

Query Your Way Out: Benchmarking Multi-Turn Knowledge Reasoning through Generalized Puzzle Solving

Xin Cheng¹ Junming Yang¹ Ning Xu¹

Abstract

Puzzles offer a natural testbed for Large Language Models (LLMs) reasoning because they pose verifiable goals under partial information and require multi-step decisions to reach a solution. However, existing puzzle benchmarks typically focus on narrow environments with rigid mechanics and tightly constrained solution targets. To bridge this gap, we introduce the Active Probing Benchmark (APB), which generalizes puzzle-solving by casting concept identification as an interactive reasoning task that integrates specialized knowledge assessment into the puzzle setting. Specifically, APB evaluates whether models can acquire information via multi-turn active probing and inference to pinpoint 960 target concepts of specialized difficulty across eight diverse academic domains. Our evaluation reveals a counter-intuitive phenomenon: most open-source reasoning models significantly underperform their standard instruction counterparts. This finding suggests that in interactive settings, these models struggle to leverage procedural reasoning to reliably access and apply specialized domain knowledge. Beyond these empirical insights, APB provides a scalable pipeline for generating benchmarks that probe knowledge reasoning under interaction across diverse settings.

1. Introduction

The rapid progress of Large Language Models (LLMs) has heightened the demand for evaluation paradigms that reliably reflect underlying model capabilities (Singh et al., 2024). Although existing benchmarks have enabled systematic comparison, their validity is increasingly constrained by practical challenges in benchmark design and upkeep. Prior work has noted that overlap between evaluation data and

large-scale pre-training corpora may confound performance estimates in some settings (Li et al., 2023; Jacovi et al., 2023; Roberts et al., 2023), motivating interest in evaluation protocols that emphasize reasoning over memorization (Xu et al., 2024; Zhou et al., 2025). Dynamic benchmarks have emerged as one response (Kiela et al., 2021; Shiral et al., 2022), including continuously updated datasets such as LatentEval (Li et al., 2024) and LiveBench (White et al., 2024). However, such approaches depend on sustained data curation and may be difficult to scale. Other paradigms, such as crowdsourced pairwise comparison in Chatbot Arena (Chiang et al., 2024), introduce variability driven by subjective or stylistic factors. Meanwhile, agent-based benchmarks (Liu et al., 2023; Zhou et al., 2023; Chan et al., 2024) primarily target domains with explicit verification signals, leaving the evaluation of general knowledge reasoning relatively underexplored.

Puzzle tasks have become a widely used evaluation paradigm for LLMs. With clearly defined objectives and verifiable endpoints, they enable assessment not only of outcomes but also of trajectory-level signals of exploratory reasoning (Giadikiaroglou et al., 2024). Prior work has often focused on specific classes of puzzles (Chollet, 2019; Lin et al., 2021; Zhang et al., 2025), including riddles, Sudoku, and related logic or combinatorial games. Recent benchmarks have advanced this approach. For instance, SPLAT (Chen et al., 2024) frames tasks as clue-seeking within pre-scripted narratives, and MTP (Badola et al., 2025) assesses models through carefully designed interactive environments. However, such environments typically define narrow action spaces that encourage models to depend on fixed strategies rather than engaging in deep knowledge reasoning. Moreover, the highly constrained goals of these scenarios restrict the evaluation method from being generalized to broader contexts. Our work extends the use of puzzle-solving by integrating multi-turn active probing with domain knowledge, aiming instead to bridge this gap and evaluate whether models can demonstrate a general understanding of any established concept.

In this work, we introduce the Active Probing Benchmark (APB), which expands the scope of puzzle-solving evaluation (Figure 1). Specifically, APB assesses knowledge

¹School of Computer Science and Engineering, Southeast University, Nanjing, China. Correspondence to: Ning Xu <xning@seu.edu.cn>.

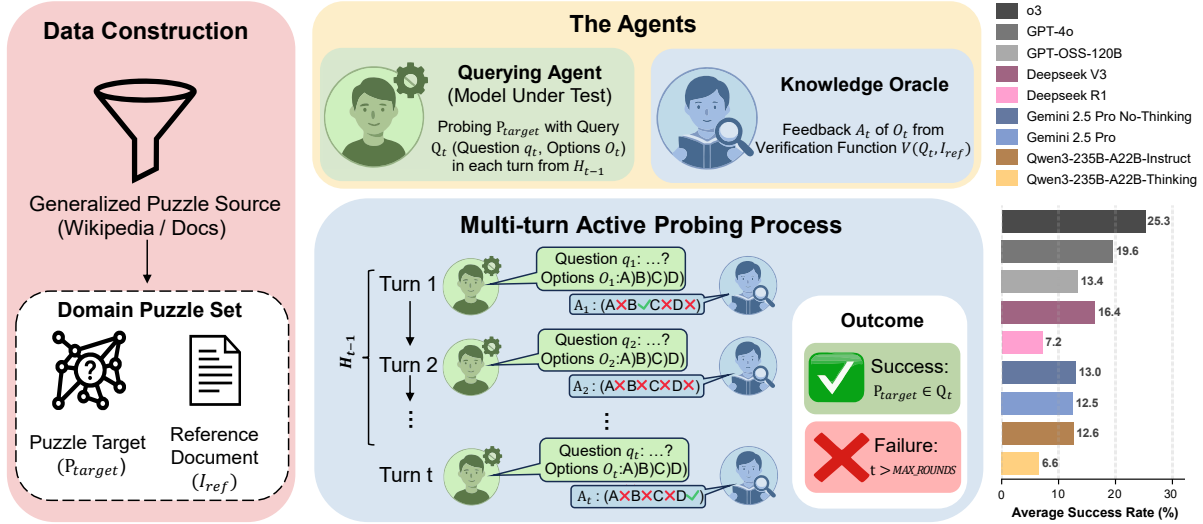


Figure 1. Overview of the Active Probing Benchmark (APB) and representative performance results. **(Left)** Data construction: each puzzle specifies a target concept P_{target} drawn from a generalized knowledge source (e.g., Wikipedia or domain documents) along with a reference document I_{ref} . **(Middle)** Multi turn active probing: the model under test iteratively asks multiple-choice queries Q_t , receives feedback A_t from a knowledge oracle, and updates its hypotheses to reduce the candidate space until the target is identified or a turn budget is exhausted. **(Right)** Performance comparison: average success rates show that several open source “Thinking” models underperform their “No-Thinking” counterparts. Full results are reported in Table 2.

reasoning under interaction: the ability to identify a target concept through active multi-turn inquiry. This setup parallels practical decision making settings such as medical diagnosis, where a physician narrows possibilities through sequential questioning. Similarly, APB requires models to iteratively ask informative questions (Figure 2), use the returned signals to refine hypotheses, and efficiently reduce the candidate space until the target concept is identified. Since Wikipedia is a major source in the pre-training corpora of foundation models, APB constructs 960 puzzles centered on concepts of specialized difficulty spanning eight academic domains, including STEM, the humanities, and health sciences. By emphasizing information seeking interaction over fixed game mechanics, APB aims to evaluate how effectively LLMs can access and apply specialized knowledge in a controlled setting with explicit objectives.

Evaluation of a wide range of leading LLMs on APB reveals the benchmark’s significant challenge where even the top-performing model GPT-5 achieved a success rate of only 33.4% (Table 2). The results uncover a counter-intuitive phenomenon: most open-source reasoning-enhanced models significantly underperform their no-thinking counterparts, a trend exemplified by models including GPT-OSS-120B, DeepSeek R1, and Qwen3-235B-A22B-Thinking (Table 2). Detailed analysis of response patterns also highlights distinct failure modes in these models. For instance, Gemini’s relative weakness stems from a tendency to frequently offer a conclusive guess which points to potential inherent flaws. Additionally, open source reasoning models tend to commit

to specific candidate targets early and then probe in a confirmatory manner, leading to inefficient probing when the true target lies outside the initial hypothesis set. These analysis suggests that in interactive settings, the models struggle to leverage procedural reasoning to reliably access and apply specialized domain knowledge.

Beyond these empirical findings, the construction of the APB introduces a scalable method to rapidly generate diagnostic benchmarks in specific specialized domain, enabling a robust evaluation of knowledge reasoning. We further showcase its application by building an IC Back-End Puzzle Benchmark, featuring 480 puzzles from four EDA command tools in the IC back-end design (Appendix A). The core contribution can be summarized as follows:

- We generalize the paradigm of puzzle-solving assessment by expanding the action and target spaces, shifting the focus from game mechanics to general knowledge reasoning across broader domains.
- We introduce the Active Probing Benchmark (APB) and a scalable pipeline for generating evaluations of specialized difficulty, instantiated by 960 puzzles derived from expert-level concepts.
- We identify a critical proficiency gap in leading open-source reasoning models, observing that under the interactive evaluation paradigm of APB, these models struggle to leverage procedural reasoning to reliably access and apply specialized domain knowledge.

