Beyond Memorization: Distinguishing Between Pattern-Based and Epistemic Reasoning in LLMs Using Epistemic Puzzles

Anonymous ACL submission

Abstract

Epistemic reasoning requires agents to infer the state of the world from partial observations and information about other agents' knowledge. Prior work evaluating LLMs on epistemic puzzles often frames failures as memorization rather than reasoning. We argue that this dichotomy is too coarse for newer models: memorization is a limiting case of *pattern-based reasoning*, where a model matches a task to a familiar template and applies the corresponding solution. We introduce a two-dimensional benchmark over DEL-style puzzles, separating narrative familiarity from inference complexity, allowing us to distinguish pattern-based from epistemic reasoning. We find that models are substantially more robust to surface form changes than prior work suggested, yet consistently struggle in asymmetric settings where familiar patterns no longer apply and success requires tracking fragmented epistemic states.

1 Introduction

Problems requiring epistemic reasoning in multi-agent systems involve inferring the state of the world by integrating one's own observations with information about other agents' knowledge (Fagin et al., 1995). Such challenges arise naturally in collaborative settings where each agent has only a partial view of the world, such as legal proceedings, education, or medical applications (Rabinowitz et al., 2018; Li et al., 2023).

Previous work evaluating LLMs' epistemic reasoning has centered on canonical problems such as the *Muddy Children* problem, reporting limited performance (Sileo and Lernould, 2023). Related benchmarks show similar performance degradation under controlled task variations: when names and numbers are altered in arithmetic word problems (Mirzadeh et al., 2025), when puzzle complexity increases (Lin et al., 2025; Shojaee et al., 2026), and when non-standard variants change the underlying rules or constraints (Liang et al., 2025). These findings are often framed as a contrast between *reasoning*, where a model derives a solution from the specific parameters of the instance, and *memorization*, where it matches the task to a familiar pretraining pattern and reproduces the associated answer. Under this view, models over-rely on memorization, which breaks down under superficial perturbations (Jiang et al., 2024; Mirzadeh et al., 2025).

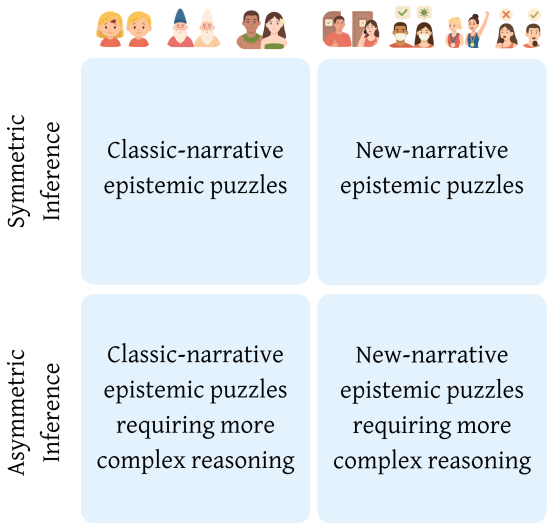
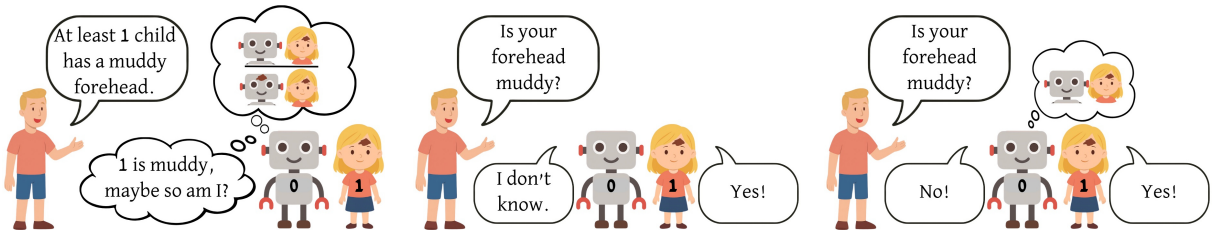


Figure 1: **Distinguishing between pattern-based and epistemic reasoning.** The horizontal Narrative axis contrasts three classic puzzles (*Muddy Children*, *Wise Men*, and *Blue-Eyed Islanders*) with four novel narratives (*Safety Inspection*, *Health Screening*, *Olympic Games*, and *Singing Contest*). The vertical Inference axis increases reasoning difficulty by shifting from standard symmetric settings, where each agent observes all others but themselves, to asymmetric, partial views.

In this work, we argue that recent advances in LLMs warrant a more nuanced treatment of memorization. Rather than viewing memorization as a binary alternative to reasoning, we treat it as one



(A) Public Announcement: Following the announcement, child 1 eliminates the possibility that both children are clean.

(B) Round 1: Child 0 cannot rule out possibilities based on the public announcement, while child 1 knows she is muddy.

(C) Round 2: Child 0 knows he is clean, since if he were muddy, child 1 would have remained uncertain.

Figure 2: An example of the *Muddy Children* puzzle with two children, one of whom is muddy, where each child observes the other child's forehead but not their own. Thought bubbles depict the possibilities child 0 considers.

form of a broader strategy that we define as *pattern-based reasoning*. Under this strategy, a model recognizes a problem as an instance or variant of a familiar template and applies the corresponding solution pattern. The central challenge is therefore not to reason directly over the specific parameters of the instance, but to identify the familiar structure to which it can be mapped. Under this view, memorization is the limiting case in which the model has encountered the same instance previously.

We therefore ask: *Can pattern-based and epistemic reasoning be distinguished in LLMs, and can performance under each be quantified?* We study this question through a controlled family of DEL-style puzzles that instantiate core epistemic mechanisms, allowing us to isolate narrative familiarity from inference complexity.

Instead of seeking specific perturbations that "break" a model, we propose a two-dimensional matrix taxonomy (Figure 1) that systematically disentangles pattern-based reasoning from epistemic reasoning. The horizontal dimension varies *narrative familiarity*, spanning seven distinct narratives: three classic puzzles - *Muddy Children* (Van Ditmarsh et al., 2007), *Wise Men* (d'Avila Garcez et al., 2009) and *Blue-Eyed Islanders* (Stuhlmüller and Goodman, 2014), compared with four novel stories - *Safety Inspection*, *Health Screening*, *Olympic Games*, and *Singing Contest*. These are designed to keep the exact reasoning needed to solve epistemic puzzles while changing the setting, entities, and surface narrative.

