ACTIVE HYPOTHESIS ANALYSIS FOR LANGUAGE MODEL REASONING UNDER AMBIGUITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reasoning from text about relationships among variables is challenging when multiple structural interpretations are compatible with the evidence. Given a set of correlations and conditional independencies, a reasoning model must determine whether a proposed relation is entailed without knowing which interpretation is correct. We introduce *Active Hypothesis Analysis for Language Models under Ambiguity* (AHA), an inference-time procedure that explicitly represents competing interpretations and actively tests them. AHA combines a frozen language model with a deterministic controller: the model proposes candidate interpretations and answers targeted local questions, while the controller selects questions by expected information gain and maintains a posterior over hypotheses. We provide a kernel-based result relating representation similarity to differences in hypothesis scores, and a finite-budget guarantee for posterior concentration under candidate coverage, informative queries, and calibrated responses. On Corr2Cause, AHA achieves 93.2 ± 0.4 F1, compared with 51.6 for budget-matched self-consistency and 74.3 for random-query Bayesian aggregation using the same GLM-5.1 backbone. Local-question accuracy ranges from 62.4% to 85.9%, with concentration degrading as response noise approaches random guessing. We further introduce Extended Corr2Cause, an all-negative rejection stress test. Results on both benchmarks show that explicit uncertainty can direct inference-time computation toward questions that distinguish competing interpretations.

1 INTRODUCTION

Reasoning with language is challenging, not only because a problem may require many steps, but because the available information may support multiple plausible interpretations. A model can therefore produce a coherent reasoning trace while still committing to an interpretation not warranted by the evidence. This creates a key challenge: when several interpretations remain plausible, the model must determine which uncertainties matter for the conclusion and how to resolve them.

Additional inference-time computation can improve reasoning. Chain-of-thought prompting generates intermediate steps (Wei et al., 2022), self-consistency aggregates sampled solutions (Wang et al., 2023), and reasoning-oriented models and test-time scaling allocate larger inference budgets (DeepSeek-AI, 2025; Muennighoff et al., 2025; Snell et al., 2024). These approaches change the amount or organization of reasoning, but leave an important design question open: *when a model is uncertain between competing hypotheses, what should the next unit of inference-time computation test?*

In this work, we view *generating alternative hypotheses* and *distinguishing among alternative hypotheses* as different reasoning problems. Suppose a model constructs several interpretations of the same premise. Generating another complete explanation may provide a further plausible account without resolving the ambiguity among the interpretations. Instead, the model can identify a local question on which the candidate explanations disagree. Answering this question can provide targeted evidence with which competing interpretations can be reweighted. Useful inference-time computation is thus not merely more reasoning, but computation directed at resolving decision-relevant uncertainty. We study this question through *structured textual reasoning*. Each input contains natural-language premises and a hypothesis to accept or reject. The benchmarks used here, Corr2Cause (Jin et al., 2024) and a new extension, Extended Corr2Cause, express correlations and conditional independencies

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.

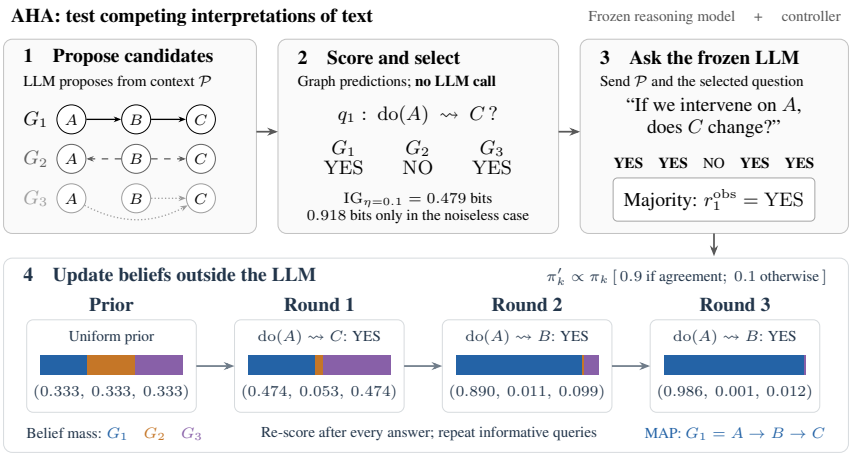


Figure 1: **Active testing of three structured textual interpretations.** Questions are linguistic checks, not physical interventions. The illustrative candidates need incomplete context; they do not form one MEC. At assumed aggregated error $\eta = 0.1$, exact posteriors are $(9, 1, 9)/19$, $(81, 1, 9)/91$, and $(729, 1, 9)/739$. The second query predicts (YES, NO, NO). Five votes do not establish the assumed error rate, and crossing a 95% posterior threshold is not a certified stopping guarantee.

in text. Despite the causal vocabulary, the interface is linguistic: models receive text and return judgments, with no numerical samples or real interventions, so we make no claim of discovering real-world causal mechanisms.

We introduce *Active Hypothesis Analysis* (AHA), an inference-time procedure for this problem. AHA is not a multi-agent system: one frozen language model proposes structured candidates and answers local intervention-style questions; a deterministic controller evaluates candidate predictions, chooses a question by expected information gain, and performs Bayesian reweighting. Figure 1 illustrates this active "twenty questions" calculation over three candidate interpretations. The questions are linguistic reasoning probes, not interventions on a physical system. Our theory has two parts: a bounded-norm RKHS inequality limiting fixed score margins for closely represented candidates, and a finite-budget bound on posterior concentration under a reliable response channel and sufficient query coverage. Neither creates information missing from the prompt, and neither is a universal impossibility theorem.

What we find and why it matters. With the same frozen GLM-5.1 backbone, AHA reaches 93.2 F1 on Corr2Cause, compared with 34.2 for direct prediction, 51.6 for budget-matched self-consistency, and 74.3 for random-query Bayesian aggregation. Together with the isolated-question and noise studies, this indicates that the useful component is not merely more calls: candidate disagreement must direct those calls, and the local response channel must remain informative. AHA is thus a testable inference principle – maintain alternatives, query their disagreements, preserve uncertainty – not a claim that self-querying creates unavailable evidence.

Contributions. **(1)** We define text-based reasoning as discrimination among structured candidate interpretations and separate benchmark labels, deterministic candidate predictions, and model responses. **(2)** We introduce AHA, a one-model inference procedure whose explicit posterior selects the next local question. **(3)** We prove conditional results for direct score margins and finite-budget posterior concentration, with their assumptions stated separately. **(4)** We evaluate the mechanism through same-backbone controls, acquisition ablations, label-wise results, isolated local questions, repeated runs, and noise sensitivity.

2 Related Work: Reasoning Models and Inference Procedures

Traces, decomposition, and aggregation. CoT (Wei et al., 2022) and least-to-most prompting (Zhou et al., 2023) organize a solution into intermediate steps. Self-consistency (Wang et al., 2023) changes answer selection by aggregating sampled reasoning paths. AHA instead maintains *persistent*

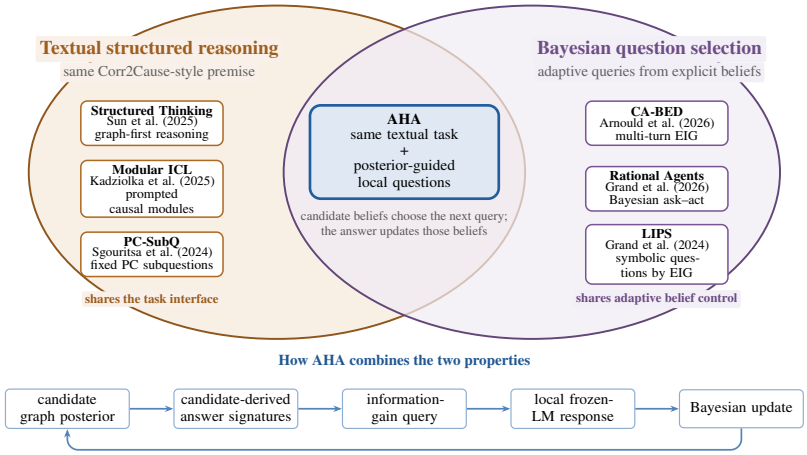


Figure 2: **AHA combines the two closest lines of work.** Corr2Cause-specific approaches share the textual structured-reasoning interface but do not maintain an explicit candidate posterior that adaptively chooses the next question (Sgouritsa et al., 2024; Sun et al., 2025; Kadziolka & Salehkaleybar, 2025). Bayesian information-seeking approaches choose questions from hypotheses or beliefs, but study different interactive tasks (Grand et al., 2024; 2026; Arnould et al., 2026). AHA uses posterior disagreement among textual candidate interpretations to choose a local query, then updates the posterior from the frozen model's answer. The figure compares mechanisms, not scores.

candidate beliefs and chooses the next subquestion from disagreement among those candidates; it does not equate the most frequently sampled explanation with the most supported interpretation.

Search, feedback, and agents. Tree of Thoughts (Yao et al., 2023a) and Graph of Thoughts (Besta et al., 2024) structure candidate reasoning computations. ReAct (Yao et al., 2023b) combines reasoning with actions; Self-Refine (Madaan et al., 2023) and Reflexion (Shinn et al., 2023) use feedback to revise outputs or memory. LATS (Zhou et al., 2024) combines tree search, model evaluation and environmental feedback. These methods already place substantial computation outside a single generated answer. AHA's distinction is therefore not "external control" alone, but a particular finite-candidate state, deterministic prediction interface, and likelihood-based update.

Training and test-time scaling. DPO (Rafailov et al., 2023), RL for reasoning (DeepSeek-AI, 2025), reasoning-trace training (Muennighoff et al., 2025) and adaptive test-time allocation (Snell et al., 2024) are complementary to AHA. Our cross-backbone comparisons do not isolate the effect of reasoning-oriented training.

