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). GPT is evaluated on classification and attribute rating performance 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 generally indistinguishable from human evaluators. Our evidence indicates that labeling results do not depend on the exact prompting strategy used, and that GPT is not relying on training data contamination or inferring attributes from other attributes. 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. Our analysis documents 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, Liz Bourgeois, Ronnie Chatterji, Hemanth Bharatha Chakravarthy, Kevin Chen, Zoey Chen, Tess Cotter, Karina Dotson, Rob Friedlander, Zoë Hitzig, Nathan Lane, Jimmy Lin, Pamela Mishkin, Venkatesh Murthy, 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.

 This version of the paper was released February 2026, and is also listed as NBER working paper 34834.

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 di erent 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 work ow 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 speci cally. Indeed, in our prompt variation section, we nd that the prompt does not appear to matter much: GABRIEL's standard prompts do not yield di erent results from seemingly simplistic prompts. A major di erence 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 speci c 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.

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 e ciently, 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 di cult 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 o ering a set of di cult 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 rm'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-speci c 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 oor 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 ne-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 de nitions. 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 speci c 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 nish 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 de nes 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 would cut labeling costs by 17 ;500 relative to a human crowdsourcer paid $15 / hour when using the cheapest (but still high performing) model, and by 700 even when using the most premium model see Table 38 in Appendix A. This can turn a years-long project costing millions for a single labeling run into something doable in minutes on a shoestring budget.

Table 1: Cost estimates of rating each text on ten attributes (e.g. promoting family values , isolationist , focused on human su ering). 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 di ers greatly in how it asks the LLM to process the data.

Table 2 overviews the various functions supported by the toolkit, which range from classi cation to pairwise ranking to feature discovery to de-identi cation 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 " ne brushwork".
gabriel.classify	Classi es 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 identi ers.	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	Identi es similarities / di erences between paired items. Output = list of di erences.	Contrast op-eds from di erent 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 scienti c theories and lters the cream of the crop.	Procure novel theories on in ation 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 les into a spreadsheet for use in GABRIEL.	Image directory converted into spreadsheet of le 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 speci cs.
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 un nished 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 Paci c. 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 e ort 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 in uence 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 de ne 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 rst 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 de ne 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 quanti cations 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 (de ned so that a high score means they promote this idea), confrontational rhetoric , optimistic about technological progress , etc. Verbatim de nitions 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 de ned 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 (5000 words each) for every year from 1880 to 2022. Each transcript is unique to a speci c congressperson in the speci ed year. Even after sampling, this still provides a very substantial level of granularity into evolution over time.

Attribute	De nition
international interventionism	Supports active engagement in foreign a airs, 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-speci c 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 de nitions for Congressional remarks analysis. See Appendix A for all measured attributes.

As we can see with theinternational 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 oors 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 di erent 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 con ict 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 di er 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 con ict, 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 intensi cation of long-running trends toward a ective 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	De nition
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 nd 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 nal 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 rst 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 speci c 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	De nition
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 scienti c 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 de nitions.

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 et 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 (de nition 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 nd rugged individualism and America's natural beauty most celebrated in the rural West (attributes from Table 19 in Appen x 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 lmed 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 a ords a manyfold increase in usable data for social scientists; unstructured text analysis o ers 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 quanti es where di erent ideas are emphasized.

Figure 6: Rugged individualism in school curricula follows a di erent 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 exibility 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 awless. 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 speci c manifestations. First, we want to ascertain that the LLM's measurement is accurate, that is that it correctly re ects 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 di erent 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 de ned 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 di erent prompts or attribute de nitions yield di erent measurements? For GPT measurements to be reliable, we need the noise (randomness in exact wording) of the prompt to not a ect 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 de nition has a subtle but salient di erence, GPT should respect that di erence where it applies.

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

1. Contamination risk. Are performance estimates in ated 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 cuto . 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 rst 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 rst 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 di erent 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:

ˆ 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 rst 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 lms 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. De nitions are included in Appendix A in Table 26 for the 21 vDem variables. We attempt to match the original de nitions given to human labelers wherever possible. When these are not provided, we still measure the same attributes as the humans but we de ne the attributes as closely as we can match the available information.

On this initial set of tests, we nd that GABRIEL measurements are very similar to human labels. We nd 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 dropo in accuracy after the model training cuto . 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 re ecting 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 de nition.

Figure 9: Predicting Metacritic rating given review. GABRIEL dramatically outperforms older zero shot techniques. Identical performance before / after training cuto 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 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 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 a liated 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 classi cation 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 ling is a pleading or a judgment, or whether a given political speech isanti 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 $F_1 = \frac{2\,\text{precision}\,\text{recall}}{\text{precision}+\text{recall}}$ where 0 means total failure and 1 means perfect performance.

