PREDICTING REAL-WORLD QUANTITIES WITH LLMS THROUGH LEARNED AGGREGATION OF SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) are increasingly used to estimate expectations of real-world quantities, yet their estimates are often inaccurate. Prior methods address this through case analysis, asking the LLM to estimate various more specific quantities (conditioned on additional facts) and aggregating the results. However, these more specific estimates may still be inaccurate, and it is not always clear how to aggregate them appropriately. We thus introduce LAoS, Learned Aggregation of Scenarios, which *learns* to predict the target quantity from a collection of specific LLM estimates. LAoS is an attention-based feedforward network that considers the various textual LLM queries (via embeddings) and the LLM's responses to them. On held-out conditional probability queries, LAoS outperforms prevalence-weighted averaging and repeated direct elicitation, also demonstrating generalizability across domains. Controlled stress tests with overlapping and unevenly distributed scenarios further demonstrate that LAoS learns to correct for these issues, as well as correcting for LLM estimation errors.

1 INTRODUCTION

Large language models (LLMs) are increasingly being used to estimate real-world quantities, such as event probabilities (Cruz et al., 2024; Nafar et al., 2026). However, LLMs' estimates can be inaccurate, miscalibrated, and inconsistent with other estimates from the same LLM (Cruz et al., 2024; Zhu & Griffiths, 2024). Estimates are also sensitive to the phrasing of the question (Paleka et al., 2025). LLMs surely "know" a great deal about the world, but extracting accurate estimates from them may require more than a single direct question.

We focus here on **analysis by scenario**: to estimate the average price of laptops in a country, we could separately consider its education, home, corporate, and gaming markets. This forces us to consider the relative sizes of these markets and their typical laptop specifications and prices. In general, the idea is to estimate the target quantity in various **scenarios** that provide additional information, and aggregate these estimates. **Case analysis** refers to the simple version where the scenarios ("cases") form an exhaustive partition of the possible outcomes, as in the laptop example. BIRD (Feng et al., 2025) has explored this method. Alternatively, one might try to *sample* scenarios that refine the original question (Ozturkler et al., 2023; Hou et al., 2024) (Ozturkler et al., 2023). In this version, the sampled scenarios are intended to be plausible but do not systematically cover all possibilities. Both approaches generally allow **inference-time scaling** (Snell et al., 2025), which refers to an LLM workflow that can improve its answers by running for longer. Given a query budget n, such a scalable workflow queries the LLM n times, either sequentially or in parallel, before computing a final answer.

One potential advantage of these approaches is that the LLM may be more accurate when queried with more concrete details (Prystawski et al., 2023; Feng et al., 2025). Another potential advantage is the purported "wisdom of crowds": the errors on different scenarios might tend to cancel out in aggregation (Reia & Fontanari, 2021). Humans may use some form of scenario analysis to estimate probabilities in unfamiliar situations (Johnson-Laird et al., 1999). LLMs have been observed to do the same within their chains of thought, such as when predicting the expectation of consecutive dice rolls (Didolkar et al., 2025).

As we outline in §A, it is challenging to design a principled approach to aggregating over sampled scenarios. Thus, we ask whether an effective aggregation function could instead be *learned* with

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

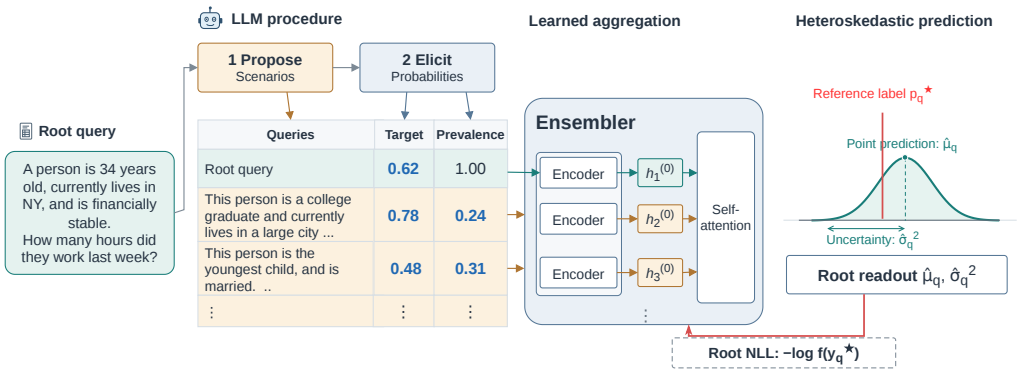


Figure 1: LAoS workflow. The LLM proposes scenarios and estimates their prevalences and conditional expectations of the target quantity, together with a direct root estimate. The ensembler combines text and numerical inputs to predict a distribution over the root target, supervised by the reference value. Figure 2 presents the architecture for the ensembler.

supervision. We introduce LAoS, Learned Aggregation of Scenarios, which learns to aggregate LLM-generated scenarios using supervision from labeled probability queries. Given a condition c and target quantity Y, an LLM proposes scenarios $c_1, \ldots, c_N \subseteq c$ and estimates their prevalences $\Pr(c_i \mid c)$ and the conditional expectations $\mathbb{E}[Y \mid c_i]$. A neural **ensembler** then processes these estimates together with embeddings of the scenario descriptions to predict the target expectation $\mathbb{E}[Y \mid c]$ and its uncertainty. Self-attention allows the ensembler to compare scenarios, so its use of each estimate can depend on redundancy and disagreement within the collection. Training against the target probabilities teaches the ensembler how these scenario-level estimates relate to the desired answer, enabling it to correct biased estimates, downweight unreliable (high-variance) estimates, and upweight scenarios that cover undersampled regions of the outcome space.

We evaluate LAoS on conditional probability queries derived from the General Social Survey (Davern et al., 2024), with variables held out from training. Our main results show improved accuracy over baselines: direct elicitation, repeated direct elicitation, and prevalence-weighted aggregation. We demonstrate that performance improves, albeit slowly, as we scale up the number of scenarios (§3.3). LAoS also improves the coherence of related probability estimates (§3.4). Finally, we check that the ensembler can cope with poorly chosen sets of scenarios (biased or overlapping), finding larger improvements over prevalence-weighted aggregation under these controlled perturbations (§3.5). We also perform ablations to check that our ensembler architecture is well-chosen (§J).

2 LAoS: Probabilistic Inference through Calibrated Case Analysis

LAoS answers conditional expectation queries expressed in natural language by aggregating estimates of the expectation under finer-grained scenarios proposed by a LLM. We describe the task formulation in §2.1, the generation of scenarios in §2.2, the architecture of our ensembler in §2.3.1, and its training objective in §2.3.2.

2.1 Task Formulation

Let q denote a query about some real-valued random *target* variable Y, given some real-world *scenario* c. It asks us to estimate the following conditional expectation:

$$y^\star = \mathbb{E}[Y \mid c] \tag{1}$$

In our setting, q is formulated in natural language, perhaps using a template that combines natural-language descriptions of Y and c.

2.2 GENERATING SCENARIOS

To estimate the answer to q, we first use an LLM to refine c into more specific scenarios $c_1, \ldots, c_N$. This gives us refined queries $q_1, \ldots, q_N$ where q_i is a natural-language request for $\mathbb{E}[Y \mid c_i]$.

In practice, we ask the LLM for additional details d_i to concatenate with c to get $c_i = c + d_i$. To make sure that $d_1, \ldots, d_N$ are diverse, we generate them all with a single LLM prompt that includes c. Our prompts are given in §F.

As a special case, we force d_1 to always be empty, so that $c_1 = c$ (the **root scenario**) and $q_1 = q$ (the **root query**).

For each i separately, we prompt a LLM again to obtain $p_i \approx \Pr(c_i \mid c) = \Pr(d_i \mid c)$, an estimate of the **prevalence** of scenario c_i, as well as $y_i \approx \mathbb{E}[Y \mid c_i]$, an estimate of the answer to query q_i.

We then embed each natural-language q_i to obtain a vector $\boldsymbol{e}_i$.[1] The concatenation of the embedding, the prevalence p_i, and the target y_i for each c_i then forms the sequence of input vectors to our ensembler.

$$\mathcal{X} = ([\boldsymbol{e}_i; y_i; p_i])_{i=1}^{N}. \tag{2}$$

2.3 SCENARIO ENSEMBLER

As it is challenging to design a principled approach to aggregate over scenario estimates, we propose an ensembler that is trained on queries q for which the true conditional expectation $y^\star$ is known. We describe our transformer-like architecture in §2.3.1 and Figure 2(a), including how we construct input representations and extract output predictions. We then give our regularized training objective in §2.3.2.

2.3.1 ARCHITECTURE

Recall that $\mathcal{X} = ([\boldsymbol{e}_i; y_i; p_i])_{i=1}^{N}$ denotes the properties of query q that are the inputs to the ensembler. We use $\boldsymbol{h}_i^{(\ell)} \in \mathbb{R}^D$ to denote the representation of scenario i at layer ℓ of the ensembler, and $H^{(\ell)} = [\boldsymbol{h}_1^{(\ell)}, \ldots, \boldsymbol{h}_N^{(\ell)}] \in \mathbb{R}^{D \times N}$ to denote the entire layer.

Embedding Layer We use a linear function with dropout to reduce the dimensionality of each query's embedding:

$$\boldsymbol{v}_i^{(0)} = \text{dropout}(W\boldsymbol{e}_i + \boldsymbol{b}) \in \mathbb{R}^{D-2} \tag{3}$$

This lets the model learn to focus on more relevant dimensions of the text embedding. We now set

$$\boldsymbol{h}_i^{(0)} = [y_i; \text{logit}(p_i); \boldsymbol{v}_i^{(0)}] \in \mathbb{R}^D \tag{4}$$

Note that we do not use positional embeddings, as the queries are unordered. The only special query is the first one—the root query, which is already marked by the fact that $p_i = 1$.

In equation (4) and elsewhere, we use a clipped version of the logit transform (taking $\epsilon = 10^{-8}$):

$$\text{logit}(p) = \log \frac{p_\epsilon}{1 - p_\epsilon} \qquad p_\epsilon = \min(\max(p, \epsilon), 1 - \epsilon) \tag{5}$$

For our general presentation, we have defined $y = \mathbb{E}[Y \mid c]$. However, when Y is a binary variable (as in our present experiments), so that the conditional expectation $\mathbb{E}[Y \mid c]$ is a probability, we redefine $y = \text{logit}(\mathbb{E}[Y \mid c])$. Thus, the first dimension of equation (4) is actually a logit.

Encoder Layers Each encoder block is a self-attention layer followed by a multi-layer perceptron, each additively contributing to a residual stream:

$$\begin{aligned} \widetilde{H}^{(\ell)} &= H^{(\ell-1)} + \text{SelfAttn}_\ell(H^{(\ell-1)}), \\ H^{(\ell)} &= \widetilde{H}^{(\ell)} + \text{FFN}_\ell(\widetilde{H}^{(\ell)}). \end{aligned} \tag{6}$$

[1] We use the first 1024 dimensions of Qwen3-Embedding-8B.

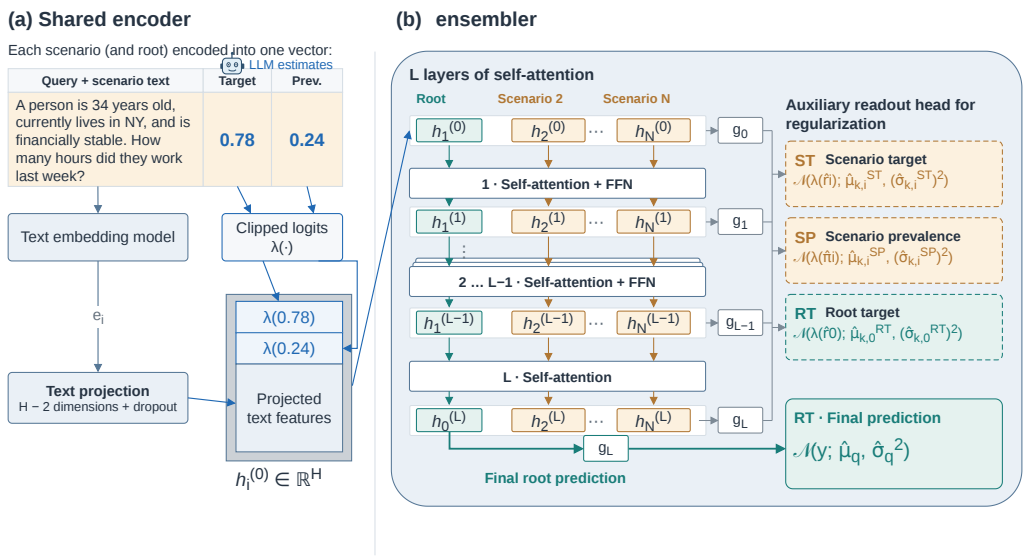


Figure 2: LAoS architecture. (a) The encoder projects text features before concatenating the target and prevalence estimates. (b) Self-attention exchanges information among root and scenario tokens. Readouts at the encoder output and after each of the five attention layers reconstruct the LLM inputs; the final root target readout predicts the reference value. Each supervised channel outputs a Gaussian.

where $\mathrm{SelfAttn}_\ell$ applies multi-head self-attention, and FFN_ℓ applies a feed-forward network to each token. Both operations include pre-layer normalization and output dropout.

The first attention layer ($\ell = 1$) omits pre-attention normalization to preserve the scale of the two logits in equation (4). The final attention layer ($\ell = L$) omits the feed-forward network.

Readout Heads LAoS estimates $y = \mathbb{E}[Y \mid c]$ while also capturing uncertainty about this estimate. That is, it predicts a probability distribution $p_{\boldsymbol{\theta}}$ over y. The family of $p_{\boldsymbol{\theta}}$ depends on the datatype of y, and the readout head outputs its parameters $\boldsymbol{\theta}$.

