OPENESTIMATE: EVALUATING LLMS ON PROBABILISTIC ESTIMATION WITH REAL-WORLD DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Decisions in the real world rely on noisy, limited data. Language models (LMs), with broad pretrained knowledge, can help decision-makers by offering informed Bayesian priors that guide better choices. However, the extent to which LMs can provide reliable priors remains poorly understood. We introduce OPENESTIMATE, a benchmark that asks LMs to express beliefs as Bayesian priors over real-world quantities from labor economics, private markets, and public health. We assess these priors for both accuracy and calibration, benchmarking them against statistical baselines built by sampling from the true distribution. Across six frontier LMs, LM-elicited priors are often inaccurate and overconfident: they seldom beat posteriors formed from five real observations. Performance improves modestly depending on how uncertainty is elicited from the model, but is largely unaffected by changes in temperature, reasoning effort, or system prompt. Given LMs' weak performance, OPENESTIMATE offers an important foundation for building systems that can reason under uncertainty and know when to doubt themselves.

1 INTRODUCTION

What is the average total funding raised by venture-backed companies outside the United States? Even finance experts struggle to answer this without first stitching together heterogeneous sources of evidence. Many deals go undisclosed, reporting is noisy, and sample sizes are often small. As a result, analysts must reason under uncertainty, blending background knowledge, intuition, and limited data to form probability distributions over plausible values rather than precise point estimates. Similar probabilistic estimation problems arise in domains such as public health and labor economics, where consequential decisions—how to allocate capital or which policies to adopt—hinge on quantities that are inherently uncertain.

In such settings, decisions are guided by Bayesian priors: probability distributions that capture initial beliefs about unknown quantities. Well-calibrated priors make downstream inference more reliable, while poorly calibrated ones can distort conclusions even when more data arrive. Language models (LMs), pretrained on vast corpora, are natural candidates for supplying such priors. Their broad background knowledge across domains could, in principle, be distilled into structured probability distributions that humans or statistical models then update with real-world data.

Assessing this potential requires benchmarks that directly test LMs' ability to generate well-calibrated distributions over uncertain quantities in realistic settings. Existing benchmarks rarely probe this skill: some focus on deterministic problem-solving with single correct answers (Hendrycks et al., 2021), others on forecasting questions whose outcomes eventually leak into training data, and another set on structured "guesstimation" methods such as Xia et al. (2024) that emphasize intermediate modeling rather than direct distributional estimation.

To fill this gap, we introduce OPENESTIMATE, a benchmark designed to evaluate LMs on complex probabilistic estimation tasks. Each task specifies a real-world variable derived from public health, finance, or labor economics datasets, such as *average total funding raised by companies outside the US* or the *average weight of US adults with diabetes*. Models are asked to parameterize explicit probability distributions (Gaussians for continuous quantities and Betas for proportions) over these variables. Performance is evaluated in terms of both (i) accuracy—whether predicted distributions concentrate near the ground truth—and (ii) calibration—whether stated confidence levels align with observed frequencies.

Domain	Dataset	Variable	+1	+2	+3	Total	Example Marginal Variable
Labor Economics	Glassdoor	1	16	20	6	43	Midpoint salary
Finance	Pitchbook	4	17	20	20	61	Total funding
Human Health	NHANES	14	20	20	20	74	Total cholesterol

Table 1: Distribution of benchmark variables across domains. Columns indicate the number of marginal variables and conditional variables with one, two, or three conditioning attributes. Each additional condition increases contextual specificity.

Using OPENESTIMATE, we evaluate quality of Bayesian priors elicited from frontier LMs. We find that these models are far from omniscient: in terms of accuracy and calibration, they often perform no better—or even worse—than just five random draws from an empirical distribution. Further, no model stands out as particularly accurate or well-calibrated across domains. However, models still demonstrate an ability to assign systematically higher likelihoods to true outcomes. This indicates that while models may struggle with absolute accuracy, their probabilistic estimates nonetheless capture non-trivial domain structure. This latent signal suggests that, while current priors are weak, they hold potential as a foundation for methods that refine or adapt LMs into trustworthy tools for probabilistic estimation. To support future research and reproducibility, we release our code, benchmark dataset, and evaluation framework.

2 THE OPENESTIMATE BENCHMARK

In this section, we describe how we built the OPENESTIMATE benchmark. We begin by defining estimation targets as variables derived from large-scale datasets in labor economics, finance, and public health (Section 2.1). We then explain how models are asked to specify their priors as Gaussian or Beta distributions parameterized from natural language prompts (Section 2.2). Finally, we outline the evaluation metrics used to assess the accuracy and calibration of these priors (Section 2.3).

2.1 DEFINING ESTIMATION TARGETS

To evaluate LM probabilistic estimation skills, we must define variables (like *average funding amount for companies outside of the US*) that are unlikely to appear in LMs' pretraining data yet estimable with background knowledge. Crucially, we need access to the ground-truth values of these variables in order to measure performance. Because much of human knowledge is already contained in pretraining corpora, creating variables that meet these criteria typically requires collecting new data experimentally, which is often costly and time-consuming.

To address this challenge, we design a variable generation procedure to create *derived variables*: quantities that can be computed directly from large-scale observational datasets where ground truth is available but that do not correspond to well-documented facts likely to appear in pretraining corpora. We sample these variables from Glassdoor [1], Pitchbook (PitchBook Data, 2024), and NHANES (for Disease Control & Prevention, 2018) datasets, which cover topics spanning across labor economics, private markets, and human health.

