VERDICT: Unveiling the Capability Threshold in Multi-Agent Logical Reasoning

Jasper T

Open University
jasper.tay@ou.ac.uk

Abstract

We introduce **VERDICT** (Verifiable Evaluation of Reasoning In Complex Truth-puzzles), a logical reasoning benchmark designed to rigorously evaluate the deductive capabilities of LLMs. VERDICT introduces **higher-dimensional complexity** through a 999-puzzle suite featuring six distinct character types across three difficulty tiers. We employ six LLMs using three prompting strategies: One-Shot, Chain-of-Thought (CoT), and Multi-Agent Debate (MAD). Our results reveal a critical **"Capability Threshold"**: while MAD significantly boosts accuracy for capable models like Gemini-2.5-Pro, it fails completely for weaker models, inducing "hallucination loops" where agents reinforce each other's errors. These findings suggest that while multi-agent architectures can act as "reasoning multipliers," they are not a universal remedy for fundamental reasoning deficits.

Introduction

Logical Reasoning is widely used in a plethora of fields like Mathematics and Philosophy to show the relationships between two or more objects. This can come in the form of various statement types. Well-Designed logical puzzles can serve as a valuable mechanism to understand the depth of Large Language Models (LLMs) reasoning and their abilities to recognize relationships among different parties when given a contextual puzzle (Li, Wang, and Zhang 2024; Giadikiaroglou et al. 2024). Previous literature has used puzzles contexts like **Knights and Knaves**, **Constraint Satisfaction Problems** (CSPs) , **River Crossing** and **Tower of Hanoi** puzzles (Lin et al. 2025). Of all these puzzle types, Knights and Knaves (K&K) is one of the toughest and most popular puzzle type due to its ability to detect LLM memorization (Xie et al. 2024). In each puzzle, models are challenged to logically identify each character type like Knights who always tells the truth or Knaves who always lies. This is popularized in the book "What is the Name of This Book?" which introduced not only 'Knights' and 'Knaves' but also other characters like 'Spies' and 'Alternators' (Smullyan 1978). It requires the usage of various reasoning abilities like suppositional, deductive and abductive reasoning to solve each puzzle (Byrne and Handley 1997). To the best of our research, even though there are numerous work on deductive logical reasoning there is not a wide variety of work focusing on Knights and Knave (Giadikiaroglou et al. 2024). We found that many past literature focuses on 2 dimension puzzles like *Knights-Knave* and 3 dimension puzzles like *Knights-Knave-Spy* (Mondorf and Plank 2024; Wu, Li, and Li 2025). Furthermore, past literature only uses open source like *LLaMA-2-7b* or/and smaller closed models like *GPT-5-Nano* and *Gemini-2.5-Flash* which might have caused poor performance on such benchmarks since intrinsic reasoning capability is the crux to achieve a higher accuracy rate (Shojaee et al. 2025). Additional studies employing reinforcement learning have demonstrated that models continue to exhibit poor performance on abductive reasoning tasks (Cai et al. 2025) A usage of various prompting methods like Chain-of-Thought (CoT)and Multi-Agent Debate (MAD) can sparsely improve models' performance (Saparov and He 2023; Dziri et al. 2023). Past studies bave also indicated that 'tit for tat' debate may not necessarily improve performance as it can lead to polarization and Judges in MAD may be bias towards agents that are of the same model (Wu, Li, and Li 2025).

Therefore, this begs the question: How do frontier Open and Closed-Source models LLM perform on a greater dimension (or difficulty) of K&K puzzle sets with a variety of prompting techniques? Thus, **VERDICT** aims to address the following Research Questions (RQs):

1. **(RQ1 Model Performance)** How will current frontier models perform when introduced with a greater number of character types? What are the differences between the reasoning and the accuracy of Open and Closed Source models?
2. **(RQ2 Prompting Techniques)** Will using a variety of prompting techniques help to improve the accuracy and reasoning of the individual models? Specifically, will Multi-Agent Debate (MAD) help in higher dimension problems?

