
Amortising Bayesian Experimental Design for Sequential Information Gathering in LLMs

Anonymous Authors¹

Abstract

Large language models (LLMs) exhibit strong reasoning and world-knowledge capabilities, yet often struggle to gather information effectively across the multi-turn interactions required in sequential decision-making settings. We introduce Amortised Sequential Information Gathering (ASIG), a fine-tuning approach that amortises Bayesian Experimental Design (BED) into LLM policies via a multi-turn extension of Group Relative Policy Optimisation with an Expected Information Gain reward. Evaluated on the 20 Questions task, ASIG more than doubles the success rate of the 7B base model and reduces inference cost by over $25\times$ relative to BED-LLM, a competitive inference-time baseline. Applied to MediQ, a medical diagnosis benchmark unseen during training, ASIG improves information-seeking performance at the 7B scale, suggesting that the learned strategies can transfer out of distribution. Our findings show that amortising BED into LLM policies provides an effective and computationally efficient approach to sequential information gathering.

1. Introduction

Effective information gathering in sequential settings is fundamental to applications ranging from healthcare to designing experiments in scientific domains. LLMs are increasingly being deployed in such domains, as they are high capacity models encoding extensive world knowledge (Petroni et al., 2019), and have shown promise in reasoning and decision-making tasks on diverse action spaces expressed in natural language (Yao et al., 2023). However, despite these capabilities, LLMs have been shown to struggle with

multi-turn reasoning and planning in various settings. One example is the “lost in conversation” phenomenon reported by Laban et al. (2026), which demonstrates that LLMs often struggle to (1) recognize when additional information is needed to complete a task and (2) effectively acquire that information during multi-turn conversations. These limitations are particularly problematic for experimental design and sequential decision making, where maintaining and updating coherent beliefs over extended interactions is key.

A growing body of work has explored methods for improving the decision-making and information-gathering capabilities of LLMs in sequential settings. Broadly, there are two main classes of approaches: *weight-space optimisation* methods (Mazzaccara et al., 2024; Tajwar et al., 2025; Chi et al., 2024; Yun et al., 2025; Auzina et al., 2026), which fine-tune the model parameters, and *inference-time optimisation* methods (Choudhury et al., 2025; Kobalczyk et al., 2025; Grand et al., 2025), which add post-processing procedures during inference. While demonstrating strong improvement in the information-gathering performance of LLMs in sequential settings, inference-time optimisation methods suffer from high computational cost during inference and must be tailored to a specific task or environment. On the other hand, weight-space optimisation methods yield fast inference time behaviour, but it remains to be seen whether these approaches improve general information gathering capabilities or just fit to the task at hand.

In this work, we introduce Amortised Sequential Information Gathering (ASIG), a weight-space optimisation framework that amortises Bayesian Experimental Design (BED) (Chaloner & Verdinelli, 1995) into the weights of an LLM, enabling sequential information gathering without additional inference-time optimisation. We treat the LLM as a BED agent in a general sequential setting, where it iteratively selects experiments to maximally reduce uncertainty over a latent target. We then fine-tune the model using Expected Information Gain (EIG) as the reward signal within a multi-turn variant of Group Relative Policy Optimisation (GRPO) (Shao et al., 2024). The reward combines EIG, which incentivises experiments that maximally reduce uncertainty over candidate targets, with an outcome-based signal that directly reinforces task success, trained

¹Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the FoGen Workshop at ICML 2026. Do not distribute.

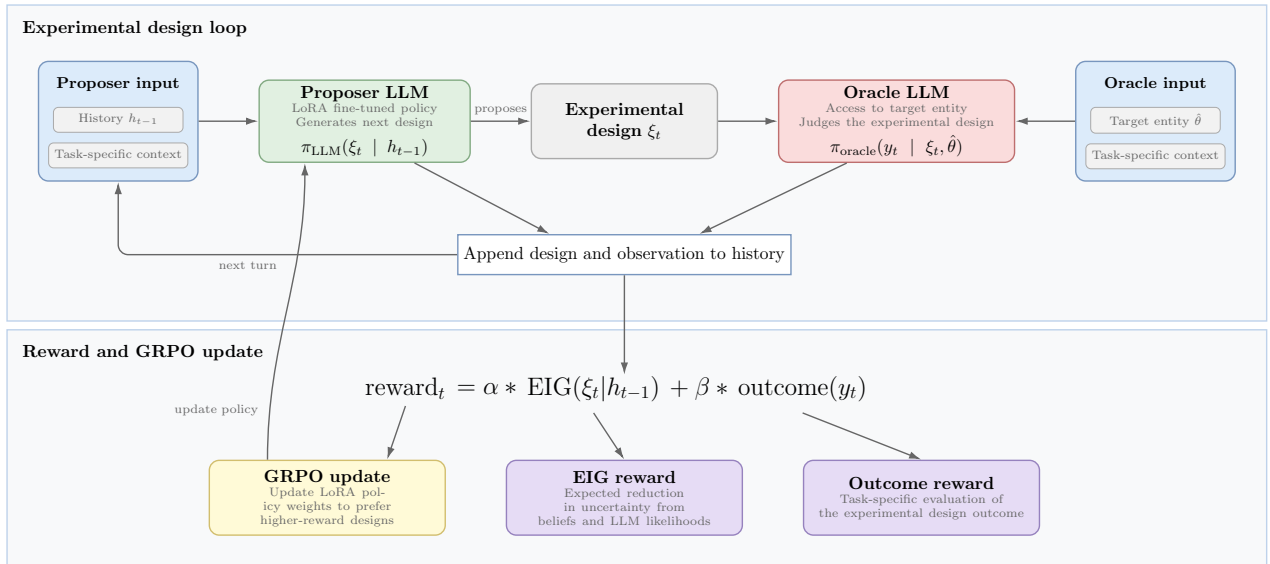


Figure 1. High-level overview of our experimental design loop. At each turn, the Proposer LLM uses the task-specific context and interaction history to generate a new experimental design. The Oracle LLM, which has access to the hidden target entity, evaluates this design and returns an observation. Both the design and observation are appended to the history for the next turn. Finally, we compute a reward combining the Expected Information Gain (EIG) and a task-specific outcome score, using GRPO to update the Proposer LLM’s policy weights.

via a learning curriculum across multi-turn rollouts. As a controlled instantiation of this framework, we train on the 20 Questions game, where binary outcomes make EIG tractable via a Rao-Blackwellised estimator and the capability being trained (asking maximally informative questions) is inherently domain-agnostic. We evaluate on held-out 20 Questions categories and assess generalisation to MediQ (Li et al., 2024), exploring whether amortised BED behaviour transfers to real-world sequential information-gathering settings.

We summarise our contributions as follows:

1. We introduce ASIG, a novel fine-tuning method to amortise BED information gathering into LLMs.
2. We extend GRPO to support multiple reward streams with distinct temporal horizons, enabling joint optimisation of turn-level EIG and multi-horizon task success.
3. We demonstrate that ASIG substantially improves 20 Questions performance over the base model, more than doubling the success rate in the 7B model class, while outperforming a competitive inference-time baseline in two out of four settings at over 25× lower inference cost.
4. We show that the information-seeking behaviour learned by ASIG transfers to MediQ, a medical diagnosis benchmark, at the 7B scale without any in-domain fine-tuning.

2. Related Work

2.1. Inference-Time Optimisation Methods

Inference-time optimisation methods improve sequential decision making and question asking in LLMs by optimising information gathering directly at test time, without modifying model weights. In this setting, LLMs are treated as decision making agents in a sequential problem with a hidden target, where the next action is selected by explicitly maximizing EIG (or an EIG proxy) given the interaction history. BED-LLM (Choudhury et al., 2025) adopts this approach on the 20 Questions task. They use the LLM to propose candidate questions, define a uniform prior over the set of possible targets, and use the questions and possible targets to estimate EIG, eventually selecting the question with the highest score at each turn. Related methods adopt similar formulations, using sampling based EIG estimation for task disambiguation (Kobalczuk et al., 2025), or maintaining explicit Bayesian belief states with Monte Carlo estimation of the EIG (Grand et al., 2025).

