LEARNING WHAT TO ASK AND WHAT TO BELIEVE: DISTILLING A VARIATIONAL BAYESIAN AGENT INTO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Interactive tasks such as eliciting a user's preferences or diagnosing a hidden mechanism require an agent to track its uncertainty over a latent state, ask informative questions and act on the answers. Language models have no explicit mechanism for this, whereas Bayesian information seeking provides one through posterior inference and expected information gain (EIG), and we ask whether it can be distilled into an 8B language model. Our teacher, a lightweight Bayesian agent with a sequential variational belief and EIG query selection, matches or approaches exact and particle references on three task families. We generate synthetic interactions with the teacher on episodes disjoint from the benchmark and fine-tune Llama-3.1-8B-Instruct to reproduce its queries and its belief state as a structured scratchpad. On three Multi-Turn Puzzles tasks the student raises Circuit Decoding from 0.00 to 0.93 success, lowers the normalized rank of the recommended film from 0.50 to 0.10 and cuts word guessing from 32 to 3.4 attempts, exceeding the strongest reported language models on the two comparable tasks. Exposing the belief state closes much of the gap where the decision is an implicit computation rather than a read-out of the interaction; in matched teacher contrasts, EIG supervision consistently outperforms random querying, whereas greater inference fidelity does not uniformly improve the distilled student; a 70B language-model teacher does not match the Bayesian teacher in downstream student performance, and the learned skill does not transfer to an unseen task family. Variational Bayesian information seeking can thus be amortized into a language model by transferring the belief state that drives its questions and decisions.

1 INTRODUCTION

Many useful interactions are inference problems: to recommend a film, a system must find out what a user values; to debug a device, an engineer must probe it. In each case there is a *latent state* that the agent cannot observe but must know to act well: the weights the user places on the attributes of a film, or the gates hidden inside the device. The agent uncovers it through a sequence of *queries*, each of which reveals a little about the state, and then commits to a *decision* that is only as good as what it has learned. Doing this well requires the two abilities that sequential Bayesian experimental design studies together (Lindley, 1956; Chaloner & Verdinelli, 1995; Rainforth et al., 2024): a belief over the latent state that is updated with every answer, and the choice of the query whose answer is expected to be most informative.

Language models (LMs) do neither reliably. On the Multi-Turn Puzzles benchmark (Badola et al., 2025), whose tasks isolate this competence, the strongest reported frontier model scores only 40% on the hidden circuits, and Llama-3.1-8B-Instruct, the open model we study, almost 0%; LM performance is known to degrade in multi-turn settings (Laban et al., 2026). Inference-time wrappers place an explicit Bayesian computation around the LM (Choudhury et al., 2026; Hu et al., 2024; Handa et al., 2024) but leave the model unchanged and multiply the cost of every turn; training-based approaches improve the model itself (Andukuri et al., 2024; Zhang et al., 2025; Chen et al., 2025; Mazzaccara et al., 2024), but their supervision comes from the LM or from heuristics, so nothing guarantees that the imitated behavior is close to optimal.

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.

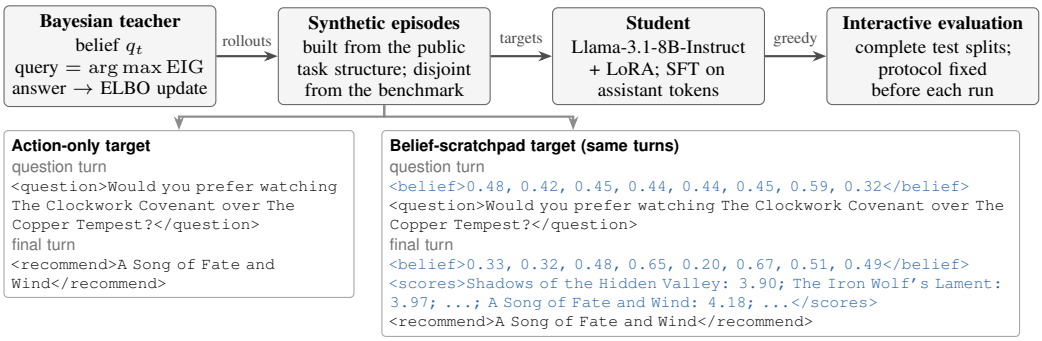

Figure 1: Overview. A Bayesian teacher interacts with synthetic episodes disjoint from the benchmark; its interactions are serialized as *actions only* or as *belief scratchpad plus action* and distilled into Llama-3.1-8B-Instruct, which is then evaluated interactively on the complete benchmark test splits. The targets are shown for two turns of one movie-recommendation dialogue of the teacher; the scratchpad adds the `<belief>` and `<scores>` lines (blue; formats for all tasks in Appendix D.4).

We take a third route (Figure 1). Our teacher is a lightweight, non-neural Bayesian agent that updates a variational belief at every turn by maximizing an evidence lower bound (ELBO) with the previous posterior as prior, and asks the query with the highest expected information gain (EIG) under that belief. It is cheap (Appendix C.6) and verifiable against Bayesian references, and because it has an explicit belief it can supervise *how* it reaches a decision and not only *what* it decides. We validate it on three task families with continuous, structured-binary and finite-symbolic latent states, generate teacher interactions on synthetic episodes disjoint from the benchmark, distill them into Llama-3.1-8B-Instruct with LoRA fine-tuning (Hu et al., 2022), and evaluate on the complete test splits under a protocol fixed before each run.

Distilled from this teacher, the student closes most of the gap between the base model and the teacher. With belief-scratchpad targets, Circuit Decoding rises from 0.00 to 0.93 Success@18 (teacher 1.00), the normalized rank of the recommended film in Movie Recommendation falls from 0.496 to 0.100 (teacher 0.071), and Word Guess falls from 31.97 to 3.37 attempts (teacher 2.29). On the two tasks for which the benchmark reports comparable language-model results, the student also improves on the strongest reported model: 97.4% versus 39.8% circuit-wise accuracy on Circuit Decoding and a normalized rank of 0.10 versus 0.16 on Movie Recommendation.

Our contributions are threefold. First, we show that a lightweight variational Bayesian information-seeking teacher can provide effective supervision for an 8B language model on three interactive-inference tasks with continuous, structured-binary, and finite-symbolic latent states, while remaining verifiable against exact or near-exact Bayesian references. Second, we show that distilling actions alone is not always sufficient: it suffices on Circuit Decoding, where the queried bits constitute the answer and the action-only student already reaches 0.92 Success@18, but when the final decision requires inference beyond the observed interaction, action-only students gather much information yet fail to convert it into the decision, and supervising the teacher's intermediate belief and task-specific decision state substantially improves the student. Third, through matched teacher and student comparisons, we separate the effects of query selection, inference fidelity, and teacher type, finding that the query rule has the largest and most consistent effect among the Bayesian-teacher components we vary, that greater inference fidelity does not necessarily produce a better distilled student, and that a 70B language model used as teacher on the same episodes yields clearly weaker students than the Bayesian teacher. The learned competence remains task-specific: leave-one-task-out experiments show little transfer of either the information-seeking policy or its information-gain criterion to the held-out task (Section 7, Appendix G).

2 Problem Formulation and Background

Interactive inference as sequential Bayesian experimental design. An episode hides a latent state $\theta^* \in \Theta$ drawn from a known prior $p(\theta)$. At turn $t = 1, \dots, T$ the agent selects a query a_t from an admissible set $\mathbb{A}_t$ fixed by the task's interaction rules and observes $y_t \sim p(y \mid \theta^*, a_t)$; the history is $h_t = (a_1, y_1, \dots, a_t, y_t)$. After the budget the agent returns a decision $d(h_T)$ scored

against θ^*. An information-seeking agent therefore supplies a *belief update* from h_t to a distribution $b_t(\theta)$, a *query rule* from b_t to $a_{t+1} \in \mathbb{A}_{t+1}$, and a *read-out* from b_T to d. Exact filtering gives $b_t(\theta) \propto b_{t-1}(\theta)\, p(y_t \mid \theta, a_t)$, Bayesian experimental design gives the query rule (Lindley, 1956; Chaloner & Verdinelli, 1995), and the read-out is the Bayes decision for the task's score. Our tasks have deterministic or near-deterministic observation models and small budgets, so every component can be made exact or near-exact.

Sequential variational belief updating. The teacher keeps a belief q_t in a fixed variational family $\mathbb{Q}$ and updates it after every observation by maximizing the per-turn evidence lower bound with the previous belief as prior,

$$q_t = \arg\max_{q \in \mathbb{Q}} \mathbb{L}_t(q), \qquad \mathbb{L}_t(q) = \mathbb{E}_{q(\theta)}\big[\log p(y_t \mid \theta, a_t)\big] - D_{\mathrm{KL}}\big(q(\theta) \,\|\, q_{t-1}(\theta)\big), \tag{1}$$

with $q_0 = p(\theta)$ when the prior lies in $\mathbb{Q}$ and its projection onto $\mathbb{Q}$ otherwise. Up to a constant, $\mathbb{L}_t(q) = -D_{\mathrm{KL}}\big(q \,\|\, \tilde{p}_t\big)$ with $\tilde{p}_t \propto q_{t-1}(\theta)\, p(y_t \mid \theta, a_t)$, so q_t is the reverse-KL projection of the one-step Bayesian update onto $\mathbb{Q}$, the recursion of streaming variational Bayes (Broderick et al., 2013; Nguyen et al., 2018): a step is exact when $\mathbb{Q}$ contains $\tilde{p}_t$, and the recursion recovers exact filtering when every posterior lies in $\mathbb{Q}$. The streaming form costs a constant amount per turn and yields exactly the per-turn belief that the scratchpad records; its small, fixed parameterization is what makes the belief writable into a training target (Appendix D.4). Because the ELBO requires a finite log-likelihood, the teacher's observation model for Movie Recommendation softens the benchmark's deterministic answer rule: with $u(i) = \theta^\top \phi_i$ the utility of film i under its attribute vector ϕ_i, the answer Yes to the pair (i, j) has the Bradley–Terry probability $\sigma\big((u(i) - u(j))/\tau\big)$, where σ is the logistic function and $\tau = 0.1$ a temperature, so that the rule recovers the benchmark's answer for non-tied pairs as $\tau \to 0$; a No Preference answer, returned only on an exact tie, is treated as uninformative, which discards the equality it conveys (Appendix C.2). The other two tasks have deterministic observation models, which we also smooth; their categorical families contain the exact update, so the only errors are the optimizer's and a small residual mass on refuted hypotheses.

Expected information gain. Given the current belief, the teacher asks the admissible query with the highest expected information gain,

$$a_{t+1} \in \arg\max_{a \in \mathbb{A}_{t+1}} \mathrm{EIG}_t(a), \qquad \mathrm{EIG}_t(a) = \mathrm{H}\big[p_t(y \mid a)\big] - \mathbb{E}_{q_t(\theta)} \mathrm{H}\big[p(y \mid \theta, a)\big], \tag{2}$$

where H is the entropy and $p_t(y \mid a) = \mathbb{E}_{q_t(\theta)}\, p(y \mid \theta, a)$ is the predictive under the current belief (Houlsby et al., 2011), computed exactly for the categorical families and by Monte Carlo with 64 posterior samples for the Gaussian family. Ties are broken uniformly at random under a fixed seed on Circuit Decoding and Movie Recommendation, and on Word Guess by posterior probability and then vocabulary order (Appendix C). The admissible set excludes the queries already asked on Circuit Decoding and Word Guess, whose deterministic answers leave nothing to learn from repetition; on Movie Recommendation it is the full set of pairs, because under the softened likelihood an asked pair keeps a positive expected gain and may be asked again, a mismatch with the deterministic respondent that costs the turn (Appendix C.2). For a deterministic observation the second term vanishes and the rule reduces to generalized binary search (Dasgupta, 2004; Nowak, 2011), and for binary feedback to uncertainty sampling, which asks the query whose predictive probability is closest to $1/2$. The rule is greedy: it values the next observation alone, without lookahead or regard for the final score.

Tasks. We evaluate on the three interactive-inference tasks of the Multi-Turn Puzzles benchmark (Badola et al., 2025) whose latent state, query space and observation model are fully specified (Table 1; Appendix B). They share the structure above but span different inference regimes (a structured binary state with deterministic bits, a continuous state with graded answers, and a finite symbolic space with multi-outcome feedback), so one teacher and one recipe are tested across latent-state and observation structures rather than within a single task.

In **Circuit Decoding** three hidden Boolean circuits over three inputs are built from AND, OR and NOT gates under a public inventory, θ is their joint truth table, each query reveals one of its bits, and after 18 of the 24 possible queries the remaining bits are deduced from the tables being realizable under the inventory. The teacher's family is a categorical over the tables enumerated from that inventory, the same hypothesis space and uniform prior as the exact reference; it differs from exact Bayes only in the update, where a smoothed likelihood and a fixed number of ELBO gradient steps leave a small residual mass on refuted tables (Appendix C.3). The read-out is the MAP table; the

Table 1: The three tasks and the teacher's variational family for each (1,000 / 300 / 400 test episodes).

	Movie Recommendation	**Circuit Decoding**	**Word Guess**
Latent state θ	weights $\in [0, 0.9]^8$ (continuous)	24-bit truth table (structured binary)	one of 40 words (finite symbolic)
Prior $p(\theta)$	uniform on the grid $\{0, 0.1, \ldots, 0.9\}^8$	uniform over the tables realizable under the inventory	uniform over the 40 words
Query a	pair of seen films (190)	(circuit, input) (24)	vocabulary word (40)
Observation y	Yes / No / No Preference	one bit	5 colors ($\leq$ 243 patterns)
Budget T	10	18	40 (minimize attempts)
Decision d	one of 40 unseen films	all 24 bits	the secret word
Benchmark metric	normalized rank of the recommended film	normalized circuit-wise accuracy	normalized number of attempts
Also reported	Success@10	Success@18	raw attempts; Success@3
Teacher family $\mathbb{Q}$	mean-field Gaussian on a range transform	categorical over grammar-enumerated tables	categorical over the 40 words

benchmark scores normalized circuit-wise accuracy, and we also report Success@18, exact recovery of all three circuits.

In **Movie Recommendation** a simulated user with hidden weights over eight film attributes scores films by a linear utility, answers pairwise comparisons of seen films (Yes, No or No Preference), and finally receives a recommendation from the unseen films. The family is a mean-field Gaussian on a logistic transform of the weight range, and the read-out is the film with the highest utility under the posterior mean. The benchmark scores $100 \times (1 - \text{normalized rank})$ of the recommendation; we also report Success@10, one if it is the user's best unseen film.

In **Word Guess** one word of a small vocabulary is secret and each attempt, a vocabulary word, is answered with per-letter feedback. The family is a categorical over the vocabulary, a family that contains the exact posterior, so the variational and exact updates coincide. The benchmark's normalization of the attempt count is not published, so we report raw attempts and Success@3 (Appendix B.5).

3 Distilling a Bayesian Teacher into a Language Model

The variational information-seeking teacher. The teacher, ELBO+EIG, combines the update of Eq. 1, the rule of Eq. 2 and a task-specific read-out; the variational family and the likelihood change between tasks (Table 1; all hyperparameters are in Appendix C). For *Movie Recommendation* the belief is a mean-field Gaussian over the eight weights, mapped into the benchmark's range, with a Bradley–Terry likelihood for the pairwise answers; the ELBO and the EIG are estimated by Monte Carlo, and the teacher recommends the unseen film with the highest utility under the posterior mean. For *Circuit Decoding* the belief is a categorical distribution over every joint truth table that the public gate inventory can realize, enumerated from the grammar alone; the ELBO and the EIG are computed exactly, and the read-out is the most probable table. For *Word Guess* the belief is a categorical distribution over the vocabulary, the EIG of a guess is the entropy of the feedback partition it induces, and the teacher commits to the most probable word once the belief has concentrated on it. Four *reference teachers* run on the same episodes: ELBO+Random and ELBO+Uncertainty keep the variational belief and change the query rule; Exact Bayes+EIG (circuit, word) and Particle Bayes+EIG (movie) keep the EIG rule and replace the variational belief by exact or near-exact inference.