In this paper, we introduce **VERDICT**, **V**erifiable **E**valuation of **R**easoning **I**n **C**omplex **T**ruth-puzzles, a logical reasoning benchmark which introduces higher dimension's K&K puzzles to evaluate the reasoning capability of Open- and Closed-Sourced Large Language Models. Due to restrictions of compute, we have prompted 999 puzzles to LLM of varying difficulty levels. In total, we have tested **VERDICT** on 6 models with a variety of prompting techniques while employing *LLM as a Judge* to quantify the soundness of the logical reasoning. We hope that, through this paper, **VERDICT** offers meaningful insights into the state of logical reasoning in frontier models.

Methodology

Building on standard Knights, Knaves, and Spy puzzles (Wu, Li, and Li 2025), **VERDICT** introduces three character types to increase complexity: *Sane* (Truth-teller), *Insane* (Always lies), and *Alternator* (Toggles truth/lie based on state). While Knight/Sane and Knave/Insane are behaviorally equivalent, they are treated as distinct but conceptually different to increase the combinatorial complexity of assignment spaces available (Shojaee et al. 2025). VERDICT utilizes statement types similar to *TruthQuest* (Mondorf and Plank 2024) but uniquely combines all types within single puzzles to test deep logical interpretation (Wan et al. 2024).

Dataset Distribution: The VERDICT benchmark consists of 999 uniquely generated puzzles, evenly distributed across three complexity tiers ($N = 333$ per level): Level 1 (Knight and Knave), Level 2 (adds Insane/Sane or Random/Alternator), and Level 3 (all 6 types). To ensure robustness against memorization and state-dependent bias, the ”Alternator” character type (present in Levels 2 and 3) is initialized with a balanced distribution of starting states (50%/50% Truth/Lie) across the dataset.

LLM as a Judge: Model outputs are evaluated by a ”Judge” model (GPT-5-Mini) using specific rubrics includes logical correctness, consistency and efficiency, which derives the **Overall Reasoning Score**. This includes both global and local logic correctness, allowing researchers to pinpoint sections of logical incorrectness. For the true **Accuracy Score** for each question, we have used a symbolic solver to attain the definite solution first to ensure that the Judge can accurately assess the accuracy of the models' output.

Experiments

Models Used

Overall, we used 6 models, specifically : *GPT-5.1*, *Gemini-2.5-Pro*, *LLaMa-4-Maverick*, *LLaMa-4-Scout*, *Phi-4* and *Gemma3-27B-IT*. This ensures a wide variety of frontier models used in both open and closed source models. VERDICT was tested on other frontier open source models like *LLaMa models*, *Phi-4* and *Gemma3* which should be sufficient to substantiate any findings that was posed in **RQ1**.

Puzzle Prompt Scenario

In each puzzle, human-like names are used (like Kevin, Laura) to better stimulate a realistic scenario. Each prompt comprises 2/4/6 statements (depending on the level) and each of the character type available which the model has to match. There is a mix of the statement types given in each puzzle.

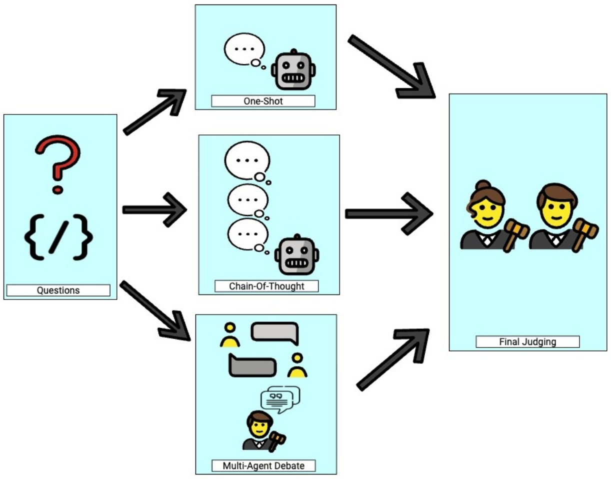

Figure 1: Caption

Puzzle Generation

To ensure that each puzzle is valid, we employed the *z3* SMT solver to model the problem as a constraint satisfaction task. Specifically, for every character $c \in C$ and type $t \in T$, a Boolean variable $V_{c,t}$ is initialized. We enforce the following constraints:

$$\forall c \left(\bigvee_{t} V_{c,t} \right) \tag{1}$$

$$\forall c, \forall i \neq j \left(\neg (V_{c,t_i} \wedge V_{c,t_j}) \right) \tag{2}$$

$$|C| = |T| \tag{3}$$

Chain-of Thought

Each CoT uses the following meta-prompt:

Think step-by-step:

1. Analyze each person's statement
2. For each person, consider what their statement means if they are each possible type
3. Look for contradictions to eliminate possibilities
4. Use logical deduction to narrow down the solution
5. Verify your answer is consistent with all statements

Therefore, the LLM's reasoning will consist of all the steps taken which can assist in further evaluation in *Local* and *Global* Logic.

Multi-Agent Debate

Multi-Agent Debate (MAD) is conducted in groups of 2 debaters and 1 moderator with a maximum of 3 rounds. All the debaters and moderator agents uses the same model to prevent bias. Each round will consist of a back and forth debate between the Affirmative and Negative which will debate until the moderator determines the final solution has been reached or the max number of rounds has been reached.

The moderator has a choice to end the debate early when it determines that the answer is found which can significantly increase performance. One difference VERDICT made was modifying the proposed 'Tit for Tat' variation in past literature to 'appropriately disagree' with the Affirmative side to prevent polarization. This allows the negative side to act as an auditor rather than a simple adversary which can filter out hallucinations early in the debate loop, reserving the Moderator's role for synthesizing verified consensus rather than arbitrating raw conflict. Below is the prompt that was modified:

> ”You are the **NEGATIVE** side, responsible for auditing the **AFFIRMATIVE** side's reasoning. Carefully check whether their proposed assignment satisfies all puzzle constraints, including whether each character's statements are consistent with their claimed type, whether any logical contradictions arise, and whether the solution is truly unique. If you find errors, clearly identify them and provide a corrected assignment; if their reasoning is sound, you may agree, but you must still verify it thoroughly. Conclude by presenting your own complete assignment.”

LLM as a Judge (Overall Reasoning Score)

After running each of the model through the benchmark, we make use of another LLM model, *GPT-5-Mini,* to evaluate the reasoning and answer based on a rubric which is shown below. This is to alleviate any form of bias the model might have to ensure a fair scoring system. Next, each of these metrics are summed to form the **Overall Reasoning Score**. For the Multi-Agent Debate, we introduced a modification to the evaluation metric. Instead of evaluating each debaters reasoning, we evaluate the moderator's reasoning for local and global logic. This is to have a comparable comparison to other prompting techniques where only the final reasoning is evaluated. Each of the graphical representation of the scores across the various models and levels are appended in the appendices.

Metrics:

Local Logic (20%) Validity of individual reasoning steps; no obvious logical mistakes.

Global Logic (20%) Overall coherence; do the steps collectively justify the answer?

Faithfulness (20%) No hallucinations; reasoning uses only facts from the puzzle.

Consistency (20%) No contradictions within the explanation.

Efficiency (20%) Reasoning is focused and avoids irrelevant tangents.

The LLM Judge will give a score between 0-20 for each of these metrics based on a rubric which is attached in the appendix.

Evaluation Protocol

Therefore, for each puzzle, the LLM will perform a two fold check for the **Accuracy Score** and the **Overall Reasoning Score**. Firstly, the Judge will take the ground-truth and check it against the models' responses to check for accuracy of the solution. This verification yields a binary output, which determines the **Accuracy Score**. Secondly, we will employ the LLM as a Judge to evaluate the logical derivation trace based on the five weighted dimensions.

Results and Discussion

Table 2 contains the average results of the *Overall Reasoning Score* and *Accuracy Score* on the **VERDICT** benchmark with 3 types of prompting techniques, One-Shot (OS), Chain-of-Thought (CoT)(CoT) and Multi-Agent Debate(MAD). We assumed One-Shot as the baseline result to compare it with other prompting techniques.