Probabilistic composition and active questions. Language Model Cascades (Dohan et al., 2022) formalize model compositions as probabilistic programs. LIPS (Grand et al., 2024) uses symbolic question programs and expected information gain; Grand et al. (2026) study Bayesian information-seeking agents; CA-BED (Arnould et al., 2026) maintains beliefs and plans informative dialogue using model-estimated likelihoods. These are close methodological precedents. AHA specializes active testing to structured interpretations of a fixed textual premise, with candidate-derived binary predictions and a calibrated-channel analysis. It does not claim to introduce Bayesian question selection.

Closest methods. Figure 2: Corr2Cause approaches share our textual input but not posterior-guided adaptive querying (Sgouritsa et al., 2024; Sun et al., 2025; Kadziolka & Salehkaleybar, 2025); LIPS, information-seeking agents and CA-BED share belief-based question selection but study other tasks (Grand et al., 2024; 2026; Arnould et al., 2026). AHA combines the two. Data-assisted frameworks (Li et al., 2025; Le et al., 2024) need a different observation interface (see appendix). We make no numerical superiority claim without a shared protocol.

3 PRELIMINARIES: TEXTUAL ENTAILMENT WITH COMPETING INTERPRETATIONS

Task and representations. An instance is $x = (\mathcal{P}, h)$: a textual premise and a proposed conclusion. Its benchmark label $\ell \in \{0, 1\}$ specifies whether the conclusion is entailed. AHA constructs a finite set $\mathcal{G} = \{G_1, \ldots, G_n\}$ of structured interpretations of the text. For our benchmarks these are symbolic DAGs over the variables named in the premise; they are reasoning artifacts, not graphs estimated from numerical observations. We distinguish three quantities throughout:

$$\ell \text{ (benchmark label)}, \qquad \hat{r}_k(q) \text{ (candidate prediction)}, \qquad r_t^{\text{obs}} \text{ (LM response)}. \tag{1}$$

Only the last is obtained by asking the model a local question. The first is used for evaluation, not supplied to the inference controller.

Binary textual checks. Let $\mathcal{V} = \{V_1, \ldots, V_d\}$ and $\mathcal{Q} = \{(V_i, V_j) : i \neq j\}$. A question asks whether V_j would be affected by a change to V_i in the system described by the text. For candidate G_k, define $\hat{r}_k(V_i, V_j) = 1$ exactly when there is a directed path from V_i to V_j. This reachability calculation is deterministic. The phrase "intervention" specifies the question's semantics; no physical intervention or new numerical observation occurs. Its interpretation as a measured distributional effect would additionally require an SCM, suitable intervention values and noncancelling effects, which are not part of our experimental interface.

Definition 1 (Candidate discrimination set)**.** *For two interpretations,* $\mathcal{D}(G_k, G_j) = \{q \in \mathcal{Q} : \hat{r}_k(q) \neq \hat{r}_j(q)\}$.

Candidate signatures need not identify every edge: adding $A \to C$ to $A \to B \to C$ does not change reachability. The analysis therefore concerns distinguishable signatures, or a candidate set separated by the chosen checks.

Entailment versus selecting one interpretation. If $\mathcal{C}(\mathcal{P})$ denotes all premise-consistent interpretations, the logical target is

$$\ell(\mathcal{P}, h) = \mathbb{1}_{\{h(G)=1 \text{ for every } G \in \mathcal{C}(\mathcal{P})\}}. \tag{2}$$

A finite candidate approximation is not a completeness certificate: a consistent counterexample can refute entailment, whereas finding one supporting DAG cannot establish it. The following definition specifies how every reported prediction here is obtained from the pool.

Definition 2 (Reported hypothesis judgment)**.** *Given the proposed pool* $\{G_1, \ldots, G_n\}$ *and cached predictions* $\hat{r}_k$*,*

$$\hat{\ell}(\mathcal{P}, h) = \mathbb{1}_{\{h(G_k)=1 \text{ for every } k=1,\ldots,n\}}. \tag{3}$$

This unanimity rule treats the pool as a proxy for $\mathcal{C}(\mathcal{P})$: a single dissenting candidate forces non-entailment. It is exact when the pool covers $\mathcal{C}(\mathcal{P})$, and biased toward rejection otherwise (Appendix F).

4 ACTIVE HYPOTHESIS ANALYSIS

Propose, predict, query, update. AHA is not a multi-agent architecture: one frozen language model proposes candidate interpretations and answers selected questions, while a deterministic external controller manages the belief state. The controller caches candidate predictions, scores discriminating questions, and sends only the selected question back to the model with the original text. An odd number M of responses may be aggregated by majority. Candidate weights are updated externally; no model gradients are computed. Retaining an explicit state makes the source of each change in belief inspectable.

Let $\eta_{\text{upd}} \in (0, 1/2)$ be the noise parameter used by the controller, and define

$$L_k(r \mid q) = (1 - \eta_{\text{upd}})\mathbb{1}_{\{r=\hat{r}_k(q)\}} + \eta_{\text{upd}}\mathbb{1}_{\{r\neq\hat{r}_k(q)\}}, \tag{4}$$

$$p_t(r \mid q) = \sum_k \pi_k^{(t-1)} L_k(r \mid q), \qquad \pi_k(r, q) = \frac{\pi_k^{(t-1)} L_k(r \mid q)}{p_t(r \mid q)}, \tag{5}$$

$$\text{IG}_t(q) = H(\boldsymbol{\pi}^{(t-1)}) - \sum_{r=0}^{1} p_t(r \mid q) H(\boldsymbol{\pi}(r, q)). \tag{6}$$

Algorithm 1 Active Hypothesis Analysis

Require: Text $\mathcal{P}$, frozen model $\mathcal{L}$, checks $\mathcal{Q}$, budget T, odd M, η_{upd}, exploration ε
Ensure: Selected structured interpretation and candidate beliefs

```
Propose, validate and deduplicate {G_1, ..., G_n} with L
if n = 0 then
    return abstain
end if
Set π_k^(0) = 1/n; cache r̂_k(q) for all k, q
for t = 1, ..., T do
    Compute IG_t(q) for each check using Eq. (6)
    With probability ε, sample q_t ~ Unif(Q);
        otherwise choose q_t ∈ arg max_q IG_t(q)
    Ask L the question (P, q_t) M times; aggregate to r_t^obs
    π_k^(t) ← π_k^(t−1) L_k(r_t^obs | q_t) / p_t(r_t^obs | q_t)
end for
return (G_{arg max_k π_k^(T)}, π^(T))
```

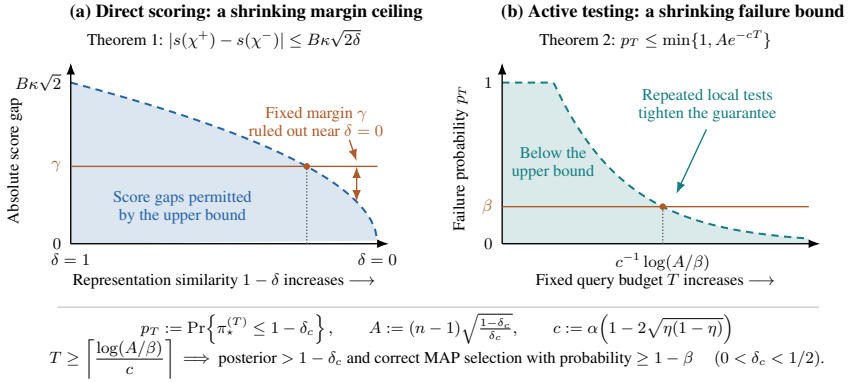


Figure 3: Overview of the two conditional guarantees for AHA. **Left:** for equal-norm candidate representations and a bounded-norm RKHS scorer, increasing representation similarity reduces the score-margin upper bound (Theorem 1). A fixed margin γ is ruled out when $B\kappa\sqrt{2\delta} < \gamma$; this does not imply classification failure. **Right:** under target support, calibrated local responses and query coverage, the posterior-concentration failure bound decreases exponentially with the fixed query budget (Theorem 2). Shading marks values below each upper bound, not proven attainable values; the horizontal axes are different and their scales are schematic. The concentration result concerns candidate selection, not an entailment certificate or recovery of a physical causal graph.

Entropy is measured in bits. The same likelihood is used for acquisition and updating. With probability $\varepsilon > 0$, select uniformly from $\mathcal{Q}$; otherwise maximize IG. This exploration branch ensures that a useful question is not permanently excluded by current beliefs.

Worked example and compute. For the three candidates in Figure 1, the first check predicts $(1, 0, 1)$. At $\eta_{\text{upd}} = 0.1$, its noisy IG is 0.479 bits, compared with 0.918 bits for a noiseless response. A YES response gives posterior $(9, 1, 9)/19$. Checking $A \rightsquigarrow B$ predicts $(1, 0, 0)$; two further YES responses yield $(81, 1, 9)/91$ and $(729, 1, 9)/739$. Each observation changes the next decision problem. For a fixed pool, reachability caching costs $O(nd^3)$ with a straightforward transitive closure, acquisition costs $O(Tn|\mathcal{Q}|)$, and response sampling uses at most TM calls plus candidate-generation calls. Thinking tokens and candidate generation must also be counted in cost comparisons.

5 Analysis: Direct Scores and Active Tests

5.1 A conditional limitation of direct candidate scores

A PSD kernel K induces features ϕ and an RKHS $\mathcal{H}$. Consider a direct score $s(\chi) = \langle w, \phi(\chi) \rangle$ for a textual premise–candidate representation $\chi = (x, y)$, with $\|w\| \le B$ and feature norms at most κ. The parameter B bounds the scorer, not the number of reasoning tokens.