The vertical dimension scales *inference complexity*, shifting from standard symmetric observations where each agent observes all others but themselves, to asymmetric observations where each agent observes a randomly sampled set of the other agents. Measuring performance across this matrix allows us to characterize how models behave as pattern-based reasoning becomes less reliable and success depends heavily on epistemic reasoning.

We evaluate various LLMs and find that while most models perform well in classic, symmetric settings, even when surface patterns are substantially modified, they struggle in asymmetric cases that require multi-step epistemic reasoning. This drop reveals distinct behavioral profiles based on how models handle narrative familiarity and structural complexity. While some models show a relative advantage on classic narratives and others on novel ones, the remaining models show no meaningful difference between the two. Beyond the benchmark itself, this distinction provides an interpretable lens on model behavior, helping reveal not only how well models perform, but also *how* their reasoning changes under different settings.

Our contributions are threefold: (1) we reinterpret "memorization" as a form of *pattern-based reasoning* (2) we introduce a matrix evaluation framework that isolates narrative familiarity from inference complexity; and (3) we show that while models often succeed under conditions well-suited to pattern-based reasoning, they struggle with complex epistemic scenarios, exposing a spectrum of distinct behavioral profiles across current LLMs.

2 Background: Dynamic Epistemic Logic

Dynamic Epistemic Logic (DEL) is a formal framework for modeling what agents know, which possibilities they consider, and how this knowledge evolves as new information is revealed. In this framework, *agents* are entities whose knowledge is being modeled, and *public announcements* rule out possibilities inconsistent with newly available information, across multiple *rounds* of communication (Baltag et al., 1998; Van Ditmarsch et al., 2008). DEL is essential for multi-agent systems as widely emphasized in the literature (Bolander and

Andersen, 2011; Belardinelli and Lomuscio, 2009; Engesser et al., 2017). It captures a distinct form of reasoning: not only deriving facts, but also tracking higher-order knowledge (i.e., knowledge about other's knowledge) and how these evolve as agents reason about each other under partial information.

We focus on canonical DEL puzzles as our baseline. These puzzles are not merely examples, but compact tests of DEL's core mechanisms - knowledge inference, higher-order knowledge, common knowledge, and knowledge dynamics - requiring models to reason from partial observations, public updates, and others' uncertainty.

Figure 2 illustrates a two-child instance of the *Muddy Children* puzzle. Each child can see the other's forehead but not their own, so their uncertainty is represented by the possible worlds in the thought bubbles. In panel 2A, both remain uncertain about themselves; the father's announcement that at least one child is muddy rules out the all-clean world. This lets child 1 infer that she is muddy, and after she announces this in panel 2B, child 0 can further narrow the possibilities and reach his own conclusion in panel 2C.

3 Distinguishing Between Pattern-Based Reasoning and Epistemic Reasoning

We aim to distinguish between two ways in which LLMs may determine the correct response of an agent in DEL-style settings: *pattern-based reasoning* and *epistemic reasoning*. Under pattern-based reasoning, the model identifies the epistemic state as a variant of a familiar problem and applies the corresponding solution. Under epistemic reasoning, by contrast, the model derives the agent's response from the observations, public announcements, and prior answers in the interaction rounds.

3.1 Task Definition

To evaluate LLMs' epistemic reasoning ability, we place the LLM in the role of an agent within a simulated world (e.g., a child in the *Muddy Children* puzzle). The model's task is to infer its own status by reasoning over the information in its prompt.

Model Input. The model receives a prompt containing the public announcement, the agents' observations, and the public history of the interaction up to the queried round.

Model Output. The model is asked regarding its own status (e.g., "Is your forehead muddy?"). It must respond with exactly one of *"Yes"*, *"No"*, or *"I don't know"*. For any given round, exactly one response is correct based on the agent's current knowledge. To verify the model's response, we develop a symbolic solver grounded in Kripke possible worlds semantics (Kripke, 1959).[1]

3.2 Benchmark

To systematically isolate pattern-based from epistemic reasoning, our benchmark evaluates models in the two orthogonal axes introduced in Figure 1: narrative familiarity and inference complexity.

The Narrative Axis (Horizontal). This axis isolates the impact of a puzzle's surface form across seven distinct narratives, as illustrated in the puzzle icons in Figure 1: the classic *Muddy Children*, *Wise Men*, and *Blue-Eyed Islanders* puzzles, alongside four novel stories we design, vastly changing the setting as well as the surface forms - *Safety Inspection*, *Health Screening*, *Olympic Games*, and *Singing Contest* puzzles. Within each row, the underlying logical structure is preserved; announced bounds, observation rules, and interaction history remain fixed, while only the narrative changes. This structural parity lets us test whether models rely on pattern-based familiarity rather than reasoning over the provided epistemic structure.[2]

The Inference Axis (Vertical). This axis isolates reasoning difficulty. Moving down a column preserves the narrative framing while changing the underlying logic: standard symmetric setups, where agents observe everyone but themselves, are replaced by more complex asymmetric scenarios. Inspired by Sileo and Lernould (2023), we formalize asymmetry using an *observation matrix* $O \in \{0,1\}^{n\times n}$, where $O_{i,j} = 1$ if and only if agent i observes agent j's status. Entries are sampled i.i.d. from $\{0,1\}$, yielding fragmented partial views that models must track.

4 Experiments

We evaluate LLMs on the axes of our benchmark.

4.1 Experimental setup

Models. We evaluate nine models from various families, including ChatGPT (Ouyang et al., 2022), Gemini (Gemini Team, 2023), Claude (Joshi,

[1] See Appendix A.2 for details and pseudocode.

[2] See Appendix A.1 for prompt examples across all seven stories and both inference types.

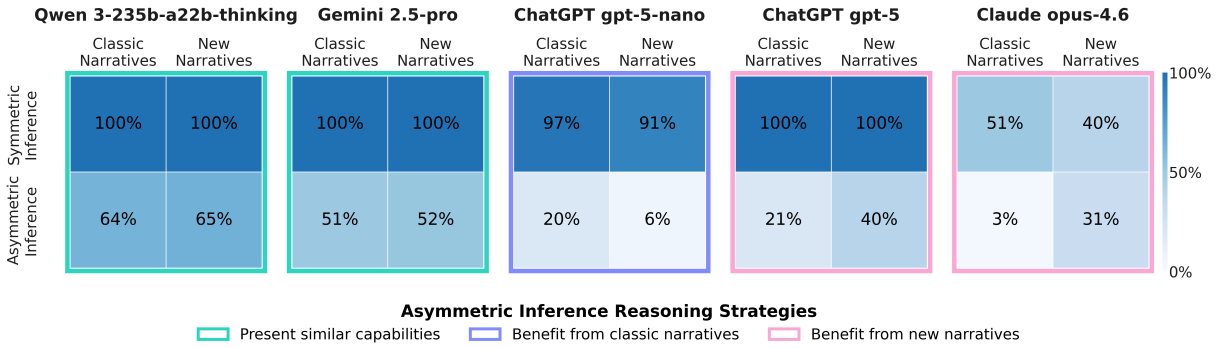


Figure 3: **Accuracy of evaluated LLMs across the two-dimensional benchmark.** While performance uniformly drops under complex asymmetric inference, models exhibit three distinct behavioral profiles based on narrative shifts in the asymmetric case: stable generalization (green), anchoring to familiar classic setups (purple), and improved performance when novel narratives strip away learned biases (pink).