Baseline Results

The general baseline results shows for all models, One-Shot degrade performances in both accuracy and reasoning as the difficulty increases which suggests that an increase depth of reasoning is necessary to solve the puzzle. Closed source frontier models like *ChatGPT-5.1* and *Gemini-2.5 Pro* performed the best in terms of **Accuracy Score** and **Overall Reasoning Score** for all three difficulty levels. This suggests advance logical reasoning capability in current frontier closed source models in both getting high **Accuracy Score** and **Overall Reasoning Score** . Open Source models have high variance levels for reasoning and accuracy levels. All open source models performed poorly on level 3 tasks but *Llama-4-Scout* in **Overall Reasoning Score**. This can be attributed to the better performance in specific metrics (attached in Appendix) like Faithfulness, Consistency and Efficiency in *Llama-4-Scout* as compared to other Open Source models. This suggests that generally, frontier Open Source models perform poorly in logical reasoning but they do not have a tendency to hallucinate and is able to be consistent with their logical reasoning.

Chain-of Thought (CoT)

Past literature has found that Chain-Of Thought(CoT) do not substantially improve LLM model task performance. However, we also found similar findings for Level 2 and 3 puzzles in VERDICT with the exception of the LLaMa models on Level 1 basic Knights and Knaves puzzles (**0.56**,**0.54** vs 0.**69**,**0.81**) on the *Accuracy score*. However, one interesting finding is frontier closed source models such as *Chatgpt-5.1* and *Gemini-2.5 Pro* faced a dropped in reasoning score while open source models have a mix reasoning score. A plausible explanation in the deterioration of the *Local* and *Global* Logic scores, suggesting that additional reasoning steps can result in a higher rate of invalid logical inferences. Open models like *Gemma3-27b* and *Phi-4* improved from One-Shot to Chain-Of Thought in all three levels while *LLaMa-4 Scout* degraded in performance in all three levels. This suggests that the effectiveness of CoT is model dependent.

Multi-Agent Debate (MAD)

Overall, a critical analysis of the Multi-Agent Debate (MAD) results reveals that this technique is not a univer-

(a) Overall Reasoning Scores (0–100)

Model	**Level 1**			**Level 2**			**Level 3**		
	OS	CoT	MAD	OS	CoT	MAD	OS	CoT	MAD
gemini-2.5-pro	98.29	**97.22**	91.56	86.85	80.88	62.66	70.83	61.04	47.24
gemma-3-27b-it	78.70	76.41	61.40	23.79	32.92	20.29	22.19	28.99	16.56
llama-4-maverick	67.05	65.16	52.54	26.98	29.34	27.11	21.49	24.68	17.13
llama-4-scout	69.64	63.71	75.97	29.87	26.35	22.54	29.81	27.97	18.89
phi-4	75.69	65.89	77.98	25.59	29.91	26.28	19.62	21.85	18.33
gpt-5.1	**98.39**	97.20	**98.15**	**87.31**	**81.21**	**74.03**	**75.37**	**69.51**	**58.26**

(b) Final Answer Accuracy (0.0–1.0)

Model	**One-Shot**			**Chain-of-Thought (CoT)**			**Multi-Agent Debate**		
	L1	L2	L3	L1	L2	L3	L1	L2	L3
gemini-2.5-pro	**1.00**	**0.98**	**0.75**	**0.99**	**0.97**	**0.72**	**1.00**	**0.94**	**0.93**
gemma-3-27b-it	0.67	0.03	0.00	0.80	0.05	0.01	0.83	0.09	0.01
llama-4-maverick	0.56	0.03	0.00	0.69	0.03	0.00	0.65	0.11	0.00
llama-4-scout	0.54	0.04	0.00	0.81	0.04	0.00	0.91	0.03	0.00
phi-4	0.72	0.04	0.00	0.71	0.05	0.00	0.78	0.08	0.00
gpt-5.1	**1.00**	0.78	0.49	1.00	0.77	0.53	**1.00**	0.76	0.55

Table 1: Experimental Results on the **VERDICT** Benchmark. Panel (a) shows the internal reasoning scores extracted from the model's trace. Panel (b) shows the strict binary accuracy of the final answer. Note that models with high reasoning scores (e.g., Kimi-k2) do not always achieve high final accuracy in strict logic settings.

sal remedy for reasoning deficits; rather, its effectiveness is highly dependent on the model's baseline competency.