Definition 3 (δ-similar representations)**.** *For nonzero features with $a_\pm = \|\phi(\chi^\pm)\|$, a pair is δ-similar when $K(\chi^+, \chi^-)/(a_+a_-) \geq 1 - \delta$, where $0 \leq \delta \leq 2$.*

Theorem 1 (Bounded-norm score-margin obstruction)**.** *For a δ-similar pair,*

$$|s(\chi^+) - s(\chi^-)| \leq B\sqrt{(a_+ - a_-)^2 + 2a_+a_-\delta}. \tag{7}$$

If $a_+ = a_- \leq \kappa$, the bound is $B\kappa\sqrt{2\delta}$. For $\delta > 0$, enforcing a fixed margin $\gamma > 0$ requires

$$B \geq \frac{\gamma}{\kappa\sqrt{2\delta}}. \tag{8}$$

The proof follows from Cauchy–Schwarz and the feature-distance identity (Appendix A). This limits a *fixed score margin*; it does not prove nonzero classification error or inevitable failure of reasoning models. SFT/DPO require controlled linearization, scorer norm and initial-score contributions before an NTK argument applies (Jacot et al., 2018). Kernel views of ICL also have separate assumptions (Han et al., 2023). Appendix A.5 states these requirements. The theorem motivates testing a different readout, not declaring every direct reasoning method impossible.

Lemma 2 in Appendix A gives sufficient token-feature conditions for $\delta = O(d^{-2})$. Shared text or hidden-state cosine similarity alone does not establish this scaling for a transformer's sequence-score NTK.

5.2 When local tests concentrate candidate beliefs

Assumption 1 (Candidate support and calibrated local responses)**.** *There is a target interpretation $G^\star$ in the fixed pool, all priors are positive, and $\mathcal{D}(G^\star, G_k) \neq \varnothing$ for every $k \neq \star$. Conditional on the complete history $\mathcal{F}_{t-1}$ and selected check q_t,*

$$\Pr(r_t^{\text{obs}} = \hat{r}_\star(q_t) \mid \mathcal{F}_{t-1}, q_t) = 1 - \eta, \qquad 0 < \eta < \tfrac{1}{2}, \tag{9}$$

and $\eta_{\text{upd}} = \eta$. If voting is used, η is the error of the aggregated *answer. This is a model of local-check reliability, not a consequence of producing a binary token or a claim of additional experimental evidence.*

Lemma 1 (Response-channel separation)**.** *Under Eq.* (9)*, a discriminating check induces response distributions with total variation $1 - 2\eta$ and Bhattacharyya coefficient $b_\eta = 2\sqrt{\eta(1-\eta)} < 1$. An undiscriminating check has coefficient one.*

The statement follows by comparing $(\eta, 1-\eta)$ with $(1-\eta, \eta)$. It concerns answer distributions, not unproved separation of neural YES/NO embeddings.

Theorem 2 (Finite-budget concentration under query coverage)**.** *Assume Assumption 1, $n \geq 2$, a uniform prior, and $\Pr(q_t \in \mathcal{D}(G^\star, G_k) \mid \mathcal{F}_{t-1}) \geq \alpha > 0$ for each wrong candidate and round. Let $c = \alpha(1 - b_\eta)$. For a fixed budget T and $0 < \delta_c < 1$,*

$$\Pr\{\pi_\star^{(T)} \leq 1 - \delta_c\} \leq (n-1)\sqrt{\frac{1-\delta_c}{\delta_c}}e^{-cT}. \tag{10}$$

For $\delta_c < 1/2$ and failure tolerance $0 < \beta < 1$, sufficient budget for correct MAP selection with posterior $> 1 - \delta_c$ is

$$T \geq \left\lceil \frac{\log((n-1)/\beta) + \frac{1}{2}\log((1-\delta_c)/\delta_c)}{\alpha[1 - 2\sqrt{\eta(1-\eta)}]} \right\rceil. \tag{11}$$

The exploration branch guarantees $\alpha \geq \varepsilon/|\mathcal{Q}|$ when every wrong candidate is distinguishable.

The proof contracts expected square-root posterior odds under query coverage, then applies Markov's inequality; Appendix A gives the full adaptive argument. The dependence on candidate count is logarithmic only at fixed α and η; for pairwise checks, the coverage lower bound can scale as $\varepsilon/[d(d-1)]$. The theorem applies at a fixed horizon, not automatically at a data-dependent early stop. The implemented procedure preserves model parameters, but frozen weights and a probability-simplex state do not remove information-theoretic ambiguity (Proposition 1). In textual reasoning, the intended benefit is better use of available model knowledge through targeted checks, subject to these explicit conditions.

Table 1: **Does the controller improve the same frozen model?** Zero-shot versus AHA; differences are percentage points. ⋆: thinking enabled. Extended F1 retains its original designation; accuracy is rejection rate. Corr2Cause accuracy is reconciled to Table 5.

(a) Corr2Cause

	Zero-shot		+AHA		Difference	
Backbone	F1	Acc.	F1	Acc.	ΔF1	ΔAcc.
LLaMA-7B	26.8	17.4	72.6	86.1	+45.8	+68.7
Gemma-3-12B	27.4	30.2	77.6	89.4	+50.2	+59.2
Qwen3.5-122B	30.1	33.4	87.1	93.9	+57.0	+60.5
Qwen3-30B⋆	32.8	36.1	92.1	96.5	+59.3	+60.4
GLM-5.1⋆	34.2	37.5	**93.2**	**97.1**	+59.0	+59.6

(b) Extended Corr2Cause, $d = 7$–24

	Zero-shot		+AHA		Difference	
Backbone	F1	Acc.	F1	Acc.	ΔF1	ΔAcc.
LLaMA-7B	37.8	39.2	52.9	54.6	+15.1	+15.4
Gemma-3-12B	42.9	44.7	56.2	57.9	+13.3	+13.2
Qwen3.5-122B	52.6	54.8	73.7	75.1	+21.1	+20.3
Qwen3-30B⋆	57.8	60.1	81.0	82.3	+23.2	+22.2
GLM-5.1⋆	59.9	62.4	**81.9**	**83.2**	+22.0	+20.8

6 Experiments: Does Active Testing Improve Reasoning?

6.1 Evaluation design

We test whether active candidate discrimination improves natural-language premise–hypothesis judgments. Corr2Cause provides 7,524 test examples with $d = 2$–6 variables and six relation templates (Jin et al., 2024); Extended Corr2Cause provides 18,000 all-negative test examples with $d = 7$–24. Here d is the number of named variables, not reasoning depth. AHA wraps six frozen backbones, using $n = 8$ candidates, at most $T = 20$ query rounds, exploration $\varepsilon = 0.1$, $M = 3$ responses per query, and nominal update noise $\eta_{\text{upd}} = 0.1$. The starred models use thinking mode. RoBERTa-Large SFT and DPO use approximately 1.3M extended training examples; the appendix records complete settings and partitions.

Four tests of the mechanism. We ask whether (Q1) the controller improves the *same* frozen model, (Q2) information-gain selection beats random questions under the same Bayesian update, (Q3) local responses are accurate enough to support belief revision, and (Q4) gains persist as the response channel becomes noisy. This separates the effects of backbone quality, additional calls, Bayesian aggregation, query acquisition, and response reliability.

Metrics and reporting. Table 1 compares the same models with and without AHA; Table 2 compares AHA with training baselines, including all six relation labels. Corr2Cause reports positive-class (entailment) F1, following Jin et al. (2024)'s convention, and accuracy. The extended test is described as all-negative, so its accuracy measures rejection of unsupported statements; always rejecting achieves 100%. Extended F1-labelled columns are retained as originally reported, not silently relabelled. Individual model breakdowns, and the two historical size-band snapshots that could not be reconciled with the reported figure, are preserved in Appendix D.

6.2 Results: targeted tests outperform repeated answers

Active control consistently improves the same frozen model. Table 1 compares each backbone with and without AHA. Zero-shot is weak throughout: even the strongest backbones reach only 32.8–34.2 F1 on Corr2Cause. The entailment label quantifies over all premise-consistent structures, but a direct answer evaluates the hypothesis against only the one the model happens to adopt. LLaMA-7B shows the resulting failure most clearly: its accuracy (17.4) falls below its F1 (26.8), which on a benchmark dominated by non-entailed hypotheses means it accepts most of them. Across all five backbones, AHA improves Corr2Cause F1 by 45.8–59.3 percentage points and extended rejection accuracy by

Table 2: **How does AHA compare with SFT and DPO?** Extended Corr2Cause comparison across methods, variable counts and labels. Panel (a): overall/size-band accuracies; panel (b): per-label gains. Scores are percentages; differences are percentage points. ‡Overall is the pooled accuracy (Table 6), not the size-band mean (Appendix D).

(a) Overall and variable-count comparison

Method	Overall‡	7–10	11–15	16–20	21–24
RoBERTa-L SFT	52.2	69.7	58.2	45.8	35.1
RoBERTa-L DPO	58.0	73.0	63.4	52.7	42.9
AHA (GLM-5.1⋆)	**83.2**	**91.4**	**88.7**	**84.2**	**79.8**
Δ (AHA − SFT)	+31.0	+21.7	+30.5	+38.4	+44.7
Δ (AHA − DPO)	+25.2	+18.4	+25.3	+31.5	+36.9

(b) Label-wise comparison: rejection accuracy by relation template

Relation label	SFT	DPO	AHA	Δ vs. SFT	Δ vs. DPO
Parent	54.7	60.6	**84.8**	+30.1	+24.2
Child	53.2	58.8	**84.1**	+30.9	+25.3
Ancestor	50.8	56.8	**81.2**	+30.4	+24.4
Descendant	50.3	56.0	**80.2**	+29.9	+24.2
Collider	46.5	51.6	**75.2**	+28.7	+23.6
Confounder	57.8	64.1	**86.0**	+28.2	+21.9

13.2–22.2 points. The strongest example is GLM-5.1, which rises from 34.2 to 93.2 Corr2Cause F1 without changing model parameters.