These approaches consistently outperform naive prompting, offering principled uncertainty handling and interpretability through explicit belief updates, while remaining applicable to off-the-shelf LLMs. However, they incur substantial inference time cost due to sampling and belief maintenance, often rely on structured hypothesis spaces or simulators (e.g. 20 Questions-type tasks), and are typically myopic and task-specific which limits their actual usability.

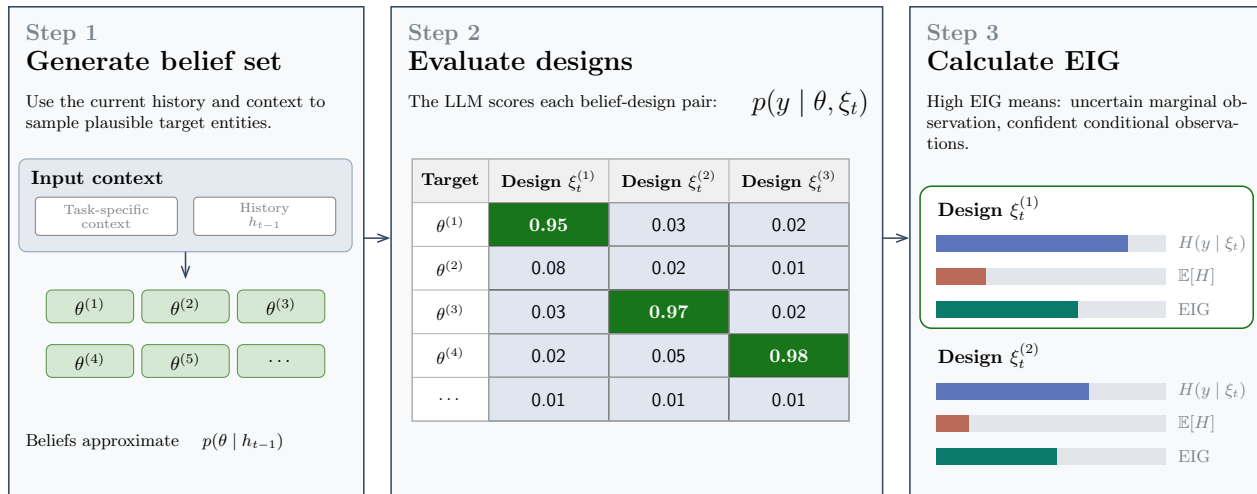


Figure 2. Overview of our step-by-step EIG calculation, following the approach of BED-LLM (Choudhury et al., 2025). 1) Plausible target entities (θ) are sampled from the current belief state. 2) For each design (ξ_t), observation likelihoods $p(y | \theta, \xi_t)$ are computed across all sampled targets. 3) These likelihoods are used to calculate the predictive entropy $H(y_t | \xi_t)$ and expected conditional entropy $\mathbb{E}[H(y_t | \theta, \xi_t)]$. High-EIG designs are those that produce uncertain marginal observations (splitting the belief space) while minimizing conditional entropy.

2.2. Weight-Space Optimisation Methods

Weight-space methods for improving LLM information-gathering divide broadly into retriever-coupled and self-contained approaches. Chi et al. (2024) and Yun et al. (2025) both couple the LLM to an external retriever network that carries the belief state, deriving EIG or rank-based rewards from the retriever’s candidate distribution. However, neither evaluates transfer beyond their training tasks, and both are restricted to settings where retrieval infrastructure exists.

Self-contained approaches remove this dependency. Mazzaccara et al. (2024) demonstrate that EIG is an effective training signal via Direct Preference Optimisation (DPO) (Rafailov et al., 2023) on a static dataset of EIG-ranked question pairs, but the fixed preference dataset cannot adapt to the model’s evolving question distribution, and generalisation is evaluated only within 20 Questions categories. Tajwar et al. (2025) takes a complementary approach, using Relative Preference Optimisation (RPO) (Pang et al., 2024) on purely outcome-based rewards with no information-theoretic component. They demonstrate zero-shot transfer across a family of simple decision-making tasks, though the authors themselves note that online RL would likely yield stronger results than their offline preference optimisation.

Δ Belief-RL (Auzina et al., 2026) is the closest antecedent to our work: an online RL method that uses the change in the model’s assigned probability to the correct target as a dense per-turn reward, without any external retriever, and demonstrates OOD generalisation to customer service and personalisation settings. Our approach shares the same

self-contained online RL paradigm, but uses EIG instead of Δ Belief(target) as the reward signal. This distinction matters because EIG captures both the uncertainty of the marginal answer distribution and the sharpness of per-belief likelihoods, rather than tracking a single scalar probability whose change may not reflect genuine information gain.

3. Preliminary

3.1. Bayesian Experimental Design

BED provides a principled framework for experimental design grounded in Bayesian decision theory, in which experiments are selected based on their expected utility under an explicit probabilistic model of the world (Chaloner & Verdinelli, 1995). In the BED formulation, the uncertainty about a quantity of interest θ is represented by a prior distribution $p(\theta)$, while a likelihood model $p(y | \theta, \xi)$ specifies how experimental outcomes y are generated under a proposed design ξ . Together, these components define a joint probabilistic model over latent variables and observations, enabling candidate experiments to be evaluated prior to data collection by computing their expected utility. This explicit generative modelling distinguishes BED from heuristic or purely discriminative approaches to experimental design, and allows prior knowledge, measurement noise, and experimental constraints to be incorporated in a coherent Bayesian manner (Rainforth et al., 2024).

Expected Information Gain. First proposed by Lindley (1956), EIG is a widely used optimality metric within BED.

Given a latent variable θ , which we seek to learn about through an experiment ξ , we define the *Information Gain* (IG) on θ from an experimental outcome y by the reduction in Shannon entropy (Shannon, 1948) from the prior to the posterior:

$$\text{IG}_\theta(\xi) = H[p(\theta)] - H[p(\theta|y, \xi)]. \quad (1)$$

EIG is then calculated as the expectation of this quantity over the marginal predictive distribution, $p(y|\xi)$:

$$\begin{aligned} \text{EIG}_\theta(\xi) &= \mathbb{E}_{p(y|\xi)}[\text{IG}_\theta(\xi, y)] \\ &= \mathbb{E}_{p(\theta)p(y|\theta, \xi)}[\log p(\theta|y, \xi) - \log p(\theta)] \\ &= \mathbb{E}_{p(\theta)p(y|\theta, \xi)}[\log p(y|\theta, \xi) - \log p(y|\xi)]. \end{aligned} \quad (2)$$

Finding the optimal experimental design thus amounts to maximising $\text{EIG}_\theta(\xi)$ with respect to ξ . However, as both $p(y|\xi)$ and $p(\theta|y, \xi)$ are intractable, EIG cannot be directly estimated from traditional Monte Carlo (MC) estimators (Rainforth et al., 2024; Foster et al., 2020). However, Rainforth et al. (2018) point out that when the outcome space \mathcal{Y} is finite, EIG can be estimated using a Rao-Blackwellised estimator:

$$\hat{\mu}_N = \sum_{y \in \mathcal{Y}} \left[\frac{1}{N} \sum_{n=1}^N p(y|\theta_n, \xi) \log p(y|\theta_n, \xi) - \hat{p}(y|\xi) \log \hat{p}(y|\xi) \right], \quad (3)$$

where $\hat{p}(y|\xi) = \frac{1}{N} \sum_{n=1}^N p(y|\theta_n, \xi)$. Since this is essentially a conventional Monte Carlo estimator with no nesting, it has the advantage of sharing the same mean squared error convergence rate ($\mathcal{O}(1/N)$) as in standard Monte Carlo.

Sequential BED. BED can be extended to sequential design settings by iteratively incorporating information gathered by prior experiments to select the next one (Rainforth et al., 2024). Formally, given a history $h_{t-1} = \{(\xi_k, y_k)\}_{k=1}^{t-1}$, the EIG of design ξ_t is given by¹

$$\text{EIG}_\theta(\xi_t|h_{t-1}) = \mathbb{E}_{p(\theta|h_{t-1})p(y_t|\theta, \xi_t)} \left[\log \frac{p(y_t|\theta, \xi_t)}{p(y_t|\xi_t)} \right],$$

where $h_0 = \emptyset$. This corresponds to updating the prior with each observation to select the next design.