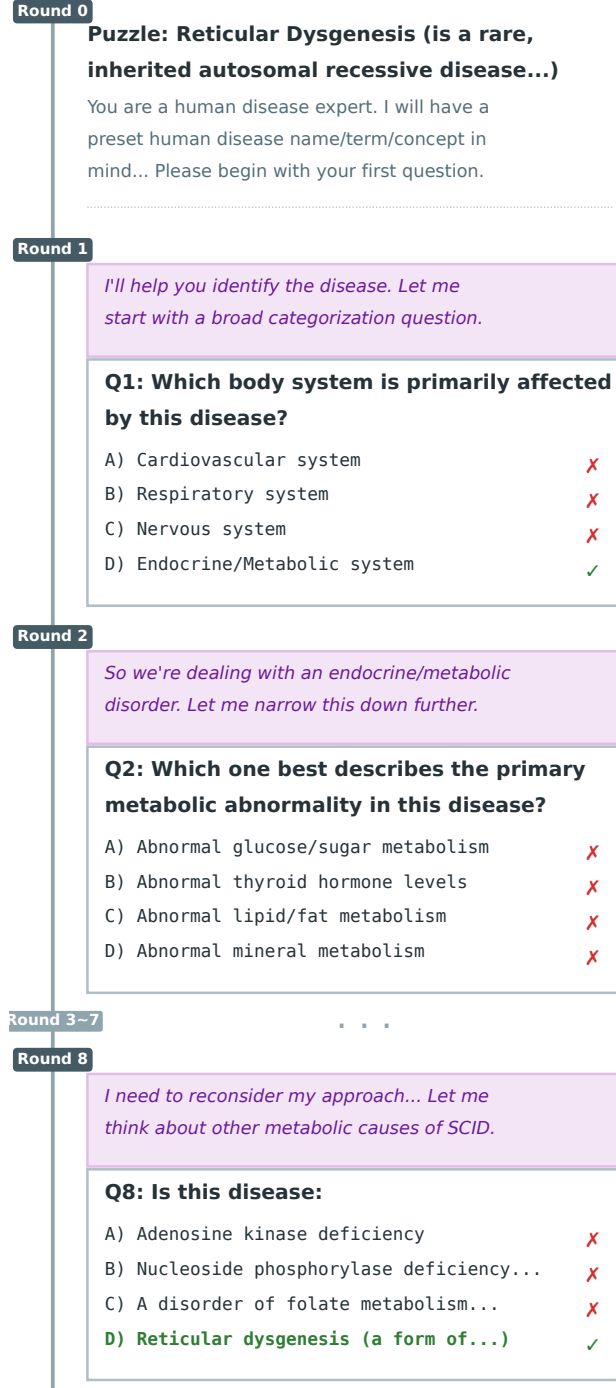


Figure 2. Example episode from the Human Disease domain in APB, where the agent successfully identifies the target concept “Reticular dysgenesis”. Across turns, the agent acquires diagnostic evidence through queries and feedback, progressively narrowing the hypothesis set until the target is determined. The resulting interaction trace provides process-level evidence of how the model accesses and applies specialized knowledge.

2. Related Work

2.1. Benchmark Evaluation

The evaluation of large language models has been largely driven by a suite of comprehensive benchmarks designed to assess their knowledge and reasoning abilities across a multitude of domains (Cobbe et al., 2021; Liu et al., 2024b; Rein et al., 2024; Clark et al., 2018; Hendrycks et al., 2021). Prominent examples such as MMLU (Hendrycks et al., 2020), BIG-Bench (Srivastava et al., 2023) and HellaSwag (Zellers et al., 2019) have become standard tools for measuring model performance (Contributors, 2023; Duan et al., 2024). A common characteristic of these benchmarks is their reliance on static, single-turn formats, typically involving multiple-choice questions or direct answer generation (Wang et al., 2024b; Suzgun et al., 2022; Yue et al., 2024; Lu et al., 2022). However, static benchmarks are vulnerable to data leakage, since portions of their test data may inadvertently be included in the massive datasets used for model pre-training (Balloccu et al., 2024; Xu et al., 2024; Ni et al., 2025; Li et al., 2023). One way to mitigate cheating is through dynamically updated benchmarks, such as LatestEval (Li et al., 2024) and LiveBench (White et al., 2024), which continuously incorporate the latest real-world data. While this approach alleviates the issue to some extent, it still fails to enable fine-grained analysis and precise evaluation of the internal knowledge a model has acquired. To address this, we propose a dynamic, multi-turn framework designed to preclude rote memorization and compel the model to externalize its deductive reasoning.

2.2. Puzzles and Games as Evaluation Paradigms

Puzzles, broadly defined as structured problems that require reasoning to reach a solution, have long served as effective probes of intelligence (Guo et al., 2024; Estermann et al., 2024; Shah et al., 2024). They constrain the search space while still demanding creativity, logical inference, and the integration of disparate knowledge. Accordingly, puzzles and games have been widely adopted as formal evaluation paradigms, from classic board games like Chess and Go (Silver et al., 2017) to text-based adventure games (Wang et al., 2024a; Gallotta et al., 2024; Zhang & Long, 2025) and interactive formats such as 20 Questions (Vinyals et al., 2019; Liang et al., 2023; Hu et al., 2018). A common feature of these settings is sustained, multi-turn interaction within a structured environment, which makes them well-suited for assessing strategic planning, logical deduction, and efficient use of information (Chen et al., 2024; Li et al., 2025; Chen et al., 2025). Motivated by this tradition, we frame evaluation as a multi-turn process that requires a model to externalize its reasoning trajectory rather than merely produce final answers. The Active Probing framework repurposes the puzzle format to evaluate how a model retrieves, integrates,

and reasons over its internalized knowledge to identify a target concept it is expected to know, thereby prioritizing the knowledge structure and logical coherence of the reasoning trajectory over simple binary success metrics.

3. Method

ACTIVE PROBING BENCHMARK is founded on a fundamental generalization of the puzzle definition, which enables the rapid construction of custom reasoning benchmarks. Section 3.1 details this generalized puzzle setting, along with our evaluation settings and metrics. Section 3.2 then describes the construction details in the APB.

3.1. Multi-turn Active Probing Process

Our evaluation method is a multi-turn interactive protocol. Its extensibility comes from a general definition of a puzzle as target concept identification through active evidence acquisition.

Generalized Puzzle Definition. Each instance specifies a hidden target concept P_{target} (e.g., a specialized Wikipedia title or a technical command) and an associated reference document I_{ref} that is accessible only to the oracle. The model must identify P_{target} by asking questions and integrating turn-level feedback rather than by reading I_{ref} directly.

Problem Formulation. Let V denote the concept space and initialize the candidate set as $S_0 = V$. The model interacts for at most $T = \text{MAX_ROUNDS}$ turns. At each turn t , the model produces a multiple-choice query $Q_t = (q_t, O_t)$ and options $Q_t = (q_t, O_t)$ including question q_t and options O_t , and receives feedback A_t . The interaction history is $H_t = \{(Q_1, A_1), \dots, (Q_t, A_t)\}$.

Interaction Mechanism. Algorithm 1 summarizes the interaction loop. At each turn t , the Querying Agent conditions on the accumulated history H_{t-1} and outputs a multiple-choice query $Q_t = (q_t, O_t)$. If $P_{\text{target}} \in Q_t$, the episode terminates with SUCCESS; otherwise, the Knowledge Oracle consults the reference I_{ref} and returns turn-level validation feedback $A_t = \mathcal{V}(Q_t, I_{\text{ref}})$ without exposing the reference text. The pair (Q_t, A_t) is appended to form H_t for the next turn, and the process continues until success or the budget T is exhausted.

3.2. The Active Probing Benchmark

To assess the specialized knowledge that models internalize from core training corpora like Wikipedia, we constructed the **Active Probing Benchmark**, featuring 960 puzzles derived from Wikipedia across eight distinct domains.

Data Construction. The benchmark targets were curated via a rigorous, multi stage filtering pipeline (Figure 3) that

Algorithm 1 APB multi-turn active probing process

Input: Round budget T , puzzle target P_{target} , reference I_{ref}

```

1:  $H_0 \leftarrow \emptyset$ 
2: for  $t = 1$  to  $T$  do
3:    $Q_t = (q_t, O_t) \leftarrow \pi_\theta(H_{t-1})$   $\triangleright$  Querying Agent: input  $H_{t-1}$ , output  $Q_t$ 
4:   if  $P_{\text{target}} \in Q_t$  then
5:     return SUCCESS
6:   end if
7:    $A_t \leftarrow \mathcal{V}(Q_t, I_{\text{ref}})$   $\triangleright$  Knowledge Oracle: option validation feedback
8:    $H_t \leftarrow H_{t-1} \cup \{(Q_t, A_t)\}$ 
9: end for
10: return FAILURE

```

distills highly specialized concepts from a dump of 6.4M English Wikipedia titles. The pipeline acts as a coarse to fine funnel: it first removes broad classes of common entities and then progressively concentrates on concepts that require advanced domain expertise. The selection process consists of the following four key stages:

- **Stage 1 (Broad filter):** We rapidly removed common entities, pop culture references, and general knowledge terms, yielding an initial pool of candidates.
- **Stage 2 (Expertise filter):** We applied prompt based criteria that retain only titles judged to require doctorate level specialized knowledge claimed in prompts, discarding undergraduate or introductory concepts.
- **Stage 3 (Domain stratification):** The remaining candidates were categorized into eight academic domains to ensure balanced coverage (e.g., biology, physics, humanities, and health sciences).
- **Stage 4 (Difficulty selection):** Within each domain, we ran a dynamic tournament style selection: from batches of roughly 150 specialized candidates, we retained only the 5–10 most challenging items according to model based difficulty scoring. This step concentrates the final benchmark on the extreme tail of the difficulty distribution.

Across stages, the pipeline retains approximately 0.14% of the original title set, producing a pool of roughly 9k advanced candidates and randomly selecting 120 targets per domain (960 puzzles total). The filtering code and prompts are released for reproducibility.¹ For each puzzle, the cleaned text from the corresponding Wikipedia article serves as the reference document I_{ref} .

Difficulty validation. To provide quantitative evidence that the selected targets correspond to specialized, long tail knowledge, we analyze Wikipedia’s link graph structure. For entities at different pipeline stages, we compute their in-link count (backlink count), which serves as a proxy for graph centrality (and correlates with common-

¹https://anonymous.4open.science/r/wikipedia_filter-7870

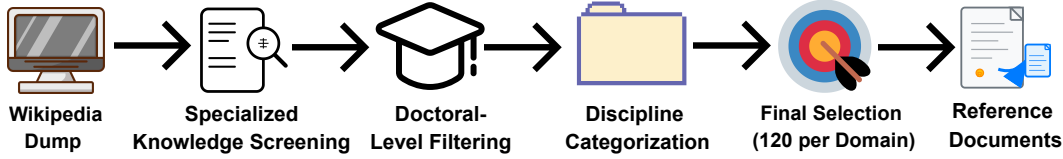


Figure 3. The multi-stage, coarse-to-fine data filtering pipeline used to construct the Active Probing Benchmark from the English Wikipedia dump.