The gain in performance is not explained by additional calls alone. On the same GLM-5.1 backbone and budget, self-consistency reaches 51.6 F1, random questions with Bayes reach 74.3, and information-gain questions with Bayes reach 93.2 (Table 3). The 18.9-point random-to-IG gap is the cleanest evidence that choosing *which uncertainty to test* matters beyond repeated generation and aggregation. Moreover maintaining explicit beliefs over hypotheses as AHA does is crucial since these determine what to ask next: posterior-guided questions beat random ones by 18.9 F1 under an identical update.

Reported gains span relation types and larger variable sets. Table 2(a) reports the reconciled comparison: SFT and DPO reach 52.2 and 58.0 overall accuracy, versus 83.2 for AHA (+31.0 and +25.2 points). This is a cross-backbone comparison; it does not isolate the effect of training. The size-band breakdown alongside it (7–10 through 21–24) is a separately sourced historical snapshot whose own mean (86.0) does not equal the pooled 83.2; a second, further-conflicting size-band snapshot (mean 78.2) is retained in the appendix as an audit record rather than merged with either. We therefore do not draw a size-dependent conclusion from the band breakdown in the main text.

Panel (b) of Table 2 uses the existing per-template accuracy entries, with gains computed as $(AHA - Baseline)$ All six labels improve over both baselines: gains are 28.2–30.9 points over SFT and 21.9–25.3 over DPO. Collider remains the lowest-scoring template for all three methods (46.5, 51.6 and 75.2), despite a 23.6-point AHA gain over DPO. Thus the reported advantage is not confined to one relation. These are relation-template rejection rates, not separate positive/negative-class results. Per-template zero-shot scores, positive-class recall and significance tests cannot be recovered from the supplied aggregates.

Most information is gained in the first few rounds. Figure 4 shows MAP accuracy rising and entropy falling most steeply at the start, with per-round information gain peaking at round 2 and declining thereafter. Early on, candidates disagree on many questions, so a single answer can split posterior mass roughly in half; once belief concentrates on a few candidates, fewer questions separate them evenly and each answer moves the posterior less. Most of the benefit therefore arrives well within the 20-round budget.

The gain depends on the reliability of local answers. Backbones answer 62.4–85.9% of isolated local questions correctly, and those with less reliable answers concentrate more slowly. The noise sweep isolates this effect: as answer error approaches $1/2$, each answer carries almost no information, and convergence slows while accuracy falls (Table 4).

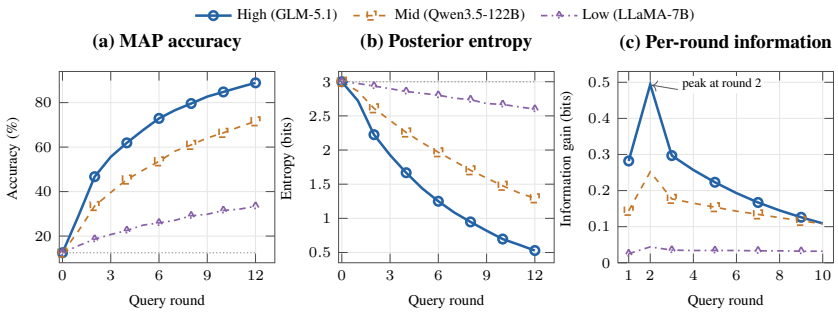


Figure 4: **How the belief state changes across query rounds.** All configurations begin from the uniform eight-candidate prior (12.5% MAP accuracy; 3 bits entropy). At round 12, high/mid/low configurations reach 88.9/71.6/33.4% accuracy and 0.53/1.29/2.60 bits of entropy. Round-2 information gain is 0.494/0.252/0.044 bits.

Table 3: **What drives the gain?** Same-backbone GLM-5.1* comparison; C2C reports positive-class F1/accuracy, Extended rejection is percent rejected on the all-negative slice. Rows A–D isolate calls, Bayesian aggregation and IG selection. Row D's accuracy is reconciled to Table 5. No $T = 0$, candidates-only row is included; Row C does not substitute for it (§6.2).

Inference configuration	**C2C F1**	**C2C accuracy**	**Extended rejection**
A: Direct prediction	34.2	37.5	62.4
B: Budget-matched self-consistency	51.6	54.3	68.8
C: Random questions + Bayes	74.3	76.9	73.5
D: **AHA: IG questions + LM responses + Bayes**	**93.2**	**97.1**	**83.2**
IG questions + synthetic noise ($\eta = 0.35$)	71.8	74.2	65.1
IG questions + reference-graph responses	97.4	98.1	96.3

Table 4: **Response-noise sensitivity on Extended Corr2Cause.** Rejection accuracy and stopping within a 20-round budget; empirical summaries, not Theorem 2 convergence times. "$>$ 20": not reached within budget. η here is not distinguished from η_{upd} (Appendix F).

Reported noise setting η	0.10	0.25	0.40	0.48
Rejection accuracy (%)	83.2	78.4	67.1	51.8
Mean rounds to convergence	9.2	12.1	19.4	$>$ 20

Limitations. The experiments remain limited to Corr2Cause-style textual entailment. Extended Corr2Cause is an all-negative rejection stress test, and the cross-backbone SFT/DPO comparison does not isolate training from inference control. A sharply peaked posterior can still be wrong under biased responses. Since Proposition 1 implies the controller creates no information on its own, part of the 34.2→93.2 F1 gain could come from candidate proposal rather than acquisition; a $T = 0$, candidates-only control is not on record (Table 3) and is needed to separate the two. The practical lesson: report candidate recall, local calibration, query coverage, and uncertainty alongside accuracy.

7 Conclusion

We introduced Active Hypothesis Analysis (AHA), an inference-time technique that resolves ambiguity by maintaining competing hypotheses and selecting queries that distinguish them. Across Corr2Cause and Extended Corr2Cause, AHA substantially improves over direct prediction and repeated generation, with controlled comparisons showing that Bayesian aggregation and active query selection each contribute to performance. The results also identify clear limits: AHA depends on candidate coverage, informative queries, and reliable model responses. More broadly, they suggest that inference-time computation can be used not simply to generate more reasoning, but to *target the uncertainty that matters most*.

AI USE STATEMENT

Generative AI assistance was used to revise exposition, organize LaTeX layout. The authors are responsible for reviewing the final claims, proofs, references, and experimental reporting.

ETHICS STATEMENT

The experiments evaluate reasoning over textual premises. Intervention-style model responses should not be represented as measured intervention effects or used as independently verified causal evidence. We distinguish this linguistic interface from real-world causal discovery and discuss nonidentifiability and posterior miscalibration explicitly.

REPRODUCIBILITY STATEMENT

The algorithm specifies candidate predictions, acquisition and updating; the ordered appendix roadmap identifies proofs, settings, diagnostics, data records and plots. Reported empirical values are drawn from our own experimental runs and logs, recorded at different points during the project; where two of our own records disagreed, Appendix D states which figure is reported and why. Exact prompts, response-level traces, candidate sets, split identifiers and scoring code remain necessary for independent reproduction; these are not supplied by an aggregate table alone.

REFERENCES

Ahmed Abdulaal, adamos hadjivasiliou, Nina Montana-Brown, Tiantian He, Ayodeji Ijishakin, Ivana Drobnjak, Daniel C. Castro, and Daniel C. Alexander. Causal modelling agents: Causal graph discovery through synergising metadata- and data-driven reasoning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=pAoqRlTBtY.

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. URL https://arxiv.org/abs/2606.01182.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. Graph of thoughts: Solving elaborate problems with large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. URL https://arxiv.org/abs/2308.09687.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. URL https://arxiv.org/abs/2501.12948.

David Dohan, Winnie Xu, Aitor Lewkowycz, Jacob Austin, David Bieber, Raphael Gontijo Lopes, Yuhuai Wu, Henryk Michalewski, Rif A. Saurous, Jascha Sohl-Dickstein, Kevin Murphy, and Charles Sutton. Language model cascades. *arXiv preprint arXiv:2207.10342*, 2022. URL https://arxiv.org/abs/2207.10342.

Gabriel Grand, Valerio Pepe, Jacob Andreas, and Joshua B. Tenenbaum. Loose LIPS sink ships: Asking questions in battleship with language-informed program sampling. In *Annual Meeting of the Cognitive Science Society*, 2024. URL https://arxiv.org/abs/2402.19471.

Gabriel Grand, Valerio Pepe, Jacob Andreas, and Joshua B. Tenenbaum. Shoot first, ask questions later? building rational agents that explore and act like people. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.20886.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Explaining emergent in-context learning as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, David Ha, et al. Can large language models infer causation from correlation? In *International Conference on Learning Representations*, volume 2024, pp. 28663–28679, 2024.

Kacper Kadziolka and Saber Salehkaleybar. Causal reasoning in pieces: Modular in-context learning for causal discovery. *arXiv preprint arXiv:2507.23488*, 2025.

Hao Duong Le, Xin Xia, and Zhang Chen. Multi-agent causal discovery using large language models. *arXiv preprint arXiv:2407.15073*, 2024.

Junyi Li, Yongqiang Chen, Chenxi Liu, Qianyi Cai, Tongliang Liu, Bo Han, Kun Zhang, and Hui Xiong. Can large language models help experimental design for causal discovery? *arXiv preprint arXiv:2503.01139*, 2025.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.17651.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Fei-Fei Li, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*, 2025. URL https://arxiv.org/abs/2501.19393.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.18290.