The variables we sample from these datasets come in two forms. Some are marginal statistics, aggregated across an entire dataset (for example, *the mean salary of data scientists*, *the median deal size of venture-backed companies*, or *the mean weight of US adults*). Others are conditional statistics, restricted to subgroups defined by up to three auxiliary attributes (for instance, *the mean salary of data scientists in Virginia*, *the median deal size of venture-backed companies in the technology sector*, or *the mean weight for adults with a diabetes diagnosis and high cholesterol*). The full breakdown of variable types across domains is shown in Table 1.

We generate conditional statistics by sampling auxiliary attributes at random from empirically observed values in the data. To avoid trivial or redundant subgroups, we follow Xia et al. (2024) in requiring that each additional conditioning attribute alters the target statistic by at least 5%. This constraint ensures that derived quantities reflect meaningful variation across subgroups rather than minor fluctuations due to sampling noise. The full sampling algorithm is described in Algorithm 1.

[1] https://www.kaggle.com/datasets/thedevastator/jobs-dataset-from-glassdoor

Algorithm 1: Sampling N_k *marginal* ($k = 0$) and *conditional* ($k = 1, 2, 3$) variables

Input: data D, auxiliary attributes $\mathcal{A}$, counts $\{N_k\}_{k=0}^{3}$, threshold τ, n minimum sample size
Output: set $\mathcal{V}$ of variables

```
V ← ∅, S ← ∅        // S tracks which attributes have already been used
for k ∈ {0, 1, 2, 3} do
    while number of variables in V with k attributes < N_k do
        sample k distinct attributes a_k ⊂ A          // a_k is a set of k attributes
        D' ← filter D by a_k                  // keep rows matching attributes in a_k
        if |D'| < n then
            continue                        // skip if filtered sample is too small
        μ̂ ← mean(Y | D')                                  // estimate mean on D'
        ŝe ← SE(μ̂; D')                          // estimate standard error on D'
        μ_0 ← mean(Y | D)                        // unconditional mean on full D
        if |μ̂ − μ_0| > τ and |μ̂ − μ_0| > ŝe and a_k ∉ S then
            add (a_k, μ̂, ŝe) to V                              // store valid variables
            add a_k to S                          // store attributes to avoid reuse
return V
```

While some variables of this kind may overlap with information already present in pretraining corpora (e.g., widely reported statistics such as overall diabetes prevalence in the United States), many others are far less likely to have been explicitly documented. In particular, conditional variants of these quantities—such as *the mean weight of adults with diabetes who also have elevated cholesterol*, or *the median deal size for companies in a specific sector with a given number of employees*—represent fine-grained combinations of attributes that are almost never reported in textual sources. By systematically varying the conditioning attributes, we generate a large set of estimation targets that remain grounded in real-world observational data yet are unlikely to be memorized facts. This design allows us to evaluate whether models can combine background knowledge with probabilistic reasoning, rather than relying on surface-level recall.

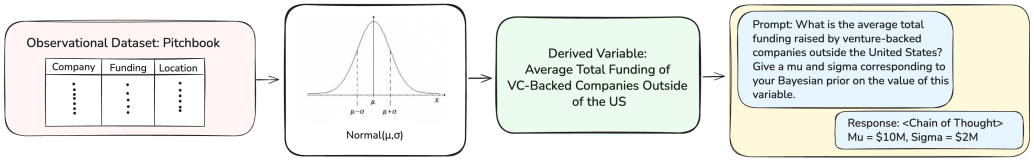


Figure 1: Variable generation and prior elicitation pipeline. We construct derived variables from large-scale observational datasets (e.g., PitchBook), specify them as statistical targets (e.g., Gaussian means), and prompt language models to provide Bayesian priors in the form of distributional parameters.

2.2 Specifying Estimates as Bayesian Priors

How do we ask LLMs to give their estimates about the likely values of these variables? One simple approach would be to evaluate models on the accuracy of their point estimates by reporting the distance (e.g. squared error) between these estimates and the ground-truth value in the data.

However, evaluation of point estimates leaves out much of what is necessary for such predictions to be useful in the real world: with simple point estimates, it's not possible to distinguish predictions that are right by chance from those that are right as a result of an accurate reasoning procedure; or conversely between predictions that are wrong but confident and predictions that are wrong but highly uncertain. Thus, rather than measuring predictions in the form of point estimates, OpenEstimate requires predictions to be specified as probability distributions or Bayesian priors on the variable of interest.

In this paper, predictions are specified via the parameters of a Gaussian or Beta distribution:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad \text{or} \quad X \sim \text{Beta}(\alpha, \beta),$$

depending on whether the target variable is continuous or a proportion.

These two forms are chosen because they arise frequently in our domains of interest– Gaussians for continuous, symmetric quantities like wages, and Betas for proportions like disease rates. In all cases, we provide the model with a brief natural language description of the variable that we want a prior distribution on, and nothing else. We tell the model which distributional form we want its prior to take on and ask it to parameterize the prior accordingly.

2.3 Evaluation Metrics

Given a prediction from the LM in the form of a Gaussian or Beta distribution, how should we evaluate its quality? We focus on two complementary dimensions of performance:

- **Accuracy**: The degree to which the model assigns high probability density (or mass) to regions close to the empirical ground-truth value.
- **Calibration**: The consistency between the model's stated uncertainty and empirical frequencies. A model is well-calibrated if events assigned probability p occur with long-run frequency p, such that nominal coverage levels of prediction intervals match their realized coverage.