2026), Mistral (Jiang et al., 2023), Qwen (Qwen Team, 2025) and OLMo (Groeneveld et al., 2024). We use the default temperature of 1 for the GPT-5 models due to API constraints and a temperature of 0 for all other models.

Dataset Generation. We generate 100 instances per puzzle narrative across the seven narratives and both inference types, yielding 1,400 unique puzzles, each with exactly 10 agents and a shared prompt structure. Models are queried exclusively at the pivotal round where an agent first accumulates sufficient information for a definitive deduction regarding their own status.

4.2 Results

Since some models fail to outperform the majority vote on the *Muddy Children* puzzle under symmetric inference and mostly reuse the same response (e.g., Olmo 3-7B-Instruct answers "No" in 99% of instances), we exclude them from the main analysis, keeping only Claude opus-4.6 for comparison.[3] Several observations can be drawn from our results.

Models demonstrate a strong ability to generalize from classic to new narratives within the symmetric setup, indicating strong pattern-based reasoning capabilities. Figure 3 shows that when moving from classic to new narratives under symmetric inference, most of the models successfully solve the new instances. This robustness to narrative modifications contrasts with prior findings, such as those of Jiang et al. (2024) and Mirzadeh et al. (2025), which observed a reversed trend where token perturbations significantly degraded model performance.

[3]See Appendix A.3 for evaluated models and full results.

In asymmetric setting, performance drops sharply, highlighting both inference complexity and a reliance on pattern-based reasoning. When the task shifts to asymmetric setups that require tracking possible worlds and can't be solved with the closed-form formula applicable in the symmetric setup, performance declines sharply across all models. While this is inherently a more complex task, both inference setups rely on the same foundational Kripke possible worlds semantics (Kripke, 1959). The difficulty arises not just from structural complexity, but from a lack of familiarity, given that novel puzzle formulations are mostly concentrated in the symmetric setting. This suggests that models have limited epistemic deduction abilities.

Under asymmetric inference, models differ in their sensitivity to narrative changes. Some models, such as Qwen 3-235b-a22b-thinking and Gemini 2.5-Pro, show consistently low performance across narratives, suggesting that the main bottleneck is the asymmetric epistemic structure itself. Others, such as ChatGPT gpt-5-nano, perform better on classic narratives, indicating stronger reliance on recognizable templates. In contrast, ChatGPT gpt-5 and Claude opus-4.6 perform better on novel narratives, suggesting that classic stories may sometimes trigger misleading associations rather than support the required inference.

5 Conclusion

We revisit the evaluation of LLMs on epistemic puzzles, arguing that the contrast between epistemic reasoning and memorization is too coarse for newer models. We show that strong performance seems to come from reducing to familiar problem structures rather than performing epistemic reasoning.

6 Limitations

Our benchmark focuses on DEL-style hidden-status puzzles, which capture central mechanisms of epistemic reasoning: partial observation, higher-order knowledge, public updates, and reasoning from others' uncertainty. However, they do not cover the full range of epistemic phenomena, such as private announcements, belief revision, deception, strategic communication, or planning under uncertainty.

References

Alexandru Baltag, Lawrence S. Moss, and Slawomir Solecki. 1998. The logic of public announcements, common knowledge, and private suspicions. In *Proceedings of the 7th Conference on Theoretical Aspects of Rationality and Knowledge (TARK VII)*, pages 43–56.

Francesco Belardinelli and Alessio Lomuscio. 2009. Quantified epistemic logics for reasoning about knowledge in multi-agent systems. *Artificial Intelligence*, 173(9-10):982–1013.

Thomas Bolander and Mikkel Birkegaard Andersen. 2011. Epistemic planning for single-and multi-agent systems. *Journal of Applied Non-Classical Logics*, 21(1):9–34.

Artur S d'Avila Garcez, Luis C Lamb, and Dov M Gabbay. 2009. *Neural-symbolic cognitive reasoning*. Springer.

Thorsten Engesser, Thomas Bolander, Robert Mattmüller, and Bernhard Nebel. 2017. Cooperative epistemic multi-agent planning for implicit coordination. *arXiv preprint arXiv:1703.02196*.

Ronald Fagin, Joseph Y Halpern, Yoram Moses, and Moshe Y Vardi. 1995. *Reasoning About Knowledge*. MIT Press, Cambridge, Massachusetts.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, and 1 others. 2024. Olmo: Accelerating the science of language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Bowen Jiang, Yangxinyu Xie, Zhuoqun Hao, Xiaomeng Wang, Tanwi Mallick, Weijie J. Su, Camillo J. Taylor, and Dan Roth. 2024. A peek into token bias: Large language models are not yet genuine reasoners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4722–4756, Miami, Florida, USA. Association for Computational Linguistics.

S. Joshi. 2026. Architectural advances and performance benchmarks of large language models in light of anthropic's claude opus 4.6. *Preprints.org*.

Saul A Kripke. 1959. A completeness theorem in modal logic1. *The journal of symbolic logic*, 24(1):1–14.

Huao Li, Yu Chong, Simon Stepputtis, Joseph Campbell, Dana Hughes, Charles Lewis, and Katia Sycara. 2023. Theory of mind for multi-agent collaboration via large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 180–192, Singapore. Association for Computational Linguistics.

Jingcong Liang, Shijun Wan, Xuehai Wu, Yitong Li, Qianglong Chen, Duyu Tang, Siyuan Wang, and Zhongyu Wei. 2025. Hardcorelogic: Challenging large reasoning models with long-tail logic puzzle games. *arXiv preprint arXiv:2510.12563*.

Bill Yuchen Lin, Ronan Le Bras, Kyle Richardson, Ashish Sabharwal, Radha Poovendran, Peter Clark, and Yejin Choi. 2025. Zebralogic: On the scaling limits of llms for logical reasoning. *arXiv preprint arXiv:2502.01100*.