Synthetic episodes and teacher interactions. Students are never trained on benchmark episodes. For each task a generator follows the benchmark's public structure and produces a few thousand episodes whose latent states are disjoint from the test split, by construction and by an explicit check against the test set. The episodes are split by latent state into training, validation and *held-out synthetic* sets (Appendix D). Every teacher is rolled out on the same episodes with the benchmark's budgets and respondent, so the students' training data differ only in the teacher's actions and beliefs.

Distillation targets: actions and beliefs. Each interaction is serialized as one chat: a fixed system prompt, the episode's public information, alternating assistant actions and user answers, and the final

decision, with actions and decisions wrapped in tags that the evaluator parses. *Action-only targets* contain nothing but the tagged action on every assistant turn. *Belief-scratchpad targets* prepend a compact serialization of the teacher's belief, so that the student writes the belief before acting on it. What is written follows the teacher's variational family: for Circuit Decoding, the posterior marginal of every truth-table bit; for Movie Recommendation, the posterior mean of the weights and, before the recommendation, the utility of every candidate film under the written weights, so that the recommendation is the argmax of a list the student has itself written; for Word Guess, the last feedback spelled out letter by letter, a verdict for every remaining candidate, and the list of survivors (formats in Appendix D.4). Under both formats the evaluator parses the tagged action and ignores everything else, so the scratchpad is never scored.

Training objective and student. The student is Llama-3.1-8B-Instruct (Grattafiori et al., 2024) with LoRA adapters (Hu et al., 2022), trained by next-token prediction on the assistant tokens only. One recipe with identical hyperparameters serves every task, teacher and target format, and the evaluated checkpoint is the one with the lowest loss on the synthetic validation split (Appendix D.7).

Controlled teacher comparisons. Episodes, recipe, student and evaluator are fixed, so a comparison between students isolates the teacher change up to training noise. We vary the *query rule* (EIG against random and uncertainty sampling, same belief), the *quality of inference* (exact or particle posterior against the variational one, same rule) and the *type* of teacher, by rolling out Llama-3.1-70B-Instruct on the same synthetic episodes as an LLM teacher: its decoding is constrained to legal actions, so its interactions are always well-formed, and its students use action-only targets and the identical recipe.

4 Experimental Setup

Evaluation episodes and conditions. All benchmark results use the full released test set for each task and the same fixed evaluation protocol for all models within that task. Training, validation and held-out synthetic episodes are generated as in Section 3 and are disjoint from the test splits; no student hyperparameter or checkpoint was chosen on the test splits, and the checks that preceded each benchmark run, against rules fixed in advance, are listed in Appendix E.5. Three choices were informed by test-split evaluations: the movie teacher's temperature and range transform were chosen (Appendix C.2), exposing beliefs followed the failure analysis of the action-only movie students (Section 5.3), and the character-level word-guess scratchpad replaced a whole-word design that failed, likely because subword tokens hide the letters of a word (Appendix D.5); the final format reproduces its benchmark results on the held-out synthetic episodes. Table 5 in Appendix A lists the conditions. The design is crossed: each of the three Bayesian teachers shared by all tasks is distilled under both target formats on all three tasks; ELBO+Uncertainty is included on Movie Recommendation, the task on which uncertainty sampling differs from EIG; and the LLM teacher, which has no belief, contributes action-only students on Circuit Decoding and Movie Recommendation. All students share one recipe (Section 3); the base model is evaluated under the action-only prompt (Appendix E.6).

Interaction protocol. Fixed for each condition before its benchmark run: one prompt template per task and target format, greedy decoding with a fixed per-turn token budget, and the benchmark's deterministic respondent. An unparsable or illegal action consumes the turn and returns an "invalid" message; a repeated legal action consumes the turn and is answered consistently; a malformed decision scores zero or the worst rank. Outputs are never repaired, re-sampled or filtered (Appendix E).

Metrics. *Task success*: the benchmark's circuit-wise accuracy and Success@18 (circuit); the benchmark's normalized-rank score and Success@10 (movie); mean attempts, with unsolved episodes counted at the full budget, and Success@3 (Word Guess, in place of the benchmark's normalized attempt count; Appendix B.5). *Information-seeking quality*: the evaluator runs a *reference posterior* on the model's own history (exact for circuit and word, a near-exact particle posterior for movie), never shown to the model, and reports the fraction of queries attaining the maximum reference EIG (*EIG-optimal rate*), the mean shortfall per query (*EIG regret*; every metric is defined in Appendix E.3), and whether the history identifies the latent state. *Belief and read-out fidelity* for scratchpad students: consistency of the written action with the written belief and distance of the written belief from the reference posterior.

Statistical analysis. All comparisons are paired on identical episodes: the exact McNemar test on discordant pairs (McNemar, 1947) and a paired bootstrap 95% confidence interval for binary out-

comes, and the Wilcoxon signed-rank test with the same bootstrap for attempt counts (Appendix E.4). Each condition is one training seed with deterministic decoding; we do not claim seed robustness.

5 RESULTS

5.1 TEACHER VALIDATION AGAINST EXACT AND PARTICLE BAYESIAN REFERENCES

Table 6 (Appendix C.7) evaluates the teachers on the full released test split of each task, using the same episodes on which the students are evaluated. On Circuit Decoding, ELBO+EIG recovers every truth table within the query budget, matches Exact Bayes+EIG in mean turns to identification (11.0), and selects an exact-EIG-optimal query on 98% of turns. On Word Guess, ELBO+EIG and Exact Bayes+EIG generate identical query sequences on all training and test episodes, because the categorical family contains the exact posterior. On Movie Recommendation, where the variational approximation is non-exact, ELBO+EIG approaches the particle reference on the rank-based benchmark score (92.9 vs. 96.5), although a gap remains in Success@10 (0.410 vs. 0.562). Random querying under the same belief performs worse than EIG on all three tasks. Uncertainty sampling ends slightly ahead of EIG at the teacher level on Movie Recommendation, but the two teachers are within noise on paired episodes (Appendix C.7): both rules act on the same mean-field belief, and it is this approximation, not the choice between the two rules, that separates them from the particle reference. Under the reference posterior the EIG teacher's questions are in fact the more informative ones. Its one measurable handicap is that no movie teacher excludes already-asked pairs, and the EIG rule, which scores a pair by its posterior predictive, re-asks an answered pair on 6% of turns, mostly among the last four questions, whereas uncertainty sampling scores pairs at the posterior mean (Eq. 3) and moves on once an answer has shifted that mean. After distillation the ordering reverses: the ELBO+EIG student leads the ELBO+Uncertainty student by 0.15 in Success@10 (Table 4; Section 5.4), because the EIG policy is the one the student partly inherits, while the uncertainty student repeats most of its questions (Section 5.5).

5.2 DISTILLATION CLOSES MOST OF THE GAP BETWEEN THE BASE MODEL AND THE TEACHER

Table 2 presents the main results. The base model never fully decodes a circuit episode, recommends the best movie at near-chance frequency, and uses most of its query budget on Word Guess. The distilled ELBO+EIG students close most of the gap between the base model and their teachers on all three tasks. On Circuit Decoding, Success@18 rises from 0 to 0.927, compared with 1.00 for the teacher. On Movie Recommendation, the benchmark score increases from 50.4 to 90.0, approaching the teacher's 92.9. On Word Guess, mean attempts fall from 31.97 to 3.37, compared with 2.29 for the teacher. These improvements over the base model are significant under paired episode-level tests (Appendix F).

On the two tasks with comparable published numbers, the distilled students also outperform the strongest language model reported by the benchmark: 97.4 versus 39.8 in circuit-wise accuracy and 90.0 versus 84.3 on the Movie Recommendation benchmark score (Appendix B.5).

5.3 ACTIONS ALONE ARE NOT ALWAYS ENOUGH: THE ROLE OF EXPLICIT BELIEFS

Action-only supervision can transfer an effective query policy without transferring the computation that turns the resulting interaction history into a decision. Table 3 compares action-only (A) and belief-scratchpad (B) students of the same teacher on identical test episodes. For the ELBO+EIG teacher on Circuit Decoding, where the queried bits are themselves the entries of the final truth table, action-only supervision is already sufficient: Success@18 changes only from 0.920 to 0.927. On the other two tasks the action-only students are above the base model but far from their teachers, and writing the belief closes most of that distance: Success@10 rises from 0.085 to 0.345 on Movie Recommendation, and mean attempts fall from 19.2 to 3.4 on Word Guess.

Figure 2a diagnoses the Movie Recommendation failure by intervening separately on the questions and on the read-out. Giving an action-only student its teacher's questions and asking only for the recommendation leaves success nearly unchanged, whereas applying a Bayes-optimal read-out to the student's own transcript recovers much of the teacher's success. The student's questions thus

Table 2: Main result on the complete test splits: base model, the belief-scratchpad students of the ELBO+EIG teacher and of the exact or particle reference, their teachers, and the strongest language model reported by the benchmark (Gemini 2.5 Pro; none for Word Guess: the normalization is unpublished and its two turn counts disagree). Benchmark score: circuit-wise accuracy (circuit) and $100 \times (1 - \text{normalized rank})$ (movie). All student–base McNemar or Wilcoxon $p < 10^{-60}$.

			ELBO+EIG		Exact / Particle Bayes+EIG		
Task	Metric	Base (8B)	Student	Teacher	Student	Teacher	Reported LM
Circuit	Success@18 ↑	0.000	**0.927**	1.000	**0.840**	1.000	–
	benchmark score (%) ↑	0.7	**97.4**	100.0	**92.8**	100.0	39.8
Movie	Success@10 ↑	0.032	**0.345**	0.410	**0.409**	0.562	–
	benchmark score (%) ↑	50.4	**90.0**	92.9	**93.1**	96.5	84.3
Word	mean attempts ↓	31.97	**3.37**	2.29	**3.10**	2.29	–
	Success@3 ↑	0.060	**0.917**	1.000	**0.917**	1.000	–

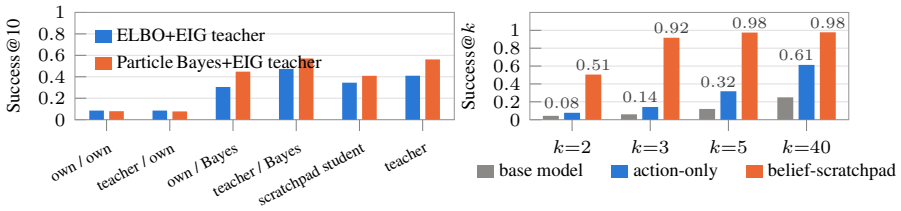

Figure 2: Action-only against belief-scratchpad distillation. (a) Movie Recommendation: success of the action-only students under their own and their teacher's questions, with their own and a Bayes-optimal read-out, against the belief-scratchpad students and the teachers. (b) Word Guess: success within k attempts for the base model and the two ELBO+EIG students. Labels in (a) give questions / read-out.

gather much of the information; what it lacks most is the read-out, a regression of the hidden weights followed by an argmax over the candidate films (Appendix F.2).

Word Guess exhibits a related but distinct failure. The action-only students usually gather enough information to isolate the secret within three attempts, but then make bookkeeping errors, repeat guesses and leave a large share of episodes unsolved (Appendix F.3). The belief scratchpad, which makes the surviving candidates and the checks behind them part of the training target, removes most of these failures (Figure 2b).

Writing the belief hurts in no final condition, although the scratchpad itself is never scored, and it helps most where the decision requires computation beyond the recorded questions and answers. Circuit Decoding is the cleanest exception: when the observations are the components of the answer, a strong action-only student already approaches its teacher, and the scratchpad helps only the two weaker circuit teachers' students, most of all the Exact Bayes+EIG student, whose final-answer errors it more than halves (Appendix F.1).

5.4 Which Part of the Teacher Matters

Table 4 reports the controlled contrasts of Section 3 on identical episodes, in the target format that works for each task. **Query rule.** In our matched Bayesian-teacher contrasts, replacing random querying with EIG produces a larger and more consistent change in student performance than replacing variational inference with exact or particle inference. On Word Guess the loss shows in Success@3 and not in the mean number of attempts, because the EIG students' rare bookkeeping errors become full-budget episodes (Section 5.5). EIG also beats uncertainty sampling where the two rules differ, reversing the teacher-level ordering of Table 6. **Quality of inference.** Exact or particle inference with the same EIG rule helps only where the approximation is loose. It helps on Movie Recommendation and makes no difference on Word Guess, where both families are exact and the two teachers' training interactions coincide, so that the two students are replicate training runs. On Circuit Decoding the effect is *negative*, although the exact teacher asks optimal queries more often

Table 3: Action-only (A) against belief-scratchpad (B) students of the same teacher on identical test episodes: score under each format and paired difference B−A with 95% bootstrap CI.

	Circuit, Success@18 ↑		Movie, Success@10 ↑		Word, attempts ↓	
Teacher	A / B	B−A	A / B	B−A	A / B	B−A
ELBO+EIG	0.920 / 0.927	+0.007 [−0.037, +0.050]	0.085 / 0.345	+0.260 [+0.228, +0.292]	19.2 / 3.4	−15.8 [−17.5, −14.0]
ELBO+Random	0.500 / 0.590	+0.090 [+0.010, +0.170]	0.105 / 0.176	+0.071 [+0.044, +0.098]	15.2 / 3.1	−12.1 [−13.4, −10.8]
ELBO+Uncertainty	–	–	0.064 / 0.194	+0.130 [+0.103, +0.157]	–	–
Exact / Particle+EIG	0.643 / 0.840	+0.197 [+0.133, +0.257]	0.079 / 0.409	+0.330 [+0.296, +0.362]	15.4 / 3.1	−12.3 [−13.8, −10.8]

Table 4: Teacher-component contrasts: paired differences between students on identical test episodes (second condition minus first) with 95% bootstrap CIs; word: Success@3 with discordant counts. Circuit uses action-only students, movie and word belief-scratchpad students (the LLM teacher row is action-only; belief-scratchpad circuit contrasts in Appendix F.1).

Contrast	Circuit	Movie	Word
Query rule: random → EIG	+0.420 [+0.350, +0.487]	+0.169 [+0.137, +0.202]	+0.205 (103/21)
Query rule: uncertainty → EIG	identical rules	+0.151 [+0.117, +0.186]	–
Inference: variational → exact or particle	−0.277 [−0.333, −0.217]	+0.064 [+0.029, +0.098]	0.000 (23/23)
Teacher type: Bayesian → LLM teacher	−0.920 [−0.950, −0.887]	−0.033 [−0.055, −0.011]	–

in training: its student imitates the teacher's queries well and fails in the read-out after the version space has collapsed. Its validation loss is twice the variational student's (Figure 4 in Appendix D.7), which suggests, but does not show, that the exact teacher's tie-breaking among EIG-optimal queries yields less predictable sequences (Appendix F.1). Writing the belief removes about two thirds of this deficit but not all of it, as expected of a read-out failure. **Type of teacher.** The LLM teacher is itself weak on the synthetic episodes while producing perfectly formed interactions, and its students inherit just that: well-formed interactions with their teacher's query quality and almost no task success (Appendix D.6). A larger general model as teacher adds little competence; the Bayesian teacher works because it supplies informative queries *and* coherent read-outs.

5.5 What the Student Learns

Students inherit their teacher's query policy. On Circuit Decoding each student's EIG-optimal rate is close to its teacher's, for strong and for weak teachers alike (Figure 5 in Appendix F), and queries are valid and never repeated until the circuit has been identified. On the continuous task the inheritance is partial: the students' questions attain the reference-EIG maximum far less often than their teachers', and some students repeat questions, so question quality is the remaining gap there (Appendix F.2). On Word Guess the inheritance is also partial: the students' guesses attain the exact-EIG maximum in five to six turns in ten, against every turn for their teachers (Figure 5). Little is lost, because on a finite hypothesis space a near-optimal guess is almost as informative: on solved episodes the students need about a quarter of an attempt more than their teacher, and the secret is the only word consistent with the feedback after at most three attempts in over 99% of their episodes (Appendix F.3).