2.3.1 Accuracy

To measure accuracy, we look at: (i) central tendency, assessed by the mean absolute error (MAE) of the distribution's mode relative to the empirical ground-truth statistic, and (ii) distributional fit, assessed by the scale-adjusted log-probability that the model assigns to the ground truth.

Our first measure assesses central tendency: does the model place the mode of its distribution close to the ground-truth statistic? To quantify this, we compute the mean absolute error (MAE) between the mode of the predicted distribution, $\hat{x}_i$, and the empirical ground-truth value x^* estimated from the full dataset:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{x}_i - x_i^*|$$

To interpret these errors across variables with different units, we benchmark LM predictions against statistical baselines derived from small empirical samples. Starting from naïve flat priors ($\alpha = 1, \beta = 1$ for Beta distributions; $\mu = 0, \sigma^2 = 100{,}000$ for Gaussians), we update with N random draws from the relevant subgroup to obtain a posterior distribution.

In addition, we construct an LM-informed baseline by performing a conjugate Bayesian update of the LM-elicited prior with the same N samples. This baseline tests not only the quality of the LM's prior in isolation but also its usefulness when combined with real data to assess the practical role of LM-elicited priors in Bayesian inference.

We summarize performance using the error ratio, defined as the LM's MAE relative to a statistical baseline:

$$\text{Error Ratio} = \frac{\text{MAE}_{\text{LLM}}}{\text{MAE}_{\text{Statistical Baseline}}}$$

An error ratio below one indicates that the LM's prediction is more accurate than simply drawing five samples from the ground truth distribution.

While this measure captures whether the distribution is centered appropriately, it ignores how much probability mass is actually assigned to the ground-truth value. We therefore also evaluate distributional fit using scale-adjusted log-probabilities. Specifically, we compute the log-probability of the ground-truth value under the model's distribution, with a correction term from the Jeffreys prior to ensure invariance to scale and parameterization ($\frac{1}{2}\log(p(1-p))$ for Betas; $\log(\sigma)$ for Gaussians). This measure reflects whether the model regards the true value as "likely" under its stated uncertainty.

Together, these two dimensions provide a fuller picture of accuracy: the error ratio tests whether models outperform simple empirical sampling in terms of central tendency, while the log-probability assesses whether the ground truth lies in a region of high probability mass under the model's stated beliefs.

2.3.2 CALIBRATION

A model is well-calibrated if the probabilities it assigns correspond to empirical frequencies: events predicted to occur with probability p should occur about p of the time. In our setting, this means that the ground-truth value should fall into each predicted quantile with the correct long-run frequency.

To measure this, we partition each model's predictive distribution into quartiles and record how often the ground-truth values fall into each bin. For a perfectly calibrated model, each quartile should contain the ground truth 25% of the time. Deviations from this ideal reflect miscalibration. Formally, we compute the expected calibration error as:

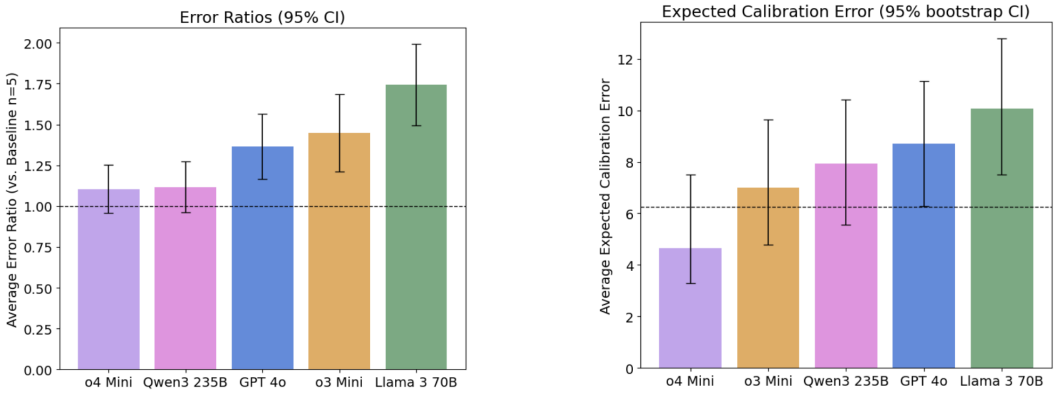
$$\text{ECE} = \frac{1}{4}\sum_{i=1}^{4} |p_i - 0.25|$$

where p_i is the empirical proportion of ground-truth values that fall into the i-th quartile of the predicted distribution across all evaluation instances. Lower values indicate better calibration, with ECE $= 0$ corresponding to perfect calibration.

3 EVALUATION

In this section, we focus on zero-shot performance under standard inference settings. We do not apply fine-tuning, retrieval augmentation, or prompt engineering beyond directly asking the model to parameterize the distribution of a variable. To contextualize the LMs' performance, we compare to the four statistical baselines described in Section 2.3.1.

We evaluate six state-of-the-art language models, including three reasoning models: Meta Llama 3.1 8B, Meta Llama 3.1 70B (Grattafiori et al., 2024), OpenAI GPT-4 (Achiam et al., 2023), OpenAI o3-mini (OpenAI, 2025a), OpenAI o4-mini (OpenAI, 2025b), and Qwen3-235B-A22B (Yang et al., 2025). We exclude Llama 3.1 8B after it fails to follow basic answer specification. We evaluate each model at a medium temperature or reasoning effort – corresponding to 0.5 for GPT-4, "medium" for o3-mini and o4-mini, 0.5 for Llama 3.1 70B Instruct Turbo, and 0.6 for Qwen3-235B-A22B. We use a standard system prompt and prior elicitation prompt which are described in full in the Appendix A.1.



