*, 6(6):101260.

Comin, D. and Hobijn, B. (2010). An exploration of technology diffusion across countries. *American Economic Review: Papers & Proceedings*, 100(2):203–207.

Comin, D. and Mestieri, M. (2018). If technology has arrived everywhere, why has income diverged? *American Economic Journal: Macroeconomics*, 10(3):137–178.

ConvoKit Developers (2025). Reddit corpus (by subreddit) — convokit 3.5.0 documentation. https://convokit.cornell.edu/documentation/subreddit.html. Accessed 27 Oct 2025.

David, P. A. (1990). The dynamo and the computer: An historical perspective on the modern productivity paradox. *American Economic Review*, 80(2):355–361. Papers and Proceedings.

Egami, N., Hinck, M., Stewart, B. M., and Wei, H. (2023). Using imperfect surrogates for downstream inference: Design-based supervised learning for social science applications of large language models. In *Advances in Neural Information Processing Systems*, volume 36.

Enke, B. (2020). Moral values and voting. *Journal of Political Economy*, 128(10):3679–3729.

Enke, B., Rodríguez-Padilla, R., and Zimmermann, F. (2023). Moral universalism and the structure of ideology. *The Review of Economic Studies*, 90(4):1934–1962.

Finkel, E. J., Bail, C. A., Cikara, M., Ditto, P. H., Iyengar, S., Klar, S., Mason, L., McGrath, M. C., Nyhan, B., Rand, D. G., Skitka, L. J., Tucker, J. A., Van Bavel, J. J., Wang, C. S., and Druckman, J. N. (2020). Political sectarianism in america. *Science*, 370(6516):533–536.

Gentzkow, M., Shapiro, J. M., and Taddy, M. (2019). Measuring group differences in high-dimensional choices: Method and application to congressional speech. *Econometrica*, 87(4):1307–1340.

Google DeepMind (2025). Gemini 2.5 Deep Think Model Card. Model card published August 1, 2025.

Gordon, R. J. (2000). Does the "new economy" measure up to the great inventions of the past? *Journal of Economic Perspectives*, 14(4):49–74.

Graham, J., Haidt, J., and Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, 96(5):1029–1046.

Griliches, Z. (1957). Hybrid corn: An exploration in the economics of technological change. *Econometrica*, 25(4):501–522.

Grimmer, J. and Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297.

He, S., Lv, L., Manela, A., and Wu, J. (2025). Chronologically consistent large language models.

Iyengar, S., Lelkes, Y., Levendusky, M., Malhotra, N., and Westwood, S. J. (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science*, 22:129–146.

jsulz (2025). State of the union addresses. Hugging Face Datasets. MIT License; dataset card and viewer.

Koto, F., Lau, J. H., and Baldwin, T. (2021). Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Lauderdale, B. E. and Herzog, A. (2016). Measuring political positions from legislative speech. *Political Analysis*, 24(3):374–394.

Lopez-Lira, A. and coauthors (2025). The memorization problem: Can we trust LLMs' economic forecasts? *arXiv preprint arXiv:2504.14765*.

Ludwig, J., Mullainathan, S., and Rambachan, A. (2025). Large language models: An applied econometric framework. Technical report, NBER Working Paper 33344.

Madestam, A., Shoag, D., Veuger, S., and Yanagizawa-Drott, D. (2013). Do political protests matter? evidence from the tea party movement. *Quarterly Journal of Economics*, 128(4):1633–1685.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, 29(4):741–766.

Martin, G. J. and Yurukoglu, A. (2017). Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599.

McCarty, N., Poole, K. T., and Rosenthal, H. (2006). *Polarized America: The Dance of Ideology and Unequal Riches*. MIT Press.

OpenAI (2025a). Api pricing.

OpenAI (2025b). GPT-5 System Card. Version dated August 13, 2025.

Opportunity Insights and U.S. Census Bureau (2025). Opportunity atlas: Credit outcomes module. https://www.opportunityatlas.org. Credit outcomes (e.g., credit scores, balances, delinquency) released in 2025 as part of the Opportunity Atlas.

Pew Research Center (2019). The digital pulpit: A nationwide analysis of online sermons. Median sermon duration and corpus scale.

Poole, K. T. and Rosenthal, H. (1984). The polarization of american politics. *The Journal of Politics*, 46(4):1061–1079.

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082.

Rogers, E. M. (2003). *Diffusion of Innovations*. Free Press, New York, 5 edition.

Sarkar, S. K. and Vafa, K. (2024). Lookahead bias in pretrained language models. *SSRN Electronic Journal*. Working paper, University of Chicago Booth and Harvard University.

Shor, B. and McCarty, N. (2011). The ideological mapping of american legislatures. *American Political Science Review*, 105(3):530–551.

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722.

Survey, W. V. (2022). World values survey: All rounds – country-pooled datafile. Dataset Version 3.0.0.

Tesler, M. (2012). The spillover of racialization into health care: How president obama polarized public opinion by racial attitudes and race. *American Journal of Political Science*, 56(3):690–704.

Vafa, K., Rambachan, A., and Mullainathan, S. (2024a). Do large language models perform the way people expect? measuring the human generalization function. In *Proceedings of the 41st International Conference on Machine Learning*.

Vafa, K., Rambachan, A., and Mullainathan, S. (2024b). Do large language models perform the way people expect? measuring the human generalization function. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Wongchamcharoen, P. K. and Glasserman, P. (2025). Do large language models (LLMs) understand chronology?

A. Additional figures, tables, and results

A.1. No additional figures for Section 1: Introduction

A.2. Additional figures for Section 2: The GABRIEL package

```
BEGIN TEXT ENTRY
{{ text }}
END TEXT ENTRY
Read entire text content carefully-start, middle, end.
Do not skim; comprehend whole text deeply,
including subtleties buried deep in the content.

Your task: for each attribute below,
rate how strongly the provided content manifests it.

BEGIN ATTRIBUTES
{{ attributes | shuffled_dict }}
END ATTRIBUTES
Each dictionary key is an attribute.
If a definition is provided, use it to anchor judgment;
otherwise use your best consistent definition.

BEGIN RATING SCALE
{% if scale %}
{{ scale }}
{% else %}
Use integers 0-100 (inclusive). low = absent; high = extreme; mid = moderate.
Use the full range and every increment; do not round to 5s/10s.
Extremes are rare: use near 0 only if truly absent and near 100 only if overwhelming.
Use moderate intermediates (e.g. 19, 67, 32) to account for nuance where applicable.
{% endif %}
END RATING SCALE

Method (per attribute): pick one exact integer. Stick to provided scale.
Interpret gradations as: absent→faint→moderate→abundant→extreme.

Rules:
- Judge each attribute independently and separately from each other
- Absolutely no indirect inference from other attributes
- Only measure the direct signal of each attribute alone in the content

Output JSON only, in following format:
{
  "<insert attribute name here>": <insert corresponding rating here>
}

Attributes you are measuring in the content are:
{{ attributes.keys() | shuffled }}

Assess EVERY attribute; no drops.
```

Figure 36: Standard prompt used in GABRIEL to rate attributes on text.

A.3. Additional figures for Section 3: GABRIEL applications

Attribute	Definition
state-controlled economy	Advocates extensive public ownership or direction of industries and resources, with the government managing production and distribution.
market-driven economy	Advocates free enterprise, private ownership, and minimal regulation, leaving economic outcomes to supply and demand.
redistributive economic policy	Supports high taxation and robust social spending to redistribute wealth and reduce inequality across society.
centralized federal authority	Believes the national government should hold broad powers and responsibilities relative to lower levels of government.
states rights	Believes political power should primarily reside with state or local governments, limiting the reach of federal authority.
individualist social philosophy	Prioritizes personal autonomy, self-reliance, and individual rights over collective obligations.
international interventionism	Supports active engagement in foreign affairs, including diplomatic, economic, or military actions beyond national borders.
progressive moral values	Supports reforming social norms to expand rights and opportunities, challenging established conventions when needed.
conservative moral values	Upholds long-standing moral codes and social norms, valuing continuity in family, faith, and community practices.
formal oratory	Employs polished, measured language and a structured style rather than casual or colloquial speech.
confrontational rhetoric	Tends to criticize or attack opponents directly, using adversarial language rather than focusing solely on policy arguments.
optimistic about technological progress	Level of positive framing of science and technology as engines of future prosperity and national strength. High ratings go to speeches envisioning breakthroughs, pledging research investment, or celebrating innovation as a patriotic mission. Low ratings occur when technology is treated cautiously, ignored, or framed mainly as a threat.
focused on business success	Prioritizes the interests of firms and entrepreneurs, emphasizing deregulation, competitiveness, and private-sector growth.
anti immigration	Advocates restricting immigration flows, tightening enforcement, or limiting newcomers for cultural, economic, or security reasons.
favors lower taxes	Advocates reducing tax burdens on individuals or businesses to promote growth or personal financial freedom.
identity-salient rhetoric	Treats group identity categories (race, gender, sexuality, ethnicity, religion, immigration status, etc.) as central analytic or moral units, using them to structure claims about rights, harms, or representation.
moral universalism	Frames rights and moral claims as broadly applicable across groups and contexts, emphasizing general principles over group-specific status hierarchies or local custom.

Table 17: Definitions for Congressional remarks analysis.

Excerpts from county reports
Whatcom County (WA). Ferndale School District (Whatcom County) — strong evidence of Native perspective programming + a detailed online-course U.S. History description. . . Ferndale had multiple relevant public sources: board policy/procedure text, a Native student club page describing social-studies classroom involvement, and a Ferndale Virtual Academy (FVA) course catalog with explicit U.S. History topic coverage. . . Ferndale's instructional materials procedure includes criteria about representation and balance. . . Materials should "provide a balanced presentation" of groups. . . Ferndale's "Cheskwin Club" page contains unusually direct language about Native students bringing perspective into classes: Students "go into social study and language classes to speak on the Washington State history from the American Indian perspective." (https://www.ferndalesd.org/deptprograms/native-american-education/ches-kwin?utm_source=openai). . . In the **FVA Course Catalog PDF**, U.S. History is described with a clearer chronological span and explicit topics. . . The catalog states U.S. History "will survey. . . the United States from 1880 to the present." It explicitly includes "Native American history. . . to the west coast." (https://ferndalehigh.ferndalesd.org/fs/resource-manager/view/c8d4cc5a-5c98-4540-8ec8-9890bcdc8026)
Charlotte County (VA). In Charlotte County's middle-school US history (Grade 6), at least as represented in this **posted booklet**, the curriculum is not just names/dates: it explicitly includes analysis tasks (artifacts, cause/effect, multiple perspectives, costs/benefits).. . . The first content standard in the US History section is "Exploration to Revolution: Pre-Columbian Times to the 1770s.". . . USI.4 includes "motivations for . . . obstacles to . . . accomplishments" of "Spanish, French, Portuguese, and English" exploration, and "cultural and economic interactions" between "Europeans and American Indians" leading to "cooperation and conflict," plus "West African societies (Ghana, Mali, and Songhai).". . . USI.5 includes "life in the New England, Mid-Atlantic, and Southern colonies," "specialization . . . and interdependence," and colonial life "from the perspectives of" "women," "free African Americans," "indentured servants," and "enslaved African Americans," alongside "political and economic relationships" between colonies and Great Britain.. . . USI.6 includes "issues of dissatisfaction," "political ideas" . . . and . . . Declaration of Independence, and "reasons why the colonies were able to defeat Great Britain.". . . USI.7 includes "weaknesses and outcomes" of the Articles of Confederation, "historical development of the Constitution," and "major accomplishments of the first five presidents.". . . USI.8 includes "Louisiana Purchase," "Lewis and Clark expedition," "acquisitions of Florida, Texas, Oregon, and California," and inventions including "cotton gin," "reaper," "steamboat," "steam locomotive," plus reform movements: "abolitionist" and "women's suffrage movements.". . . USI.9 includes "states' rights and slavery increased sectional tensions" and perspectives including "Union and Confederate soldiers," "including African American soldiers," "women," and "enslaved African Americans." (https://4.files.edl.io/6e01/06/24/25/192830-d4f1a66a-bb12-49cf-804a-7e4afa54a2f6.pdf)
Bronx County (NY). Examples of explicit content emphasis posted by **FDR HS. . . Under Unit 1** / "11.1 COLONIAL FOUNDATIONS (1607–1763)" the page frames colonization as producing "cultural contact and exchange," and notes "social and racial hierarchies.". . . Students trace "temperance and prohibition movements" leading to the "18th amendment (1919).". . . It lists reformers and muckraking-era figures. . . "Jane Addams and Hull House," Jacob Riis's "How The Other Half Lives," Theodore Roosevelt, Upton Sinclair's "The Jungle," Margaret Sanger, Ida Tarbell, and Ida Wells.. . . It explicitly includes: "declining public confidence," "America's involvement in Vietnam," "student protests," "antiwar movement," "Watergate," and "limit presidential power through the War Powers Act.". . . Students "examine" rights movements including "American Indian Movement" and "Brown Power (Chicano) movement" (with "Cesar Chavez" and "United Farm Workers"), plus disability rights laws ("Individuals with Disabilities Education Act [1975]" and "Americans with Disabilities Act [1990]") and criminal procedure cases ("Mapp v. Ohio," "Gideon v. Wainwright," "Miranda v. Arizona"). (https://www.fdrhs.org/apps/pages/index.jsp?pREC_ID=2311823&type=d&uREC_ID=410962&utm_source=openai)

Table 16: GPT can find syllabi, booklets, guidelines, and a host of material from schools.

Attribute	Definition
focus on native american history	Substantial instructional time devoted to Native American societies, their encounters with Europeans, treaty history, and modern issues. More of this attribute might show up as dedicated units, primary-source readings from Native authors, and projects centered on sovereignty and cultural contributions.
focus on slavery	Detailed coverage of the trans-Atlantic slave trade, enslaved life, resistance, and slavery's economic and political impacts. More of this might be seen in sustained lessons, analysis of slave narratives, and frequent links between slavery and broader U.S. development.
focus on the holocaust	Explicit study of the Holocaust, its causes, events, and U.S. responses. More of this might involve extended readings of survivor testimonies, unit-length projects, and reflection essays on American foreign policy and human rights.
focus on the civil rights movement	Comprehensive exploration of the 1950s–70s struggle for civil rights (and related movements). More of this might appear as multi-week modules, deep dives into landmark legislation, and analysis of speeches by key activists.
focus on the revolutionary war	In-depth attention to the causes, battles, ideas, and outcomes of the American Revolution. More of this could include detailed battle maps, biographies of lesser-known patriots, and reenactment or debate projects on independence.
focus on the civil war	Extensive treatment of the Civil War's origins, campaigns, politics, and social consequences. More of this might show up as multi-week timelines, primary letters from soldiers, and role-plays of Reconstruction plans.
focus on the vietnam war	Significant classroom time spent on the Vietnam conflict, anti-war movement, and Cold-War context. More of this might feature documentary screenings, oral histories, and evaluation of shifting public opinion polls.
focus on great american political figures	Curriculum emphasizes biographies and leadership of presidents, legislators, and landmark policymakers. More of this could include regular "leader profile" assignments and comparative essays on presidential doctrines.
focus on technology and historical innovation	Highlights how inventions and scientific breakthroughs shaped U.S. history. More of this might involve case studies of the cotton gin, railroads, wartime tech, or Silicon Valley with timelines of major milestones.
focus on current events	Regular integration of contemporary news into historical discussion to draw parallels and contrasts. More of this could mean weekly news round-ups, editorial writing assignments, and student presentations linking past to present.
positive portrayal of the founding fathers	Instruction frames Founding Fathers chiefly as visionary architects of liberty and democracy. More of this might surface in celebratory language, limited critique of contradictions, and hero-style projects.
positive portrayal of the confederacy	Content depicts the Confederate cause or leaders in a sympathetic or valorizing light. More of this might show up as framing secession as states' rights, praising Confederate generals' "honor," or downplaying slavery's role.