3.2. Group Relative Policy Optimisation

GRPO (Shao et al., 2024) is a variant of Proximal Policy Optimization (PPO) (Schulman et al., 2017) that eliminates the need for an expensive critic model to approximate the

¹This assumes that y_t is independent of h_{t-1} given θ and ξ_t , which is the case for our 20 Questions training environment.

value function and instead calculates the advantage using the relative differences between individual rollout rewards and the group mean. At each training step, GRPO samples a prompt from the prompt dataset, performs G rollouts, and then scores them with a given reward model. These advantages are then used to update the policy π_{LLM} using a clipped surrogate objective.

4. ASIG: Amortised Sequential Information Gathering

We consider a multi-turn interaction environment, in which the Proposer LLM agent is tasked with uncovering a latent target quantity $\theta \in \Theta$ through a budget of T sequential experimental designs $\xi \in \Xi$. The environment is initialised with a task-specific prompt which defines the search space. Over a trajectory $\tau = \{(\xi_k, y_k)\}_{k=1}^{k=T}$, the agent interacts with the environment, modelled by the frozen and independent Oracle LLM, through experimental designs ξ and receives its outcomes $y \in \mathcal{Y}$. The interaction history at step t is denoted by $h_{t-1} = \{(\xi_k, y_k)\}_{k=1}^{k=t-1}$, with h_0 being the initial task-specific prompt. In this setting, we treat the Proposer LLM agent as our sequential BED policy $\pi_{\text{LLM}}(\cdot|h_{t-1})$, from which experiments are sampled based on the accumulated context. We then use the Proposer LLM to generate beliefs and compute likelihoods $p_{\text{LLM}}(y_t|\theta, \xi_t)$ to estimate the EIG for each proposal. An outcome-based reward is computed from the corresponding experimental outcomes, and together these two components define the reward signal in our multi-turn GRPO fine-tuning setup. Figure 1 provides a high-level overview of the approach, while the following subsections detail the specific implementation evaluated in this work.

4.1. Training Environment

In this study, we implement our training procedure using the 20 Questions game environment, which has several desirable properties in this context. First, 20 Questions forms a natural instantiation of the BED framework in which the questions posed by the LLM are themselves the experimental designs ξ_t , and the answers are the corresponding experimental outcomes y_t . Additionally, the binary experimental outcome space (i.e. “yes” or “no”) allows for an efficient estimation of the EIG through the Rao-Blackwellised estimator in Equation (3). Finally, Zhang et al. (2024) have shown that succeeding in 20 Questions requires good planning to efficiently partition the hypothesis space before deductive reasoning can be used to produce educated guesses in later turns. This generic planning-oriented structure makes the task a useful proxy for learning capabilities that may transfer to other information gathering tasks.

The resulting environment is implemented as an interaction between the Proposer LLM, which is fine-tuned, and

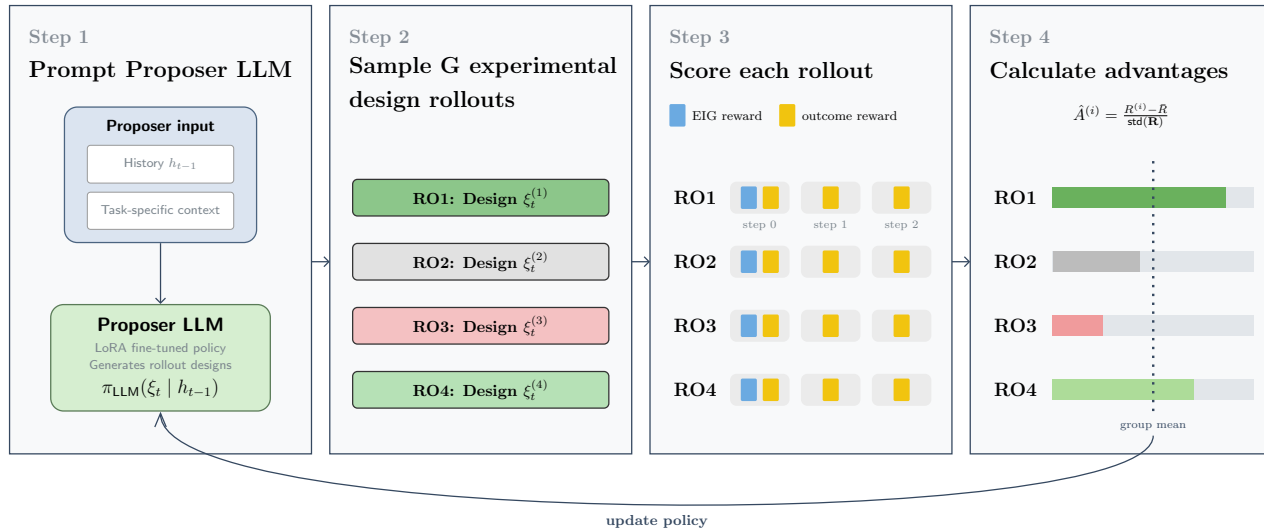


Figure 3. Overview of our multi-turn GRPO method. At each turn, the Proposer LLM generates G rollouts based on the given context and interaction history. Each rollout consists of multiple steps of experimental design and observation. We calculate the EIG reward for the first step, and the outcome-based reward across all O steps. The combined reward is used to calculate the group mean, and the individual advantages are then used to update the Proposer LLM’s policy.

a separate frozen Oracle LLM. At each turn, the Proposer receives the task category together with the accumulated question-answer history and generates a question or guess. The Oracle, which has access to the hidden target entity, evaluates this proposal and returns a binary answer, either “Yes” or “No”, unless the Proposer has correctly identified the target, in which case the Oracle returns “Correct” and the interaction terminates. For non-terminal turns, the resulting question-answer pair is appended to the history and the interaction continues until either the target is identified or the budget of $T = 20$ turns is exhausted.

4.2. Reward

Effective information gathering requires the ability to both design maximally informative experiments and identify when enough information has been acquired to identify the hidden entity and act optimally. We argue that these are two distinct skills, which motivates two separate components for the reward signal. For each experimental design (e.g. a question in 20 Questions) generated in the GRPO rollout, we calculate two reward components: EIG and outcome reward. These two signals are complementary. EIG provides dense, turn-level feedback on question quality but carries no signal on whether information gathering ultimately succeeds, while the outcome-based reward drives goal-directed behaviour but suffers from sparse credit assignment in long-horizon settings. For 20 Questions specifically, the outcome reward encourages the model to guess the hidden entity using as few questions as possible, while EIG encourages asking maximally informative questions.

The outcome reward, denoted $r_{\text{out}}(y_t)$, consists of a step-wise penalty for all questions that lead to a non-terminal state and a large positive reward if the questioner has correctly guessed the hidden entity.

To calculate the EIG for a given question ξ_t , we use the approach introduced by Choudhury et al. (2025) and visualized in Figure 2. At each turn, we sample beliefs $\theta \sim p_{\text{LLM}}(\cdot | h_{t-1})$ from the Proposer LLM to create a belief set based on the current conversation history. This set is then filtered to ensure that only coherent beliefs, which are consistent with the previous questions and answers, are retained. If the resulting belief set is too small, we perform up to three rounds of belief regeneration to ensure a large enough set. We denote the resulting belief set at turn t formed from this process as Θ_t . Afterwards, we uniformly sample N beliefs $\theta_n \in \Theta_t$ and query the Proposer LLM to obtain the likelihoods for each sampled belief, $p_{\text{LLM}}(y_t | \theta_n, \xi_t)$, which are used to estimate the question’s EIG via the Rao-Blackwellised estimator in Equation (3). We scale the EIG reward by $1/\ln(2)$ to obtain a normalized scalar, $r_{\text{EIG}}(\xi_t, \Theta_t)$, within the $[0, 1]$ range.