(a) Accuracy

(b) Calibration

Figure 2: Accuracy and calibration of LMs compared to the statistical baseline of $N = 5$ samples (dotted line). (a) Accuracy is reported as the error ratio of model predictions relative to the baseline. (b) Calibration is reported as expected calibration error (ECE). Across domains, models are at best comparable to the five-sample baseline and often worse.

Accuracy. We first evaluate the accuracy of the central tendency of each model's estimates by reporting its mean absolute error as a multiple of the mean absolute error of a statistical baseline computed using an uninformative prior, as described in 2.3.1. We find that models hardly perform better than the $N = 5$ statistical baseline, as shown in Table 2. Therefore, we use the $N = 5$ baseline as a common point of comparison across results.

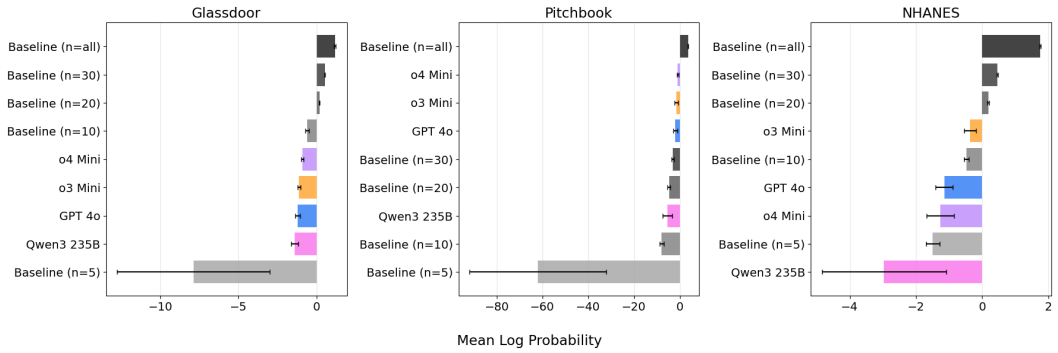


Figure 3: Log-probabilities of the ground truth under model priors across domains. Reasoning models (o3 Mini, o4 Mini) consistently allocate higher likelihoods to the true outcome than weak baselines (e.g., $N = 5$), in some cases rivaling $N = 20$–30 sample posteriors.

We find that reasoning models tend to perform as well as the statistical baseline, whereas non-reasoning models tend to perform worse than the statistical baseline (Figure 2a). However, even the top-performing models did not surpass the accuracy of an estimate derived from five real data points, suggesting that OPENESTIMATE remains challenging for frontier models.

Domain	N	% Help	Prior	Post	Base	Δ
NHANES	5	73.0%	1.319	0.928	1.000	7.2%
	10	48.6%	1.319	0.840	0.712	−17.9%
	20	41.9%	1.319	0.739	0.511	−44.4%
	30	33.8%	1.319	0.700	0.424	−65.0%
PitchBook	5	50.8%	1.768	1.208	1.000	−20.8%
	10	42.6%	1.768	1.075	0.743	−44.7%
	20	32.8%	1.768	0.944	0.550	−71.7%
	30	27.9%	1.768	0.873	0.433	−101.7%
Glassdoor	5	62.8%	1.800	0.980	1.000	2.0%
	10	60.5%	1.800	0.724	0.688	−5.2%
	20	51.2%	1.800	0.478	0.432	−10.5%
	30	27.9%	1.800	0.347	0.291	−19.2%

Table 2: Error ratios relative to a five-sample statistical baseline (N=5). "Prior" refers to the LM prior alone; "Post" to the conjugate update to the LLM prior with N samples; "Base" to the posterior from flat prior with N samples. Δ is the % improvement between Post and Base.

Although the models are far from omniscient, they nonetheless exhibit useful structure in how they allocate probability mass. Examining the log probabilities of the true outcome reveals that their elicited priors capture non-trivial knowledge by concentrating probability near the correct answer (Figure 3). For example, in Pitchbook, o4 Mini and o3 Mini are able to assign higher likelihoods to the true outcome to outperform statistical baselines up to $N = 30$ samples.

Calibration. Next, we assess model calibration.[2] We first assess calibration by computing the expected calibration error (as defined in Section 2.3.2). Models are only as calibrated as five random samples from the empricial distribution (Figure 2b) and display systematic overestimation (Figure

[2]We exclude the statistical baselines from Figure 4 in this analysis because the baselines derive their posteriors from the same dataset used to compute the ground-truth values, larger sample sizes produce extremely tight distributions centered on the ground-truth mean. This leads the ground truth to almost always fall in the middle quantiles (e.g., second or third).

4). In Pitchbook, overestimation is compounded by high rates of underestimation as well, with both tails overweighted. No single model dominates across domains.