Written beliefs are faithful and the action follows them. On Movie Recommendation the recommendation is almost always the argmax of the written scores, the written belief identifies the best film about as often as the student succeeds, and the written final belief lies close to the reference posterior mean given the same questions. On Word Guess the elimination verdicts are correct on more than 99% of turns and the written candidate set equals the exact consistent set on most turns (Appendices F.2 and F.3).

Remaining errors. On Circuit Decoding the ELBO+EIG students fail almost only after the version space collapses to one hypothesis, mostly off by one bit: a read-out slip, not a query choice. On Movie Recommendation the errors are mostly repeated or uninformative questions. On Word Guess they are rare bookkeeping errors: in about 2% of episodes a wrong verdict or a copying error drops the secret from the written list, a state the teacher's traces never contain, and the student then repeats one legal guess to the end of the budget. The random-query student has perfect bookkeeping and never dead-ends on the test split, which is why it ties the EIG students on mean attempts there while losing on Success@3 on both the test and the held-out synthetic episodes (Appendix F.3).

6 Related Work

Information seeking with language models. Language models ask suboptimal questions in twenty-questions games (Zhang et al., 2024; Bertolazzi et al., 2023), lose track of what they have learned (Laban et al., 2026), and struggle on the puzzles studied here (Badola et al., 2025). Prior remedies prompt the model to detect ambiguity and ask clarifying questions (Kuhn et al., 2022; Zhang & Choi, 2023), teach it to do so from self-generated or preference data (Andukuri et al., 2024; Zhang et al., 2025; Chen et al., 2025), or optimize question selection with an information-gain reward computed by the model itself (Mazzaccara et al., 2024; Hu et al., 2024); we instead supervise with a verifiable Bayesian agent whose questions maximize EIG under its belief.

Bayesian experimental design with language models. BED-LLM (Choudhury et al., 2026) and related methods (Handa et al., 2024; Piriyakulkij et al., 2023; Kobalczyk et al., 2025; Li et al., 2025; Arnould et al., 2026) run an EIG computation or elicitation loop around the model at test time, leaving it unchanged and paying for it at every turn. We pay it once, offline, and amortize it into the weights, as deep adaptive design does for neural policies (Foster et al., 2021; Ivanova et al., 2021; Blau et al., 2022) and concurrent work for a language-model policy with an EIG reward (Hartmann et al., 2026; Deng et al., 2026).

Learning from Bayesian models and reasoning traces. Qiu et al. (2025) fine-tune language models to mimic the predictions of a Bayesian model and report transfer to new recommendation and shopping domains; our setting additionally requires the student to choose what information to acquire, and our leave-one-task-out experiments do not exhibit analogous cross-task transfer. Zou et al. (2026) show that a drifting belief corrupts credit assignment under outcome-reward reinforcement learning, and truncate the uninformative tail of such trajectories during training, on simplified variants of two of the puzzles studied here (Appendix H); our scratchpad targets address the same failure through supervision by an explicit belief (see also Li et al., 2021; Kim & Schuster, 2023). Distilling the outputs or rationales of a stronger model (Hinton et al., 2015; Ho et al., 2023; Hsieh et al., 2023) and scratchpads or search traces (Nye et al., 2021; Wei et al., 2022; Gandhi et al., 2024; Lehnert et al., 2024; Laskin et al., 2023) shows that exposing intermediate state changes what is computed or learned; we show this for a Bayesian belief in an interactive setting (Appendix H).

7 Limitations

Task-specific competence. We trained students on two task families and evaluated them zero-shot on the third, in all three directions, under a unified interface that removes format incompatibilities. Neither the information-seeking policy nor the information-gain criterion transfers: held-out success is within noise of the base model's, one direction shows negative transfer, and every student solves its two source tasks (Appendix G); what is learned is a task-specific procedure.

Scope. The teacher needs an explicit hypothesis space and likelihood. We use one student backbone, one training seed per condition, greedy decoding and a deterministic respondent. The single replicate pair available (two word-guess students whose training data coincide) agrees closely when the students work and differs by several attempts when they fail (Appendix F.3). Question quality on the continuous task remains below the teacher's, and about two percent of word-guess episodes end in an unrecoverable bookkeeping error, a state the teacher's traces never contain.

8 Conclusion

A lightweight Bayesian agent that updates a variational belief and asks the most informative query is close to exact or particle references on three tasks where language models struggle, and can be distilled into an 8B model from synthetic interactions alone. The recipe works when the target exposes what the teacher knows, not only what it does: where the decision must be computed, a written belief closes most of the gap. Among the teacher components we vary, the information-gain rule has the largest and most consistent effect, whereas greater inference fidelity does not necessarily improve distillation; a 70B language-model teacher yields weaker students than the Bayesian teacher.

AI USE STATEMENT

In this work, we used generative AI tools (large language models used as coding and writing assistants) to implement methods, namely the teachers, the episode generators, the serialization of teacher traces into training targets, and the training and evaluation code, and to provide feedback on the experimental methodology and to draft interpretations of results, in both cases from analyses that the authors specified and reviewed. Generative AI tools were involved in producing synthetic data in two ways: the deterministic episode generators of Appendix D.1 were implemented with AI coding assistance and reviewed by the authors, and in the LLM teacher baseline of Appendix D.6, Llama-3.1-70B-Instruct is rolled out on the same synthetic episodes as an experimental comparison and its action-only dialogues train the corresponding students. No other episode, label or training target was produced by a language model. We have not used generative AI tools to develop the theoretical models or conceptual framework, to propose or refine hypotheses, to design the experiments, or to formulate or prove mathematical claims. Translation, data-set cleaning and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to identify and summarize related literature, to draft and edit parts of the manuscript for readability, to create and modify tables and figures from the experiments' output files, to run and monitor experiments, and to format references. We have reviewed all AI-assisted work: all experimental designs, decision rules and interpretations were specified or reviewed by the authors; AI-generated code was reviewed and tested by the authors, with the integrity checks, smoke runs and pre-registered validation gates described in Appendix E.5; every experimental number in the paper was traced back to the output files of the evaluator, teacher or training run that produced it, every language-model score quoted from the benchmark to the benchmark paper, and every citation surfaced with AI assistance was verified against the original source. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses synthetic interactive puzzles with simulated respondents and publicly released models and benchmark data; no human subjects or personal data are involved. Improved information-seeking ability in language models could be misused to elicit information from people more effectively; the tasks studied here are fully specified puzzles with a known latent state, and the learned competence did not transfer across the three task families (Appendix G), which limits this concern for the present work.

REPRODUCIBILITY STATEMENT

Every benchmark result in the main text is computed on the complete released test split of the corresponding task under an evaluator whose prompts, decoding settings, parsing rules and scoring are specified in Appendix E, with hashes of prompts, adapters and episode manifests verified inside every run; results on training, validation or held-out synthetic episodes, on a subset of a test split, or under the unified interface of Appendix G are labeled as such where they appear. The teachers are specified with all hyperparameters in Appendix C; the generators, splits, serialization formats and training configuration in Appendix D; the statistical procedures in Appendix E.4; software versions, hardware, runtimes and seeds in Appendix I; and the prompts in Appendix J. No code or data accompany this submission; the code, synthetic episode banks, teacher traces, training data and LoRA adapters for every condition will be released upon acceptance (Appendix I).

REFERENCES

Chinmaya Andukuri, Jan-Philipp Fränken, Tobias Gerstenberg, and Noah D. Goodman. STaR-GATE: Teaching language models to ask clarifying questions. In *Conference on Language Modeling*, 2024.

Daniel Arnould, Rashad Aziz, Zixuan Kang, Tanav Changal, Kevin Zhu, Sunishchal Dev, Gabriel Grand, and Shreyas Sunil Kulkarni. CA-BED: Conversation-aware Bayesian experimental design. *arXiv preprint arXiv:2606.01182*, 2026.

Kartikeya Badola, Jonathan Simon, Arian Hosseini, Sara Marie Mc Carthy, Tsendsuren Munkhdalai, Abhimanyu Goyal, Tomáš Kočiský, Shyam Upadhyay, Bahare Fatemi, and Mehran Kazemi. Multi-

Turn Puzzles: Evaluating interactive reasoning and strategic dialogue in LLMs. *arXiv preprint arXiv:2508.10142*, 2025.

Leonardo Bertolazzi, Davide Mazzaccara, Filippo Merlo, and Raffaella Bernardi. ChatGPT's information seeking strategy: Insights from the 20-questions game. In *Proceedings of the 16th International Natural Language Generation Conference*, 2023.

Tom Blau, Edwin V. Bonilla, Iadine Chadès, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, 2022.

Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. Streaming variational Bayes. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Maximillian Chen, Ruoxi Sun, Tomas Pfister, and Sercan Ö. Arık. Learning to clarify: Multi-turn conversations with action-based contrastive self-training. In *International Conference on Learning Representations*, 2025.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. BED-LLM: Intelligent information gathering with LLMs and Bayesian experimental design. In *International Conference on Learning Representations*, 2026. arXiv:2508.21184.

Sanjoy Dasgupta. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems*, volume 17, 2004.

Mengyi Deng, Zhiwei Li, Xin Li, Tingyu Zhu, Ying Zhao, Zhijiang Guo, and Wei Wang. Uncertainty-aware clarification in LLM agents with information gain. *arXiv preprint arXiv:2606.03135*, 2026.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential Bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

Kanishk Gandhi, Denise Lee, Gabriel Grand, Muxin Liu, Winson Cheng, Archit Sharma, and Noah D. Goodman. Stream of search (SoS): Learning to search in language. In *Conference on Language Modeling*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

Jakob Hartmann, James Harvey, Jhonathan Navott, Erik Y. Wang, Luckeciano C. Melo, Flaviu Cipcigan, Cheng Zhang, and Alessandro Abate. Amortising Bayesian experimental design for sequential information gathering in LLMs. *arXiv preprint arXiv:2607.03426*, 2026.

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

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! Outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.

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*, 2022.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Desi R. Ivanova, Adam Foster, Steven Kleinegesse, Michael U. Gutmann, and Tom Rainforth. Implicit deep adaptive design: Policy-based experimental design without likelihoods. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Ayush Rajesh Jhaveri, Anthony GX-Chen, Ilia Sucholutsky, and Eunsol Choi. Failing to falsify: Evaluating and mitigating confirmation bias in language models. *arXiv preprint arXiv:2604.02485*, 2026.

Najoung Kim and Sebastian Schuster. Entity tracking in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023.

Katarzyna Kobalczyk, Nicolas Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with LLMs. In *International Conference on Learning Representations*, 2025.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. CLAM: Selective clarification for ambiguous questions with generative language models. *arXiv preprint arXiv:2212.07769*, 2022.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, 2023.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*, 2026. arXiv:2505.06120.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. In-context reinforcement learning with algorithm distillation. In *International Conference on Learning Representations*, 2023.

Lucas Lehnert, Sainbayar Sukhbaatar, DiJia Su, Qinqing Zheng, Paul McVay, Michael Rabbat, and Yuandong Tian. Beyond A*: Better planning with transformers via search dynamics bootstrapping. In *Conference on Language Modeling*, 2024.

Belinda Z. Li, Maxwell Nye, and Jacob Andreas. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *International Conference on Learning Representations*, 2025.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. Learning to ask informative questions: Enhancing LLMs with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Cuong V. Nguyen, Yingzhen Li, Thang D. Bui, and Richard E. Turner. Variational continual learning. In *International Conference on Learning Representations*, 2018.

Robert D. Nowak. The geometry of generalized binary search. *IEEE Transactions on Information Theory*, 57(12):7893–7906, 2011.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

Wasu Top Piriyakulkij, Volodymyr Kuleshov, and Kevin Ellis. Active preference inference using language models and probabilistic reasoning. *arXiv preprint arXiv:2312.12009*, 2023.

Linlu Qiu, Fei Sha, Kelsey Allen, Yoon Kim, Tal Linzen, and Sjoerd van Steenkiste. Bayesian teaching enables probabilistic reasoning in large language models. *arXiv preprint arXiv:2503.17523*, 2025.

Tom Rainforth, Adam Foster, Desi R. Ivanova, and Freddie Bickford Smith. Modern Bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

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

Michael J. Q. Zhang and Eunsol Choi. Clarify when necessary: Resolving ambiguity through interaction with LMs. *arXiv preprint arXiv:2311.09469*, 2023.

Michael J. Q. Zhang, W. Bradley Knox, and Eunsol Choi. Modeling future conversation turns to teach LLMs to ask clarifying questions. In *International Conference on Learning Representations*, 2025.

Yizhe Zhang, Jiarui Lu, and Navdeep Jaitly. Probing the multi-turn planning capabilities of LLMs via 20 question games. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024.

Deyu Zou, Yongqiang Chen, Jianxiang Wang, Haochen Yang, Mufei Li, James Cheng, Pan Li, and Yu Gong. Reducing belief deviation in reinforcement learning for active reasoning of LLM agents. In *International Conference on Learning Representations*, 2026. arXiv:2510.12264.

A Notation and Summary of Conditions

A.1 Notation

θ, θ^*, Θ	latent state, its true value, the hypothesis space
$a_t \in \mathbb{A}_t$, y_t	query at turn t from the admissible set $\mathbb{A}_t$ (unasked queries on circuit and word; all pairs on movie), its observation
h_t	history $(a_1, y_1, \ldots, a_t, y_t)$; T the query budget
$d(h_T) \in \mathbb{D}$	terminal decision (recommendation, truth table, secret word)
$p(y \mid \theta, a)$	observation model of the task (deterministic for circuit and word)
b_t	exact Bayesian posterior given h_t; the *reference posterior* of the evaluator
$q_t \in \mathbb{Q}$	the teacher's variational belief after turn t; $\mathbb{Q}$ its family
$\mathbb{L}_t(q)$	per-turn evidence lower bound, Eq. 1, with $\beta = 1$
$\mathrm{EIG}_t(a)$	expected information gain of query a under q_t, Eq. 2
ε	smoothing of a deterministic likelihood (0.01 circuit, 10^{-6} word)
ϕ_i	attribute vector of film i (Movie Recommendation)
$\mathrm{pred}[k, a]$	answer predicted by hypothesis θ_k to query a (Circuit Decoding)

A.2 Glossary of Teachers, Students and Target Formats

ELBO+EIG	the proposed teacher: sequential variational belief (Section 2), EIG query rule (Section 2), task read-out
ELBO+Random	same belief, uniformly random admissible query
ELBO+Uncertainty	same belief, the pair whose predicted answer is closest to even odds (movie only; equals EIG on circuit)
Exact Bayes+EIG	exact posterior by hard filtering of hypotheses inconsistent with the observations; exact EIG (circuit, word)
Particle Bayes+EIG	200,000-particle posterior on the benchmark's weight grid, exact reweighting, closed-form EIG (movie)
LLM teacher	Llama-3.1-70B-Instruct, greedy decoding under a legality constraint, on the same synthetic episodes
Base model	Llama-3.1-8B-Instruct without adapters, under the action-only prompt
Action-only student	Llama-3.1-8B-Instruct + LoRA trained on action-only targets (Section 3)
Belief-scratchpad student	Llama-3.1-8B-Instruct + LoRA trained on belief-scratchpad targets
Complete test split	all released test episodes of a task (300 / 1,000 / 400)
Held-out synthetic episodes	generator episodes disjoint from training, validation and the benchmark (464 / 400 / 400)
Reference posterior	exact (or particle) posterior computed by the evaluator on the model's own history; never shown to the model
Leave-one-task-out student	one student trained on two task families and evaluated zero-shot on the third (Appendix G)

Table 5: Conditions evaluated on the complete test splits. A = action-only targets, B = belief-scratchpad targets. Uncertainty sampling coincides with EIG on Circuit Decoding; the LLM teacher has no belief and therefore no scratchpad condition.