Iman Mirzadeh, Keivan Alizadeh-Vahid, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2025. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. In *International Conference on Learning Representations*, volume 2025, pages 94743–94765.

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

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. Machine theory of mind. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4218–4227. PMLR.

Parshin Shojaee, Iman Mirzadeh, Maxwell Horton, Samy Bengio, Mehrdad Farajtabar, and 1 others.

2026. The illusion of thinking: Understanding the strengths and limitations of reasoning models via the lens of problem complexity. *Advances in Neural Information Processing Systems*, 38:108018–108059.

Damien Sileo and Antoine Lernould. 2023. MindGames: Targeting theory of mind in large language models with dynamic epistemic modal logic. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4570–4577, Singapore. Association for Computational Linguistics.

Andreas Stuhlmüller and Noah D Goodman. 2014. Reasoning about reasoning by nested conditioning: Modeling theory of mind with probabilistic programs. *Cognitive Systems Research*, 28:80–99.

Hans Van Ditmarsch, Wiebe van Der Hoek, and Barteld Kooi. 2008. *Dynamic epistemic logic*. Springer.

Hans Van Ditmarsh, Wiebe van der Hoek, and Barteld Kooi. 2007. Dynamic epistemic logic. *Synthese*.

Algorithm 1 GETSOLVERLABEL

Determines whether agent i knows their status at round j, handling both the symmetric (closed-form) and asymmetric (pre-computed history) cases.

Input: n: total agents; k: positive agents; j: query round; q: boundary value; $\tau \in \{\text{LOWER}, \text{UPPER}\}$: bound type; i: queried agent; $\mu \in \{0, 1\}$: agent i's true status; $rand$: TRUE for asymmetric inference; $\mathcal{H}$: knowledge history (only used in the asymmetric case, to avoid excessive computations.)

Output: 1 if agent i knows their status at round j, else 0

if $rand$ **then**
  return $i \in \mathcal{H}[j]$ {pre-computed by Algorithm 2}
end if
if $(\tau{=}\text{LOWER} \wedge q{=}0) \vee (\tau{=}\text{UPPER} \wedge q{=}n)$ **then**
  return 0 {uninformative bound}
end if
{If in the symmetric case, compute with closed formula}
$s \leftarrow |k - q| + 1$ {first round at which some agent gains knowledge}
if $(\tau{=}\text{LOWER} \wedge \mu{=}1) \vee (\tau{=}\text{UPPER} \wedge \mu{=}0)$ **then**
  return $j \geq s$ {positives under lower bound know first, negatives under upper bound know first}
else
  return $j \geq s{+}1$ {the other group knows one round later}
end if

Algorithm 2 Iterated Possible-Worlds Elimination

*Simulates iterative reasoning grounded in **Kripke possible worlds semantics (Kripke, 1959)**, computing for each round which agents know their own status.*

procedure GETKNOWLEDGEHISTORY(n, O, w^*, q, τ)

Input: n: total agents; $O \in \{0, 1\}^{n \times n}$: observation matrix; $w^* \in \{0, 1\}^n$: true world; q: boundary value; τ: bound type

Output: $\mathcal{H}$: list of knower sets for each round from 1 to $n{+}1$: $\mathcal{H}[1], \ldots, \mathcal{H}[n{+}1]$

$\mathcal{W} \leftarrow \{w \in \{0, 1\}^n : w \text{ consistent with } (\tau, q)\}$
$\mathcal{H} \leftarrow [\,]$
for $t = 1$ **to** $n{+}1$ **do**
  $K_t \leftarrow FindKnowers(n, O, w^*, \mathcal{W})$
  Append K_t to $\mathcal{H}$
  if $|\mathcal{W}| \leq 1$ **then**
    Fill $\mathcal{H}$'s remaining slots with K_t; **break**
  end if
  $\mathcal{W}' \leftarrow \{w \in \mathcal{W} : FindKnowers(n, O, w, \mathcal{W}) = K_t\}$
  if $\mathcal{W}' = \mathcal{W}$ **then** {steady state}
    Fill $\mathcal{H}$'s remaining slots with K_t; **break**
  end if
  $\mathcal{W} \leftarrow \mathcal{W}'$
end for
return $\mathcal{H}$

procedure FINDKNOWERS$(n, O, w^*, \mathcal{W})$

Input: n: total agents; $O \in \{0, 1\}^{n \times n}$: observation matrix; $w^* \in \{0, 1\}^n$: true world; $\mathcal{W}$: current candidate worlds

Output: K: indices of agents that know their own status in the current world based on the revealed information.

$K \leftarrow [\,]$
for $i = 0$ **to** $n{-}1$ **do**
  $V_i \leftarrow \{i' : O_{i,i'} = 1\}$ {agents observable by i}
  $\mathcal{W}_i \leftarrow \{w \in \mathcal{W} : w[V_i] = w^*[V_i]\}$ {worlds matching i's view}
  if $\mathcal{W}_i \neq \emptyset \wedge |\{w[i] : w \in \mathcal{W}_i\}| = 1$ **then**
    Append i to K
  end if
end for
return K

A Appendix - Prompt Examples, Solver Logic and Additional Results

A.1 Prompt examples

Figure 4 shows the prompt template used uniformly for the 7 puzzle stories under symmetric inference. Figures 5-11 show prompt examples for the 7 puzzle stories under the symmetric inference, all of the same parameters with the story modified. Figures 12-18 show prompt examples for the 7 puzzle stories under the asymmetric inference, all of the same parameters with the story modified.

A.2 Solver Logic

Our solver is grounded in Kripke possible worlds semantics (Kripke, 1959), and assigns a ground-

[Premise] There are n gymnasts… the coach gives each gymnast a sheet listing every other gymnast's name and status - advanced to the final or not - excluding their own status… The coach then publicly announces that at least q gymnasts made it into the final. Then, in repeated rounds, the coach asks each gymnast whether they know if they made it into the final. You are one of the gymnasts…
[Example] For example, if there are 2 gymnasts…
[Hypothesis] Suppose there are 10 gymnasts… at least 1 gymnast made it into the final. In round 1 the gymnasts who made it into the final answered that they knew if they made it into the final, while the gymnasts who did not make it answered that they did not know. In round 2, assuming that you see 1 gymnast that made it into the final and the rest didn't; did you make it into the final? Answer with "Yes", "No", or "I don't know".

Figure 4: **Reduced prompt example in the *Olympic Games* setup.** Each prompt contains a description of the protocol, a short example illustrating the required response format, and a test instance specifying the public interaction history and the agents' observations.

truth label to every problem instance, determining whether a queried agent knows their own status at a given reasoning round. The solver handles two inference settings through a unified interface (Algorithm 1).