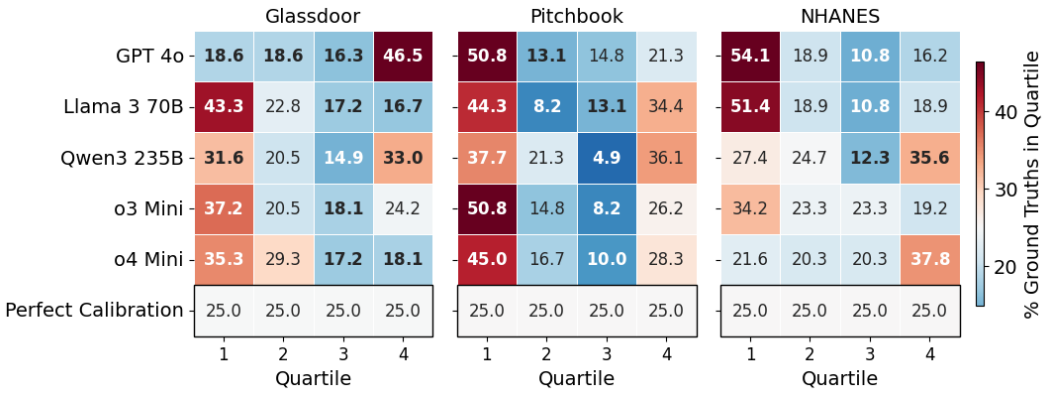


Figure 4: Heatmap describing the deviations from perfect calibration of each approach. Bolded values are statistically significant according to a per-quartile binomial test ($p < 0.05$). All approaches systematically overestimated across domains (Quartile 1 is greater than 25%). In some instances, there was high rates of both over and under-estimation (Quartile 1 and 4 are greater than 25%).

Next, we examine the cumulative distribution of ground-truth values relative to the predicted priors (Figure 5) to understand how tightly models concentrate their uncertainty. We find the best models cover 80% of the ground truth values within two to three standard deviations of the mean. However, performance is domain-dependent: in Glassdoor and NHANES, the best models cover over 80% of ground-truth values within two standard deviations, while in Pitchbook, three standard deviations are required. This suggests that even the strongest models vary substantially in how they express uncertainty across domains.

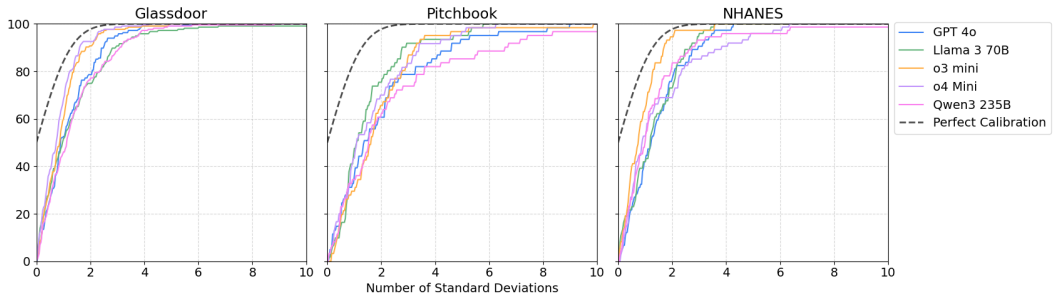


Figure 5: Cumulative distribution function displaying the percentage of ground truth values that fall within $n\sigma$ standard deviations away from the mean of the prior, where σ is the standard deviation of the prior. The dashed line represents perfect calibration for a Gaussian. The best performing models have 80% of the ground truths within 1.5-2.5 standard deviations from the prior mean.

Finally, we analyze whether model-reported uncertainty is a reliable guide to predictive accuracy (Figure 6) by comparing the standard deviation ratio to the error ratio. Ideally, models are low error and well-calibrated. In the Glassdoor domain, models appear reasonably well-calibrated but consistently less accurate than the baseline. In contrast, models in Pitchbook cluster closer to the ideal point with low error and good calibration, although o3 Mini stands out as markedly overconfident. Results in NHANES fall in between these extremes: models generally achieve lower error than in Glassdoor, but their uncertainty estimates are less well-calibrated, with several models exhibiting either under- or over-dispersion. Taken together, these results indicate that the relationship between uncertainty and accuracy is once again strongly domain-dependent.

We also assess whether predictive uncertainty aligns with accuracy by examining the rank correlation between the two. A stronger correlation between predictive uncertainty and accuracy would indicate that uncertainty is a good predictor of accuracy. However, reality is mixed: uncertainty is a good predictor of accuracy in NHANES but not necessarily in Pitchbook or Glassdoor.

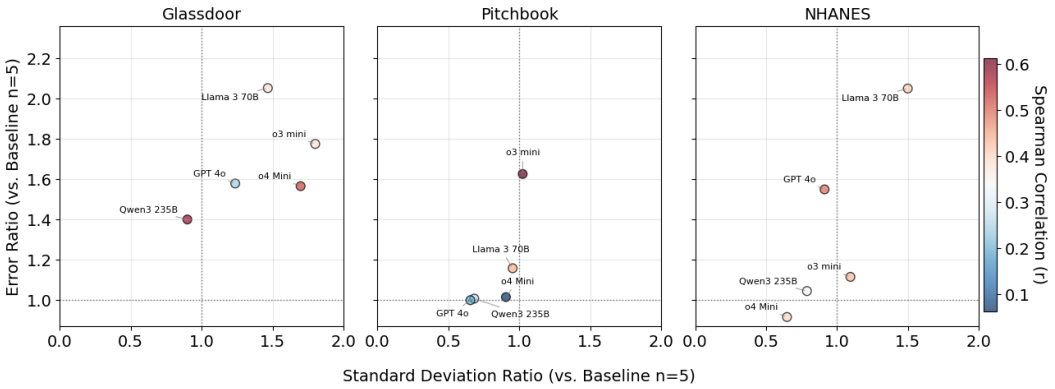


Figure 6: Relationship between uncertainty and accuracy across domains. Each point shows a model's error ratio versus its standard deviation ratio relative to the $N = 5$ baseline. Ratios closer to 1 indicate better alignment: low error and well-calibrated uncertainty. Colors indicate the Spearman correlation between predictive uncertainty and accuracy.