Condition	Circuit Decoding	Movie Recommendation	Word Guess
Base model	yes	yes	yes
ELBO+EIG student	A, B	A, B	A, B
ELBO+Random student	A, B	A, B	A, B
ELBO+Uncertainty student	(= EIG)	A, B	–
Exact / Particle Bayes+EIG student	A, B	A, B	A, B
LLM teacher student	A	A	–

B Task Specifications

Every test split in this paper is the `test` split of the benchmark's public Hugging Face release (`arianhosseini/mt_puzzles`; subsets `circuit_decoding`, `movie_recommendation` and `word_guess`) (Badola et al., 2025), downloaded once and frozen by per-row hashes in a manifest that every evaluator verifies. All teacher and student numbers reported

on a test split are computed on these released episodes; the episodes we generate (Appendix D) are used only for training, validation and the held-out checks of Appendix E.5.

B.1 Movie Recommendation

Each episode provides an ordered list of eight attribute names, 20 seen films and 40 unseen films, each with an attribute vector $\phi \in \mathbb{R}^8$ (values with one or two decimals, sparse in three attributes), and a hidden weight vector $\theta^* \in \{0, 0.1, \ldots, 0.9\}^8$. The utility of a film is $\theta^{*\top}\phi$. A question names two seen films (A, B); the deterministic user answers Yes if $u(A) > u(B)$, No if $u(A) < u(B)$ and No Preference on an exact tie (utilities are exact multiples of 10^{-3}; ties occur in 1,134 of the 190,000 seen pairs of the test split). Ten questions, then one recommendation from the unseen list. Success@10 is 1 if the recommended film attains the maximum unseen utility (7 test episodes have a tied optimum, and any optimal film counts). The test split has 1,000 episodes: 20 score tables, each served to 50 users.

B.2 Circuit Decoding

Three circuits A, B, C over inputs (x_0, x_1, x_2) are built from AND, OR and NOT gates under a public total inventory; the test split contains three inventories, AND/OR/NOT $= 3/3/0$, $3/3/1$ and $1/5/2$, with 100 episodes each. The latent state is the joint 24-bit table (bits 1–8 circuit A on inputs $000, \ldots, 111$; bits 9–16 circuit B; bits 17–24 circuit C). A query `<query>C(x0,x1,x2)</query>` is answered with the bit. Eighteen queries, then the 24-bit `<guess>`. Success@18 is exact match; we also report bitwise accuracy.

B.3 Word Guess

Each episode gives a vocabulary of 40 five-letter words, of which one is secret; the test split has 10 vocabularies, each word being the secret exactly once (400 episodes). An `<attempt>` must be a vocabulary word and is answered with five colors computed by the two-pass rule (greens first, then yellows for remaining letters, duplicates handled by count). The episode ends at the correct guess or after 40 attempts. The score is the number of attempts (unsolved $= 40$); Success@k is the fraction solved within k.

B.4 Interaction Rules and Scoring

Identical for teachers and students: an illegal or unparsable action consumes the turn and is answered "Invalid query." (or "Invalid question." / invalid attempt) with no information; a repeated legal action consumes the turn and is answered consistently; a malformed terminal decision scores 0 (circuit), the worst rank (movie) or an unsolved episode (word).

B.5 Correspondence with the Metrics Reported by the Benchmark

The benchmark scores Circuit Decoding by a normalized *circuit-wise accuracy*, which we take to be the fraction of individual circuits whose truth table is recovered exactly. We report it as the benchmark score next to Success@18, the stricter criterion that all three circuits of an episode are recovered, which is a lower bound on it. On the test split the circuit-wise accuracy is 0.7% for the base model, 97.0%, 80.7% and 86.8% for the action-only ELBO+EIG, ELBO+Random and Exact Bayes+EIG students, 97.4%, 84.4% and 92.8% for their belief-scratchpad counterparts, 2.6% for the LLM teacher student and 100.0% for the two EIG teachers. For Movie Recommendation the benchmark reports the normalized rank of the recommended film, shown as a success rate $100 \times (1 - \text{normalized rank})$ with normalized rank $= (\text{rank} - 1)/39$ (ties share the best rank; 0.5 is a random pick). For Word Guess it reports a normalized attempt count whose scorer is not published, and the two mean turn counts it gives for Gemini 2.5 Pro on this task disagree (its Table 2 lists 2.4 average turns; its Section 3.1 text states 5.37), so we report raw attempts and make no language-model comparison on this task; the corresponding cell of Table 2 is left empty. The language-model values for the other two tasks are read from the summary figure of the benchmark paper (Badola et al., 2025): Gemini 2.5 Pro, Claude 3.5 Sonnet and GPT-4o score 39.8, 9.3 and 15.4 on Circuit Decoding and 84.3, 82.0 and 62.4 on Movie Recommendation. All of these published values were obtained under the benchmark's own

prompts and each API's default sampling parameters (Section 3 of the benchmark paper), not under the evaluator of Appendix E (our prompt templates, greedy decoding, fixed token budgets), so their comparison with our students (Table 2, Section 5.2) is indicative rather than paired.

C The Bayesian Teachers

C.1 The Sequential Variational Update and Its Relation to Exact Filtering

For a single observation y_t the log marginal likelihood under the current prior q_{t-1} satisfies $\log \int q_{t-1}(\theta)p(y_t \mid \theta, a_t)\, d\theta \;\geq\; \mathbb{L}_t(q)$ for every q, with equality if and only if $q \propto q_{t-1}(\theta)\, p(y_t \mid \theta, a_t)$, that is, at the exact Bayesian update. Hence maximizing $\mathbb{L}_t$ over an unrestricted family recovers exact filtering, and maximizing it over a restricted family $\mathbb{Q}$ returns the member of $\mathbb{Q}$ that minimizes $D_{\mathrm{KL}}(q \,\|\, \tilde{b}_t)$ where $\tilde{b}_t$ is the exact update of the previous variational belief. Freezing q_t and using it as the prior of turn $t+1$ gives the recursion of streaming variational Bayes (Broderick et al., 2013); the approximation error accumulates only through the projections. For a categorical $\mathbb{Q}$ over a finite hypothesis space the projection is exact, and the remaining errors are those of the finite optimizer and of the likelihood smoothing ε (Appendices C.3 and C.4).

C.2 Movie Recommendation

Family: $q_t(z) = \mathbb{N}(\mu_t, \operatorname{diag} s_t^2)$ on $z \in \mathbb{R}^8$ with $\theta = 0.9\,\sigma(z)$, so that $\theta \in (0, 0.9)^8$ matches the benchmark's range; initial $\mu_0 = 0$, $s_0 = 1$. Likelihood: the Bradley–Terry model $p(\text{Yes} \mid \theta, (i, j)) = \sigma(\theta^\top(\phi_i - \phi_j)/\tau)$ with $\tau = 0.1$; a No Preference answer is treated as uninformative for the update. Update: the ELBO expectation is estimated with 16 reparameterized samples and maximized with Adam over $(\mu, \log s)$; EIG: Monte Carlo over 64 posterior samples for every one of the 190 pairs (asked pairs are not excluded; a re-asked pair is answered consistently and consumes the turn), ties broken uniformly. Read-out: the unseen film with the highest utility under $\theta(\mu_T)$. The range transform and the temperature raised the teacher's test-split Success@10 from 0.29 to 0.42 during development on a Bradley–Terry respondent; nothing was tuned on the deterministic-respondent protocol used for all reported numbers.

C.3 Circuit Decoding

Hypothesis space: all triples of circuits realizable with the public inventory, enumerated from the grammar (every allocation of gates to the three circuits, every tree shape and leaf order, every NOT placement), de-duplicated on the joint table: $K = 252$, 3,924 and 2,886 tables for the three inventories (prior entropies 8.0, 11.9 and 11.5 bits); the grammar was verified to reproduce all 900 circuits of the test split. Prediction matrix $\mathrm{pred} \in \{0, 1\}^{K \times 24}$. Family: $q_t = \mathrm{softmax}(\lambda_t)$, $\lambda_t \in \mathbb{R}^K$. Likelihood: $p_\varepsilon(y \mid \theta_k, a) = 1 - \varepsilon$ if $\mathrm{pred}[k, a] = y$, else ε, with $\varepsilon = 0.01$. Update: exact K-term ELBO, Adam (learning rate 0.05, 80 steps) warm-started at $\lambda = \log q_{t-1}$, $\beta = 1$. EIG: $p_t(y{=}1 \mid a) = \sum_k q_t(\theta_k)[(1-\varepsilon)\mathrm{pred}[k, a] + \varepsilon(1 - \mathrm{pred}[k, a])]$ and $\mathrm{EIG}_t(a) = \mathrm{H}_2[p_t(y{=}1 \mid a)] - \mathrm{H}_2(\varepsilon)$ over unasked queries, ties uniform. Read-out: the maximum-a-posteriori table. Because every hypothesis consistent with the data receives the same update, q_t is exactly uniform on the version space; the only error is mass left on refuted hypotheses (total variation 0.027 from the exact posterior at turn 18), and the MAP table is always a consistent hypothesis.

C.4 Word Guess

Family: categorical over the 40 words, uniform prior. Likelihood: the deterministic coloring $f(a, \theta)$ smoothed as $p_\varepsilon(y \mid \theta, a) = 1 - \varepsilon$ if $y = f(a, \theta)$, else $\varepsilon/242$, with $\varepsilon = 10^{-6}$. Update: exact expectation over the 40 hypotheses, Adam on the logits (1,600 steps, learning rate 0.5), which brings every update within $D_{\mathrm{KL}} < 10^{-4}$ of the analytic posterior. EIG: entropy of the predictive feedback partition of a candidate guess (smoothed variant), over all unguessed words; ties within 10^{-4} broken by posterior probability, then vocabulary order. Commit rule: guess the maximum-a-posteriori word once its belief exceeds 0.999. The policies receive only the vocabulary, the belief and the feedback matrix; the first guess is therefore a fixed function of the vocabulary.

C.5 Reference Teachers

ELBO+Random: the same belief; the query is drawn uniformly from the admissible set $\mathbb{A}_t$ (five paired seeds per episode). ELBO+Uncertainty (movie): uncertainty sampling under the same belief, which asks the pair whose predicted answer under the posterior-mean weights is closest to even odds,

$$a_{t+1} = \arg\min_{(i,j)\in\mathbb{A}} \left|\hat{p}_t(\text{Yes} \mid i,j) - \tfrac{1}{2}\right|, \qquad \hat{p}_t(\text{Yes} \mid i,j) = \sigma\big(\bar{\theta}_t^\top(\phi_i - \phi_j)/\tau\big), \quad \bar{\theta}_t = \mathbb{E}_{q_t}[\theta], \tag{3}$$

that is, the pair with the smallest $|\bar{\theta}_t^\top(\phi_i - \phi_j)|$; unlike Eq. 2 it uses the mean of the belief and not its spread. Asked pairs are not excluded by any movie teacher (Appendix C.7). On Circuit Decoding, where every answer is one bit, uncertainty sampling coincides with Eq. 2. Exact Bayes+EIG: hard filtering, $w_t(\theta_k) \propto w_{t-1}(\theta_k)\mathbf{1}[\text{pred}[k, a_t] = y_t]$, and exact EIG $\text{H}_2[\sum_k w_t(\theta_k)\text{pred}[k, a]]$ (circuit) or the exact partition entropy (word). Particle Bayes+EIG (movie): 200,000 particles on the grid $\{0, \ldots, 0.9\}^8$, exact reweighting under the deterministic answer model with a temperature-aware likelihood, closed-form EIG; this posterior also serves as the evaluator's reference posterior for Movie Recommendation.

C.6 Hyperparameters and Computational Cost

	Movie	Circuit	Word
Family	Gaussian, 16 parameters	categorical, $K \le 3{,}924$	categorical, 40
Optimizer	Adam, 16 ELBO samples	Adam, lr 0.05, 80 steps	Adam, lr 0.5, 1,600 steps
Likelihood smoothing	$\tau = 0.1$	$\varepsilon = 0.01$	$\varepsilon = 10^{-6}$
EIG estimate	64 samples	exact	exact (smoothed)
Wall time per episode (CPU)	3.3 s	0.25 s	< 0.7 s
Reference teacher per episode	54 s (particle)	0.003 s (exact)	< 0.1 s (exact)

C.7 Extended Teacher Results

Table 6: Teacher validation on the complete released test splits (300 / 1,000 / 400 episodes; five paired seeds on Circuit Decoding and for ELBO+Random on Word Guess, one run on Movie Recommendation). EIG-optimal rate and regret are measured under the exact posterior. ELBO+Uncertainty asks the pair whose predicted answer is closest to even odds (Eq. 3, Appendix C.5) and is run only on Movie Recommendation, where it differs from EIG.

Task	Metric	ELBO+EIG	ELBO+Rand.	ELBO+Unc.	Exact / Part.
Circuit	Success@18 ↑	**1.00**	0.70	–	**1.00**
	turns to identify ↓	11.0	17.1	–	11.0
	EIG-opt. / regret (bits)	0.98 / 0.001	0.30 / 0.33	–	1.00 / 0
Movie	Success@10 ↑	0.410	0.373	0.428	**0.562**
	benchmark score (%) ↑	92.9	91.0	93.8	**96.5**
Word	mean attempts ↓	**2.29**	3.09	–	**2.29**
	Success@3 ↑	**1.00**	0.73	–	**1.00**

Circuit Decoding, per inventory (300 test episodes × 5 paired seeds).

	Success@12			queries to identification		
Inventory (K)	ELBO+EIG	ELBO+Rand.	Exact+EIG	ELBO+EIG	ELBO+Rand.	Exact+EIG
3/3/0 (252)	0.88	0.32	0.87	9.3	16.3	9.2
3/3/1 (3924)	0.63	0.18	0.58	12.3	17.8	12.3
1/5/2 (2886)	0.82	0.25	0.84	11.5	17.3	11.3

Overall: Success@18 1.00 / 0.70 / 1.00; bitwise accuracy at 18: 1.000 / 0.972 / 1.000; posterior mass on θ^* at 18: 0.97 / 0.62 / 1.00; total variation to the exact posterior at 18: 0.027 / 0.123 / 0; exact-EIG regret 0.001 / 0.329 / 0 bits per query.

Movie Recommendation (1,000 test episodes, deterministic respondent).

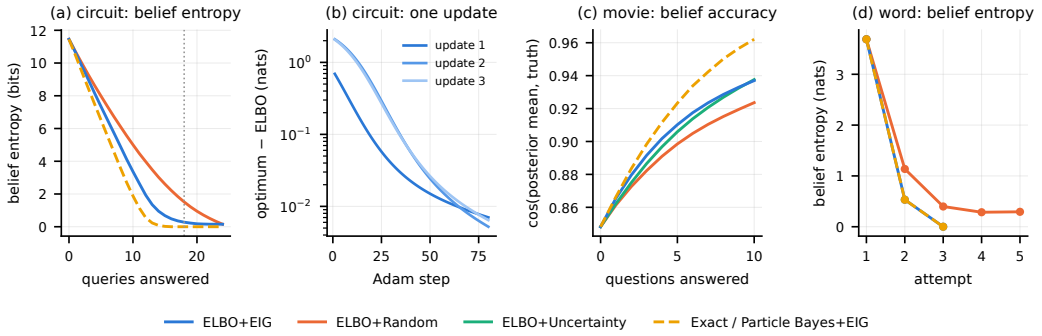

Figure 3: Inference dynamics of the teachers. (a) Circuit Decoding: mean entropy of the belief after each query on the synthetic training episodes (dotted line: the 18-query budget). (b) One variational update on a training episode: distance of the ELBO from its analytic optimum over the 80 Adam steps, for the first three updates. (c) Movie Recommendation: cosine between the posterior mean and the true weights after each question (synthetic training episodes). (d) Word Guess: mean belief entropy before each attempt on the complete test split (episodes still running; the two EIG teachers coincide).