Table 1. Topological Analysis of Concept Difficulty. In-link counts drop drastically, confirming the selection of “edge” knowledge.

Group	<i>N</i>	Mean	Med.	Type
G1: Rejected (Stage 1)	1k	379.2	500 ⁺	Common
G2: Rejected (Stage 4)	1k	86.7	19.0	Specialized
G3: Final (APB)	1k	59.3	16.0	Deep/Edge

ness). As shown in Table 1, entities rejected in Stage 1 exhibit saturated in-link counts (median > 500 , capped by the Wikipedia API), consistent with central and widely referenced concepts. In contrast, the final APB targets have much smaller in-link counts (median 16), indicating that they are substantially more peripheral in the knowledge graph and thus more specialized than the rejected groups.

3.3. The Active Probing Benchmark

Benchmark Composition. The Active Probing Benchmark is composed of 960 puzzles, which are systematically categorized into eight distinct domains: Biology, Human Disease, Chemistry, Physics, Mathematics, Computer Science, Economics, and Philosophy. These domains were chosen to ensure comprehensive coverage of major academic fields, including the humanities and social sciences, the natural and engineering sciences, and health.

Each domain comprises 120 puzzles distributed across 7 to 8 primary sub-disciplines (Figure 4) to ensure diverse topical coverage. For example, the Physics domain sources items from eight distinct subfields, with key contributions from Quantum Field Theory and Mathematical Physics. We also provide additional granularity within each subfield. For instance, the “Specialized Topics” sub-discipline covers Plasma Physics, Atomic, Molecular, and Optical (AMO) Physics, and Particle and Nuclear Physics. This multi-topic setup is a fundamental aspect of our benchmark across all domains, supporting a comprehensive evaluation of expert knowledge. Full details are described in the Appendix D.

4. Experiments and Analyses

4.1. Setup

This section outlines the key parameter configurations and the set of models evaluated in our experiments.

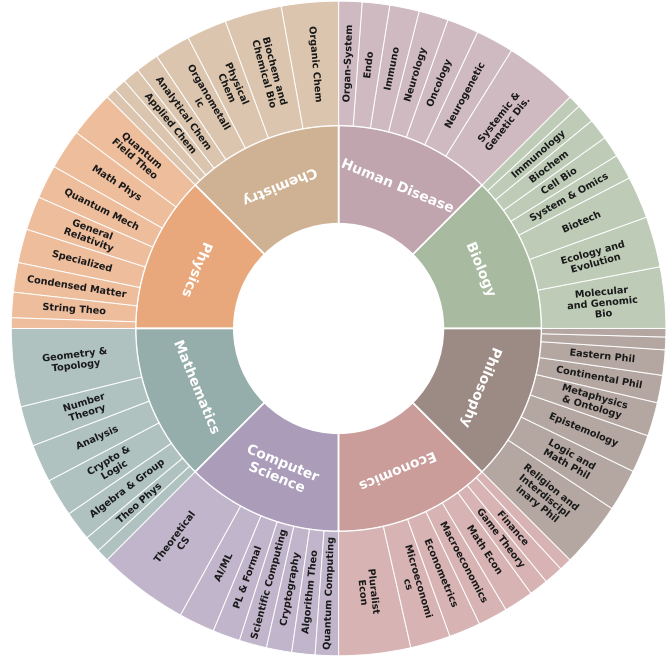


Figure 4. Hierarchical distribution of puzzle targets across domains and sub-disciplines. Eight distinct color families represent the primary domains, while individual segments denote specific sub-topics. The area of each segment is proportional to the number of questions in that category. Exact distributions are provided in Appendix D.

Maximum Rounds. We set MAX_ROUNDS to 15. This decision is justified by empirical data (Table 3), which shows that a successful session concludes in 9.78 turns on average (9.15 for the top-performing GPT-5). A limit of 15 thus provides sufficient opportunity for models to converge without being cost-prohibitive. Simultaneously, this cap manages the escalating token costs associated with including the full conversation history in each turn. While increasing the limit to 20 would have nearly doubled the evaluation cost.

Temperature. We standardized the temperature at 0.2 to ensure reproducibility. Our ablation study (Appendix B.2) confirms the robustness of this choice, demonstrating that alternative settings (e.g., higher temperatures for reasoning models) yield negligible gains and do not alter the performance hierarchy. GPT-5 used its default “medium” effort,

Table 2. Success Rate (%) for various models across eight domains. The best performance in each domain is highlighted in **bold**. The bottom row, ‘Domain Avg.’, summarizes the average performance for each domain.

Model	Bio	Dis	Chem	Phys	Math	CS	Econ	Phil	Avg.
GPT-5	34.2	37.5	33.3	34.2	34.2	25.0	39.2	30.0	33.4
Claude Opus 4	27.5	40.8	29.2	26.7	18.3	19.2	31.7	27.5	27.6
Claude Sonnet 4	29.2	35.0	23.3	23.3	17.5	18.3	42.5	21.7	26.4
o3	28.3	33.3	21.7	29.2	20.0	10.0	35.0	25.0	25.3
GPT-4o	26.7	24.2	20.0	17.5	11.7	10.8	25.0	20.8	19.6
Grok-4	14.2	30.8	16.7	13.3	10.0	13.3	20.8	21.7	17.6
DeepSeek V3	21.7	22.5	13.3	15.0	11.7	11.7	18.3	16.7	16.4
GPT-OSS-120B	19.2	21.7	12.5	10.8	5.8	8.3	20.8	8.3	13.4
Gemini 2.5 Pro No-Thinking	9.2	28.3	12.5	6.7	5.8	7.5	16.7	17.5	13.0
Qwen3-235B-A22B-Instruct	5.8	27.5	5.0	11.7	5.8	10.0	15.0	20.0	12.6
Gemini 2.5 Pro	7.5	26.7	11.7	13.3	1.7	8.3	17.5	13.3	12.5
DeepSeek R1	5.8	21.7	5.0	1.7	5.0	3.3	10.0	5.0	7.2
Qwen3-235B-A22B-Thinking	7.5	11.7	5.8	3.3	1.7	5.0	9.2	8.3	6.6
Domain Avg.	18.2	27.8	16.2	15.9	11.5	11.6	23.2	18.1	17.8

verified as optimal in our analysis.”

Knowledge Oracle. We designated DeepSeek V3 as the Knowledge Oracle responsible for providing correctness feedback. This model was chosen for its strong instruction-following capabilities and consistent outputs at a low temperature. To ensure our main findings are not sensitive to this specific choice, we conducted a detailed ablation study (Appendix B.1). This includes both a subset analysis and a comprehensive cross-validation using Claude Sonnet 4 on the full 960-puzzle benchmark, confirming that the relative performance hierarchy of models remains stable ($\rho = 0.96$) despite absolute scoring shifts.

Evaluated Models. Our evaluation on the APB covers a comprehensive suite of 13 leading closed-source and open-source LLMs to assess their capabilities. We list all the evaluated models below:

- **Closed-Source models:** GPT-5, o3 and GPT-4o (Hurst et al., 2024), Claude Opus 4 and Claude Sonnet 4, Gemini 2.5 Pro and Gemini 2.5 Pro No-Thinking (Comanici et al., 2025), Grok-4.
- **Open-Source models:** Deepseek V3 (Liu et al., 2024a) and Deepseek R1 (Guo et al., 2025), GPT-OSS-120B (Agarwal et al., 2025), Qwen3-235B-A22B-Instruct and Qwen3-235B-A22B-Thinking (Yang et al., 2025).

4.2. Main Results

The main evaluation results are presented in Table 2, covering detailed success rate assessments across 8 domains: Biology (Bio), Human Disease (Dis), Chemistry (Chem),

Physics (Phys), Mathematics (Math), Computer Science (CS), Economics (Econ), and Philosophy (Phil).

Analysis of Main Results. The evaluation results demonstrate pronounced disparities across 8 domains (Table 2). In the strongest-performing domains, Human Disease (27.8%) and Economics (23.2%) achieved the highest average success rates. The high accuracy in Human Disease can be attributed to its comprehensive and well-structured classification system, while the strong performance in Economics may stem from the relatively lower depth and complexity of its specialized knowledge. In the most challenging domains, Mathematics (11.5%) and Computer Science (11.6%) recorded the lowest success rates, where even GPT-5, the strongest model, attained only 34.2% and 25.0% respectively. Intermediate performance was observed in Biology, Chemistry, Physics, and Philosophy.

From a model-centric perspective, the results reveal a clear performance hierarchy. While GPT-5 leads the pack, its predecessor GPT-4o (19.6%) demonstrates remarkable longevity, effectively separating the top-performing models and the rest. When compared to other closed-source offerings, the relatively modest results from the Gemini series point to specific weaknesses for this reasoning task. Furthermore, the struggles of DeepSeek R1 (7.2%) and Qwen3-235B-A22B-Thinking (6.6%) underscore the current limitations and vulnerabilities within the open-source reasoning model landscape.

4.3. Detailed Analyses

This section provides a detailed analysis of model performance based on Table 3, focusing on the behavioral patterns observed during the knowledge reasoning process.

