EXACT POSTERIOR DISTILLATION: TEACHING LANGUAGE MODELS WHAT TO BELIEVE AND WHOM TO TRUST

Anonymous authors
Paper under double-blind review

ABSTRACT

AI agents gather evidence from sources that may be incomplete or misleading. As evidence arrives, they need to revise both their confidence in possible answers and their trust in individual sources. We introduce **BayesClue**, a synthetic environment in which models infer a hidden answer from reports by reliable, noisy, and adversarial sources. The process used to synthesize the data allows us to compute exact posterior probabilities over answers and source types after each report. Our proposed method of exact posterior distillation (EPD) trains models to predict these probabilities. Across eight models from four families, EPD reduces belief error under changes in evidence length, candidate count, and source proportions. In a ten-model comparison, all five belief-training variants achieve lower belief error than final-answer SFT. EPD also helps models switch to the answer favored by new evidence, even in sequences longer than those seen during training, and improves average performance on external evidence judgments and contextual appropriateness. These results suggest that explicit supervision of uncertainty can complement training on final answers.

1 INTRODUCTION

LLM-based AI agents typically use data from multiple sources, including retrieved documents, stored memories, and tool responses (Lewis et al., 2020; Xu et al., 2024). These sources can differ in reliability and provide incomplete, outdated, or conflicting information. Therefore, to formulate a final response, an LLM should be able to determine which conclusions the evidence supports and decide which sources to trust.

Evidence gathering is a two-way inference problem: beliefs about the answer depend on which sources the agent trusts, while trust changes as new evidence arrives. Suppose an assistant is asked whether a user attended a meeting. The calendar has no entry, but a travel record shows a trip to its location. Learning that the calendar often omits work events should weaken the evidence against attendance. A correct final answer alone does not show whether the assistant made this adjustment or trusted the right source.

Standard pre-training and post-training objectives provide limited direct supervision for reasoning about uncertain answers and unreliable sources. Pre-training teaches models to predict text without explicit supervision of source reliability. Post-training teaches models to produce correct or preferred responses (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024), with process supervision extending this training to intermediate reasoning steps (Lightman et al., 2024; Uesato et al., 2022). These objectives do not directly supervise how beliefs about answers and source reliability should change with new evidence. We ask whether explicitly teaching these updates helps models reason about conflicting evidence.

In this work, we introduce **BayesClue**, a synthetic environment in which models infer a hidden answer from sources of unknown reliability (Figure 1). The known generative process gives exact posterior probabilities over candidate answers and source types after each report. We further propose *Exact Posterior Distillation* (EPD), a training procedure that uses these marginals as targets. We show that this training helps models judge how strongly the evidence supports each answer and change their preferred answer when new information points elsewhere. Although EPD is trained only in

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.

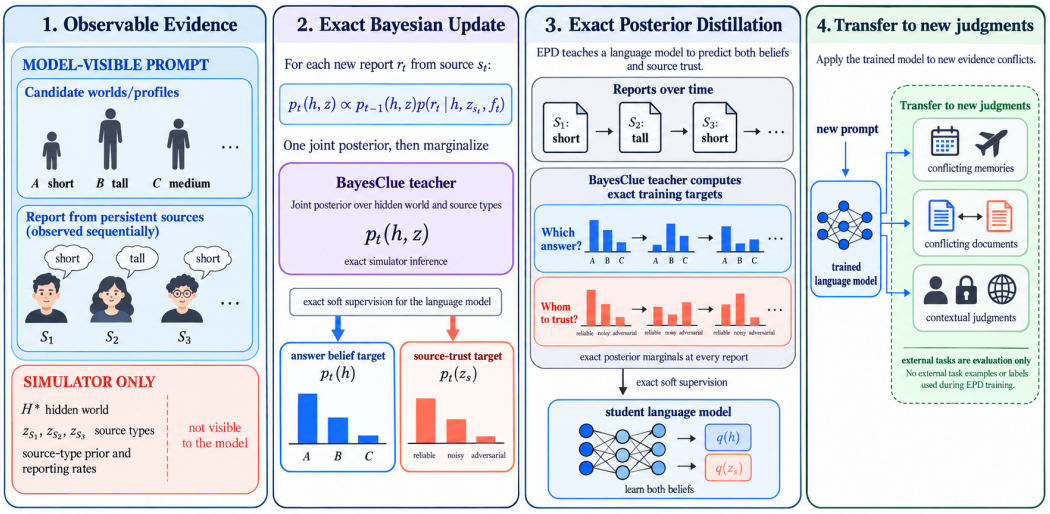

Figure 1: **Learning what to believe and whom to trust.** In **BayesClue**, named sources provide clues about a hidden answer. The simulator updates probabilities over the answer and each source's reliability after every clue, and EPD learns to predict both. We then test whether this training helps with conflicting evidence and contextual judgments. The distributions are schematic; the true answer, source types, and source proportions are hidden from the model.

the synthetic environment, the learned behavior transfers to real tasks, where trained models resolve conflicting information from documents and memories more accurately.

Our contributions are threefold:

1. **A synthetic environment with exact belief targets. BayesClue** generates sequential evidence from persistent reliable, noisy, and adversarial sources and computes exact posterior targets over candidate answers and source types. It enables controlled studies of how beliefs should change as evidence arrives.
2. **Exact Posterior Distillation** – a post-training procedure that teaches LLMs to estimate how likely each answer is and which sources to trust as new evidence arrives. It uses the exact probabilities computed by **BayesClue** as training targets, rather than only the final correct answer.
3. **Evidence of generalization and transfer.** We find lower belief error across model families, and models trained with EPD switch to the answer favored by new evidence even in sequences longer than those seen in training. Although trained only on synthetic data, EPD also transfers to real tasks: external benchmarks on conflicting memories, facts, and documents, as well as contextual-integrity judgments. We compare EPD with reliability prompting, final-answer supervised fine-tuning (SFT), and GRPO. EPD achieves lower belief error than SFT on all ten evaluated models and larger average gains than prompting on several real tasks that require weighing conflicting evidence. It also outperforms GRPO in resolving conflicting memories.

2 Related Work

Supervision beyond a final answer. Outcome and preference objectives are trained from responses or comparisons between them (Ouyang et al., 2022; Rafailov et al., 2023; Shao et al., 2024), while process supervision adds judgments of intermediate steps (Lightman et al., 2024; Uesato et al., 2022), and knowledge distillation trains a student on a teacher's soft predictions (Hinton et al., 2015; Müller et al., 2019; Yuan et al., 2020). EPD introduces a specific teacher and target: exact inference over a hidden answer and the reliability of its sources after every report. **BayesClue** supplies the controlled

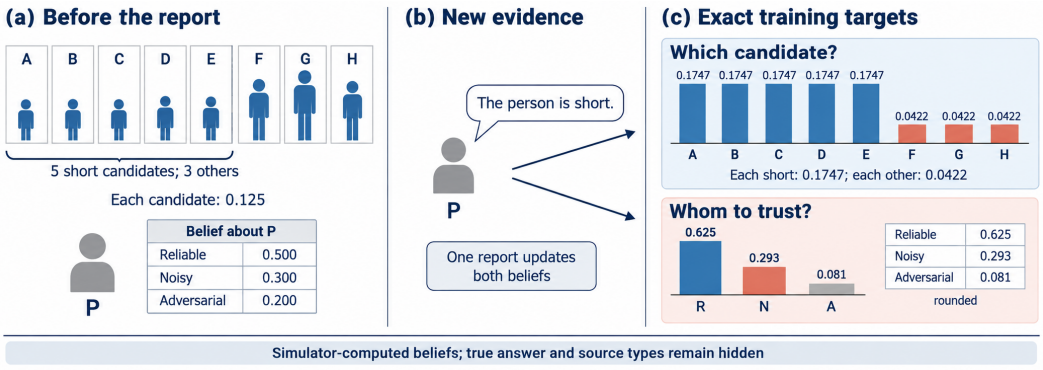

Figure 2: **One report changes what to believe and whom to trust.** Five of the eight candidates are short. When source P reports "short," the simulator updates the probabilities of the candidates and of P's possible types. These probabilities provide the training targets for EPD. Section D gives the calculation.

environment needed to compute those targets. Together, they make coupled answer-and-source beliefs directly available for language model training.

Belief tracking and conflicting evidence. Beliefs are a standard representation of uncertainty in partially observed systems (Kaelbling et al., 1998). Recent work studies Bayesian updating and confidence in language models (Kim et al., 2025; Imran et al., 2025; Kadavath et al., 2022; Geng et al., 2024). Bayesian teaching trains LLMs to imitate a Bayesian assistant (Qiu et al., 2026); we directly supervise posterior distributions over both answers and source reliability. Knowledge-conflict studies test how models reconcile passages with each other or with their parametric knowledge (Chen et al., 2022; Zhou et al., 2023; Xu et al., 2024). Our work connects these questions through training: **BayesClue** defines what the updated beliefs should be, EPD teaches them, and external evaluations measure whether the resulting behavior transfers.

Contextual integrity and safety judgments. Contextual integrity relates appropriate information use to the roles and setting in which it occurs (Nissenbaum, 2004). CASEBench uses this framework to evaluate safety judgments for requests presented in different contexts (Sun et al., 2025). We use it to test a further consequence of belief supervision: whether training on uncertain evidence improves agreement with human judgments in a setting where appropriateness depends on context. No CASEBench examples or safety labels are used in EPD training.

3 BAYESCLUE AND EXACT POSTERIOR DISTILLATION

BayesClue comprises a set of synthetic detective cases in which an agent must identify the correct candidate from clues provided by sources of unknown reliability. A case forms an *episode* with K candidates and m named *sources*. Candidate *worlds* are profiles of categorical *features*, such as cloak color, height, mark, and guild, and exactly one of them is the hidden truth of the case.

A clue is a *report* that names a source, a feature, and a claimed value. For example, source P may report that the hidden candidate is short. Every source has a hidden *type* that stays fixed within an episode: reliable sources usually report the true feature value, noisy sources make mistakes, and adversarial sources usually report a wrong value. Source names are reused, but their types are resampled between episodes. As reports arrive, the agent must update both its confidence in the candidates and its trust in the sources.

Constructing training targets. After each report, the simulator computes two kinds of training targets: the probability that each candidate is correct and, for each source, the probabilities that it is reliable, noisy, or adversarial. These probabilities account for all reports observed so far. For example, in Figure 2, source P reports that the hidden candidate is "short." This increases the probabilities of short candidates and changes the probabilities of P's possible types. Other candidates remain

Table 1: **What each training variant teaches the model.** The "rand" variants vary the proportions of reliable, noisy, and adversarial sources across episodes. The fixed variant uses proportions of $(0.5, 0.3, 0.2)$ throughout training.