Teacher	Success@10	norm. rank	benchmark %	rank of true best	No Pref. rate
ELBO+EIG	0.410 ± 0.016	0.071	92.9	3.40	0.011
ELBO+Random	0.373 ± 0.015	0.090	91.0	3.88	0.006
ELBO+Uncertainty	0.428 ± 0.016	0.062	93.8	3.07	0.014
Particle Bayes+EIG	0.562 ± 0.016	0.035	96.5	2.38	0.031

On the synthetic training episodes the same teachers reach 0.434 / 0.385 / 0.444 / 0.546, with reference-EIG-optimal rates 0.142 / 0.005 / 0.015 / 0.841 and reference-EIG regret 0.055 / 0.203 / 0.122 / 0.001 nats per question. Paired on the 1,000 test episodes, ELBO+Uncertainty minus ELBO+EIG in Success@10 is $+0.018$ (191 / 173 discordant episodes, exact McNemar $p = 0.37$, 95% bootstrap CI $[-0.018, +0.055]$); the two teachers' per-turn success curves stay within 0.02 of each other at every turn, and their question sets are nearly disjoint (mean Jaccard overlap 0.035). No movie teacher excludes asked pairs: the repeated-question rates are 0.064 / 0.022 / 0.014 / 0.024 (ELBO+EIG / ELBO+Random / ELBO+Uncertainty / Particle Bayes+EIG), and 84% of the ELBO+EIG repeats fall in the last four questions.

Word Guess (400 test episodes). Mean attempts 2.29 ± 0.03 / 3.09 ± 0.03 / 2.29 ± 0.03; Success@1/2/3: 0.025/0.682/1.000 for both EIG teachers, 0.024/0.317/0.733 for ELBO+Random; Success@5 1.000 / 0.967 / 1.000. Paired: ELBO+EIG saves 0.79 attempts against ELBO+Random (95% bootstrap CI $[0.72, 0.87]$, cluster CI over vocabularies $[0.74, 0.85]$, Wilcoxon $p = 5.7 \times 10^{-52}$; better on 328, worse on 47, tied on 25 episodes) and produces guess sequences identical to Exact Bayes+EIG on all 400 episodes.

Inference dynamics. Figure 3 shows how the teachers' beliefs evolve. On Circuit Decoding the entropy of the variational belief falls from 11.5 bits (the mean prior entropy over the 3,397 training episodes, which are dominated by the two large inventories) to 0.28 bits after 18 EIG queries, against 1.52 bits with random queries and 0.00 for exact inference with EIG queries; the total-variation distance between the variational belief and the exact posterior is about 0.03 at the final turn (0.027 on the test split; 0.12 under random queries, whose refuted hypotheses are eliminated more slowly). Within one update, the 80 Adam steps bring the ELBO to within about 0.01 nats of its analytic optimum, the log marginal likelihood of the answer under the previous belief, so the per-turn projection is essentially exact and the residual error is the smoothing ε. On Movie Recommendation the cosine between the posterior mean and the true weights rises from 0.85 (prior) to 0.94 after ten EIG questions, 0.92 with random questions and 0.96 for the particle reference; the remaining gap between the Gaussian and the particle belief is the approximation the student inherits. On Word Guess the belief entropy drops from 3.69 nats (uniform over 40 words) to 0.53 after one EIG attempt and to zero after two on almost every episode, against 1.14 and 0.40 nats with random attempts; the variational update matches the analytic categorical posterior to within a KL divergence of 6×10^{-5} on every one of the 6,755 updates.

D SYNTHETIC DATA AND DISTILLATION DETAILS

D.1 EPISODE GENERATORS AND SEPARATION FROM THE BENCHMARK

Circuit. A layered generator over the public grammar with the three test inventories produces 4,296 episodes with distinct joint truth tables; each table is checked against the frozen test manifest (overlap 0). **Movie.** 220 score tables $\times$ 20 users $=$ 4,400 episodes; the generator follows the format of the public rows (eight attributes, 20 seen and 40 unseen films, one- or two-decimal scores sparse in three attributes) with attribute budgets calibrated to the public rows; users are drawn on the same weight grid; the generator rejects the 1,000 test weight vectors and the 20 test score tables, and every episode is checked against the frozen test manifest (weight-vector and score-table overlap 0). Utilities are computed exactly (rounded to nine decimals) so that ties are handled as in the benchmark. **Word.** 110 vocabularies $\times$ 40 secrets $=$ 4,400 episodes drawn from a pool of 1,933 five-letter words that excludes every word of the 10 test vocabularies; every synthetic vocabulary is checked to contain no test word.

D.2 DATA SPLITS AND DATASET STATISTICS

	Circuit	Movie	Word
Episodes (train / val / held-out synthetic)	3,397 / 435 / 464	3,600 / 400 / 400	3,600 / 400 / 400
Split unit	truth table (by inventory)	score table	vocabulary
Messages per dialogue	40	24	2×attempts
Tokens per dialogue, action-only	949 (all)	$\approx$ 3,570 (max 3,624)	$\leq$ 971
Tokens per dialogue, belief scratchpad	3,516 (all)	max 4,709	max 1,388
Supervised tokens per teacher, action-only	0.78M	0.82M	0.07–0.10M
Supervised tokens per teacher, belief scratchpad	8.85M	–	2.83M
Teacher success on training episodes (ELBO+EIG)	1.000	0.434	2.31 attempts

D.3 DIALOGUE SERIALIZATION FOR ACTION-ONLY TARGETS

One dialogue is one chat: the system prompt of Appendix J, a user message with the public episode information (the inventory; the attribute names and film lists; the vocabulary), then alternating assistant actions and user answers, the terminal user prompt, and the assistant's decision. Only assistant messages are supervised. Circuit example (abridged):

```
user:      Circuit Decoding configuration: ... components: AND=3, OR=3, NOT=0.
           Ask your first query.
assistant: <query>C(0,0,1)</query>
user:      0
assistant: <query>C(1,0,1)</query>
...        (18 queries in total)
user:      Your query budget is exhausted. Submit your final answer now. ...
assistant: <guess>001101110001001100110111</guess>
```

A validator checks every compiled dialogue for format, causality (no future action visible in any prefix), absence of hidden-state leakage (the latent state, circuit wiring, utilities and diagnostics never appear) and fidelity to the teacher trace.

D.4 BELIEF-SCRATCHPAD FORMATS

Movie. Before every question: `<belief>w1, ..., w8</belief>` with the teacher's posterior mean mapped to weight space, two decimals (the prior line is 0.45 repeated eight times). Before the recommendation: `<belief>...</belief>`, then `<scores>TITLE: S; ...</scores>` listing every unseen film in the given order with its utility under the written weights (two decimals), then `<recommend>`. In the training targets the recommendation is the argmax of the written scores in 99.3 to 99.6% of dialogues (rounding accounts for the rest).

Circuit. Before every query and before the final guess: `<belief>b1, ..., b24</belief>` with the teacher's posterior marginal that each of the 24 truth-table bits equals one, two decimals, followed by `<query>` or `<guess>`; the final instruction states that a guess bit is one exactly when its belief entry exceeds 0.50. The recorded teacher traces store the teacher's actions but not its beliefs, so the marginals are recomputed by replaying each teacher's own inference on its recorded interactions

and checked against the recorded per-turn entropy, mass on the true table and number of consistent tables. The guess agrees with the written final belief in every ELBO+EIG and Exact Bayes+EIG dialogue and in 62% of the ELBO+Random dialogues: that teacher identifies the circuit in only 56% of training episodes and submits a posterior sample, which we leave unrepaired.

Word. From the second attempt on: `<feedback>S grey, P grey, I grey, T grey, E green</feedback>` pairing each letter of the previous guess with its color; `<check>` listing every candidate of the previous `<candidates>` line, spelled letter by letter, with `ok` or the first violated constraint (`E green@5 -> out`); `<candidates>` listing the survivors spelled letter by letter; then `<attempt>`. The first turn writes the full vocabulary as candidates. All lines are compiled deterministically from the frozen teacher traces, so the scratchpad is exact by construction.

D.5 Scratchpad Designs That Did Not Work

The first word-guess scratchpad listed the surviving candidates as whole words, without the feedback and check lines. Its student imitated the first guess and the format perfectly but wrote a wrong candidate list from the second turn on (equal to the exact set on 2 of 98 second turns; containing the secret in 16), then committed to the last surviving word and repeated it to the budget: 31.9 attempts and 22% solved on 100 test episodes, against 2.29 for its teacher. The written list was a coarse function of the vocabulary and the color counts, not of letter positions, which Llama's tokenizer never fully exposes (`ENACT` is `EN|ACT`, `FROST` is `F|RO|ST`). The letter-spelled format of Appendix D.4 fixed this (3.09 attempts, 98% solved on the same 100 episodes). The 100 episodes are part of the 400 reported in the paper; the format was fixed once, before the remaining 300 episodes were evaluated, and was run without further change; the complete test split and the 400 held-out synthetic episodes agree (3.37 against 3.37 attempts).

D.6 LLM-Teacher Baseline

Llama-3.1-70B-Instruct (bf16, tensor parallel 4, vLLM (Kwon et al., 2023)) is rolled out greedily on the same training and validation episodes as the Bayesian teachers, under the same prompts, with a token-trie logits processor that admits only legal actions (unasked legal queries; a valid 24-bit guess; legal pairs of seen films; unseen titles), so its interactions are always well-formed. Its trace quality is measured with the same reference posterior as everything else. Circuit (3,397 training episodes): Success@18 0.002, bitwise accuracy 0.686, EIG-optimal 0.300, regret 0.340 bits; it identifies the circuit by turn 18 in 45% of episodes (0.19 GPU-hours on four A100s). Movie (3,600 episodes): Success@10 0.078, normalized rank 0.409, reference-EIG-optimal 0.002, regret 0.269 nats; it opens with the same pair in 38% of episodes. Its action-only dialogues are compiled and its students trained with the identical recipe; their validation losses are the lowest of all students (0.016 circuit, 0.137 movie), because a greedy language model is the most predictable teacher, not the best one.

D.7 Training Configuration and Compute

Student	Llama-3.1-8B-Instruct, bf16, SDPA attention, gradient checkpointing
Adapters	LoRA $r = 32$, $\alpha = 64$, dropout 0.05 on q, k, v, o, gate, up, down projections
Objective	next-token cross-entropy on assistant tokens (system, user and answer tokens masked)
Optimizer	AdamW (Loshchilov & Hutter, 2019) (fused), learning rate 2×10^{-4}, cosine schedule, 3% warm-up, weight decay 0.01
Batch	2 dialogues per GPU $\times$ 2 GPUs $\times$ 4 accumulation steps $=$ 16 dialogues per update
Epochs	3; checkpoint with the lowest validation loss on the synthetic validation split
Sequence length	2,048 (circuit action-only), 4,096 (circuit scratchpad; movie action-only; word), 6,144 (movie scratchpad); no truncation
Seed	0 for every condition
Wall time (2×A100)	circuit action-only 32 min, scratchpad 1.9 h; movie action-only 2.6 h, scratchpad 3.9 h; word action-only 26 min, scratchpad 47 min

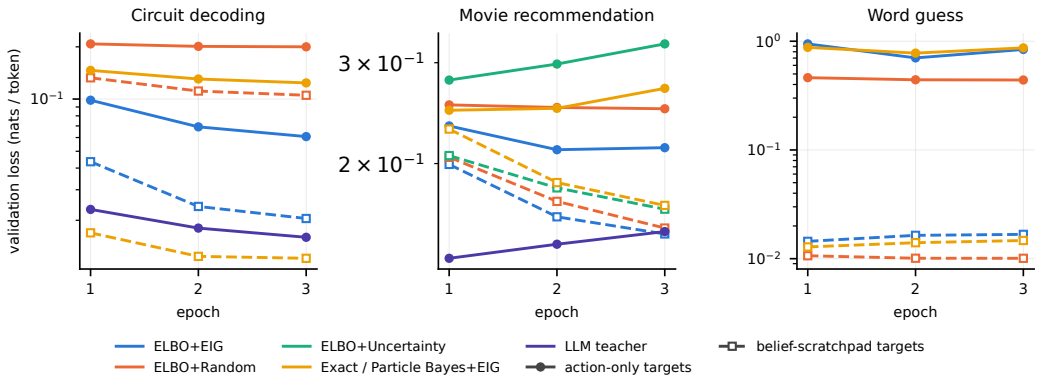

Figure 4: Validation loss per epoch (synthetic validation split, assistant tokens) of every distilled student, one panel per task. Solid lines with filled markers: action-only targets; dashed lines with open markers: belief-scratchpad targets. Validation loss measures how predictable a teacher's interactions are, not how good they are: the LLM teacher students have the lowest loss and the worst benchmark scores, the action-only exact-inference circuit student has twice the loss of the corresponding variational student, and the word action-only students' losses are more than an order of magnitude above the scratchpad students', whose targets contain the deterministic bookkeeping. Losses are not comparable across target formats within a task (different supervised tokens). The checkpoint with the lowest validation loss is the one evaluated.

With adapter parameters ψ and a dialogue $x_{1:L}$ whose assistant-token positions form the set $\mathcal{T}$, the objective is the next-token cross-entropy on those positions only,

$$\mathcal{L}_{\mathrm{SFT}}(\psi) = -\sum_{\text{dialogues}} \sum_{t \in \mathcal{T}} \log p_\psi(x_t \mid x_{<t}), \tag{4}$$

averaged over the assistant tokens of a batch; system, user and answer tokens receive no loss. Validation losses (epochs 1/2/3): circuit ELBO+EIG 0.099/0.069/0.061, ELBO+Random 0.208/0.201/0.200, Exact Bayes+EIG 0.146/0.131/0.124, LLM teacher best 0.016; circuit scratchpad ELBO+EIG 0.044/0.024/0.020, ELBO+Random 0.133/0.111/0.105, Exact Bayes+EIG 0.017/0.012/0.012; movie action-only ELBO+EIG 0.233/0.211/0.213, scratchpad ELBO+EIG 0.199/0.161/0.151; word scratchpad ELBO+EIG 0.0144/0.0164/0.0167 (best at epoch 1), word action-only ELBO+EIG 0.945/0.702/0.840, ELBO+Random 0.462/0.442/0.440, Exact Bayes+EIG 0.878/0.777/0.869. Validation loss orders teachers by the predictability of their action sequences, not by their quality.

Figure 4 shows the validation loss of every student per epoch.

E Evaluation Protocol and Statistical Methods

E.1 Protocol Specification and Integrity Checks

Each evaluator is a single program per task that loads the base model and, for student conditions, one LoRA adapter; runs the deterministic respondent; uses the frozen prompt template of the condition's target format, greedy decoding (temperature 0), and fixed per-turn token budgets (circuit 32 per query and 64 for the guess for action-only, 224 and 256 for scratchpad; movie 48 and 32 for action-only, 96 and 640 for scratchpad; word 256 for action-only, 1,024 for scratchpad). Before any episode runs, the program verifies SHA-256 hashes of the test manifest, the prompt templates, the adapters and the frozen training data, and aborts on any mismatch; a four-episode smoke run precedes the full run. Runs are resumable per episode and outputs are immutable once complete. Where a condition's episodes had already been evaluated by an earlier run with byte-identical evaluator code, they were reused after row-for-row verification and a fresh in-run reproduction of the smoke episodes.

E.2 Handling of Invalid, Repeated and Malformed Outputs

As in Appendix B.4. Outputs are parsed by a strict parser: exactly one tag of the expected kind; for scratchpad conditions the parser skips well-formed scratchpad blocks and parses the tagged action after them, so a malformed scratchpad does not itself invalidate a turn.

E.3 Definitions of the Reported Metrics

Notation follows Appendix A: an episode has true state θ^*, query turns $t = 1, \dots, T$ with the model's action $\hat{a}_t$ and answer y_t, history h_t and terminal decision d. Unless stated otherwise, a rate is the mean of a per-turn or per-episode indicator over all episodes of a condition.