We observed significant improvements in **Accuracy Scores** only when the underlying model already demonstrated at least ”average” performance in One-Shot or Chain-of-Thought (CoT) settings. For open-source models like *LLaMA*, *Gemma*, and *Phi*, which started with a baseline accuracy of **0.00** on Level 3 puzzles in One-Shot settings, MAD failed to yield any improvement, with scores remaining static at **0.00**. This suggests that if a model cannot generate a initial hypothesis, the debate agents essentially hallucinate at one another which is supported by previous literature. Unlike *Gemini-2.5-Pro*, which had a baseline competence of **0.75** to build upon, these weaker models appeared to enter a ”negative feedback loop” where the auditing agent failed to identify logical inconsistencies, effectively validating the errors of the proposing agent. Thus, multi-agent architectures seem to act **less as creators** of reasoning and more as **multipliers of existing latent ability**. For models that cleared this threshold, specifically closed-source frontier models like *Gemini-2.5-Pro*, MAD proved highly effective for accuracy. Gemini's accuracy on Level 3 surged from **0.75** (One-Shot) to **0.93** (MAD). Therefore, this phenomenon of **”Capability Threshold”** is a limitation of MAD especially in logical reasoning.

However, this boost came with a notable trade-off: a deterioration in the **Overall Reasoning Score**. We observed a stark degradation in the reasoning metrics for *Gemini-2.5-Pro* and *GPT-5.1* across all three levels, with the sharpest declines occurring in Levels 2 and 3. This creates a paradox where models become more accurate but apparently ”worse” at reasoning. We hypothesize this is largely due to our evaluation methodology, which assesses the **Moderator's** final summary rather than the full debate transcript. The Moderator often synthesizes the correct answer efficiently, discarding the granular step-by-step logic that the Judge model rewards. Future work should investigate this discrepancy by evaluating the individual debaters' reasoning chains separately to determine if the ”reasoning degradation” is a true loss of logic or simply an artifact of summarization.

Conclusion

Overview

In conclusion, we presented VERDICT, a novel benchmark for assessing the logical reasoning limits of LLMs through complex ”Knights and Knaves” puzzles. By integrating a symbolic *Z3* solver into our generation and evaluation pipeline, we ensured a rigorous ground-truth standard often missing in natural language reasoning tasks. Our comprehensive evaluation of open and closed-source models yields two key insights:

First, we identify a distinct **”Capability Threshold”** for Multi-Agent Debate. Contrary to the assumption that debate universally improves reasoning, our results show it acts as a ”reasoning multiplier” rather than a generator. For models below this threshold , debate is counter-productive, leading to ”destructive interference” where agents converge on in-correct consensus.

Secondly, while closed-source frontier models demonstrate superior handling of higher-dimensional logic, open-source models exhibit a sharp performance drop-off as complexity increases. This suggests that current open-weights training has not yet bridged the gap in intrinsic deductive reasoning. We hope VERDICT will encourage researchers to developing methods that can bootstrap reasoning in smaller models without relying on the computational overhead of multi-agent systems.

Research Questions (RQs)

In **VERDICT**, we also addressed our 2 RQs poised. For **RQ1**, both open and close-source models degrade as the difficulty(levels) increases. However, there is a stark difference in model performance between close and open source models. This is a potential research direction that can be undertaken to improve open source models in complex logical reasoning. For **RQ2**, the improvement in accuracy and reasoning scores are not consistent due to the changes in prompting techniques. MAD assists models that have a substantial reasoning capability to improve like *Gemini-2.5 Pro*. Other prompting techniques like Chain-Of Thought sparsely improves performance and reasoning but it can degrade performance when in simpler puzzles when the solutions are more straightforward.

Limitations

While **VERDICT** advancements over previous benchmarks, several limitations remain that warrant further exploration.

Temperature of LLMs

Consistent with standard evaluation methodologies, we conducted a single pass for each model at a temperature of 0. While this ensures reproducibility, it may not fully capture minor non-deterministic variations in model output.