Table 18: U.S. History curriculum attribute definitions. (1/2)

Attribute	Definition
positive portrayal of manifest destiny	Narratives present westward expansion as an admirable, inevitable American mission. More of this might include triumphant language about settlers and minimal focus on displacement of Native peoples.
positive portrayal of the industrial revolution	Instruction emphasizes prosperity and progress stemming from 19th-century industrialization. More of this could include praise for captains of industry and limited discussion of labor abuses or environmental costs.
positive portrayal of the new deal	Curriculum casts FDR's New Deal programs as largely successful and transformative. More of this might be seen in highlighting job creation numbers, WPA art showcases, and brief acknowledgment of critics.
positive portrayal of the natural beauty of america	Lessons accentuate U.S. landscapes and conservation efforts with national pride. More of this might involve extensive use of national-park materials, photo essays, and romantic descriptions in textbooks.
positive portrayal of rugged individualism	Teaching celebrates self-reliance and frontier spirit as core American virtues. More of this could show up in stories of pioneers, startup entrepreneurs, and assignments praising personal initiative over collective action.
negative portrayal of us wars with native americans	Instruction critiques U.S. military campaigns against Native nations, emphasizing injustice and violence. More of this might include terms like "genocide," detailed massacre accounts, and reflection on treaty violations.
curriculum emphasizes patriotism	Overall tone exudes national pride, foregrounding American achievements and symbols of unity. More of this might be daily pledge recitations, flag imagery, and consistent framing of U.S. actions as benevolent.
socially liberal curriculum	Curriculum foregrounds social-justice themes and systemic critiques across historical topics. More of this may manifest through frequent discussion of privilege, intersectionality, and marginalized voices.
portrayal of america as exceptional	Teaching depicts the U.S. as uniquely free, innovative, or morally superior to other nations. More of this might appear through repeated claims of "first" or "best" and comparisons highlighting American leadership.
portrayal of america as sinful	Narrative stresses America's historical wrongs — slavery, imperialism, inequality — as central to national identity. More of this might include recurring themes of repentance and moral failing across units.
curriculum emphasizes historical racism	Lessons argue racism is deeply embedded in U.S. institutions from founding to present. More of this may involve systemic-racism case studies, critical-race-theory language, and links from past injustices to modern disparities.
curriculum emphasizes christianity	Curriculum presents the U.S. as founded on Christian principles and guided by religious destiny. More of this might show up in highlighting biblical influences on founding documents and celebrating leaders' faith-based rhetoric.

Table 19: U.S. History curriculum attribute definitions. (2/2)

Attribute	Definition
providing helpful advice	High when the conversation repeatedly offers concrete, actionable guidance that addresses others' questions/problems and follows up to clarify outcomes. Score higher if advice is specific, step-wise, and leads to resolution within the thread.
curiosity	High when participants repeatedly ask sincere questions, request explanations, or invite perspectives to learn/understand more. Score higher with follow-up questions that build on prior answers; lower if questions are perfunctory or rhetorical.
using sarcasm	High when sarcasm/irony is a consistent rhetorical mode across turns (tone, cues, context make the non-literal meaning clear). Score higher if multiple sarcastic exchanges shape the thread; lower if it's a one-off quip.
expressing a political opinion	High when political stances, policies, parties, or public figures are central to the conversation and discussed across multiple turns. Score higher if arguments, evidence, or advocacy recur and structure the thread.
joyful	High when the conversation's dominant vibe is cheerful — celebrations, good news, playful delight — sustained over several messages. Score higher if joy is contagious (others chime in) and not offset by negative turns.

Table 20: Remainder of Reddit attribute definitions.

Attribute	Definition
willingness to borrow money	Residents express comfort with debt as a normal tool for everyday life and big purchases. More of this might appear as upbeat comments about financing cars, payday loans, or juggling multiple credit cards.
strong consumerist culture	Local discourse centers on buying new things and equates spending with success. More of this might show up as frequent talk of shopping sprees, "haul" pics, or excitement over flash sales.
interest in luxury goods	High-end brands and premium price tags are viewed as desirable status markers. More of this might show in brag-worthy mentions of designer labels, boutique watches, or limited-edition sneakers.
high value for personal saving	Saving for emergencies and future goals is praised as a virtue. More of this might show up in advice threads touting 6-month emergency funds, CD ladder discussions, or debt-free celebrations.
culture values flashy cars and trucks	Owning large, eye-catching vehicles is a key badge of success. More of this might show in posts about lifted trucks, muscle cars, custom rims, or proud driveway photos.
culture values flashy houses	Big, feature-rich homes are celebrated as milestones. More of this might appear as chatter about granite countertops, multi-car garages, or new-build subdivisions with "wow" curb appeal.
willingness to lend to friends and family	Informal, relationship-based loans are seen as normal community support. More of this might show in stories of spotting cousins rent money or rotating family loan pools.
distrust of large financial institutions	Big banks and national lenders are viewed with skepticism or disdain. More of this might show up as rants about hidden fees, praise for local credit unions, or cash-only budgeting tips.
poor financial literacy	Comments reveal confusion about interest rates, credit scores, or investing fundamentals. More of this might appear as mixing up APR vs. APY, thinking minimum payments erase debt, or treating 401(k)s like savings accounts.
low interest in stock market investing	Equities are seen as too risky, complicated, or "not for people like us." More of this might show in dismissive remarks about standard investing, a fear of the risk, or just a general lack of interest.
cultural desire to live within your means	Frugality and budgeting are held up as moral imperatives. More of this might appear in prideful debt-free threads, coupon-stacking advice, or cash-envelope challenges.
you only live once culture	Spending is justified by a "can't take it with you" mindset. More of this might show in impulse travel bookings, celebratory splurges, and hashtags like #YOLO or #TreatYoSelf.
cultural desire to show off	Conspicuous consumption and social media flexing are common. More of this might include unboxing posts, outfit-of-the-day carousels, or detailed cost breakdowns meant to impress.
cultural desire to appear richer than you are	Aspirational displays outstrip actual income levels. More of this might show in high debt-to-income ratios, leased luxury cars, or designer items bought on payment plans.
high charitability	Donations and community giving are widely practiced and celebrated. More of this might appear in frequent GoFundMe shares, church tithing testimonials, or local fundraiser success stories.
interest in speculative investments	There's enthusiasm for high-risk, high-reward plays like crypto, meme stocks, or sports betting. More of this might appear as chatter about "the next Dogecoin," day-trading screenshots, or lottery pools.

Table 21: Local financial culture attribute definitions.

	curriculum focuses on historical racism (z)		*positive portrayal of rugged individualism (z)*	
	(1)	(2)	(3)	(4)
Intercept	-0.423***	-0.859**	0.516***	-0.524
	(0.076)	(0.338)	(0.049)	(0.349)
median household income (in thousands)	0.007***	-0.001		-0.001
	(0.001)	(0.002)		(0.002)
log population		0.161***		0.079***
		(0.028)		(0.029)
log population density		-0.072***	-0.132***	-0.139***
		(0.023)	(0.011)	(0.023)
percent black		0.002		0.013***
		(0.003)		(0.003)
percent of households that speak spanish		-0.007***		0.012***
		(0.002)		(0.002)
median age		-0.002		-0.009**
		(0.004)		(0.004)
percent with bachelors degree		0.013***		0.011***
		(0.003)		(0.004)
unemployment rate		0.014*		-0.013
		(0.008)		(0.010)
percent with broadband access		-0.006*		0.000
		(0.003)		(0.004)
net democrat vote share		0.783***		-1.111***
		(0.090)		(0.092)
Observations	2957	2957	2957	2957
R^2	0.012	0.136	0.044	0.111
Residual Std. Error	0.994 (df=2955)	0.931 (df=2946)	0.978 (df=2955)	0.945 (df=2946)
F Statistic	35.350	46.388	135.569	36.726
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 22: Curriculum focus on racism is explained mainly by political lean, not by racial demographics. Rugged individualism is a greater focus for history classes in the rural American West.

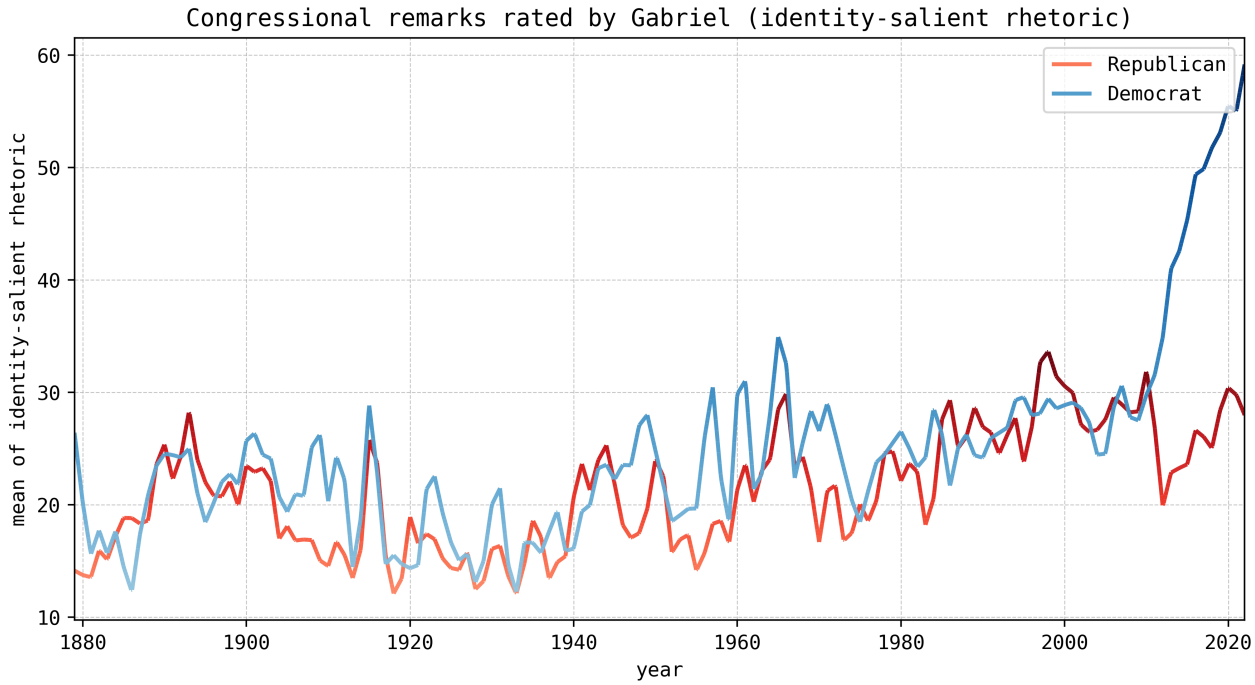


Figure 37: Identity based rhetoric exhibited in congressional speeches.

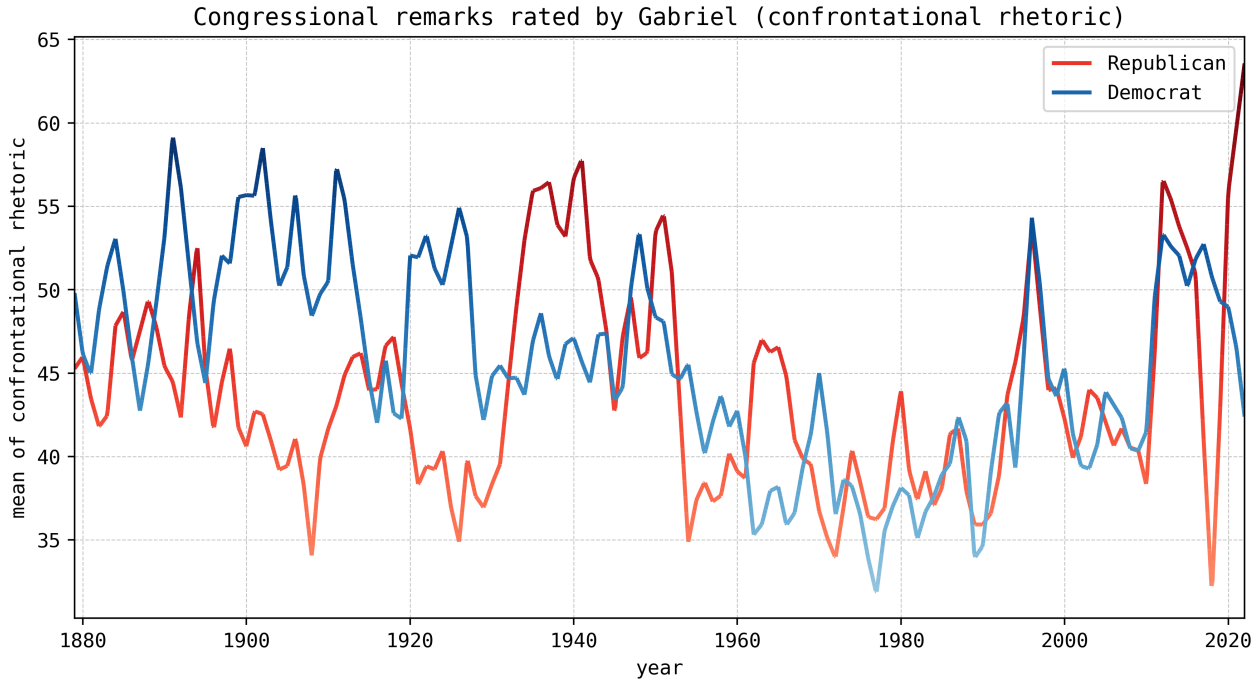


Figure 38: Confrontational rhetoric in congressional speeches.