3.1 Ablations

We investigate how inference-time settings influence the quality of elicited priors, focusing on three factors: (i) temperature or reasoning effort, (ii) system prompt, and (iii) elicitation protocol. To isolate their effects, we evaluate both a reasoning model (OpenAI o4-mini) and a non-reasoning model (OpenAI gpt-4o). The full set of results is shown in Appendix A.2.

Across models and domains, elicitation protocol is by far the most consequential factor, while temperature and system prompt have negligible effects. Because prior specification is central to our task, we tested three distinct elicitation strategies. *Direct elicitation* asks models to provide distribution parameters without additional structure. *Quantile elicitation* requests specific percentiles, encouraging models to reason explicitly about uncertainty ranges. *Mean–variance elicitation* separates point estimates from dispersion, prompting reflection on confidence levels.

As shown in Figure 7, direct elicitation consistently yields the best performance for reasoning models, whereas quantile elicitation is superior for non-reasoning models. We hypothesize that reasoning models are able to map parameter-level prompts onto coherent distributions, while non-reasoning models benefit from the scaffolding imposed by percentile queries.

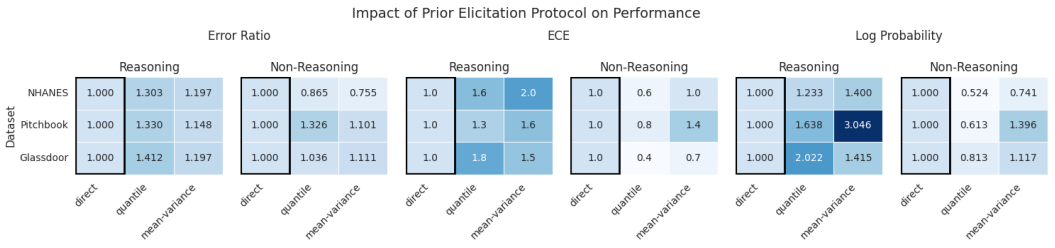


Figure 7: Effect of elicitation protocol (direct, quantile, mean–variance) on error ratio, expected calibration error (ECE), and log probability across reasoning and non-reasoning models. Direct elicitation is most effective for reasoning models, while quantile elicitation benefits non-reasoning models.

4 Related Work

Our work intersects with three major lines of language model research: evaluating probabilistic reasoning as a mathematical skill, structuring probabilistic reasoning for better estimation, and applications to forecasting.

Evaluating probabilistic reasoning. One line of research examines how well LMs perform at problem-solving tasks involving structured probabilistic models. For example, Paruchuri et al. (2024) evaluate models' probabilistic reasoning given simple idealized distributions; Nafar et al. (2025) tests models' ability to provide probabilistic estimates given a Bayesian network; and Jin et al. (2023) examine the models' causal reasoning given probabilities. Collectively, these studies frame probabilistic reasoning as a mathematical exercise with clearly defined inputs and well-specified outputs. By contrast, our benchmark targets real-world estimation problems, where the relevant information must be inferred rather than provided and the ground truth itself may be ambiguous or unavailable.

Structuring probabilistic reasoning. Another line of work proposes structures for LM-based probabilistic reasoning to improve performance. Using "guesstimation" questions similar to ours, Xia et al. (2024) prompt LMs to propose relevant random variables and moment constraints, and then fits a log-linear distribution that satisfies these constraints.Feng et al. (2024) take a similar approach, and evaluate a multi-step process in which LMs brainstorm relevant factors, make coarse probabilistic assessments, and construct an approximate Bayesian network for inference.

These approaches extend beyond single-variable reasoning by introducing latent structure and explicit intermediate steps. However, the focus for both of these works is on answering discrete multiple-choice questions, such as those where the LM must select the most likely explanation or outcome. Our benchmark, by contrast, emphasizes continuous and potentially open-ended variables: models must explicitly place probability distributions over possible outcomes. While our evaluation does not impose an explicit reasoning structure on the LM, future work could explore how structured approaches of this kind might be adapted to improve performance in our setting.

Language model-based forecasting. Recent studies have also evaluated LMs' forecasting capabilities (Karger et al., 2024; Halawi et al., 2024; Ye et al., 2024; Chang et al., 2024; Schoenegger et al., 2025). These works also test whether models can synthesize heterogeneous evidence into well-calibrated estimates, but they focus on making predictions about real-world future events. In contrast to our benchmark, the outcomes of forecasting questions are, by design, highly likely to appear in LMs' training data after they resolve; they thus perpetually become "stale" and must be replaced with new questions, as noted by Karger et al. (2024). By focusing on questions that require reasoning about fine-grained cross of tabular datasets, rather than future events, OPENESTIMATE questions are designed to remain challenging over time.