MAD design choice

To ensure fair comparability with single-agent baselines like One-Shot and CoT, our evaluation of Multi-Agent Debate (MAD) was restricted to the moderator's final reasoning. This design choice, however, obscures the quality of individual debater responses and potential polarization dynamics, which future work could assess to provide a more holistic view of the debate process.

LLM as a Judge

Relying solely on GPT-5-Mini as the Judge may introduce evaluation biases such as a preference for outputs from similar architectures or assigning higher reasoning scores to longer reasoning chains. Additionally, there might be instances where LLM hallucinates when comparing the output of the LLM to the ground-truth given. To mitigate this, future iterations of VERDICT can employ ensemble judging to ”cancel-out” architecture specific biases and tapping on symbolic solvers to judge the LLM reasoning trace. This can mitigate the subjective assessment of the current judgement.

Language limitations

Finally, the current benchmark consists exclusively of English-language puzzles, potentially favoring models with English-centric training data and highlighting the need for multilingual or culturally neutral extensions to test cross-lingual logical reasoning.

References

Byrne, R. M. J.; and Handley, S. J. 1997. Reasoning Strategies for Suppositional Deductions. *Cognition*, 62(1): 1–49.

Cai, C.; Zhao, X.; Liu, H.; Jiang, Z.; Zhang, T.; Wu, Z.; Hwang, J.-N.; and Li, L. 2025. The Role of Deductive and Inductive Reasoning in Large Language Models. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16780–16790. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Dziri, N.; Lu, X.; Sclar, M.; Li, X.; Jiang, L.; Lin, B. Y.; Welleck, S.; West, P.; Bhagavatula, C.; Bras, R. L.; Hwang, J.; Sanyal, S.; Ren, X.; Ettinger, A.; Harchaoui, Z.; and Choi, Y. 2023. Faith and Fate: Limits of Transformers on Compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 70293–70332. Curran Associates, Inc.

Giadikiaroglou, P.; Lymperaiou, M.; Filandrianos, G.; and Stamou, G. 2024. Puzzle solving using reasoning of large language models: A survey. arXiv:2402.11291.

Li, Y.; Wang, H.; and Zhang, C. 2024. Assessing logical puzzle solving in large language models: Insights from a minesweeper case study. arXiv:2311.07387.

Lin, B. Y.; Bras, R. L.; Richardson, K.; Sabharwal, A.; Poovendran, R.; Clark, P.; and Choi, Y. 2025. ZebraLogic: On the Scaling Limits of LLMs for Logical Reasoning. arXiv:2502.01100.

Mondorf, P.; and Plank, B. 2024. Liar, Liar, Logical Mire: A Benchmark for Suppositional Reasoning in Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 7114–7137.

Saparov, A.; and He, H. 2023. Language Models are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR)*.

Shojaee, P.; Mirzadeh, I.; Alizadeh, K.; Horton, M.; Bengio, S.; and Farajtabar, M. 2025. The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity. arXiv:2506.06941.

Smullyan, R. M. 1978. *What is the Name of This Book?: The Riddle of Dracula and Other Logical Puzzles*. Englewood Cliffs, NJ: Prentice-Hall.

Wan, Y.; Wang, W.; Yang, Y.; Yuan, Y.; Huang, J. T.; He, P.; Jiao, W.; and Lyu, M. R. 2024. A & B == B & A: Triggering Logical Reasoning Failures in Large Language Models. *arXiv preprint arXiv:2401.00757*.

Wu, H.; Li, Z.; and Li, L. 2025. Can LLM Agents Really Debate? A Controlled Study of Multi-Agent Debate in Logical Reasoning. arXiv:2511.07784.

Xie, C.; Huang, Y.; Zhang, C.; Yu, D.; Chen, X.; Lin, B. Y.; Li, B.; Ghazi, B.; and Kumar, R. 2024. On Memorization of Large Language Models in Logical Reasoning. arXiv:2410.23123.

Xie, T.; Gao, Z.; Ren, Q.; Luo, H.; Hong, Y.; Dai, B.; Zhou, J.; Qiu, K.; Wu, Z.; and Luo, C. 2025. Logic-RL: Unleashing LLM Reasoning with Rule-Based Reinforcement Learning. arXiv:2502.14768.