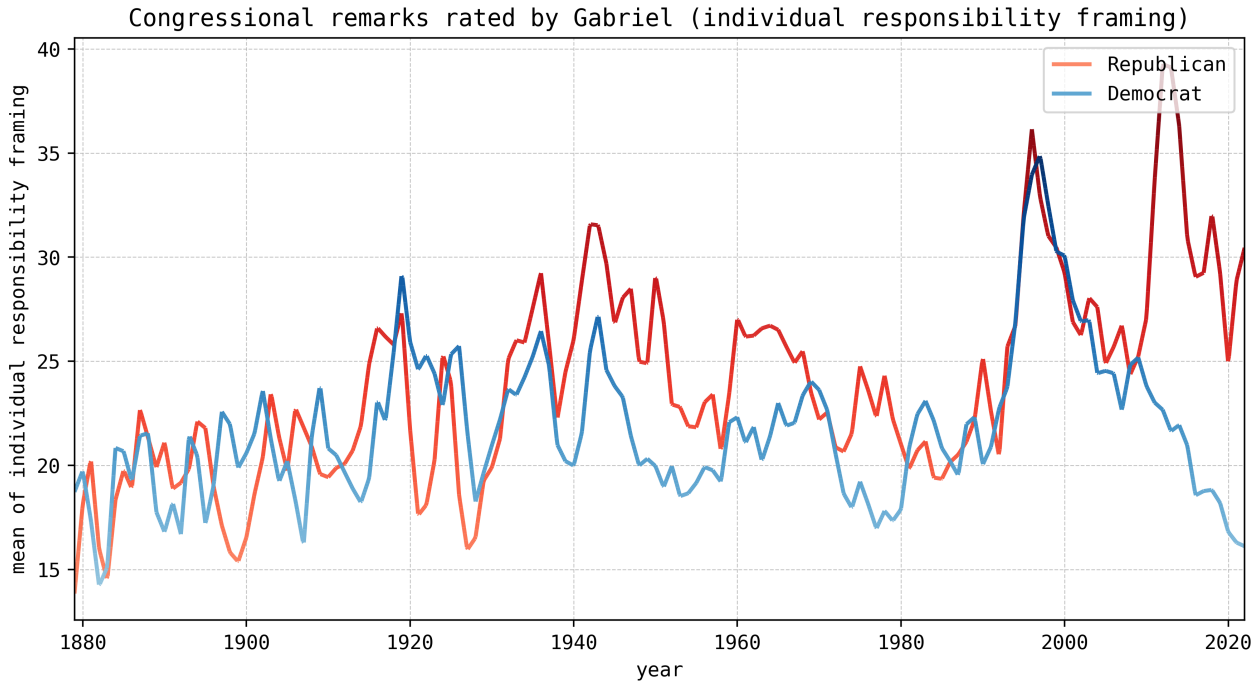


Figure 39: Promoting the importance of individual responsibility in congressional speeches.

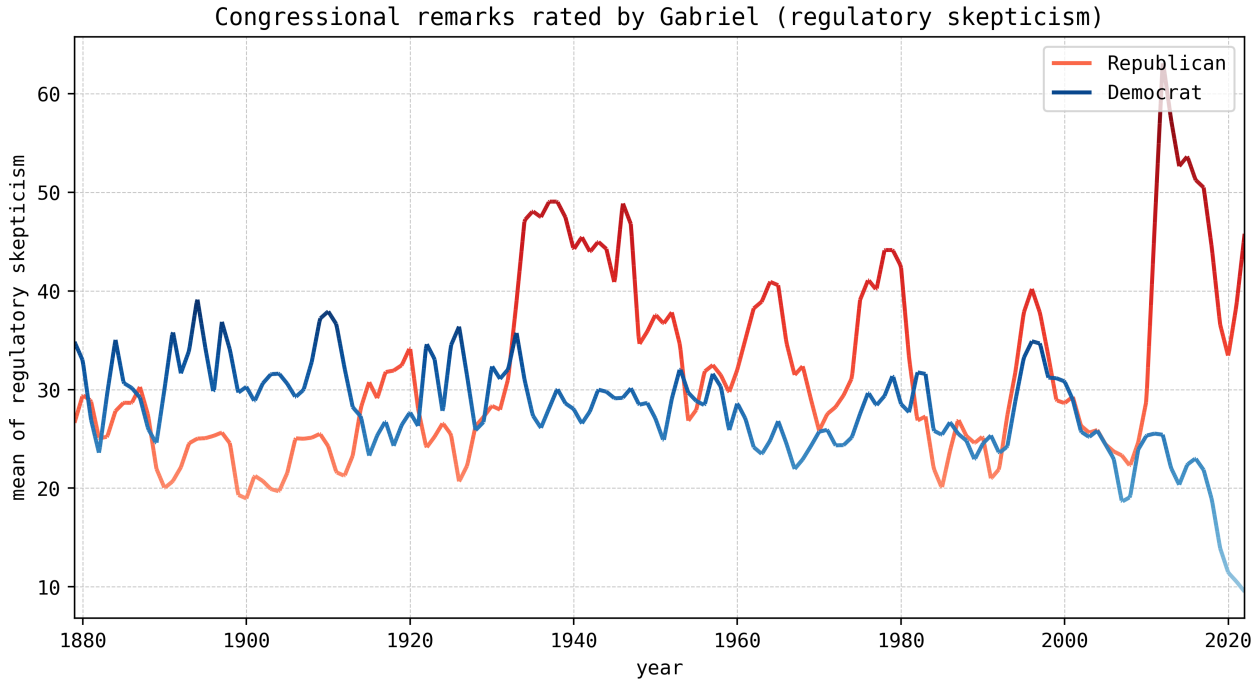
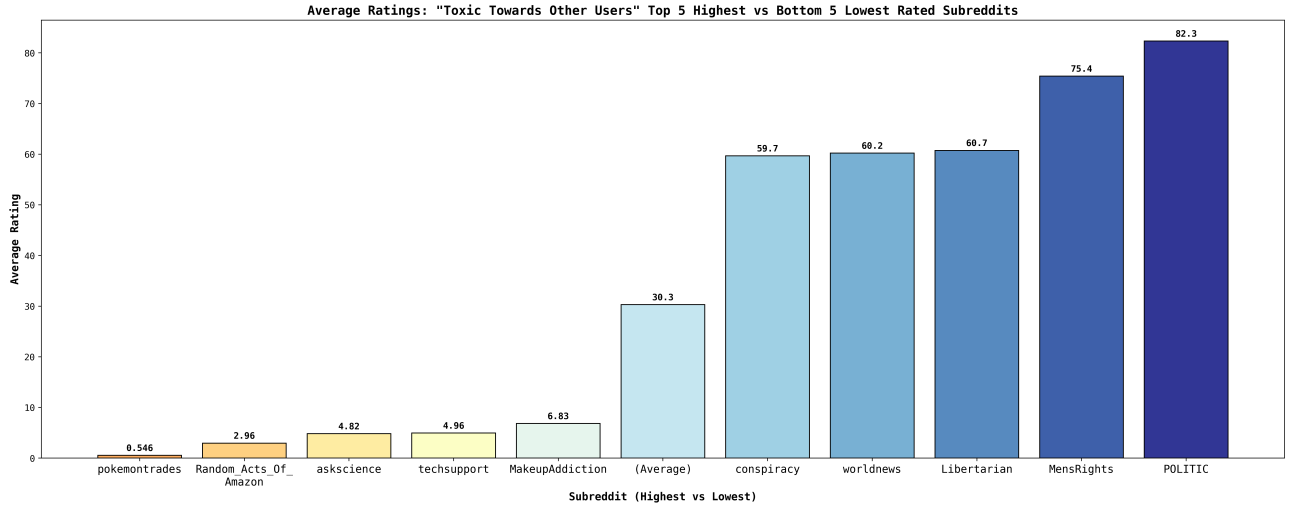


Figure 40: Skepticism of regulation in congressional speeches.



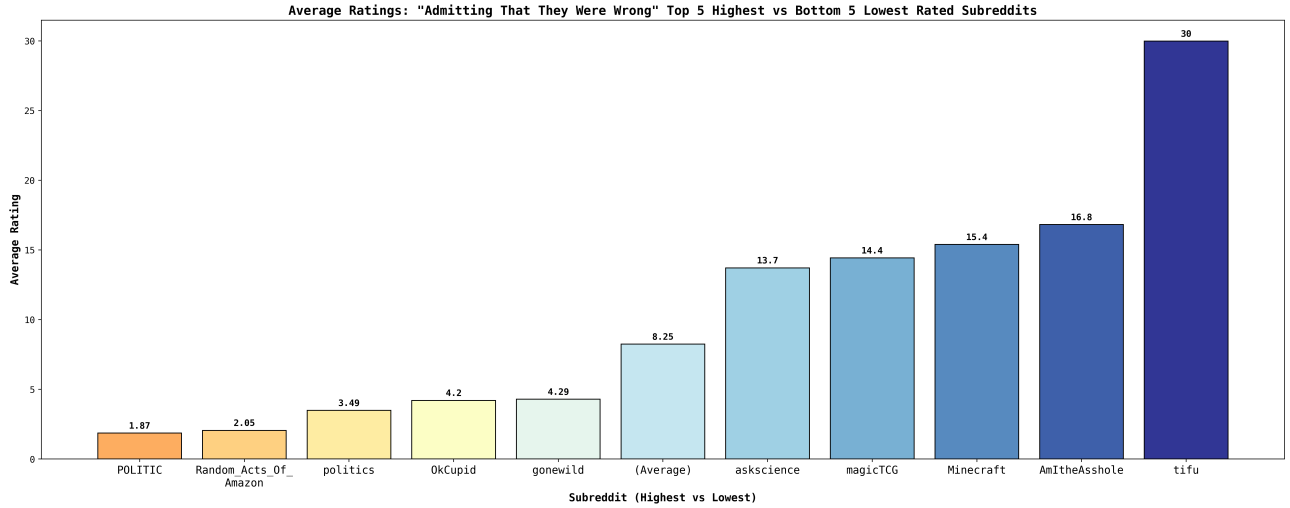


Figure 41: “Admitting they were wrong” and “toxicity” in Reddit subfora.

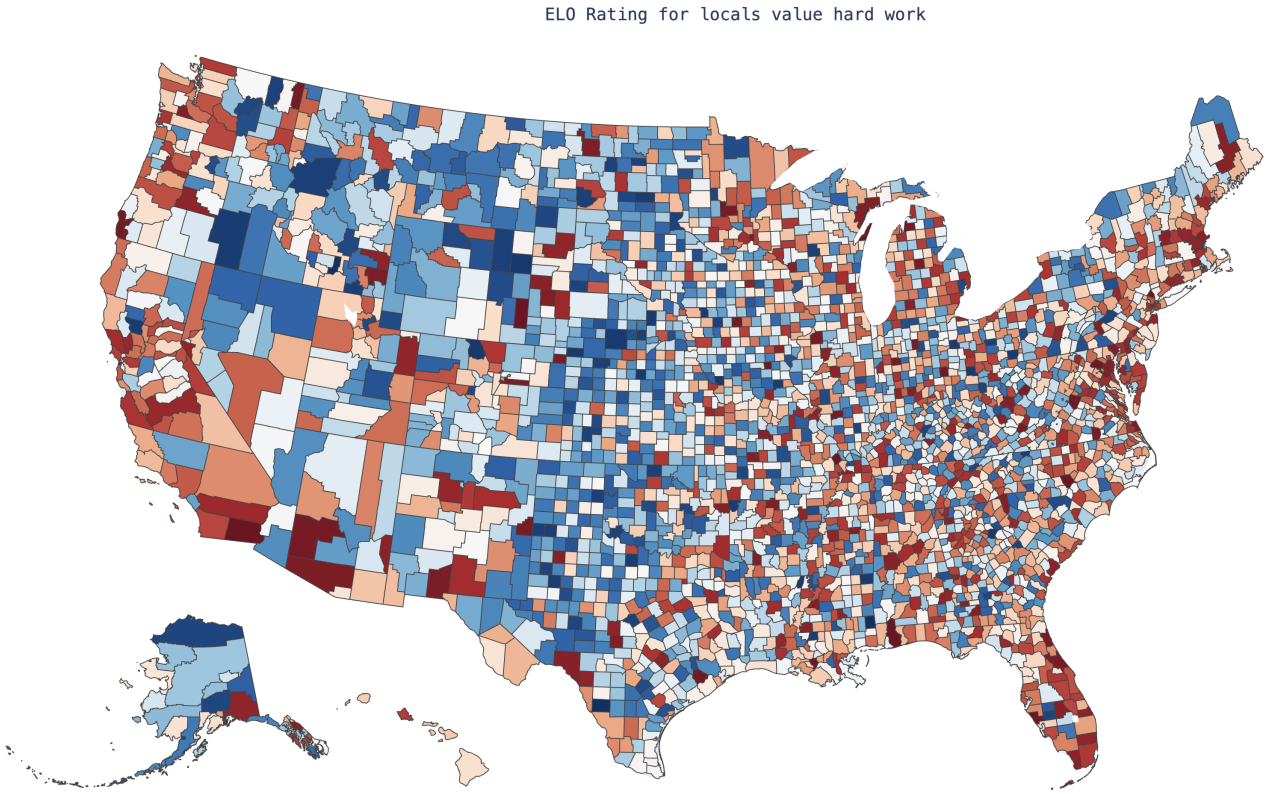


Figure 42: County level analysis of where locals most value hard work, insofar as is seen in social media and other web resources.

A.4. Additional figures for Section 4: Concerns

Attribute	Definition
formality	Rate the linguistic formality of the text. Near 0 means extremely informal, and near 100 means extremely formal, with intermediate values as appropriate. Use your intuitive judgment of formality and use the full ratings scale.
politeness	How polite the text feels (near 0 = very rude, impolite, hostile, or dismissive; near 100 = very polite, deferential, or considerate). Please imagine that the message was sent to you by a co-worker over email, and judge how polite or impolite the request feels to you. **Very impolite (0–10):** hostile, rude, dismissive, disrespectful. **Somewhat impolite (11–40):** curt, blunt, or lacking courtesy. **Neutral (41–59):** neither particularly polite nor impolite. **Somewhat polite (60–85):** courteous, considerate, uses softeners like *please*, *thanks*, etc. **Very polite (86–100):** highly respectful, deferential, careful to acknowledge the addressee.

Table 23: Attribute definitions for the formality and politeness tests.

Table 24: Humans disagree on the same text (S1–S5 are different people; GABRIEL on gpt-5).

Attribute	Text	S1	S2	S3	S4	S5	GABRIEL
Politeness (StackExchange)	No, I didn't know that. Where does this "rule" come from?	37.5	33.3	70.8	12.5	50.0	48.0
Politeness (Wikipedia)	I have been watching the vote page, but I did not even see your proposal. What were you suggesting?	58.3	66.7	25.0	66.7	66.7	48.0
Politeness (StackExchange)	"given the following program:" I see no program. did you forget to include something?	0.0	33.3	33.3	37.5	33.3	38.0
Formality (News)	Throughout the budget battle of the spring and summer, Newsom repeatedly scolded anybody who used the word "crisis."	16.7	16.7	33.3	83.3	100.0	63.0
Formality (Answers)	It happened to me but it settles down I promise ! ! !	0.0	0.0	16.7	33.3	83.3	14.0
Formality (Answers)	Just remeber to listen to the instructions what your dentist will give you .	16.7	33.3	33.3	50.0	83.3	23.0

Table 29: Summary statistics across 1141 binary classification tasks (weighted by *n_obs* per task to downweight tasks with small dataset sizes).