Arm	Supervision
EPD-RAND	Exact world and source marginal distributions.
SAMPLE-RAND	One frozen label sampled from each EPD target.
MAP-RAND	The most probable label of each EPD target.
EPD-FIXED	Exact marginals under one fixed source mixture.
EPD-WORLD-ONLY	EPD-RAND world targets; source losses masked.

possible because P may be mistaken or adversarial. The effect of a report depends on the source's inferred type: a report from a likely adversarial source can support alternatives to its claim.

To compute these targets exactly, the simulator considers every combination of a candidate world h and an assignment of types $\mathbf{z}$ to the sources. Let $p_t(h, \mathbf{z})$ be the probability of such a combination after the first t reports, called its *posterior probability*. When source s_t reports a value r_t for feature f_t, Bayes' rule gives

$$p_t(h, \mathbf{z}) \;\propto\; p_{t-1}(h, \mathbf{z})\, p(r_t \mid h, z_{s_t}, f_t). \tag{1}$$

Here, z_{s_t} is the type of the reporting source. Normalizing the right-hand side gives the updated joint posterior, and the training targets are two kinds of marginals of it. The *world marginal* sums over all source-type assignments and gives the probability of each candidate world. The *source marginal* for a source sums over worlds and over the types of the other sources, and gives the probability that this source is reliable, noisy, or adversarial. By default, an episode has $K = 8$ candidate worlds, $m = 3$ sources, and eight reports. Every report therefore adds one world distribution and three source distributions, so a full episode provides 32 training targets: eight world and 24 source distributions. Section D provides details on the generator and inference procedure.

Exact Posterior Distillation. EPD teaches LLMs to predict the exact probabilities computed by the simulator. Each training example presents the candidate profiles, source identities, and all reports observed so far, followed by one of two questions: which candidate is correct (a *world query*), or whether a specified source is reliable, noisy, or adversarial (a *source query*). Each possible answer is represented by a single-token label. We take the model's next-token probabilities for these labels and normalize them to sum to one, obtaining its predicted distribution over candidates or source types.

Training brings this predicted distribution closer to the corresponding simulator target by minimizing their Kullback-Leibler (KL) divergence. For N training examples, let x_i be the input to the model, p_i^* the exact target distribution, and $q_\theta(\cdot \mid x_i)$ the distribution predicted by the model with parameters θ. The training objective is

$$\mathcal{L}_{\mathrm{EPD}}(\theta) = \frac{1}{N} \sum_{i=1}^{N} D_{\mathrm{KL}}(p_i^* \,\|\, q_\theta(\cdot \mid x_i))\,. \tag{2}$$

This objective supervises the probability assigned to every candidate or source type, preserving the uncertainty in the simulator's target distribution. We train LoRA adapters (Hu et al., 2022) on this objective alone. We do not penalize the model for drifting from the base model's predictions, and we do not add an auxiliary objective that preserves its other capabilities.

The model sees only the candidate profiles, the source names, and the reports observed so far. The simulator additionally knows two hidden quantities. The first is the set of reporting rates, the numerical probabilities with which sources of each type report the true value. The second is the *source mix*, the probabilities with which sources are assigned the reliable, noisy, and adversarial types. The mix is fixed at $(0.5, 0.3, 0.2)$ in the fixed-mix condition and drawn anew for every episode in the varied-mix condition. Neither the mix nor the reporting rates appear in the prompt.

4 EXPERIMENTAL SETUP

Models. We evaluate models from the `Qwen`, `OLMo`, `Granite`, and `Gemma` families, with three independent training runs per model. Section A lists all models.

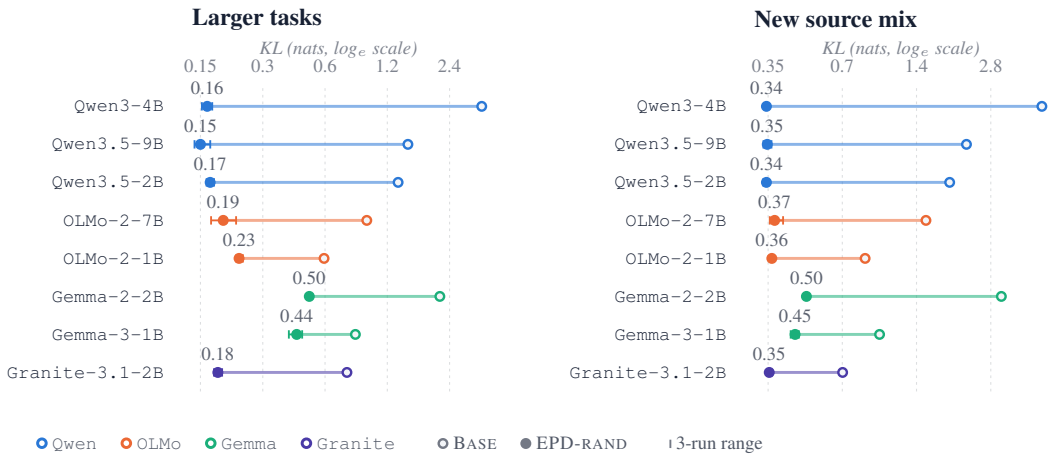

Figure 3: **EPD improves belief accuracy across four model families.** Hollow markers show the base models and filled markers show EPD-RAND; bars span three training runs. Lower KL means more accurate beliefs, and both axes use a logarithmic scale. Left: average error on standard cases, longer histories (16 reports), and larger candidate sets (12 candidates), compared with eight reports and eight candidates during training. Right: average error across three tests that change the proportions of reliable, noisy, and adversarial sources, including cases dominated by adversarial sources or by a single source type.

Training and comparison arms. We train each model for 2,000 updates at effective batch size 32, followed by 400 updates on new examples at batch size 64. To isolate the effect of the training target, EPD-RAND, SAMPLE-RAND, and MAP-RAND use identical prompts, examples, ordering, batches, and answer positions, differing only in their target distributions. Separate controls use fixed source proportions or omit source-query losses; the latter also reduces the amount of supervision (Table 1). Unless otherwise specified, EPD refers to EPD-RAND.

We compare against prompting and alternative post-training methods. PROMPT adds an instruction to consider source reliability at evaluation time. ANSWER-SFT trains on the realized hidden world using full-vocabulary cross-entropy, unlike MAP, which targets the most probable posterior label. It uses the same episodes, updates, batches, and seeds as the belief arms, but requires four passes because source-query records are omitted. GRPO (Shao et al., 2024) rewards generated world labels that match the posterior's leading answer and matches EPD in update count and adapter capacity. Sections C and L provide further details.

Internal evaluation. We measure *belief error* as forward KL, in nats, from the exact target distribution to the model's answer distribution; lower is better. We test generalization by extending episodes from eight to 16 reports, increasing the candidate count from eight to 12, and varying the proportions of reliable, noisy, and adversarial sources. Models share held-out episodes. We compare each trained model with its own base and report world and source errors separately where available.

External evaluation. We test transfer without training on external benchmark examples. Multisource-membench tests conflicts among personal memories (Yang et al., 2026), CONFACT tests factual conflicts (Ge et al., 2025), and RAMDocs evaluates judgments and generation from conflicting documents (Wang et al., 2025). FaithEval-CF tests faithfulness of a model to counterfactual context (Ming et al., 2025), RGB provides additional generation-based checks (Chen et al., 2024), and CASEBench tests context-dependent safety probabilities (Sun et al., 2025), probing transfer to related forms of context-sensitive reasoning. Cross-model results average training runs within each model before aggregating across models.

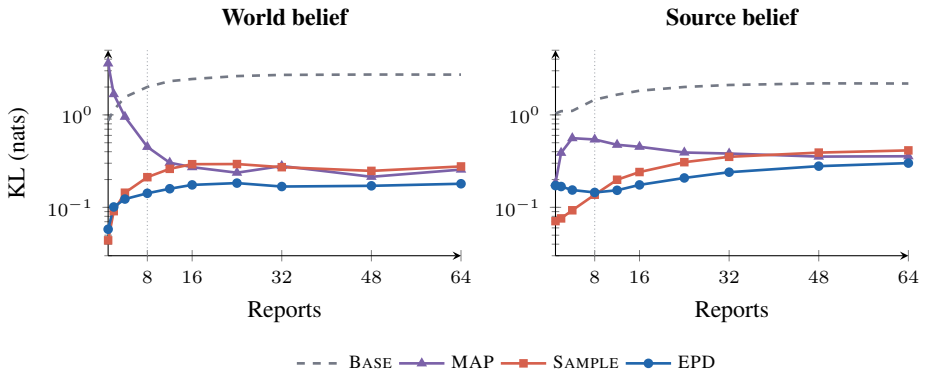

Figure 4: **Belief accuracy persists as more evidence arrives.** We evaluate `Qwen3.5-4B` on 200 episodes with up to 64 reports. The vertical line marks the eight-report training limit. The two plots show errors in predictions about the answer and source reliability; lower KL is better. Trained curves average three runs. We separately test whether new evidence prompts the model to change its answer (Section H).

5 RESULTS

5.1 LEARNING BELIEFS UNDER NEW CONDITIONS

Models learn more accurate beliefs across families EPD improves held-out belief prediction for all eight models (Figure 3). With longer histories and more candidates, the mean per-model relative reduction in KL is approximately 77%. The effect spans all four families, despite large differences in base-model performance. The gains also persist when we change the probabilities that sources are reliable, noisy, or adversarial. These tests include an adversarial-heavy population and mixtures dominated by one source type, showing that the improvement extends to different source populations.

The improvement is not explained by learning to choose the leading answer alone. At a common checkpoint, EPD predicts beliefs more accurately than MAP supervision under both structural changes and changes in the source mix (Table 6). Preserving uncertainty over alternatives therefore matters. Removing source targets also weakens robustness, consistent with the value of learning about the evidence as well as the answer (Section C).

Varying whom to trust improves robustness Training with different proportions of reliable, noisy, and adversarial sources helps models adapt when these proportions change at test time (Section G). Training with fixed proportions gives the lowest belief error on longer histories and larger candidate sets, provided the source proportions remain unchanged. However, relative to its performance on those tests, it degrades more when the source proportions change.