Model	PL	IG	IO	IS	SR	TR
GPT-5	300	0.32	6.16	3.61	9.15	13.02
Claude Opus 4	494	0.48	6.41	4.11	10.23	13.60
Claude Sonnet 4	565	0.44	5.13	3.14	9.81	13.54
o3	357	0.52	6.32	3.59	9.66	13.67
GPT-4o	473	0.81	6.57	4.34	10.68	14.06
Grok-4	1930	1.15	4.87	3.08	9.72	14.04
DeepSeek V3	551	2.29	5.31	2.68	9.96	14.07
GPT-OSS-120B	409	1.21	6.13	3.64	9.53	14.23
Gemini 2.5 Pro No-Thinking	908	2.12	5.51	3.82	10.48	14.37
Qwen3-235B-A22B-Instruct	1396	1.43	7.15	4.75	9.20	14.21
Gemini 2.5 Pro	906	2.20	5.28	3.70	10.31	14.37
DeepSeek R1	574	5.13	3.63	2.22	10.20	14.66
Qwen3-235B-A22B-Thinking	1495	2.98	4.15	2.91	8.21	14.48
Avg.	797	1.62	5.59	3.51	9.78	14.02

Table 3. Detailed behavioral analysis metrics. **PL**: Probing Length, **IG**: Incorrect Guess rounds, **IO**: Incorrect Option rounds, **IS**: Incorrect Streak rounds, **SR**: Success Rounds, **TR**: Total Rounds. Blue bars visualize the normalized magnitude of each metric within its column to highlight behavioral patterns.

Probing Length. The Probing Length (PL) measures the average character length of the agent’s output per turn. This output typically includes not only the multiple-choice question but also the model’s explicit reasoning based on prior rounds. Our analysis reveals that increased verbosity does not correlate with better performance on this task. Indeed, efficiency in knowledge probing rounds appears to be a hallmark of the top models like GPT-5 (PL = 300). This trend generally holds for the other top-performing models with relatively concise probing lengths. Furthermore, we observe that PL is often strongly correlated with a model’s developer family, suggesting that different series of models possess inherent and consistent styles for expressing their reasoning process in this interactive setting. For instance, both Gemini models exhibit a similarly high verbosity (PL $\approx$ 907), as do the DeepSeek models (PL $\approx$ 560) and the OpenAI models ($299 < \text{PL} < 500$). This suggests that different series of models possess inherent and consistent styles for expressing their probing process in this interactive setting.

Incorrect Guess Rounds. The Incorrect Guess Rounds (IG) metric quantifies the average number of times an agent

makes a conclusive but incorrect guess of the target concept per session. We interpret a high IG as an indicator of model overconfidence, where the agent prematurely commits to a wrong answer despite having insufficient information. Our results show a strong negative correlation between IG and Success Rate, suggesting that early incorrect commitments are associated with persistent mislocalization in the knowledge space, which limits the usefulness of subsequent probing. For instance, the top-performing model GPT-5 has the lowest IG (0.32), while the struggling DeepSeek R1 has the highest (5.13). From a strategic standpoint, such actions represent a low-utility approach by invalidating only a single hypothesis. In a context like Human Disease, a propensity for making frequent, incorrect assertions is a harmful failure mode that could lead to detrimental outcomes, underscoring the importance of penalizing such overconfident behavior in model evaluation.

Incorrect Probing and Error Correction. We analyze two metrics related to probing errors: Incorrect Option rounds (IO) and Incorrect Streak rounds (IS). IO measures the average number of rounds where all four proposed options are incorrect, suggesting inaccurate concept grounding during knowledge localization. IS measures the longest consecutive sequence of such rounds, reflecting the model’s ability to recover from a flawed reasoning path. However, these metrics are not simple indicators of poor performance. An ‘all-incorrect’ round can be highly informative for a capable agent, as it allows for significant pruning of the search space and a major strategy correction. This can paradoxically help a strong model quickly locate an obscure concept. Thus, these metrics offer insights into a model’s strategic behavior. The case of DeepSeek R1 is particularly illustrative: despite having the lowest IO (3.63) and IS (2.22), it achieves one of the lowest success rates. This suggests that its probing strategy is ineffective at leveraging information and asking questions to find the P_{target} , a strategic failure that helps explain its high rate of premature guesses, as reflected in its top-ranking IG.

Overall Efficiency. We evaluate efficiency using two complementary metrics: Average Success Rounds (SR) and Average Total Rounds (TR). SR measures the average number of rounds in successful sessions only, reflecting speed on solvable problems. TR provides a more holistic score by including failed sessions, which are counted as MAX_ROUNDS (15), thus blending success rate with efficiency. Relying solely on SR can be misleading for models with low success rates, as it only reflects performance on easier tasks they could solve. TR, therefore, offers a more comprehensive view. For example, GPT-5’s top performance gives it the best TR (13.02) and leading SR (9.15), confirming its position as the most effective model overall. This combined metric also reveals nuances in reasoning efficiency. For instance, despite Claude Opus 4 having a higher Suc-

cess Rate than Claude Sonnet 4, Sonnet’s lower TR (13.54 vs. 13.60) and SR (9.81 vs 10.23) suggest it has a relative advantage in the efficiency of its reasoning process.

4.4. Error Analysis

Open-Source Reasoning Models. To understand the relatively weak performance of open source reasoning models, we analyzed the publicly available chain-of-thought traces for GPT-OSS-120B, DeepSeek R1, and Qwen3-235B-A22B-Thinking. We observe a recurring pattern that is best characterized as a failure of knowledge localization during active probing. Rather than using early turns to acquire broadly discriminative evidence, these models often anchor on several concrete candidate targets for P_{target} at the beginning of an episode (Figure 5). Subsequent multiple-choice queries are then constructed to seek confirmation for these candidates, instead of eliciting information that would re-partition the broader candidate space.

This premature anchoring has two downstream effects. First, when the true P_{target} is not among the initial hypotheses, the model’s questions tend to target attributes that are irrelevant to the correct region of the domain knowledge space, yielding feedback that is weakly informative for correction. As a result, the interaction trace exhibits repetitive cycles of querying around the wrong neighborhood, with many low-utility terms and little reduction of the remaining uncertainty. Second, the same anchoring increases the likelihood of conclusive but unsupported commitments, producing more frequent incorrect final guesses. This is consistent with the comparatively high IG values observed for these models and with their lower success rates overall.

These traces suggest that the performance gap is driven less by a lack of procedural reasoning steps and more by difficulty in using multi-turn feedback to maintain and update a well-grounded hypothesis set. In other words, the models often fail to treat early turns as evidence acquisition for concept identification, and instead collapse prematurely onto specific candidates, which limits their ability to recover when the initial localization is wrong.

Gemini Series Models. The Gemini 2.5 Pro models exhibit a distinct behavioral pattern that helps explain their relatively weak performance. A primary characteristic is their high Probing Length (PL $\approx$ 907), far exceeding other top closed-source models. We observe that this verbosity stems from a tendency to repeatedly summarize and elaborate on prior turns before issuing the next query. From a knowledge probing perspective, this behavior shifts effort from acquiring new discriminative evidence to reprocessing self-generated context, which can dilute the influence of objective turn-level feedback. As a result, the model may persist in a mislocalized hypothesis set and continue probing within an irrelevant region of the knowledge space, culmi-

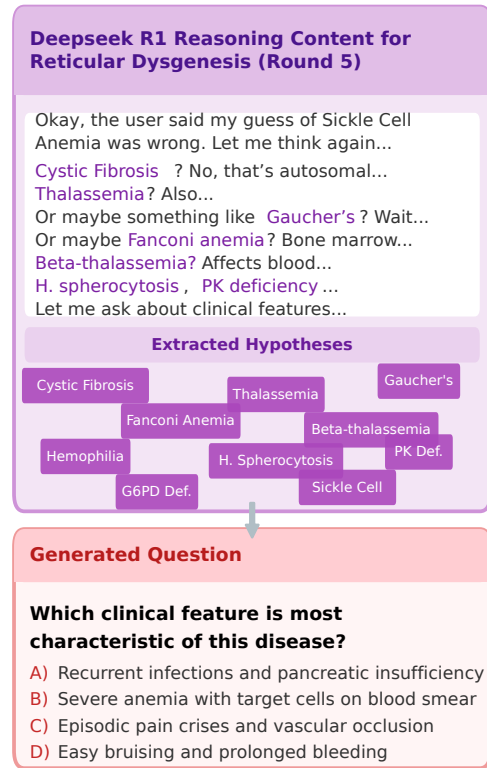


Figure 5. DeepSeek R1 reasoning trace for the target ‘Reticular dysgenesis’ (Round 5). **Top:** The internal reasoning content shows the model brainstorming specific disease candidates following an incorrect guess. **Middle:** The extracted hypotheses visualize the diverse set of conditions considered during this step. **Bottom:** The generated follow-up question focuses on differentiating clinical features.

nating in more frequent unsupported final commitments, reflected by their notably high IG (2.1–2.2) relative to other leading closed-source models.

5. Conclusion

We introduced APB, an interactive benchmark that extends puzzle evaluation from narrow game-like settings to knowledge reasoning over specialized concepts. By casting concept identification as multi-turn information seeking with verifiable endpoints, APB provides process-level evidence of how models acquire and use domain knowledge. Experiments show that APB is challenging even for strong systems (GPT-5 reaches 33.4% success), and that several open source reasoning models lag behind their no-thinking counterparts, consistent with difficulties in grounding multi-turn inference in specialized knowledge. Beyond these results, APB offers a practical construction pipeline for producing challenging diagnostic evaluations in additional specialized domains, helping broaden the empirical study of knowledge reasoning under interaction.

Impact Statement

The primary objective of this work is to evaluate and better understand the limitations of interactive knowledge reasoning in large language models, contributing to more trustworthy and informative model assessment. APB is intended for research use and reports objective success criteria together with qualitative failure analyses to reduce over-interpretation of single metrics. This study was conducted independently and adheres to standard ethical research principles, including considerations of bias, fairness, privacy, security, and legal compliance.