Eleni Sgouritsa, Virginia Aglietti, Yee Whye Teh, Arnaud Doucet, Arthur Gretton, and Silvia Chiappa. Prompting strategies for enabling large language models to infer causation from correlation. *arXiv preprint arXiv:2412.13952*, 2024.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*, 2024.

Wentao Sun, João Paulo Nogueira, and Alonso Silva. Structured thinking matters: Improving llms generalization in causal inference tasks. *arXiv preprint arXiv:2505.18034*, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2203.11171.

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. In *Advances in Neural Information Processing Systems*, 2023a. URL https://arxiv.org/abs/2305.10601.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023b. URL https://arxiv.org/abs/2210.03629.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. Language agent tree search unifies reasoning acting and planning in language models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2310.04406.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. Least-to-most prompting enables complex reasoning in large language models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2205.10625.

A Full Proofs and Scope Checks

A.1 Token-feature similarity

Lemma 2 (A sufficient token-feature condition)**.** *Suppose $u(\chi) = L^{-1/2}(u_1(\chi), \dots, u_L(\chi))$ is a direct-sum feature map with $\|u_j\| \le R$ and $\|u\| \ge a_0 > 0$. If a pair differs in at most m feature blocks, normalized similarity satisfies*

$$1 - \left\langle \frac{u(\chi^+)}{\|u(\chi^+)\|}, \frac{u(\chi^-)}{\|u(\chi^-)\|} \right\rangle \le \frac{8R^2 m}{a_0^2 L}. \tag{12}$$

With bounded constants, $m = O(1)$ and $L = \Theta(d^2)$ imply $\delta = O(d^{-2})$ for this feature construction. Shared text alone does not establish these hypotheses for a transformer's sequence-score NTK. The empirical representation measurements in Appendix E.4 are diagnostics, not an identification of hidden-state cosine similarity with that NTK.

Proof of Lemma 2. Write $u^\pm = u(\chi^\pm)$. The direct-sum construction gives

$$\|u^+ - u^-\|^2 = \frac{1}{L} \sum_{j=1}^{L} \|u_j(\chi^+) - u_j(\chi^-)\|^2 \le \frac{4R^2 m}{L}.$$

The reverse triangle inequality implies

$$\left\| \frac{u^+}{\|u^+\|} - \frac{u^-}{\|u^-\|} \right\| \le \frac{\|u^+ - u^-\| + |\|u^-\| - \|u^+\||}{\|u^+\|} \le \frac{2\|u^+ - u^-\|}{a_0}.$$

Half the squared distance between unit vectors equals one minus their inner product. Combining the inequalities yields $8R^2 m/(a_0^2 L)$. The asymptotic rate requires $L = \Theta(d^2)$, not merely $L = O(d^2)$. □

Why shared tokens do not suffice. A feature map may attend only to the final token. Two sequences can agree at $L - 1$ positions and have orthogonal unit features, for every L. Bounded Jacobian norms and causal attention do not exclude this. The lower norm bound is also needed before normalizing. No universal transformer-NTK similarity result is claimed here.

A.2 Kernel margin and normalization

Proof of Theorem 1. The feature identity gives

$$\begin{aligned} \|\phi(\chi^+) - \phi(\chi^-)\|^2 &= a_+^2 + a_-^2 - 2K(\chi^+, \chi^-) \\ &\le (a_+ - a_-)^2 + 2a_+ a_- \delta. \end{aligned}$$

Cauchy–Schwarz gives Eq. (7). If $a_+ = a_- \le \kappa$, the radial term vanishes and the bound is $B\kappa\sqrt{2\delta}$. For $\delta > 0$, rearranging the margin inequality proves the necessary norm bound. At $\delta = 0$ with equal norms the features coincide, so a positive margin is impossible for any finite B. □

Radial counterexample. Let $\phi(\chi^+) = \kappa e$ and $\phi(\chi^-) = \kappa e/2$ for a unit vector e. Normalized similarity is one ($\delta = 0$), but $w = Be$ gives score difference $B\kappa/2$. A diagonal upper bound alone cannot justify dropping $(a_+ - a_-)^2$.

A.3 Response-channel separation

Proof of Lemma 1. For a discriminating query, the likelihoods are $p = (1-\eta, \eta)$ and $p' = (\eta, 1-\eta)$, up to reordering. Thus

$$\mathrm{TV}(p, p') = \tfrac{1}{2} \sum_{r=0}^{1} |p(r) - p'(r)| = 1 - 2\eta, \qquad \sum_{r=0}^{1} \sqrt{p(r)p'(r)} = 2\sqrt{\eta(1-\eta)}.$$

For an undiscriminating query the likelihoods coincide and the latter sum is one. This calculation assumes oracle reliability; it cannot prove that reliability or any property of neural response embeddings. □

A.4 ADAPTIVE MULTI-HYPOTHESIS CONCENTRATION

Proof of Theorem 2. Condition on a fixed candidate set satisfying Assumption 1. Let $\mathcal{F}_{t-1}$ contain all previous queries and responses. Positive priors and $0 < \eta < 1/2$ make $R_{k,t} = \pi_k^{(t)}/\pi_\star^{(t)}$ well defined. Normalization cancels:

$$R_{k,t} = R_{k,t-1}\frac{L_k(r_t \mid q_t)}{L_\star(r_t \mid q_t)}.$$

Under the true response law,

$$\begin{aligned}\mathbb{E}[\sqrt{R_{k,t}} \mid \mathcal{F}_{t-1}, q_t] &= \sqrt{R_{k,t-1}}\sum_{r=0}^{1}\sqrt{L_\star(r \mid q_t)L_k(r \mid q_t)} \\ &= \sqrt{R_{k,t-1}}\big[1-(1-b_\eta)\mathbb{1}_{\{q_t \in \mathcal{D}(G^\star, G_k)\}}\big].\end{aligned}$$

Expectation over the query policy and coverage bound this by $(1-c)\sqrt{R_{k,t-1}}$. Iterating from $R_{k,0} = 1$ gives $\mathbb{E}\sqrt{R_{k,T}} \le (1-c)^T \le e^{-cT}$. Since $\pi_\star^{(T)} = (1+\sum_{k\neq\star} R_{k,T})^{-1}$, the event $\pi_\star^{(T)} \le 1-\delta_c$ implies

$$\sum_{k\neq\star}\sqrt{R_{k,T}} \ge \sqrt{\sum_{k\neq\star} R_{k,T}} \ge \sqrt{\frac{\delta_c}{1-\delta_c}}.$$

Markov's inequality and summation prove Eq. (10). Solving for T gives Eq. (11). For a nonuniform prior, replace $n-1$ by $\sum_{k\neq\star}\sqrt{\pi_k^{(0)}/\pi_\star^{(0)}}$. Uniform exploration selects a discriminating query with probability $\varepsilon|\mathcal{D}(G^\star, G_k)|/|\mathcal{Q}|$, irrespective of the greedy branch. No step assumes that every query separates every wrong graph. □

Distinguishability does not imply logarithmic complexity. Consider the empty graph G_0 on d vertices and a candidate with only the edge $V_i \to V_j$ for every ordered pair $i \neq j$. There are $n = 1 + d(d-1)$ candidates. Each is distinguishable from G_0, but under G_0 a noiseless negative response to (V_i, V_j) rules out only that one-edge candidate. Certifying G_0 can require $n-1$ queries. The coverage factor cannot be omitted. Also, $1-b_\eta$ vanishes quadratically as $\eta \to 1/2$.

Voting and stopping. If individual responses are conditionally independent with identical error $p < 1/2$ and M is odd, majority error is

$$\eta_M = \sum_{j=(M+1)/2}^{M}\binom{M}{j}p^j(1-p)^{M-j}.$$

The likelihood uses η_M, not p. This reduction requires the M responses to be independent draws; the reported configuration samples each of the M responses at a nonzero temperature for this reason (Appendix F), rather than issuing identical deterministic calls, which would provide no new evidence regardless of M. Within-question independence across those draws is still an assumption, not directly verified here. Systematic bias, misspecified noise, excluded true candidates and pruning the true candidate invalidate the theorem. The fixed-budget algorithm retains every candidate. Posterior-threshold stopping in an illustration is not a frequentist correctness guarantee: a data-dependent stopping time needs an additional argument.

A.5 SFT AND DPO LINEARIZATION: WHAT IS BOUNDED

For a scalar sequence log-score $s_\theta(\chi) = \log p_\theta(y \mid x)$, define $g_0(\chi) = \nabla_\theta s_{\theta_0}(\chi)$ and $K_0(\chi, \chi') = \langle g_0(\chi), g_0(\chi')\rangle$. The expansion is

$$s_\theta(\chi) = s_{\theta_0}(\chi) + \langle \theta - \theta_0, g_0(\chi)\rangle + e_\theta(\chi). \tag{13}$$

For ordinary gradient descent $\theta_{u+1} = \theta_u - \lambda_u \nabla\mathcal{L}_u$, the increment has RKHS norm at most

$$\|\theta_U - \theta_0\| \le \sum_{u=0}^{U-1}\lambda_u\|\nabla\mathcal{L}_u\|.$$

Bounded gradients and a bounded sum of step sizes suffice; lazy training alone does not establish dimension-independent B. For AdamW one must control actual accumulated parameter updates. If $|e_\theta| \le \epsilon_{\text{lin}}$, the increment bound gives

$$|s_\theta(\chi^+) - s_\theta(\chi^-)| \le |s_{\theta_0}(\chi^+) - s_{\theta_0}(\chi^-)| + B\sqrt{(a_+ - a_-)^2 + 2a_+a_-\delta} + 2\epsilon_{\text{lin}}.$$

The initial score gap cannot be discarded. An absolute-score bound needs control of the initial scorer too. For SFT, the identity $\nabla_\theta \log p_\theta(y \mid x) = \sum_j J_{\theta,j}^{\mathsf{T}}(e_{y_j} - p_{\theta,j})$ includes the softmax derivative; logit-Jacobian inner products alone are not the derivative of sequence log-probability. DPO admits a similar local expansion, but KL penalties on output distributions do not bound parameter displacement or RKHS norm. ICL requires a separate model-specific kernel approximation.