When each y is an unbounded real number, we can use a *Gaussian* readout that yields the parameters of a normal distribution over $\mathbb{R}$: $\boldsymbol{\theta} = (\mu, \sigma)$ where $\sigma = \mathrm{softplus}(a) + \delta$, $\delta > 0$, and (μ, a) is predicted from $\boldsymbol{h}_0^{(L)}$—the top-layer representation of the root query—with a multi-layer perceptron. This is deep heteroskedastic regression.

On one of our datasets, Y is a binary variable. Recall that in this case we redefine $y = \mathrm{logit}(\mathbb{E}[Y \mid c])$, an unbounded real. We then predict a distribution over this logit using a Gaussian readout.[2]

On our other dataset, the supervised answer $y^\star$ is always given as 0 or 1. Thus, we use a *Bernoulli* readout that yields the parameter of a Bernoulli distribution over $\{0, 1\}$: $\boldsymbol{\theta} = (p)$ where $p = \sigma(l)$ and l is again predicted from $\boldsymbol{h}_0^{(L)}$ with a multi-layer perceptron.

Other distribution families could be used as appropriate. For example, if $y \geq 0$, then we could have the readout head output the parameters of a log-normal distribution.

2.3.2 Training

[2]More generally, in settings where Y ranges over k possible outcomes (e.g., married/single/divorced) and thus $\mathbb{E}[Y \mid c]$ is a probability distribution, we could use a *Logistic Normal* readout that predicts the parameters (μ, Σ) of a logistic normal distribution over the $(k-1)$-simplex. For $k = 2$, this reduces to our practice above, except that the log-likelihood loss in §2.3.2 would be slightly different due to the change of variable.

Prediction Loss For each training query q with supervised answer $y^\star$, we incur loss

$$\mathcal{L}_{\text{final}} = -\log p_{\boldsymbol{\theta}}(y^\star) \tag{7}$$

where $p_{\boldsymbol{\theta}}$ is the predicted distribution over y.

Auxiliary Loss We add an auxiliary loss $\mathcal{L}_{\text{Aux}}$ to enhance the retention of the two numerical input features p_i and y_i with a reconstruction loss at every layer. Details are provided in §B. Our final objective is thus

$$\mathcal{L} = \mathcal{L}_{\text{final}} + \mathcal{L}_{aux} \tag{8}$$

2.4 Separating Mean and Variance Supervision

For NLL loss, joint mean and variance training can become overconfident as the mean overfits to the training queries: increasingly small residuals reward increasingly small predicted variances (Jordahn et al., 2026). For a fixed mean μ and label y, writing $v = \sigma^2$ gives

$$\frac{\partial \,\mathrm{NLL}}{\partial v} = \frac{v - (y-\mu)^2}{2v^2}, \qquad v_{\text{opt}} = (y-\mu)^2. \tag{9}$$

Consequently, learning the smaller training-error scale underestimates uncertainty when errors on unseen queries are larger. The inverse-variance factor in the mean gradient also couples uncertainty to which queries are fitted most strongly (Seitzer et al., 2022).

We separate direct mean and variance supervision using a fixed split of training queries into $\mathcal{D}_\mu$ and $\mathcal{D}_\sigma$. All instances of the same root query are assigned to the same subset. For every final readout head, we backpropagate either the variance or the mean update gradients, depending on the split:

$$\nabla_{(\mu,\sigma)} \,\mathrm{NLL}_{\text{split}}(q; y, \mu, \sigma) = \begin{cases} \left(\dfrac{\mu - y}{\sigma^2},\, 0\right), & q \in \mathcal{D}_\mu, \\ \left(0,\, \dfrac{1}{\sigma} - \dfrac{(y-\mu)^2}{\sigma^3}\right), & q \in \mathcal{D}_\sigma. \end{cases} \tag{10}$$

So mean and variance learning update the ensembler jointly but never update on the same datapoint, enabling it to make better mean estimates informed by calibrated variance estimates. We demonstrate in §J.6 that LAoS achieves more accurate predictions with detached mean and variance gradients when trained with NLL loss, compared to naive heteroskedastic readout and homoskedastic prediction with a fixed $\sigma^2 = 1$.

3 Experiments

3.1 Experiment Setup

Data We evaluate our method on the General Social Survey Dataset (GSS) and Com2Sense (C2S) (Davern et al., 2024; Singh et al., 2021). On GSS, we curate a training set from 94 variables as target variable Y, each coupled with around 18 root scenarios c, forming a total of 1685 conditional expectations queries. Test set are 169 queries curated from 9 target variables with the same pipeline. C2S contains comparisons about everyday scenarios. We use 2301 queries across 130 target variables for training, and 1521 queries across 86 target variables for evaluation. Examples for both datasets are given in §H.

LLMs & Pretrained Embeddings We report our main experiment results on Qwen3.6-27B (Qwen Team, 2026). In §3.4, we show that our method improves prediction coherency for Gemma3-12B, Gemma3-27B, Qwen3.5-27B, and Qwen3.6-27B. We use Qwen3-Embedding-8B to embed scenario text descriptions. Full configurations for LAoS can be found in §C.

Inference Budget & Metric We compare against other inference time scaling methods at matched inference budget, from 1 up to 64 LLM calls per question. Under each budget, we report the Kullback–Leibler (KL) divergence, mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) for each tested method.

3.2 BASELINES

We compare our method against several baseline methods:

- **(Repeated) direct prompting** We query the LLM to directly estimate the target expectation, potentially repeating multiple times and taking their average.
- **Prevalence-weighted averaging** We query the LLM to estimate the target quantity under *scenarios*, and average scenario estimates by their estimated prevalence, namely $(\sum_i p_i \cdot y_i)/\sum_i p_i$.
- **LLM as an ensembler** We prompt the LLM with c_i (the scenario description) alongside p_i and y_i (the raw LLM estimates), asking for an estimate of the target quantity informed by such scenario-level beliefs.
- **BIRD** builds an ad hoc Bayesian model for each question with case analysis (Feng et al., 2025).
- **LORA**: Wang et al. (2025) performs LoRA fine-tuning on LLMs with a diverse probabilistic QA dataset, demonstrating improved probabilistic estimates across multiple domains unseen during training, including C2S. We finetune Qwen3.6-27B using their released schema as a reference for methods beyond inference-time scaling. We do not finetune the LLM on C2S or GSS directly due to the limited dataset size.

3.3 MAIN RESULTS

Table 1 and Figure 3 summarize our main results. With an ensembler trained on the corresponding dataset, LAOS achieves the lowest KL, MSE, MAE, and RMSE on both benchmarks at all three reported budgets. At 64 LLM calls, LAOS reaches a KL of 0.5362 on COM2SENSE, reducing KL by 25.1% relative to repeated prompting and by 27.8% relative to prevalence-weighted averaging. It also outperforms LoRA and BIRD, which achieve KL values of 0.5774 and 0.5844, respectively. On GSS, LAOS achieves a KL of 0.0723, a 31.9% reduction relative to prevalence-weighted averaging (0.1062), the strongest baseline.

LAOS benefits from increased inference-time budget, with larger inference-scaling gains in both domains compared to naive baselines(Figure 3). On Com2Sense, Increasing the budget from 16 to 64 calls reduces its KL from 0.5592 to 0.5362 (4.1%) and RMSE from 0.4307 to 0.4193 (2.6%). Over the same range, repeated prompting reduces KL by only 0.8% and RMSE by 0.2%, while prevalence-weighted averaging reduces KL by 0.7% but increases RMSE. BIRD improves substantially between 16 and 32 calls, but gains little thereafter (KL decreases by only 0.3% from 32 to 64 calls, compared with 2.5% for LAOS). On GSS, performance improvements from inference-scaling are smaller across methods compared to Com2Sense. LAOS's KL falls from 0.0757 at 16 calls to 0.0722 at 32 calls and remains nearly unchanged at 64 calls (0.0723). Prevalence-weighted averaging and LLM-as-ensembler does not scale on GSS. Prevalence-weighted averaging improves initially, but its KL rises from 0.0960 at 32 calls to 0.1062 at 64 calls. The LLM ensembler performs worse when given 30 scenarios compared to when given only 7 scenarios, with KL increasing from 0.1407 to 0.1596.

3.4 LAOS IMPROVES ESTIMATION COHERENCY

LLMs are reported to yield incoherent estimates on related quantities, violating basic laws of probability (Zhu & Griffiths, 2024; Paleka et al., 2025). In this section, we show that our method alleviates this issue and produces more coherent estimates. On the GSS dataset, we report

- **Order invariance** whether the LLM is coherent when we change the order in which we list the conditions in the prompt, namely $\mathbb{E}[Y \mid A, s] = \mathbb{E}[Y \mid s, A]$
- **Chain rule** whether a direct estimate of a joint probability agrees with the product of two separately elicited conditional probabilities, namely $P(s_1, s_2 \mid A) = P(s_1 \mid A)P(s_2 \mid A, s_1)$.
- **Model disagreement (reference)** whether two LLMs agree on an estimate, as a reference to interpret the scale of incoherency

Table 2 presents results for Qwen3.5-27B, Qwen3.6-27B, Gemma-3-12B, and Gemma-3-27B, measured as mean absolute differences between the log-odds of the paired estimates. Raw LLM estimates exhibit substantial incoherence, with chain-rule gaps exceeding the disagreement between different models within each family. For Qwen3.6 and Qwen3.5, the chain-rule gaps are 0.635 and 0.654, respectively, compared with 0.420 between the two models. Gemma-3-12B and Gemma-3-27B exhibit larger chain-rule gaps of 1.137 and 1.190, compared with 0.771 between the two

Method	16 LLM calls				32 LLM calls				64 LLM calls			
	KL	MSE	MAE	RMSE	KL	MSE	MAE	RMSE	KL	MSE	MAE	RMSE
Com2Sense test												
LLM direct prompting	0.8931	0.2275	0.3038	0.4769	0.8931	0.2275	0.3038	0.4769	0.8931	0.2275	0.3038	0.4769
Repeated prompting	0.7212	0.2072	0.3015	0.4552	0.7190	0.2075	0.3019	0.4555	0.7156	0.2065	0.3008	0.4545
LLM ensembler	0.8850	0.2343	0.3084	0.4840	0.9021	0.2311	0.3022	0.4807	0.8966	0.2343	0.3045	0.4841
Prevalence-weighted scenario avg	0.7479	0.2171	0.3247	0.4659	0.7357	0.2198	0.3342	0.4688	0.7424	0.2266	0.3439	0.4761
BIRD	0.7641	0.2193	0.4264	0.4683	0.5863	0.2029	0.4189	0.4504	0.5844	0.2003	0.4249	0.4476
LORA	0.5774	0.1907	0.3520	0.4366	0.5774	0.1907	0.3520	0.4366	0.5774	0.1907	0.3520	0.4366
LAoS (Trained on GSS)	0.5917	0.1993	0.3670	0.4465	0.5883	0.1985	0.3634	0.4456	0.5876	0.1988	0.3619	0.4459
LAoS (Trained on C2S)	**0.5592**	**0.1855**	**0.2823**	**0.4307**	**0.5498**	**0.1807**	**0.2720**	**0.4251**	**0.5362**	**0.1758**	**0.2650**	**0.4193**
GSS test												
LLM direct prompting	0.1538	0.0560	0.1751	0.2367	0.1538	0.0560	0.1751	0.2367	0.1538	0.0560	0.1751	0.2367
Repeated prompting	0.1221	0.0469	0.1667	0.2165	0.1178	0.0451	0.1635	0.2123	0.1165	0.0445	0.1624	0.2109
LLM ensembler	0.1407	0.0505	0.1663	0.2248	0.1507	0.0535	0.1733	0.2312	0.1596	0.0556	0.1798	0.2358
Prevalence-weighted scenario avg	0.0999	0.0388	0.1497	0.1969	0.0960	0.0370	0.1472	0.1923	0.1062	0.0405	0.1571	0.2013
BIRD	0.1558	0.0654	0.2133	0.2558	0.1487	0.0646	0.2127	0.2541	0.1526	0.0656	0.2135	0.2561
LORA	0.1696	0.0712	0.2319	0.2669	0.1696	0.0712	0.2319	0.2669	0.1696	0.0712	0.2319	0.2669
LAoS (Trained on GSS)	**0.0757**	**0.0306**	**0.1354**	**0.1748**	**0.0722**	**0.0292**	**0.1339**	**0.1708**	**0.0723**	**0.0292**	**0.1357**	**0.1710**
LAoS (Trained on C2S)	0.1856	0.0777	0.2284	0.2788	0.1538	0.0624	0.1968	0.2498	0.1509	0.0595	0.1877	0.2438

Table 1: Prediction error on COM2SENSE and GSS at three LLM-call budgets. Our method achieves the lowest KL, MSE, MAE, and RMSE across tested budgets. **Bold** indicates the best performance, with underline indicating no statistical significance from best entry under paired permutation test.

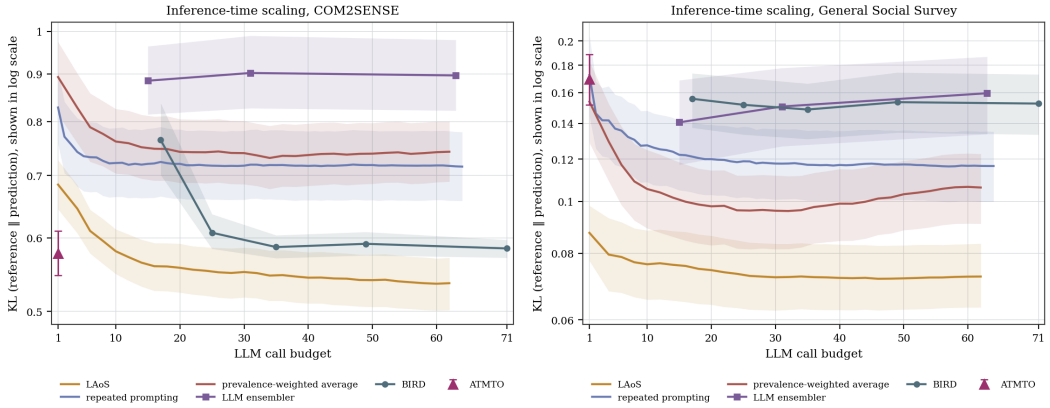


Figure 3: Prediction error versus LLM-call budget on COM2SENSE and GSS. LAOS achieves the lowest KL among tested methods, and improves with additional computation in both domains.

models. Simply changing the order of the conditions also produces gaps of 0.186–0.312 across the four models.