References

- Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*, 2025.
- Badola, K., Simon, J., Hosseini, A., Carthy, S. M. M., Munkhdalai, T., Goyal, A., Kočíšký, T., Upadhyay, S., Fatemi, B., and Kazemi, M. Multi-turn puzzles: Evaluating interactive reasoning and strategic dialogue in llms. *arXiv preprint arXiv:2508.10142*, 2025.
- Balloccu, S., Schmidová, P., Lango, M., and Dušek, O. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*, 2024.
- Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., Starace, G., Liu, K., Maksin, L., Patwardhan, T., et al. Mle-bench: Evaluating machine learning agents on machine learning engineering. *arXiv preprint arXiv:2410.07095*, 2024.
- Chen, G., Xu, W., Zhang, H., Chan, H. P., Liu, C., Bing, L., Zhao, D., Luu, A. T., and Rong, Y. Finereason: Evaluating and improving llms’ deliberate reasoning through reflective puzzle solving. *arXiv preprint arXiv:2502.20238*, 2025.
- Chen, Q., Zhang, B., Wang, G., and Wu, Q. Weak-eval-strong: Evaluating and eliciting lateral thinking of llms with situation puzzles. *Advances in Neural Information Processing Systems*, 37:79642–79665, 2024.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J. E., et al. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning*, 2024.
- Chollet, F. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- Comanici, G., Bieber, E., Schaeckermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blstein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.
- Contributors, O. Opencompass: A universal evaluation platform for foundation models, 2023.
- Duan, H., Yang, J., Qiao, Y., Fang, X., Chen, L., Liu, Y., Dong, X., Zang, Y., Zhang, P., Wang, J., et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.
- Estermann, B., Lanzendörfer, L., Niedermayr, Y., and Wattenhofer, R. Puzzles: A benchmark for neural algorithmic reasoning. *Advances in Neural Information Processing Systems*, 37:127059–127098, 2024.
- Gallotta, R., Todd, G., Zammit, M., Earle, S., Liapis, A., Toggelius, J., and Yannakakis, G. N. Large language models and games: A survey and roadmap. *IEEE Transactions on Games*, 2024.
- Giadikiaroglou, P., Lymperaious, M., Filandrianos, G., and Stamou, G. Puzzle solving using reasoning of large language models: A survey. *arXiv preprint arXiv:2402.11291*, 2024.
- Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.
- Guo, K., Nan, B., Zhou, Y., Guo, T., Guo, Z., Surve, M., Liang, Z., Chawla, N., Wiest, O., and Zhang, X. Can llms solve molecule puzzles? a multimodal benchmark for molecular structure elucidation. *Advances in Neural Information Processing Systems*, 37:134721–134746, 2024.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

- Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.
- Hu, H., Wu, X., Luo, B., Tao, C., Xu, C., Wu, W., and Chen, Z. Playing 20 question game with policy-based reinforcement learning. *arXiv preprint arXiv:1808.07645*, 2018.
- Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.
- Jacovi, A., Caciularu, A., Goldman, O., and Goldberg, Y. Stop uploading test data in plain text: Practical strategies for mitigating data contamination by evaluation benchmarks. *arXiv preprint arXiv:2305.10160*, 2023.
- Kiela, D., Bartolo, M., Nie, Y., Kaushik, D., Geiger, A., Wu, Z., Vidgen, B., Prasad, G., Singh, A., Ringshia, P., et al. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.
- Li, K., Chen, X., Song, T., Zhou, C., Liu, Z., Zhang, Z., Guo, J., and Shan, Q. Solving situation puzzles with large language model and external reformulation. *arXiv preprint arXiv:2503.18394*, 2025.
- Li, Y., Guerin, F., and Lin, C. An open source data contamination report for large language models. *arXiv preprint arXiv:2310.17589*, 2023.
- Li, Y., Guerin, F., and Lin, C. Latesteval: Addressing data contamination in language model evaluation through dynamic and time-sensitive test construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18600–18607, 2024.
- Liang, T., He, Z., Huang, J.-t., Wang, W., Jiao, W., Wang, R., Yang, Y., Tu, Z., Shi, S., and Wang, X. Leveraging word guessing games to assess the intelligence of large language models. *arXiv preprint arXiv:2310.20499*, 2023.
- Lin, B. Y., Wu, Z., Yang, Y., Lee, D.-H., and Ren, X. Riddlesense: Reasoning about riddle questions featuring linguistic creativity and commonsense knowledge. *arXiv preprint arXiv:2101.00376*, 2021.
- Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.
- Liu, H., Zheng, Z., Qiao, Y., Duan, H., Fei, Z., Zhou, F., Zhang, W., Zhang, S., Lin, D., and Chen, K. Mathbench: Evaluating the theory and application proficiency of llms with a hierarchical mathematics benchmark. *arXiv preprint arXiv:2405.12209*, 2024b.
- Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022.
- Ni, S., Kong, X., Li, C., Hu, X., Xu, R., Zhu, J., and Yang, M. Training on the benchmark is not all you need. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 24948–24956, 2025.
- Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.
- Roberts, M., Thakur, H., Herlihy, C., White, C., and Dooley, S. To the cutoff... and beyond? a longitudinal perspective on llm data contamination. In *The Twelfth International Conference on Learning Representations*, 2023.
- Shah, K., Dikkala, N., Wang, X., and Panigrahy, R. Causal language modeling can elicit search and reasoning capabilities on logic puzzles. *Advances in Neural Information Processing Systems*, 37:56674–56702, 2024.
- Shirali, A., Abebe, R., and Hardt, M. A theory of dynamic benchmarks. *arXiv preprint arXiv:2210.03165*, 2022.
- Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.
- Singh, A. K., Kocyigit, M. Y., Poulton, A., Esiobu, D., Lomeli, M., Szilvasy, G., and Hupkes, D. Evaluation data contamination in llms: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.
- Suzgun, M., Scales, N., Schärli, N., Gehrmann, S., Tay, Y., Chung, H. W., Chowdhery, A., Le, Q. V., Chi, E. H.,

- Zhou, D., , and Wei, J. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.
- Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. Alphastar: Mastering the real-time strategy game starcraft ii. *DeepMind blog*, 2:20, 2019.
- Wang, R., Todd, G., Xiao, Z., Yuan, X., Côté, M.-A., Clark, P., and Jansen, P. Can language models serve as text-based world simulators? *arXiv preprint arXiv:2406.06485*, 2024a.
- Wang, Y., Ma, X., Zhang, G., Ni, Y., Chandra, A., Guo, S., Ren, W., Arulraj, A., He, X., Jiang, Z., et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024b.
- White, C., Dooley, S., Roberts, M., Pal, A., Feuer, B., Jain, S., Shwartz-Ziv, R., Jain, N., Saifullah, K., Dey, S., et al. Livebench: A challenging, contamination-limited llm benchmark. *arXiv preprint arXiv:2406.19314*, 2024.
- Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*, 2024.
- Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.
- Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.
- Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- Zhang, J. and Long, Y. Learning to play like humans: A framework for llm adaptation in interactive fiction games. *arXiv preprint arXiv:2505.12439*, 2025.
- Zhang, Z., Chen, Z., Zhang, Z., Sun, Y., Tian, Y., Jia, Z., Li, C., Liu, X., Min, X., and Zhai, G. Puzzlebench: A fully dynamic evaluation framework for large multimodal models on puzzle solving. *arXiv preprint arXiv:2504.10885*, 2025.
- Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., et al. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*, 2023.
- Zhou, X., Weyssow, M., Widyasari, R., Zhang, T., He, J., Lyu, Y., Chang, J., Zhang, B., Huang, D., and Lo, D. Lessleak-bench: A first investigation of data leakage in llms across 83 software engineering benchmarks. *arXiv preprint arXiv:2502.06215*, 2025.

Appendix

A. Rapid Custom Benchmark Construction

Our framework provides a scalable and generalizable method for rapidly constructing diagnostic benchmarks in any specialized domain. The process is straightforward and consists of two main stages:

- **Target Identification and Curation:** First, a domain of interest and a corresponding source corpus of knowledge are identified (e.g., technical manuals, API documentation, academic articles). From this corpus, a list of key concepts, terms, commands, or entities is systematically extracted to serve as the puzzle targets (P_{target}).
- **Reference Document Generation:** For each puzzle target, a "golden" reference document (I_{ref}) is established (e.g., the specific documentation page for a command). This document serves as the ground truth for the environment's feedback during the multi-turn evaluation.

A.1. Case Study: The IC Back-End Puzzle Benchmark

To demonstrate the efficacy of this method, we applied it to the niche, industrial domain of IC back-end design. We used the official documentation for four widely-used EDA command sets—Synopsys ICC, ICC2, and Cadence Innovus (both common and legacy)—as our source corpus.

Table 4. Success rates (%) on the IC Back-End Puzzle Benchmark. Tools are grouped by vendor: Synopsys (ICC, ICC2) and Cadence Innovus (Common, Legacy). "NT" denotes "No-Thinking". **Qwen3 models refer to the Qwen3-235B-A22B variants.** Best results in each category are **bolded**.

Model	Synopsys		Cadence Innovus		Avg.
	ICC	ICC2	Common	Legacy	
Claude Opus 4	27.5	14.2	16.7	22.5	20.2
GPT-5	25.8	15.0	14.2	20.0	18.8
grok-4	15.0	10.8	18.3	24.2	17.1
o3	20.0	15.0	12.5	12.5	15.0
Claude Sonnet 4	14.2	10.8	9.2	18.3	13.1
Gemini 2.5 Pro	10.8	8.3	5.0	16.7	10.2
Gemini 2.5 Pro NT	10.8	8.3	6.7	12.5	9.6
GPT-4o	6.7	7.5	7.5	7.5	7.3
DeepSeek V3	10.8	4.2	6.7	5.8	6.9
GPT-OSS-120B	8.3	7.5	4.2	4.2	6.0
Qwen3 Instruct	4.2	5.8	3.3	4.2	4.4
Qwen3 Thinking	6.7	2.5	3.3	5.0	4.4
DeepSeek R1	4.2	0.8	2.5	2.5	2.5
Domain Avg.	12.7	8.5	8.5	12.0	10.4

Following our process, we curated 480 puzzles by sampling 120 command names from each of the four sets. This

benchmark exemplifies how our framework can be used to quickly create robust, expert-level evaluations that test practical knowledge in any specialized domain.