Task success and benchmark scores. On Circuit Decoding $d \in \{0,1\}^{24}$ is the submitted table and $d^{(c)}$, $\theta^{*(c)}$ are the eight-bit blocks of circuit $c \in \{A, B, C\}$; a malformed guess counts as all bits wrong:

$$\begin{aligned} \text{Success@18} &= \mathbf{1}[d = \theta^*], \qquad \text{bit acc.} = \frac{1}{24}\sum_{i=1}^{24} \mathbf{1}[d_i = \theta^*_i], \\ \text{circuit-wise acc.} &= \frac{1}{3}\sum_{c\in\{A,B,C\}} \mathbf{1}\big[d^{(c)} = \theta^{*(c)}\big], \end{aligned} \tag{5}$$

and the benchmark score is $100\times$ the mean circuit-wise accuracy. On Movie Recommendation, with $u(i) = \theta^{*\top}\phi_i$ over the 40 unseen films and d the recommended film,

$$\begin{aligned} \text{rank}(d) &= 1 + \big|\{i :\ u(i) > u(d)\}\big|, \qquad \text{normalized rank} = \frac{\text{rank}(d) - 1}{39}, \\ \text{Success@10} &= \mathbf{1}\big[u(d) = \max_i u(i)\big], \end{aligned} \tag{6}$$

so tied films share the best rank and any optimal film counts; the benchmark score is $100 \times (1 -$ mean normalized rank), and a malformed recommendation has normalized rank 1 and success 0. On Word Guess, N is the number of attempts up to and including the correct one, invalid and repeated attempts included, with $N = 40$ for an unsolved episode; mean attempts is the mean of N and Success@$k = \mathbf{1}[N \le k]$.

Reference posterior. On the model's own history, and without ever showing it to the model, the evaluator maintains

$$b_t(\theta) \ \propto\ p(\theta)\prod_{s\le t} p(y_s \mid \theta, \hat{a}_s) \tag{7}$$

over the turns with a legal action, with the task's deterministic likelihood and uniform prior: exactly, over the K grammar-enumerated tables (circuit) or the 40 words (word), where b_t is uniform on the *version space* $V_t = \{\theta :\ p(y_s \mid \theta, \hat{a}_s) = 1 \text{ for all } s \le t\}$ (the consistent set C_t on Word Guess); and on Movie Recommendation by the particle posterior of Appendix C.5, $N_p = 200{,}000$ particles $\theta^{(n)}$ drawn uniformly from the grid $\{0, 0.1, \dots, 0.9\}^8$ with seed 0 and weights

$$w_t^{(n)} \ \propto\ \prod_{s\le t} \sigma\Big(z_s\,\frac{u^{(n)}(i_s) - u^{(n)}(j_s)}{\tau}\Big), \qquad \tau = 0.1, \tag{8}$$

where $u^{(n)}(i) = \theta^{(n)\top}\phi_i$, (i_s, j_s) is the pair asked at turn s and $z_s = +1$ for Yes and -1 for No; No Preference answers and invalid turns contribute no factor, and the answer to a repeated pair is applied once more.

Reference EIG, EIG-optimal rate and EIG regret. At turn t the admissible set $\mathbb{A}_t$ of Eq. 2 is the set of unasked queries (circuit), unguessed words (word) or all 190 pairs (movie). The reference EIG of $a \in \mathbb{A}_t$ is Eq. 2 evaluated under b_{t-1} instead of q_{t-1},

$$\text{EIG}_t^{\text{ref}}(a) = \text{H}\big[b_{t-1}(y \mid a)\big] - \mathbb{E}_{b_{t-1}(\theta)}\text{H}\big[p(y \mid \theta, a)\big], \qquad b_{t-1}(y \mid a) = \mathbb{E}_{b_{t-1}(\theta)}\, p(y \mid \theta, a), \tag{9}$$

which for a deterministic likelihood is the entropy of the partition of b_{t-1} induced by the possible answers: the binary entropy of the predicted bit, in bits, on Circuit Decoding, and the entropy over

the feedback patterns, in nats, on Word Guess; on Movie Recommendation it is evaluated under the particle weights, in nats. With $\mathrm{EIG}_t^* = \max_{a \in \mathbb{A}_t} \mathrm{EIG}_t^{\mathrm{ref}}(a)$,

$$o_t = \mathbf{1}\big[\mathrm{EIG}_t^{\mathrm{ref}}(\hat{a}_t) \geq \mathrm{EIG}_t^* - \epsilon\big], \qquad r_t = \mathrm{EIG}_t^* - \mathrm{EIG}_t^{\mathrm{ref}}(\hat{a}_t), \tag{10}$$

with $\epsilon = 10^{-9}$ bits (circuit), 10^{-9} nats (word) and 10^{-6} nats (movie), so that ties count as optimal. The *EIG-optimal rate* is the mean of o_t and the *EIG regret* the mean of r_t over the counted query turns of all episodes. Wasted turns: on Circuit Decoding an invalid or repeated query is scored with $\mathrm{EIG}_t^{\mathrm{ref}}(\hat{a}_t) = 0$, so $o_t = 0$ and $r_t = \mathrm{EIG}_t^*$; on Movie Recommendation an invalid question is scored the same way and a repeated pair like any legal pair under b_{t-1}; on Word Guess invalid and repeated attempts are excluded from both means (they enter N).

Identification and posterior mass. The posterior mass on the true state is $b_t(\theta^*)$. A history identifies the state when the version space is a singleton,

$$\begin{aligned} \text{identified@}T &= \mathbf{1}\big[|V_T| = 1\big] \ \text{(circuit, word)}, \\ \text{turns to identify} &= \min\{t : \theta_s^{\mathrm{MAP}} = \theta^* \text{ for all } s \geq t\}, \end{aligned} \tag{11}$$

where the second quantity, reported for the circuit teachers, is the first turn after which the teacher's maximum-a-posteriori table θ_s^{MAP} stays correct through the last of its 24 diagnostic queries (25 if it never does). On Movie Recommendation the analogous quantities use the mass on the true best film,

$$p_{\mathrm{best},t} = \sum_n w_t^{(n)}\, \mathbf{1}\Big[\arg\max_i u^{(n)}(i) \in B^*\Big], \qquad \text{identified@10} = \mathbf{1}\big[p_{\mathrm{best},10} \geq 0.9\big], \tag{12}$$

with B^* the set of unseen films attaining $\max_i u(i)$. The *oracle read-out* of Appendix F is the posterior mode of the best film under b_{10} (movie) and, on Word Guess, the indicator that $|C_t| = 1$ for some $t \leq k$.

Interaction quality. The *invalid rate* is the fraction of query turns whose output fails the strict parser or names an illegal action, and the *repeat rate* the fraction whose output is a legal action already taken in the episode.

Belief and read-out fidelity (scratchpad students). A scratchpad is *well formed* when its belief block parses under the strict parser. On Circuit Decoding the written marginals are $\hat{b}_t \in [0,1]^{24}$ and

$$\begin{aligned} \mathrm{MAE}_t &= \frac{1}{24}\sum_{i=1}^{24} \big|\hat{b}_{t,i} - b_t(\theta_i{=}1)\big|, \\ \text{guess consistent} &= \mathbf{1}\Big[d_i = \mathbf{1}[\hat{b}_{T,i} > 0.5] \text{ for all } i \text{ with } |\hat{b}_{T,i} - 0.5| > 0.005\Big], \end{aligned} \tag{13}$$

where $b_t(\theta_i{=}1)$ is the exact marginal of bit i under the reference posterior given the student's own queries. On Movie Recommendation, with written weights $\hat{\theta}_t$ and written scores $\hat{s}(i)$ of the unseen films: recommendation consistent $= \mathbf{1}[d = \arg\max_i \hat{s}(i)]$; written belief optimal $= \mathbf{1}[\arg\max_i \hat{\theta}_T^\top \phi_i \in B^*]$; distance to the reference $= \|\hat{\theta}_T - \mathbb{E}_{b_T}[\theta]\|_2$ and to the truth $= \|\hat{\theta}_T - \theta^*\|_2$. On Word Guess, for turns $t \geq 2$ with written candidate list $\hat{C}_t$: candidates exact $= \mathbf{1}[\hat{C}_t = C_t]$; secret retained $= \mathbf{1}[\theta^* \in \hat{C}_t]$; feedback line correct $= 1$ when the written (letter, color) pairs equal the true feedback of the previous attempt; check verdicts correct $=$ the fraction of listed candidates whose written verdict (kept or eliminated) agrees with the two-pass feedback rule applied to the previous attempt.

E.4 Statistical Tests and Confidence Intervals

All comparisons are paired on identical episodes. Binary metrics: exact binomial McNemar test on discordant pairs; paired bootstrap of the mean difference with 10,000 resamples of episodes (seed 0), reporting the 2.5 and 97.5 percentiles. Attempts: two-sided Wilcoxon signed-rank test on non-zero differences and the same paired bootstrap; a cluster bootstrap resamples the 10 vocabularies. Per-condition CIs of a proportion are Wilson intervals or the same bootstrap.

E.5 Validation Checks Prior to Benchmark Evaluation

The checks against rules fixed in advance that preceded a benchmark run were run on 200 held-out synthetic episodes for the belief-scratchpad students of Movie Recommendation and Circuit Decoding and on the first 100 test episodes of Appendix D.5 for the ELBO+EIG letter-scratchpad student of Word Guess; the circuit action-only students had teacher-level and smoke checks only. No other student was checked against a pre-specified rule before its benchmark run; every run had the integrity and smoke checks of Appendix E. Movie (200 held-out episodes, rule: paired gain of the scratchpad student over its action-only counterpart $\geq +0.10$ and recommendation equal to the argmax of the written scores in $\geq 80\%$ of episodes): ELBO+EIG 0.320 against 0.105 ($+0.215$ [$+0.140$, $+0.290$]), ELBO+Random 0.230 against 0.130, ELBO+Uncertainty 0.235 against 0.080, Particle Bayes+EIG 0.415 against 0.110; the subsequent benchmark values 0.345 / 0.176 / 0.194 / 0.409 preserve the ordering, and each lies inside or next to the held-out interval. Word (rule applied on the first 100 test episodes, the same 100 on which the scratchpad format was fixed, Appendix D.5; rule: mean attempts ≤ 6 and exact candidate tracking on turns ≥ 2 in $\geq 80\%$): the letter-scratchpad student passed the first criterion (3.09 attempts, 98 of 100 solved) and the second on non-repeated turns (0.81; 0.53 when the repeated turns of the two unsolved episodes are pooled in), which we report as written. Circuit, action-only students: the teacher-ordering check on the synthetic episodes (ELBO+EIG $\approx$ Exact Bayes+EIG $\gg$ ELBO+Random) and the training smoke checks were the only pre-benchmark rules. Circuit, belief-scratchpad students (200 held-out episodes, all seven conditions on the same episodes; rule for the ELBO+EIG pair, fixed before any of these students was trained: a well-formed belief on $\geq 95\%$ of query turns, a parsable guess in $\geq 95\%$ of episodes, and Success@18 no more than 0.10 below the action-only student): ELBO+EIG 0.910 against 0.920 for its action-only counterpart (-0.010 [-0.055, $+0.035$]), ELBO+Random 0.530 against 0.485 ($+0.045$ [-0.055, $+0.150$]), Exact Bayes+EIG 0.735 against 0.625 ($+0.110$ [$+0.015$, $+0.205$], McNemar $p = 0.030$); beliefs were well formed on at least 99.9% of query turns and the written final marginals lie within a mean absolute error of 0.007, 0.034 and 0.015 per bit of the exact posterior given the student's own queries. The action-only students approximately reproduce their benchmark values on these episodes (0.920, 0.485, 0.625 against 0.920, 0.500, 0.643). Word students were additionally evaluated on all 400 held-out synthetic episodes after the benchmark run (Appendix F.3).

E.6 Base-Model Behavior under Each Prompt Format

Circuit: 493 unparsable queries, 97 of 300 malformed final guesses, bitwise accuracy 0.380, EIG-optimal 0.239; the base's queries happen to identify the circuit in 33% of episodes and it never reads the table out. Movie, action-only prompt: every question is invalid (untagged or malformed); the recommendation parses in 100% of episodes and is near chance (0.032; normalized rank 0.496). Movie, scratchpad prompt: the base copies the prior belief line, writes an untagged question and an all-zero scores line without a tagged recommendation, and scores 0 by construction. Word: 17% invalid attempts (six-letter words, non-words), 58% repeats, 32.0 attempts, yet its guesses isolate the secret within three attempts under the exact consistency check in 80% of episodes: the base fails more at reading the feedback out than at choosing informative words.

F Additional Distillation Results

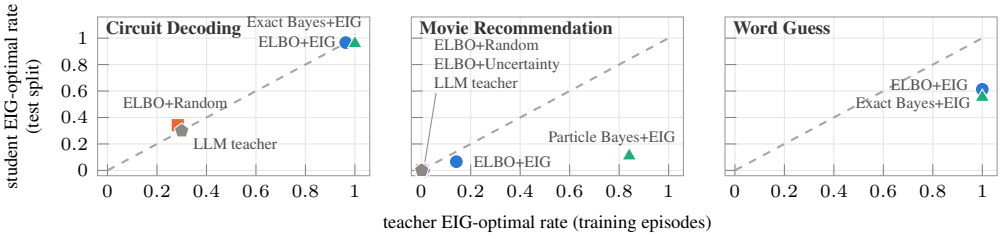

Figure 5: Students inherit their teacher's information-seeking policy on Circuit Decoding and only partly on Movie Recommendation and Word Guess: the teacher's EIG-optimal rate on the training episodes against its student's on the complete test split, one panel per task, one labeled point per teacher (belief-scratchpad students; the LLM teacher's students are action-only).

F.1 CIRCUIT DECODING

Condition	Success@18	bit acc.	final parse	valid query	EIG-opt.	regret	identified@18	val. loss
Base	0.000	0.380	0.677	0.909	0.239	0.355	0.330	–
ELBO+Random student	0.500	0.955	1.000	1.000	0.310	0.310	0.567	0.200
ELBO+EIG student	**0.920**	**0.996**	1.000	1.000	**0.954**	**0.014**	**0.997**	0.061
Exact Bayes+EIG student	0.643	0.974	1.000	1.000	0.896	0.066	0.897	0.124
LLM teacher student	0.000	0.619	1.000	1.000	0.300	0.340	–	0.016
ELBO+Random student (B)	0.590	0.972	1.000	1.000	0.342	0.291	0.713	0.105
ELBO+EIG student (B)	**0.927**	**0.996**	1.000	0.996	**0.966**	**0.008**	**0.993**	0.020
Exact Bayes+EIG student (B)	0.840	0.989	1.000	1.000	0.959	0.027	0.963	0.012

Rows marked (B) are the belief-scratchpad students and all other rows are action-only students; validation losses are not comparable across the two formats because the scratchpad tokens are supervised. The paragraphs below describe the action-only students unless stated otherwise. Per inventory ($3/3/0$, $3/3/1$, $1/5/2$), Success@18: ELBO+EIG 1.00 / 0.84 / 0.92; Exact Bayes+EIG 0.85 / 0.52 / 0.56; ELBO+Random 0.56 / 0.42 / 0.52; LLM teacher student 0 / 0 / 0. **Final-answer errors.** ELBO+EIG student: 24 failures, 23 after the version space had collapsed to one table (off by one bit in 17, two in 5, three in 1) and 1 with two hypotheses left; regret 0.041 in failures against 0.012 in successes. Exact Bayes+EIG student: 107 failures, 89 after collapse (36 off by one bit, 45 by two, 8 by three or four), 18 with two hypotheses left; queries are exactly optimal on turns 1–5 and 0.80–0.95 optimal afterward. ELBO+Random student: 150 failures, 93 with at least two consistent hypotheses left at the final guess (77 with exactly two) and 57 read-out errors after collapse. Opening queries are deterministic per inventory under greedy decoding (the prompt differs only by inventory before the first query). **Complete paired table** (same 300 episodes; first-named minus second-named condition; discordant counts first-only / second-only): ELBO+EIG vs. ELBO+Random $+0.420$ $[+0.350, +0.487]$, $143/17$, $p = 5.3 \times 10^{-26}$; Exact Bayes+EIG vs. ELBO+EIG -0.277 $[-0.333, -0.217]$, $9/92$, $p = 1.8 \times 10^{-18}$; Exact Bayes+EIG vs. ELBO+Random $+0.143$ $[+0.067, +0.220]$, $94/51$, $p = 4.5 \times 10^{-4}$; ELBO+EIG vs. base $+0.920$ $[+0.887, +0.950]$, $276/0$; every student vs. base: no discordant pair in the base's favor; LLM teacher student vs. base $0/0$, $p = 1$. **Inheritance of the query policy**: EIG-optimal rate of the teacher on the training episodes against its action-only student's on the test split, 0.963 / 0.954 (ELBO+EIG), 0.284 / 0.310 (ELBO+Random), 1.000 / 0.896 (Exact Bayes+EIG) and 0.300 / 0.300 (LLM teacher).