LAOS improves estimation coherency across all four models, reducing both order and chain-rule gaps. Order gaps decrease by 17.9–46.5%, with the largest reduction for Gemma-3-12B, whose gap falls from 0.312 to 0.167. For the chain rule, LAOS reduces the gap from 0.635 to 0.427 for Qwen3.6 and from 0.654 to 0.457 for Qwen3.5, improvements of 32.8% and 30.1%, respectively. Gemma-3-27B also improves from 1.190 to 1.048 (11.9%), while the improvement for Gemma-3-12B is smaller, from 1.137 to 1.108 (2.6%).

3.5 LAOS YIELDS ROBUST PREDICTIONS EVEN WITH POORLY CHOSEN TEST-TIME SCENARIOS

Our method is designed to work with overlapping or biased scenarios through learned aggregation. In this section, we stress-test this ability on GSS using scenarios that are deliberately biased or heavily overlapping. All methods use Qwen3.6-27B estimates and receive the same scenarios at each count. The generation procedures are described in §G, with examples in §I. Besides prevalence-weighted averaging, we compare against *single-correction* LAOS, an ablation trained on a single scenario per query. This ablation learns to correct individual LLM estimates without learning how to aggregate estimates across scenarios. At inference, it corrects each scenario estimate separately,

	Order invariance				Chain rule				Model comparison	
	Qwen3.6	Qwen3.5	Gemma 12B	Gemma 27B	Qwen3.6	Qwen3.5	Gemma 12B	Gemma 27B	Qwen3.6 vs 3.5	Gemma 12B vs 27B
n	982	982	982	982	997	997	997	997	1979	1979
raw LLM	0.263 [0.236, 0.292]	0.236 [0.210, 0.262]	0.312 [0.280, 0.343]	0.186 [0.160, 0.211]	0.635 [0.602, 0.668]	0.654 [0.623, 0.685]	1.137 [1.086, 1.191]	1.190 [1.142, 1.236]	**0.420** [0.398, 0.442]	0.771 [0.733, 0.809]
LAOS	**0.216** [0.203, 0.229]	**0.165** [0.155, 0.175]	**0.167** [0.157, 0.179]	**0.144** [0.132, 0.156]	**0.427** [0.407, 0.448]	**0.457** [0.437, 0.479]	**1.108** [1.082, 1.135]	**1.048** [1.024, 1.074]	0.616 [0.598, 0.633]	**0.592** [0.572, 0.612]
noise floor	0.173	0.149	0.060	0.030	0.219	0.194	0.077	0.035	0.171	0.069

Table 2: Mean absolute log-odds gaps on GSS, with 95% bootstrap confidence intervals over queries. Order and chain-rule tests compare equivalent formulations; model comparison measures within-family disagreement on their pooled queries. The noise floor is the gap between repeated elicitations of the same query. Lower of raw LLM and LAOS is bold.

	Overlapping scenarios (n=997)		Biased proposal (n=860)	
Method	k=10	k=30	k=10	k=30
Prevalence-weighted avg.	0.2462	0.1933	0.1760	0.1722
Single-correction LAOS	0.1802	0.1508	0.2018	0.2004
LAOS	**0.1325**	**0.1328**	**0.1316**	**0.1333**

Table 3: Stress-test MAE at 10 and 30 scenarios, using Qwen3.6-27B estimates on GSS. All methods omit the direct root estimate. Single-correction LAOS corrects scenario estimates individually before prevalence averaging. Lower is better; column minima are bold.

and we average the corrected estimates using the LLM-predicted prevalences as weights to produce the final estimate.

Table 3 and Figure 4 show that LAOS outperforms both baselines under both adversarial settings. With 30 overlapping scenarios, LAOS achieves an MAE of 0.1328, compared with 0.1933 for prevalence-weighted averaging and 0.1508 for single-correction LAOS. Both baselines gain considerable improvement when given a second set of overlapping situation, namely $k = 11$ through 20. LAOS do not scale significantly with more scenarios, compared to when scenarios are diverse (Figure 3). Under biased proposals, LAOS achieves an MAE of 0.1333 at 30 scenarios, compared with 0.1722 for prevalence-weighted averaging and 0.2004 for single-correction LAOS.

4 Related Work

Large language models (LLMs) are increasingly being used to estimate real-world quantities (Chen et al., 2020; Gu et al., 2024; Nafar et al., 2026; Selby et al., 2025). This is partially motivated by their broad knowledge and predictive capabilities (Zhang et al., 2020; Paruchuri et al., 2024; Ashokkumar et al., 2026). A major type of "knowledge about the real world" is the probability of real-world events. LLMs are able to supply probabilistic predictions in such settings as priors for Bayesian workflows (Selby et al., 2025; Nafar et al., 2026). Yet, LLMs are also incoherent and uncalibrated (Zhu & Griffiths, 2024; Lovering et al., 2024). Specifically, they yield contradictory predictions on related quantities such as complementary events (Zhu & Griffiths, 2024; Freedman & Toni, 2025), and are highly variable under changes of prompt phrasing (Ren et al., 2025). This motivates methods that improve the LLM to give more accurate and coherent estimates (Nan et al., 2026; Zhu et al., 2025; Yamin et al., 2026). A straightforward approach is to fine-tune the LLM on real-world quantity or probability estimation datasets, though the LLM do not always generalize to domains that the training set do not cover (Wang et al., 2025; Suh et al., 2025). There's also works that build probabilistic models on top of language models, translating LLM outputs to coherent probabilistic knowledge (Wong et al., 2023).

Test-time scaling aims to improve language model performance by allocating additional computation during inference (Snell et al., 2025; Brown et al., 2024). This computation can support longer reasoning trajectories, iterative refinement, or the generation of multiple candidate solutions that are subsequently selected or aggregated (Wei et al., 2022; Khot et al., 2023; Ozturkler et al., 2023; Sanders & Van Durme, 2026; Brown et al., 2024). Various kinds of inference strategies have been proposed, such as majority voting, best-of-n, and weighted voting (Wang et al., 2023; Wu et al., 2025). The aggregation of samples of LLM answers into a corrected answer can be done by an auxiliary ensembler trained on held-out data (Qi et al., 2025). We propose such an ensembler for

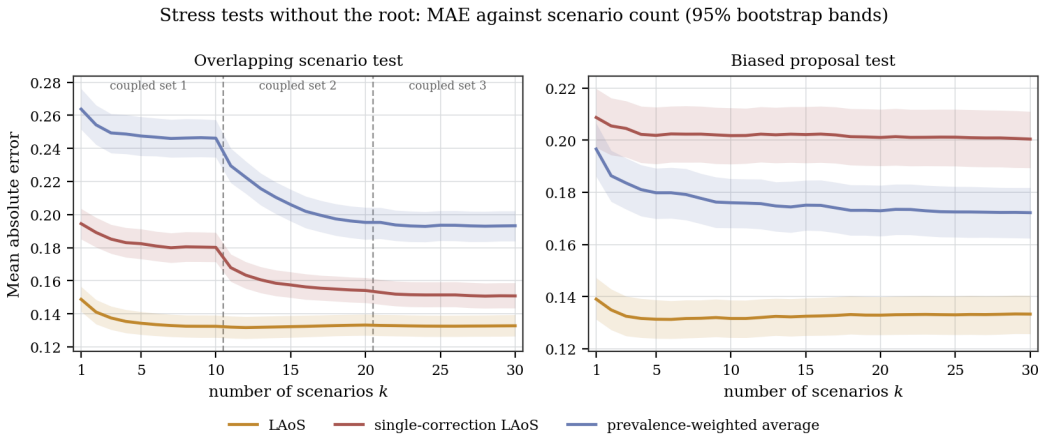


Figure 4: KL versus scenario count in GSS stress tests, using Qwen3.6-27B estimates without the direct root estimate: overlapping scenarios (left; 997 queries) and biased proposals (right; 860 queries). All methods use the same scenario prefixes. Vertical dashed lines separate groups of ten related scenarios. Bands are pointwise 95% bootstrap confidence intervals over queries (2,000 resamples).

LLM's sampled analysis under *scenarios*, which are not as coupled as repeated samples of the direct target question, thus may saturate less with scaling.

An instance of test-time scaling is case analysis, which asks a target query under more specific scenarios and combines the resulting conclusions into a final answer (Didolkar et al., 2025). Many LLM workflows have been proposed that exploits case-analysis to elicit more accurate answers from the LLM (Chen et al., 2026; Xia et al., 2024). To name a few, Maieutic Prompting recursively generates explanations supporting alternative answers and uses a weighted satisfiability solver to infer truth assignments that best satisfy their logical relations (Jung et al., 2022). Probabilistic Tree-of-thought Reasoning decomposes a complex query into a tree of sub-questions and answers them from leaves to root, then aggregates using confidence estimates (Cao et al., 2023). BIRD builds an ad hoc Bayesian model for each question with case analysis (Feng et al., 2025). These workflows either restrict the considered scenarios to exhaustive partitions of the outcome space, i.e. cases, where aggregation is straightforward, or use an LLM to handle the aggregation for non-exhaustive and overlapping scenarios. Our experiments show that aggregation by prompting an LLM is prone to errors and do not scale with computation, and that constraining scenarios to cases limits performance. This thus motivates our method, which learns the aggregation of scenario estimates with a Transformer-like module.

5 Conclusion

We introduced LAoS, probabilistic inference through calibrated case analysis, which learns to ensemble LLM probability estimates under proposed scenarios into a calibrated and more informed prediction of the target probabilistic quantity, using supervision on queries with label. The LLM remains frozen, with the ensembler the only training component, making the method light-weight. Aggregation requires neither disjoint scenarios nor a known proposal distribution. On Com2Sense, LAoS achieves the lowest KL and RMSE among the evaluated methods, reducing KL by 36% relative to direct prompting at the largest budget. On GSS, we demonstrate that LAoS outperforms prevalence weighting and independent correction of scenario estimates under controlled overlap and proposal bias. LAoS also reduces order-invariance violations across four LLMs without explicit consistency constraints. Chain-rule violations remain, suggesting a role for constraints across related queries in future work.

6 AI USE STATEMENT

We used AI tools to assist with code implementation, identifying related work, and polishing the language of author-written text. AI tools were not used to generate or synthesize research ideas or to draft manuscript sections. The authors retain full responsibility for the implementation, references, and content of the manuscript.

7 REPRODUCIBILITY STATEMENT

Both datasets are publicly available (Davern et al., 2024; Singh et al., 2021); §3 describes how queries and train/test splits are constructed, and §H shows examples. All models are open-weight: Qwen3.6-27B (with Gemma3 and Qwen3.5 in §3.4) as the LLM and Qwen3-Embedding-8B for embeddings. Scenario generation and estimate elicitation are described in §2.2, with every prompt template reproduced verbatim in §F, including those of the LLM-ensembler and stress-test baselines (§G). The LAoS architecture and training objective are given in §2.3.1, §2.3.2 and §2.4; the full training configuration, hyperparameter search, and hypergradient tuning of regularization weights are in §§C–E.

REFERENCES

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, 656(8126):115–122, 2026. doi: 10.1038/s41586-026-10742-x.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. URL https://arxiv.org/abs/2407.21787.

Shulin Cao, Jiajie Zhang, Jiaxin Shi, Xin Lv, Zijun Yao, Qi Tian, Lei Hou, and Juanzi Li. Probabilistic tree-of-thought reasoning for answering knowledge-intensive complex questions. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 12541–12560, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.835. URL https://aclanthology.org/2023.findings-emnlp.835/.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. Uncertain natural language inference. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8772–8779, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.774. URL https://aclanthology.org/2020.acl-main.774/.

Yufei Chen, Yiran Zhao, Xiaogang Xu, Qipeng Xie, Jiafei Wu, and Zhe Liu. LEAP: Likelihood elicitation and aggregation for LLM-based probabilistic forecasting. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, 2026. URL https://arxiv.org/abs/2609.01337. To appear.

André F. Cruz, Moritz Hardt, and Celestine Mendler-Dünner. Evaluating language models as risk scores. In *Advances in Neural Information Processing Systems*, volume 37, pp. 97378–97407. Curran Associates, Inc., 2024. doi: 10.52202/079017-3089. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/b0a4b3e384b4554e65a47ad1f6b0310a-Paper-Datasets_and_Benchmarks_Track.pdf.

Michael Davern, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. General Social Survey 1972–2022. [Machine-readable data file]. Principal Investigator, Michael Davern; Co-Principal Investigators, Rene Bautista, Jeremy Freese, Pamela Herd, and Stephen L. Morgan. NORC ed. Chicago: NORC at the University of Chicago. 1 datafile (Release 4) and 1 codebook (2022 Release 4), 2024. URL https://gss.norc.org.

Aniket Didolkar, Nicolas Ballas, Sanjeev Arora, and Anirudh Goyal. Metacognitive reuse: Turning recurring LLM reasoning into concise behaviors. *arXiv preprint arXiv:2509.13237*, 2025. URL https://arxiv.org/abs/2509.13237.

Yu Feng, Ben Zhou, Weidong Lin, and Dan Roth. BIRD: A trustworthy Bayesian inference framework for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fAAaT826Vv.

Gabriel Freedman and Francesca Toni. Exploring the potential for large language models to demonstrate rational probabilistic beliefs. In *The International FLAIRS Conference Proceedings*, volume 38. University of Florida George A Smathers Libraries, 2025. doi: 10.32473/flairs.38.1.138892.

Bowen Gu, Rishi J. Desai, Kueiyu Joshua Lin, and Jie Yang. Probabilistic medical predictions of large language models. *npj Digital Medicine*, 7(1), 2024. doi: 10.1038/s41746-024-01366-4.