4.3. Multi-Turn GRPO

We extend the default GRPO formulation (Shao et al., 2024) to the multi-turn setting. At each training step, a starting state consisting of the task category and the interaction history up to that turn, h_{t-1} , is sampled from the prompt dataset outlined in Section 4.5. From this state, the Proposer LLM generates G independent policy rollouts by sampling multiple consecutive question-answer pairs within each roll-

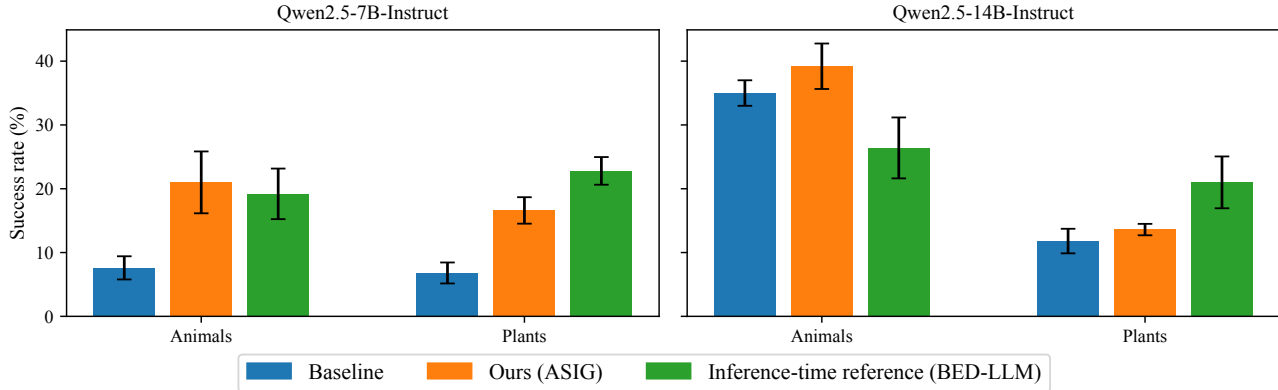


Figure 4. Mean success rate (± 1 std) over 5 episodes of Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct on the 20 Questions task across the animals and plants categories. On the 7B model, ASIG more than doubles the success rate of the baseline and outperforms BED-LLM in the animals category. On the 14B model, ASIG improves over the baseline by 4.2 and 1.8 percentage points on animals and plants, respectively. BED-LLM performs better than ASIG on the 14B plants category but falls substantially below both ASIG and the baseline on animals, suggesting limited robustness across settings despite using an order of magnitude more inference-time compute.

out. A high-level overview is shown in Figure 3. For each rollout, we compute the scalar reward described in Section 4.2 using

$$R_t = \frac{\alpha}{\mathcal{Z}_E} \sum_{k=0}^{E-1} \gamma_{\text{EIG}}^k r_{\text{EIG}}(\xi_{t+k}, \Theta_{t+k}) + \frac{\beta}{\mathcal{Z}_O} \sum_{k=0}^{O-1} \gamma_{\text{out}}^k r_{\text{out}}(y_{t+k}), \quad (4)$$

where α, β are weighting constants, $\gamma_{\text{EIG}}, \gamma_{\text{out}} \in (0, 1]$ are discount factors, and E, O correspond to the rollout horizon for the EIG and outcome-based reward, respectively. The normalization constants \mathcal{Z}_E and \mathcal{Z}_O ensure that different rollout horizons for EIG and outcome-based rewards do not lead to varying reward magnitudes. For a horizon H and discount factor γ , we define

$$\mathcal{Z}(H, \gamma) = \sum_{k=0}^{H-1} \gamma^k = \begin{cases} H & \text{if } \gamma = 1 \\ \frac{1 - \gamma^H}{1 - \gamma} & \text{otherwise.} \end{cases} \quad (5)$$

In our experiments, we use $E = 1$ and $O = 3$, i.e. the EIG reward is calculated only with respect to the first question in each rollout, whereas the outcome reward also considers the two subsequent questions. This favours questions that maximize EIG at the current turn, while increasing the probability of a correct guess within three turns. After obtaining the rewards for each question, we calculate the average group return and individual rollout advantages, which are used to update the LoRA (Hu et al., 2022) adapters of the questioner model.

4.4. Learning Curriculum

In preliminary experiments, we observed that the Proposer LLM initially rarely makes specific guesses. To encourage this behaviour, we employ a learning curriculum strategy in which the model is first trained without the EIG reward ($\alpha = 0$), before subsequently introducing the EIG signal. This curriculum increases the frequency of hypothesis-seeking questions, which contain specific guesses, relative to constraint-seeking questions that probe the characteristics of the hidden entity.

4.5. GRPO Prompt Dataset

At the start of each training episode, we randomly sample n target entities from a category-entity list to generate a 20 Questions self-play dataset between the Proposer and Oracle. Each state in this dataset, consisting of the entity’s category and the Q/A history up to question $t - 1$, then serves as input for GRPO rollouts starting at question t .

5. Experiments

5.1. Setup

To evaluate our method, we train on the 20 Questions task and assess both in-distribution and out-of-distribution performance. We use Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct as Proposer LLMs and Qwen2.5-72B-Instruct as the Oracle LLM (Yang et al., 2025). For training, we construct a list of 1,000 target entities spanning 20 categories. This list was generated with Claude Opus 4.7 (Anthropic, 2026) using Prompt 5 and manually reviewed to ensure quality. We train a LoRA adapter on this dataset following the approach in Section 4, with hyperparameters listed in Table 3.

We evaluate in-distribution performance on two held-out categories and out-of-distribution performance on MediQ (Li et al., 2024), a medical information-seeking benchmark. We further verify that improved information-seeking capabilities do not come at the cost of degraded general language abilities, as measured on three standard LLM benchmarks.

5.2. 20 Questions

We evaluate the fine-tuned models on two held-out categories, animals and plants, each comprising 100 entities not seen during training. The animal entities are drawn from Choudhury et al. (2025) and the plant entities are a subset of those used by Auzina et al. (2026). The full lists are provided in Boxes 1 and 2. We compare against the instruction-tuned base model and our reimplementations of BED-LLM (Choudhury et al., 2025; Filbry, 2026).

Success rate. Figure 4 shows the success rate across model sizes and evaluation categories, measured as the fraction of games in which the questioner identifies the hidden entity within the 20 question budget. On the 7B model, ASIG more than doubles the success rate over the baseline and surpasses BED-LLM on the animals category. On the 14B model, ASIG improves over the baseline by 4.2 and 1.8 percentage points on animals and plants, respectively. BED-LLM performs better on the 14B plants category but degrades on the animals category, falling below both ASIG and the baseline, suggesting that it does not generalise robustly across settings despite requiring an order of magnitude more inference-time compute.

Figure 7 in the Appendix further shows that ASIG identifies the target entity earlier than the baseline, requiring fewer questions on average. Qualitative examples comparing baseline and ASIG on the same target entity are shown in Boxes 3 and 4 in the Appendix.

Behavioural analysis. To better understand the strategies learned by ASIG, we classify each question as either constraint-seeking (asking about the entity’s characteristics to narrow down candidates) or hypothesis-seeking (making a specific guess about the target), following the taxonomy of Mazzaccara et al. (2024). Figure 5 shows the fraction of hypothesis-seeking questions across turns for the 7B model. ASIG consistently exceeds the baseline across both categories, committing to explicit guesses earlier in the game. For the 14B model, ASIG’s hypothesis-seeking rate closely tracks the baseline (Figure 8 in the Appendix). This likely reflects the stronger prior capabilities of the 14B model, which requires less behavioural adaptation to benefit from fine-tuning. BED-LLM achieves the highest overall rate, but this is partly driven by hard-coded guessing logic that forces explicit guesses once the belief set falls below three candidates, which ASIG does not rely on.

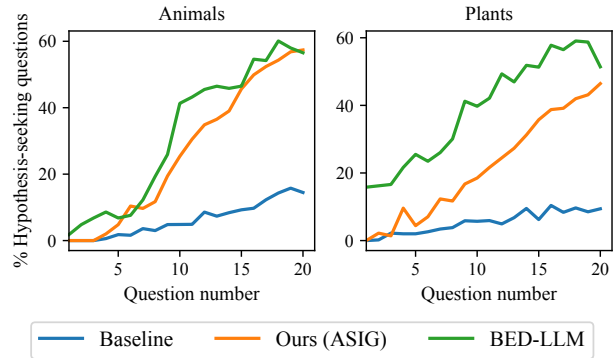


Figure 5. Fraction of questions classified as hypothesis-seeking at each turn for Qwen2.5-7B-Instruct across the animals and plants categories. ASIG consistently and substantially exceeds the baseline across both categories, suggesting that fine-tuning encourages the model to commit to explicit guesses earlier in the game. BED-LLM shows the highest rate overall, though this partly reflects its hard-coded guessing rules rather than an emergent strategy as in ASIG.