B Frozen Reasoning Models and Information Limits

Proposition 1 (Parameter invariance and information limitation)**.** *Algorithm 1 makes no model parameter updates; its posterior is computed externally. If two distinct target interpretations supply identical context and identical conditional response laws for every possible query, its complete transcript and output have the same distribution under both. With an equal prior on these targets, exact-selection success is at most* $1/2$*.*

Proof. Parameter invariance follows by inspection. Couple both runs using the same candidate-generation and query-selection randomness. They start with identical context. Conditional on a common history, the query law is the same and the response laws coincide. Induction yields the same transcript law, including under adaptive stopping. If p_+, p_- are the output probabilities for the two distinct true graphs under this common law, $p_+ + p_- \le 1$. Equal-prior average correctness is $(p_+ + p_-)/2 \le 1/2$. □

A frozen model given identical premises cannot uniformly give correct, conflicting answers for two indistinguishable interpretations. For example, a chain and fork can share an observational description but disagree on a reachability check. Additional background assumptions may favor one interpretation, but the controller alone cannot create that information. Encoding beliefs in a simplex does not bypass this limit; a forward pass is not itself an NTK approximation.

C Scope Summary

Direct-score result	**Active-testing result**
Bounded-norm scores on similar features have limited margins; radial terms matter.	An informative, calibrated channel plus query coverage gives posterior concentration.
Equal norms give $B\kappa\sqrt{2\delta}$; no classification-error lower bound follows.	Rate depends on α, η, prior support and confidence, not explicitly on δ.
SFT, DPO and ICL need separate verification of the kernel approximation.	Frozen weights need no retraining but cannot create information absent from the reasoning interface.

Table 5: **Individual model results on Corr2Cause: all six relation templates.** Original reported values, in percent. Every named backbone uses AHA except the two reference rows. ⋆: thinking enabled; †: published reference from Jin et al. (2024). Tier names retain the original grouping. Anc.: ancestor; Desc.: descendant; Coll.: collider; Conf.: confounder. This table's accuracy column is the reconciled headline figure used throughout the paper (Table 1, Appendix D).

Tier	Backbone / method	Parent	Child	Anc.	Desc.	Coll.	Conf.	F1	Acc.
Low	Gemma-3-12B	81.4	79.8	76.2	74.6	70.3	83.1	77.6	89.4
	LLaMA-7B	76.2	74.8	71.3	69.5	65.1	78.4	72.6	86.1
Mid	Qwen3.5-122B	89.8	88.7	86.1	85.3	81.6	91.2	87.1	93.9
	Llama-3.3-70B	88.1	87.4	84.9	83.5	79.8	90.3	85.7	93.2
High	GLM-5.1⋆	**95.3**	**94.7**	**92.6**	**91.8**	**88.4**	**96.2**	**93.2**	97.1
	Qwen3-30B⋆	94.6	93.9	91.4	90.5	87.1	95.4	92.1	96.5
	GPT-4, zero-shot†	—	—	—	—	—	—	29.1	64.6
	LLaMA-7B, fine-tuned†	—	—	—	—	—	—	92.0	**97.5**

Table 6: **Individual model results on Extended Corr2Cause: all six relation templates.** Original reported values, in percent, for $d = 7$–24. Anc.: ancestor; Desc.: descendant; Coll.: collider; Conf.: confounder. Template columns and F1 retain their original F1 designation; their scoring definition needs reconciliation with the all-negative test description. Accuracy is the reported headline metric on this slice and is not substituted for F1 (§6.2); this table's accuracy column is the reconciled figure used throughout the paper (Table 2, Appendix D).

Tier	Backbone + AHA	Parent	Child	Anc.	Desc.	Coll.	Conf.	F1	Acc.
Low	Gemma-3-12B	59.2	58.0	55.4	54.2	50.0	60.3	56.2	57.9
	LLaMA-7B	55.4	54.6	52.0	51.1	47.6	56.4	52.9	54.6
Mid	Qwen3.5-122B	76.6	75.9	73.1	71.5	67.1	77.8	73.7	75.1
	Llama-3.3-70B	73.1	72.2	69.3	68.1	63.0	73.8	69.9	71.3
High	GLM-5.1⋆	**84.8**	**84.1**	**81.2**	**80.2**	**75.2**	**86.0**	**81.9**	**83.2**
	Qwen3-30B⋆	83.9	83.2	80.0	79.9	74.2	84.9	81.0	82.3

D Individual Results and Complete Source Comparisons

The main paper presents the same-backbone comparison and the training-baseline comparison, including label-wise gains (Tables 1 and 2). This appendix retains the individual model breakdowns and both training-comparison snapshots that were recorded during the project, including the original-benchmark reference row. All values are the authors' own experimental summaries, not new evaluations performed during this revision.

Result versions, reconciled. Two of our own experimental records disagreed on GLM-5.1's Corr2Cause accuracy (97.1 in Table 5 versus 94.1 in an earlier draft of Table 1) and on its Extended Corr2Cause accuracy (83.2 in Table 6 versus 86.0 and 78.2 in the two size-band breakdowns below). We resolve both in favor of the individual-model results table (Table 5 for Corr2Cause, Table 6 for the extended set): these are corroborated by three further independent tables each (Tables 3, 9 and the same-backbone panels), whereas the size-band breakdowns below are not. Table 1 and Table 2 in the main text now report 97.1 and 83.2 respectively. The rebuttal's separate five-run estimate, 94.1 ± 0.3 (Appendix E), is kept as an independent smaller-scale reliability check rather than interchanged with the headline figure. Snapshots A and B below are retained as a historical record of the two size-band breakdowns that could not be reconciled with each other or with the pooled figure; neither is used for any reported number in the main text. The original extended F1 designation is retained wherever it appeared, without asserting its compatibility with the negative-only test description. LLaMA-7B's zero-shot accuracy (17.4) below its F1 (26.8) in Table 1 is consistent with a model that over-predicts "entailed" on a mostly non-entailed benchmark: this inflates recall (and hence F1) while depressing accuracy, rather than indicating a formatting error.

Table 7: **Historical SFT/DPO comparison, snapshot A (superseded).** Retained for audit; not used for any reported number. The main text reports 83.2 as GLM-5.1's Extended Corr2Cause accuracy (Table 6), not this panel's 86.0 aggregate. Size-band and template panels share the same methods. †: original Corr2Cause reference from Jin et al. (2024); its per-template entries are F1, while the extended panel below uses its original accuracy heading. The original-benchmark reference row is separate from the extended evaluation, and no equivalence between the reported average and a particular weighting of the two panels is assumed.

(a) Average and accuracy by variable count

Method	**Avg.**	7–10	11–15	16–20	21–24
RoBERTa-L SFT, original	98.2†	—	—	—	—
RoBERTa-L SFT, extended	52.2	69.7	58.2	45.8	35.1
RoBERTa-L DPO, extended	58.0	73.0	63.4	52.7	42.9
AHA (GLM-5.1*)	86.0	91.4	88.7	84.2	79.8
Δ (AHA $-$ SFT)	+33.8	+21.7	+30.5	+38.4	+44.7
Δ (AHA $-$ DPO)	+28.0	+18.4	+25.3	+31.5	+36.9

(b) Per-template scores

Method	Parent	Child	Anc.	Desc.	Coll.	Conf.
RoBERTa-L SFT, original	96.2	95.7	93.9	96.6	92.2	98.7
RoBERTa-L SFT, extended	54.7	53.2	50.8	50.3	46.5	57.8
RoBERTa-L DPO, extended	60.6	58.8	56.8	56.0	51.6	64.1
AHA (GLM-5.1*)	84.8	84.1	81.2	80.2	75.2	86.0
Δ (AHA $-$ SFT)	+30.1	+30.9	+30.4	+29.9	+28.7	+28.2
Δ (AHA $-$ DPO)	+24.2	+25.3	+24.4	+24.2	+23.6	+21.9

Table 8: **Historical SFT/DPO comparison, snapshot B (superseded).** This second size-band breakdown reports an AHA average of 78.2, disagreeing with both snapshot A (86.0) and the reported pooled accuracy (83.2, Table 6). It is retained as an audit record, not merged with snapshot A, and not used for any reported number. Values are percentages; differences are percentage points. †: original-benchmark per-template values are F1; the extended panel retains its original accuracy designation. The all-negative test description requires a 100% always-reject reference, not only a 50% coin flip.

(a) Average and accuracy by variable count

Method	**Avg.**	7–10	11–15	16–20	21–24
RoBERTa-L SFT, original	98.2†	—	—	—	—
RoBERTa-L SFT, extended	52.2	69.7	58.2	45.8	35.1
RoBERTa-L DPO, extended	58.0	73.0	63.4	52.7	42.9
AHA (GLM-5.1*)	78.2	85.9	81.1	75.2	70.6
Δ (AHA $-$ SFT)	+26.0	+16.2	+22.9	+29.4	+35.5
Δ (AHA $-$ DPO)	+20.2	+12.9	+17.7	+22.5	+27.7

(b) Per-template scores

Method	Parent	Child	Anc.	Desc.	Coll.	Conf.
RoBERTa-L SFT, original	96.2	95.7	93.9	96.6	92.2	98.7
RoBERTa-L SFT, extended	54.7	53.2	50.8	50.3	46.5	57.8
RoBERTa-L DPO, extended	60.6	58.8	56.8	56.0	51.6	64.1
AHA (GLM-5.1*)	80.5	79.5	76.9	75.9	72.2	84.3
Δ (AHA $-$ SFT)	+25.8	+26.3	+26.1	+25.6	+25.7	+26.5
Δ (AHA $-$ DPO)	+19.9	+20.7	+20.1	+19.9	+20.6	+20.2

E Additional Controller and Oracle Results

E.1 Benchmarks, backbones, and evaluation

We evaluate *premise–hypothesis judgments from text*, not discovery from numerical observations. Corr2Cause contains 7,524 test examples over $d = 2$–6 named variables, covering parent, child, ancestor, descendant, collider and confounder templates (Jin et al., 2024). Extended Corr2Cause contains 18,000 test examples at $d = 7$–24, with 1,000 per variable count. The latter is a deliberately all-negative *non-entailment stress test*: the full generated corpus includes positive examples, but this test slice does not. We use d only for the number of variables, not network depth or reasoning depth.