Bairu Hou, Yujian Liu, Kaizhi Qian, Jacob Andreas, Shiyu Chang, and Yang Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 19023–19042. PMLR, 2024. URL https://proceedings.mlr.press/v235/hou24b.html.

P. N. Johnson-Laird, Paolo Legrenzi, Vittorio Girotto, Maria Sonino Legrenzi, and Jean-Paul Caverni. Naive probability: A mental model theory of extensional reasoning. *Psychological Review*, 106(1):62–88, 1999. doi: 10.1037/0033-295x.106.1.62.

Mikkel Jordahn, Jonas Vestergaard Jensen, James Harrison, Michael Riis Andersen, and Mikkel N. Schmidt. Practical deep heteroskedastic regression. *arXiv preprint arXiv:2603.01750*, 2026. URL https://arxiv.org/abs/2603.01750.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1266–1279, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.82. URL https://aclanthology.org/2022.emnlp-main.82/.

Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_nGgzQjzaRy.

Charles Lovering, Michael Krumdick, Viet Dac Lai, Seth Ebner, Nilesh Kumar, Varshini Reddy, Rik Koncel-Kedziorski, and Chris Tanner. Language model probabilities are not calibrated in numeric contexts. *arXiv preprint arXiv:2410.16007*, 2024. URL https://arxiv.org/abs/2410.16007.

Jelena Luketina, Mathias Berglund, Klaus Greff, and Tapani Raiko. Scalable gradient-based tuning of continuous regularization hyperparameters. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pp. 2952–2960. PMLR, 2016.

Aliakbar Nafar, K. Brent Venable, Zijun Cui, and Parisa Kordjamshidi. Extracting probabilistic knowledge from large language models for Bayesian network parameterization. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Fy3Byg3CVo.

Yang Nan, Qihao Wen, Jiahao Wang, Pengfei He, Ravi Tandon, Yong Ge, and Han Xu. Interpretable probability estimation with LLMs via Shapley reconstruction. *arXiv preprint arXiv:2601.09151*, 2026. URL https://arxiv.org/abs/2601.09151.

Batu Ozturkler, Esmeralda S. Whitammer, Zhen Wang, and Nebojsa Jojic. ThinkSum: Probabilistic reasoning over sets using large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1216–1239, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.68. URL https://aclanthology.org/2023.acl-long.68/.

Daniel Paleka, Abhimanyu Pallavi Sudhir, Alejandro Alvarez, Vineeth Bhat, Adam Shen, Evan Wang, and Florian Tramèr. Consistency checks for language model forecasters. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r5IXBlTCGc.

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. URL https://arxiv.org/abs/2406.12830.

Ben Prystawski, Michael Li, and Noah Goodman. Why think step by step? Reasoning emerges from the locality of experience. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70926–70947. Curran Associates, Inc., 2023. doi: 10.52202/075280-3107. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e0af79ad53a336b4c4b4f7e2a68eb609-Paper-Conference.pdf.

Jianing Qi, Xi Ye, Hao Tang, Zhigang Zhu, and Eunsol Choi. Learning to reason across parallel samples for LLM reasoning. *arXiv preprint arXiv:2506.09014*, 2025. URL https://arxiv.org/abs/2506.09014.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Sandro M. Reia and José F. Fontanari. Wisdom of crowds: much ado about nothing. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(5):053402, 2021. doi: 10.1088/1742-5468/abfa1f.

Kevin Ren, Santiago Cortes-Gomez, Carlos Miguel Patiño, Ananya Joshi, Ruiqi Lyu, Jingjing Tang, Alistair Turcan, Khurram Yamin, Steven Wu, and Bryan Wilder. Predicting language models' success at zero-shot probabilistic prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18337–18363, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.995. URL https://aclanthology.org/2025.findings-emnlp.995/.

Kate Sanders and Benjamin Van Durme. Bonsai: Interpretable tree-adaptive grounded reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32884–32892. Association for the Advancement of Artificial Intelligence (AAAI), 2026. doi: 10.1609/aaai.v40i39.40569.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=aPOpXlnV1T.

David Selby, Yuichiro Iwashita, Kai Spriestersbach, Mohammad Saad, Dennis Bappert, Archana Warrier, Sumantrak Mukherjee, Koichi Kise, and Sebastian Vollmer. Had enough of experts? Quantitative knowledge retrieval from large language models. *Stat*, 14(2), 2025. doi: 10.1002/sta4.70054.

Shikhar Singh, Nuan Wen, Yu Hou, Pegah Alipoormolabashi, Te-lin Wu, Xuezhe Ma, and Nanyun Peng. COM2SENSE: A commonsense reasoning benchmark with complementary sentences. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 883–898, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.78. URL https://aclanthology.org/2021.findings-acl.78/.

Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025. URL https://arxiv.org/abs/2502.16761.

Liaoyaqi Wang, Zhengping Jiang, Anqi Liu, and Benjamin Van Durme. Always tell me the odds: Fine-grained conditional probability estimation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=xhDcG8qtw9.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://arxiv.org/abs/2201.11903.

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, 2023. URL https://arxiv.org/abs/2306.12672.

Yangzhen Wu, Zhiqing Sun, Shanda Li, Sean Welleck, and Yiming Yang. Inference scaling laws: An empirical analysis of compute-optimal inference for problem-solving with language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2408.00724.

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. URL https://arxiv.org/abs/2412.02081.

Khurram Yamin, Jingjing Tang, Santiago Cortes-Gomez, Amit Sharma, Eric Horvitz, and Bryan Wilder. When agents say one thing and do another: Validating elicited beliefs from LLMs. *arXiv preprint arXiv:2602.06286*, 2026. URL https://arxiv.org/abs/2602.06286.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. Do language embeddings capture scales? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4889–4896, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.439. URL https://aclanthology.org/2020.findings-emnlp.439/.

Jian-Qiao Zhu and Thomas L. Griffiths. Incoherent probability judgments in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024. URL https://escholarship.org/uc/item/2r68p98c.

Jian-Qiao Zhu, Haijiang Yan, and Thomas L. Griffiths. Recovering event probabilities from large language model embeddings via axiomatic constraints. *arXiv preprint arXiv:2505.07883*, 2025. URL https://arxiv.org/abs/2505.07883.

A Principled aggregation is difficult

In this paper, we take an engineering approach and simply learn an effective aggregation function, using a flexible neural model that we train end-to-end. Might there be a more principled way to aggregate given estimates from a given set of scenarios $c_1, \ldots, c_N$?

Consider taking a weighted average of the scenario-specific estimates. It is tempting to give lower weight w_i to scenarios that are less probable. However, that would be wrong. Highly specific scenarios, despite being extremely improbable, may deserve high weight because they are representative of an entire region.

This suggests an constrained optimization approach that tries to ensure that different regions of the outcome space are covered appropriately. Suppose we have a *partition* of the outcome space into cases $z_1, \ldots, z_M$ with known probabilities $p(z_j)$; for example, we might partition the outcomes by

the values of some variable. Then we would like the total weight of the scenarios that fall into each case to be close to that case's real-world probability:

$$(\forall 1 \le j \le M) \sum_{i=1}^{N} p(z_j \mid c_i) \cdot w_i \approx p(z_j) \tag{11}$$

There are many possible partitions, each of which contributes additional constraints like (11). We could try to choose a weight vector satisfying $\sum_i w_i = 1$ and all of these constraints.

This is not easy to make practical, since it requires finding suitable partitions into cases and assessing whether an instance of scenario c_i is likely to fall into case z_j (the probability $p(z_j \mid c_i)$).

Furthermore, the above idea assumes that the LLM's estimates are accurate. We would like to debias the LLM's estimates, which means learning when to adjust them up or down. And even unbiased or debiased estimates may have high variance. Thus, we would like to learn when the LLM is unreliable and downweight those scenarios accordingly.

Covariance among estimates introduces further complications. Covariance may arise from overlapping scenarios, correlated Y values across related scenarios, or correlated LLM errors across related scenarios.

To address all these questions, we have considered using a Gaussian process to jointly model the true Y values for the infinitely many individual points, the true conditional expectations of Y in the N scenarios, and the LLM estimates of those N conditional expectations.

However, such a method would be quite expensive, and would still rely on dubious embeddings of natural language questions. We therefore thought it prudent to begin with a less articulated approach that simply trains a generic deep network end-to-end, in hopes that it would sort out all of the issues above.

B Auxiliary Loss

Auxiliary Reconstruction Loss The two dimensions at the start of equation (4) are crucial to preserve. Thus, we take pains to ensure that $\boldsymbol{h}_i^{(\ell)}$ retains this information. We do this by attaching additional Gaussian readout heads to $\boldsymbol{h}_i^{(\ell)}$, which try to predict these two scalars. Specifically, after each encoder layer ℓ, and for each scenario c_i, we attach readout heads (defined in §2.3.1) to reconstruct its numerical input features y_i and $\mathrm{logit}(p_i)$. For *non-root scenarios* (i.e. c_i for $i > 1$), we attach two readout layers ST (for scenario target) to predict y_i and SP (for scenario prevalence) to predict p_i, both heterosckedastically. For the *root scenario* c_1, we attach a readout layer RT (for root target) to predict y_1 for all but the last layer, since the last layer for root is attached to final readout and is supervised by the main prediction loss. Figure 2 shows a schematic of the different readout layers.

Each layer uses its own parameters for the auxiliary readout heads.

$$\mathcal{L}_{\text{ST}}(p_{\text{ST}}{}_i^{\ell}, y_i) := -\log p_{\text{ST}}{}_i^{\ell}(y_i) \qquad 1 \le \ell \le L \tag{12}$$

$$\mathcal{L}_{\text{SP}}(p_{\text{SP}}{}_i^{\ell}, p_i) := -\log p_{\text{SP}}{}_i^{\ell}(p_i) \qquad 1 \le \ell \le L \tag{13}$$

$$\mathcal{L}_{\text{RT}}(p_{\text{RT}}{}_i^{\ell}, y_1) := -\log p_{\text{RT}}{}_i^{\ell}(y_1) \qquad 1 \le \ell < L \tag{14}$$

Final Objective Our final objective is

$$\mathcal{L}(\mathbf{p}_{\text{ST}}, \mathbf{p}_{\text{SP}}, \mathbf{p}_{\text{RT}}, p_{\text{FINAL}}; \mathbf{p}, \mathbf{y}, y^\star) = \mathcal{L}_{\text{final}}(p_{\text{FINAL}}, y^\star) + \mathcal{L}_{\text{aux}}(\mathbf{p}_{\text{ST}}, \mathbf{p}_{\text{SP}}, \mathbf{p}_{\text{RT}}; \mathbf{p}, \mathbf{y}) \tag{15}$$

$$\mathcal{L}_{\text{aux}}(\mathbf{p}_{\text{ST}}, \mathbf{p}_{\text{SP}}, \mathbf{p}_{\text{RT}}; \mathbf{p}, \mathbf{y}) = \sum_{\ell} \sum_{i>1} \lambda_{\ell,\text{ST}} \mathcal{L}_{\text{ST}}(p_{\text{ST}}{}_i^{\ell}, y_i) + \lambda_{\ell,\text{SP}} \mathcal{L}_{\text{SP}}(p_{\text{SP}}{}_i^{\ell}, p_i) \tag{16}$$

$$+ \sum_{\ell<L} \lambda_{\ell,\text{RT}} \mathcal{L}_{\text{RT}}(p_{\text{RT}}, y_1) \tag{17}$$

We assign a nonnegative weight $\lambda_{l,e}$ to each auxiliary task for $e \in \{\text{ST}, \text{SP}, \text{RT}\}$, and choose these weights online through hyper-gradient (Luketina et al., 2016).

C TRAINING CONFIGURATION

Encoder	concatenation of target logit, prevalence logit and text embedding
Inter-scenario layers	4 self-attention layers, 4 heads, feed-forward width 256, dropout 0.2
Hidden width	512
Read-out	a prediction head on every token; the answer is read from the root token
Embeddings	Qwen3-Embedding-8B, last-token pooling, dimension 1024
Input scaling	none (inputs are never standardized)
Optimizer	Adam, learning rate 7×10^{-5}, weight decay 10^{-5}
Batch size	32, gradient clipping 0.75
Schedule	up to 200 epochs, early stopping with patience 10
Validation	10% of target variables held out whole (group-disjoint)
Mean/variance split	30% of queries supervise the scale, the rest the mean

In the stress tests LAOS draws a random prefix of the scenarios each epoch; single-correction LAOS sets the scenario count to zero; the no-root variant reads out by cross-attention over the scenario states alone and therefore carries no input-reconstruction loss.

ATMTO uses LoRA with rank 16 and $\alpha = 32$ on the attention and MLP projections, 100 added bin tokens, a soft target formed by discretizing a Gaussian of standard deviation 0.05, learning rate 2×10^{-5} and maximum sequence length 256. We depart from the released recipe in three declared ways: a warmup ratio of 0.03 in place of a fixed 5,000-step warmup, which would never complete at our training lengths; one epoch over their 314k examples rather than four; and no rank-consistency margin term.