5 LIMITATIONS AND FUTURE WORK

Our current tasks use primarily Gaussian/Beta parametric forms and a fixed set of domains; expanding to heavy-tailed, multimodal, and discrete/ordinal targets, as well as structured priors (mixtures, hierarchies), is an important next step. Ground truths are estimated from finite samples, and while cross-sectional by design, residual leakage cannot be ruled out. We evaluate zero-shot models without retrieval or fine-tuning; studying training-time interventions for uncertainty awareness and domain adaptation is complementary. Methodologically, future evaluations could incorporate proper scoring rules (e.g., log score, CRPS), Wasserstein distances, and decision-centric metrics that measure the value of LM priors for downstream choices. Finally, interactive pipelines—retrieval to surface relevant evidence, multi-pass critique to detect unit/base-rate errors, and posterior-over-prior ensembling—may turn brittle priors into robust, uncertainty-aware estimates.

6 CONCLUSION

We introduced OPENESTIMATE, a benchmark and evaluation framework for assessing language models on open-ended probabilistic estimation with real-world tabular data. The benchmark (i) defines a task where models must express beliefs as full probability distributions, (ii) elicits priors through several protocols, and (iii) evaluates performance along accuracy, calibration, and uncertainty—against statistical baselines that use only a handful of true samples. By focusing on cross-sectional quantities from domains such as public health (NHANES), labor economics (Glassdoor), and private markets (PitchBook), OPENESTIMATE probes reasoning under uncertainty while limiting direct lookup and information leakage.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

He Chang, Chenchen Ye, Zhulin Tao, Jie Wu, Zhengmao Yang, Yunshan Ma, Xianglin Huang, and Tat-Seng Chua. A comprehensive evaluation of large language models on temporal event forecasting. *arXiv preprint arXiv:2407.11638*, 2024.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. Bird: A trustworthy bayesian inference framework for large language models. *arXiv preprint arXiv:2404.12494*, 2024.

Centers for Disease Control and Prevention. National health and nutrition examination survey (nhanes), 2017–2018. U.S. Government, National Center for Health Statistics, 2018. URL `https://www.cdc.gov/nchs/nhanes`. Data set, accessed via `https://www.cdc.gov/nchs/nhanes`; includes questionnaires, datasets, and documentation.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. *Advances in Neural Information Processing Systems*, 37: 50426–50468, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL `https://arxiv.org/abs/2009.03300`.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: Assessing causal reasoning in language models. *Advances in Neural Information Processing Systems*, 36: 31038–31065, 2023.

Ezra Karger, Houtan Bastani, Chen Yueh-Han, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip E Tetlock. Forecastbench: A dynamic benchmark of ai forecasting capabilities. *arXiv preprint arXiv:2409.19839*, 2024.

Aliakbar Nafar, Kristen Brent Venable, Zijun Cui, and Parisa Kordjamshidi. Extracting probabilistic knowledge from large language models for bayesian network parameterization. *arXiv preprint arXiv:2505.15918*, 2025.

OpenAI. Openai o3-mini system card. System card, OpenAI, January 2025a. Published January 31, 2025.

OpenAI. Openai o3 and o4-mini system card. System card, OpenAI, April 2025b. Published April 16, 2025.

Akshay Paruchuri, Jake Garrison, Shun Liao, John Hernandez, Jacob Sunshine, Tim Althoff, Xin Liu, and Daniel McDuff. What are the odds? language models are capable of probabilistic reasoning. *arXiv preprint arXiv:2406.12830*, 2024.

PitchBook Data. Pitchbook database. Accessed via Wharton Research Data Services (WRDS), 2024. URL `https://wrds-www.wharton.upenn.edu/`. Accessed December 29, 2024.

Philipp Schoenegger, Peter S Park, Ezra Karger, Sean Trott, and Philip E Tetlock. Ai-augmented predictions: Llm assistants improve human forecasting accuracy. *ACM Transactions on Interactive Intelligent Systems*, 15(1):1–25, 2025.

Shepard Xia, Brian Lu, and Jason Eisner. Let's think var-by-var: Large language models enable ad hoc probabilistic reasoning. *arXiv preprint arXiv:2412.02081*, 2024.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. Mirai: Evaluating llm agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024.

A Appendix

A.1 Zero-Shot Estimation

We tested Llama 3 8B but excluded it from our analysis because it incorrectly followed instructions pertaining to units and had an average error that was orders of magnitude larger than the other models due to this mistake.

System Prompt.

For Glassdoor, the LM's system prompt was:

```
You are a helpful assistant that can answer questions about the labor market.
```

For Pitchbook, the LM's system prompt was:

```
You are a helpful assistant.
```

For NHANES, the LM's system prompt was:

```
You are a helpful assistant that can answer questions about human health.
```

Statistical Baselines.

In the NHANES analysis, the $N =$ all baseline has non-zero MAE because limited effective sample sizes for some variables prevented the posterior from fully converging to the ground truth estimate, leaving it partially shrunk toward the flat prior.

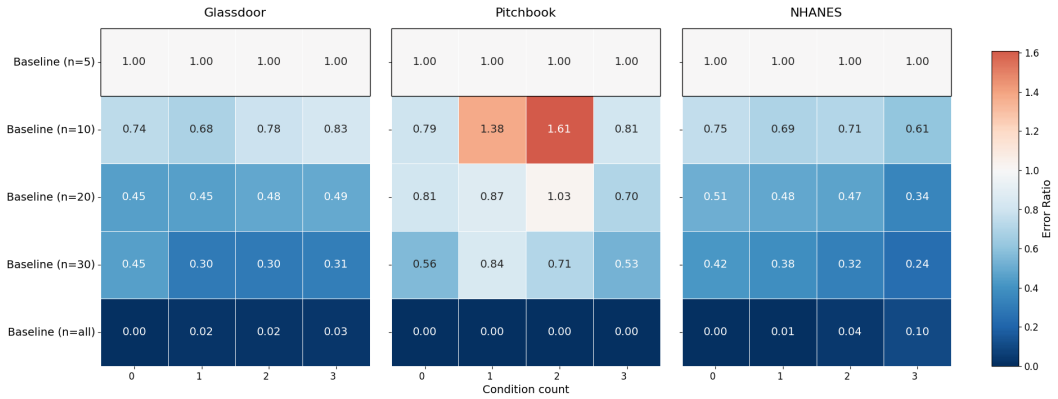


Figure 8: Error ratios for all statistical baselines as the number of auxiliary attributes increases.