Inference-Time Efficiency. By amortising the Bayesian experimental design computation into the model weights, ASIG incurs the same inference cost as the base LLM. Notably, this makes ASIG between 25x and 36x faster than BED-LLM (Figure 6), which must perform explicit belief sampling and EIG maximisation at every turn. This efficiency gain makes ASIG considerably more applicable to real-world settings where latency and compute are constrained.

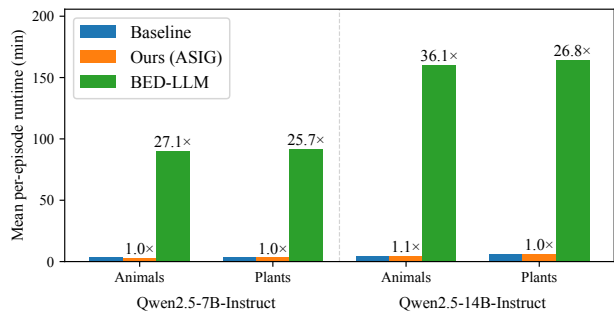


Figure 6. Mean per-episode wall-clock runtime of the baseline, ASIG, and BED-LLM across all model and category settings. Runtime ratios relative to the baseline are annotated above each ASIG and BED-LLM bar. ASIG matches the baseline runtime, while BED-LLM is between 25x and 36x slower, as it requires repeated belief sampling and EIG estimation at every turn rather than a single forward pass.

5.3. Generalization to Medical Information Seeking

To assess out-of-distribution performance, we evaluate our fine-tuned models on MediQ (Li et al., 2024), a benchmark that simulates a clinical interaction in which a doctor elicits information from a patient through follow-up questions be-

Table 1. MediQ results for the baseline and ASIG on Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct. For the 7B model, ASIG improves accuracy and voluntary accuracy while reducing overconfidence, at a minor cost in underconfidence. Results are more mixed for the 14B model, consistent with the weaker behavioural adaptation observed in the hypothesis-seeking analysis.

Model	Variant	Accuracy (%) \uparrow	Voluntary accuracy (%) \uparrow	Overconfidence rate (%) \downarrow	Underconfidence rate (%) \downarrow
Qwen2.5-7B-Instruct	Baseline	46.7	50.1	36.1	10.6
	Ours (ASIG)	47.6	51.8	33.4	11.7
Qwen2.5-14B-Instruct	Baseline	54.1	64.6	20.7	16.5
	Ours (ASIG)	54.2	63.0	22.2	16.4

fore answering a multiple-choice clinical question. At each turn, a doctor agent uses an abstention module to decide whether sufficient information has been gathered to answer or whether to ask another question. We use our fine-tuned Qwen2.5-7B and Qwen2.5-14B models from the 20 Questions task as the doctor and Qwen2.5-72B-Instruct as the patient, which responds based on a provided patient record.

We compare each fine-tuned model to its base counterpart across four metrics: accuracy (overall success rate), voluntary accuracy (success rate on cases where the model chose to commit to an answer rather than being forced to at the question budget), overconfidence rate (proportion of cases where the model committed to an incorrect answer), and underconfidence rate (proportion of cases where the model continued asking questions despite already having sufficient information to answer correctly).

Our 7B model improves accuracy and voluntary accuracy by 0.9 and 1.7 percentage points, respectively. The overconfidence rate drops from 36.1% to 33.4%, at a slight cost of 1.1 percentage points in underconfidence. Results are more mixed for the 14B model: accuracy remains essentially unchanged, voluntary accuracy decreases slightly from 64.6% to 63.0%, and overconfidence increases marginally by 1.5 percentage points.

Overall, ASIG transfers well to the medical domain at the 7B scale. The gains in voluntary accuracy and reduction in overconfidence align with our earlier finding that fine-tuning encourages more deliberate hypothesis-seeking, here reflected in the model’s improved judgement about when to commit to an answer. The mixed 14B results are consistent with the weaker behavioural adaptation observed in the hypothesis-seeking analysis, suggesting that the larger model’s stronger prior capabilities leave less room for fine-tuning to induce meaningful behavioural change.

5.4. Capability Retention

To ensure that fine-tuning does not degrade performance on other tasks or lead to failure modes such as catastrophic forgetting, we evaluate the baseline and fine-tuned models on three standard LLM benchmarks: ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), and MMLU (Hendrycks et al., 2021a;b).

Table 2. LLM benchmark scores for the baseline and ASIG on Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct across ARC-Challenge, HellaSwag, and MMLU. ASIG causes no meaningful degradation across any benchmark or model size, confirming that general language capabilities are preserved.

Model	Variant	ARC Challenge	HellaSwag	MMLU
Qwen2.5-7B	Baseline	43.6	65.4	68.7
	Fine-tuned	43.8	65.4	68.8
Qwen2.5-14B	Baseline	50.0	75.6	76.8
	Fine-tuned	49.8	75.6	76.8

As shown in Table 2, the fine-tuned models achieve near-identical scores to the base models across all three benchmarks, confirming that improved information-seeking abilities do not come at the cost of general language capabilities.

6. Conclusion

We introduced ASIG, a multi-turn GRPO-based fine-tuning method that amortises BED into LLM policies for sequential information gathering. By shifting information-seeking behaviour from inference-time optimisation into model weights, ASIG improves task performance while substantially reducing the computational overhead of sequential BED at deployment time.

Across two model scales, fine-tuning on 20 Questions consistently improves information-gathering performance on held-out categories, outperforming BED-LLM, a competitive inference-time BED baseline, in two of four settings while reducing inference cost by over 25 \times overall. We further show partial transfer to MediQ, where ASIG improves clinical reasoning accuracy at the 7B scale without measurable degradation on standard language-understanding benchmarks. These results suggest that amortising information-seeking behaviour into model weights provides a practical and scalable approach to test-time information acquisition in LLMs.

Future work includes disentangling the effects of EIG and outcome-based rewards, evaluating broader sequential information-gathering benchmarks, and extending ASIG to reasoning models that generate intermediate reasoning steps.

References

- Anthropic. Claude 4.7 Sonnet, 2026. URL <https://claude.ai/>.
- Auzina, I. A., Strüber, J., Hernández-Gutiérrez, S., Goel, S., Prabhu, A., and Bethge, M. Intrinsic Credit Assignment for Long Horizon Interaction, February 2026. URL <http://arxiv.org/abs/2602.12342>. arXiv:2602.12342 [cs].
- Chaloner, K. and Verdinelli, I. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273–304, August 1995. ISSN 0883-4237, 2168-8745. doi: 10.1214/ss/1177009939. URL <https://projecteuclid.org/journals/statistical-science/volume-10/issue-3/Bayesian-Experimental-Design-A-Review/10.1214/ss/1177009939.full>.
- Chi, Y., Lin, J., Lin, K., and Klein, D. CLARINET: Augmenting Language Models to Ask Clarification Questions for Retrieval, April 2024. URL <http://arxiv.org/abs/2405.15784>. arXiv:2405.15784 [cs].
- Choudhury, D., Williamson, S., Goliński, A., Miao, N., Smith, F. B., Kirchhof, M., Zhang, Y., and Rainforth, T. BED-LLM: Intelligent Information Gathering with LLMs and Bayesian Experimental Design, October 2025. URL <http://arxiv.org/abs/2508.21184>. arXiv:2508.21184 [cs].
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge, March 2018. URL <http://arxiv.org/abs/1803.05457>. arXiv:1803.05457 [cs].
- Filbry, L. BED-LLM-reproduction/model.py at main · LinusFilbry/BED-LLM-reproduction, 2026. URL <https://github.com/LinusFilbry/BED-LLM-reproduction/blob/main/model.py>.
- Foster, A., Jankowiak, M., O’Meara, M., Teh, Y. W., and Rainforth, T. A Unified Stochastic Gradient Approach to Designing Bayesian-Optimal Experiments. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 2959–2969. PMLR, June 2020. URL <https://proceedings.mlr.press/v108/foster20a.html>.
- Grand, G., Pepe, V., Andreas, J., and Tenenbaum, J. B. Shoot First, Ask Questions Later? Building Rational Agents that Explore and Act Like People, October 2025. URL <http://arxiv.org/abs/2510.20886>. arXiv:2510.20886 [cs].
- Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D., and Steinhardt, J. Aligning AI With Shared Human Values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring Massive Multitask Language Understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*, 2022. URL <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Kobalczyk, K., Astorga, N., Liu, T., and Schaar, M. v. d. Active Task Disambiguation with LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL <https://openreview.net/forum?id=JAMxRSXLFz>.
- Laban, P., Hayashi, H., Zhou, Y., and Neville, J. LLMs Get Lost In Multi-Turn Conversation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL <https://openreview.net/forum?id=VKGTGGcwl6>.
- Li, S. S., Balachandran, V., Feng, S., Ilgen, J. S., Pierson, E., Koh, P. W., and Tsvetkov, Y. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, December 2024. doi: 10.52202/079017-0908. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/32b80425554e081204e5988ab1c97e9a-Abstract-Conference.html.
- Lindley, D. V. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, December 1956. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177728069. URL <https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-27/issue-4/On-a-Measure-of-the-Information-Provided-by-an-Experiment/10.1214/aoms/1177728069.full>.
- Mazzaccara, D., Testoni, A., and Bernardi, R. 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. URL <https://aclanthology.org/2024.findings-emnlp.291/>.