D HYPERPARAMETER TUNING

We tune LAOS on the GSS validation set in several stages:

Optimizer and regularization strength.

learning rate	weight decay	dropout	val. MAE
2×10^{-4}	10^{-5}	0.4	**0.759** $\pm$ 0.182
2×10^{-4}	10^{-3}	0.2	0.762 $\pm$ 0.166
2×10^{-4}	10^{-2}	0.4	0.763 $\pm$ 0.154
2×10^{-5}	10^{-5}	0.2	0.770 $\pm$ 0.146
2×10^{-4}	10^{-2}	0.2	0.774 $\pm$ 0.145
$7 \times 10^{-5\dagger}$	10^{-5}	0.2	0.776 $\pm$ 0.193
7×10^{-5}	10^{-3}	0.4	0.786 $\pm$ 0.173
7×10^{-5}	10^{-3}	0.2	0.792 $\pm$ 0.145

The remaining ten settings of the $3 \times 3 \times 2$ *grid span 0.794–0.829.* † *Setting used in the paper.*

Network size.

attention layers	hidden dim	batch	val. MAE
2	256	32	**0.765** $\pm$ 0.143
2	512	32	0.774 $\pm$ 0.175
2	512	64	0.782 $\pm$ 0.157
4	256	32	0.795 $\pm$ 0.201
4	256	64	0.795 $\pm$ 0.147
4†	512	32	0.799 $\pm$ 0.149
2	256	64	0.809 $\pm$ 0.119
4	512	64	0.811 $\pm$ 0.163

† *Setting used in the paper.*

Readout task Position k	Encoder 0	Block 1 1	Block 2 2	Block 3 3	Block 4 4	Final 5
Scenario target	0.023	0.028	0.064	0.125	0.088	**1.153**
Scenario prevalence	0.028	0.008	0.090	0.109	**0.163**	0.047
Root target	0.031	0.004	0.075	0.125	0.140	1 (fixed)

Table 4: Hypergradient-tuned weights for three readout tasks at six positions. All free weights supervise reconstruction of LLM inputs. The final root target instead uses the reference value and has fixed weight one. Scenario losses are averaged: with $N = 30$, each scenario's contribution to the shared-network gradient is scaled by $w/30$, giving aggregate coefficient w for that task.

The reconstruction weights are tuned separately by hypergradient (§E). Against an unregularized control, the tuned weights improve validation MAE by 0.0322 (95% CI $[0.0067, 0.0577]$, $p = 0.020$). Weights are averaged across the folds that did not produce them, so they are never evaluated on the fold that selected them, and are then held fixed for every model in the paper.

E Hyperparameter tuning of regularization weights using hypergradient

The regularization weights control gradients into the shared network. They do not scale the optimization of the readout heads. We implement this using

$$\mathrm{GS}_w(h) = wh + (1 - w)\,\mathrm{sg}(h). \tag{18}$$

The forward value equals h, while its derivative with respect to h is w. Each task evaluates its head on the corresponding gradient-scaled representation. Its head therefore receives the full gradient even when $w = 0$. The mean/variance stop-gradient rule in Equation equation 10 is applied at the head outputs.

Let $C(\theta, \psi) = |\mathcal{Q}|^{-1} \sum_{q \in \mathcal{Q}} (y_q^\star - \widehat{\mu}_q)^2$ be the root squared error on a held-out meta batch. For a training batch, we form the weighted shared-network gradient $G_\theta(\mathbf{w})$, including the mean/variance masks. Let Δ_{Adam} denote the update using the current optimizer state, weight decay, and global gradient clipping. The one-step look-ahead is

$$\begin{aligned} \theta'(\mathbf{w}) &= \theta - \Delta_{\mathrm{Adam}}(G_\theta(\mathbf{w})), \\ F(\mathbf{w}) &= C(\theta'(\mathbf{w}), \psi), \\ \mathbf{w} &\leftarrow [\mathbf{w} - \eta_w \nabla_{\mathbf{w}} F(\mathbf{w})]_+, \qquad w_{L-1,R} \leftarrow 1. \end{aligned} \tag{19}$$

The derivative includes the dependence of clipping and the Adam update on $\mathbf{w}$, with optimizer history held constant. The look-ahead updates the shared network while holding head parameters at their current values. It differentiates the one-step objective rather than the converged training solution. We initialize the free weights at 0.03 and update them every two network steps with $\eta_w = 0.1$.

Table 4 reports the weights averaged over three weight-learning runs. The six positions consist of the tokenizer output, four Transformer blocks, and the final self-attention readout. Thus, $L = 6$ counts supervised positions, not Transformer blocks.

The final scenario-target weight is the largest learned coefficient. Scenario-prevalence weights are largest in the later Transformer blocks and smaller at the final readout. Root reconstruction also receives more weight in deeper intermediate positions than at the encoder or first block. These are learned coefficients, not estimates of each task's fraction of the total gradient magnitude.

F Prompt templates

Templates are reproduced verbatim; placeholders in braces are filled with the Known facts c, the statement Y (written $\{\texttt{B}\}$) and, where applicable, the scenario d_i (written $\{\texttt{X}\}$). Long few-shot blocks are elided in the middle, and the final turn is always shown in full.

F.1 SCENARIO GENERATION AND PROBABILITY ELICITATION

Scenario generation.

```
[system]
You will be given a situation: a target event and some facts that you
    know. You need to prepare to act appropriately in this situation,
    but you don't know the details. Describe one plausible way to
    flesh out the situation that would make the target event more
    likely, but not certain. You may describe things that could cause
    the target event, things that could result from it, or things that
    tend to go along with it for some other reason, such as a common
    cause. The goal is not to entertain the reader but to consider
    realistic possibilities so that you can plan, so the details
    should be representative and fit the given facts. Do not restate
    the facts you were given, and do not include anything that makes
    the target event certain. For instance, if the target event is
    'the person is married', do not write 'the person has a spouse',
    but you may write 'the person has several siblings and really
    loves kids'. Reply with one or two sentences of new details only.
    Here are some example outputs that are considered appropriate and
    plausible:

[user]
The target event: the respondent has a valid passport. You also know
    that the respondent lives in a small rural town.

[assistant]
The respondent's daughter is marrying someone from Ireland next year,
    and the family has started talking about traveling to the wedding.

[user]
The target event: taking the stairs would get someone to the tenth
    floor sooner than taking the elevator. You also know that the
    speed of getting from the lobby to the tenth floor is being
    compared between taking the elevator and taking the stairs. It is
    8:55 on a weekday morning.

[assistant]
Dozens of office workers are lined up for the building's only
    elevator, which stops on almost every floor, while the stairwell
    door is right beside the lobby entrance.

[user]
The target event: the respondent's answer to the question "How often
    do you go fishing?" (answer options: "Several times a year",
    "About once a year", "Less often", "Never") would be "Several
    times a year". You also know that the respondent lives in the
    Northeast.

[...]
(3 further few-shot turns elided)

[user]
The target event: {B}. You also know that {A}
```

Target probability, $\mathbb{E}[Y \mid c]$ and $\mathbb{E}[Y \mid c, d_i]$.

```
[system]
You estimate how likely a statement is to be true, given some facts
    that you know. The facts may describe an adult who took part in a
    survey in the United States around 2022, or an everyday situation.
    Read the statement carefully, including any negation, and take
    every given fact into account. When the facts make the statement
```

```
    nearly certain or nearly impossible, give a value close to 1 or 0.
    Return a single decimal probability between 0 and 1 and nothing
    else. Here are some example estimates that are considered accurate
    and proper:

[user]
You know that the speed of getting from the lobby to the tenth floor
    is being compared between taking the elevator and taking the
    stairs. The elevator is out of service for maintenance. What is
    the likelihood that taking the stairs would get someone to the
    tenth floor sooner than taking the elevator? Answer:

[assistant]
0.9650

[user]
You know that the respondent lives in a small rural town. What is the
    likelihood that the respondent has a valid passport? Answer:

[assistant]
0.3310

[user]
You know that the comfort of sleeping outdoors on a summer night is
    being compared between a tent and a hammock. The campsite's only
    flat spot is hard, uneven ground. What is the likelihood that
    sleeping in the hammock would be more comfortable than sleeping in
    the tent? Answer:

[...]
(7 further few-shot turns elided)

[user]
You know that {A}. What is the likelihood that {B_fact}? Answer:
```

Scenario prevalence, $p(d_i \mid c)$.

```
[system]
You estimate how likely a scenario is, given some circumstances that
    you know. The circumstances may describe an adult who took part in
    a survey or an everyday situation. The scenario adds details;
    judge how likely it is that all of them hold. A more detailed
    scenario is by its nature less likely than a more general one: for
    example, a person being a 28-year-old man is less likely than a
    person being a man. Return a single decimal probability between 0
    and 1 and nothing else. Here are some example estimates that are
    considered accurate and proper:

[user]
Circumstance: the speed of getting from the lobby to the tenth floor
    is being compared between taking the elevator and taking the
    stairs. It is 8:55 on a weekday morning. Scenario: Dozens of
    office workers are lined up for the building's only elevator,
    which stops on almost every floor, while the stairwell door is
    right beside the lobby entrance. Likelihood of the scenario given
    the circumstance:

[assistant]
0.0850

[user]
Circumstance: the respondent is 25 years old, and the respondent lives
    in Manhattan. Scenario: The respondent rents an apartment with at
    least one roommate. Likelihood of the scenario given the
```

```
    circumstance:

[assistant]
0.5230

[user]
Circumstance: the respondent lives in the Northeast. Scenario: The
    respondent lives in a small town on the coast of Maine and keeps a
    motorboat at a mooring near the harbor. Likelihood of the scenario
    given the circumstance:

[...]
(5 further few-shot turns elided)

[user]
Circumstance: {A}. Scenario: {X}. Likelihood of the scenario given the
    circumstance:
```

F.2 OVERLAPPING-SCENARIO GENERATION

Scenario generation.

```
[system]
You will be given a survey question about a respondent, its possible
    answers, and one fact already known about the respondent. Write 10
    short scenarios about this respondent.

Step 1 -- choose ONE new core situation. The core must:
- be a NEW fact that goes beyond the known fact. It must not restate,
    elaborate, or describe a cause or consequence of the known fact;
- be clearly relevant to how the respondent would answer the question
    (it may make either answer more likely);
- leave the answer genuinely uncertain: it should shift the odds, not
    settle the question. A fact that makes one answer true or false by
    itself is NOT acceptable;
- be plausible for someone who fits the known fact;
- if cores already used for this question are listed, be substantively
    different from every one of them: a different aspect of the
    respondent's life, not a rewording, narrower version, or close
    variant of an earlier core.

Step 2 -- write all 10 scenarios as heavy variations of that single
    core: restate it in different words, add or drop one minor detail,
    or describe a close cause or consequence of the CORE. The 10
    scenarios must OVERLAP HEAVILY: someone who has read any one of
    them should learn little new from the other nine. Every scenario
    must also leave the answer uncertain.

Bad examples:
- Known fact: the respondent's parents are divorced. "the respondent
    grew up alternating between two households" only elaborates the
    known fact.
- Question: Does the respondent live with their parents? "the
    respondent rents an apartment with unrelated roommates" settles
    the answer.

Rules:
- Each scenario is one sentence stating facts about "the respondent",
    consistent with the known fact.
- Never restate or elaborate the known fact.
- Do not state or directly imply the answer to the question.
- Output only a JSON object of the form {"core": "<the core in one
    sentence>", "scenarios": [<10 strings>]}, nothing else.
```

```
[user]
Question: Is the respondent currently married? [yes; no]
Known fact: the respondent is 45 years old.

[assistant]
{"core": "the respondent lives with a long-term partner and two
    school-age children", "scenarios": ["the respondent lives in a
    suburban house with a long-term partner and two school-age
    children", "the respondent shares a home in the suburbs with their
    partner and their two kids", "the respondent and their partner of
    fifteen years are raising two children together", "the respondent
    has two children and lives with the children's other parent", "the
    respondent's household consists of a long-term partner and two
    school-age children", "the respondent moved to the suburbs with
    their partner to raise their two children", "the respondent co-
    owns a suburban home with a long-term partner", "the respondent
    spends most weekends on family activities with their partner and
    two kids", "the respondent and a long-term partner share childcare
    duties for their two children", "the respondent has lived with a
    partner and two children in the same suburban home for a decade"]}

[user]
Question: Is the respondent currently married? [yes; no]
Known fact: the respondent is 45 years old.
Cores already used for this question -- choose a clearly different
    one:
- the respondent lives with a long-term partner and two school-age
    children

[assistant]
{"core": "the respondent works as a traveling consultant and is away
    from home most weeks", "scenarios": ["the respondent works as a
    management consultant who travels to client sites most weeks",
    "the respondent spends four nights a week in hotels for their
    consulting job", "the respondent's consulting work keeps them on
    the road for most of the month", "the respondent is a traveling
    consultant who is rarely home on weeknights", "the respondent
    flies to a different city for work nearly every Monday", "the
    respondent has held a travel-heavy consulting job for over a
    decade", "the respondent's job as a consultant requires frequent
    overnight trips", "the respondent is usually home only on weekends
    because of consulting travel", "the respondent has accumulated
    top-tier airline status from constant work travel", "the
    respondent works long weeks away from home as an itinerant
    consultant"]}

[user]
Question: Does the respondent live with their parents? [yes; no]
Known fact: the respondent is 24 years old.

[...]
(5 further few-shot turns elided)

[user]
Question: {B_question} [{domain}]
Known fact: {A}{previous_cores}
```