As shown in Figure 9, we find that error ratios do not always increase with the number of auxiliary attributes: LLMs can be less accurate on simple aggregates (e.g., average cholesterol across all adults) than on more constrained subgroups, but in other cases, accuracy degrades with added specificity (e.g., salaries of data scientists generally vs. within Virginia-based companies of a certain revenue).

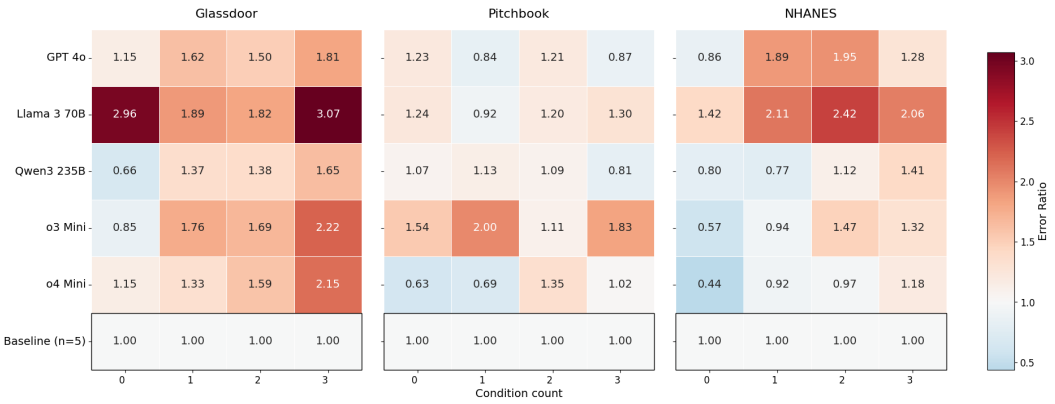


Figure 9: Error ratios across domains as the number of auxiliary attributes increases.

A.2 Ablations

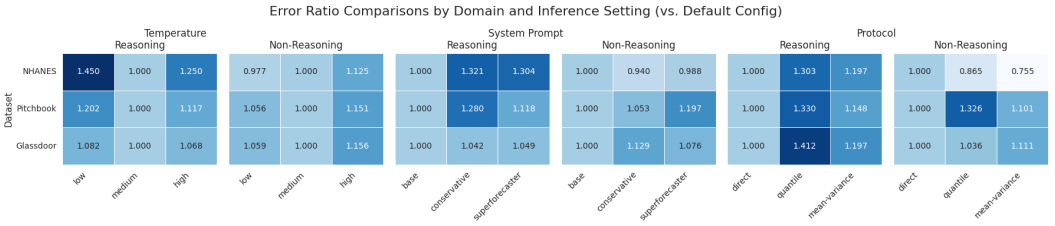


Figure 10: We examine the impact of changing temperature, system prompt, and elicitation protocol on error ratio.

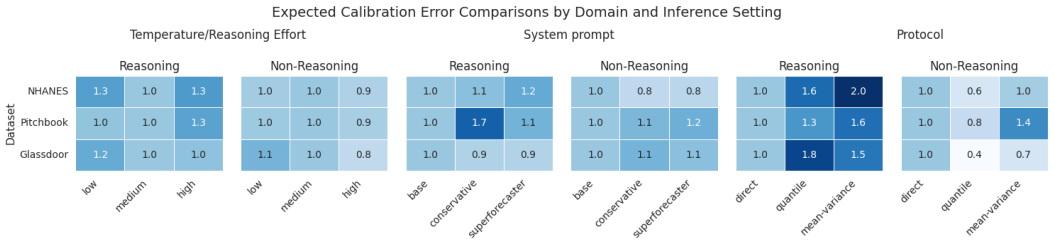


Figure 11: We examine the impact of changing temperature, system prompt, and elicitation protocol on expected calibration error.

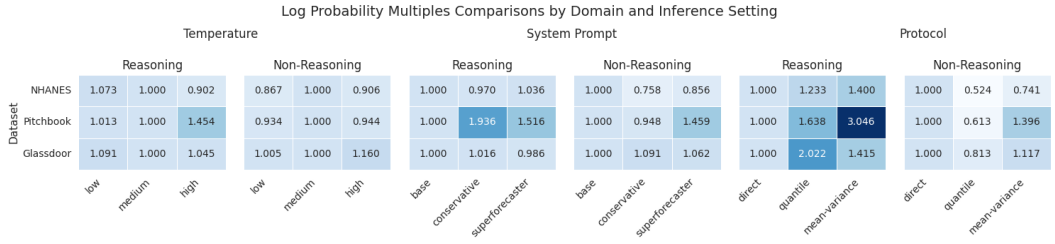


Figure 12: We examine the impact of changing temperature, system prompt, and elicitation protocol on log probabilities.