Models retain accurate beliefs and revise answers as evidence accumulates. Useful evidence may arrive long after a model forms its initial judgment. To test whether the model can continue to use it, we evaluate `Qwen3.5-4B` on histories of up to 64 reports, compared with eight during training. Under EPD, error in the predicted probabilities over candidate answers stays nearly flat, while the base model's error grows (Figure 4). Predictions about source reliability become less accurate as histories grow longer, but EPD still outperforms both MAP and SAMPLE at 64 reports. Each query includes all evidence available so far, so this evaluation tests how the model interprets longer histories.

As evidence accumulates, a model must also be willing to change its mind. We therefore examine episodes in which a report arriving after the eight-report training limit changes the simulator's most likely answer. We select these revision events using only the simulator's probabilities, before inspecting model predictions.

EPD responds to these changes more reliably than MAP or SAMPLE. When the simulator's probability for a candidate changes by at least 0.05, EPD moves that candidate's probability in the same

Table 2: **EPD gives the largest average gain on all three evidence-judgment benchmarks.** Values show changes from each method's reference model in percentage points; higher is better. Blue indicates improvement and orange indicates decline, with darker colors showing larger changes. C denotes CONFACT macro-F1, M MemBench accuracy, and R RAMDocs discrimination. Prompting is compared with the standard prompt. † marks an unchanged, uninformative probe; ‡ marks results that use a separate base for each method (Section J).

	EPD			PROMPT			ANSWER-SFT		
Model	C	M	R	C	M	R	C	M	R
`Qwen3-4B`	+1.5	+0.1	0.0†	+1.2	+1.4	0.0†	+2.3	−0.7	0.0†
`Qwen3.5-2B`	+9.4	−6.2	+19.8	−9.3	+1.3	+6.2	+7.6	−4.9	+14.7
`Qwen3.5-4B`	−18.3‡	+6.2	+6.2‡	−1.7	+2.5	+2.5	−1.2‡	+1.4	+5.6‡
`Qwen3.5-9B`	−10.4	−0.2	+3.8	+0.6	+1.1	0.0	+4.1	−5.4	+8.5
`OLMo-2-1B`	+8.7	+8.3	+13.0	+0.6	−10.9	+2.9	+23.5	+1.6	−4.3
`OLMo-2-7B`	+20.6	+20.1	+12.2	0.0	+2.0	+3.3	+20.2	+13.3	+2.2
`Granite-3.1-2B`	+12.1	+3.6	+10.1	+1.3	+4.8	+2.5	−9.3	−8.9	−6.3
`Gemma-2-2B`	+14.3	−1.6	−3.0	+2.2	−3.8	+0.4	−11.1	−1.5	+4.4
`Gemma-3-1B`	+28.5	+4.8	+5.1	0.0	+8.3	0.0	+0.5	+8.0	+3.4
`Gemma-3-4B`	+26.9	+17.3	+17.6	0.0	+2.7	−0.8	0.0	+14.7	+13.3

direction 90.2% of the time, compared with 83.6% for MAP. We refer to this measure as *direction accuracy*. EPD also selects the newly favored answer in 64.3% of revision events, compared with 57.5% for MAP. These results show that the model can revise its beliefs beyond the training horizon, although most selected revisions occur shortly beyond it, at reports 9–16. Section H gives the event distribution and uncertainty estimates.

5.2 BELIEF TRAINING HELPS MODELS INTERPRET EVIDENCE IN OTHER TASKS

Across eight models, training on **BayesClue** improves how models judge evidence outside the simulator (Table 2 and section J). The clearest gains are on CONFACT and RAMDocs, which test factual conflict and the ability to distinguish supported answers from misleading ones. Both improve on average, with 95% bootstrap intervals across models that exclude zero (Section J). Models also improve on average at answering questions from personal records and following supplied context when it contradicts familiar facts. These gains vary across models, and some benefit more from other forms of supervision.

Multisource-membench provides a close connection to the training task. It asks models to answer questions about a person using records from several sources, which may be incomplete or conflicting. On `Qwen3.5-4B`, EPD raises accuracy from 60.8% to 67.0%, compared with 59.0% when training omits questions about source reliability (Section C). This comparison suggests that learning about sources is useful, although omitting those questions also reduces the amount of supervision. When only the most confident 80% of predictions are retained, accuracy rises from 63.8% for the base to 72.0% for EPD.

What changes in individual predictions? The examples in Table 3 show how training changes the model's interpretation of conflicting evidence. In the memory example, EPD reconciles a missing calendar entry with a recorded trip and selects the correct answer. In the document example, it makes the supported year the most likely answer, but also assigns more probability to the misleading year. Choosing the correct answer therefore does not necessarily mean that the full probability distribution has improved.

Predictions agree more closely with human safety judgments. We also test transfer to CASEBench, where human annotators judge requests in light of the sender, recipient, and communication context (Sun et al., 2025; Nissenbaum, 2004). EPD receives neither CASEBench examples nor safety labels during training.

Table 4: **Agreement with human judgments.** Average CASEBench error across nine models. Scores are percentages; lower is better.

Metric	BASE	EPD	Change (pp)
Brier	13.60	9.59	−4.01
ECE	23.46	17.52	−5.94

Table 3: **EPD moves probability toward the supported answer.** Two examples from `Qwen3.5-2B`, with evidence and answer options shortened for readability. Values are the probabilities assigned to the listed answers. Bold marks the model's choice, and ⋆ marks the correct answer.

Evidence	Answer	BASE	EPD
Memory (item F3). No work entry in the calendar; another source records a 125-mile trip involving Garcia.	Day truly off	8.6%	11.1%
	Both occurred ⋆	41.2%	**68.9%**
	Worked despite no entry	**50.2%**	20.0%
Documents. One passage dates the Cathedral of Saint Augustine to 1856; a near-duplicate says 1900; a third is unrelated.	1856 ⋆	34.8%	**36.2%**
	1900	14.5%	31.9%
	Uncertain	**50.7%**	31.9%

After training, the models' predicted safety probabilities are closer, on average, to the proportion of annotators who judge each request safe (Table 4). Mean Brier error falls by 4.0 percentage points, with a 95% bootstrap interval across models of $[-7.7, -0.4]$. Calibration error against majority judgments, measured by ECE, falls by 5.9 points, although its interval includes zero. Six models improve on both measures (Section K). These results show better average agreement with human judgments.

5.3 COMPARISONS WITH PROMPTING AND ALTERNATIVE POST-TRAINING METHODS

We next ask whether instructing models to consider source reliability, or training them to give correct answers, produces similar benefits. The reliability instruction improves memory accuracy on eight of ten models, but its benefits are less consistent across other tasks (Table 2 and section C). For example, it raises `Qwen3.5-2B`'s memory accuracy by 1.3 percentage points while lowering CONFACT macro-F1 by 9.3 points. Prompting can therefore help, but this instruction does not consistently reproduce the improvements from belief training.

SFT and EPD learn different aspects of the task. Across ten models, ANSWER-SFT has higher held-out belief KL than every one of the five belief arms (Figure 5a,Table 12). It also increases KL over the base on nine models. The exception is `Qwen3-4B`, where KL falls from 3.432 to 1.648 but remains above all belief arms.

SFT remains competitive when models generate answers. On RAMDocs, it achieves higher recall than every belief-training method for five of eight models, although this does not necessarily represent an improvement over the base. Belief training thus has a clear advantage for estimating probabilities, while the relative benefits on external tasks depend on the model and evaluation (Section C).

Rewarding correct answers is not enough to reproduce the gains from belief supervision. Does learning which answer earns a reward also teach a model how strongly to believe it? In our GRPO experiments, the benefits are limited. Despite matching EPD in update count and adapter capacity, GRPO increases mean belief error from .733 to .792 on `Granite-3.1-2B` and only modestly reduces it, from .717 to .683, on `OLMo-2-1B`. Each result averages three runs across six held-out conditions.

Table 5: **Answering questions from conflicting memories.** MemBench accuracy (%) after EPD or GRPO training. Higher is better; trained scores average three runs.

Model	BASE	EPD	GRPO
`Granite-3.1-2B`	46.99	**50.59**	49.21
`OLMo-2-1B`	44.94	**53.24**	39.72

The difference also extends to memory questions: EPD exceeds GRPO by 1.38 and 13.52 percentage points on MemBench, respectively (Table 5). Together, these results favor directly teaching models how to distribute belief over alternatives, rather than relying on rewards for the leading answer to produce that behavior. Section L gives the full results.

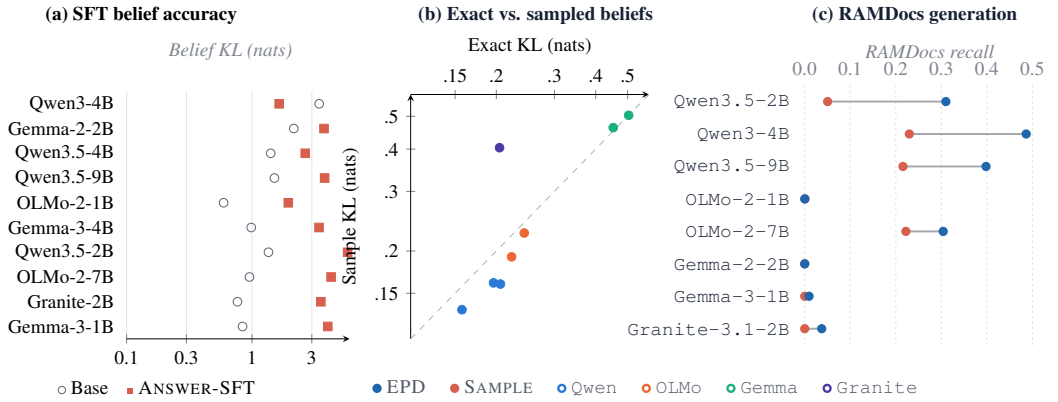

Figure 5: **Training targets affect beliefs and generated answers differently.** (a) Belief error before and after answer-SFT across ten models. (b) Belief error with exact or sampled targets across eight models, averaged over standard cases, longer histories, and larger candidate sets. The dashed line marks equal error. (c) RAMDocs generation recall with exact or sampled targets. Lower belief error and higher recall are better. Trained results average three runs.

5.4 Does Training Require the Full Posterior?

EPD teaches models the probability of every possible answer. Is this more effective than training on a single answer sampled from those probabilities? To test this, SAMPLE-RAND replaces each posterior distribution with one sampled label, fixed before training, while keeping the rest of the training setup unchanged. Unlike choosing the most likely answer, sampling allows less likely alternatives to appear as training targets.

We find no clear average advantage for exact targets in belief accuracy (Figure 5b). The mean difference is small, its uncertainty interval includes zero, and the results vary substantially across models. Thus, these experiments do not establish that providing the full distribution is necessary to achieve the observed level of belief accuracy.