F.3 LLM ENSEMBLER

Ensembler, root estimate shown.

```
[system]
```

```
You combine several probability estimates into one final estimate. You
    are given a statement, some facts that are known to be true, and a
    set of scenarios that describe different ways the situation might
    look in more detail. For each scenario you are given how likely
    that scenario is given the known facts, and how likely the
    statement would be if that scenario held. The scenarios were
    written automatically: they need not cover every possibility, they
    may overlap or restate the same consideration several times, and
    they need not be representative of the known facts. Weigh them
    accordingly rather than treating them as a complete list of
    separate cases. Return a single decimal probability between 0 and
    1 that the statement is true given only the known facts, and
    nothing else. Here are some example estimates that are considered
    accurate and proper:

[user]
Known facts: the respondent is 25 years old, and the respondent lives
    in Manhattan.
Statement: the respondent does not own a car.
Estimate from the known facts alone: 0.7960
Final estimate of the likelihood that the respondent does not own a
    car:

[assistant]
0.7960

[user]
Known facts: the warmth of a cabin overnight is being compared between
    keeping the wood stove burning and letting it go out. It is
    snowing heavily outside.
Statement: the cabin would stay warmer overnight if the wood stove
    were allowed to go out.
Estimate from the known facts alone: 0.0270
Scenarios:
1. (likelihood 0.0120) The stove's chimney is partly blocked, so smoke
    leaks into the cabin whenever the fire is burning and the door has
    to be propped open to clear the air.
   Estimate of the statement if this scenario held: 0.7300
2. (likelihood 0.0090) The chimney is obstructed and the cabin door
    must be kept open for air whenever the stove is lit.
   Estimate of the statement if this scenario held: 0.7100
3. (likelihood 0.0080) Soot has built up in the flue, forcing the
    occupants to open the door to vent smoke while the stove burns.
   Estimate of the statement if this scenario held: 0.7200
Final estimate of the likelihood that the cabin would stay warmer
    overnight if the wood stove were allowed to go out:

[assistant]
0.0380

[user]
Known facts: the respondent lives in the Northeast.
Statement: the respondent's answer to the question "How often do you
    go fishing?" (answer options: "Several times a year", "About once
    a year", "Less often", "Never") would be "Several times a year".
Estimate from the known facts alone: 0.0910
Scenarios:
1. (likelihood 0.0250) The respondent lives in a dense urban
    neighborhood and does not own a car.
   Estimate of the statement if this scenario held: 0.0300
2. (likelihood 0.0060) The respondent lives in a small town on the
    coast of Maine and keeps a motorboat at a mooring near the harbor.
   Estimate of the statement if this scenario held: 0.7400
3. (likelihood 0.0180) The respondent works long hours on a weekday
    schedule and spends weekends catching up on housework.
```

```
   Estimate of the statement if this scenario held: 0.0400
4. (likelihood 0.0040) The respondent spends most weekends at a family
    camp on a lake in New Hampshire.
   Estimate of the statement if this scenario held: 0.6100
5. (likelihood 0.0320) The respondent has school-age children and
    plans family outings around their sports schedule.
   Estimate of the statement if this scenario held: 0.0700
6. (likelihood 0.0090) The respondent grew up fishing with a parent
    and still keeps the same rods in the garage.
   Estimate of the statement if this scenario held: 0.4800
7. (likelihood 0.0210) The respondent lives in an apartment building
    with no storage for outdoor equipment.
   Estimate of the statement if this scenario held: 0.0350
8. (likelihood 0.0150) The respondent is retired and has free time on
    weekday mornings.
   Estimate of the statement if this scenario held: 0.2100
9. (likelihood 0.0070) The respondent belongs to a local rod and gun
    club.
   Estimate of the statement if this scenario held: 0.8200
10. (likelihood 0.0280) The respondent's main hobby is following
    professional sports on television.
   Estimate of the statement if this scenario held: 0.0500
Final estimate of the likelihood that the respondent's answer to the
    question "How often do you go fishing?" (answer options: "Several
    times a year", "About once a year", "Less often", "Never") would
    be "Several times a year":

[...]
(3 further few-shot turns elided)

[user]
Known facts: {A}.
Statement: {B_fact}.
Estimate from the known facts alone: {ROOT}
{SCENARIO_BLOCK}Final estimate of the likelihood that {B_fact}:
```

Ensembler, scenarios only.

```
[system]
You combine several probability estimates into one final estimate. You
    are given a statement, some facts that are known to be true, and a
    set of scenarios that describe different ways the situation might
    look in more detail. For each scenario you are given how likely
    that scenario is given the known facts, and how likely the
    statement would be if that scenario held. The scenarios were
    written automatically: they need not cover every possibility, they
    may overlap or restate the same consideration several times, and
    they need not be representative of the known facts. Weigh them
    accordingly rather than treating them as a complete list of
    separate cases. Return a single decimal probability between 0 and
    1 that the statement is true given only the known facts, and
    nothing else. Here are some example estimates that are considered
    accurate and proper:

[user]
Known facts: the warmth of a cabin overnight is being compared between
    keeping the wood stove burning and letting it go out. It is
    snowing heavily outside.
Statement: the cabin would stay warmer overnight if the wood stove
    were allowed to go out.
Scenarios:
1. (likelihood 0.0120) The stove's chimney is partly blocked, so smoke
    leaks into the cabin whenever the fire is burning and the door has
    to be propped open to clear the air.
```

```
   Estimate of the statement if this scenario held: 0.7300
2. (likelihood 0.0090) The chimney is obstructed and the cabin door
    must be kept open for air whenever the stove is lit.
   Estimate of the statement if this scenario held: 0.7100
3. (likelihood 0.0080) Soot has built up in the flue, forcing the
    occupants to open the door to vent smoke while the stove burns.
   Estimate of the statement if this scenario held: 0.7200
Final estimate of the likelihood that the cabin would stay warmer
    overnight if the wood stove were allowed to go out:

[assistant]
0.1000

[user]
Known facts: the respondent lives in the Northeast.
Statement: the respondent's answer to the question "How often do you
    go fishing?" (answer options: "Several times a year", "About once
    a year", "Less often", "Never") would be "Several times a year".
Scenarios:
1. (likelihood 0.0250) The respondent lives in a dense urban
    neighborhood and does not own a car.
   Estimate of the statement if this scenario held: 0.0300
2. (likelihood 0.0060) The respondent lives in a small town on the
    coast of Maine and keeps a motorboat at a mooring near the harbor.
   Estimate of the statement if this scenario held: 0.7400
3. (likelihood 0.0180) The respondent works long hours on a weekday
    schedule and spends weekends catching up on housework.
   Estimate of the statement if this scenario held: 0.0400
4. (likelihood 0.0040) The respondent spends most weekends at a family
    camp on a lake in New Hampshire.
   Estimate of the statement if this scenario held: 0.6100
5. (likelihood 0.0320) The respondent has school-age children and
    plans family outings around their sports schedule.
   Estimate of the statement if this scenario held: 0.0700
6. (likelihood 0.0090) The respondent grew up fishing with a parent
    and still keeps the same rods in the garage.
   Estimate of the statement if this scenario held: 0.4800
7. (likelihood 0.0210) The respondent lives in an apartment building
    with no storage for outdoor equipment.
   Estimate of the statement if this scenario held: 0.0350
8. (likelihood 0.0150) The respondent is retired and has free time on
    weekday mornings.
   Estimate of the statement if this scenario held: 0.2100
9. (likelihood 0.0070) The respondent belongs to a local rod and gun
    club.
   Estimate of the statement if this scenario held: 0.8200
10. (likelihood 0.0280) The respondent's main hobby is following
    professional sports on television.
   Estimate of the statement if this scenario held: 0.0500
Final estimate of the likelihood that the respondent's answer to the
    question "How often do you go fishing?" (answer options: "Several
    times a year", "About once a year", "Less often", "Never") would
    be "Several times a year":

[assistant]
0.1300

[user]
Known facts: the time it takes to get across the city at rush hour is
    being compared between riding a bicycle and driving a car.
Statement: riding a bicycle would get someone across the city sooner
    than driving a car.
Scenarios:
1. (likelihood 0.1800) Traffic is backed up on every main artery
    through downtown.
```

```
   Estimate of the statement if this scenario held: 0.8600
2. (likelihood 0.1200) The route has a continuous protected bike lane
    from end to end.
   Estimate of the statement if this scenario held: 0.8200
3. (likelihood 0.0900) Heavy rain is falling and the streets are
    slick.
   Estimate of the statement if this scenario held: 0.2400
4. (likelihood 0.1500) The city centre has a congestion charge and few
    parking spaces.
   Estimate of the statement if this scenario held: 0.7900
5. (likelihood 0.0700) The trip is more than fifteen miles each way.
   Estimate of the statement if this scenario held: 0.1500
6. (likelihood 0.1100) Several streets are closed for construction,
    forcing detours for cars.
   Estimate of the statement if this scenario held: 0.8400
7. (likelihood 0.0600) The cyclist is carrying a large box of
    materials.
   Estimate of the statement if this scenario held: 0.1800
8. (likelihood 0.1000) The route is mostly flat with few stops.
   Estimate of the statement if this scenario held: 0.7100
9. (likelihood 0.0800) There is a steep hill halfway along the route.
   Estimate of the statement if this scenario held: 0.3100
10. (likelihood 0.1300) Parking near the destination typically takes
    ten minutes to find.
   Estimate of the statement if this scenario held: 0.7700
11. (likelihood 0.0500) It is below freezing and there is ice on the
    road.
   Estimate of the statement if this scenario held: 0.1200
12. (likelihood 0.0900) The driver has access to a highway that
    bypasses the city centre.
   Estimate of the statement if this scenario held: 0.1700
13. (likelihood 0.1400) Traffic lights along the route are timed for
    cars moving at speed.
   Estimate of the statement if this scenario held: 0.3300
14. (likelihood 0.0400) The cyclist is an experienced commuter who
    rides this route daily.
   Estimate of the statement if this scenario held: 0.6800
15. (likelihood 0.0600) The car is a rideshare that must first travel
    to the pickup point.
   Estimate of the statement if this scenario held: 0.7400
16. (likelihood 0.1000) The bicycle can use a riverside path that cars
    cannot access.
   Estimate of the statement if this scenario held: 0.8800
17. (likelihood 0.0700) The cyclist would need to change clothes on
    arrival.
   Estimate of the statement if this scenario held: 0.4000
18. (likelihood 0.0500) A major sporting event has just let out
    nearby.
   Estimate of the statement if this scenario held: 0.8900
19. (likelihood 0.1200) The trip is about three miles.
   Estimate of the statement if this scenario held: 0.7600
20. (likelihood 0.0300) The bicycle has a flat tyre that needs fixing
    first.
   Estimate of the statement if this scenario held: 0.0800
21. (likelihood 0.0800) The driver knows several side streets that
    avoid the worst junctions.
   Estimate of the statement if this scenario held: 0.2900
22. (likelihood 0.0600) Strong headwinds are blowing along the whole
    route.
   Estimate of the statement if this scenario held: 0.2200
23. (likelihood 0.1100) The destination is inside a pedestrian zone
    where cars must stop at the edge.
   Estimate of the statement if this scenario held: 0.8500
24. (likelihood 0.0400) The car is already parked on the street
    outside.
```

```
   Estimate of the statement if this scenario held: 0.3400
25. (likelihood 0.0900) The rush-hour peak has just passed and traffic
    is beginning to thin.
   Estimate of the statement if this scenario held: 0.4300
26. (likelihood 0.0500) The cyclist is not confident riding in mixed
    traffic and takes a longer quiet route.
   Estimate of the statement if this scenario held: 0.3600
27. (likelihood 0.0700) There is a bike-share station at both ends of
    the trip.
   Estimate of the statement if this scenario held: 0.6900
28. (likelihood 0.0300) The road has a dedicated bus and bike lane
    that cars may not use.
   Estimate of the statement if this scenario held: 0.8700
29. (likelihood 0.1000) The driver would have to drop off a passenger
    on the way.
   Estimate of the statement if this scenario held: 0.6600
30. (likelihood 0.0600) An accident has closed one lane of the main
    bridge on the route.
   Estimate of the statement if this scenario held: 0.8300
Final estimate of the likelihood that riding a bicycle would get
    someone across the city sooner than driving a car:

[assistant]
0.5600

[user]
Known facts: {A}.
Statement: {B_fact}.
{SCENARIO_BLOCK}Final estimate of the likelihood that {B_fact}:
```

F.4 ATMTO

```
### Question: Given the premise "{premise}", how likely is it that the
hypothesis "{hypothesis}" is true?

### Answer: <|label_level_K|>
```

G ADVERSARIAL SCENARIO GENERATION

Overlapping scenarios. For each of 997 GSS queries, GPT-4.1 generates three distinct core situations and ten closely related variations of each, yielding 30 scenarios. Variations of the same core situation are grouped together in the scenario sequence. At each scenario count k, all methods receive the same first k scenarios. The first ten scenarios therefore describe variations of a single core situation, with the second and third core situations introduced at $k = 11$ and $k = 21$, respectively. This construction adds heavily overlapping scenarios within each group before introducing a different core situation.

Biased proposals. For each of 860 GSS queries with original condition A, we provide an additional survey fact H only to the scenario writer. The additional fact biases which scenarios are proposed, while the target remains $\mathbb{E}[Y \mid c]$. The probability elicitation prompts retain the original context A and the generated scenario; H is not supplied separately. Across queries, the mean absolute difference $|\,\mathbb{E}[Y \mid c, H] - \mathbb{E}[Y \mid c]|$ is 0.141, indicating how much the reference probability would change if H were included in the known facts. As in the overlap test, all methods receive the same scenario prefixes.

H LAOS INFERENCE EXAMPLES

Here we give five example questions processed by LAOS.