- 495 Pang, R. Y., Yuan, W., Cho, K., He, H., Sukhbaatar, S., and
 496 Weston, J. Iterative reasoning preference optimization.
 497 In *Proceedings of the 38th International Conference on*
 498 *Neural Information Processing Systems*, volume 37 of
 499 *NIPS '24*, pp. 116617–116637, Red Hook, NY, USA,
 500 2024. Curran Associates Inc. ISBN 979-8-3313-1438-5.
- 501 Petroni, F., Rocktäschel, T., Miller, A. H., Lewis, P., Bakhtin,
 502 A., Wu, Y., and S. Riedel. Language Models as Knowl-
 503 edge Bases? In *In: Proceedings of the 2019 Conference*
 504 *on Empirical Methods in Natural Language Processing*
 505 *(EMNLP), 2019, 2019.*
- 507 Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Man-
 508 ning, C. D., and Finn, C. Direct preference optimization:
 509 your language model is secretly a reward model. In *Pro-*
 510 *ceedings of the 37th International Conference on Neural*
 511 *Information Processing Systems, NIPS '23*, pp. 53728–
 512 53741, Red Hook, NY, USA, 2023. Curran Associates
 513 Inc.
- 515 Rainforth, T., Cornish, R., Yang, H., Warrington, A., and
 516 Wood, F. On Nesting Monte Carlo Estimators. In *Pro-*
 517 *ceedings of the 35th International Conference on Ma-*
 518 *chine Learning*, pp. 4267–4276. PMLR, July 2018. URL
 519 [https://proceedings.mlr.press/v80/ra-](https://proceedings.mlr.press/v80/rainforth18a.html)
 520 [inforth18a.html](https://proceedings.mlr.press/v80/rainforth18a.html).
- 522 Rainforth, T., Foster, A., Ivanova, D. R., and Smith, F. B.
 523 Modern Bayesian Experimental Design. *Statistical Sci-*
 524 *ence*, 39(1):100 – 114, 2024. doi: 10.1214/23-STS915.
 525 URL <https://doi.org/10.1214/23-STS915>.
- 527 Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and
 528 Klimov, O. Proximal Policy Optimization Algorithms,
 529 August 2017. URL [http://arxiv.org/abs/17](http://arxiv.org/abs/1707.06347)
 530 [07.06347](http://arxiv.org/abs/1707.06347). arXiv:1707.06347 [cs].
- 531 Shannon, C. E. A mathematical theory of communication.
 532 *The Bell System Technical Journal*, 27(3):379–423, July
 533 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.t
 534 b01338.x. URL [https://ieeexplore.ieee.or](https://ieeexplore.ieee.org/abstract/document/6773024)
 535 [g/abstract/document/6773024](https://ieeexplore.ieee.org/abstract/document/6773024).
- 537 Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X.,
 538 Zhang, H., Zhang, M., Li, Y. K., Wu, Y., and Guo,
 539 D. DeepSeekMath: Pushing the Limits of Mathemat-
 540 ical Reasoning in Open Language Models, April 2024.
 541 URL <http://arxiv.org/abs/2402.03300>.
 542 arXiv:2402.03300 [cs].
- 544 Tajwar, F., Jiang, Y., Thankaraj, A., Rahman, S. S., Kolter,
 545 J. Z., Schneider, J., and Salakhutdinov, R. Training a
 546 Generally Curious Agent. In *Forty-second International*
 547 *Conference on Machine Learning, 2025.* URL [https:](https://openreview.net/forum?id=UeB3Hdrhda)
 548 [/openreview.net/forum?id=UeB3Hdrhda](https://openreview.net/forum?id=UeB3Hdrhda).
- Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B.,
 Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu,
 J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang,
 K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M.,
 Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia,
 T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan,
 Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5
 Technical Report, January 2025. URL [http://arxi](http://arxiv.org/abs/2412.15115)
 v.org/abs/2412.15115. arXiv:2412.15115 [cs].
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan,
 K. R., and Cao, Y. ReAct: Synergizing Reasoning and
 Acting in Language Models. In *The Eleventh Interna-*
 tional Conference on Learning Representations, 2023.
 URL [https://openreview.net/forum?id=](https://openreview.net/forum?id=WE_vluYUL-X)
 WE_vluYUL-X.
- Yun, D., Schouten, M., and Papadopoulos, D. Sher-
 lock Your Queries: Learning to Ask the Right Ques-
 tions for Dialogue-Based Retrieval, October 2025. URL
<http://arxiv.org/abs/2510.18659>.
 arXiv:2510.18659 [cs].
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi,
 Y. HellaSwag: Can a Machine Really Finish Your Sen-
 tence? In *Proceedings of the 57th Annual Meeting of the*
Association for Computational Linguistics, 2019.
- Zhang, Y., Lu, J., and Jaitly, N. Probing the Multi-turn
 Planning Capabilities of LLMs via 20 Question Games.
 In *ICLR 2024 Workshop: How Far Are We From AGI,*
 2024. URL [https://openreview.net/forum](https://openreview.net/forum?id=dhylNBeusb)
[?id=dhylNBeusb](https://openreview.net/forum?id=dhylNBeusb).

A. Hyperparameters

Table 3. List of hyperparameters. All values are shared across the 7B and 14B model variants. Hyperparameters for BED-LLM and EIG calculation are taken from (Choudhury et al., 2025).

Symbol	Name	Value
BED-LLM inference		
N	Belief set size	15
M	Candidate questions per turn	15
τ_f	Belief filter threshold	0.2
T_{gen}	Generation temperature (beliefs & candidates)	1.3
T_a	Answer temperature (likelihood)	0.7
—	Max regeneration rounds	3
—	Monte Carlo samples for EIG	15
—	Retain consistent old beliefs	true
Reward		
α	EIG reward weight curriculum	$0.0 \xrightarrow{\text{epoch } 20} 1.0$
β	Outcome reward weight	1.0
$r_{\text{out}}(\text{Correct})$	Correct-guess reward	+5
$r_{\text{out}}(\text{Yes/No})$	Step penalty	-1
N	Belief set size (reward model)	15
τ_f	Belief filter threshold (reward model)	0.2
T_a	Answer temperature (reward model)	0.7
E	EIG rollout horizon	1
O	Outcome rollout horizon	3
γ_{eig}	EIG discount factor	1.0
γ_{out}	Outcome discount factor	0.9
GRPO training		
—	Epochs	40
—	Initial learning rate	2×10^{-6}
β_{KL}	KL penalty coefficient	0.0
G	Group size (completions per prompt)	8
—	Per-device batch size	2
—	Gradient accumulation steps	8
—	Max completion length (tokens)	128
—	Sampled entities per epoch	100
—	Dataset refresh (epochs)	1
—	Questions budget per game	20
LoRA		
r	Rank	32
α_{LoRA}	Scaling factor	64
—	Dropout	0.05
—	Target modules	all attention + MLP projections

B. Prompts

Prompt 1: Proposer Agent System Prompt

You are playing a game of 20 Questions. Using all previous questions and answers:

Generate the best question to help identify the target {category}. The question should be phrased so the answer is Yes or No. Print only the question - no numbering, punctuation, or extra text.