The methods differ more when generating answers. Exact targets yield higher RAMDocs recall on the `Qwen` models and `OLMo-2-7B`, with broadly similar patterns on RGB (Figure 5c). Other models show small differences or perform near the floor. These results suggest that the choice of training target can affect generation even when average belief accuracy is similar. Exact targets remove the variation introduced by sampling labels, but our experiments do not establish this as the cause of the generation differences (Section F).

6 Conclusion

We introduced **BayesClue** and Exact Posterior Distillation to teach language models what to believe and whom to trust as evidence arrives. Supervising probabilities over answers and source reliability improves belief prediction across model families, supports answer revision beyond the training horizon, and transfers to evidence judgments outside the simulator. SFT remains competitive on generation, suggesting that belief and response training may be complementary.

Our simulator assumes persistent source types and categorical reports, while real sources can change with topic and time. The long-horizon study uses one base model, and external gains vary across models. Testing belief supervision with changing sources and in agents that gather their own evidence is a natural next step.

Reproducibility Statement

Sections D to F specify the generative model, exact targets, and objective identities. Sections A to C and I document the model cohorts, training comparisons, and scoring conventions. The long-horizon

selection procedure and GRPO settings are given in Sections H and L; supplementary tables report the per-model results.

AI USE STATEMENT

We used coding agents to assist with implementation and analysis, and generative AI tools for light writing assistance and the schematic illustrations in Figures 1 and 2. The authors reviewed the AI-assisted work, verified numerical claims against the experiment results, and take responsibility for the final manuscript.

REFERENCES

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2292–2307, 2022.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17754–17762, 2024.

Ziyu Ge, Yuhao Wu, Daniel Wai Kit Chin, Roy Ka-Wei Lee, and Rui Cao. Resolving conflicting evidence in automated fact-checking: A study on retrieval-augmented LLMs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 9656–9664, 2025.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 6577–6595, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Sohaib Imran, Ihor Kendiukhov, Matthew Broerman, Aditya Thomas, Riccardo Campanella, Rob Lamb, and Peter M. Atkinson. Are LLM belief updates consistent with Bayes' theorem? In *ICML 2025 Workshop on Assessing World Models*, 2025.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Minsu Kim, Sangryul Kim, and James Thorne. From evidence to belief: A Bayesian epistemology approach to language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 10578–10611, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations (ICLR)*, 2024.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. FaithEval: Can your language model stay faithful to context, even if "The Moon is Made of Marshmallows". In *International Conference on Learning Representations (ICLR)*, 2025.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Helen Nissenbaum. Privacy as contextual integrity. *Washington Law Review*, 79(1):119–157, 2004.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744, 2022.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *Nature Communications*, 17: 1238, 2026. doi: 10.1038/s41467-025-67998-6.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741, 2023.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangzhi Sun, Xiao Zhan, Shutong Feng, Phil Woodland, and Jose Such. CASE-Bench: Context-aware SafEty benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 57938–57960, 2025.

Jonathan Uesato, Nate Kushman, Ramana Kumar, H. Francis Song, Noah Y. Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. Retrieval-augmented generation with conflicting evidence. In *Second Conference on Language Modeling (COLM)*, 2025.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for LLMs: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 8541–8565, 2024.

Tiancheng Yang, Matthias Schonlau, and Ilia Sucholutsky. Selective QA over conflicting multi-source personal memory: A diagnostic testbed and method comparison. *arXiv preprint arXiv:2605.30087*, 2026.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3903–3911, 2020.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14544–14556, 2023.

A Models and Study Composition

The multi-model study covers `Qwen3-4B`, `Qwen3.5-2B`, `Qwen3.5-9B`, `OLMo-2-1B`, `OLMo-2-7B`, `Granite-3.1-2B`, `Gemma-2-2B`, `Gemma-3-1B`, and `Gemma-3-4B`. Each replicated training arm uses three independent runs from the same base. Belief accuracy, target comparisons, and four-task transfer averages use eight models, excluding `Gemma-3-4B`; CASEBench contains all nine. `Qwen3.5-4B` is used for the long-horizon study and a separate memory comparison; it is excluded from the eight- and nine-model averages. Prompting and answer-SFT belief evaluations cover all ten models. Table 2 reports all three evidence-judgment endpoints for all ten models, including additional results for `Gemma-3-4B` and `Qwen3.5-4B`. The detailed eight-endpoint SFT transfer table covers five models, while the generation ranking covers eight. GRPO has complete evaluations for `Granite-3.1-2B` and `OLMo-2-1B`, with three runs each.

Table 6: **Retaining uncertainty improves belief accuracy at a matched checkpoint.** Entries are the mean reduction in marginal-belief KL under EPD-RAND relative to MAP-RAND across eight models.

Evaluation	KL reduction (nats)
Structural changes	.288
Source-mix changes	.385

Table 7: **Exact targets give the lowest structural-shift belief error among the matched arms.** IID and structural conditions at phase-two step 400. Each cell averages marginal-belief KL equally over three conditions and three training runs (one evaluation for the base). These are fixed-checkpoint results; the fidelity plot uses validation-selected checkpoints. Lower is better.

Model	Base	EPD	Sample	MAP	Fixed	World-only
`Qwen3-4B`	3.432	0.206	0.160	0.567	0.040	0.448
`Qwen3.5-2B`	1.353	0.196	0.161	0.637	0.022	0.361
`Qwen3.5-9B`	1.509	0.157	0.134	0.683	0.018	0.257
`OLMo-2-1B`	0.594	0.243	0.226	0.423	0.105	0.318
`OLMo-2-7B`	0.955	0.222	0.192	0.591	0.042	0.369
`Granite-3.1-2B`	0.766	0.205	0.404	0.444	0.053	0.490
`Gemma-2-2B`	2.154	0.503	0.503	0.486	0.053	0.486
`Gemma-3-1B`	0.841	0.452	0.463	0.655	0.193	0.496

Table 8: **Randomized-mix training holds up best when the source mix changes.** Source-mixture conditions at phase-two step 400. Each cell averages marginal-belief KL equally over three conditions and three training runs (one evaluation for the base). These are fixed-checkpoint results; the fidelity plot uses validation-selected checkpoints. Lower is better.

Model	Base	EPD	Sample	MAP	Fixed	World-only
`Qwen3-4B`	4.513	0.386	0.342	0.827	0.415	0.525
`Qwen3.5-2B`	1.908	0.383	0.348	0.916	0.409	0.512
`Qwen3.5-9B`	2.231	0.361	0.331	1.016	0.408	0.470
`OLMo-2-1B`	0.866	0.389	0.368	0.633	0.448	0.625
`OLMo-2-7B`	1.529	0.403	0.354	0.851	0.423	0.675
`Granite-3.1-2B`	0.702	0.400	0.464	0.702	0.449	0.546
`Gemma-2-2B`	3.091	0.501	0.501	0.704	0.431	0.548
`Gemma-3-1B`	0.991	0.479	0.479	0.734	0.478	0.521

Tables 7 and 8 give the per-model results at phase-two step 400. Fixed-mix training yields low structural-shift error, but performs less well when the source mix changes.

B TRAINING RECORDS AND CONTROLLED COMPARISONS

Inputs and readouts. A record contains the candidate feature table, the named sources, the complete report prefix, and one query. World queries ask which candidate is correct; source queries ask for the type of one named source. No previously predicted belief is fed back into a later record. At each of the eight training prefixes, there is one world record and one record for each of the three sources. Thus, each episode produces 32 records, of which eight concern the world. The source types and realized world remain fixed within an episode, but are resampled for a new episode. The source names alone do not identify their types.

Let $\mathcal{A}(x)$ be the designated single-token labels for a query and $\ell_\theta(a; x)$ the model's logit at its answer position. The belief readout is

$$q_\theta(a \mid x) = \frac{\exp \ell_\theta(a; x)}{\sum_{b \in \mathcal{A}(x)} \exp \ell_\theta(b; x)}, \qquad a \in \mathcal{A}(x). \tag{3}$$

There are K labels for a world query and three for a source query. The loss concerns this answer position rather than a generated explanation. It constrains relative probability among the designated labels; it does not constrain the total probability the model assigns to those labels within its vocabulary.

Targets. Table 9 makes the target transformations explicit. For SAMPLE-RAND, the sampled label is stored once per record. Revisiting the record does not resample it. For MAP-RAND, the label is the leading answer under the posterior, which need not equal the episode's realized hidden world. Answer-SFT instead uses that realized world at every world-query prefix, including ambiguous prefixes where another candidate has larger posterior probability.

Table 9: **The arms differ only in their target and its normalization.** Targets and output normalization. p_t^H and $p_t^{Z_s}$ are the exact marginals after report t; e_y is a point mass on label y. The vocabulary-normalized Answer-SFT loss differs from the label-normalized belief losses.

Arm	World target	Source target	Normalization
EPD-RAND	p_t^H	$p_t^{Z_s}$	Query labels
SAMPLE-RAND	e_y, $y \sim p_t^H$	e_z, $z \sim p_t^{Z_s}$	Query labels
MAP-RAND	$e_{\arg\max p_t^H}$	$e_{\arg\max p_t^{Z_s}}$	Query labels
EPD-FIXED	p_t^H, fixed mix	$p_t^{Z_s}$, fixed mix	Query labels
EPD-WORLD-ONLY	p_t^H	Loss masked	Query labels
ANSWER-SFT	$e_{h_{\text{true}}}$	Records removed	Full vocabulary

Budget and repetition. The belief arms use a single pass per phase. The first phase has 2,000 updates at effective batch 32; the second uses fresh records for 400 updates at effective batch 64. This gives 64,000 and 25,600 record exposures, respectively. At 32 records per episode, these counts correspond to 2,000 and 800 complete episodes. Removing the source records leaves 16,000 and 6,400 world records for Answer-SFT. Four passes recover the same number of record exposures and optimizer updates. This matches the total number of supervised answer positions, but repeats each world record four times. The adapters have rank 64 and alpha 64, with no base-model anchor ($\lambda = 0$). The reported EPD learning rate is 10^{-4}, compared with 10^{-6} for GRPO.

Run configuration and checkpoint selection. The training generator uses $K = 8$ candidates, three sources, and eight reports. Each phase-one run contains 2,000 episodes and each phase-two run contains 800 fresh episodes. The fixed source-type mix is $(0.5, 0.3, 0.2)$; the randomized mix is drawn independently for each episode from $\mathrm{Dirichlet}(1, 1, 1)$ and is not shown in the prompt. The three training lineages use separate episode seeds. A record-level seed fixes each sampled-target label before training.