Example 1 (GSS). *Known facts c:* the respondent is not religious at all.
Statement Y: the respondent received a flu shot in the past year

reference $\mathbb{E}[Y \mid c]$	0.571
LLM direct answer	0.385
prevalence-weighted average	0.858
LAOS	**0.635**

The six most prevalent of the thirty scenarios:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent has a young child who requires annually documented flu vaccinations for daycare attendance.	0.285	0.725
The respondent works for a large corporation that mandates annual immunizations, including the flu vaccine, as a condition of employment.	0.150	0.925
The respondent works as a high school teacher, and the school district mandated that all staff members receive a flu vaccine to retain their positions.	0.125	0.925
The respondent works as a school teacher and was required to provide proof of vaccination to the district nurse this past fall.	0.065	0.650
The respondent works as a nurse at a local hospital, where annual vaccination records are strictly monitored as a condition of employment.	0.045	0.850
The respondent works as an elementary school teacher and was required to show proof of vaccination by their school district during the spring health check.	0.045	0.650

Example 2 (GSS). *Known facts c:* the respondent think that in certain situations they would approve of a policeman striking an adult male citizen.
Statement Y: the respondent believes that a Muslim clergyman who preaches hatred of the United States should be allowed to speak in their community

reference $\mathbb{E}[Y \mid c]$	0.519
LLM direct answer	0.421
prevalence-weighted average	0.759
LAOS	**0.558**

The six most prevalent of the thirty scenarios:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent holds a strong commitment to free speech principles and believes that all viewpoints, even offensive ones, should be protected unless they incite immediate violence.	0.450	0.725
The respondent feels that the free speech protections in the Constitution are more important than their personal comfort or political views.	0.450	0.725
The respondent believes that allowing all lawful speech, even when deeply offensive, is essential to preserving constitutional rights and preventing government overreach.	0.450	0.725
The respondent strongly values free speech and believes that counter-speech and debate are the only proper ways to combat hateful ideas, regardless of their content.	0.350	0.825
The respondent places a very high value on free speech principles, believing that society is better served by publicly confronting and debating hateful ideas rather than suppressing them.	0.350	0.725
The respondent places a very high value on free speech principles and tends to support controversial speakers, even when their views are offensive or extreme.	0.150	0.785

Example 3 (GSS). *Known facts c:* the respondent believes that someone who believes that Blacks are genetically inferior can be allowed to make a speech in the respondent's community claiming that Blacks are inferior.
Statement Y: the respondent think that in certain situations they would approve of a policeman

striking an adult male citizen

reference $\mathbb{E}[Y \mid c]$	0.737
LLM direct answer	**0.725**
prevalence-weighted average	0.633
LAOS	0.765

The six most prevalent of the thirty scenarios:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent holds a rigid view on free speech and believes that any form of state censorship or physical intervention by authorities violates individual liberties, even in the face of hate speech or violent…	0.650	0.050
The respondent holds a strong view that public order and the ability of police to enforce authority are paramount, even if it requires the use of physical force against non-compliant adults.	0.650	0.850
The respondent holds a generally lenient view of authority and tends to prioritize social order over individual rights in high-tension situations.	0.650	0.825
The respondent feels that strong, immediate physical intervention is justified to prevent a serious crime, even if it results in injury to the perpetrator.	0.650	0.850
The respondent strongly prioritizes public safety and order, believing that police officers should have broad discretion to use physical force to subdue violent suspects or restore calm during chaotic…	0.650	0.854
The respondent holds a strong laissez-faire attitude toward civil liberties and believes that authority figures, including police, should have broad discretion to use physical force in the field.	0.650	0.850

Example 4 (COM2SENSE). *Known facts c:* the convenience of obtaining a medicinal permit is being compared to the convenience of national relocation. Access to required medication is widely available in all states in the country.
Statement Y: national relocation is more convenient than a medicinal permit

reference $\mathbb{E}[Y \mid c]$	0.000
LLM direct answer	**0.012**
prevalence-weighted average	0.798
LAOS	0.145

The six most prevalent of the thirty scenarios:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
Filling out the application for the medicinal permit involves a lengthy, multi-step bureaucratic process, whereas national relocation entails simply packing up and moving without any regulatory hurdles.	0.050	0.985
Movement across state lines requires no special documentation, whereas obtaining a medicinal permit involves a lengthy application process and strict verification.	0.050	0.985
The respondent has a flexible remote job and no significant family ties in their current state, making the move logistically simple.	0.035	0.050
The medicinal permit requires submitting blood work and attending an in-person interview with a specialist, whereas state change of address forms can be filed online in a few minutes.	0.035	0.985
The applicant currently lives in a state with strict residency verification laws that require repetitive in-person visits to government offices, whereas moving to a neighboring state with streamlined digital…	0.035	0.850
The medicinal permit application requires notarized medical records and a six-week postal review process, whereas national relocation involves simply buying a ticket and moving, with no bureaucratic wait times…	0.035	0.985

Example 5 (COM2SENSE). *Known facts c:* the comfort level of sitting on a tile floor versus a carpet floor is being compared. The carpet floor is plush and of high quality, providing more

cushioning and warmth than a tile floor.
Statement Y: it's more comfortable to sit on a tile floor rather than a carpet floor

reference $\mathbb{E}[Y \mid c]$	0.000
LLM direct answer	**0.050**
prevalence-weighted average	0.743
LAoS	0.234

The six most prevalent of the thirty scenarios:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
It is the middle of summer and the room is stuffy because the air conditioning has broken down.	0.150	0.650
The carpet is thick with long fibers that trap dust mites and trigger severe allergic reactions, while the smooth tile can be easily wiped clean to minimize irritants.	0.050	0.150
It is the height of summer and the room is extremely hot and humid, while the tile floor has been cooled by air conditioning blowing directly onto it.	0.045	0.725
The carpet has not been vacuumed in months and is covered in a thick layer of pet hair, dog fur, and fine dust that irritates the skin, while the tile was recently wiped down with disinfectant.	0.035	0.852
The room is extremely hot and humid, causing the carpet to feel sticky and trap sweat against the skin, while the tiled surface remains cool and dry.	0.035	0.850
The tile floor has been thoroughly cleaned and is dust-free, whereas the deep-pile carpet is filled with accumulated dust, pet dander, and hidden insect debris that would irritate the skin.	0.035	0.650

I Inference examples with LAoS under controlled stress tests

§3.5 describes both stress tests.

I.1 Overlapping scenarios

Each question's thirty scenarios form three coupled sets of ten, each set written as small variations of one core situation.

Overlap example 1. *Known facts c:* The respondent's situation is that their parents are divorced or separated.
Question: Does the respondent believe a person has the right to end their own life if they have dishonored their family?
Core situation of the set shown: the respondent attends religious services nearly every week

reference $\mathbb{E}[Y \mid c]$	0.842
LLM direct answer	0.421
prevalence-weighted average	0.612
LAoS	**0.874**
sum of the ten prevalences below	2.22

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
the respondent attends church services nearly every week	0.385	0.785
the respondent goes to worship services at their local church almost every Sunday	0.185	0.785
the respondent rarely misses a weekly religious service	0.185	0.785
the respondent is a regular churchgoer who attends services about once a week	0.215	0.785
the respondent attends weekly services and belongs to a small Bible study group at their church	0.185	0.785
the respondent has attended the same congregation almost every weekend for years	0.185	0.725
the respondent considers weekly church attendance an important part of their routine	0.185	0.785
the respondent goes to religious services about four times a month	0.125	0.725
the respondent attends Sunday services regularly and volunteers at church events	0.185	0.785
the respondent is an active member of a congregation and attends services most weeks	0.385	0.785

Overlap example 2. *Known facts c:* The respondent finds it neither easy nor difficult for their household to make ends meet.
Question: Is the respondent a child of another adult who lives in the same household with them?
Core situation of the set shown: the respondent is a married adult in their forties raising children

reference $\mathbb{E}[Y \mid c]$	0.605
LLM direct answer	0.920
prevalence-weighted average	0.957
LAOS	**0.701**
sum of the ten prevalences below	1.65

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
the respondent is a married adult in their forties raising school-age children	0.185	0.990
the respondent is 44, married, and has two children in middle school	0.085	0.990
the respondent and their spouse are in their forties and raising three kids	0.085	0.990
the respondent is a parent in their forties who has been married for fifteen years	0.185	0.990
the respondent is a married parent in their early forties with young children	0.185	0.990
the respondent is in their mid-forties and spends most evenings helping their kids with homework alongside their spouse	0.085	0.990
the respondent is a 46-year-old married parent of two teenagers	0.185	0.990
the respondent is in their forties, married, and busy raising a family	0.185	0.980
the respondent and their spouse, both in their forties, have children in elementary school	0.285	0.980
the respondent is a married adult in their forties whose children are still in school	0.185	0.990

Overlap example 3. *Known facts c:* The respondent is in favor of sex education in public schools.
Question: Does the respondent support removing a book from the public library if it claims that Black people are genetically inferior?
Core situation of the set shown: the respondent is a strong believer in free speech who has donated to a civil liberties organization

reference $\mathbb{E}[Y \mid c]$	0.597
LLM direct answer	0.925
prevalence-weighted average	0.967
LAOS	**0.877**
sum of the ten prevalences below	3.90

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
the respondent strongly values free speech and has donated to a civil liberties organization	0.285	0.985
the respondent is a longtime member of a civil liberties group that defends freedom of expression	0.285	0.925
the respondent considers freedom of speech one of the most important constitutional rights	0.680	0.925
the respondent gives money each year to an organization that fights government censorship	0.285	0.985
the respondent closely follows court cases about freedom of expression and supports civil liberties groups	0.412	0.925
the respondent describes themselves as a free speech advocate and supports civil liberties causes	0.685	0.985
the respondent has contributed to a nonprofit that defends First Amendment rights	0.285	0.925
the respondent believes strongly in open public debate and has supported civil liberties groups financially	0.285	0.985
the respondent is passionate about protecting free expression and belongs to a civil liberties organization	0.285	0.925
the respondent regularly follows and supports the work of groups that defend free speech rights	0.412	0.985

Overlap example 4. *Known facts c:* The respondent strongly believes that it is the government's responsibility to pay individuals' medical bills.
Question: Does the respondent believe that the differences in jobs, income, and housing between Black people and white people are due to Black people not having the educational opportunities needed to rise out of poverty?
Core situation of the set shown: the respondent attended an underfunded public high school in a low-income neighborhood

reference $\mathbb{E}[Y \mid c]$	0.608
LLM direct answer	0.785
prevalence-weighted average	0.575
LAOS	**0.591**
sum of the ten prevalences below	3.56

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
the respondent attended an underfunded public high school in a low-income neighborhood	0.385	0.785
the respondent went to a public high school that had outdated textbooks and overcrowded classrooms	0.412	0.785
the respondent grew up in a poor neighborhood where the local schools lacked basic resources	0.412	0.785
the respondent's high school struggled with low funding and a shortage of experienced teachers	0.412	0.785
the respondent was educated in a public school district with some of the lowest per-student spending in the state	0.285	0.785
the respondent attended high school in a low-income area where few classmates went on to college	0.285	0.725
the respondent went to an under-resourced public school that had trouble keeping teachers	0.412	0.785
the respondent grew up attending schools in a struggling district with limited course offerings	0.385	0.785
the respondent's public high school was in a poor part of town and had very little funding	0.285	0.785
the respondent remembers their high school lacking science labs and up-to-date materials because of budget shortfalls	0.285	0.785

Overlap example 5. *Known facts c:* The respondent strongly disagrees that family life suffers when the woman has a full-time job.

Question: Would the respondent provide emotional support to a close family member or friend who decided to have an abortion?
Core situation of the set shown: the respondent is a woman in her thirties with a large circle of close female friends

reference $\mathbb{E}[Y \mid c]$	0.950
LLM direct answer	**0.725**
prevalence-weighted average	0.533
LAOS	0.629
sum of the ten prevalences below	2.44

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
the respondent is a woman in her thirties who has a large circle of close female friends	0.285	0.785
the respondent is a 34-year-old woman who stays in frequent contact with many close girlfriends	0.285	0.785
the respondent is a woman in her thirties whose close friends often come to her for advice	0.215	0.785
the respondent is in her thirties and has a tight-knit group of female friends she has known since college	0.185	0.825
the respondent is a woman in her mid-thirties who talks with her closest female friends almost every day	0.285	0.785
the respondent is a thirty-something woman with a wide network of close women friends	0.285	0.785
the respondent is a woman in her thirties who regularly meets up with a group of close female friends	0.285	0.785
the respondent is a 36-year-old woman who is the person her friends confide in during hard times	0.041	0.785
the respondent is a woman in her thirties with several close female friends she considers like sisters	0.285	0.785
the respondent is a woman in her early thirties who maintains many close friendships with other women	0.285	0.785

I.2 BIASED PROPOSAL

For each question a hidden fact H is revealed only to the scenario writer. Every question the model is then scored on states A alone, and the target remains $\mathbb{E}[Y \mid c]$. The survey rate $\mathbb{E}[Y \mid c, H]$ shows the direction and size of the pull the scenarios inherit.

Biased example 1. *Known facts c (in every scored prompt):* The respondent is neutral on whether the government should be responsible for improving the standard of living of poor Americans.
Hidden fact H (scenario writer only): The respondent is strongly pro-choice regarding abortion.
Question: Would the respondent help pay for an abortion if a close family member or friend decided to have one?

reference $\mathbb{E}[Y \mid c]$	0.448
survey rate $\mathbb{E}[Y \mid c, H]$	0.776
prevalence-weighted average	0.612
single-correction LAOS	0.731
LAOS	**0.423**

Four of the scenarios the writer produced:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent lives in a high-income household with significant disposable income and has a history of financially supporting their socially liberal peers.	0.085	0.420
The respondent has a close friend who recently faced an unplanned pregnancy and expressed financial anxiety, which has made the respondent more acutely aware of the practical barriers to accessing abortion…	0.042	0.725
The respondent has a close friend who is currently experiencing an unplanned pregnancy and has expressed concern about the high costs associated with the procedure.	0.042	0.425
The respondent views reproductive autonomy as a fundamental individual right and has previously expressed willingness to provide financial support to friends in difficult personal circumstances.	0.412	0.785

Biased example 2. *Known facts c (in every scored prompt):* The respondent has never been divorced or legally separated.
Hidden fact H (scenario writer only): The respondent voted for Joe Biden in the 2020 presidential election.
Question: Does the respondent believe that a pregnant woman should be allowed a legal abortion if she is unmarried and does not want to marry the father?

reference $\mathbb{E}[Y \mid c]$	0.574
survey rate $\mathbb{E}[Y \mid c, H]$	0.862
prevalence-weighted average	0.823
single-correction LAOS	0.864
LAOS	**0.659**

Four of the scenarios the writer produced:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent lives in a major urban area with a diverse population and identifies as politically moderate to liberal, often engaging with social justice issues on social media.	0.285	0.785
The respondent is a 34-year-old woman who works in the education sector in a mid-sized city in the Northeast. She is currently in a committed relationship but has expressed concerns about the rising cost of…	0.001	0.825
The respondent identifies as a liberal, places a high value on bodily autonomy, and has expressed support for reproductive rights in previous discussions with friends and family.	0.285	0.925
The respondent identifies as a progressive Democrat, regularly attends services at a liberal mainline Protestant church in the Northeast, and has spoken favorably about reproductive rights in previous…	0.041	0.925