Symmetric inference. When all agents observe all others (i.e., the observation matrix O has $O_{i,i'} = 1$ for all $i \neq i'$), the epistemic state of every agent can be derived analytically. Given the public announcement bound (τ, q) and the true number of positive agents k, the first round at which any agent gains knowledge is $s = |k - q| + 1$. Under a lower bound, positive agents ($\mu = 1$) know at round s and negative agents at round $s + 1$; the roles are reversed under an upper bound. PRODUCESOLVERLABEL (Algorithm 1) encodes this closed-form rule directly, returning 1 if and only if the query round j meets or exceeds the agent's knowledge threshold. If the bound is uninformative (e.g., "at least 0" or "at most n"), no agent can ever know, and the label is 0.

Asymmetric inference. When agents observations of each other are detailed through an arbitrary binary matrix $O \in \{0, 1\}^{n \times n}$, there is no closed form. Algorithm 2 implements *iterated possible world elimination*, pre-computes a knowledge history $\mathcal{H}$ before any label is queried. Starting from the set $\mathcal{W}$ of all worlds consistent with the public announcement, the procedure GETKNOWLEDGEHISTORY iterates round by round:

A.3 Experimental Setup, Evaluated Models and Additional Results

Problem Instance structure. We define a problem instance using the tuple (n, k, q, j, O). In this formulation, n represents the total number of agents (e.g., the number of children in the *Muddy Children* puzzle), and k denotes the count of positive agents possessing the hidden status (e.g., being "muddy"). The parameter q represents the lower bound value in the public announcement (e.g., "at least q are muddy"), while j specifies the target reasoning round at which the agent's knowledge is evaluated. Finally, $O \in \{0, 1\}^{n \times n}$ is an observation matrix where the entry $O_{i,i'} = 1$ if agent i observes the status of agent i', and $O_{i,i'} = 0$ otherwise.

An example of this parameterization is illustrated in Figure 2. There, the number of agents is $n = 2$, the number of agents possessing the hidden status (muddy) is $k = 1$, the boundary is a lower bound of value $q = 1$, and the observations are as in the matrix $O = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, meaning each child observes the other child, but not themselves.

Our experiments use a fixed number of agents $n = 10$.

Symmetric Inference. We covered all valid combinations of marked agents $k \in [0, 10]$, bound value $q \in [0, k]$, and reasoning rounds of inference. The diagonal of the observation matrices is 0 ($O_{i,i} = 0$) and the rest is filled with 1 ($O_{i,j} = 1, i \neq j$).

Asymmetric Inference. We use a lower bound $q = 8$, reasoning round $j = 3$ in all puzzles, and observation matrices sampled randomly from $\{0, 1\}^{n \times n}$ and marked agents both allowing inference in reasoning round 3.

Post-process and verification. In post processing, we evaluate only the first non-empty line of each response, since the prompt instructs the model to provide its answer on the first line and any explanation on a new line. We accept exactly one of "Yes", "No", and "I don't know", while allowing minor formatting variation such as surrounding whitespace, a final period, and apostrophe variants in "don't".

Results. Table 2 lists all models included in our evaluation, together with their parameter counts when known. Table 1 reports results for models

Table 1: Accuracy (%) of omitted models over the *Muddy Children* puzzle with symmetric inference.

Model	Accuracy (%)
ChatGPT gpt-4	52%
Gemini 2.5 Flash Lite	48%
OLMo 3-7B-Instruct	46%
Mistral Small-24B-Instruct-2501	52%

Symmetric Inference - The Muddy Children Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n children playing in the mud. Their father calls them to the house and arranges them in a semicircle, so that each child can clearly see every other child, but not themselves. The father then publicly announces that at least q children have muddy foreheads. Then, in repeated rounds, the father asks each child whether they know if their own forehead is muddy. In each round, the children may rely only on the round number, the answers from the previous rounds, and the observed muddy/clean status of the other children.
You are one of the children, and in round j, you need to answer the question "Is your forehead muddy?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 children, and the father announces that at least 2 have muddy foreheads, then in round 1, you should respond "Yes" for both children.
Answer the following question:
[Hypothesis] Suppose there are 10 children and the father publicly announces that at least 1 child has a muddy forehead. In round 1 the muddy children answered that they knew if they had mud on their forehead, while the clean children answered that they did not know. In round 2, assuming that you see 1 child with a muddy forehead and the rest are clean; is your forehead muddy? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 5: Prompt example - symmetric inference, the *Muddy Children* puzzle.

Model	# Parameters
ChatGPT gpt-5	?
ChatGPT gpt-5-nano	?
ChatGPT gpt-4	?
Claude opus-4.6	?
Gemini 2.5-Pro	?
Gemini 2.5-Flash-Lite	?
Qwen 3-a22b-thinking	235B
Mistral Small-Instruct	24B
OLMo 3 Instruct	7B

Table 2: **Evaluated models.** Parameter counts are listed only when publicly disclosed by the provider.

omitted from our main analysis due to their low performance on the *Muddy Children* puzzle under symmetric inference. All four models score near random chance: although the prompt offers three options (*yes*, *no*, *I don't know*), we evaluate only rounds where the correct answer is determinately *yes* or *no*, placing the effective random baseline at 50%. Table 3 reports macro F1 scores for the main-analysis models across all four benchmark quadrants. In the symmetric setup, the label distribution is 55 Yes and 45 No; in the asymmetric setup, it is 60 Yes and 40 No. For each quadrant, macro F1 is computed separately per puzzle over the two ground-truth classes (*Yes* and *No*), then averaged across puzzles. Since models are queried only at the pivotal round, *I don't know* is never a correct answer and is excluded from the average. While the absolute values naturally differ from the accuracy results in Figure 3, as the two metrics capture different aspects of model performance, the macro-F1 scores exhibit the same qualitative patterns: a consistent drop from symmetric to asymmetric settings across all models, and the same relative ordering of behavioral profiles observed in the accuracy results.