The belief adapters use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), weight decay 0.01 on non-bias parameters, gradient clipping at norm 1, and a cosine learning-rate schedule with 3% warmup in each phase. The maximum training sequence length is 2,048 tokens. Training uses bfloat16 and SDPA. LoRA uses rank and alpha 64, zero dropout, and all linear modules except inactive vision modules in

Table 10: **Belief training and answer-SFT are matched on updates, batches, and runs.** Training budget per run. Phase two follows phase one. Counts refer to supervised records, not input tokens or FLOPs. Masking source losses in the world-only control reduces its number of supervised positions.

Setting	Belief arms	Answer-SFT
Phase-one updates	2,000	2,000
Phase-two updates	400	400
Effective batch, phases one / two	32 / 64	32 / 64
Passes per phase	1	4
Source-query supervision	Yes[†]	No
Independent runs per model	3	3
Models in SFT comparison	10	10

[†]Except for EPD-WORLD-ONLY, which masks these losses.

vision-language bases. Microbatches and accumulation give the effective batches in Table 10; the 9B model uses smaller microbatches. Phase two continues the phase-one adapter. These optimizer settings apply to the belief arms. Answer-SFT matches their data, update counts, effective batch sizes, and training seeds.

Three 500-episode validation sets use the ordinary, randomized-mix, and reversed-mix regimes. A stable 500-record subset of each set is used to select the phase-one final checkpoint or a phase-two checkpoint at steps 50, 100, ..., 400 by the lowest world-plus-source marginal KL. The belief-fidelity display uses these selected checkpoints. Target comparisons use phase-two step 400 for every arm.

Held-out episodes were generated before training finished. The six conditions contain 500 episodes each: ordinary ($K = 8, T = 8$, fixed mix $(.5, .3, .2)$), longer horizon ($K = 8, T = 16$), larger candidate set ($K = 12, T = 8$), randomized mix ($\mathrm{Dirichlet}(1, 1, 1)$), reversed fixed mix $(.2, .3, .5)$, and an extreme mix drawn from $\mathrm{Dirichlet}(.2, .2, .2)$. Each condition is evaluated on the same stable subset of at most 2,000 records per model and arm. The ordinary, longer-horizon, larger-candidate, randomized, reversed, and extreme conditions use episode seeds 80000, 80004, 80005, 80001, 80002, and 80003, respectively. KL averages records within a condition; structural and source-mix summaries then average their three conditions equally.

What the controls isolate. The exact, sampled, and MAP randomized-mix arms share records, prompts, order, batches, and label sets. Their comparison changes only the target distribution. The fixed-mix comparison changes the distribution of episodes as well as their posterior targets. The world-only comparison changes both supervision content and quantity. Answer-SFT changes target identity, output normalization, query coverage, and repetition; its matched update budget does not remove these differences. The GRPO comparison matches adapter capacity and update count but has a different learning rate and generates completions during training. We therefore do not treat either comparison as matching training FLOPs.

C Prompting and Supervised Fine-Tuning

Prompting. The reliability prompt is one direct instruction to consider source reliability, rather than an exhaustive search over prompting strategies. The prompting evaluation covers ten models. Changes compare the reliability prompt with a separately evaluated standard prompt. The standard-prompt reference differs slightly from the base-model evaluation used for training comparisons; the largest MemBench difference is 1.36 percentage points on `Gemma-3-1B`. Each prompt delta uses its own reference. Table 11 includes FaithEval faithful accuracy.

SFT protocol. ANSWER-SFT uses full-vocabulary cross-entropy on the realized hidden world, with source-query records removed. It shares frozen episodes, optimizer updates (2,000 then 400), effective batches (32 then 64), and training seeds with the belief arms. Removing source rows leaves one quarter of the records, so matching updates requires four passes over the world records. Held-out scores are reported in Table 12.

Table 11: **A reliability instruction helps on memory questions but not consistently elsewhere.** Reliability prompting minus the standard prompt, in percentage points. Columns report CONFACT macro-F1, MemBench forced accuracy, RAMDocs discrimination and generation recall, and FaithEval faithful accuracy. Higher is better. These are descriptive changes; no cross-model significance claim is made.

Model	CONFACT	MemBench	RAM discr.	RAM recall	FaithEval
Qwen3-4B	+1.2	+1.4	0.0	+0.7	0.0
Qwen3.5-2B	−9.3	+1.3	+6.2	−3.2	−0.2
Qwen3.5-4B	−1.7	+2.5	+2.5	+0.2	−0.6
Qwen3.5-9B	+0.6	+1.1	0.0	+0.8	−2.6
OLMo-2-1B	+0.6	−10.9	+2.9	0.0	−3.8
OLMo-2-7B	0.0	+2.0	+3.3	+1.3	−1.8
Granite-3.1-2B	+1.3	+4.8	+2.5	−1.5	+0.6
Gemma-2-2B	+2.2	−3.8	+0.4	−0.9	−0.2
Gemma-3-1B	0.0	+8.3	0.0	+3.9	+1.4
Gemma-3-4B	0.0	+2.7	−0.8	+1.2	−0.8

Table 12: **Every belief arm reaches lower belief error than answer-SFT on all ten models.** Held-out marginal-belief KL; trained scores average three runs on frozen evaluation sets, and lower is better. The range spans the five belief arms, not a confidence interval, and is reported to two decimals in the source summary. These are separate evaluations from the long-horizon study.

Model	Base	Belief arms	Answer-SFT
Qwen3-4B	3.432	0.05–0.38	1.648
Gemma-2-2B	2.154	0.06–0.50	3.752
Qwen3.5-4B	1.410	0.02–0.37	2.658
Qwen3.5-9B	1.509	0.02–0.38	3.797
OLMo-2-1B	0.594	0.11–0.35	1.948
Gemma-3-4B	0.986	0.05–0.39	3.425
Qwen3.5-2B	1.354	0.03–0.43	5.786
OLMo-2-7B	0.955	0.05–0.42	4.276
Granite-3.1-2B	0.766	0.06–0.47	3.542
Gemma-3-1B	0.841	0.19–0.53	4.020

Interpreting the comparison. The ten-model result concerns marginal-belief KL at the evaluation interface. Answer-SFT does not directly train source readouts, and its full-vocabulary objective differs from the restricted-label loss used by the belief arms. It is therefore a test of whether conventional answer training transfers to these readouts, not an isolated comparison of soft and hard targets. MAP-RAND and SAMPLE-RAND provide the closer target controls because they retain the prompts, query types, and label normalization.

SFT transfer. Table 13 gives eight transfer endpoints for five models; Table 2 expands the three evidence-judgment endpoints to additional models. MemBench forced accuracy and selective accuracy at 80% coverage are reported separately. Qwen3-4B's unchanged FaithEval and RAMDocs probe scores are uninformative comparisons, not wins. Several generation scores for small models are near floor.

Source supervision on Qwen3.5-4B. EPD-WORLD-ONLY masks source-query losses and therefore receives fewer supervised targets than EPD. Its result does not isolate supervision content from supervision quantity. The Qwen3.5-4B comparison below uses its own base and standard-prompt evaluations.

EPD also reduces memory risk–coverage AURC from .272 to .220 and raises accuracy at 80% coverage from .638 to .720. ANSWER-SFT reaches .578 recall and .830 precision on RAMDocs generation. These generation endpoints are distinct from the candidate-label probes in Table 2.

On CASEBench, this model's Brier error falls from .184 to .067 and ECE from .347 to .082 under EPD. These results are additional to, and excluded from, the nine-model summary in Table 4.

Table 13: **Answer-SFT changes external evidence judgments inconsistently across models.** SFT minus its replication base in percentage points, averaged over three training runs. Higher is better. Column names abbreviate the model families; sizes identify the five models. FaithEval uses faithful accuracy. † denotes an uninformative unchanged probe.

Metric	Qwen3 4B	Granite 2B	OLMo-2 1B	Gemma-3 1B	Gemma-2 2B
MemBench forced accuracy	−0.7	−8.9	+1.6	+8.0	−1.5
MemBench accuracy at 80% coverage	−2.2	−12.3	+4.7	+10.1	−1.1
CONFACT macro-F1	+2.3	−9.3	+23.5	+0.5	−11.1
FaithEval-CF faithful accuracy	0.0†	−35.4	+16.7	+0.5	+0.4
RAMDocs generation recall	−8.3	−9.7	+0.9	−0.2	+3.7
RAMDocs discrimination	0.0†	−6.3	−4.3	+3.4	+4.4
RGB generation (hard)	+4.0	−30.0	+1.3	−3.1	+1.0
RGB generation (easy)	+0.6	−67.7	+3.1	−18.6	+4.6

Table 14: **EPD gives the largest memory-arbitration gain, and dropping source supervision removes it.** Memory arbitration on `Qwen3.5-4B`. Forced accuracy is shown as a percentage; higher is better. The world-only control receives fewer supervised targets than EPD.

Condition	MemBench forced accuracy (%)
BASE	60.8
PROMPT	63.2
ANSWER-SFT	62.2
EPD	67.0
EPD-WORLD-ONLY	59.0

D GENERATOR AND EXACT INFERENCE

Candidate worlds are distinct profiles sampled from four categorical features: cloak (four values), height (three), mark (three), and guild (three). The true world is uniform over the candidate set. The default setup uses eight candidates and three sources, named P, Q, and R. Sources report in round-robin order; each report's feature is drawn uniformly with replacement. A declarative wording template is drawn independently of source type. Each source's type is sampled once per episode and persists across reports.

Table 15: **Source type determines how often a report is true.** Emission probabilities for a feature with d values. Wrong reports distribute probability uniformly among the other $d-1$ values.

Source type	True value	Each wrong value
Reliable	95%	$5\%/(d-1)$
Noisy	65%	$35\%/(d-1)$
Adversarial	2%	$98\%/(d-1)$

For a known source mixture π, the initial probability of a world–source assignment is

$$p_0(h, \mathbf{z} \mid \pi) = \frac{1}{K} \prod_{s=1}^{m} \pi_{z_s}. \tag{4}$$

The simulator enumerates the $K3^m$ assignments, applies Equation (1), and normalizes in log space. The world marginal sums over all type assignments. The marginal for source s sums over worlds and the other sources' types:

$$p_t^H(h) = \sum_{\mathbf{z}} p_t(h, \mathbf{z}), \qquad p_t^{Z_s}(z) = \sum_{h} \sum_{\mathbf{z}: z_s = z} p_t(h, \mathbf{z}). \tag{5}$$

Although both targets come from one joint posterior, fitting them separately does not uniquely identify that joint distribution.