Belief-scratchpad students (same 300 episodes, evaluated in a separate run and paired by episode). Per inventory, Success@18: ELBO+EIG 1.00 / 0.89 / 0.89; Exact Bayes+EIG 1.00 / 0.70 / 0.82; ELBO+Random 0.79 / 0.46 / 0.52. Format contrasts, scratchpad minus action-only (discordant counts, solved only by the scratchpad student / only by the action-only student): ELBO+EIG $+0.007$ $[-0.037, +0.050]$, $22/20$, $p = 0.88$; ELBO+Random $+0.090$ $[+0.010, +0.170]$, $89/62$, $p = 0.034$; Exact Bayes+EIG $+0.197$ $[+0.133, +0.257]$, $82/23$, $p = 6.0 \times 10^{-9}$. Among the scratchpad students: ELBO+EIG vs. ELBO+Random $+0.337$ $[+0.277, +0.393]$, $109/8$, $p = 8.9 \times 10^{-24}$; Exact Bayes+EIG vs. ELBO+EIG -0.087 $[-0.137, -0.037]$, $16/42$, $p = 8.6 \times 10^{-4}$; Exact Bayes+EIG vs. ELBO+Random $+0.250$ $[+0.187, +0.313]$, $95/20$, $p = 7.1 \times 10^{-13}$; every student vs. base: no discordant pair in the base's favor. The written belief is well formed on at least 99.8% of query turns and 99.7% of final turns, the written final marginals lie within a mean absolute error of 0.005, 0.028 and 0.011 per bit of the exact posterior given the student's own queries (ELBO+EIG, ELBO+Random, Exact Bayes+EIG), and the guess equals the rounded written belief in 97%, 76% and 100% of episodes. The final-answer errors keep their character. ELBO+EIG student: 22 failures, 20 after the version space had collapsed (14 off by one bit). Exact Bayes+EIG student: 48 failures against 107 with action-only targets, 40 after collapse (27 off by one bit) and 8 with two hypotheses left; its queries are exactly optimal on turns 1–5 and 0.89 optimal afterward. ELBO+Random student: 123 failures, 64 with at least two consistent hypotheses left (53 with exactly two) and 59 read-out errors after collapse. The scratchpad makes the read-out error visible: in all 40 after-collapse failures of the Exact Bayes+EIG student and in 15 of the 20 for the ELBO+EIG student, the erroneous bit is already wrong in the written final belief, which the guess then copies (40 and 12 episodes). The only invalid or repeated queries of any circuit student are 22 of the 5,400 turns of the ELBO+EIG scratchpad student, all at turn 14 or later and after the circuit had been identified. Inheritance of the query policy (the circuit points of Figure 5), teacher on the training episodes / scratchpad student on the test split: 0.963 / 0.966 (ELBO+EIG), 0.284 / 0.342 (ELBO+Random), 1.000 / 0.959 (Exact Bayes+EIG).

F.2 MOVIE RECOMMENDATION

All students on the test split.

Target	Teacher	Success@10 [95% CI]	rank	bench. %	repeat	pairs	EIG-opt.	regret	p(best)
–	Base	0.032 [0.022, 0.043]	0.496	50.4	0.000	0.00	0.000	0.554	0.075
A	ELBO+EIG	0.085 [0.068, 0.103]	0.396	60.4	0.509	4.91	0.025	0.279	0.172
A	ELBO+Random	0.105 [0.086, 0.124]	0.390	61.0	0.801	1.99	0.000	0.408	0.112
A	ELBO+Uncertainty	0.064 [0.049, 0.080]	0.423	57.7	0.583	4.17	0.000	0.331	0.156
A	Particle Bayes+EIG	0.079 [0.063, 0.096]	0.398	60.2	0.040	9.43	0.024	0.180	0.263
A	LLM teacher	0.052 [0.039, 0.066]	0.433	56.7	0.005	9.91	0.001	0.276	0.203
B	ELBO+EIG	0.345 [0.316, 0.375]	0.100	90.0	0.301	6.99	0.066	0.162	0.247
B	ELBO+Random	0.176 [0.153, 0.200]	0.208	79.2	0.798	2.02	0.000	0.399	0.118
B	ELBO+Uncertainty	0.194 [0.170, 0.219]	0.208	79.2	0.587	4.13	0.000	0.332	0.159
B	Particle Bayes+EIG	**0.409** [0.379, 0.439]	**0.069**	**93.1**	0.044	9.49	0.110	0.078	0.317

Action-only contrasts (paired, 1,000 episodes): ELBO+EIG vs. ELBO+Random -0.020 $[-0.038, -0.003]$ ($p = 0.035$); ELBO+EIG vs. ELBO+Uncertainty $+0.021$ ($p = 0.028$); Particle Bayes+EIG vs. ELBO+EIG -0.006 ($p = 0.65$); LLM teacher vs. ELBO+EIG -0.033 $[-0.055, -0.011]$ ($p = 4.7 \times 10^{-3}$). **Scratchpad against action-only on identical episodes**: ELBO+EIG $+0.260$ $[+0.228, +0.292]$ ($300/40$ discordant, $p = 2.1 \times 10^{-50}$); ELBO+Random $+0.071$ $[+0.044, +0.098]$; ELBO+Uncertainty $+0.130$ $[+0.103, +0.157]$; Particle Bayes+EIG $+0.330$ $[+0.296, +0.362]$ ($365/35$, $p = 2.1 \times 10^{-70}$); against the action-only base: $+0.313$, $+0.144$, $+0.162$, $+0.377$ $[+0.345, +0.409]$.

Read-out decomposition (Section 5.3). A student is given a teacher's ten questions and answers and produces only the recommendation; the oracle read-out is the posterior mode of the best unseen film under the reference posterior on the same transcript. Own questions + own read-out / own + oracle / teacher questions + own read-out / teacher + oracle / teacher + teacher read-out: ELBO+EIG 0.085 / 0.305 / 0.085 / 0.473 / 0.410; ELBO+Random 0.105 / 0.195 / 0.108 / 0.413 / 0.373; ELBO+Uncertainty 0.064 / 0.262 / 0.060 / 0.485 / 0.428; Particle Bayes+EIG 0.079 / 0.448 / 0.077 / 0.571 / 0.562; LLM teacher student 0.052 / 0.330 / 0.048 / 0.571 / 0.562 (particle prefix). Under an oracle read-out the teacher ordering re-emerges. **Attribute-permutation probe** (200 test episodes): the attribute rows of the 40 unseen films are permuted among the unchanged titles; a utility-reading model follows the attributes to their new title, whereas a title-reading model keeps the title. Action-only students keep the title in 7–25% of episodes (base 38–42%), follow the attributes in 4–22%, and move to a third title in 66–81%; success stays at chance under the permutation. **Belief fidelity** (scratchpad students): belief and scores lines parse on $\geq 98\%$ of turns; recommendation $=$ argmax of written scores in 0.976 / 0.999 / 0.930 / 0.966 of episodes; argmax of written belief is optimal in 0.338 / 0.177 / 0.201 / 0.414 (Success@10 0.345 / 0.176 / 0.194 / 0.409); Euclidean distance of the written final belief to the reference posterior mean 0.210 / 0.087 / 0.351 / 0.135, to the true weights 0.633 / 0.748 / 0.785 / 0.532 (reference mean to true weights 0.591 / 0.739 / 0.683 / 0.506). **Repeated questions**: no movie teacher excludes asked pairs (teacher repeat rates 0.014–0.064, Appendix C.7), yet the ELBO+Random and ELBO+Uncertainty students repeat 58–80% of their questions under both formats, the ELBO+EIG student 30–51% and the particle student 4%.

F.3 WORD GUESS

Test split (400 episodes) and **held-out synthetic episodes** (400).

Episodes	Condition	attempts [95% CI]	median	S@2	S@3	S@5	unsolved	invalid	repeat
test	Base	31.97 [30.56, 33.36]	40	0.043	0.060	0.120	300	0.172	0.577
test	ELBO+EIG A	19.16 [17.52, 20.84]	9	0.077	0.142	0.318	155	0.047	0.560
test	ELBO+Random A	15.23 [13.94, 16.56]	9	0.075	0.130	0.280	79	0.018	0.243
test	Exact Bayes+EIG A	15.37 [13.97, 16.83]	8	0.085	0.155	0.312	98	0.014	0.432
test	ELBO+EIG B	3.37 [2.87, 3.95]	2	0.505	**0.917**	0.975	9	0.028	0.221
test	ELBO+Random B	**3.15** [3.03, 3.27]	3	0.295	0.713	0.963	**0**	0.000	0.005
test	Exact Bayes+EIG B	**3.10** [2.69, 3.58]	2	0.507	**0.917**	0.983	6	0.045	0.137
synthetic	ELBO+EIG A	17.69 [16.13, 19.28]	9	0.083	0.158	0.273	134	0.007	0.534
synthetic	ELBO+Random A	14.92 [13.67, 16.15]	10	0.068	0.125	0.245	71	0.006	0.188
synthetic	Exact Bayes+EIG A	15.89 [14.45, 17.37]	9	0.087	0.155	0.318	105	0.010	0.427
synthetic	ELBO+EIG B	3.37 [2.86, 3.97]	2	0.522	0.917	0.970	9	0.013	0.232
synthetic	ELBO+Random B	3.78 [3.27, 4.36]	3	0.357	0.750	0.958	9	0.000	0.237
synthetic	Exact Bayes+EIG B	3.00 [2.62, 3.45]	2	0.527	0.915	0.988	5	0.033	0.129

Teachers on the same episodes: 2.29 / 3.09 / 2.29 attempts (test), 2.31 / 3.20 / 2.31 (synthetic). On solved episodes alone the EIG students take 2.53 and 2.54 attempts against 3.15 for ELBO+Random; the difference in the means is entirely the 40-attempt dead ends. **Paired contrasts** (test split): ELBO+EIG B vs. base -28.60 $[-30.09, -27.14]$, Wilcoxon $p = 2.1 \times 10^{-65}$; B vs. A for ELBO+EIG -15.78 $[-17.54, -14.04]$, Success@3 $313/3$ discordant, for ELBO+Random -12.08 $[-13.42, -10.78]$, $239/6$, and for Exact Bayes+EIG -12.27 $[-13.80, -10.82]$, $311/6$ (held-out synthetic episodes: -14.32, -11.14, -12.90); ELBO+EIG B vs. ELBO+Random B $+0.23$ $[-0.29, +0.81]$ attempts, Success@3 $103/21$, McNemar $p = 3.5 \times 10^{-14}$; Exact Bayes+EIG B vs. ELBO+EIG B -0.27 $[-0.98, +0.41]$, $p = 0.63$, Success@3 $23/23$. On the synthetic episodes the EIG-vs-random attempts contrast is -0.41 $[-1.15, +0.34]$ and Success@3 $90/23$ ($p = 1.5 \times 10^{-10}$): the sign of the mean flips because the random student dead-ends nine times in one vocabulary, while Success@3 tells the same story on both sets. **Scratchpad diagnostics** (test split, turns ≥ 2): candidates equal to the exact consistent set 0.723 / 1.000 / 0.843 (ELBO+EIG / ELBO+Random / Exact Bayes+EIG); secret retained 0.792 / 1.000 / 0.897; feedback line correct 0.997 / 1.000 / 0.995; check verdicts correct 0.993 / 1.000 / 0.992; exact-EIG-optimal rate over the counted attempts 0.613 / 0.471 / 0.553, against 1.000 for the two EIG teachers on their training episodes (Figure 5). **Unrecoverable episodes.** ELBO+EIG: 9 of 400; in seven the secret is dropped from the written list at turn two or three (a wrong verdict or a copying error between the check and candidates lines) and one legal guess is repeated for the rest of the budget; one is a read-out slip with a correct one-word list; one is a run of invalid attempts. Exact Bayes+EIG: 6 of 400 (four with the secret dropped, one read-out slip, one run of letter-spaced invalid attempts after the list emptied). ELBO+Random: none on the test split; nine on the synthetic episodes, all in one vocabulary, all with exact bookkeeping and a repeated opener (a read-out failure with a correct list). The teacher's traces never contain a wrong list, so the students have seen no example of a correction, and a dropped secret is rarely restored: on the test split it re-enters the written list in one of the fifteen episodes in which it is dropped, and three of those fifteen are solved. **Oracle read-out**: the secret is the only consistent word after at most three attempts in 0.995 / 0.892 / 1.000 of the ELBO+EIG / ELBO+Random / Exact Bayes+EIG scratchpad students' episodes, 0.960 / 0.910 / 0.935 for the action-only students and 0.795 for the base; of the action-only students' 155 / 79 / 98 unsolved episodes, 150 / 74 / 89 had isolated the secret. **Replicate runs.** The ELBO+EIG and Exact Bayes+EIG teachers choose identical guesses on every synthetic training and validation episode, so under either target format their two students are trained on identical dialogues with the identical recipe and seed; they differ only through non-deterministic GPU arithmetic. This is the only replicate pair in the paper. With belief-scratchpad targets the replicates differ by 0.27 attempts $[-0.98, +0.41]$ with identical Success@3 ($23/23$ discordant) and candidate-set accuracies of 0.72 and 0.84; with action-only targets they differ by 3.79 attempts $[+2.22, +5.26]$ (155 against 98 unsolved episodes). The action-only ELBO+EIG student is also 3.93 attempts $[+2.38, +5.46]$ worse than the action-only ELBO+Random student, a gap of the same size as the replicate gap, while the second replicate is not separated from the ELBO+Random student ($+0.14$ $[-1.26, +1.51]$; Success@3 $p \geq 0.28$ for all three pairs): among students that fail at the read-out, differences of a few attempts are within run-to-run variation, and we draw no conclusion about the query rule from them.

F.4 Example Interactions

Movie Recommendation, belief-scratchpad target of the ELBO+EIG teacher (training episode; the dialogue of Figure 1; abridged): the first turn writes the prior `<belief>0.45, ..., 0.45</belief>` and asks about the pair with the largest predicted disagreement; after the answer "No" the belief moves to `0.48, 0.42, 0.45, 0.44, 0.44, 0.45, 0.59, 0.32`; the final turn writes the belief and forty scores, and recommends the argmax. Word Guess, belief-scratchpad student: see the target format in Appendix D.4, which the student reproduces with the exact set on 72% of turns. Circuit Decoding, action-only student: eighteen valid distinct queries and a 24-bit guess in every episode.

G Cross-Task Generalization

G.1 Research Questions and Pre-Specified Decision Rules

Two questions were fixed before any cross-task student was trained. Q1: does the information-seeking *policy* distilled on two task families transfer zero-shot to the third (judged by reference-EIG regret,

identification rate and task success on the held-out task, paired against the base model on identical episodes)? Q2: does the information-gain *criterion* transfer, that is, can a student trained to rank candidate queries by their EIG on two tasks rank candidates on the third better than a shuffled-target control and the base model? The decision rule for Q2 required a positive result in at least two of the three directions, with confidence intervals excluding zero against both controls.

G.2 Leave-One-Task-Out Protocol

For each held-out task a student is trained on $1{,}800$ episodes from each of the two other tasks (frozen subsets of the training banks; $128/840/832$ circuit episodes per inventory, 10 users of each of 180 score tables, 20 secrets of each of 90 vocabularies) with the recipe of Section 3, using the belief-scratchpad targets for movie and word and the action-only targets for circuit, and a loss that weights the two source tasks equally per micro-batch. No data from the held-out task is read during training. Source-task checks evaluate every student on both of its source tasks.