Biased example 3. *Known facts c (in every scored prompt):* The respondent think that in certain situations they would approve of a policeman striking an adult male citizen.
Hidden fact H (scenario writer only): The respondent believe that a Muslim clergyman who preaches hatred of the United States shpuld be allowed to teach in a college or university.
Question: Does the respondent believe that a Muslim clergyman who preaches hatred of the United States shpuld be allowed to speak in the respondent's community?

reference $\mathbb{E}[Y \mid c]$	0.519
survey rate $\mathbb{E}[Y \mid c, H]$	0.907
prevalence-weighted average	0.694
single-correction LAOS	0.759
LAOS	**0.431**

Four of the scenarios the writer produced:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent lives in a large, diverse metropolitan area in the Northeast and places a high value on academic freedom and free speech principles, often citing constitutional protections as a priority in…	0.041	0.725
The respondent identifies as a libertarian and strongly prioritizes absolute protection of free speech rights under the First Amendment, often viewing any form of government censorship or restriction on public…	0.185	0.825
The respondent identifies as holding strong libertarian views on free speech, believing that the government should rarely, if ever, restrict expression regardless of the content. They have previously expressed…	0.185	0.825
The respondent generally identifies as a strong civil libertarian who prioritizes freedom of speech and academic freedom as absolute values, rarely making exceptions even for views they find deeply offensive…	0.085	0.825

Biased example 4. *Known facts c (in every scored prompt):* The respondent is employed, and the safety of workers is often a high priority with management at work.
Hidden fact H (scenario writer only): The respondent is employed, and the respondent always get to do a number of different things on their job.
Question: The respondent is employed. Is the respondent currently part of a self-managed team?

reference $\mathbb{E}[Y \mid c]$	0.371
survey rate $\mathbb{E}[Y \mid c, H]$	0.562
prevalence-weighted average	0.564
single-correction LAOS	0.737
LAOS	**0.433**

Four of the scenarios the writer produced:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent works in a modern tech company that has recently transitioned to an agile methodology, where cross-functional groups are given autonomy over their project milestones and daily task assignments,…	0.042	0.850
The respondent works in a mid-sized manufacturing or logistics firm that has recently transitioned to a lean management philosophy to improve efficiency and workplace safety. Their role involves ongoing…	0.043	0.350
The respondent works in a tech startup where the organizational structure is flat and agile, relying heavily on cross-functional pods to drive product development.	0.042	0.725
The respondent holds a mid-level operational role in a logistics or manufacturing firm that recently implemented agility-focused restructuring to improve workflow efficiency, where job descriptions are broad…	0.041	0.425

Biased example 5. *Known facts c (in every scored prompt):* The respondent's situation is that their parents are divorced or separated.
Hidden fact H (scenario writer only): The respondent has never experienced people acting as if they are afraid of them.
Question: What is the respondent's sex?

reference $\mathbb{E}[Y \mid c]$	0.529
survey rate $\mathbb{E}[Y \mid c, H]$	0.635
prevalence-weighted average	0.722
single-correction LAOS	0.833
LAOS	**0.620**

Four of the scenarios the writer produced:

scenario d_i	$\hat{p}(d_i \mid c)$	$\widehat{\mathbb{E}}[Y \mid c, d_i]$
The respondent had to help care for their younger siblings after the parents split, often taking on a nurturing role at home.	0.185	0.550
The respondent typically enjoys shopping for clothing and spends time asking friends for relationship advice.	0.412	0.625
The respondent enjoys shopping at a beauty supply store with her female friends.	0.350	0.950
The respondent works as a nurse and frequently volunteers at community health fairs.	0.013	0.885

J ABLATIONS

We ablate six components of LAOS on GSS. All models share the setup of the GSS model in Table 1: Gaussian NLL training on GSS with 10% of target variables held out for early stopping, 30 scenarios per query, and reconstruction weights learned during each run by hypergradient (§E). Each ablated model changes one component and is trained once with the same seed. Early stopping keeps the checkpoint with the lowest held-out loss from the fifth epoch on; without this floor, the base model keeps its first-epoch checkpoint, which reaches KL 0.1224 at 64 calls. We evaluate on the 371 distinct GSS test queries. Table 5 reports all metrics at the main-text budgets. The figures give KL when a model trained with 30 scenarios receives only its first k scenarios at inference, with pointwise 95% bootstrap bands over queries (2,000 resamples). Single training runs are sensitive to checkpoint selection: moving the checkpoint of single-correction LAOS from epoch 3 to epoch 6 changes its KL at 64 calls from 0.1101 to 0.1663.

	16 LLM calls				32 LLM calls				64 LLM calls			
Model	KL	MSE	MAE	RMSE	KL	MSE	MAE	RMSE	KL	MSE	MAE	RMSE
LAOS	**0.0757**	**0.0306**	**0.1354**	**0.1748**	**0.0722**	**0.0292**	**0.1339**	**0.1708**	**0.0723**	**0.0292**	**0.1357**	**0.1710**
w/o learned aggregation	0.0994	0.0390	0.1492	0.1974	0.1129	0.0435	0.1556	0.2085	0.1309	0.0491	0.1654	0.2217
w/o learned correction	**0.0731**	**0.0287**	**0.1279**	**0.1696**	**0.0697**	**0.0281**	**0.1283**	**0.1675**	**0.0780**	**0.0316**	**0.1379**	**0.1776**
target estimates $\widehat{r}_i$=1	0.2032	0.0755	0.2118	0.2748	0.1424	0.0571	0.1823	0.2389	0.1210	0.0495	0.1726	0.2225
prevalences $\widehat{\pi}_i$=1	0.0890	0.0359	0.1465	0.1895	0.0883	0.0357	0.1462	0.1888	0.0878	0.0355	0.1459	0.1883
shuffled embeddings	0.0888	0.0335	0.1411	0.1830	0.0801	**0.0302**	**0.1350**	**0.1737**	**0.0762**	**0.0288**	**0.1328**	**0.1697**
fixed variance (σ^2=1)	**0.0690**	**0.0274**	**0.1318**	**0.1656**	**0.0760**	**0.0302**	**0.1363**	**0.1739**	**0.0809**	**0.0320**	**0.1410**	**0.1790**
70/30 mean/variance split	**0.0767**	0.0318	**0.1365**	0.1783	**0.0777**	0.0322	0.1372	0.1793	0.0835	0.0345	**0.1410**	0.1857

Table 5: Ablation studies. Bold marks the best value in a column and every value not statistically significant.

J.1 LEARNED AGGREGATION

Single-correction LAOS keeps learned correction but removes learned aggregation. It is trained on queries without scenarios, corrects the direct estimate and each scenario estimate independently, and averages the corrected probabilities with the prevalence weights, giving the direct estimate weight one. Its KL is 0.1335, 0.1481 and 0.1663 at 16, 32 and 64 calls, against 0.0910, 0.0921 and 0.0975 for LAOS ($p < 0.001$). It is lowest with no scenarios (0.1093) and rises with every added scenario (Figure 5). Scenario estimates skew above the reference, and correcting them one at a time does not remove that shared bias: the mean prediction rises from 0.75 at $k = 0$ to 0.81 at $k = 30$, against a reference mean of 0.65.

J.2 LEARNED CORRECTION

The converse ablation keeps learned aggregation but removes correction. A shared head scores every token at the final layer, and the prediction is the softmax-weighted average of the LLM's own estimates in probability space, $\widehat{y} = \sum_i w_i \widehat{r}_i$ over the direct and scenario estimates, so it cannot leave the range of those estimates. KL is 0.1173, 0.1063 and 0.0998, above LAOS at every budget but significantly only at 16 calls ($p = 0.002$). The gap narrows as scenarios are added (Figure 6), and the learned weights stay ahead of prevalence weighting of the same estimates (0.1267, 0.1227 and 0.1302).

J.3 LLM RAW TARGET PREDICTIONS

Setting every target estimate $\widehat{r}_i$, including the direct estimate, to one removes the LLM's answers while keeping prevalences and embeddings. KL rises to 0.3613, 0.1812 and 0.1223 ($p \leq 0.003$). Without scenarios the model predicts close to one (KL 1.60); KL falls with each added scenario and drops below direct prompting (0.1871) from $k = 15$ on (Figure 7). Scenario texts and prevalences alone therefore carry information about the reference, but they do not replace the estimates.

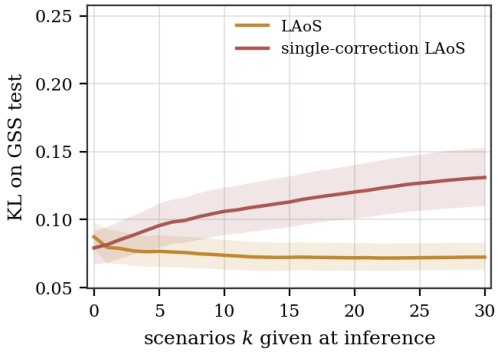


Figure 5: Learned aggregation: LAoS and single-correction LAoS.

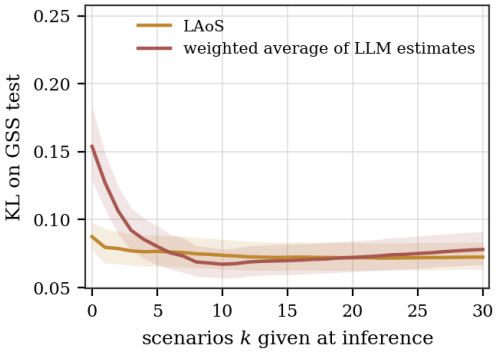


Figure 6: Learned correction: LAoS and a learned weighted average of the LLM's estimates.

J.4 LLM RAW PREVALENCE PREDICTIONS

Setting every prevalence $\widehat{\pi}_i$ to one (the query's own prevalence is already one) raises KL to 0.1054, 0.1047 and 0.1046, significantly at 16 and 32 calls ($p = 0.015$ and 0.047). From three scenarios on, KL stays between 0.1044 and 0.1054 (Figure 8).

J.5 LLM EMBEDDINGS

We pool the text embeddings of all training and test queries, 31 per query (30 scenarios and the query itself), and reassign them with one random permutation, so each query receives embeddings of unrelated queries and scenarios. Estimates and prevalences keep their positions, and both training and test use the shuffled embeddings. KL is 0.1015 at 16 calls, above LAoS ($p = 0.007$), but 0.0935 at 32 calls and 0.0895 at 64 calls, where it is below LAoS's 0.0975 ($p = 0.024$; Figure 9). On GSS, the embeddings help when few scenarios are available; with 30 scenarios, the numerical estimates suffice.

J.6 HETEROSKEDASTICITY

We compare LAoS's separated supervision with (a) a fixed variance $\sigma^2 = 1$, under which the NLL reduces to squared error, with the mean trained on the same 70% of queries, and (b) naive heteroskedastic training, in which every query supervises mean and variance jointly without stopping gradients; (b) is the GSS model of Table 1. Both alternatives reach lower KL than separated supervision at 32 and 64 calls: 0.0820 and 0.0843 with a fixed variance ($p = 0.014$ and 0.022) and 0.0855 and 0.0841 with joint training ($p = 0.084$ and 0.006), against 0.0921 and 0.0975 (Figure 10). In this setting, separating mean and variance supervision does not improve accuracy on GSS. Its mean prediction drifts upward as scenarios are added, from 0.60 at $k = 1$ to 0.74 at $k = 30$ (reference mean 0.65), consistent with its KL rising after $k = 2$.

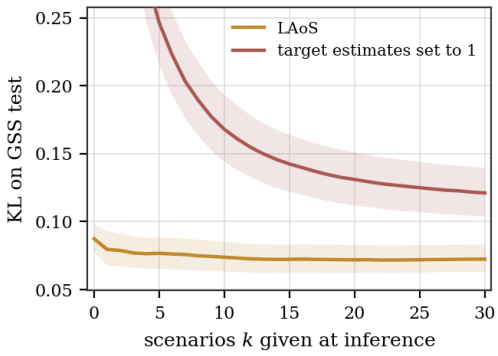


Figure 7: LLM raw target predictions: LAOS and LAOS with every target estimate set to one, which is off scale for $k \leq 10$ (KL up to 1.60).

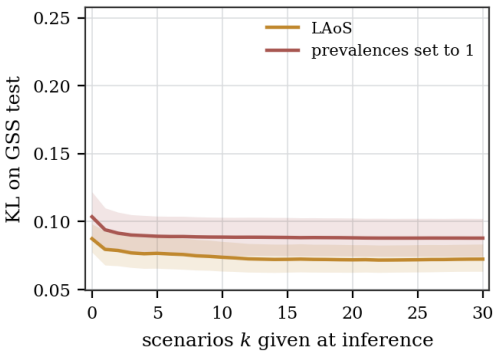


Figure 8: LLM raw prevalence predictions: LAOS and LAOS with every prevalence set to one.

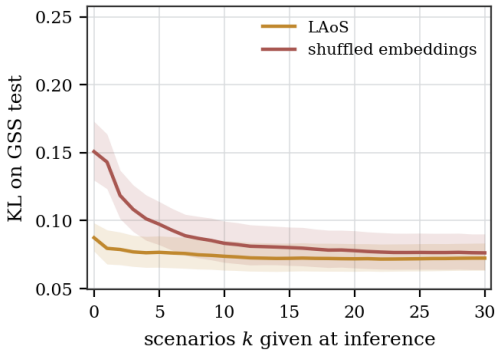


Figure 9: LLM embeddings: LAOS and LAOS trained and tested with shuffled embeddings.

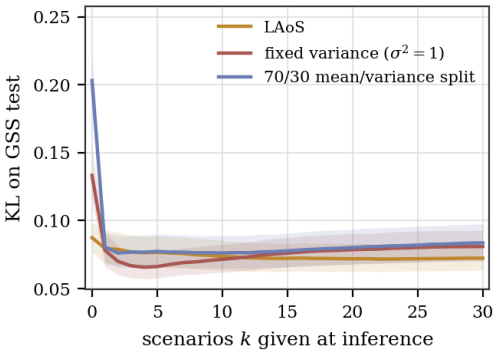


Figure 10: Heteroskedasticity: LAOS, a fixed variance, and joint mean and variance training.