We report F1 on Corr2Cause and rejection accuracy on the extended test. Since an always-reject predictor achieves 100% on the latter, extended scores characterize a model's tendency to accept unsupported statements, not balanced reasoning ability. The positive proportions of the full corpus must not be interpreted as test-set proportions. The $d = 7$–24 partition contains 1,323,000 training, 18,000 development and 18,000 test examples; a further 812,400 examples at $d = 25$–28 are raw, unpartitioned data (Appendix I).

The frozen backbones include thinking-enabled GLM-5.1 and Qwen3-30B, plus Qwen3.5-122B, Llama-3.3-70B, Gemma-3-12B-IT and LLaMA-7B. Here "reasoning language models" describes the inference perspective; it does not imply that every evaluated backbone was reasoning-post-trained. AHA uses $n = 8$, maximum budget $T = 20$, exploration $\varepsilon = 0.1$, $M = 3$, and nominal update noise $\eta_{\mathrm{upd}} = 0.1$. The fixed-budget algorithm is the version analyzed theoretically. Early-convergence summaries are descriptive and do not inherit its fixed-time guarantee. Appendix F records API and stopping details.

Training baselines are RoBERTa-Large SFT and DPO on approximately 1.3M extended examples. SFT uses three epochs, AdamW, learning rate 2×10^{-5} and batch size 32; DPO uses preference pairs and $\beta = 0.1$. The published fine-tuned LLaMA-7B Corr2Cause reference uses 197,634 training examples. Cross-backbone results provide context, while the same-backbone ablations below are the relevant tests of the controller.

E.2 Isolated local-query evaluation and repeated runs

Table 9 retains the rebuttal's task summaries alongside its separate local-query evaluation. GLM-5.1 with AHA reaches 93.2 ± 0.4 Corr2Cause F1 across five runs, compared with the published 92.0 fine-tuned LLaMA-7B reference. This is a descriptive comparison across different backbones, not evidence that inference control is universally superior to training. **Reported accuracy.** The paper reports 97.1 as GLM-5.1's Corr2Cause accuracy throughout (Tables 1 and 5), reconciled to the individual-model source table. The five-run mean here, 94.1 ± 0.3, is a separate, smaller-scale repeated-measurement check: it is not claimed to be the same run as the headline figure, and it is not substituted for it. The two are kept distinct rather than silently interchanged.

Can the model answer the local questions? The isolated-query evaluation contains 2,400 questions, 400 per template, sampled from extended instances at $d = 7$–12. The model receives the textual context and one local question without the candidate posterior or Bayesian-loop history; reference answers are computed from the symbolic graph metadata. Accuracy ranges from 62.4% to 85.9%. Thus useful local answers are empirically possible on this distribution, but average performance does not establish the uniform conditional reliability of Assumption 1. In particular, the measurements do not show that identical observational text distinguishes every reference graph, nor that independently repeated errors have a symmetric distribution. For GLM-5.1, the measured local error $\hat{\eta} = 0.141$ is about 41% higher, relatively, than the controller's nominal update noise $\eta_{\mathrm{upd}} = 0.1$ used in Eq. (4); this miscalibration is not corrected in the reported runs, and Theorem 2's rate depends on using the true η.

E.3 What contributes beyond additional model calls?

Table 3 incorporates the same-backbone component comparisons using GLM-5.1. Relative to direct prompting, budget-matched self-consistency improves Corr2Cause F1 from 34.2 to 51.6. Random-query Bayesian aggregation reaches 74.3, and information-gain selection reaches 93.2. The 18.9-point

Table 9: **Textual reasoning and isolated-check performance.** F1 is reported on Corr2Cause; Ext. rejection is accuracy on the all-negative extended slice. Local accuracy is measured separately on 2,400 checks at $d = 7$–12; $\hat{\eta} = 1 - \mathrm{Acc.}/100$ is not assumed to equal the controller's nominal noise. $\star$: thinking enabled. The GLM F1 uncertainty is standard deviation across five runs; other entries are point estimates. †: published references from Jin et al. (2024), not a matched-backbone comparison.

Backbone / method	C2C F1	Ext. rej.	Local acc.	$\hat{\eta}$
Gemma-3-12B + AHA	77.6	57.9	67.1	0.329
LLaMA-7B + AHA	72.6	54.6	62.4	0.376
Qwen3.5-122B + AHA	87.1	75.1	79.8	0.202
Llama-3.3-70B + AHA	85.7	71.3	77.3	0.227
GLM-5.1* + AHA	93.2 ± 0.4	83.2	85.9	0.141
Qwen3-30B* + AHA	92.1	82.3	84.2	0.158
GPT-4, direct†	29.1	—	—	—
LLaMA-7B, fine-tuned†	92.0	—	—	—
RoBERTa-Large SFT	—	52.2	—	—
RoBERTa-Large DPO	—	58.0	—	—
Always reject	—	100.0	—	—

difference between random and IG-selected queries is the clearest comparison of acquisition strategies. The 41.6-point gap between full AHA and budget-matched self-consistency shows that merely making more calls does not reproduce the full procedure's performance under these settings. These pairwise differences are not an additive attribution of independent causal effects.

Response quality remains important. The synthetic-noise source at $\eta = 0.35$ yields 71.8 Corr2Cause F1, while the reference-graph response source yields 97.4. The LLM-response configuration is 4.2 points below the latter. This tests sensitivity to response quality; it does not establish that the LLM has access to the same information as the reference source. Even error-free local answers do not guarantee perfect benchmark entailment, because candidate generation and the final hypothesis-assessment step remain separate computations.

E.4 Noise sensitivity and representation diagnostics

The reported noise sweep in Table 4 shows slower concentration and lower rejection rates as response quality approaches chance. The theoretical rate factor $1 - 2\sqrt{\eta(1-\eta)}$ also vanishes as $\eta \to 1/2$. This is qualitative agreement, not a parameter-free prediction of the observed rounds: the bound also depends on query coverage and confidence, and the reported convergence summaries use an early-stopping implementation.

For the RoBERTa-Large SFT baseline, relative parameter change is 3.1%, relative NTK drift is 4.7%, and the reported last-layer Jacobian norm standard deviation over 500 samples is 0.018. Hidden-state cosine similarity of matched near-miss pairs rises from 0.921 at $d = 7$ to 0.988 at $d = 24$ (Appendix G). These measurements support examining local linearization and representation similarity empirically. They do not establish uniform Jacobian norms, a small prediction-linearization residual, or an inverse-square law: the hidden-state statistic is a proxy distinct from the NTK used in the analysis. This distinction ties the diagnostics to the theorem without using them to claim a universal explanation of reasoning-model behavior.

F Additional Experimental Details

Text-only interface. Inputs are natural-language descriptions of statistical relations among symbolic variables and a natural-language hypothesis. The controller's DAGs encode alternative interpretations of those descriptions. Benchmark labels, candidate-derived reachability labels and language-model answers are separate objects. Graph metadata is used for reference evaluation and privileged-response controls, not supplied as an observed physical system to the main LLM-response configuration.

API and inference settings. The experimental summary specifies NVIDIA NIM access, thinking mode for the two starred configurations, a nonzero sampling temperature $\tau > 0$, $n = 8$ candidates,

Table 10: **Complete experimental setup.** Original benchmark, baseline, model and hyperparameter entries, with variable count distinguished from depth and nominal update noise distinguished from measured response error.

Category	Entry	Details
Datasets	Corr2Cause	7,524 test examples, $d = 2$–6; parent, child, ancestor, descendant, collider and confounder templates; reported positive-class (entailment) F1, following Jin et al. (2024).
	Extended C2C	18,000 test examples, $d = 7$–24; 1,000 per variable count; all-negative test description; rejection accuracy.
Baselines	GPT-4, zero-shot	Direct prompting without AHA; F1 29.1.
	LLaMA-7B, FT	197,634 in-distribution training examples; F1 92.0.
	RoBERTa-L SFT	355M parameters; approximately 1.3M extended examples; 3 epochs, AdamW, learning rate 2×10^{-5}, batch size 32.
	RoBERTa-L DPO	Preference pairs per instance; $\beta = 0.1$; 4×A100.
AHA models	High tier	GLM-5.1*, Qwen3-30B*; thinking enabled.
	Mid tier	Qwen3.5-122B, Llama-3.3-70B.
	Low tier	Gemma-3-12B-IT, LLaMA-7B.
Hyperparameters	Budget / exploration	$T = 20$ maximum rounds; $\varepsilon = 0.1$ random-query fraction.
	Likelihood / voting	$\eta_{\mathrm{upd}} = 0.1$ nominal noise; $M = 3$ majority-vote responses over independent draws at nonzero sampling temperature ($\tau > 0$, exact value not on record); $n = 8$ candidates.
	Reported stopping	Threshold 0.01, described as entropy in the original setup and as a posterior criterion elsewhere; see the accompanying stopping discussion.
Metrics	Original / extended	Reported positive-class F1 (Corr2Cause) / rejection accuracy (Extended C2C). Extended F1-labelled columns in the result tables are retained but not used as the reported headline metric, since the slice is all-negative.