We evaluated 13 leading LLMs on this benchmark, with the results presented in Table 4. The findings underscore the benchmark's significant challenge, with an average success rate of only 10.4% across all models. A clear performance hierarchy emerged, with Claude Opus 4 (20.2%), GPT-5 (18.8%), and Grok-4 (17.1%) forming a distinct top tier. Notably, reasoning-enhanced models like DeepSeek R1 (2.5%) and Qwen3 235B Thinking (4.4%) performed very poorly, reinforcing our main paper's finding of a gap between procedural reasoning and applied domain knowledge. The domain-specific results also reveal non-uniform knowledge; for instance, Grok-4 excels on Innovus commands but is weaker on ICC, suggesting that model knowledge of these specialized tools is often fragmented. This case study validates our method's ability to create challenging benchmarks that effectively differentiate model capabilities in highly specialized, real-world domains.

A.2. Broader Applications and Future Directions

The methodology detailed in this section is designed for broad applicability beyond the technical domains demonstrated. We encourage the research community to leverage our scalable process to construct new benchmarks across a wide array of fields. For example, puzzle targets could be drawn from the humanities (e.g., famous figures, artworks, film titles) or the natural sciences (e.g., celestial bodies, materials, biological species).

The creation of such diverse benchmarks would facilitate a more holistic evaluation of an LLM's all-around knowledge reasoning abilities. Moreover, the resulting multi-turn interaction data would provide a rich resource for deeper, cross-domain pattern recognition, enabling a more comprehensive exploration of model reasoning strategies and behaviors.

B. Ablation Study

To validate the robustness of our evaluation framework, we conducted a comprehensive ablation study identifying potential sources of variance in our experimental pipeline. Our objective is to confirm that the observed model performance tiers are intrinsic to the agents' capabilities rather than artifacts of the evaluation configuration. Specifically, we address two critical dimensions: (1) the reliability of the **Knowledge Oracle** (Judge Consistency), and (2) the sensitivity of performance to **hyperparameters** and inference-time compute (Temperature and Reasoning Effort).

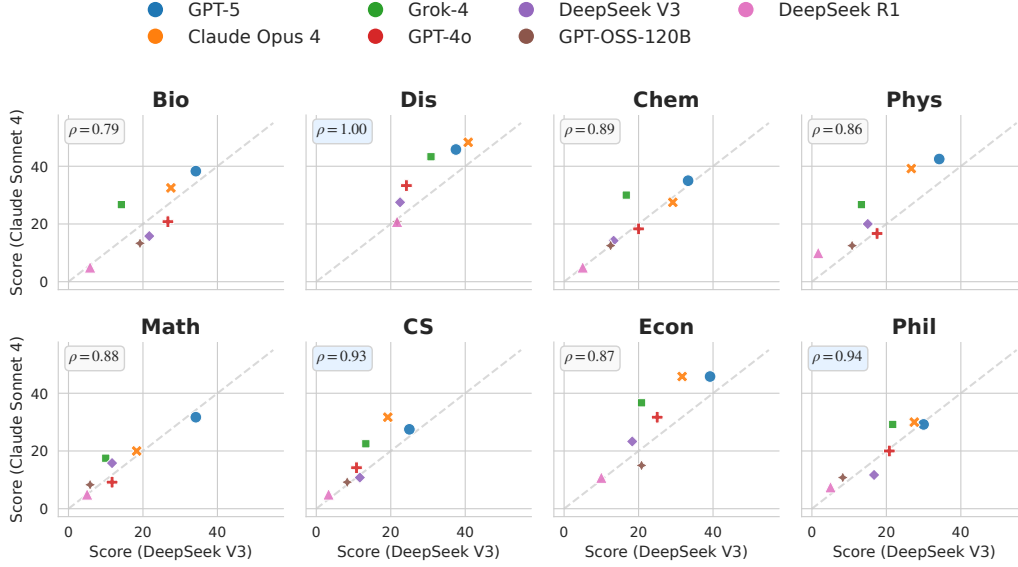


Figure 6. Ranking consistency analysis (DeepSeek V3 vs. Claude Sonnet 4) across eight domains. Each point represents a model’s performance in a specific domain. The distinct clustering of model colors and the high Spearman’s rank correlation (ρ) within each domain confirm that relative model rankings remain robust despite the absolute score shift caused by the stricter judge.

B.1. Cross-Validation with Claude Sonnet 4

We employed Claude Sonnet 4 to cross-validate the main results against DeepSeek V3. This evaluation covered the full core benchmark (960 puzzles) across seven representative models, covering the spectrum from top-tier proprietary models to open-source and reasoning-specific models.

Rank Stability Across Different Judges Figure 6 visualizes the domain-level performance correlation between the two judges. While absolute success rates tend to be higher under Claude Sonnet 4 (as indicated by the data points generally lying above the $y = x$ diagonal), the relative performance ordering remains remarkably stable:

- **Visual Consistency:** As shown in the facet plots, models form distinct clusters. Top-performing models (e.g., GPT-5, Claude Opus 4) consistently appear in the upper-right quadrant, while lower-tier models remain in the lower-left across all domains.
- **High Correlation:** The quantitative evidence supports this visual trend. We observe a Spearman rank correlation of $\rho = 0.96$ for the overall model rankings, indicating near-identical leaderboard structures. This consistency remains robust at the granular level, with an average correlation of $\rho = 0.89$ across individual domains (exemplified by $\rho = 1.00$ in Disease and $\rho = 0.94$ in Philosophy).
- **Robust Failure Modes:** “Reasoning-enhanced” models (DeepSeek R1, GPT-OSS-120B) remain at the bot-

tom regardless of the judge. Notably, even with the more capable Claude Sonnet 4 as the judge, DeepSeek R1 achieves only 8.6% accuracy compared to 7.2% under V3. This confirms that the failure of these models stems from their internal strategies (i.e., the concrete-to-abstract fallacy), not judge-induced noise.

Explanation for Spikes and Judge Differences Our case analysis identified two primary factors driving the observed differences between judges: context robustness and semantic validation style.

- **Grok-4’s Outlier Performance (Context Robustness):** Grok-4 exhibits an anomalously high probing length (avg. 1930 characters/turn). DeepSeek V3 struggles to maintain precise instruction-following when parsing such verbose contexts. In contrast, Claude Sonnet 4 effectively filters through Grok-4’s lengthy outputs to extract the core logic, explaining why Grok-4 shows the most dramatic improvement (+11.5%) under the stronger judge.
- **Judging Characteristics (Association vs. Precision):** DeepSeek V3 tends toward *associative validation*—occasionally accepting highly related “attributes” as the target “entity” itself. In contrast, Claude Sonnet 4 enforces stricter semantic precision.

In conclusion, replacing DeepSeek V3 with Claude Sonnet 4 acts primarily as a **linear scaling factor**. It uniformly “lifts” the scores of capable models but does not fundamentally

alter the leaderboard. The high rank correlation confirms that DeepSeek V3 remains a valid, cost-effective proxy for differentiating model tiers.

B.2. Robustness to Sampling Parameters and Reasoning Effort

To ensure that our main findings are not artifacts of sub-optimal hyperparameters, we conducted a targeted ablation study on a representative subset of the benchmark ($N = 120$). We isolate the effects of sampling temperature on open-source reasoning models and the impact of test-time compute scaling on proprietary models.

Temperature Sensitivity in Open-Source Models A potential hypothesis for the underperformance of open-source reasoning models (e.g., DeepSeek R1, Qwen3) is that the default low-temperature settings ($T = 0.2$) might constrain their exploratory capabilities. To test this, we evaluated DeepSeek R1 at its recommended high temperature ($T = 1.0$) and Qwen3 Thinking at a moderate setting ($T = 0.6$).

As detailed in Table 5, the results indicate that temperature is not the bottleneck.

- **DeepSeek R1:** While increasing the temperature to $T = 1.0$ yielded a marginal improvement (6.7% $\rightarrow$ 8.3%), the model remains firmly in the bottom tier.
- **Qwen3:** Increasing the temperature resulted in a slight performance fluctuation (11.7% $\rightarrow$ 10.8%), offering no statistical benefit.

These findings reinforce our conclusion that the failure modes of these models—specifically the concrete-to-abstract fallacy—are structural rather than parameter-dependent.

Table 5. Ablation results on sampling temperature and reasoning effort ($N = 120$). The “Tier” column reflects the relative ranking within the full benchmark context.

Model	Configuration	Success Rate	Tier
<i>Effect of Temperature (Open-Source Reasoning)</i>			
DeepSeek R1	Temp = 0.2 (Default)	6.7%	Bottom
	Temp = 1.0 (Recommended)	8.3%	Bottom
Qwen3 Thinking	Temp = 0.2 (Default)	11.7%	Low
	Temp = 0.6	10.8%	Low
<i>Effect of Reasoning Effort (GPT-5)</i>			
GPT-5	Low Effort	33.3%	Top
	Medium Effort (Default)	38.3%	Top
	High Effort	34.6%	Top

Scaling Laws of Test-Time Compute We further investigated the relationship between reasoning effort and performance using GPT-5. By varying the inference budget across

Low, Medium (Default), and High settings, we observed a distinct **inverted-U trajectory**: performance peaks at Medium effort (38.3%) and declines at High effort (34.6%).