Symmetric Inference - The Wise Men Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
A king in a distant country searches for a new advisor, and calls n wise men to his court for a final test. The king places a hat - either white or blue - on each wise man's head, and arranges the wise men in a circle so that each wise man can clearly see every other wise man, but not themselves. The king then publicly announces that at least q wise men are wearing blue hats. Then, in repeated rounds, the king asks each wise man whether they know if their own hat is blue. In each round, the wise men may rely only on the round number, the answers from the previous rounds, and the observed hat colors of the other wise men.
You are one of the wise men, and in round j, you need to answer the question "Are you wearing a blue hat?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 wise men, and the king announces that at least 2 wise men are wearing blue hats, then in round 1, you should respond "Yes" for both wise men.
Answer the following question:
[Hypothesis] Suppose there are 10 wise men and the king publicly announces that at least 1 wise man is wearing a blue hat. In round 1 the wise men wearing blue hats answered that they knew if their own hat was blue, while the wise men wearing white hats answered that they did not know. In round 2, assuming that you see 1 wise man wearing a blue hat and the rest are wearing white hats; are you wearing a blue hat? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 6: Prompt example - symmetric inference, the *Wise Men* puzzle.

Model	Classic Sym.	Classic Asym.	New Sym.	New Asym.
Qwen 3-235b-a22b-thinking	1.0000	0.5823	0.9974	0.6107
Gemini 2.5-pro	0.9966	0.5032	1.0000	0.5951
ChatGPT gpt-5-nano	0.9864	0.2787	0.9415	0.0839
ChatGPT gpt-5	1.0000	0.3375	1.0000	0.5616
Claude opus-4.6	0.6482	0.0438	0.5091	0.2482

Table 3: **Macro F1 scores across the four benchmark quadrants.** For each quadrant, macro F1 is computed separately per puzzle over the two ground-truth classes (*Yes* and *No*), then averaged across puzzles. Since models are queried only at the pivotal round, *I don't know* is never a correct answer and is excluded from the average.

Symmetric Inference - The Blue-Eyed Islanders Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n islanders living on a remote island, where talking about eye color is forbidden. A foreigner arrives on the island, captivated by the islanders' eye colors and eager to discuss the topic with them. He then gathers the islanders in a circle, so that each islander can clearly see every other islander's eye color, but not their own. The foreigner then publicly announces that at least q islanders have blue eyes. Then, in repeated rounds, the foreigner asks each islander whether they know if their own eyes are blue. In each round, the islanders may rely only on the round number, the answers from the previous rounds, and the observed eye colors of the other islanders.
You are one of the islanders, and in round j, you need to answer the question "Do you have blue eyes?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 islanders, and the foreigner announces that at least 2 islanders have blue eyes, then in round 1, you should respond "Yes" for both islanders.
Answer the following question:
[Hypothesis] Suppose there are 10 islanders and the foreigner publicly announces that at least 1 islander has blue eyes. In round 1 the blue-eyed islanders answered that they knew if their own eyes were blue, while the brown-eyed islanders answered that they did not know. In round 2, assuming that you see 1 islander with blue eyes and the rest have brown eyes; do you have blue eyes? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 7: Prompt example - symmetric inference, the *Blue-Eyed Islanders* puzzle.

Symmetric Inference - The Olympic Games Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n gymnasts competing in an Olympic qualification round. At the end of the round, the coach gives each gymnast a sheet listing every other gymnast's name and status - advanced to the final or not - excluding their own status. The gymnasts can't share the given information with others. The coach then publicly announces that at least q gymnasts made it into the final. Then, in repeated rounds, the coach asks each gymnast whether they know if they made it into the final. In each round, the gymnasts may rely only on the round number, the answers from the previous rounds, and the observed status of the other gymnasts.
You are one of the gymnasts, and in round j, you need to answer the question "Did you make it into the final?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 gymnasts, and the coach announces that at least 2 gymnasts made it into the final, then in round 1, you should respond "Yes" for both gymnasts.
Answer the following question:
[Hypothesis] Suppose there are 10 gymnasts and the coach publicly announces that at least 1 gymnast made it into the final. In round 1 the gymnasts who made it into the final answered that they knew if they made it into the final, while the gymnasts who did not make it answered that they did not know. In round 2, assuming that you see 1 gymnast that made it into the final and the rest didn't; did you make it into the final? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 8: Prompt example - symmetric inference, the *Olympic Games* puzzle.

Symmetric Inference - The Health Screening Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n patients attending a routine health screening at a clinic. After reviewing the preliminary test results, the nurse gives each patient a sheet listing every other patient's name and test result - clear or needs follow-up - excluding their own result. The patients can't share the given information with others. The nurse then publicly announces that at least q patients received a clear result. Then, in repeated rounds, the nurse asks each patient whether they know if their own result is clear. In each round, the patients may rely only on the round number, the answers from the previous rounds, and the observed test results of the other patients.
You are one of the patients, and in round j, you need to answer the question "Did you receive a clear result?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 patients, and the nurse announces that at least 2 patients received a clear result, then in round 1, you should respond "Yes" for both patients.
Answer the following question:
[Hypothesis] Suppose there are 10 patients and the nurse publicly announces that at least 1 patient received a clear result. In round 1 the patients who received a clear result answered that they knew if their result was clear, while the patients who need a follow-up answered that they did not know. In round 2, assuming that you see 1 patient who received a clear result and the rest need a follow-up; did you receive a clear result? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 9: Prompt example - symmetric inference, the *Health Screening* puzzle.

Symmetric Inference - The Safety Inspection Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n tenants living in an apartment building undergoing a safety inspection. After inspecting all units, the safety warden clips a colored lanyard around each tenant's neck before they gather in the lobby: green if their unit passed, red if their unit requires repairs. The warden arranges the tenants in a circle in the lobby so that each tenant can clearly see every other tenant's lanyard, but not their own. The safety warden then publicly announces that at least q tenants have green lanyards. Then, in repeated rounds, the safety warden asks each tenant whether they know if their own unit passed. In each round, the tenants may rely only on the round number, the answers from the previous rounds, and the observed lanyards of the other tenants.
You are one of the tenants, and in round j, you need to answer the question "Did your unit pass the inspection?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 tenants, and the safety warden announces that at least 2 tenants have green lanyards, then in round 1, you should respond "Yes" for both tenants.
Answer the following question:
[Hypothesis] Suppose there are 10 tenants and the safety warden publicly announces that at least 1 tenant has a green lanyard. In round 1 the tenants with green lanyards answered that they knew if their unit passed the inspection, while the tenants with red lanyards answered that they did not know. In round 2, assuming that you see 1 tenant with a green lanyard and the rest have red lanyards; did your unit pass the inspection? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

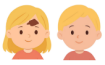
Figure 10: Prompt example - symmetric inference, the *Safety Inspection* puzzle.