Likelihood and log-space update. For candidate h, let $v_f(h)$ be its value for feature f, let d_f be the number of possible feature values, and let $\rho_z \in \{0.95, 0.65, 0.02\}$ be the reporting accuracy of type z. Then

$$L_t(h, \mathbf{z}) = \begin{cases} \rho_{z_{s_t}}, & r_t = v_{f_t}(h), \\ (1 - \rho_{z_{s_t}})/(d_{f_t} - 1), & r_t \neq v_{f_t}(h). \end{cases} \tag{6}$$

Writing $u_t(h, \mathbf{z}) = \log p_{t-1}(h, \mathbf{z}) + \log L_t(h, \mathbf{z})$, the normalized update is

$$\log p_t(h, \mathbf{z}) = u_t(h, \mathbf{z}) - \operatorname{logsumexp}_{h', \mathbf{z}'} u_t(h', \mathbf{z}'). \tag{7}$$

The candidate table, source order, selected feature, and wording are conditioned on when computing the posterior. Their sampling does not depend on the hidden world or source types, so they add no assignment-dependent likelihood factor. Repeated reports are independent conditional on the world, source types, and feature schedule. They are generally dependent after the hidden variables are marginalized out.

The default state space has $8 \cdot 3^3 = 216$ assignments; the 12-candidate condition has 324. Each report updates every assignment and then extracts the four marginals. Straightforward enumeration takes $O(TmK3^m)$ time for all targets and $O(K3^m)$ working memory. This describes the enumeration used here, rather than a lower bound on exact inference for the model's conditional-independence structure.

Why the joint state is needed. Even with independent source-type priors, observing a report couples the reporting source's type to the candidate world. Evidence from one source can subsequently alter beliefs about another source through the shared world. Updating each source marginal independently would discard this dependence. An adversarial source is also informative: because its true-value probability is only 0.02, its claim can favor alternative feature values once its type is inferred.

Worked first report. Suppose five of eight candidates are short, and source P reports "short." Under the fixed mixture $(0.5, 0.3, 0.2)$ and a three-valued height feature, the likelihood is $0.5(0.95) + 0.3(0.65) + 0.2(0.02) = 0.674$ for a short candidate and $0.5(0.025) + 0.3(0.175) + 0.2(0.49) = 0.163$ for any other candidate. The posterior probability of each short candidate is $0.674/(5 \cdot 0.674 + 3 \cdot 0.163) \approx 0.1747$; each other candidate receives approximately 0.0422. For P's type, the report likelihoods are 0.603125, 0.471875, and 0.19625. Multiplying by the type prior and normalizing gives approximately $(0.625, 0.293, 0.081)$. Thus, a single report changes both the world and source marginals. Subsequent updates use the joint posterior, including the dependence between the world and P's type; substituting independent marginal estimates would generally be incorrect.

E Hidden Priors and the Visible-Information Reference

EPD-RAND draws an episode's source mixture from $\mathrm{Dirichlet}(1, 1, 1)$. The generator description also specifies flooring mixture components at 10^{-3} and renormalizing. The oracle posterior conditions on this realized mixture, whereas the model must predict from the visible prompt x.

This hidden, varying mixture can produce an irreducible component of expected oracle KL: the same visible evidence can correspond to different oracle targets under different mixtures. Fixed reporting rates and a fixed mixture do not create this information gap merely because they are absent from the prompt; a predictor can in principle learn those constants.

Let p_{or} denote an oracle marginal and define $p_{\mathrm{vis}}(\cdot \mid x) = \mathbb{E}[p_{\mathrm{or}}(\cdot) \mid x]$ under the same generative distribution. Then

$$\begin{aligned} \mathbb{E} D_{\mathrm{KL}}(p_{\mathrm{or}} \| q(x)) = {} & \mathbb{E} D_{\mathrm{KL}}(p_{\mathrm{or}} \| p_{\mathrm{vis}}(\cdot \mid x)) \\ & + \mathbb{E} D_{\mathrm{KL}}(p_{\mathrm{vis}}(\cdot \mid x) \| q(x)). \end{aligned} \tag{8}$$

The first term is an information gap. The second is the model's error relative to the best distribution available from the prompt. This is a population identity under a matched generative distribution, not a per-example decomposition of observed errors.

Derivation of the information gap. Fix a query type and condition on its visible input x. For any predicted distribution $q(\cdot \mid x)$ with positive probability wherever p_{vis} is positive,

$$\mathbb{E}[D_{\text{KL}}(p_{\text{or}} \| q) - D_{\text{KL}}(p_{\text{or}} \| p_{\text{vis}}) \mid x] \tag{9}$$

$$= \sum_a \mathbb{E}[p_{\text{or}}(a) \mid x] \log \frac{p_{\text{vis}}(a \mid x)}{q(a \mid x)} = D_{\text{KL}}(p_{\text{vis}} \| q). \tag{10}$$

Averaging over x gives the stated identity. Nonnegativity of KL makes p_{vis} optimal for expected oracle KL. It is the posterior mean of the oracle target conditional on the observed reports, rather than an average over mixtures drawn independently of those reports. The identity holds separately for world and source queries and therefore for a fixed weighted average of their losses. Under a distribution shift, the conditional expectation must be formed under the new distribution; a predictor optimal for the training mixture distribution need not remain optimal.

For an *unmodified* symmetric Dirichlet prior with concentration c, integrating out the mixture gives the following probability for a particular ordered source-type assignment with counts n_j:

$$p(\mathbf{z}) = \frac{\Gamma(3c)}{\Gamma(3c+m)} \prod_{j \in \{R,N,A\}} \frac{\Gamma(c+n_j)}{\Gamma(c)}. \tag{11}$$

This initializes exact enumeration without conditioning on a realized mixture. Flooring and renormalizing the mixture changes its distribution, so the expression is not an exact integrated reference for that modified generator. A numerical visible-information evaluation must match the implemented mixture transformation. We do not estimate this information gap numerically; the reported KL values use the oracle targets.

F Exact and Sample Objectives

If Y is drawn from the posterior target p^*, then

$$\mathbb{E}_{Y \sim p^*}[-\log q_\theta(Y \mid x)] = -\sum_y p^*(y) \log q_\theta(y \mid x). \tag{12}$$

Under the usual interchange of differentiation and this finite sum, the expected gradients also agree. KL differs from the right-hand side only by the target entropy, which is constant with respect to the model. Exact supervision thus averages out target-label sampling noise. SAMPLE-RAND uses a single frozen draw per record, so its finite empirical objective can differ from the exact objective even after many passes through the same records. Neither the expectation identity nor variance reduction guarantees equal downstream behavior after finite optimization.

Matched-checkpoint comparison. At phase-two step 400, sampled targets yield lower mean KL over the ordinary, longer-history, and larger-candidate conditions for six of eight models (Table 7). The Gemma-2-2B difference is only 0.000020 nats, effectively a tie. Larger gains from exact targets on Granite offset smaller losses on other models, giving a mean exact-minus-sampled difference of -0.007 nats with a 95% interval of $[-0.067, 0.030]$ (Figure 5b). These results do not establish an average advantage for exact targets.

Gradient variance. For one record, write q for the restricted-label softmax and ℓ for its logits. The exact-target gradient is $q - p^*$. The sampled-label gradient is $q - e_Y$, with

$$\mathbb{E}[q - e_Y] = q - p^*, \qquad \text{Cov}(q - e_Y) = \text{diag}(p^*) - p^*(p^*)^\top. \tag{13}$$

Its trace is $1 - \sum_a p^*(a)^2$, which vanishes for a point-mass target. Exact targets remove this label-sampling contribution at a fixed record and parameter setting. Minibatch sampling, episode selection, and optimization still introduce variation; with frozen sampled labels, revisiting a record repeats its particular target error.

Relation to realized-answer supervision. For a world query, the realized hidden world is itself a draw from the posterior when conditioning on the evidence and generator parameters. Realized-answer cross-entropy is therefore a proper population objective for the corresponding world posterior.

The same hidden world is reused at all prefixes of an episode, however, so these realized labels are dependent across prefixes. Moreover, the implemented SFT loss normalizes over the vocabulary $\mathcal{V}$. If P_θ denotes that distribution and $M_\theta(x) = \sum_{a \in \mathcal{A}(x)} P_\theta(a \mid x)$, then for a designated answer label y,

$$-\log P_\theta(y \mid x) = -\log q_\theta(y \mid x) - \log M_\theta(x). \tag{14}$$

The second term trains the model to place probability mass on the answer-label set, whereas the belief objective only specifies relative probabilities within that set. Together with the missing source targets and the repeated world records, this distinguishes Answer-SFT from the sampled-posterior control. Its empirical disadvantage is not a proof that hard labels cannot recover calibrated world probabilities with sufficient data and suitable optimization.

G Prior-Shift Comparison

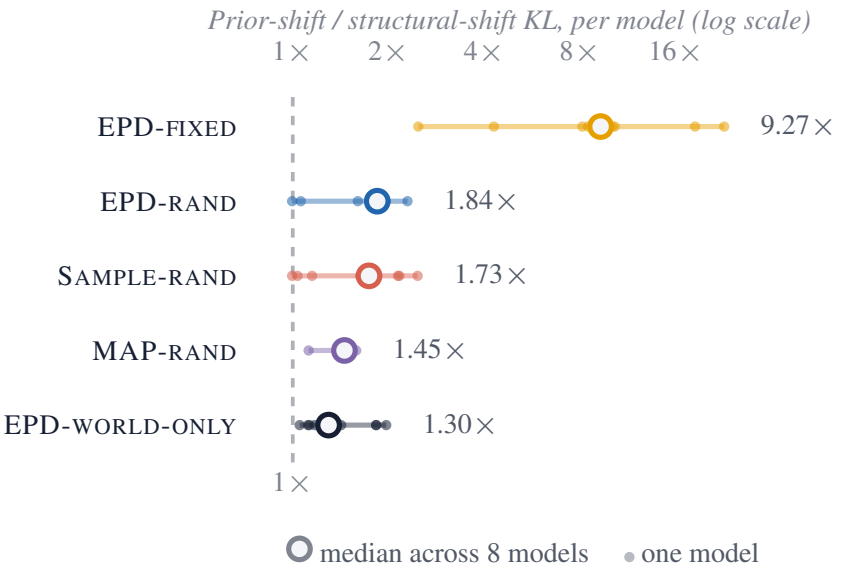

Figure 6: **Fixed-mix training degrades most when the source mix changes.** Prior-shift KL divided by structural-shift KL for each training arm. Small marks are individual models; large marks are cross-model medians over eight models. Ratios describe relative degradation and should be considered alongside absolute fidelity.