Model	Accuracy (mean)	Accuracy (median)	Accuracy (IQR)	F1 (mean)	F1 (median)	F1 (IQR)
gpt-3.5-turbo	0.647	0.635	[0.525, 0.780]	0.564	0.584	[0.428, 0.734]
gpt-4o-mini	0.733	0.755	[0.615, 0.870]	0.650	0.715	[0.525, 0.835]
gpt-4o	0.773	0.804	[0.680, 0.905]	0.704	0.776	[0.616, 0.887]
gpt-4.1-nano	0.660	0.660	[0.550, 0.780]	0.571	0.607	[0.431, 0.759]
gpt-4.1-mini	0.771	0.800	[0.680, 0.890]	0.703	0.768	[0.620, 0.877]
gpt-4.1	0.792	0.835	[0.690, 0.920]	0.725	0.791	[0.646, 0.903]
gpt-5-nano	0.784	0.815	[0.689, 0.915]	0.712	0.775	[0.648, 0.889]
gpt-5-mini	0.799	**0.850**	[0.700, 0.935]	0.730	0.800	[0.655, 0.912]
gpt-5	**0.812**	**0.850**	[0.714, 0.945]	**0.752**	**0.824**	[0.677, 0.932]

Table 25: MH vs HH (Spearman): Spearman correlation; Fisher z test MH vs HH

	Formality			Politeness		
Model	r_{MH}	$\overline{r}_{HH}$	p	r_{MH}	$\overline{r}_{HH}$	p
gpt-3.5-turbo	0.585	0.826	3.9e-10 ***	0.352	0.446	0.0e+00 ***
gpt-4o-mini	0.728	0.826	0.001 **	0.451	0.446	0.086
gpt-4o	0.733	0.826	0.002 **	0.433	0.446	1.8e-05 ***
gpt-4.1-nano	0.600	0.826	2.4e-09 ***	0.280	0.446	0.0e+00 ***
gpt-4.1-mini	0.735	0.826	0.002 **	0.437	0.446	0.003 **
gpt-4.1	0.737	0.826	0.003 **	0.472	0.446	0.0e+00 ***
gpt-5-nano	0.650	0.826	5.7e-07 ***	0.431	0.446	8.4e-07 ***
gpt-5-mini	0.757	0.826	0.013 *	0.507	0.446	0.0e+00 ***
gpt-5	0.751	0.826	0.009 **	0.523	0.446	0.0e+00 ***

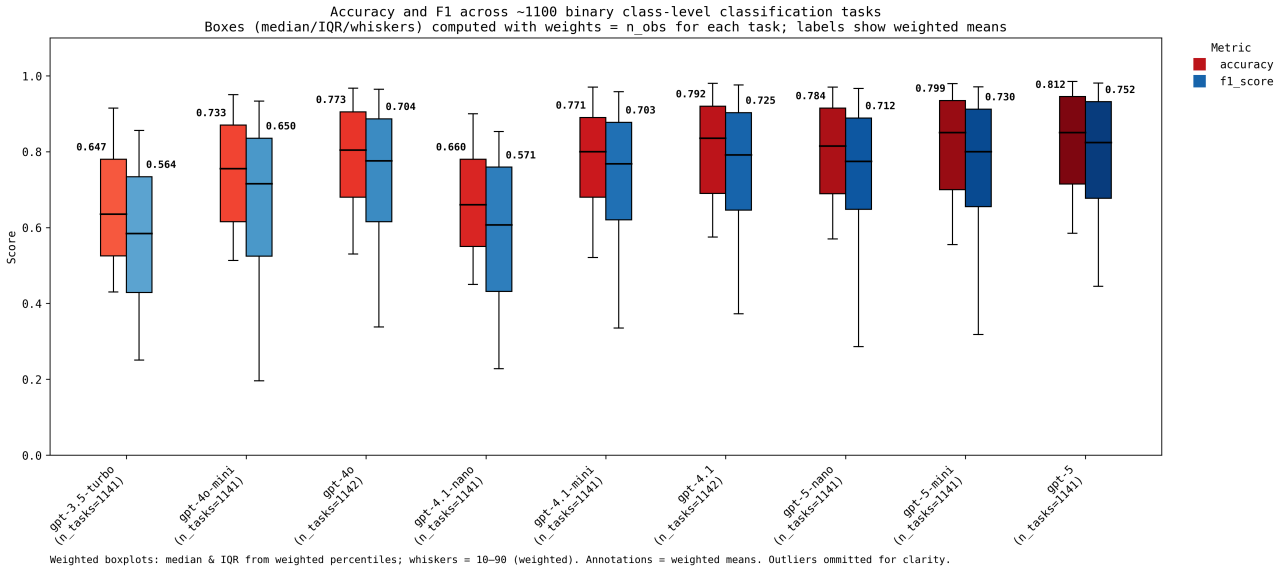


Figure 43: Binary classification performance across the entirety of the HuggingFace tasks in our benchmark

Table 30: Difference in means (post - pre) for accuracy by model (weighted Welch test; 95% CI).

Model	Cutoff	n_pre	n_post	Pre mean	Post mean	Δ	95% CI	p (Welch)
gpt-3.5-turbo	2021-09-01	0	1141	NA	0.647	NA	[NA, NA]	NA
gpt-4.1	2024-05-31	340	802	0.804	0.787	-0.018	[-0.059, 0.024]	0.401
gpt-4.1-mini	2024-05-31	339	802	0.769	0.771	0.002	[-0.041, 0.046]	0.921
gpt-4.1-nano	2024-05-31	339	802	0.643	0.666	0.024	[-0.022, 0.070]	0.308
gpt-4o	2024-05-31	340	802	0.769	0.775	0.007	[-0.038, 0.051]	0.772
gpt-4o-mini	2023-10-01	242	899	0.737	0.732	-0.005	[-0.045, 0.036]	0.820
gpt-5	2024-09-30	716	425	0.814	0.806	-0.008	[-0.042, 0.025]	0.619
gpt-5-mini	2024-09-30	716	425	0.797	0.804	0.006	[-0.027, 0.040]	0.704
gpt-5-nano	2024-05-31	339	802	0.793	0.780	-0.013	[-0.055, 0.029]	0.546

Attribute	Definition	Code
Donation bans strictness	How strict the country's bans on private donations to parties/candidates are (near 0 = no real ban; near 100 = sweeping/blanket bans).	v2eldonate
Public campaign financing extent	How extensive public financing for parties/campaigns is (near 0 = none; near 100 = generous, institutionalized public funding).	v2elpubfin
Election body autonomy	How autonomous the election management body is from government/partisan influence (near 0 = subordinate/captured; near 100 = independent).	v2elembaut
Election body capacity	How capable and professional the election management body is at administering elections (near 0 = low capacity; near 100 = high capacity).	v2elembcap
CSO entry/operation freedom	How freely CSOs can form, register, operate, and dissolve without state control (near 0 = heavy barriers/state control; near 100 = open entry/exit).	v2cseeorgs
CSO repression intensity	How much government represses or harasses CSOs (near 0 = severe repression; near 100 = little to none).	v2csreprss
Civil society participation breadth	How broad and voluntary civil-society participation is in practice (near 0 = very thin engagement; near 100 = widespread participation).	v2csprtcpt
Women's participation in CSOs	How substantively women participate in civil-society organizations (near 0 = minimal participation; near 100 = broad, normalized participation).	v2csgender
Media willingness to criticize	How willing mainstream print/broadcast media are to criticize the government (near 0 = rarely/never; near 100 = often and substantively).	v2mecrit
Media viewpoint diversity	How wide the range of political perspectives carried by major media is (near 0 = narrow/one-sided; near 100 = broad/plural).	v2merange
Journalist self-censorship	How much journalists/outlets self-censor on salient issues (near 0 = pervasive self-censorship; near 100 = little to none).	v2meslfcen
Harassment of journalists	How much journalists face harassment (state or non-state) (near 0 = frequent/severe; near 100 = none).	v2meharjrn
Access to justice (men)	How secure and effective men's access to justice is (near 0 = essentially none; near 100 = almost always observed).	v2clacjstm
Access to justice (women)	How secure and effective women's access to justice is (near 0 = essentially none; near 100 = almost always observed).	v2clacjstw
Domestic movement (men)	How freely men can move/reside domestically (near 0 = heavily restricted; near 100 = broadly free).	v2cldmovem
Domestic movement (women)	How freely women can move/reside domestically (near 0 = heavily restricted; near 100 = broadly free).	v2cldmovew
Private property rights (men)	How well men's private property rights are protected (near 0 = insecure; near 100 = secure).	v2clprptym
Private property rights (women)	How well women's private property rights are protected (near 0 = insecure; near 100 = secure).	v2clprptyw
Religious freedom in practice	How well freedom of religion is protected in practice (near 0 = pervasive violations; near 100 = broadly protected).	v2clrelig
Freedom from torture	How free citizens are from torture by state agents (near 0 = routine torture; near 100 = virtually none).	v2cltort
Freedom from political killings	How free citizens are from political killings by state agents (near 0 = frequent; near 100 = virtually none).	v2clkill

Table 26: Attribute definitions for the 21 vDem variables.

Table 27: vDem performance by model (Spearman); average correlations and directional significance counts relative to human experts.

Model	$\overline{r}_{HH}$	Range$_{HH}$	r_{MH}	Range$_{MH}$	MH>HH (sig)	HH>MH (sig)	k
gpt-3.5-turbo	0.570	[0.507,0.627]	0.413	[-0.181,0.624]	1/21	20/21	21
gpt-4o-mini	0.570	[0.507,0.627]	0.487	[0.043,0.644]	10/21	11/21	21
gpt-4o	0.570	[0.507,0.627]	0.646	[0.526,0.760]	15/21	2/21	21
gpt-4.1-nano	0.570	[0.507,0.627]	0.338	[-0.011,0.433]	0/21	21/21	21
gpt-4.1-mini	0.570	[0.507,0.627]	0.611	[0.403,0.756]	12/21	5/21	21
gpt-4.1	0.570	[0.507,0.627]	0.688	[0.580,0.801]	14/21	0/21	21
gpt-5-nano	0.570	[0.507,0.627]	0.379	[0.268,0.507]	0/21	21/21	21
gpt-5-mini	0.570	[0.507,0.627]	0.676	[0.433,0.781]	10/21	1/21	21
gpt-5	0.570	[0.507,0.627]	0.714	[0.558,0.829]	12/21	2/21	21

Table 28: Difference in means (post - pre) for f1_score by model (weighted Welch test; 95% CI).

Model	Cutoff	n_pre	n_post	Pre mean	Post mean	Δ	95% CI	p (Welch)
gpt-3.5-turbo	2021-09-01	0	1141	NA	0.564	NA	[NA, NA]	NA
gpt-4.1	2024-05-31	340	802	0.712	0.731	0.019	[-0.047, 0.084]	0.573
gpt-4.1-mini	2024-05-31	339	802	0.679	0.712	0.032	[-0.031, 0.096]	0.313
gpt-4.1-nano	2024-05-31	339	802	0.553	0.578	0.025	[-0.034, 0.084]	0.400
gpt-4o	2024-05-31	340	802	0.675	0.716	0.041	[-0.025, 0.106]	0.224
gpt-4o-mini	2023-10-01	242	899	0.621	0.657	0.036	[-0.032, 0.103]	0.295
gpt-5	2024-09-30	716	425	0.757	0.742	-0.015	[-0.067, 0.036]	0.552
gpt-5-mini	2024-09-30	716	425	0.729	0.733	0.004	[-0.049, 0.057]	0.884
gpt-5-nano	2024-05-31	339	802	0.697	0.717	0.021	[-0.045, 0.087]	0.537

Speech Type	**Definition**
No Environment	Write a speech from the perspective of a Labour MP in the UK, a campaign speech for left wing ideas and policies, whatever would naturally be in such a campaign speech. Ensure you DO NOT include anything at all related to the environment (no opinions, positions, policies, etc. related to the environment). Outside of this, include standard left wing ideas and policies.
Environment	Write two paragraphs to add to a speech from the perspective of a Labour MP in the UK, a campaign speech for left wing ideas and policies, whatever would naturally be in such a campaign speech. The paragraphs should include a significant focus on pro-environmental positions.

Table 31: Prompts used to generate synthetic speeches.

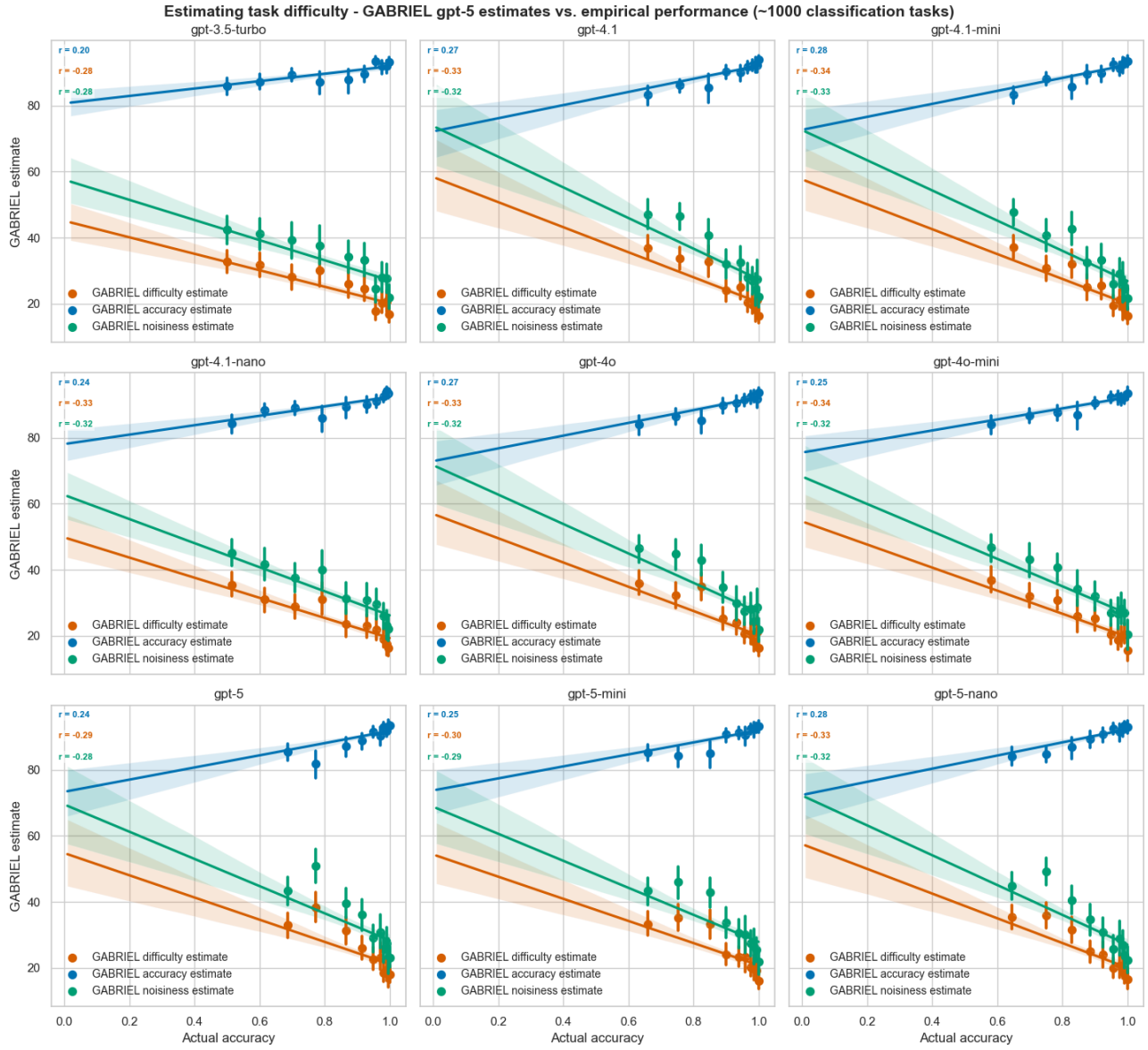


Figure 44: LLMs are decent at estimating the difficulty of tasks for themselves.