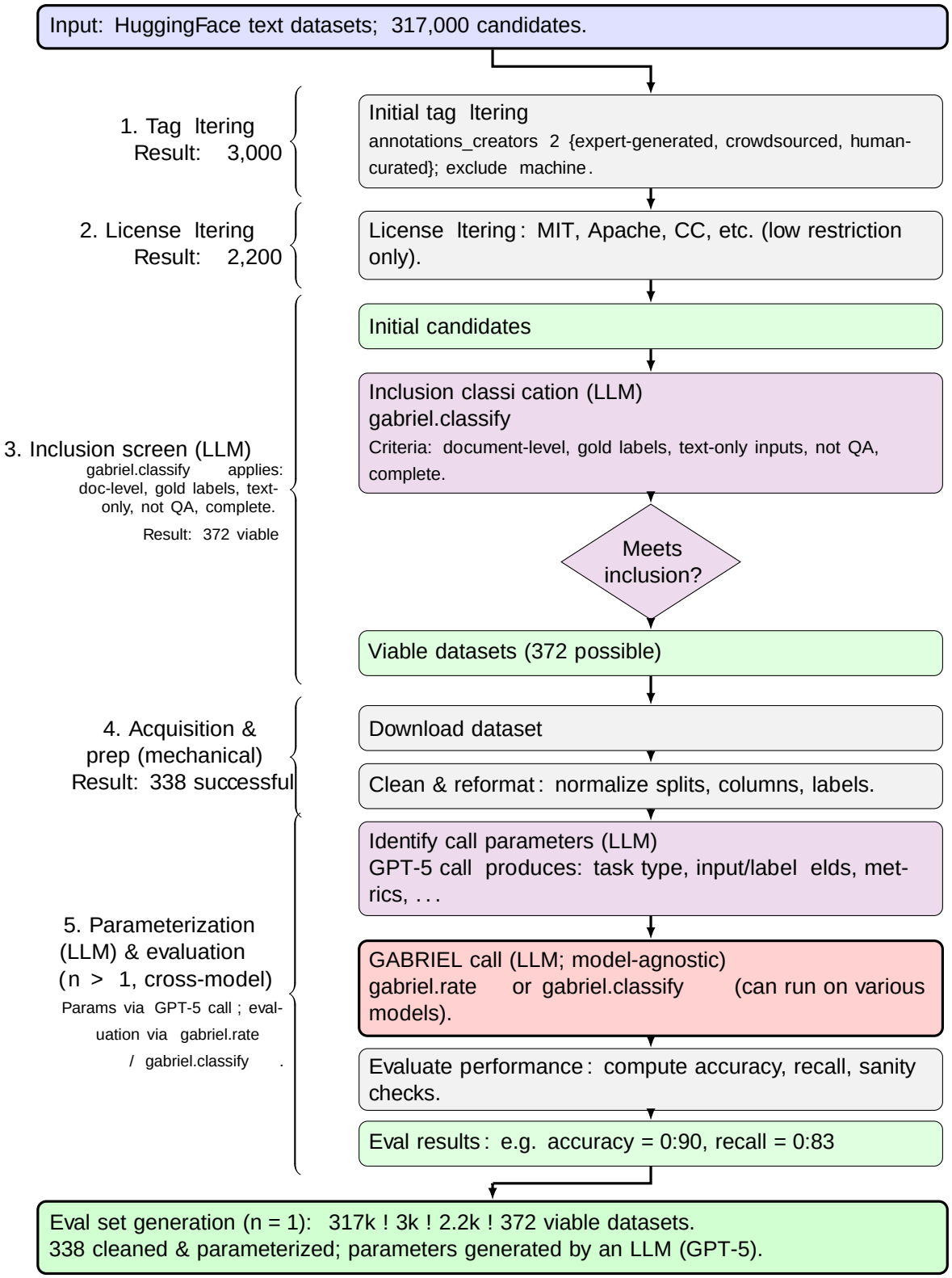

Figure 12: Pipeline schematic. Purple cells: LLM-based steps; gray: mechanical ltering; green: dataset states; red: main GABRIEL execution.

Figure 13: Classi cation accuracy (percentage agreement with humans) across over >1000 di erent classi cation tasks in >300 di erent 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 / classi cation task. This indicates general usability, unlike prior machine learning approaches.

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

score for these binary classi cation 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 (0:7 0:9+ for classi cation 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 re ecting real research and its use of human labeling.

This high accuracy is consistent across very di erent types of text and very di erent 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 exibility 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 speci c 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 cuto 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 de nition 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 nd 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 cuto , more on this in the next section), this co-movement is strong evidence that GABRIEL's variables capture real structure in the world.

Attribute	De nition
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 nancing cars, payday loans, or juggling multiple credit cards.

Table 7: Attribute de nition 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.