The fixed-prior arm has the largest ratio in the displayed comparison. Its strong structural-shift performance partly explains the size of this ratio. The randomized-prior arms sacrifice some specialization to one source population while varying the mixtures encountered during training. These results do not isolate the mechanism by which that variation improves robustness.

H Long Horizons and Selected Revision Events

This study uses `Qwen3.5-4B` and three independent adapter training runs, evaluated at phase-two step 400. The generator uses eight worlds, three sources, a fixed mixture $(0.5, 0.3, 0.2)$, 64-report episodes, and seed 91000. The ordinary cohort consists of the first 200 episodes and is evaluated at reports 1, 2, 4, 8, 12, 16, 24, 32, 48, and 64. Each queried state supplies one world and three source records. The curves in Figure 4 use only this ordinary cohort.

Revision events are selected when the oracle world MAP changes after report eight and the before/after posterior has total-variation distance at least 0.20. At most one event is kept per episode: the largest qualifying change. The first 200 qualifying episodes occur among the first 651 generated episodes. The ordinary and revision cohorts share 56 episode IDs, so they are not independent samples. There are 160 events at reports 9–16, 35 at reports 17–31, and five at reports 32–64.

Table 16: **EPD keeps the lowest long-horizon error and revises its answer most often.** `Qwen3.5-4B` readouts and revision outcomes. KL is evaluated at report 64 on ordinary episodes. Direction and new-leader scores use the selected revision cohort. Adapter entries average three training runs; each base condition is one distinct model result.

Condition	World KL	Source KL	Direction (%)	New leader (%)
BASE	2.725	2.183	48.9	12.0
PROMPT	2.757	1.751	51.8	12.5
EPD	0.180	0.301	90.2	64.3
MAP	0.256	0.357	83.6	57.5
SAMPLE	0.277	0.413	78.8	53.8
ANSWER-SFT	4.959	4.505	53.6	18.0

Direction accuracy considers world labels whose oracle probability changes by at least 0.05, and requires the model's change to have the same nonzero sign. There are 562 such label components across 200 events. Components within an event are dependent; bootstrap intervals resample events, not components. New-leader success requires the model's post-report MAP to match the oracle's new MAP. The EPD-minus-MAP direction difference is 6.6 percentage points, with a paired episode-bootstrap 95% interval of $[4.7, 8.6]$. The corresponding new-leader difference is 6.8 points, with interval $[2.2, 11.7]$. These intervals condition on one base model and three trained adapters.

At report 64, the oracle's mean maximum world probability is 0.9997, and none of the ordinary world posteriors has maximum probability below 0.9. Long-history success therefore differs from performance on ambiguous revision states. On revision states after the selected report, world KL is 0.340 for EPD, 0.756 for MAP, and 0.502 for SAMPLE. ANSWER-SFT was not trained to emit these probability readouts: its high KL measures transfer to this interface, rather than its free-form answer quality. Every query includes the full transcript, and no records were dropped for length in this evaluation.

Complete horizon trajectories. Tables 17 and 18 report all evaluated prefixes. The ordinary cohort contains 8,000 records: 200 episodes, ten prefixes, and four queries per prefix. Revision readouts add 1,600 records from 200 events, two prefixes, and four queries. No records were dropped for length; the longest checked prompt had 1,150 tokens against an 8,192-token limit. Horizon tables use only ordinary episodes, so the event-selected prefixes do not change the composition of the curves.

Table 17: **World-belief error stays flat under EPD out to 64 reports.** World-belief KL over the ordinary long-horizon cohort. Each column uses the same 200 episodes, and adapter entries average three runs. Training ends at eight reports.

Reports	1	2	4	8	12	16	24	32	48	64
BASE	0.864	1.125	1.558	2.004	2.315	2.442	2.627	2.706	2.730	2.725
PROMPT	0.779	0.995	1.470	1.981	2.347	2.455	2.640	2.688	2.757	2.757
EPD	0.058	0.101	0.123	0.142	0.159	0.175	0.183	0.168	0.171	0.180
MAP	3.599	1.677	0.954	0.451	0.304	0.272	0.237	0.280	0.214	0.256
SAMPLE	0.044	0.091	0.144	0.212	0.261	0.293	0.294	0.272	0.248	0.277
ANSWER-SFT	6.576	6.980	7.021	6.616	6.506	6.280	6.149	5.858	5.325	4.959

I EVALUATION METRICS AND AGGREGATION

Belief scores. For record i, the forward marginal KL is

$$D_i = \sum_{a \in \mathcal{A}(x_i)} p_i^*(a) \log \frac{p_i^*(a)}{q_\theta(a \mid x_i)},$$

with natural logarithms and $0 \log 0 = 0$. World KL averages world records; source KL averages source-query records. At a prefix with three source queries, an unweighted mean over all four records gives weight $1/4$ to the world and $3/4$ to the source average. This is not the KL between

Table 18: **Source-belief error grows with history length but remains lowest under EPD.** Source-belief KL over the ordinary long-horizon cohort. Each column uses the same 200 episodes, and adapter entries average three runs. Training ends at eight reports.

Reports	1	2	4	8	12	16	24	32	48	64
BASE	1.035	1.096	1.105	1.461	1.655	1.826	2.003	2.105	2.192	2.183
PROMPT	0.260	0.270	0.369	0.689	0.954	1.155	1.420	1.596	1.726	1.751
EPD	0.172	0.168	0.154	0.145	0.153	0.175	0.208	0.240	0.279	0.301
MAP	0.184	0.389	0.560	0.542	0.475	0.452	0.392	0.382	0.354	0.357
SAMPLE	0.071	0.076	0.093	0.137	0.199	0.241	0.308	0.352	0.391	0.413
ANSWER-SFT	5.008	4.957	4.508	4.398	4.569	4.674	4.578	4.629	4.590	4.505

joint world–source distributions. Top-1 agreement asks whether the model's leading label matches the oracle's leading label, rather than necessarily the realized world or source type. It ignores how probability is distributed over the remaining labels.

Relative reduction and model averages. If $D_{b,j}$ and $D_{a,j}$ are base and adapter errors for model j, its relative reduction is $1 - D_{a,j}/D_{b,j}$. The reported cross-model relative reduction averages these per-model ratios. It is different from dividing the mean adapter error by the mean base error. For transfer, we first average the three adapter runs within a model and then compute its change from the corresponding base. Model-bootstrap intervals resample those model-level changes. They describe the evaluated model cohort; they do not treat three runs of a shared base as three independent model families.

External endpoints. Table 19 records the endpoints used in the paper. Forced-choice discrimination, calibration, and free generation answer different questions. A model can choose the supported candidate more often while generating fewer supported answers, or improve its leading choice while assigning more probability to a misinformation alternative. We therefore retain separate scores rather than combine them into one transfer number.

Table 19: **Each external endpoint answers a different question, so we report them separately.** External metrics and their interpretation. Percentage-point changes refer to metrics reported on a 0–1 scale, except RAMDocs discrimination, which ranges from -1 to 1.

Benchmark	Endpoint	Interpretation
MemBench	Forced accuracy	Correct answer selection with full coverage.
	Selective accuracy at 80%	Accuracy when retaining the most confident 80% of predictions.
	AURC	Area under the error–coverage curve; lower is better.
CONFACT	Macro-F1	Equal-weight mean of classwise F1 scores for claim verification.
RAMDocs	Discrimination	Candidate-label probe distinguishing supported answers from misinformation.
	Generation recall / precision	Supported-answer coverage and correctness in generated text.
FaithEval-CF	Faithful accuracy	Accuracy following counterfactual supplied context.
RGB	Generation gold-answer rate	Gold-answer rate with one correct and two wrong documents (hard), or two correct and one wrong (easy).
CASEBench	Soft Brier / majority ECE	Probability error against human safe rates / calibration against majority judgments.

Benchmark procedures. The RAMDocs generation evaluation uses the first 500 items, greedy decoding, and a 1,024-token output limit. Answers are extracted after the requested answer marker and normalized before checking whether the gold answer is contained in the response. Recall counts recovered gold answers; precision counts correct predicted answers. The separate RAMDocs conflict

probe has 243 eligible items with a supported and a planted answer. Their order is randomized, the model makes a forced one-token choice, and discrimination is its gold-pick rate minus its misinformation-pick rate.

CONFACT uses 611 ModC examples and forced one-token yes/no responses. The two answer orders are both evaluated and averaged; macro-F1 weights the two classes equally. FaithEval-CF uses 500 counterfactual examples and reports faithful accuracy. MemBench evaluates 120 personas with 18 questions each for three benchmark seeds; each run averages the macro accuracy across question IDs and seeds. At 80% coverage, questions are ranked by the model's confidence, with deterministic tie handling. These procedures apply to the EPD transfer evaluation.

Selective prediction. At coverage c, a confidence-based selector retains a fraction c of the evaluated predictions. Selective accuracy measures correctness on that retained set; selective risk is one minus that accuracy. AURC integrates risk as coverage varies. At full coverage, selective and forced accuracy coincide, but the 80% endpoint measures a different subset. The SFT comparison uses the explicit 80% operating point. Prompting results use macro-averaged MemBench forced accuracy.

Dependence and uncertainty. Prefixes from one episode share a hidden world and source types. Labels in one revision event also share an evidence transition. For the revision intervals we average adapter runs within events, then resample the 200 event IDs with replacement 20,000 times using RNG seed 3. Direction accuracy is recomputed from the pooled informative-label counts in each resample. These intervals condition on the evaluated model and adapters. They do not estimate variation across base models. Point estimates from the five-model SFT transfer table are descriptive; no paired significance test between SFT and EPD is claimed.

J External Transfer by Model

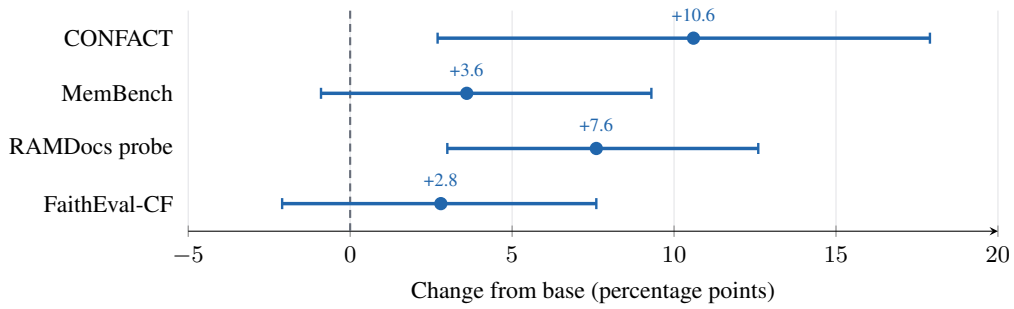

Figure 7: **EPD improves evidence judgments on average across eight models.** Mean EPD transfer across eight models, with reported 95% model-bootstrap intervals. CONFACT uses macro-F1, FaithEval-CF faithful accuracy, MemBench forced accuracy, and RAMDocs discrimination. Intervals describe variation across models, not a paired comparison with SFT or prompting.