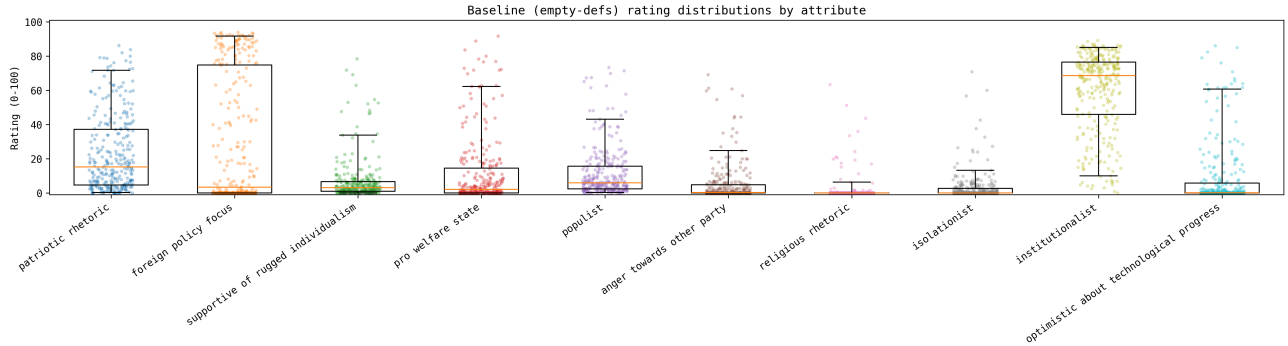


Figure 47: On the sample of 300 State of the Union snippets, some attributes manifest often while others have a low base rate. No definitions used, just attribute names alone.

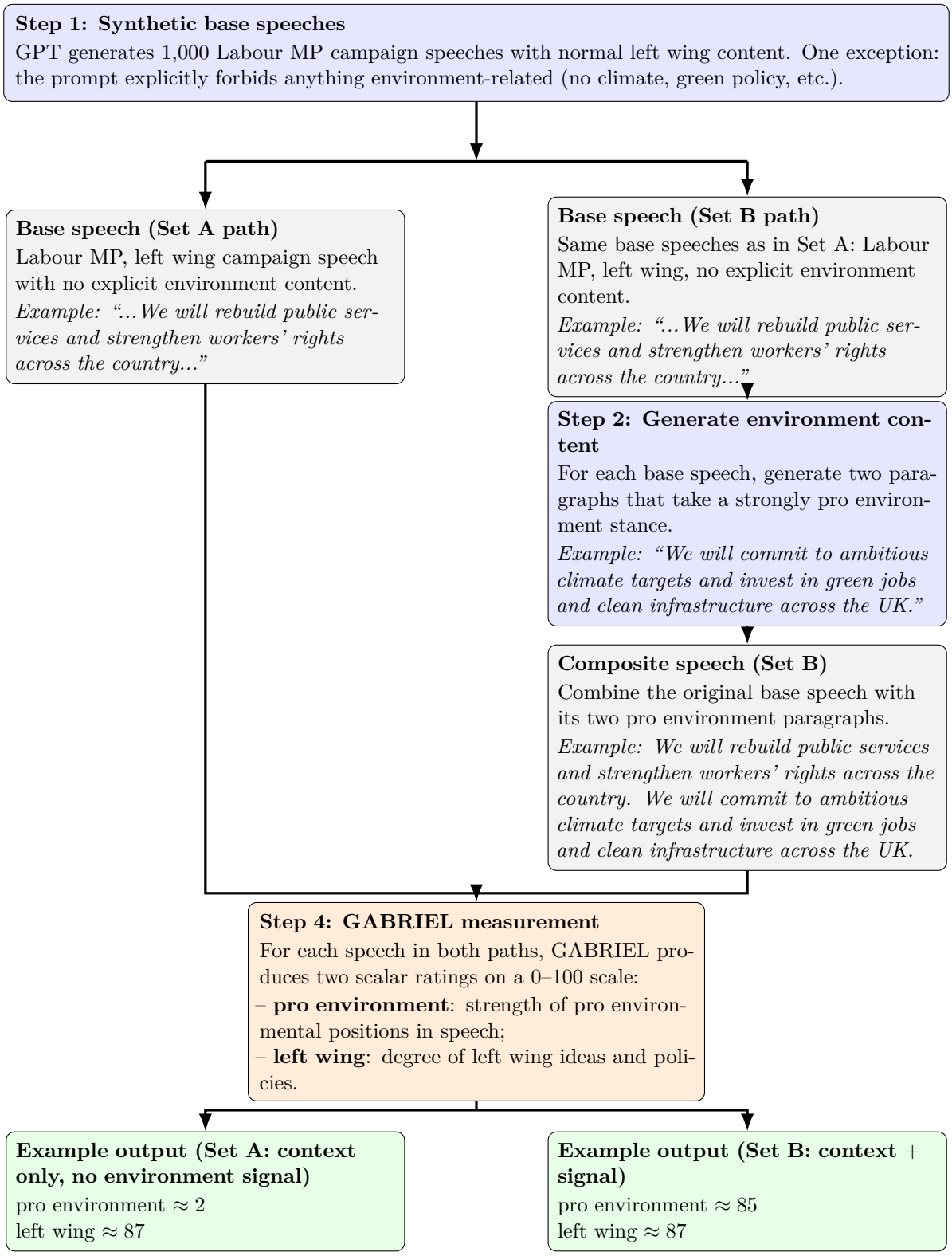


Figure 45: Schematic of the synthetic speech test for inference bias. Speech samples are only for illustration, in practice both base speeches and environmental content are much longer.

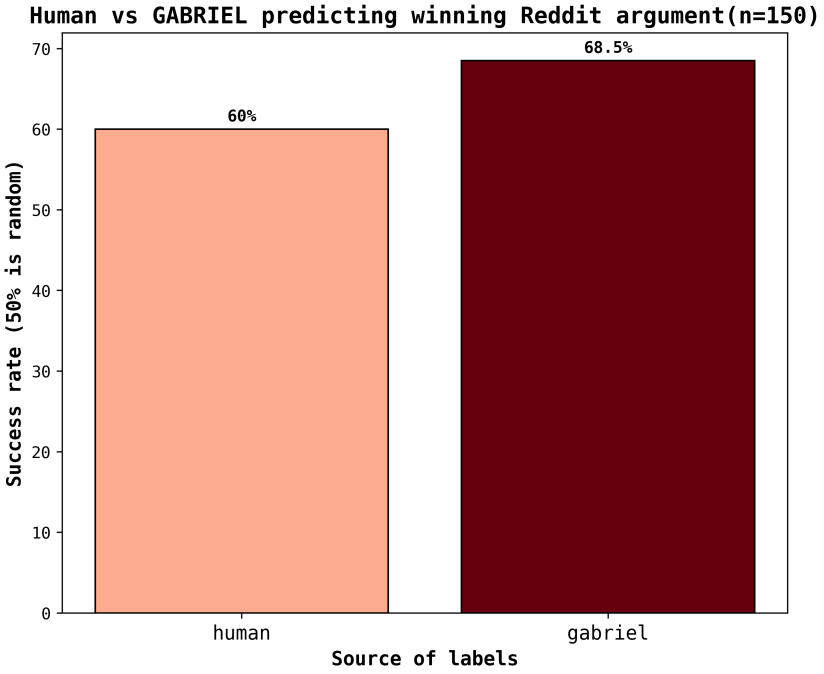


Figure 46: GABRIEL and a human labeler both try to estimate persuasiveness of an argument. GPT outperforms the human, indicating human / AI disagreement does not necessarily mean AI inferiority.

	Dependent variable: original rating
	(1)
debiased rating	0.900***
	(0.0071)
Intercept	5.435***
	(0.3351)
Observations	3,143
R^2	0.856
Note:	* p<0.1; ** p<0.05; *** p<0.01

Table 32: The correlation is important because most applications rely most on the relative ordering of entity ratings, not the raw value.

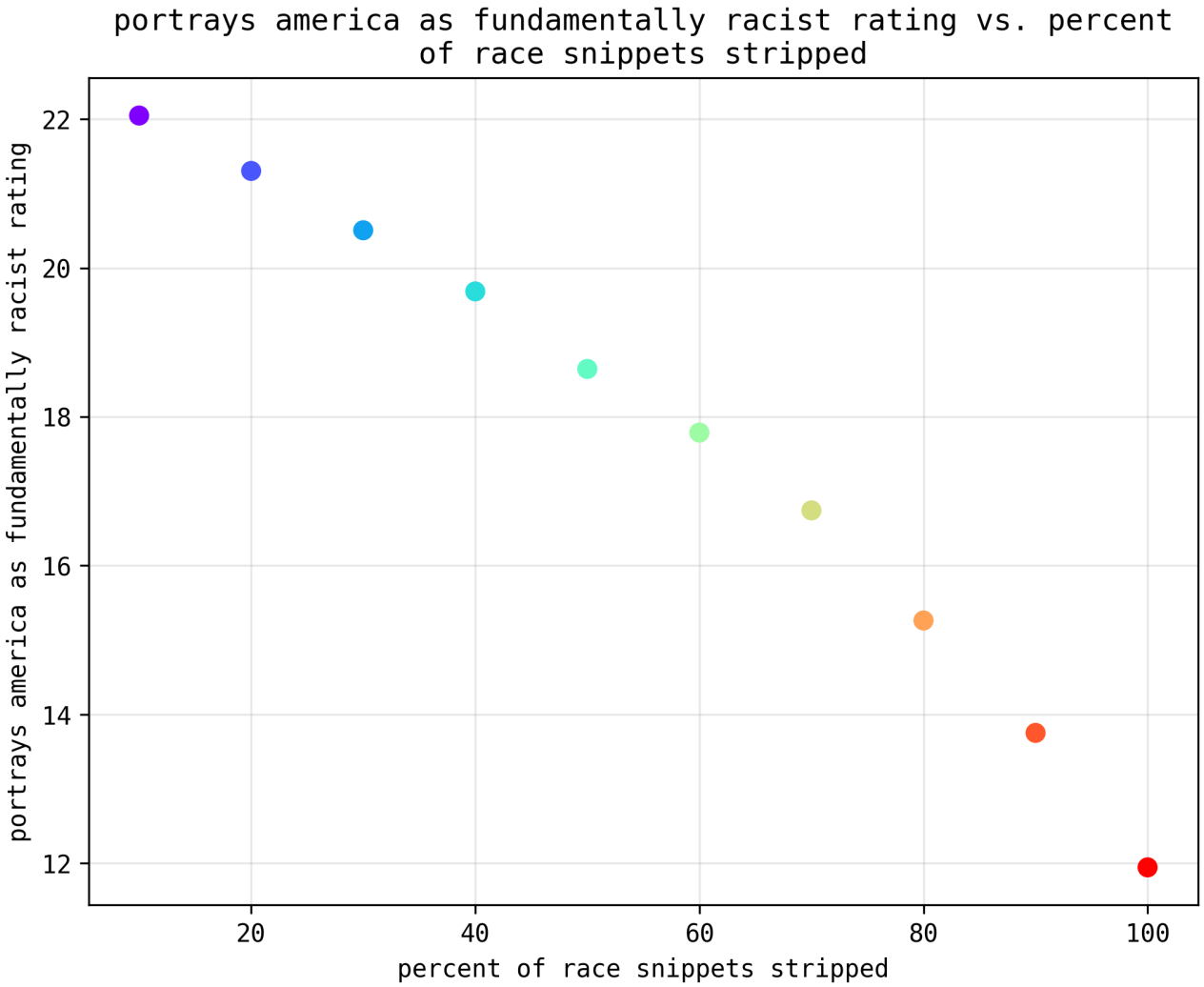


Figure 48: Increasing the percent of signal stripped correlates linearly with ratings attenuation. Note that 100 percent does not mean all race signal was removed, due to incomplete signal stripping.

A.4.1. Testing look-ahead bias on synthetic earnings calls

Look-ahead bias is a specific version of shortcut inference where knowledge of outcomes interferes with the LLM's measurements on historical text (Sarkar and Vafa, 2024). We build a test case where this effect would be particularly likely to manifest, if it is a significant concern. For 160 S&P 500 companies which we have yearly returns data in 2014, we synthetically generate 30 fake earnings call snippets per company (using `gabriel.whatever`). Each earnings call snippet is assigned a random number between 0 and 100. This "optimism injection" dictates how optimistic the snippet is (near 0 makes the snippet very pessimistic about the company's prospects; near 100 very optimistic). The snippets include the real company's name, concern its real business, and explicitly mention they are from Q4 2013. The substance is all fake, calibrated by the "optimism injection". We measure `good prospects for future stock performance` (using `gabriel.rate`) on each snippet. We wish for this to be measured purely on the content of the synthetic snippet, but GPT's broad world knowledge could allow it to infer a rating from how the company actually performed the following year.

Attribute	Definition
`good prospects for future stock performance`	How positive expected stock performance should be over the next 12 months based on the company's prospects as expressed in the earnings call snippet. High means a prediction of strong stock gains, low means a prediction of weak stock performance for the company.

Table 33: Attribute measured by GABRIEL in synthetic earnings call snippets.