Prompt 2: Oracle Answerer System Prompt

You are playing the oracle answerer in a game of 20 Questions. The questioner has to guess a secret entity from the category: {category}. Your secret entity is: {entity}

CRITICAL INSTRUCTIONS:

1. If the question directly identifies, names, or guesses "{entity}" as the answer, respond ONLY with the word "Correct"
2. For yes/no questions about properties or characteristics, respond ONLY with "Yes" or "No"
3. Never add explanations, extra words, or punctuation

Examples:

- "Is it {entity}?" -> "Correct"
- "Is it a {entity}?" -> "Correct"
- "Are you thinking of {entity}?" -> "Correct"
- "I think it's {entity}" -> "Correct"
- "Is the {category} a {entity}?" -> "Correct"
- "Is it bigger than a car?" -> "Yes" or "No" (depending on {entity})
- "Can you find it indoors?" -> "Yes" or "No" (depending on {entity})
- "Is it man-made?" -> "Yes" or "No" (depending on {entity})
- "Is it alive?" -> "Yes" or "No" (depending on {entity})

Prompt 3: Likelihood Model System Prompt

You are playing the answerer in a game of 20 Questions. Your chosen entity is: {hypothesis}

When asked a question, you must reply exactly "Yes" or "No", depending on if your chosen entity fulfills the question.

Prompt 4: Question Classifier System Prompt

You are classifying questions from a 20 Questions game.

C - constraint-seeking: narrows the space of possible entities by asking about properties or categories (e.g. "Is it a mammal?", "Is it bigger than a car?")

H - hypothesis-seeking: asks whether the target IS a specific entity (e.g. "Is it a cat?", "Is it the Eiffel Tower?")

U - unknown: the question is ambiguous or does not fit either category

Reply with ONLY the single letter: C, H, or U.

Prompt 5: Training Dataset Generation Prompt

Role: You are a curator building a high-quality entity dataset for a 20 Questions game.

Task: Generate ~50 entities for each of the categories listed below. Each entity must be guessable through ~20 yes/no questions by an average adult player.

Categories & quantities:

- Musical instrument
- Fruit
- Organ
- Profession
- City
- Clothing
- Game
- Furniture
- Author
- Vegetable
- Sport
- Vehicle
- Electronic device
- TV show
- Celebrity
- Food chain
- University course
- Monument
- Cooking equipment
- River

Quality criteria - each entity must be:

1. Recognizable - known to a general adult audience, not requiring specialist knowledge.
2. Concrete & unambiguous - a single, well-defined referent (e.g., "elephant," not "large mammal").
3. Distinguishable - possesses a clear set of yes/no-answerable properties (size, habitat, function, era, material, etc.).
4. Non-trivial - avoid items so generic they're guessed in 3 questions ("dog") or so obscure they're unguessable.
5. Diverse within category - span subtypes, regions, eras, and difficulty levels; no near-duplicates.

Output format: CSV

For reference, attached are example lists for the animals and plants category. Generate entities with a similar level of difficulty.

C. 20 Questions Evaluation Datasets

Animals dataset

- | | | | |
|----------------------|-----------------------|-----------------------|--------------------------|
| • African elephant | • Platypus | • European badger | • Vicuña |
| • Bengal tiger | • Rhinoceros | • Moose | • Wildebeest |
| • Bald eagle | • Tasmanian devil | • African grey parrot | • Rock hyrax |
| • Blue whale | • Wombat | • Scarlet macaw | • Yak |
| • Red kangaroo | • Sloth | • Black mamba | • Zebra |
| • Giant panda | • Blue-ringed octopus | • Albatross | • Blue dragon nudibranch |
| • Snow leopard | • Manatee | • Humpback whale | • Chinchilla |
| • Green sea turtle | • Narwhal | • Dugong | • Dhole |
| • American alligator | • Sea otter | • Anaconda | • Electric eel |
| • Bottlenose dolphin | • Coral snake | • Kookaburra | • Flying fox |
| • Emperor penguin | • King cobra | • Coyote | • Gharial |
| • Great white shark | • Harpy eagle | • Brown bear | • Horseshoe crab |
| • Golden poison frog | • Lemur | • Golden jackal | • Indigo bunting |
| • Honey bee | • Koala | • Capybara | • Jerboa |
| • Monarch butterfly | • Aye-aye | • Ibex | • Kakapo |
| • Okapi | • Snowy owl | • Japanese macaque | • Lionfish |
| • Chimpanzee | • Elk | • Kiwi | • Markhor |
| • Arctic fox | • Wolverine | • Leafcutter ant | • Nautilus |
| • Komodo dragon | • Caracal | • Mantis shrimp | • Olive baboon |
| • Giraffe | • Cassowary | • Ocelot | • Pika |
| • Cheetah | • Quokka | • Peregrine falcon | • Quoll |
| • Hammerhead shark | • Pangolin | • Quetzal | • Rosy boa |
| • Axolotl | • Saiga antelope | • Raccoon | |
| • Orca | • Galápagos tortoise | • Sand cat | |
| • Puffin | • Sumatran orangutan | • Tarantula | |
| • Red panda | • Red-eyed tree frog | • Uakari | |

Box 1. Animals dataset: 100 entities used in our evaluation. Dataset taken from Choudhury et al. (2025).

Plants dataset

• Rose	• Aspen	• Lilac	• Impatiens
• Oak	• Sycamore	• Ranunculus	• Mimosa
• Pine	• Hawthorn	• Azalea	• Forsythia
• Palm	• Holly	• Gladiolus	• Delphinium
• Redwood	• Marigold	• Hibiscus	• Bellflower
• Elm	• Lily	• Daffodil	• Aster
• Cedar	• Primrose	• Petunia	• Trillium
• Juniper	• Hyacinth	• Snowdrop	• Hydrangea
• Hickory	• Lotus	• Begonia	• Zinnia
• Willow	• Crocus	• Freesia	• Agave
• Bamboo	• Anemone	• Cornflower	• Aloe
• Hemlock	• Carnation	• Hellebore	• Cactus
• Poplar	• Iris	• Dahlia	• Saguaro
• Birch	• Bluebell	• Peony	• Fern
• Cypress	• Lavender	• Violet	• Philodendron
• Magnolia	• Orchid	• Camellia	• Cyclamen
• Eucalyptus	• Poppy	• Hollyhock	• Lupin
• Maple	• Pansy	• Alyssum	• Rudbeckia
• Fir	• Dandelion	• Geranium	• Yarrow
• Beech	• Sunflower	• Gardenia	• Foxglove
• Mahogany	• Narcissus	• Scilla	• Wisteria
• Spruce	• Calendula	• Primula	• Clematis
• Dogwood	• Fuchsia	• Tuberosa	• Honeysuckle
• Yew	• Jasmine	• Poinsettia	• Allium
• Larch	• Chrysanthemum	• Amaryllis	• Viburnum

Box 2. Plants dataset: 100 entities used in our evaluation. Dataset based on Auzina et al. (2026).

D. Extended Results

Table 4. Success rate (mean \pm std over episodes) for the baseline, ASIG, and BED-LLM on Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct across the animals and plants categories. ASIG consistently improves over the baseline across all settings. BED-LLM outperforms ASIG on plants but falls below both ASIG and the baseline on the 14B animals category, suggesting it does not generalise robustly across model sizes and categories.