This observation leads to two critical insights:

1. **Diminishing Returns:** In the constrained paradigm of strategic questioning, excessive test-time compute does not linearly translate to better outcomes. “Overthinking” may lead to hallucinated constraints or overly convoluted search paths.
2. **Strategy Quality over Quantity:** Crucially, even at the “Low” effort setting, GPT-5 (33.3%) significantly outperforms the best open-source reasoning models. This suggests that the *quality* of the underlying reasoning strategy—how the model prunes the search space—is a more determinant factor than the raw *quantity* of inference steps or token generation.

C. Token Usage

This section details the token consumption patterns, averaged per puzzle, for both the Querying Agent (QA) and the Knowledge Oracle (KO). It is important to note that a single puzzle evaluation consists of multiple rounds of interaction.

The token cost of the Knowledge Oracle (KO) is characterized by highly efficient outputs but demanding inputs. The KO’s completion tokens are minimal and stable across all tests (averaging 460 tokens), as its responses are restricted to a fixed format providing only correctness judgments. This ensures a fair and consistent evaluation. However, its prompt token requirements are substantial, averaging 18,347 tokens per puzzle, because the KO must process the full reference document (I_{ref}) in each round to provide accurate feedback.

Similarly, the QA prompt costs are significant, averaging 18,826 tokens per puzzle, as the full interaction history is included in the context for each new turn. The vast majority of performance variance and economic cost, however, comes from the SA’s completion tokens, which reflect model verbosity (Table 6). Notably, some of the poorest-performing open-source reasoning models were paradoxically the most computationally expensive. GPT-OSS-120B was the most costly, consuming over 72,000 total QA tokens on average, followed by Qwen3 235B Thinking (approx. 57,600 tokens) and DeepSeek R1 (approx. 43,300 tokens). Collectively, these three open-source models averaged 57,784 QA tokens per puzzle—almost exactly double the 28,705-token average of the remaining models. This contrasts sharply with top-performing models like GPT-5, which achieved the highest success rate while using less than half the tokens of GPT-OSS-120B. This suggests that excessive verbosity is not only an unsuccessful strategy but also an economically inefficient one for this task.

Table 6. A summary of token usage for representative models. "Comp." stands for Completion, "Tok." for Tokens. KO output is low due to fixed formatting.

Model	QA Prompt	QA Comp.	QA Total Tok.	KO Comp. Tok.
<i>(Best Performance)</i>				
GPT-5	9,821	24,301	34,122	430
<i>(Most Verbosity)</i>				
GPT-OSS-120B	14,034	58,471	72,505	473
Qwen3 235B Think.	33,414	24,170	57,584	473
Deepseek R1	15,242	28,021	43,263	458
<i>(Most Concise)</i>				
GPT-4o	13,441	1,513	14,954	466
Average	18,826	16,583	35,416	460

D. Benchmark Composition

Table 7. Biology Domain Composition.

Sub-Discipline	Count (%)
Molecular and Genomic Biology	29 (24.2%)
Genes, Proteins, and Nucleic Acids	16 (13.3%)
Gene Expression and Regulation	7 (5.8%)
DNA Structure and Dynamics	5 (4.2%)
Genetics and Inheritance	1 (0.8%)
Ecology and Evolution	24 (20.0%)
Medical Biology and Pathology	10 (8.3%)
Evolutionary Biology and Genetics	5 (4.2%)
Biophysics and Theoretical Biology	5 (4.2%)
Ecology and Environmental Science	4 (3.3%)
Biotechnology and Bioengineering	20 (16.7%)
Molecular and Genetic Techniques	9 (7.5%)
Other Techniques and Models	5 (4.2%)
Imaging and Microscopy	3 (2.5%)
Drug Discovery and Delivery	3 (2.5%)
Systems and Omics Biology	12 (10.0%)
Systems-Level Concepts and Models	8 (6.7%)
"-omics" Fields	4 (3.3%)
Cellular and Subcellular Biology	10 (8.3%)
Cellular Processes	3 (2.5%)
Cell Death	3 (2.5%)
Organelles and Subcellular Structures	2 (1.7%)
Cytoskeleton and Molecular Motors	2 (1.7%)
Biochemistry and Metabolism	10 (8.3%)
Proteins and Enzymes	8 (6.7%)
Metabolic Pathways	2 (1.7%)
Immunology	9 (7.5%)
Core Concepts and Processes	5 (4.2%)
Signaling and Receptors	4 (3.3%)
Neuroscience	6 (5.0%)
Molecular and Cellular Neuroscience	2 (1.7%)
Systems and Cognitive Neuroscience	2 (1.7%)
Neuropharmacology	1 (0.8%)
Clinical Neuroscience	1 (0.8%)

Table 8. Human Disease Domain Composition.

Sub-Discipline	Count (%)
Systemic & Structural Genetic Disorders	35 (29.2%)
Multi-system Genetic Syndromes	14 (11.7%)
Skeletal and Overgrowth Syndromes	13 (10.8%)
Syndromes with Major Dermatologic Features	4 (3.3%)
Genomic and Inheritance Disorders	4 (3.3%)
Neurogenetic Disorders	18 (15.0%)
Neurodevelopmental Syndromes	13 (10.8%)
Hereditary Neuromuscular and Neurodegenerative Diseases	5 (4.2%)
Oncology	15 (12.5%)
Hematologic Malignancies	6 (5.0%)
Solid Tumors	6 (5.0%)
Hereditary Cancer Syndromes	2 (1.7%)
Transmissible Cancers	1 (0.8%)
Neurology	14 (11.7%)
Neurodegenerative and Movement Disorders	4 (3.3%)
Epilepsy and Seizure Disorders	3 (2.5%)
Other Neurological Conditions	3 (2.5%)
Neuroinflammatory and Demyelinating Diseases	2 (1.7%)
Congenital and Structural Brain Disorders	2 (1.7%)
Immuno & Rheum	14 (11.7%)
Autoinflammatory and Autoimmune Syndromes	7 (5.8%)
Systemic Vasculitides	4 (3.3%)
Primary Immunodeficiencies	2 (1.7%)
Systemic and Catastrophic Conditions	1 (0.8%)
Endo & Metab	13 (10.8%)
Inborn Errors of Metabolism	11 (9.2%)
Disorders of Mineral and Porphyrin Metabolism	2 (1.7%)
Organ-System	11 (9.2%)
Hematologic Disorders	4 (3.3%)
Gastroenterology and Hepatology	3 (2.5%)
Dermatologic Disorders	1 (0.8%)
Pulmonology	1 (0.8%)
Infectious and Transmissible Diseases	1 (0.8%)
Other Conditions	1 (0.8%)

Table 9. Chemistry Domain Composition.

Sub-Discipline	Count (%)
Organic Chemistry	27 (22.5%)
Named Reactions and Rearrangements	11 (9.2%)
Synthesis, Reagents, and Specific Compounds	8 (6.7%)
Reaction Mechanisms and Concepts	5 (4.2%)
Stereochemistry and Molecular Structure	3 (2.5%)
Biochemistry and Chemical Biology	27 (22.5%)
Enzymes and Cofactors	11 (9.2%)
Biological Processes and Pathways	8 (6.7%)
Bio-analytical and Chemical Biology Tools	6 (5.0%)
Bio-inorganic Chemistry	2 (1.7%)
Physical Chemistry	19 (15.8%)
Quantum Chemistry and Spectroscopy	6 (5.0%)
Electrochemistry	4 (3.3%)
Thermodynamics and Calorimetry	3 (2.5%)
Kinetics and Dynamics	3 (2.5%)
Condensed Matter and Surface Chemistry	3 (2.5%)
Inorganometallics and Organometallics	17 (14.2%)
Organometallic Compounds	8 (6.7%)
Specific Inorganic Compounds	6 (5.0%)
Bonding and Molecular Structure	3 (2.5%)
Analytical Chemistry	11 (9.2%)
Spectroscopy and Spectrometry	5 (4.2%)
Instrumentation and Databases	3 (2.5%)
Microscopy and Surface Analysis	2 (1.7%)
Chromatography and Separation	1 (0.8%)
Applied Chemistry	8 (6.7%)
Photochemistry and Photophysics	2 (1.7%)
Magnetochemistry	2 (1.7%)
Industrial and Applied Chemistry	2 (1.7%)
Radiochemistry and Exotic Matter	1 (0.8%)
Theoretical Physics in Chemistry	1 (0.8%)
Materials Chemistry	6 (5.0%)
Key Concepts and Materials	6 (5.0%)
Theoretical Chemistry	5 (4.2%)
Concepts and Models	3 (2.5%)
Methods and Techniques	2 (1.7%)

Table 10. Physics Domain Composition.

Sub-Discipline	Count (%)
Quantum Field Theory	22 (18.3%)
Foundational and Axiomatic QFT	9 (7.5%)
Gauge Theories and the Standard Model	6 (5.0%)
Non-Perturbative and Topological QFT	3 (2.5%)
Supersymmetry and Beyond	2 (1.7%)
QFT in Curved Spacetime	2 (1.7%)
Mathematical Physics	19 (15.8%)
Differential Geometry and Topology	6 (5.0%)
Group Theory and Algebra	5 (4.2%)
Functional Analysis and Operator Theory	5 (4.2%)
Integrable Systems and Exact Solutions	2 (1.7%)
Twistor Theory	1 (0.8%)
Quantum Mechanics	16 (13.3%)
Foundational Concepts and Formalism	6 (5.0%)
Quantum Systems and Dynamics	4 (3.3%)
Quantum Computation and Information	4 (3.3%)
Quantum States and Phenomena	2 (1.7%)
General Relativity	16 (13.3%)
Alternative and Extended Theories of Gravity	7 (5.8%)
Formalisms and Solutions	5 (4.2%)
Astrophysical Relativity	3 (2.5%)
Spacetime and Black Holes	1 (0.8%)
Specialized Topics in Physics	16 (13.3%)
Atomic, Molecular, and Optical (AMO) Physics	7 (5.8%)
Particle and Nuclear Physics	5 (4.2%)
Plasma Physics	4 (3.3%)
Condensed Matter	14 (11.7%)
Quantum Materials and Phases	6 (5.0%)
Electronic and Spintronic Effects	4 (3.3%)
Many-Body and Superconducting Systems	2 (1.7%)
Statistical and Soft Matter	2 (1.7%)
String Theory	12 (10.0%)
Core Concepts and Dualities	6 (5.0%)
Advanced and Related Theories	5 (4.2%)
Cosmological Applications	1 (0.8%)
Classical Physics and Electrodynamics	5 (4.2%)
Core Concepts and Phenomena	5 (4.2%)

Table 11. Mathematics Domain Composition.