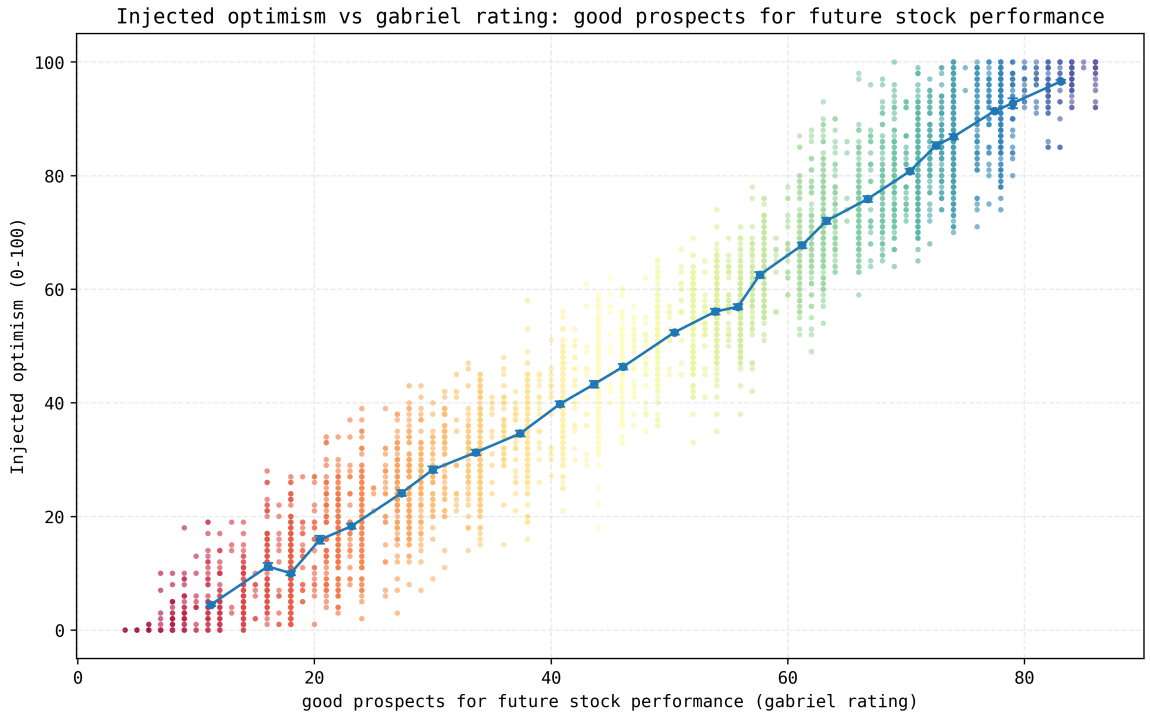


Figure 49: Our measurement of `good prospects for future stock performance` is closely correlated with the random "optimism injection". This indicates the measurement is grounded in the text.

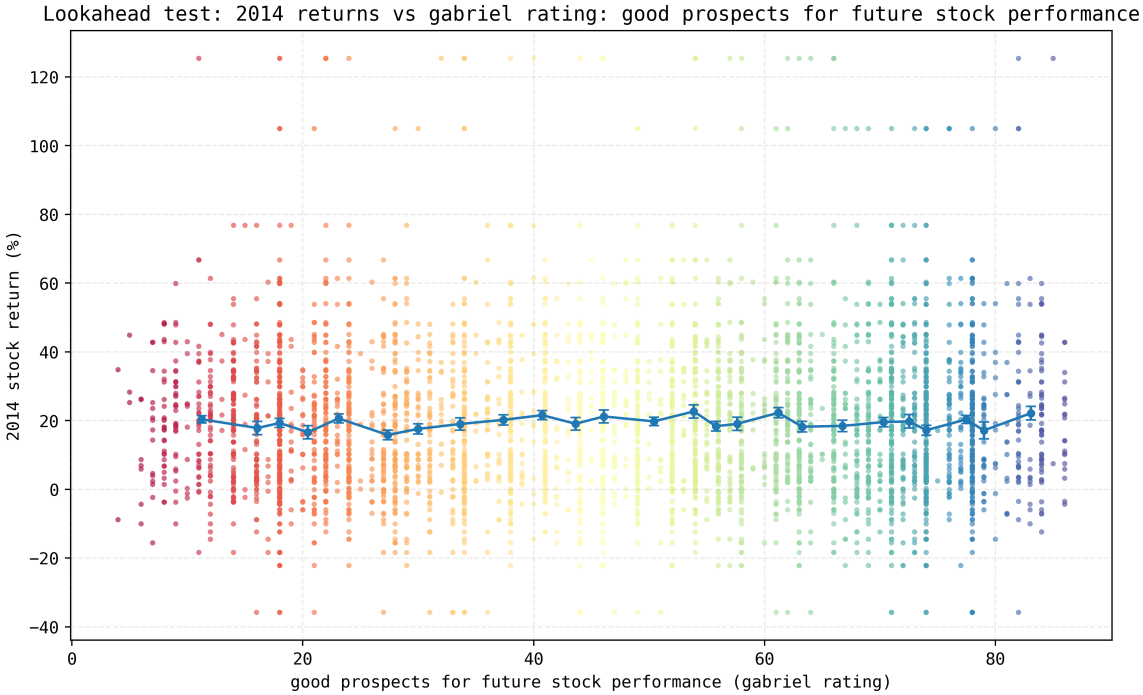


Figure 50: There is no evidence of significant look-ahead bias: our measurement of `good prospects for future stock performance` is highly correlated with synthetically controlled text optimism but uncorrelated with real returns the following year.

We find that `good prospects for future stock performance` is highly correlated with the "optimism injection" but not correlated at all with real returns the following year. This does not prove there is no look-ahead bias, but it does indicate such effects are small, if they exist. GPT likely knows subsequent real performance for these companies. But when we ask it to measure or predict based on the text, it is strongly anchored in the text. Next we test shortcut inference more broadly.

A.5. Additional figures for Section 5: Technology adoption

Attribute	Definition
year of invention	The singular, precise year that this technology was first invented or the first working prototype was built. The year of first prototype/invention is the year that the first working prototype of this technology was created (NOT some unrealistic concept work but a working, legitimate prototype); it will usually be the same year that this technology is commonly agreed to have been invented. The prototype may have been subpar in performance and capability, and it may not have been publicly revealed at the time. Careful consideration of the internal development and earliest versions of prototypes must be done to ensure you are isolating the true earliest working prototype and not just an early product. This is usually NOT the year of the first product release, but the year of the invention/first working prototype which comes before, sometimes well before. Report the precise year of invention alone (no rounding).
year of wide adoption by target audience	The single, specific year that marks widespread adoption and use of this technology by its target audience. By this year, the tech has clearly become widely adopted and accepted as normal by its target audience, and it has reached a level of polish sufficient for it to be genuinely useful and productively beneficial to users in its intended role. Importantly, the technology could still grow and improve after this year (e.g., the iPhone was widely adopted by 2012 even though it improved iteratively afterwards). This is the year it was complete and functional enough to have merited AND succeeded at acquiring mass adoption by its target audience.
inventor	The full name of the person most responsible for inventing this technology. Report the inventor name alone.
institution of invention	The institution where this technology was first invented. Report the institution name alone. Report "unknown" if you do not know, or if there was no specific institution.
institution type	The type of institution where this technology was first invented—report one of "independent inventor," "corporate," "university," "government," or "military."
country of invention	The country within which this technology was first invented (NOT the birthplace of the inventor, but where the invention was first made). Report the country name in standard modern English ("United States," "China," "United Kingdom," "Japan," etc.), even if the country went by a different name at the time of invention.
subregion of invention	The subregion/prefecture/province/state within the country where this technology was first invented. Report the subregion name in standard modern English ("California," "Bavaria," "New York," "Kanagawa," "Texas," etc.), even if it went by a different name at the time of invention.
unit price in current USD	The price of one unit of this technology, as it was at the time of initial wide adoption, converted as best you can approximate to current USD (after logical conversion and inflation adjustment). Report the estimated price alone, without the currency symbol.
prerequisite technologies	The specific essential technologies that had to be invented or developed first before this technology could have existed. This could be a short or long list depending on the technology. It should include all major technologies directly or indirectly necessary for this one. Report all prerequisite technologies as a comma-separated list (A, B, C, etc).

Table 34: Definitions for extracted attributes about each technology.

Attribute	Definition
instructions for all attributes	Be meticulous, accurate, precise, and truthful in providing the information requested. If you do not know any given attribute, report "unknown." Be especially careful and precise in determining **year of invention** and **year of wide adoption**. If you only have a narrow range, interpolate the most plausible single year based on evidence. Do not round years; report the exact year. Truthfulness and accuracy are paramount — triple-check your work.
examples and guidance	Examples: iPhone – 2005 (invention), 2012 (wide adoption); Nuclear Fission Bomb – 1942 (invention), 1952 (wide adoption); Incandescent Light Bulb – 1838 (invention), 1924 (wide adoption); VCR – 1986 (wide adoption). Use these examples to guide your reasoning about invention versus adoption.
initial filtering condition	The entity must be a historically significant technology. Decide whether each entity is an actual technology that was invented or fabricated by human inventors, or whether it is not a technology (in which case do not include it in output). Historically significant means the technology has had a meaningful impact on the world, of significance in technological history. It is not a technology if it is a concept, broad field of research, a scientific theory, a history, a natural commodity, etc. It is a technology if it is a tangiable device, product, or equipment which would normally considered technology or a tool (e.g. iPhone, steel plow, cotton gin, etc). It is also a technology if it is a material or chemical fabrication (e.g. drugs, building materials, GMO strains of corn, etc). It is also a technology if it is a specific process by which things are made (e.g. Bessemer steel process, assembly line, interchangeable parts, specific invented agricultural practices/setups, etc). It is a technology if it is software, a fabrication, a tangiable invented tech or tool, a specific invented manufacturing/agricultural/service process, etc. Examples of non-technologies: 'Manhattan project', 'corn', 'chair', 'history of the automobile', 'coal', 'astronomy', 'crop', 'bacteria', 'enginnering', 'agricultural machinery', 'nuclear science', 'drought tolerance', 'reusablility', 'information science', 'energy return on investment', 'boolean algebra', 'agroecology', etc are NOT technologies. Examples of insufficiently historically significant or overly niche tech that should not be included: 'ZZZ (video game)', 'HMCS Royal Mount (K677)', 'Xbox One Elite controller', 'Xinmin-Tongliao high-speed railway', 'Samsung Galaxy Core Advance', 'Asus Eee PC 1201N', 'KSR Bengaluru-Dharwad Vande Bharat Express', 'Nexium IV', 'JAC Tongyue EV', 'Toyota 3S-FE engine', etc should NOT be included. On the other hand, 'nuclear fission bomb', 'GMOs', 'GMO corn', 'jeans', 'cotton candy', 'strip mining', 'Ford Model T', 'infrared telescope', 'Wikipedia', 'MOSFET', 'Google', 'UNIX', 'methylphenidate', 'nylon', 'steel plow', 'fish farming', 'IR8 rice', 'botnet', 'CSS', 'social media', 'online dating', 'bra', 'internet forum', 'object oriented programming', 'book', 'suede', 'calendar', 'assembly line', 'smartphone portrait photography mode', 'lava lamp', 'containerization', 'electrification', 'electromagnet', 'laser', etc are technologies. Ensure you capture all historically significant technologies, missing none. Do not include any non-technologies or non-historically significant technologies in output. Have a high bar for historical significance. Also avoid niche, specific product editions of what is actually a broader generic class of technology. Pay attention to the examples given above to aid your judgement. Only include the technologies clearly meeting the above criteria in output, verbatim.

Table 35: Various instructions used in the pipeline.

Attribute	Definition
dramatically increased worker productivity	The 'dramatically increased worker productivity' attribute is high when the technology in question ultimately contributed to significant gains in the economic productivity of each US worker using it. The tech should have made people much better at completing their work, and trickled down to benefiting the productivity of other workers. Simple tools like the plow or the sewing machine or complex tools like a cotton gin or electricity would score highly. These techs massively increased the productivity of their users. A technology with the greater increase on economic productivity would have greatly improved what workers are capable of in a way that has been historically documented (not just conjectured), while not majorly creating negatives or distractions which impede productivity (e.g. smartphone). This should be considered throughout the lifespan of the technology, not just at the beginning or the end. Even if the tech started slow or has since become obsolete, as long as it had a recorded dramatic increase in total factor productivity for workers at some point in history and/or majorly contributed to the advent of later technologies which did so, it should score highly. Do not underweight 19th and early 20th century tech; be sure to consider their massive impacts at the time and not have recency bias.
easy to update	The 'easy to update' attribute is high when the original version of the technology, near the time of invention, was easy to modify, improve, and fix by the producer of the tech. It would also be high if the user of the technology could easily modify and upgrade the tech. A technology with more ease of updating would be more flexible to modification, easier to build upon, and without high capital costs for creating new versions.
requires highly specialized training	The 'requires highly specialized training' attribute is high when the tech in question, near the time of invention, could only be operated by specially trained and/or highly skilled workers.
intense competition to develop	The 'intense competition to develop' attribute is high when the technology in question involved significant rivalries or competitive efforts among individuals, companies, or nations during its development phase.
inventor was highly eccentric	The 'inventor was highly eccentric' attribute is high when the creator or key developer of the technology was known for unconventional behavior, odd mannerisms, carelessness about profits, or generally not fitting into society well.
long time to install	The 'long time to install' attribute is high when the technology in question, near the time of invention, took a lot of time to install before use by the user.
expensive	The 'expensive' attribute is high when the tech in question, near the time of invention, was very costly to acquire.
productive usage requires business reorganization	The 'productive usage requires business reorganization' attribute is high when using the tech in its early days, especially by businesses and for productive purposes, required changing the informational and organizational structure of the business.
productive usage requires new infrastructure	The 'productive usage requires new infrastructure' attribute is high when implementing the tech by the user, near the time of invention, required new buildings, rooms, warehouses, or roads.

Table 36: Definitions for attributes / characteristics measured on each technology. (1/2)

Attribute	Definition
high fixed costs in product development	The 'high fixed costs in product development' attribute is high when bringing the technology from concept to a viable first product required very large upfront spending before any meaningful revenue: e.g., costly R&D, prototyping, tooling, specialized facilities, regulatory approvals, and engineering headcount.
reliant on network effects	The 'reliant on network effects' attribute is high when the technology's usefulness to each user rises sharply with the number of other users (or complementary nodes), such that early adoption stalled without a critical mass.
early mass manufacturing challenges	The 'early mass manufacturing challenges' attribute is high when, near the time of invention, scaling production from lab/prototype to volume faced serious hurdles: low yields, tight tolerances, hard-to-standardize processes, expensive QA, immature tooling, or scarce skilled process engineers.
well financed	The 'well financed' attribute is high when the technology's early development was backed by deep and complete funding.
strong geopolitical incentives	The 'strong geopolitical incentives' attribute is high when countries or blocs had clear national-security, prestige, deterrence, or strategic-trade reasons to develop and deploy the technology quickly, producing races, state subsidies, secrecy/export controls, and mission-driven programs.
reliant on academic research	The 'reliant on academic research' attribute is high when key enabling insights, methods, datasets, or proofs-of-concept emerged from universities or public labs, and progress depended on academic breakthroughs to unblock productization.
easy to use	The 'easy to use' attribute is high when, near the time of early commercialization, typical users could achieve competent operation with minimal instruction, low cognitive load, and little specialized equipment.
widespread excitement by target audience	The 'widespread excitement by target audience' attribute is high when the intended audience (not necessarily the general public) showed strong enthusiasm during development or early launch—evidenced by preorders, pilot waitlists, trade-press buzz, professional endorsements, and rapid trials.
reliant on complex supply chains	The 'reliant on complex supply chains' attribute is high when the technology, at introduction, required multi-tier, geographically dispersed suppliers with specialized inputs (materials, precision components, tooling, software), such that coordination, lead times, and vendor concentration were binding constraints.
largely iterative	The 'largely iterative' attribute is high when the technology was primarily an incremental improvement (performance, cost, reliability, form factor) on an existing lineage rather than a novel paradigm.
large and bulky	The 'large and bulky' attribute is high when the tech in question, near the time of invention, was physically large and bulky.
contains many parts	The 'contains many parts' attribute is high when the technology in question, near the time of invention, consists of numerous components, subsystems, or parts. A technology with a higher score in this attribute would be characterized by its complexity and the integration of multiple elements to function effectively.
useless in early days	The 'useless in early days' attribute is high for technologies that, at the time of their initial development, had little to no practical application or discernible utility.