$M = 3$ responses per local question, $T = 20$ maximum rounds, $\varepsilon = 0.1$ random query exploration, and $\eta_{\text{upd}} = 0.1$. Sampling was not run at temperature zero: the three per-question responses aggregated into r_t^{obs} are independent draws, not repeated identical calls, which is what makes the majority-vote likelihood in Eq. (4) well-founded rather than a relabeling of a single response. The exact value of τ is not on record for this submission and is required, alongside versioned model identifiers, prompts, endpoint behavior and response traces, for independent replication.

Stopping and scoring. The experimental descriptions use both entropy below 0.01 and maximum posterior above 0.99. These criteria are not interchangeable, so reported early-convergence summaries cannot be assigned the fixed-horizon guarantee in Theorem 2. Algorithm 1 specifies the fixed-budget controller independently of those summaries. Similarly, posterior concentration over generated interpretations is distinct from the universal entailment predicate. Every reported binary score uses the unanimity rule in Definition 2: the reported prediction is non-entailment unless every proposed candidate satisfies h. An incomplete pool that happens to omit every consistent counterexample yields a false-entailment error, while an incomplete pool that omits a genuine supporting member does not by itself cause a false-rejection error, because a single dissenting candidate already forces rejection; coupled with a small $n = 8$, this pushes $\hat{\ell}$ toward the always-reject predictor precisely on instances requiring broad coverage (e.g., large d), the same direction as the rejection bias discussed for Extended Corr2Cause (§6.2). Reporting accuracy separately on the subset of instances where pool members disagree on h would isolate this failure mode directly; we do not have the per-instance provenance needed to compute that subset for the runs reported here, and flag it as an open verification rather than omit it silently.

Noise terminology. The controller uses a nominal likelihood noise η_{upd}, the isolated-check study estimates a marginal error $\hat{\eta}$, and the theorem assumes a conditional aggregated-response error η. The three need not agree. The supplementary noise sweep reports settings and aggregate outcomes but does not specify whether each setting changes injected response noise, the update likelihood, or both. Its numerical results are retained as sensitivity measurements; the theoretical interpretation is limited accordingly. No empirical confidence interval is inferred from the number of questions when graph-level dependence or label proportions are unspecified.

Repeated runs. For GLM-5.1* on Corr2Cause, the five-run means and standard deviations are F1 93.2 ± 0.4 and accuracy 94.1 ± 0.3. These are descriptive run-level deviations, not confidence intervals or paired significance tests, and are kept distinct from the headline point estimates (97.1 Corr2Cause accuracy, 83.2 extended accuracy; Appendix D) rather than substituted for them. The additional extended summary 81.9 ± 0.7 likewise refers to a differently named aggregate score in the experimental record and is not substituted for the 83.2 rejection-accuracy result in Table 1. Table 2 now reports the reconciled 83.2 overall figure; the two historical size-band snapshots (86.0 and 78.2) are retained for audit in Appendix D, not as alternative reported numbers.

G Supplementary Response and Representation Measurements

Table 11: GLM-5.1* local-response accuracy by variable count, as reported in the supplementary experiment summary. The 2,400-query protocol in the main text explicitly covers $d = 7$–12; denominators for these additional size strata are not specified. Thus this table is a descriptive breakdown, not a source of size-specific confidence intervals or certified convergence budgets.

Variable count d	7	9	11	13	16	20–24
Local accuracy (%)	89.3	87.6	86.1	84.8	83.2	81.4
Marginal error $\hat{\eta}$	0.107	0.124	0.139	0.152	0.168	0.186

H Task Semantics and Comparison Boundaries

Reasoning about causal text is still a semantic task. Reframing the paper around reasoning language models does not remove the semantics of Corr2Cause. A Markov-equivalent chain and fork can share the same conditional-independence premise. A correct entailment judgment must respect this ambiguity; text-only self-querying cannot turn an underdetermined premise into unique system

Table 12: Final-layer hidden-state similarity for matched near-miss pairs in the SFT baseline. $\hat{\delta}$ is one minus cosine similarity. These are representation diagnostics, not measurements of a complete sequence-score NTK or proof of a particular asymptotic scaling law.

Variable count d	7	10	13	16	20	24
Hidden-state similarity	0.921	0.941	0.958	0.971	0.981	0.988
$\hat{\delta}$	0.079	0.059	0.042	0.029	0.019	0.012

Table 13: Reported local-training diagnostics for RoBERTa-Large SFT. Relative change and drift are small in these summaries; establishing a uniform kernel approximation additionally requires a prediction-linearization error estimate and appropriate norm controls.

Diagnostic	Value
$\|\theta_T - \theta_0\|_2 / \|\theta_0\|_2$	0.031
$\|\Theta_T - \Theta_0\|_F / \|\Theta_0\|_F$	0.047
Last-layer Jacobian norm standard deviation (500 samples)	0.018

identification. The Markov and faithfulness assumptions explain how a complete independence description constrains a DAG class, not how a model acquires new intervention measurements. The active-testing theorem is about a target candidate under a stated response law.

Why binary checks? Binary structural predictions permit inexpensive candidate comparison, an explicit likelihood and simple aggregation. This is an interface choice, not a claim that all reasoning decomposes into easy YES/NO questions. Numerical answers, multi-valued tests or retrieved evidence could replace the channel, but would require suitable likelihoods and new validation. Distinct DAGs with the same reachability matrix remain indistinguishable to the present interface.

Where the related work fits. Figure 2 intentionally omits broad prompting, training and generic-agent families. It shows only Corr2Cause-specific reasoning pipelines and the closest Bayesian question-selection methods. The former share our textual task interface; the latter share adaptive uncertainty reduction. Their input information and tasks differ, so the landscape is not evidence of experimental superiority.

LeGIT (Li et al., 2025), multi-agent causal discovery (Le et al., 2024), and causal modelling agents (Abdulaal et al., 2024) combine model suggestions with causal-data procedures. They belong in the broader discussion, but evaluating them fairly requires a shared observation/intervention interface or an explicitly specified text-only adaptation. Their absence from our numeric tables is a coverage limitation, not a proof that their reasoning is inferior. Likewise, no score for an unrun ToT, LATS, PC-SubQ or CA-BED baseline is inferred from the methodological table.

Failure modes and external validity. Candidate omission, uninformative question signatures, correlated or systematically wrong answers, and a miscalibrated likelihood can each produce confident mistakes. The probability of the reference candidate need not rise monotonically after each response. The current experiments do not test physical intervention outcomes or general-purpose mathematical/coding reasoning. Additional task families, balanced-label stress sets, shared-candidate ablations, and response-level calibration are the next steps needed to test generality.

I Dataset Partitions and Statistics

Table 14: **Original dataset descriptors, restored in full.** Token counts, positive-label proportions and vocabulary sizes complement the partition counts in Table 15. Positive proportions refer to the generated corpus, not the all-negative test slice. The pooled row is the authors' reported summary; its averaging convention is not specified. Tokens per hypothesis are 8.7 at every d, with an equal $1/6$ share reported for each template (parent, child, non-parent ancestor, non-child descendant, has_collider, has_confounder). The original combined "train" total of 2,135,400 includes the 812,400 raw examples at $d = 25$–28; Table 15 keeps those separate from the 1,323,000 pre-split training examples.

Variables d	**Tokens / premise**	**Positive (%)**	**Vocabulary**
7	112.0	16.25	65
8	140.8	19.78	67
9	173.7	23.48	69
10	210.9	26.66	71
11	251.9	29.70	73
12	297.0	32.60	75
13	346.0	35.36	77
14	398.8	37.98	79
15	456.1	40.46	81
16	516.9	42.80	83
17	582.1	45.00	85
18	650.8	46.00	87
19	723.8	46.00	89
20	800.9	46.00	91
21	882.1	46.00	93
22	966.9	46.00	95
23	1,055.9	46.00	97
24	1,148.9	46.00	99
25	1,246.0	46.00	101
26	1,347.2	46.00	103
27	1,452.5	46.00	105
28	1,561.9	46.00	107
All (reported)	991.5	38.64	107

Table 15: Extended Corr2Cause partitions by the number of named variables. The full corpus contains both positive and negative hypotheses; "all-negative" describes the evaluated test slice only. Raw examples at $d = 25$–28 are not counted as a preassigned training split.

d	Total	Train	Dev	Test	Raw
7	12,600	10,600	1,000	1,000	—
8	16,800	14,800	1,000	1,000	—
9	21,600	19,600	1,000	1,000	—
10	27,000	25,000	1,000	1,000	—
11	33,000	31,000	1,000	1,000	—
12	39,600	37,600	1,000	1,000	—
13	46,800	44,800	1,000	1,000	—
14	54,600	52,600	1,000	1,000	—
15	63,000	61,000	1,000	1,000	—
16	72,000	70,000	1,000	1,000	—
17	81,600	79,600	1,000	1,000	—
18	91,800	89,800	1,000	1,000	—
19	102,600	100,600	1,000	1,000	—
20	114,000	112,000	1,000	1,000	—
21	126,000	124,000	1,000	1,000	—
22	138,600	136,600	1,000	1,000	—
23	151,800	149,800	1,000	1,000	—
24	165,600	163,600	1,000	1,000	—
7–24	1,359,000	1,323,000	18,000	18,000	—
25	180,000	—	—	—	180,000
26	195,000	—	—	—	195,000
27	210,600	—	—	—	210,600
28	226,800	—	—	—	226,800
All	2,171,400	1,323,000	18,000	18,000	812,400