Symmetric Inference - The Singing Contest Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n singers competing in a singing contest. Each singer receives a score from the judges, which determines whether they are included on this year's list of rising singing talents. At the end of the contest, the host sticks a sticker on each of the singers' backs; X if the singer didn't get into the list and V if they did. the host then arranges the singers so that each singer can clearly see every other singer's sticker, but not their own. The host then publicly announces that at least q singers have V stickers. Then, in repeated rounds, the host asks each singer whether they know if they made it into the list. In each round, the singers may rely only on the round number, the answers from the previous rounds, and the observed stickers of the other singers.
You are one of the singers, and in round j, you need to answer the question "Did you make it into the list of rising singing talents?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 singers, and the host announces that at least 2 singers have V stickers, then in round 1, you should respond "Yes" for both singers.
Answer the following question:
[Hypothesis] Suppose there are 10 singers and the host publicly announces that at least 1 singer has a V sticker. In round 1 the singers with V stickers answered that they knew if they made it into the list, while the singers with X stickers answered that they did not know. In round 2, assuming that you see 1 singer with a V sticker and the rest have X stickers; did you make it into the list of rising singing talents? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 11: Prompt example - symmetric inference, the *Singing Contest* puzzle.

Asymmetric Inference - The Muddy Children Puzzle



[Premise] The following is a problem with parameters that may vary between instances:
There are n children playing in the mud. Their father calls them to the house and arranges them in a certain order that creates an observation matrix. Each entry (i,k) in the observation matrix indicates whether child i knows the status of child k. A value of 1 means child i knows child k's status, while a value of 0 means they do not know. The father then publicly announces that at least q children have muddy foreheads. Then, in repeated rounds, the father asks each child whether they know if their own forehead is muddy. In each round, the children may rely only on the round number, the answers from the previous rounds, and the observed muddy/clean status of the other children.
You are one of the children, and in round j, you need to answer the question "Is your forehead muddy?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 children, and the father announces that at least 2 children have muddy foreheads, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both children.
Answer the following question:
[Hypothesis] Suppose there are 10 children indexed by 0-9, and the father publicly announces that at least 8 children have muddy foreheads. In round 1, children 4, 5, 6, and 7 answered that they knew if they had mud on their forehead, while the rest answered that they did not know what was on their forehead. In round 2, children 2, 3, 4, 5, 6, and 7 answered that they knew if they had mud on their forehead, while the rest answered that they did not know what was on their forehead. In round 3, assuming that you are child 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that children 2, 3, 4, 5, and 7 have muddy foreheads and child 9 has a clean forehead; is your forehead muddy? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

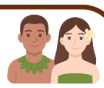
Figure 12: Prompt example - asymmetric inference, the *Muddy Children* puzzle.

Asymmetric Inference - The Wise Men Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
A king in a distant country searches for a new advisor, and calls n wise men to his court for a final test. The king places a hat - either white or blue - on each wise man's head, and arranges the wise men in a certain order that creates an observation matrix. Each entry (i,k) in the observation matrix indicates whether wise man i knows the hat color of wise man k. A value of 1 means wise man i knows wise man k's hat color, while a value of 0 means they do not know. The king then publicly announces that at least q wise men are wearing blue hats. Then, in repeated rounds, the king asks each wise man whether they know if their own hat is blue. In each round, the wise men may rely only on the round number, the answers from the previous rounds, and the observed hat colors of the other wise men.
You are one of the wise men, and in round j, you need to answer the question "Are you wearing a blue hat?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 wise men, and the king announces that at least 2 wise men are wearing blue hats, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both wise men.
Answer the following question:
[Hypothesis] Suppose there are 10 wise men indexed by 0-9, and the king publicly announces that at least 8 wise men are wearing blue hats. In round 1, wise men 4, 5, 6, and 7 answered that they knew if their own hat was blue, while the rest answered that they did not know if their own hat was blue. In round 2, wise men 2, 3, 4, 5, 6, and 7 answered that they knew if their own hat was blue, while the rest answered that they did not know if their own hat was blue. In round 3, assuming that you are wise man 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that wise men 2, 3, 4, 5, and 7 are wearing blue hats and wise man 9 is wearing a white hat; are you wearing a blue hat? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 13: Prompt example - asymmetric inference, the *Wise Men* puzzle.

Asymmetric Inference - The Blue-Eyed Islanders Puzzle



[Premise] The following is a problem with parameters that may vary between instances:
There are n islanders living on a remote island, where talking about eye color is forbidden. A foreigner arrives on the island, captivated by the islanders' eye colors and eager to discuss the topic with them. He then arranges the islanders in a certain order that creates an observation matrix. Each entry (i,k) in the observation matrix indicates whether islander i knows the eye color of islander k. A value of 1 means islander i knows islander k's eye color, while a value of 0 means they do not know. The foreigner then publicly announces that at least q islanders have blue eyes. Then, in repeated rounds, the foreigner asks each islander whether they know if their own eyes are blue. In each round, the islanders may rely only on the round number, the answers from the previous rounds, and the observed eye colors of the other islanders.
You are one of the islanders, and in round j, you need to answer the question "Do you have blue eyes?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 islanders, and the foreigner announces that at least 2 islanders have blue eyes, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both islanders.
Answer the following question:
[Hypothesis] Suppose there are 10 islanders indexed by 0-9, and the foreigner publicly announces that at least 8 islanders have blue eyes. In round 1, islanders 4, 5, 6, and 7 answered that they knew if their own eyes were blue, while the rest answered that they did not know if their own eyes were blue. In round 2, islanders 2, 3, 4, 5, 6, and 7 answered that they knew if their own eyes were blue, while the rest answered that they did not know if their own eyes were blue. In round 3, assuming that you are islander 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that islanders 2, 3, 4, 5, and 7 have blue eyes and islander 9 has brown eyes; do you have blue eyes? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 14: Prompt example - asymmetric inference, the *Blue-Eyed Islanders* puzzle.

Asymmetric Inference - The Olympic Games Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n gymnasts competing in an Olympic qualification round. At the end of the round, the coach gives each gymnast a sheet listing some gymnast's name and status - advanced to the final or not - that creates an observation matrix. The gymnasts can't share the given information with others. Each entry (i,k) in the observation matrix indicates whether gymnast i knows the status of gymnast k. A value of 1 means gymnast i knows gymnast k's status, while a value of 0 means they do not know. The coach then publicly announces that at least q gymnasts made it into the final. Then, in repeated rounds, the coach asks each gymnast whether they know if they made it into the final. In each round, the gymnasts may rely only on the round number, the answers from the previous rounds, and the observed status of the other gymnasts.
You are one of the gymnasts, and in round j, you need to answer the question "Did you make it into the final?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 gymnasts, and the coach announces that at least 2 gymnasts made it into the final, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both gymnasts.
Answer the following question:
[Hypothesis] Suppose there are 10 gymnasts indexed by 0-9, and the coach publicly announces that at least 8 gymnasts made it into the final. In round 1, gymnasts 4, 5, 6, and 7 answered that they knew if they made it into the final, while the rest answered that they did not know if they made it into the final. In round 2, gymnasts 2, 3, 4, 5, 6, and 7 answered that they knew if they made it into the final, while the rest answered that they did not know if they made it into the final. In round 3, assuming that you are gymnast 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that gymnasts 2, 3, 4, 5, and 7 made it into the final and gymnast 9 did not make it into the final; did you make it into the final? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 15: Prompt example - asymmetric inference, the *Olympic Games* puzzle.