G.3 Native and Unified Action Interfaces

A format-compatibility pilot evaluated each in-domain student on the two other tasks under those tasks' native prompts. The rule fixed in advance required the student's action to be parsable on at least 90% of turns for every student–task pair; four of six pairs passed (0.946–1.000) and the word student failed on circuit (0.772) and movie (0.289). To remove format as a confound, all cross-task results below use a *unified interface*: a canonical outer wrapper `<action>...</action>` around the task's native action, with a matching prompt for every task, under which every cross-task student emits a well-formed action on at least 99% of turns. The base model under the canonical prompts truncates its final circuit guess in 49% of episodes at the frozen final-token budget; this affects only the base's read-out metrics, not its policy metrics, and no budget was changed after seeing results.

G.4 Zero-Shot Transfer of the Information-Seeking Policy

Held-out task	Metric	Base	Cross-task student	In-domain student
Circuit (300)	Success@18 ↑	0.000	0.000	0.920
	EIG regret (bits) ↓	0.325	0.540	0.014
	identified@18 ↑	0.457	0.047	0.997
Word (400)	attempts ↓	30.55	29.96	3.37
	Success@3 ↑	0.092	0.075	0.917
	EIG regret (nats) ↓	0.104	0.118	0.167
Movie (1,000)	Success@10 ↑	0.022	0.028	0.345
	normalized rank ↓	0.595	0.518	0.100
	EIG regret (nats) ↓	0.248	0.240	0.162

Paired against the base on identical episodes: circuit regret $+0.215$ $[+0.196,\ +0.235]$ (worse), identified -0.410 $[-0.470,\ -0.350]$; word attempts -0.59 $[-2.13,\ +0.93]$ ($p = 0.51$), regret $+0.013$ (worse); movie Success@10 $+0.006$ $[-0.006,\ +0.018]$, regret -0.008 $[-0.013,\ -0.004]$ (3% of the base's). The cross-task students repeat actions on 39% (circuit) and 50% (word) of turns once the history exceeds the source tasks' horizon, and have no read-out for the held-out task. Source-task checks: on both of its sources every cross-task student is within 18 points of the in-domain student on Success@18 (circuit 0.797 and 0.747) and Success@10 (movie 0.307 and 0.292), and needs 2.87 and 4.48 attempts on Word Guess against the in-domain student's 3.37, so the held-out results are interpretable.

G.5 Transfer of the Information-Gain Criterion: Candidate-Ranking Formulation

Each ranking example presents a state (history under the unified interface) and eight legal candidate actions labeled with random letters; the target is the softmax of the candidates' reference EIG at temperature τ (*soft targets*, primary), the argmax (*hard targets*, variant), or the soft targets permuted within the candidate set (*shuffled*, the decisive control with identical compute). Students are trained jointly on the two source tasks' policy dialogues and ranking examples. Held-out banks are built from the test episodes only: $4{,}800$ (circuit), $6{,}138$ (word) and $16{,}000$ (movie) examples from EIG-greedy

and random rollouts, of which 3,660 / 1,746 / 16,000 have non-trivial EIG spread. Metrics on these contrast examples, for a state with candidate EIGs $g_1, \dots, g_8$ under the reference posterior, extremes $g_{\max} > g_{\min}$, the student's scores $z_1, \dots, z_8$ (the logits of the eight candidate letters at the answer position) and its choice $\hat{k} = \arg\max_k z_k$:

$$\text{normalized EIG regret} = \frac{g_{\max} - g_{\hat{k}}}{g_{\max} - g_{\min}} \ \text{(primary)}, \qquad \text{top-1} = \mathbf{1}\big[g_{\hat{k}} \geq g_{\max} - \epsilon\big],$$
$$\text{NDCG@8} = \frac{\sum_{r=1}^{8} \tilde{g}_{\pi(r)} / \log_2(r+1)}{\sum_{r=1}^{8} \tilde{g}_{(r)} / \log_2(r+1)}, \qquad \tilde{g}_k = \frac{g_k - g_{\min}}{g_{\max} - g_{\min}}, \tag{14}$$

where π orders the candidates by decreasing z, $\tilde{g}_{(r)}$ is the r-th largest normalized EIG and ϵ is the task's tie tolerance of Appendix E.3, so top-1 counts a tied optimum as correct. A uniform-random choice has expected normalized regret $\frac{1}{8}\sum_k (g_{\max} - g_k)/(g_{\max} - g_{\min})$, the reference line quoted with the results. Confidence intervals come from an episode-clustered paired bootstrap with 10,000 resamples.

G.6 Results

Held-out task (contrast examples)	Base	Shuffled	Soft	Hard	Policy student	In-domain
Circuit (3,660)	0.457	0.418	0.509	0.410	0.500	0.337
Word (1,746)	0.385	0.450	0.400	–	0.398	0.421
Movie (16,000)	0.457	0.448	0.458	–	0.445	0.437

Normalized EIG regret (lower is better; uniform random choice scores 0.42–0.45). Paired differences, soft-target student minus base and minus shuffled control: circuit $+0.052$ [$+0.033$, $+0.071$] and $+0.091$ [$+0.068$, $+0.114$] (worse); word $+0.014$ [-0.005, $+0.033$] and -0.050 [-0.072, -0.029]; movie $+0.002$ [-0.005, $+0.009$] and $+0.011$ [$+0.003$, $+0.018$]. Source checks pass in every direction (soft-target students reach regret 0.001–0.235 on their own sources against 0.395–0.474 for the base). Robustness variants (permuted letters, task name removed, history removed, anonymized candidates), run for the base model and the soft-target student in each direction, reveal no evidence of a shortcut: permuting the candidate letters or removing the task name changes regret by at most 0.02; removing the history changes it by at most 0.03 except for the circuit soft-target student ($0.509 \to 0.634$), which therefore relies on the history; and anonymized payloads bring both models to the uniform-random level (base / soft-target student against uniform random: circuit 0.421 / 0.423 vs. 0.419; word 0.448 / 0.461 vs. 0.442; movie 0.446 / 0.449 vs. 0.446). Zero of three directions meet the pre-specified rule.

G.7 Analysis of Negative Transfer on Circuit Decoding

Circuit Decoding is the only held-out cell with real dynamic range (the in-domain student is 0.120 below the base), and there transfer is negative: the soft-target student appears to carry a content-based preference for dense, novel inputs that suits movie and word openings but not circuit ones. At the first turn its chosen query sets 2.47 of the three input bits on average, against 1.40 for the highest-EIG candidate and 1.52 for a random one, with regret 0.59 against 0.37 for the base, whereas the hard-target and shuffled students choose inputs of random density (1.55 and 1.52 ones) and beat the base at the first turn (regret 0.29 and 0.30). On word the base already ranks well and even the in-domain student is below it; on movie every condition is within 0.02 of the base.

G.8 Deviations from the Pre-Specified Plan

(i) The switch from native to unified interfaces followed the failed compatibility pilot and was made before any held-out result under the unified interface was seen. (ii) The joint policy-plus-ranking objective replaced a ranking-only objective after a letter-only pilot never left the content-blind marginal; the change was decided before any held-out ranking result of the joint objective was seen. (iii) One seed per condition.

G.9 Discussion

Neither the policy nor its criterion transfers zero-shot across these three task families under two-source training. The result is consistent with the students learning task-specific procedures (which the

in-domain results show they learn well) rather than a portable notion of information value; whether broader task distributions or a shared belief language would change this is open.

H Extended Related Work

Information seeking with language models. Beyond the works cited in Section 6, falsification and rule-discovery benchmarks (Jhaveri et al., 2026) document a related failure, a tendency not to seek disconfirming evidence (confirmation bias), which an EIG rule counteracts by construction. **Bayesian experimental design with LMs.** Inference-time systems differ in what the LM supplies: a prior over hypotheses, a likelihood, or the candidate queries; our teacher requires none of these from the LM and is therefore verifiable, at the price of needing a task model. **Learning probabilistic reasoning from Bayesian models.** Qiu et al. (2025) distill the predictions of a Bayesian model and report transfer to new domains; our interactive setting adds the query rule and finds that, where the decision is an implicit computation beyond the recorded interaction, the belief must be written for the decision to be learned, and that transfer across our task families does not occur. **Distillation of algorithms and traces.** Algorithm distillation (Laskin et al., 2023) and search-trace training (Gandhi et al., 2024; Lehnert et al., 2024) teach transformers the dynamics of a procedure; our belief scratchpad is the analogous exposure of a Bayesian filter's state. **Amortized experimental design.** Policy-based designers (Foster et al., 2021; Ivanova et al., 2021; Blau et al., 2022) are trained by maximizing an EIG-based objective; our students are trained by imitation of an explicit designer, which makes the quality of the supervision measurable, and concurrent work (Hartmann et al., 2026; Deng et al., 2026) takes the reward-based route with an LM policy. **Reinforcement learning for active reasoning.** The closest reward-based work on belief tracking is that of Zou et al. (2026), who train Qwen2.5-7B-Instruct with outcome rewards (PPO, GRPO, GSPO) and truncate a training trajectory once a task-specific proxy indicates that the agent's belief has stopped improving, which stabilizes credit assignment. Three of their five tasks are adapted from the benchmark used here, but they are not the released tasks, so their numbers cannot be set against ours: Circuit Decoding becomes the identification of two hidden circuits from a list of ten candidate structures within ten turns, on which Gemini 2.5 Pro reaches 92% against 40% on the released task (Badola et al., 2025), and Movie Recommendation uses five questions and an exact-match score on a regenerated split; Word Guess is not included. The two routes are complementary: outcome reward with belief control needs no model of the task, whereas a Bayesian teacher needs one and in return supplies verifiable supervision for the query and for the belief.

I Reproducibility Details

I.1 Software, Hardware and Runtime

Teachers, generators and evaluators: Python 3.11, NumPy and PyTorch (CPU, teachers only). Training and evaluation: PyTorch 2.8 (CUDA 12.8), Transformers 4.57, PEFT 0.17, bf16, SDPA attention, on NVIDIA A100 80 GB GPUs; the LLM teacher with vLLM 0.11. Runtimes: teacher rollouts on the synthetic episodes take under one CPU-hour per teacher on circuit and word and about 4 CPU-hours on movie (3.3 s per episode, Appendix C.6; particle reference $\approx$ 66 CPU-hours); training as in Appendix D.7; evaluation 26 s per episode for circuit students, 10–27 s for action-only movie and word conditions, 120 s per episode for movie scratchpad students and 85–97 s for word scratchpad students.

I.2 Seeds and Determinism

Training seed 0 for every condition; teacher seeds 0–4 paired by episode where the teacher is stochastic; bootstrap seed 0. Greedy decoding on the same GPU type reproduced every re-run episode bit-exactly: $300/300$ circuit episodes for the base and action-only ELBO+EIG conditions evaluated twice, $100/100$ episodes in two independent replications of word-guess conditions, and the smoke episodes regenerated inside every reused evaluation (Appendix E). Training is not bit-reproducible: the two word-guess students trained on identical dialogues (Appendix F.3) end with validation losses of 0.0144 and 0.0128 (belief scratchpad) and 0.702 and 0.777 (action-only).

I.3 Licenses of Models and Datasets

Llama 3.1 models are used under the Llama 3.1 Community License; the benchmark is used under its release license. The word pool is a public list of five-letter English words.

J Prompts

J.1 System and Turn Prompts per Task

Circuit Decoding, system prompt.

```
You are playing Circuit Decoding, a multi-turn puzzle.

There are 3 hidden Boolean circuits, named A, B and C. Every circuit takes the
same 3 binary inputs (x0, x1, x2) and produces one output bit. The circuits are
built only from AND, OR and NOT gates. You are told only the total gate
inventory shared by the three circuits (how many AND, OR and NOT gates they use
in total); the wiring of each circuit is hidden.

Your goal is to recover the complete truth table of every circuit.

Rules:
1. On each turn you may ask exactly one query: the output of one circuit on one
   input assignment. Write it as <query>C(x0,x1,x2)</query>, where C is A, B or
   C and x0,x1,x2 are bits, for example <query>B(0,1,1)</query>.
2. The environment answers each query with a single bit, 0 or 1.
3. There are 24 possible queries (3 circuits x 8 input assignments). Each query
   may be asked at most once.
4. You have a budget of 18 queries. After the 18th answer you will be asked to
   submit your final answer.
5. On query turns respond with exactly one <query>...</query> tag and nothing else.
```

Circuit Decoding, final turn.

```
Your query budget is exhausted. Submit your final answer now.

Give the complete truth tables of circuits A, B and C as one 24-bit string
inside <guess>...</guess>: bits 1-8 are the outputs of circuit A on inputs
(x0,x1,x2) = 000, 001, 010, 011, 100, 101, 110, 111 in that order, bits 9-16
are the outputs of circuit B on the same inputs, and bits 17-24 are the outputs
of circuit C. For example, <guess>000000000000000000000000</guess> would mean
that every circuit outputs 0 on every input. Respond with exactly one
<guess>...</guess> tag and nothing else.
```

Movie Recommendation, system prompt (action-only).

```
You are playing Movie Recommendation, a multi-turn puzzle.

A user has a hidden preference function over 8 movie attributes:
score(movie) = w1*a1 + w2*a2 + ... + w8*a8, where a1..a8 are the movie's
attribute scores (given to you) and w1..w8 are the user's unknown non-negative
weights. You will be given two lists of movies with their attribute scores:
20 "seen" movies that you may ask the user about, and 40 "unseen" movies from
which you must finally recommend the one the user would score highest.

Rules:
1. On each turn ask exactly one question of the form <question>Would you prefer
   watching A over B?</question>, where A and B are two different movies from
   the seen list, written exactly as listed.
2. The user answers Yes, No or No Preference according to their hidden
   preference function.
3. You have a budget of 10 questions. After the 10th answer you will be asked
   for your recommendation.
4. On question turns respond with exactly one <question>...</question> tag and
   nothing else.
```

Movie Recommendation, final turn (action-only).

```
Your question budget is exhausted. Recommend now: reply with exactly one
<recommend>TITLE</recommend> tag containing the title of one movie from the
unseen list, written exactly as listed, and nothing else.
```

Word Guess, first user message (the vocabulary replaces `[VOCAB]`).

```
Let's play word guess. The secret word is from the following vocabulary: [VOCAB]
At each attempt, I will give feedback to you and specify which tiles are grey,
yellow and green. Green implies that the letter is in the word and in the
correct position, yellow implies that the letter is in the word but in the
wrong position and grey implies that the letter is not in the word.

For e.g., if the secret word is SPARK and you guessed PROOF, then the feedback
would be:
yellow,yellow,grey,grey,grey.

Your goal is to correctly guess the secret word in the minimum number of
attempts and will be evaluated on the number of attempts you take to arrive at
the secret word. Think step by step to ensure that you're taking the most
optimal actions to guess the secret word.

All of your guesses should be structured as follows:
<attempt>Your guess</attempt>

Make sure that all of your attempts are strictly from the vocabulary specified
above.
Start with your first attempt and make sure to follow the format specified above.
```

J.2 Action-Only and Belief-Scratchpad Templates

The movie belief-scratchpad prompt replaces rule 4 of the action-only system prompt with:

```
4. Before every question, write your current estimate of the user's weights as
   <belief>w1, w2, w3, w4, w5, w6, w7, w8</belief> (eight numbers between 0.00
   and 0.90 with two decimals, in the attribute order given; before the first
   question use your prior estimate), then exactly one <question>...</question>
   tag, and nothing else. Update the estimate after each answer.
```

and the final turn with:

```
Your question budget is exhausted. Recommend now: first write your final
estimate of the user's weights as <belief>w1, w2, w3, w4, w5, w6, w7, w8</belief>,
then <scores>TITLE: S; TITLE: S; ...</scores> listing every unseen movie in the
order given with S = its score under your estimated weights (two decimals), then
exactly one <recommend>TITLE</recommend> tag containing the title of the unseen
movie with the highest score, written exactly as listed, and nothing else.
```

The word-guess scratchpad uses the same user prompt; the scratchpad lines are learned from the targets alone (Appendix D.4).

J.3 Unified-Interface Prompts for the Cross-Task Experiments

The unified interface wraps each task's native action in `<action>...</action>` and adds one sentence to each system prompt stating that every reply must consist of exactly one such tag containing the task's action; scratchpad lines, where used, precede the tag. The full templates will be part of the code release upon acceptance.