CONFACT uses macro-F1 in this aggregate display, FaithEval-CF faithful accuracy, MemBench forced accuracy, and RAMDocs discrimination. These metrics should not be averaged in raw units. The main text reports each task separately so that a gain on one benchmark does not conceal a regression on another. The underlying external evaluation also includes generation, which is a separate endpoint from these candidate-label probes.

Table 2 compares EPD, prompting, and SFT on three evidence-judgment benchmarks. EPD gives larger gains than SFT for six of ten models on CONFACT and seven of ten on MemBench. SFT also improves evidence judgments on several models. These are descriptive comparisons, not paired significance tests between methods. The marked `Qwen3.5-4B` CONFACT and RAMDocs entries are calculated from stage-level results relative to each method's own base evaluation. CONFACT base macro-F1 is 0.5471 for EPD and 0.5512 for SFT; the respective changes are -18.3 and -1.2 percentage points. RAMDocs base discrimination is 0.0453 for EPD and 0.0576 for SFT. This base difference exceeds the 0.6-point gap between their reported improvements, so we do not rank the

	CONFACT	Membench	RAMDocs probe	FaithEval-CF
Qwen3-4B	+0.015	+0.001	+0.000	+0.000
Qwen3.5-9B	-0.104	-0.002	+0.038	+0.079
Qwen3.5-2B	+0.094	-0.062	+0.198	-0.016
OLMo-2-7B	+0.206	+0.201	+0.122	+0.084
OLMo-2-1B	+0.087	+0.083	+0.130	+0.149
Gemma-2-2B	+0.143	-0.016	-0.030	+0.016
Gemma-3-1B	+0.285	+0.048	+0.051	-0.001
Granite-3.1-2B	+0.121	+0.036	+0.101	-0.091
mean Δ	**+0.106** [+0.027,+0.179]	**+0.036** [-0.009,+0.093]	**+0.076** [+0.030,+0.126]	**+0.028** [-0.021,+0.076]

Figure 8: **Transfer gains are positive on average but vary by model and task.** EPD-RAND minus base for the eight-model external cohort. Positive values improve the indicated task metric. The bottom row contains mean changes and reported 95% model-bootstrap intervals. The model cohort is specified in Section A.

methods on this RAMDocs comparison. Small RAMDocs differences should also be interpreted cautiously given variation across training runs. The additional models in this table do not change the eight-model averages or bootstrap intervals above. FaithEval faithful accuracy is reported in the separate eight-model transfer display. GRPO reports RAMDocs gold-pick rate rather than discrimination, so the two endpoints are not compared directly.

K CASEBENCH CALIBRATION

Each model condition is evaluated on 900 items. EPD entries average three independent adapter training runs. Soft Brier uses human safe-rate targets; ECE evaluates calibration against majority labels. Their targets and interpretation therefore differ. The human safe-rate target is taken from the released item data. The model probability is normalized over the designated safe and unsafe label tokens. ECE uses ten equal-width bins on the resulting prediction confidence in $[0.5, 1]$.

Let $u_i \in [0, 1]$ be the human safe-rate target and p_i the model's predicted safe probability. The soft Brier score is

$$\text{Brier}_{\text{soft}} = \frac{1}{n}\sum_{i=1}^{n}(p_i - u_i)^2.$$

It scores distance from the human response rate, including disagreement among annotators. This differs from averaging squared error against individual binary annotations, which adds the target-dependent term $n^{-1}\sum_i u_i(1 - u_i)$. ECE instead compares binned confidence with empirical correctness against majority labels. Consequently, a Brier improvement need not imply the same change in ECE, majority accuracy, or sensitivity to paired context changes.

The calibration result does not establish a uniform benefit from source-belief supervision on safety judgments. Six models improve on both metrics, two regress on both, and one has mixed results. In particular, an improved probability score need not imply increased responsiveness to context. We reserve that claim for an evaluation with a verified definition of context sensitivity and its alignment with the human judgments.

Other belief-training controls. Tables 21 and 22 report the five training arms on the same nine-model cohort. These comparisons show that EPD's advantage over the base does not imply that it is the best arm for every model. CASEBench is a transfer evaluation; its labels are not used to construct the BayesClue targets.

Table 20: **Six of nine models improve on both CASEBench measures after EPD training.** CASEBench calibration for all nine replication models. Scores are shown as percentages of their 0–1 scales; lower is better. `Qwen3.5-4B` is excluded from these rows.

	Soft Brier (%)		ECE (%)	
Model	BASE	EPD	BASE	EPD
`Qwen3-4B`	19.03	13.90	31.18	23.62
`Qwen3.5-2B`	9.81	5.82	15.85	4.55
`Qwen3.5-9B`	7.89	5.05	8.91	5.79
`OLMo-2-1B`	29.75	15.09	47.63	28.19
`OLMo-2-7B`	10.47	10.71	27.85	24.20
`Granite-3.1-2B`	6.14	7.89	3.75	16.05
`Gemma-2-2B`	13.80	7.19	28.10	12.93
`Gemma-3-1B`	20.12	10.51	35.60	20.79
`Gemma-3-4B`	5.41	10.16	12.26	21.50

Table 21: **No single belief arm is best on CASEBench for every model.** Soft Brier for all belief-training controls on CASEBench, shown as percentages of the 0–1 score. Lower is better. Adapter entries average three independent runs, each evaluated on 900 items. World-only masks source losses; fixed uses a fixed source mix.

Model	Base	EPD	Sample	MAP	Fixed	World-only
`Qwen3-4B`	19.03	13.90	6.49	15.47	10.48	10.53
`Qwen3.5-2B`	9.81	5.82	4.68	6.60	4.30	4.62
`Qwen3.5-9B`	7.89	5.05	3.76	7.91	4.08	4.52
`OLMo-2-1B`	29.75	15.09	15.82	21.96	12.67	25.88
`OLMo-2-7B`	10.47	10.71	8.37	18.77	7.01	12.50
`Granite-3.1-2B`	6.14	7.89	10.33	13.80	9.33	8.02
`Gemma-2-2B`	13.80	7.19	7.59	24.45	12.67	11.02
`Gemma-3-1B`	20.12	10.51	9.83	19.16	7.91	13.31
`Gemma-3-4B`	5.41	10.16	9.08	14.78	8.55	9.28

Table 22: **Calibration gains under belief training are model-dependent.** Majority-label ECE for all belief-training controls on CASEBench, shown as percentages of the 0–1 score. Lower is better. Adapter entries average three independent runs, each evaluated on 900 items. World-only masks source losses; fixed uses a fixed source mix.

Model	Base	EPD	Sample	MAP	Fixed	World-only
`Qwen3-4B`	31.18	23.62	6.69	25.70	16.30	15.28
`Qwen3.5-2B`	15.85	4.55	11.65	7.12	14.07	8.57
`Qwen3.5-9B`	8.91	5.79	9.27	8.76	7.93	6.03
`OLMo-2-1B`	47.63	28.19	29.23	39.22	23.46	43.66
`OLMo-2-7B`	27.85	24.20	20.02	36.70	14.72	31.43
`Granite-3.1-2B`	3.75	16.05	15.58	29.70	19.44	16.62
`Gemma-2-2B`	28.10	12.93	12.66	41.81	22.54	23.16
`Gemma-3-1B`	35.60	20.79	19.56	34.36	13.29	23.73
`Gemma-3-4B`	12.26	21.50	16.43	30.51	19.99	21.87

L GRPO WITH MATCHED UPDATES AND ADAPTER CAPACITY

The completed `Granite-3.1-2B` and `OLMo-2-1B` evaluations each contain three independent runs, with seeds 100, 101, and 102. Training uses 2,400 GRPO updates, rank-64/alpha-64 LoRA adapters, and a learning rate of 10^{-6}. Episodes come from the same frozen phase-one data used by EPD. The reward is one when the generated final world label equals the oracle MAP answer; no distributional or source-type target is supplied. Update count and adapter capacity match EPD, but learning rates and computation per update differ. In particular, GRPO generates completions during training, so matching updates is not matching FLOPs.

Table 23: **Rewarding the leading answer leaves belief error near the base.** GRPO belief error across all six held-out conditions. Each trained score averages three runs on a frozen evaluation capped at 2,000 records per condition. Lower KL is better. Scores average world and source marginal errors, not joint-distribution KL.

	`Granite-3.1-2B`		`OLMo-2-1B`	
Condition	BASE	GRPO	BASE	GRPO
IID	.6627	.7439	.5031	.4904
16 reports	.9230	.9891	.7286	.7142
12 candidates	.7090	.8124	.5094	.4973
Random prior	.6933	.7356	.7645	.7192
Reversed prior	.3579	.3675	.7499	.6747
Extreme prior	1.0515	1.1011	1.0452	1.0042

On `Granite`, all three runs increase KL on five of six conditions. The exception is one run under the reversed prior, where KL falls by .0121 nats; the mean across runs still increases. On `OLMo`, all three runs reduce KL in every condition. These reductions range from .0121 to .0752 nats after averaging runs. We do not infer a mechanism such as formatting improvement or noise reduction from these scores alone.

Table 24: **GRPO's external scores after matched training.** Averaged over three runs. All values are percentages. Higher is better except misinformation rates. RAMDocs gold-pick rate is not the discrimination metric used in Table 2; FaithEval uses faithful accuracy across these evaluations. These are absolute scores; we do not infer changes from bases in other evaluation cohorts.

Metric	`Granite-3.1-2B`	`OLMo-2-1B`
CONFACT accuracy	79.76	22.48
CONFACT macro-F1	50.22	19.85
MemBench forced accuracy	49.21	39.72
RAMDocs gold-pick rate	20.16	47.19
RAMDocs generation recall	15.63	11.78
FaithEval faithful accuracy	80.77	49.10
RGB gold-pick rate	92.25	53.49
RGB misinformation, 2 correct / 1 wrong ↓	8.91	13.05
RGB misinformation, 1 correct / 2 wrong ↓	51.55	39.40

`OLMo-2-1B`'s low CONFACT accuracy warrants caution when interpreting its external performance, but does not by itself establish a format-compliance failure.