Model	Category	Baseline	Ours (ASIG)	BED-LLM
Qwen2.5-7B-Instruct	Animals	7.6 \pm 1.8	21.0 \pm 4.8	19.2 \pm 4.0
	Plants	6.8 \pm 1.6	16.6 \pm 2.1	22.8 \pm 2.2
Qwen2.5-14B-Instruct	Animals	35.0 \pm 2.0	39.2 \pm 3.6	26.4 \pm 4.8
	Plants	11.8 \pm 1.9	13.6 \pm 0.9	21.0 \pm 4.1

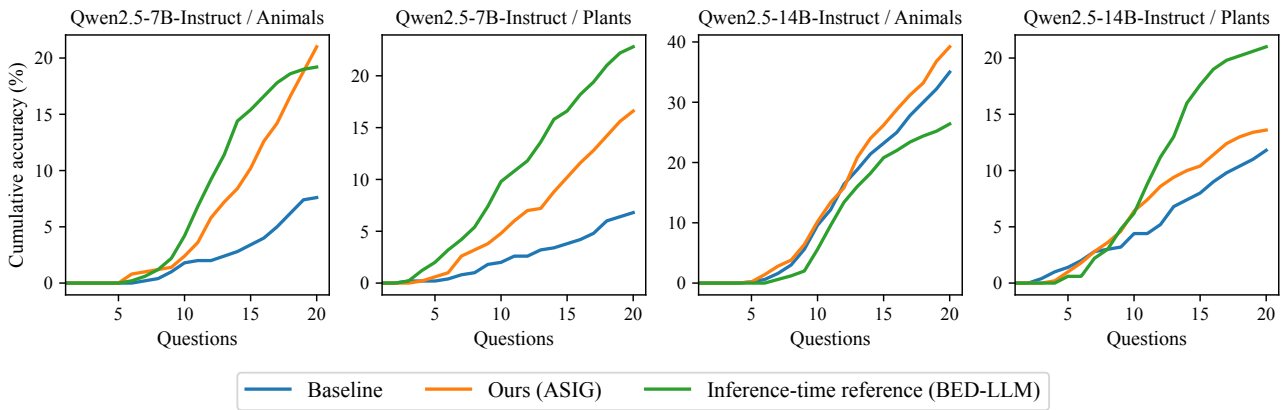


Figure 7. Cumulative fraction of games correctly identified within N questions (average over 5 episodes) for Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct across the animals and plants categories. ASIG outperforms the baseline at almost every turn and across the four settings. It also narrows the gap to or even outperforms the inference-time reference BED-LLM, despite using only a fraction of the inference-time compute.

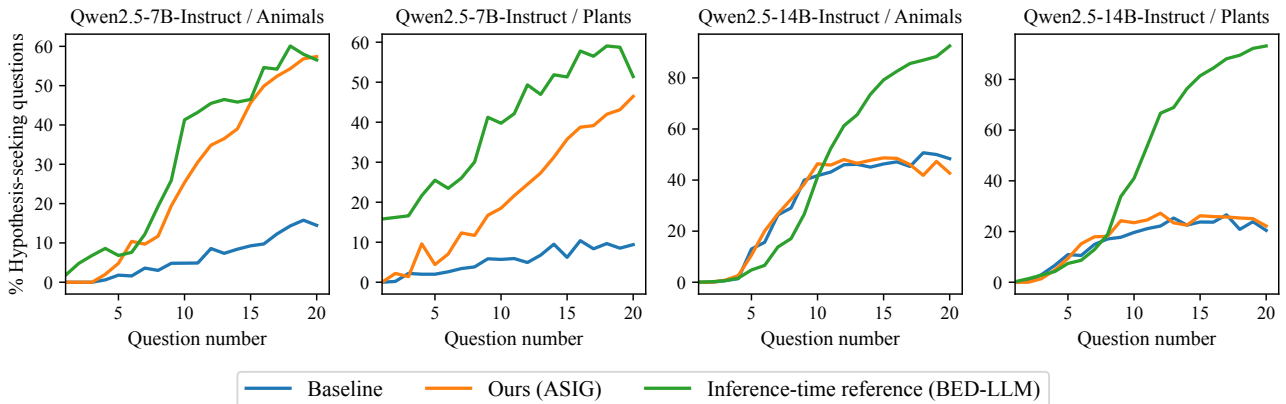


Figure 8. Fraction of questions classified as hypothesis-seeking at each turn for Qwen2.5-7B-Instruct and Qwen2.5-14B-Instruct across the animals and plants categories. All methods become more hypothesis-seeking as the game progresses. For the 7B model, ASIG consistently and substantially exceeds the baseline, suggesting that fine-tuning encourages earlier explicit guesses. For the 14B model, ASIG’s hypothesis-seeking rate closely tracks the baseline across both categories, indicating a weaker behavioural shift. BED-LLM shows the highest rate overall, though this partly reflects its hard-coded guessing rules rather than an emergent strategy as in ASIG.

E. Example Games

Target: `Fir`

Baseline	Failed (20 Q)	Ours (ASIG)	Solved (16 Q)
1. Does the plant require direct sunlight for at least 6 hours a day?	✗	1. Does the plant have leaves?	✓
2. Does the plant have spines or thorns on its stems or leaves?	✗	2. Is the plant a deciduous tree?	✗
3. Is the plant native to a tropical or subtropical climate?	✗	3. Does the plant have flowers?	✗
4. Does the plant shed its leaves annually?	✗	4. Is the plant a fern?	✗
5. Does the plant have flowers that are predominantly one solid color?	✗	5. Does the plant have thorns?	✗
6. Does the plant have compound leaves with multiple leaflets?	✗	6. Is the plant a grass or a grain?	✗
7. Does the plant produce fruit that splits open when ripe?	✗	7. Is the plant a shrub?	✗
8. Does the plant have a distinctive odor when crushed?	✓	8. Is the plant a vine?	✗
9. Is the odor sweet and reminiscent of almond?	✗	9. Is the plant a conifer?	✓
10. Does the plant have glossy leaves?	✗	10. Is the plant a deciduous conifer?	✗
11. Does the plant have variegated leaves with different colors?	✗	11. Is the plant an evergreen conifer?	✓
12. Does the plant produce small, inconspicuous flowers?	✗	12. Does the plant have needle-like leaves?	✓
13. Does the plant have large, broad leaves?	✗	13. Is the plant a pine?	✗
14. Does the plant have needle-like leaves?	✓	14. Is the plant a cedar?	✗
15. Is the plant a conifer?	✓	15. Is the plant a spruce?	✗
16. Is the plant evergreen?	✓	16. Is the plant a fir?	✓
17. Does the plant produce cones?	✓		
18. Is the plant a pine tree?	✗		
19. Does the plant produce winged seeds?	✓		
20. Is the plant a maple tree?	✗		

Box 3. Example 20 Questions game with target `Fir`. The baseline fails to converge within the 20-question budget, while our fine-tuned 7B model identifies the target in 16 questions.

935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989

Target: Honey bee			
Baseline	Failed (20 Q)	Ours (ASIG)	Solved (11 Q)
1. Is the animal domesticated?	✗	1. Is the animal a mammal?	✗
2. Does the animal live primarily on land?	✓	2. Is the animal cold blooded?	✓
3. Is the animal a predator?	✗	3. Does the animal live primarily on land?	✓
4. Does the animal live in herds or groups?	✓	4. Is the animal larger than a typical house cat?	✗
5. Is the animal an herbivore?	✓	5. Does the animal have wings?	✓
6. Is the animal larger than a car?	✗	6. Can the animal fly?	✓
7. Is the animal a type of deer or antelope?	✗	7. Is the animal a bird?	✗
8. Does the animal have horns or antlers?	✗	8. Is the animal an insect?	✓
9. Is the animal a pig or related to pigs?	✗	9. Is the insect a butterfly or moth?	✗
10. Is the animal a type of sheep or goat?	✗	10. Is the insect a bee or wasp?	✓
11. Is the animal a type of rodent?	✗	11. Is the insect a honey bee?	✓
12. Is the animal a horse or related to horses?	✗		
13. Is the animal a type of elephant or relative to elephants?	✗		
14. Is the animal a type of camel or related to camels?	✗		
15. Is the animal a type of primate?	✗		
16. Is the animal a type of wild boar or wild pig?	✗		
17. Is the animal a type of buffalo or bison?	✗		
18. Is the animal a type of llama or alpaca?	✗		
19. Is the animal a type of rhinoceros?	✗		
20. Is the animal a type of hippopotamus?	✗		

Box 4. Example 20 Questions game with target Honey bee. The baseline fails to converge within the 20-question budget, while our fine-tuned 14B model identifies the target in 11 questions.