Sub-Discipline	Count (%)
Geometry & Topology	37 (30.8%)
Algebraic Topology & Homotopy Theory	12 (10.0%)
General & Geometric Topology	12 (10.0%)
Differential & Geometric Analysis	11 (9.2%)
K-Theory	2 (1.7%)
Number Theory	19 (15.8%)
Arithmetic & Algebraic Geometry	12 (10.0%)
Algebraic Number Theory & Class Field Theory	5 (4.2%)
Analytic Number Theory	2 (1.7%)
Analysis	18 (15.0%)
Operator Theory & Operator Algebras	9 (7.5%)
Real, Complex, & Harmonic Analysis	8 (6.7%)
Control Theory	1 (0.8%)
Cryptography & Logic	18 (15.0%)
Cryptography	7 (5.8%)
Theoretical Computer Science & Algorithms	6 (5.0%)
Logic & Set Theory	5 (4.2%)
Algebra & Group	14 (11.7%)
Homological & Abstract Algebra	8 (6.7%)
Representation Theory	4 (3.3%)
Group Theory	2 (1.7%)
Theo Phys	8 (6.7%)
Quantum Mechanics & Quantum Field Theory	3 (2.5%)
General Relativity & Spacetime Physics	2 (1.7%)
Plasma & Fluid Dynamics	1 (0.8%)
Physical Chemistry	1 (0.8%)
Dynamical Systems	1 (0.8%)
Statistics	6 (5.0%)
Probability & Statistics	4 (3.3%)
Stochastic Analysis	2 (1.7%)

Table 12. Computer Science Domain Composition.

Sub-Discipline	Count (%)
Theoretical Computer Science	43 (35.8%)
Mathematical Foundations	25 (20.8%)
Logic in Computer Science	12 (10.0%)
Computability and Complexity Theory	6 (5.0%)
Artificial Intelligence and Machine Learning	17 (14.2%)
Bayesian and Probabilistic Methods	5 (4.2%)
Learning Theory and Models	4 (3.3%)
Specialized Domains and Applications	4 (3.3%)
Logic, Reasoning, and Knowledge Representation	3 (2.5%)
Robotics and Embodied AI	1 (0.8%)
Programming Languages and Formal Methods	13 (10.8%)
Formal Verification	4 (3.3%)
Automata and Formal Languages	4 (3.3%)
Type Theory and Program Analysis	3 (2.5%)
Formal Semantics and Rewriting Systems	2 (1.7%)
Scientific Computing	13 (10.8%)
Computational Physics and Engineering	6 (5.0%)
Numerical Methods	5 (4.2%)
Optimization and Control	2 (1.7%)
Cryptography	12 (10.0%)
Cryptosystems and Primitives	4 (3.3%)
Cryptanalysis	4 (3.3%)
Secure Protocols	2 (1.7%)
Foundational Concepts	2 (1.7%)
Algorithmic Theory	11 (9.2%)
Graph and String Algorithms	3 (2.5%)
Algorithmic Information Theory	3 (2.5%)
Data Structures and Compression	2 (1.7%)
Combinatorial Optimization	2 (1.7%)
Algorithmic Game Theory	1 (0.8%)
Quantum Computing	11 (9.2%)
Quantum Algorithms and Information	4 (3.3%)
Quantum Foundations and Simulation	4 (3.3%)
Quantum Complexity and Models	2 (1.7%)
Interdisciplinary Applications	1 (0.8%)

Table 13. Economics Domain Composition.

Sub-Discipline	Count (%)
Pluralist Econ	34 (28.3%)
Behavioral and Experimental Economics	8 (6.7%)
Public Economics and Social Choice Theory	8 (6.7%)
Marxian Economics	7 (5.8%)
Other Theories and Applied Fields	6 (5.0%)
Economic History and History of Thought	5 (4.2%)
Microeconomics	19 (15.8%)
Consumer and Firm Theory	13 (10.8%)
General Equilibrium and Market Design	6 (5.0%)
Econometrics	15 (12.5%)
Models and Methods	11 (9.2%)
Statistical Concepts	4 (3.3%)
Macroeconomics	14 (11.7%)
Economic Models and Growth	8 (6.7%)
Monetary Policy and Critiques	6 (5.0%)
Math/Comp Econ	14 (11.7%)
Models and Methods	14 (11.7%)
Game Theory	9 (7.5%)
Concepts and Models	9 (7.5%)
Finance	9 (7.5%)
Asset Pricing and Risk Management	9 (7.5%)
International Economics	6 (5.0%)
Trade and Finance	6 (5.0%)

Table 14. Philosophy Domain Composition.

Sub-Discipline	Count (%)
Philosophy of Religion and Interdisciplinary	31 (25.8%)
History of Philosophy and Specific Texts	10 (8.3%)
Philosophy of Religion and Theology	9 (7.5%)
Other Interdisciplinary Concepts	8 (6.7%)
Philosophy of Language	4 (3.3%)
Logic and Philosophy of Mathematics	20 (16.7%)
Modal and Non-Classical Logics	7 (5.8%)
Logic in Computation and Language	7 (5.8%)
Set Theory and Mathematical Foundations	6 (5.0%)
Epistemology & Phil Sci	18 (15.0%)
Epistemology	10 (8.3%)
Philosophy of Science	4 (3.3%)
Philosophy of Physics	4 (3.3%)
Metaphysics & Ontology	16 (13.3%)
Core Concepts	9 (7.5%)
Theories of Reality	4 (3.3%)
Metaphysics of Mind	3 (2.5%)
Continental Philosophy	13 (10.8%)
Phenomenology and Existentialism	5 (4.2%)
Post-Structuralism and Deconstruction	4 (3.3%)
Lacanian Psychoanalysis	4 (3.3%)
Eastern Philosophy	12 (10.0%)
Indian Philosophy (Hindu Schools)	6 (5.0%)
Indian Philosophy (Buddhist Schools)	5 (4.2%)
Indian Religious/Philosophical Concepts	1 (0.8%)
Social Philosophy	6 (5.0%)
Legal and Social Philosophy	5 (4.2%)
Political Philosophy	1 (0.8%)
Mind	4 (3.3%)
Core Concepts	3 (2.5%)
Psychoanalytic Philosophy	1 (0.8%)

E. Active Probing Prompt

Querying Agent Prompt

You are a {domain} expert. I will have a preset {domain} name/term/concept in mind, and you need to continuously ask questions and reason until you can identify the name/term/concept. You can ask questions to narrow down the possibilities. Your questions must be in the form of a single, four-option multiple-choice question at a time. You can only ask one multiple-choice question at a time. Each question is crucial and should be based on your reasoning, professional knowledge, and the previous rounds of questions and answers. Please try to identify the term/concept in as few rounds as possible.

I have now preset a name/term/concept. Please begin with your first question.

Figure 7. The system prompt is for the querying agent. In each round, it poses strategic queries to quickly identify the P_{target} by interacting with an environment. The {domain} placeholder is replaced with the specific domain name (e.g., Human Disease, Physics) for each evaluation.

Knowledge Oracle Prompt

You are a {domain} expert. You will have a preset {domain} name/term/concept in mind, and I need to identify it by asking questions and using reasoning. I can ask questions to narrow down the scope, and my method of questioning will be a single, four-option multiple-choice question at a time. The specific details regarding my question are all within the <qa_response> part. You will respond to my questions based on some “golden information” (reference material).

The name/term/concept you have preset is: {puzzle_name}

Reference material: <reference_material> {puzzle_reference} </reference_material>

In your reply, you must include a <pure_parse>...</pure_parse> block and a <pure_answer>...</pure_answer> block.

For the <pure_parse>...</pure_parse> block:

You will only judge the correctness of the options. For example, when option A, D are correct, and option B, C are incorrect:

<pure_parse>A is correct, B is incorrect, C is incorrect, D is correct</pure_parse>

When all options are incorrect, reply with <pure_parse>A is incorrect, B is incorrect, C is incorrect, D is incorrect</pure_parse>

If the <qa_response> part provides only an incorrect final guess without including a multiple-choice question, reply with <pure_parse>Your guess is incorrect, please ask a new question</pure_parse> and <pure_answer>None</pure_answer>.

The <pure_parse> block is forbidden from containing the name of the preset name/term/concept or any other content from the <reference_material> part.

For the <pure_answer>...</pure_answer> block:

If the message you receive in the <qa_response> part (or one of the options in the <qa_response> part) explicitly states the name of your preset name/term/concept (or its alias), then include <pure_answer>Your preset name</pure_answer> in your reply (Here, “Your preset name” is replaced with the actual name that appears). If it does not, then include <pure_answer>None</pure_answer>.

All of your replies must include both of these blocks.

Figure 8. The system prompt for the Knowledge Oracle. This agent acts as an oracle, providing structured and non-leaking feedback to the strategic agent based on the P_{target} and I_{ref} .