Asymmetric Inference - The Health Screening Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n patients attending a routine health screening at a clinic. After reviewing the preliminary test results, the nurse gives each patient a sheet listing some patients' names and test result - clear or needs follow-up - that creates an observation matrix. The patients can't share the given information with others. Each entry (i,k) in the observation matrix indicates whether patient i knows the status of patient k. A value of 1 means patient i knows patient k's status, while a value of 0 means they do not know. The nurse then publicly announces that at least q patients received a clear result. Then, in repeated rounds, the nurse asks each patient whether they know if their own result is clear. In each round, the patients may rely only on the round number, the answers from the previous rounds, and the observed test results of the other patients.
You are one of the patients, and in round j, you need to answer the question "Did you receive a clear result?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 patients, and the nurse announces that at least 2 patients received a clear result, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both patients.
Answer the following question:
[Hypothesis] Suppose there are 10 patients indexed by 0-9, and the nurse publicly announces that at least 8 patients received a clear result. In round 1, patients 4, 5, 6, and 7 answered that they knew if their result was clear, while the rest answered that they did not know if their result was clear. In round 2, patients 2, 3, 4, 5, 6, and 7 answered that they knew if their result was clear, while the rest answered that they did not know if their result was clear. In round 3, assuming that you are patient 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that patients 2, 3, 4, 5, and 7 received a clear result and patient 9 needs a follow-up; did you receive a clear result? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 16: Prompt example - asymmetric inference, the *Health Screening* puzzle.

Asymmetric Inference - The Safety Inspection Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n tenants living in an apartment building undergoing a safety inspection. After inspecting all units, the safety warden clips a colored lanyard around each tenant's neck before they gather in the lobby: green if their unit passed, red if their unit requires repairs. The warden arranges the tenants in a certain order in the lobby that creates an observation matrix. Each entry (i,k) in the observation matrix indicates whether tenant i knows the status of tenant k. A value of 1 means tenant i knows tenant k's status, while a value of 0 means they do not know. The safety warden then publicly announces that at least q tenants have green lanyards. Then, in repeated rounds, the safety warden asks each tenant whether they know if their own unit passed. In each round, the tenants may rely only on the round number, the answers from the previous rounds, and the observed lanyards of the other tenants.
You are one of the tenants, and in round j, you need to answer the question "Did your unit pass the inspection?" based on what you see, the round number, the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 tenants, and the safety warden announces that at least 2 tenants have green lanyards, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both tenants.
Answer the following question:
[Hypothesis] Suppose there are 10 tenants indexed by 0-9, and the safety warden publicly announces that at least 8 tenants have green lanyards. In round 1, tenants 4, 5, 6, and 7 answered that they knew if their unit passed the inspection, while the rest answered that they did not know if their unit passed the inspection. In round 2, tenants 2, 3, 4, 5, 6, and 7 answered that they knew if their unit passed the inspection, while the rest answered that they did not know if their unit passed the inspection. In round 3, assuming that you are tenant 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that tenants 2, 3, 4, 5, and 7 have green lanyards and tenant 9 has a red lanyard; did your unit pass the inspection? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 17: Prompt example - asymmetric inference, the *Safety Inspection* puzzle.

Asymmetric Inference - The Singing Contest Puzzle

[Premise] The following is a problem with parameters that may vary between instances:
There are n singers competing in a singing contest. Each singer receives a score from the judges, which determines whether they are included on this year's list of rising singing talents. At the end of the contest, The host sticks a sticker on each of the singers' backs; X if the singer didn't get into the list and V if they did. the host then arranges the singers in a certain order that creates an observation matrix. Each entry (i,k) in the observation matrix indicates whether singer i knows the status of singer k. A value of 1 means singer i knows singer k's status, while a value of 0 means they do not know. The host then publicly announces that at least q singers have V stickers. Then, in repeated rounds, the host asks each singer whether they know if they made it into the list. In each round, the singers may rely only on the round number, the answers from the previous rounds, and the observed stickers of the other singers.
You are one of the singers, and in round j, you need to answer the question "Did you make it into the list of rising singing talents?" based on what you see, the round number, and the answers from the previous rounds. Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.
[Example] For example, if there are 2 singers, and the host announces that at least 2 singers have V stickers, then in round 1, given the following observation matrix:
[0,1]
[1,0]
you should respond "Yes" for both singers.
Answer the following question:
[Hypothesis] Suppose there are 10 singers indexed by 0-9, and the host publicly announces that at least 8 singers have V stickers. In round 1, singers 4, 5, 6, and 7 answered that they knew if they made it into the list, while the rest answered that they did not know if they made it into the list. In round 2, singers 2, 3, 4, 5, 6, and 7 answered that they knew if they made it into the list, while the rest answered that they did not know if they made it into the list. In round 3, assuming that you are singer 0 and observe the world as described in the following observation matrix:
[0, 0, 1, 1, 1, 1, 0, 1, 0, 1]
[0, 0, 1, 0, 1, 0, 1, 0, 0, 0]
[0, 1, 0, 0, 1, 1, 1, 1, 1, 0]
[0, 0, 1, 0, 1, 1, 0, 1, 1, 0]
[0, 1, 1, 0, 1, 1, 1, 1, 0, 1]
[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
[1, 0, 0, 1, 0, 1, 1, 0, 1, 0]
[1, 0, 1, 0, 0, 0, 0, 1, 0, 1]
[1, 1, 1, 1, 0, 1, 0, 1, 0, 0]
[0, 0, 1, 1, 0, 0, 0, 0, 1, 0]
Specifically, you see that singers 2, 3, 4, 5, and 7 have V stickers and singer 9 has an X sticker; did you make it into the list of rising singing talents? Answer with "Yes", "No", or "I don't know". Write your answer without additional text and, on a new line, explain your reasoning.

Figure 18: Prompt example - asymmetric inference, the *Singing Contest* puzzle.