Table 37: Definitions for attributes / characteristics measured on each technology. (2/2)

A.6. No additional figures for Section 6: Conclusion

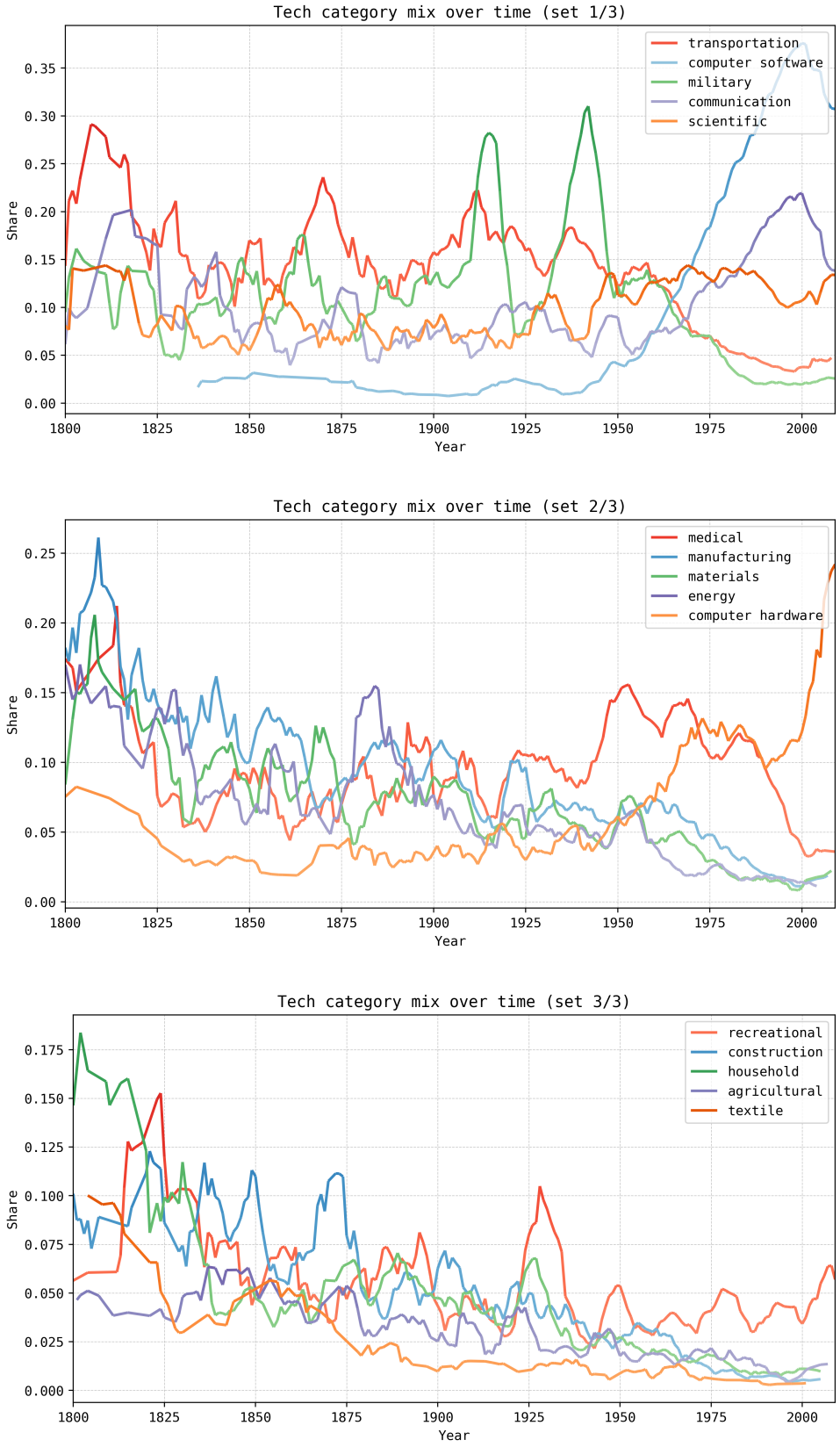


Figure 51: The nature of technologies has evolved greatly over time.

B. Best practices for using GPT as a measurement tool

Below, we outline some best practices for using GPT / GABRIEL as part of social science research methods. These are derived from our experience using GPT in dozens of projects, and also informed by our validation experiments.

Dream big and apply liberally — there is so much qualitative data and so many angles to assess it from. GPT should not only aid your existing research methods — its existence should alter the research questions themselves. Intelligent language models mean that many more hypotheses are testable, because a much larger corpus of human behavioral data is measurable. This is particularly true for researchers with less resources: the text and image data already exists online, and the models are cheap. Deploying GPT to measure attributes on a dataset is easy and near costless. This makes GPT ideal for prototyping and preliminary investigation on new research questions, gaining valid measurements of complex human concepts in short order. This allows more iteration and exploration by the researcher.

GPT is a measurement tool; you must still use its measurement soundly. Using GPT in research can sometimes feel like black magic. GPT usage can give the impression that shortcuts were taken or statistical validity is lessened.

We dispute this. It is essential to appropriately scope what exactly GPT is offering here. In this paper, we employ GABRIEL only as a tool to measure and quantify attributes, and we confirm that these measurements are valid and unbiased.

However, just like any other datapoint in a spreadsheet, valid measurements from GPT can be used in a statistically invalid manner. These validity problems might relate to GPT use (see later section on p-hacking) but more often it is the whole host of traditional validity concerns that any quantitatively rigorous paper might encounter. It is important to treat the validity of GPT measurements as separate from overall statistical validity. GPT is not a shortcut past concerns that would apply to any other measured data. On the other hand, GPT evaluations should be respected as scientific measurements, as human labeling might be and not completely distinct from what a voltmeter does. GABRIEL ratings should be treated as any other measured data would be — a valid measurement (which we show is generally true) used in a valid manner (which is highly dependent on the specific data and research method).

Peruse samples and understand your corpus before applying GPT. It is hard to know what attributes to measure without first reading a number of samples from the dataset. A proclaimed set of political tweets could be long and discursive or brief and poorly filtered. The shorter each text is, the simpler the attributes should be — or multiple texts could be grouped together by author to reduce noise (e.g. grouping 15 tweets by a single user into a single "text" datapoint for GABRIEL). The samples should inform what is possible to measure, and how an attribute should be defined to measure what you are interested in.

Play with your data inside ChatGPT first; if your task doesn't work there, it might not work at scale with GABRIEL. As mentioned before, using GABRIEL (or any similar implementation of GPT) to measure attributes or parse qualitative data is not a technically complicated task. Simply taking a few samples from your dataset, feeding them into ChatGPT or another chatbot, and asking the chatbot to perform the task you are interested in should give you similar results to GABRIEL.

GABRIEL is an exercise in standardization and scale; there is no secret sauce beyond the capabilities of today's AI models. This makes chatbots a good playground to tweak attribute choice and definitions. A rule of thumb is to use ChatGPT — or a GABRIEL run on a small, manually inspected sample — to ensure measurement is possible and to isolate the quantity you desire.

Small scale manual validation helps ensure GPT is measuring what you want to measure, but it is not necessary in many cases. We show that GPT measurements are broadly accurate, at least to a human standard and perhaps beyond. We find this to be true across hundreds of very different tasks with no priming or any sort of manual intervention. These results (and others which demonstrate GPT as a general purpose comprehension machine) are powerful indicators that some essential practices from prior machine learning usage in research are perhaps not necessary anymore.

We believe manual validation on small samples of data is still useful, even if only on a handful of datapoints to be sure attributes are defined and measured in the way you want them to. But for most cases, our broad validation exercise indicates accuracy can be expected — just as if the tasks were assigned to a human annotator — without comprehensive manual validation. We leave it to the judgment of the researcher on whether what they are measuring is similar enough the broad range of comprehension tasks we demonstrate high accuracy on. Most labeling tasks likely fall under this umbrella.

p-hacking is a much bigger concern when the cost of measuring many attributes is so low — leave out a test set and document all experiments to avoid. GPT makes the cost of measuring attributes very low. It is easy to test out dozens or even hundreds of attributes; it is also trivial to run one attribute with many different definitions. Intentionally or not, this presents a major concern for p-hacking, where hundreds of attributes or attribute variance are run, and only those reaching statistical significance are reported.

This is a novel problem but not unprecedented. There are numerous other situations where a glut of measurements exist and can be cherry picked (e.g. Census or World Bank variables). We again put forward that GPT is a measurement tool and just like those other situations, the measurement must be used in a statistically valid way.

Here we recommend three important approaches to consider. First, this problem has been a major one in computer science and machine learning for a while. ML involves high dimensional relationships between inputs and outputs. A classic problem when training ML models is overfitting, where there are too many parameters/coefficients in the model for the amount of data. Overfitting means that

the parameters / coefficients of the model "memorize" the training data, rather than learning useful heuristics that generalize well. This leads to models which are extremely accurate on the training data, but do not perform well on other data sampled from the same population.

A key insight from machine learning is the use of separate training and testing data. ML researchers almost always leave out a portion of their overall data corpus from all training steps. They only test their model's performance on this data after training is completed. If the model is overfit, it will perform significantly worse on the test data it has never seen before. If it is not overfit, it has learned useful, generalizable heuristics that apply to the whole population.

We recommend a similar strategy when using GPT to analyze qualitative data, especially if quantifying many attributes. If possible, it is best to conduct all experiments, attribute selection, and significance tests on a "training" random subsample of the corpus (say, 50%). Then, only after attributes are finalized should they be run on the left out "testing" set, and the significance of these results should be reported. A third validation sample can be used during "training", also left out but used periodically to ensure there is no overfitting.

Like with ML models, our concern here is a high dimensional parameter space — too many attributes could just be reflecting statistical noise and not significance. Overly refined attributes could just be learning quirks of the training data, not the population. Leaving out a test set for evaluation assuages these concerns.

Our second recommendation is to document all attributes that are tested, any significant variants in attribute definitions, and track the overall count of attributes measured. This mainly applies to situations where many attributes are measured, there are few datapoints, or the relevant regression/significance test is not strong.

Third, we recommend considering the appropriate significance threshold for your situation. For example, if hundreds of attributes are measured (which may be correct and interesting), a stricter significance threshold might be in order.

Don't spend too much time trying to find the perfect prompt — but ensure your attributes are clear, parsimonious, and well defined. We have found — and we show in our validation section — that GPT is robust to different ways of writing the same prompt. Like a human, it understands the idea of what is being asked and is not overly anchored to the precise text. Good prompting is helpful, and concise prompting keeps costs down on large runs. However, we find it unnecessary to fixate on precise prompt wording.

Leverage scale and parallelization. The truly unique power of GPT here is not intelligence, but cheap intelligence at enormous scale and speed. This allows enormous qualitative corpora to be analyzed. The value proposition of a GPT is quite enhanced if your approach scales well to thousands of GPT instances reviewing thousands of documents separately. In particular, parallelization dramatically speeds up the process. GABRIEL is built so that hundreds of separate GPT calls are

running simultaneously, each rating attributes on its own passage from the corpus. It is also better practice to use generalizable prompts where different entities/passages/attributes can be substituted in programmatically. Prompt templates, parallelization, and structured outputs like JSON allows GPT's intelligence to scale to very large corpora.

GPT excels at descriptive statistics and filtering data. Even before any analysis, passing a qualitative dataset through GABRIEL is a good way to understand the data and get descriptive trends. Measuring attributes (e.g. "formality" or "uses misinformation") or extracting information (e.g. "year of invention" or "university affiliation") on a qualitative dataset can reveal important trends over space, time, and type.

A major advantage of GPT based descriptive statistics is that they are parsimonious. Unlike standard metadata (e.g. number of posts, word count, etc), GPT can measure interesting and interpretable signals. These descriptive statistics are often novel in their own right, and are useful in guiding further inquiry. They also help cut the data on conceptual lines — if you have a large dataset of parliamentary speeches and you wish only to analyze those pertaining to environmentalism, you can classify and screen for that first and subset to those speeches.

Rule of thumb: could you measure what you are asking GPT to measure? A simple question to ask yourself when asking GPT to measure something is whether you would be able to measure that same thing yourself, given copious time on any given datapoint. It is true GPT can notice subtle patterns a human would miss in the noise, if only because it can be run across a much larger dataset. Nevertheless, this rule of thumb is a good place to start when conceiving of a GPT measurement.

C. Methods explained in detail

C.1. Systematic replication pipeline

Here, we detail the exact steps used to generate our large-scale replication corpus. See Figure 12 for a schematic.

1. Begin with all 317,000 text databases hosted on HuggingFace.
2. We filter these to those tagged as having either "expert-generated" or "crowdsourced" or "human-curated" annotations, then filter only to those with open licenses (MIT, Apache, CC, etc).
3. We then run a classification layer to check whether the dataset as hosted on HuggingFace is:
   (a) Document-level data
   (b) With gold-standard annotations
   (c) Text-only inputs (no multimodality)

(d) Not a QA task (like a multiple question quiz — this is not what we want to evaluate)

(e) Complete as hosted on the platform.

4. We then download, clean, and reformat each candidate dataset.

5. We dynamically generate attribute names and definitions which best match the variables of interest based on the context.

6. We replicate the labels using GABRIEL and evaluate the performance of our model.

C.2. Validity of Wikipedia

For our purposes, Wikipedia is largely representative. Research shows that article authorship is broadly spread around the world. While a Western bias does exist and 15% of authors are from the US, this distribution is in line with the global GDP distribution, which might matter most in sourcing technologies. We observe that the existence of a Wikipedia article is strongly predicted by that entity being of historical significance. Top cited economists are almost all on Wikipedia; there is a strong linear decline in Wikipedia likelihood as citation count decreases (Figure 52). This validates Wikipedia as a good source of important entities, since it is predicted here by a ground truth source of importance (citations).

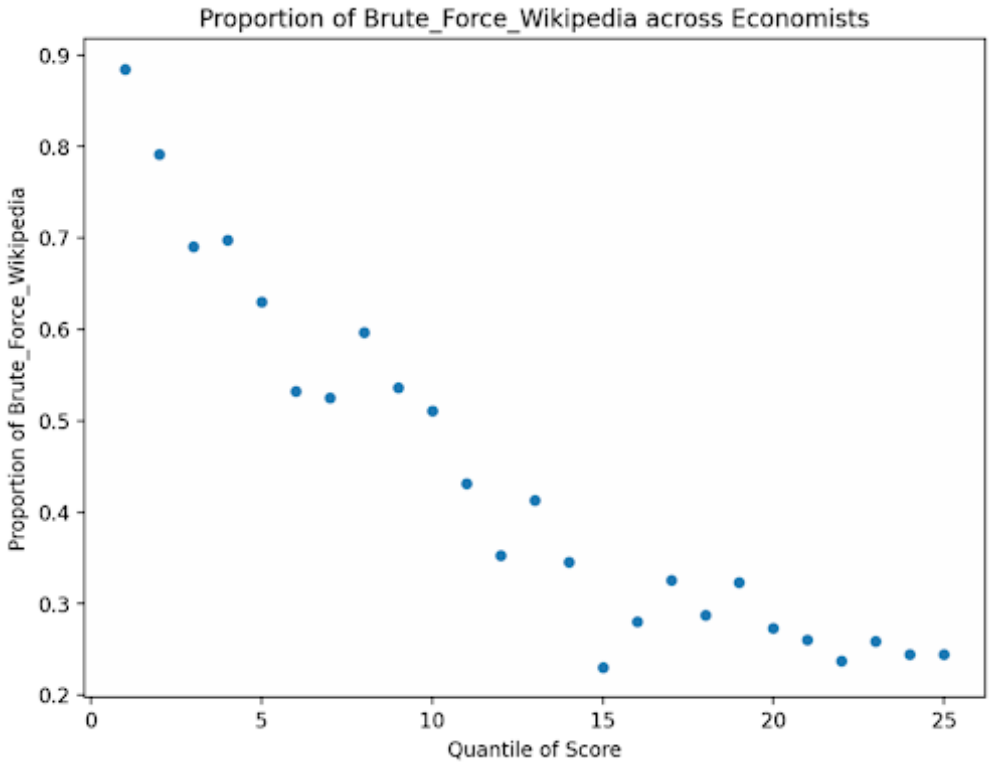


Figure 52: Matching economist names to Wikipedia articles - more cited economists are more present

Finally, in Figure 53, we apply a similar method to Supreme Court cases. We find that reference count

by lower courts to prior Court cases — a real world signal of historical significance — is well correlated with the likelihood of a Wikipedia article about that Court case.

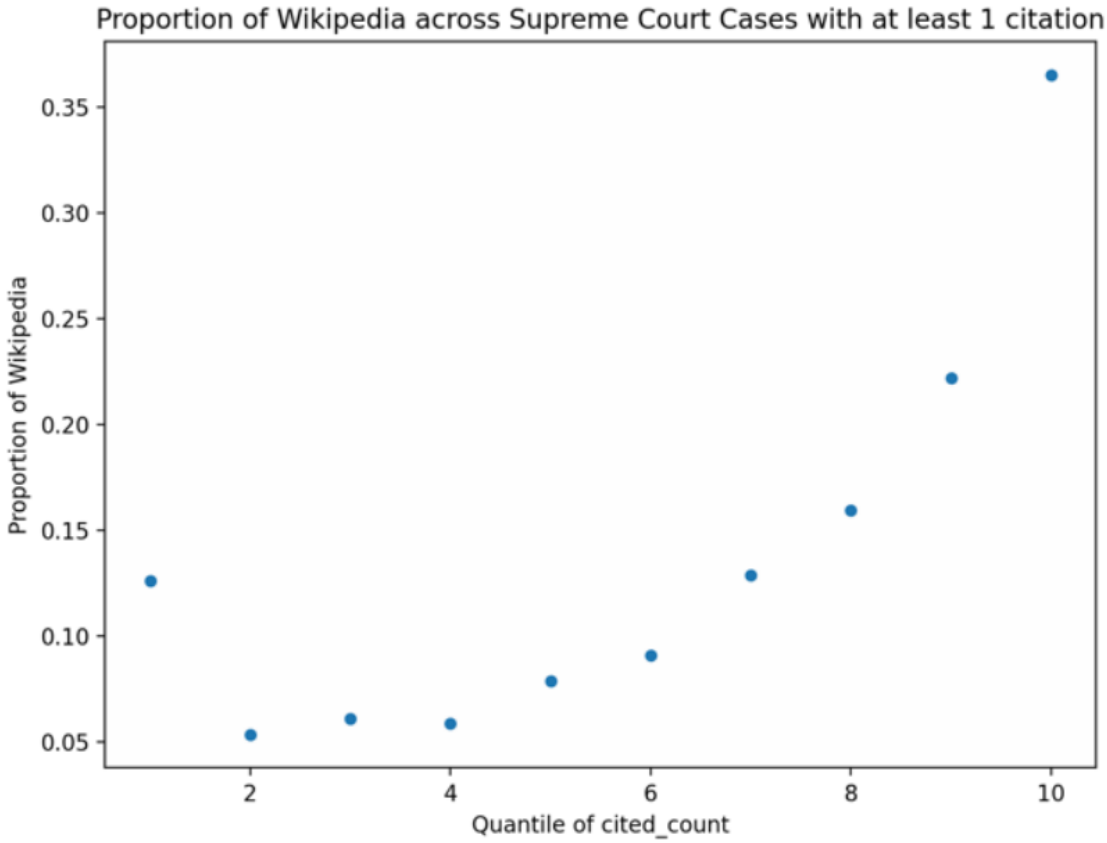


Figure 53: Matching Supreme Court cases to Wikipedia articles - more cited cases are more present

Importantly, this data set allows us to analyze the distribution over time. Ideally, we want an unbiased dataset over the Industrial era where the likelihood of a case appearing on Wikipedia is strongly correlated with its reference count and not correlated with its year. We do see a significant correlation with year, meaning that newer cases are more likely to have Wikipedia articles. However, the distribution is still quite representative of the overall time period and recency bias does not dominate. Further filtering attempts to address these remaining biases later.

D. Statistical tests

Setup and weighting. We analyze model performance on classification tasks with per–observation weights given by the number of underlying instances, $w_i = n_i$. We retain rows with finite metric values and positive weights only. For each model j, we define its knowledge cutoff date c_j and split observations by timestamp t_i into *pre* ($t_i \le c_j$) and *post* ($t_i > c_j$) groups. When reporting sample sizes we show both the raw counts $(n_{\text{pre}}, n_{\text{post}})$ used in estimation and, internally for inference, the Kish effective sizes.

D.0.1. Pre–post difference in weighted means (Welch test)

For a given performance metric $x_i \in [0,1]$ (F1 or accuracy), we estimate for each model j the pre and post *weighted* means

$$\bar{x}_w^{\text{pre}} = \frac{\sum_{i\in\text{pre}} w_i x_i}{\sum_{i\in\text{pre}} w_i}, \qquad \bar{x}_w^{\text{post}} = \frac{\sum_{i\in\text{post}} w_i x_i}{\sum_{i\in\text{post}} w_i},$$

and the difference $\Delta_j = \bar{x}_w^{\text{post}} - \bar{x}_w^{\text{pre}}$.

We construct a two–sided Welch (unequal–variance) test and confidence interval for Δ_j using unbiased weighted variances and effective sample sizes. For group $g \in \{\text{pre}, \text{post}\}$ with weights $\{w_i\}$ and values $\{x_i\}$,

$$s_{w,g}^2 \;=\; \frac{\sum_{i\in g} w_i \,(x_i - \bar{x}_{w,g})^2}{\sum_{i\in g} w_i \;-\; \frac{\sum_{i\in g} w_i^2}{\sum_{i\in g} w_i}}, \qquad n_g^{\text{eff}} \;=\; \frac{\left(\sum_{i\in g} w_i\right)^2}{\sum_{i\in g} w_i^2}.$$

The variance of the weighted mean is approximated by $s_{w,g}^2/n_g^{\text{eff}}$, yielding

$$\text{Var}(\Delta_j) \;\approx\; \frac{s_{w,\text{pre}}^2}{n_{\text{pre}}^{\text{eff}}} + \frac{s_{w,\text{post}}^2}{n_{\text{post}}^{\text{eff}}}, \quad \text{SE}(\Delta_j) = \sqrt{\text{Var}(\Delta_j)}.$$

Degrees of freedom follow the Welch–Satterthwaite approximation,

$$\text{df} \;=\; \frac{\left(\frac{s_{w,\text{pre}}^2}{n_{\text{pre}}^{\text{eff}}} + \frac{s_{w,\text{post}}^2}{n_{\text{post}}^{\text{eff}}}\right)^2}{\frac{\left(\frac{s_{w,\text{pre}}^2}{n_{\text{pre}}^{\text{eff}}}\right)^2}{n_{\text{pre}}^{\text{eff}}-1} \;+\; \frac{\left(\frac{s_{w,\text{post}}^2}{n_{\text{post}}^{\text{eff}}}\right)^2}{n_{\text{post}}^{\text{eff}}-1}}.$$

We report two–sided p–values for $H_0 : \Delta_j = 0$ and $100(1-\alpha)\%$ confidence intervals $\Delta_j \pm t_{1-\alpha/2,\text{df}} \cdot \text{SE}(\Delta_j)$ (with $\alpha = 0.05$). If either group's effective size is too small, the test and CI are not reported (set to NA). Cutoff dates and $(n_{\text{pre}}, n_{\text{post}})$ are shown alongside each estimate.

D.0.2. Descriptive model summaries (weighted moments and quantiles)

Independently of pre/post splits, we summarize each model's distribution across classification tasks using *weighted* moments and quantiles. For metric values $\{x_i\}$ and weights $\{w_i\}$ we compute the weighted mean and median, as well as the interquartile interval $\text{IQR} = [Q_{0.25}, Q_{0.75}]$ from weighted percentiles. Weighted quantiles are obtained by sorting x_i, forming normalized cumulative weights $C_k = \frac{\sum_{i\le k} w_i}{\sum_i w_i}$, and linearly interpolating the desired probability level p on C_k to recover Q_p. In tables we present, for each model, the weighted mean, the weighted median, and the IQR as $[Q_{0.25}, Q_{0.75}]$; the best (largest) mean and median are highlighted in bold. All summaries restrict to rows with finite metric values and positive weights. We filter to observations whose task label includes "classification" and use $w_i = n_i$ for weighting.

E. Cost comparisons

For the SOTU line, we tokenize the entirety of the source corpus of SOTUS from jsulz (2025) and find $T_{\text{in}} = 2{,}286{,}149$ input tokens across $N = 240$ addresses. We also compute total words directly from the corpus text, yielding 1,973,012 words. We fix 250 output tokens per address ($T_{\text{out}} = 60{,}000$), allowing for small amounts of reasoning.

Model costs follow current pricing:

$$\text{Cost} = \left(\tfrac{T_{\text{in}}}{10^6}\right) p_{\text{in}} + \left(\tfrac{T_{\text{out}}}{10^6}\right) p_{\text{out}},$$

e.g., for `gpt-5-mini`: $(2.286149) \cdot \$0.25 + (0.060) \cdot \$2.00 = \$0.6915 \approx \$0.69$ (OpenAI, 2025a).

For the human benchmark we compute total words directly from the corpus text, divide by the meta-analytic adult silent reading speed (non-fiction $\approx$ 238 wpm), add $\sim$ 2 minutes per item for rubric/entry, and apply a \$15/hour target wage plus a 20% platform fee—yielding $\sim$ 146 labor-hours and $\approx$ \$2,631 for a single-rater pass (Brysbaert, 2019).

For sermons, we assume 100,000 items at $\sim$5,000 words each (based on Pew's nationwide study of online sermons reporting a median $\sim$37-minute sermon, i.e., $\sim$5–6k words). To translate words to tokens for model-pricing calculations, we use the common heuristic 1 token $\approx$ 0.75 words, implying $\approx$ 666.7M input tokens and 25M output tokens (250 per item), priced with the same formula (Pew Research Center, 2019).

$$\begin{aligned}\text{human cost (SOTU)} = {} & \underbrace{1{,}973{,}012}_{\text{words}} \times \frac{1}{\underbrace{238}_{\text{wpm}}} \times \frac{1}{60} \times \underbrace{15}_{\$/\text{hour}} \times \underbrace{1.2}_{\text{platform (+20\%)}} \\ & + \underbrace{240}_{N} \times \underbrace{2}_{\text{entry min/item}} \times \frac{1}{60} \times 15 \times 1.2 \\ \approx {} & \$\mathbf{2{,}631}.\end{aligned}$$

$$\begin{aligned}\text{human cost (sermons, 100,000 items)} = {} & \underbrace{500{,}000{,}000}_{\text{words (100,000}\times\text{5,000)}} \times \frac{1}{\underbrace{238}_{\text{wpm}}} \times \frac{1}{60} \times \underbrace{15}_{\$/\text{hour}} \times \underbrace{1.2}_{\text{platform (+20\%)}} \\ & + \underbrace{100{,}000}_{N} \times \underbrace{2}_{\text{entry min/item}} \times \frac{1}{60} \times 15 \times 1.2 \\ \approx {} & \$\mathbf{690{,}252}.\end{aligned}$$

$$
\begin{aligned}
\text{model cost (SOTU; nano)} &= \underbrace{240}_{N} \times \underbrace{\frac{2{,}286{,}149}{240}}_{\text{avg input tok/item}} \times \frac{p_{\text{in}}^{\text{nano}}}{10^6} + \underbrace{240}_{N} \times \underbrace{250}_{\text{output tok/item}} \times \frac{p_{\text{out}}^{\text{nano}}}{10^6} \\
&= \underbrace{2{,}286{,}149}_{T_{\text{in}}} \times \frac{p_{\text{in}}^{\text{nano}}}{10^6} + \underbrace{60{,}000}_{T_{\text{out}}} \times \frac{p_{\text{out}}^{\text{nano}}}{10^6} \\
&= 2{,}286{,}149 \times \frac{0.05}{10^6} + 60{,}000 \times \frac{0.4}{10^6} \\
&\approx \$\mathbf{0.14}.
\end{aligned}
$$

$$
\begin{aligned}
\text{model cost (sermons; nano)} &= \underbrace{100{,}000}_{N} \times \underbrace{\left(\frac{5{,}000}{0.75}\right)}_{\text{input tok/item} \approx 6{,}667} \times \frac{p_{\text{in}}^{\text{nano}}}{10^6} + \underbrace{100{,}000}_{N} \times \underbrace{250}_{\text{output tok/item}} \times \frac{p_{\text{out}}^{\text{nano}}}{10^6} \\
&= \underbrace{666{,}666{,}667}_{T_{\text{in}}\ (\text{approx})} \times \frac{p_{\text{in}}^{\text{nano}}}{10^6} + \underbrace{25{,}000{,}000}_{T_{\text{out}}} \times \frac{p_{\text{out}}^{\text{nano}}}{10^6} \\
&= 666{,}666{,}667 \times \frac{0.05}{10^6} + 25{,}000{,}000 \times \frac{0.4}{10^6} \\
&\approx \$\mathbf{43}.
\end{aligned}
$$

Table 38: Input cost (USD) per 1M tokens: human vs. GPT models. Full detail on the calculations is provided in Appendix E.

Method / Model	**Cost per 1M input tokens**
Human (wage-only; 238 wpm, $15/h)	$662.50
Human (+20% platform)	$795.00
GPT-5 nano (API input rate)	$0.05
GPT-5 mini (API input rate)	$0.25
GPT-5 (API input rate)	$1.25