Appendix

Prompts

This appendix details the exact prompts used for the single-agent reasoning methods and the automated evaluation Judge.

A. Reasoning Method Prompts

A.1. One-Shot Prompting

One-Shot: System Prompt

```
You are a logic expert. You MUST return ONLY valid JSON that follows the schema.
```

One-Shot: User Prompt

```
Solve the following puzzle and return the output in this exact schema:

{
  "label": {
    "<name>": "<type>"
  },
  "reasoning": "<your reasoning>"
}

Now solve this puzzle and return ONLY the JSON result:

STORY:
{puzzle}
```

A.2. Chain-of-Thought(CoT) Prompting

CoT: System Prompt

```
You are a logic expert. You solve puzzles by thinking step-by-step.
You MUST return ONLY valid JSON that follows the schema.
```

CoT: User Prompt

```
Solve the following logic puzzle using Chain-of-Thought (CoT) reasoning.

Think step-by-step:
1. Analyze each person’s statement
2. For each person, consider what their statement means if they are each possible type
3. Look for contradictions to eliminate possibilities
4. Use logical deduction to narrow down the solution
5. Verify your answer is consistent with all statements

Return your answer in this exact JSON schema:

{
  "steps": [
    "Step 1: Analyze statement X...",
    "Step 2: If person Y is type Z, then...",
    "Step 3: This contradicts statement A, so...",
    ...
  ],
  "reasoning": "Complete explanation of your logical deduction process",
  "label": {
    "<name>": "<type>",
    ...
  }
}

The "steps" array should contain your detailed step-by-step reasoning.
The "reasoning" should be a summary of your overall approach.
The "label" should be your final answer mapping each person to their type.

PUZZLE:
{puzzle}

Now solve this puzzle step-by-step and return ONLY the JSON result:
```

B. Judge Evaluation Prompts

B.1. Single-Agent Evaluation

Judge: System Prompt

```
You are a strict logician evaluating the QUALITY OF REASONING on Knights-and-Knaves style logic puzzles. You are given:

1. The puzzle description.
2. The ground-truth solution.
3. A model’s final answer and explanation.

Your job is NOT to re-solve the puzzle from scratch, but to Judge how good the model’s reasoning is, assuming you know the ground truth.

Rate each dimension carefully, following these definitions:

* logic_local (20%): Are individual reasoning steps locally valid (no obvious logical errors)?
* logic_global (20%): Taken together, do the steps justify the final answer? Are there major leaps?
* faithfulness (20%): Does the explanation rely only on the puzzle text? No invented statements or facts.
* consistency (20%): Is the explanation internally consistent (no contradictions)?
* efficiency (20%): Is the reasoning focused on useful deductions instead of irrelevant tangents?

From each dimension (worth 20 percent each), give a score between 0 and 20 on:
* overall_reasoning: Your holistic assessment of reasoning quality ignoring final correctness.

SCORING GUIDELINES (be strict and discriminating):
- 20: Flawless reasoning, every step is rigorous and clear
- 18-19: Excellent, minor presentation issues only
- 15-17: Good, but has small logical gaps or redundancies
- 12-14: Adequate, but with notable logical weaknesses
- 8-11: Significant logical errors or confusion
- 4-7: Major flaws, barely coherent reasoning
- 0-3: Completely wrong or nonsensical

MOST RESPONSES SHOULD SCORE BETWEEN 12-18, NOT AT THE EXTREMES.
Be discriminating - perfect scores (20) should be rare.

You must also check whether the final answer matches the ground truth and set final_answer_correct to 0 or 1.

Respond ONLY with a JSON object in this exact schema, no commentary:
{
  "final_answer_correct": 0 or 1,
  "logic_local": any number between 0 and 20,
  "logic_global": any number between 0 and 20,
  "faithfulness": any number between 0 and 20,
  "consistency": any number between 0 and 20,
  "efficiency": any number between 0 and 20,
  "overall_reasoning": sum of the above five dimensions (0-100)
}
```

Judge: User Prompt Template

```
PUZZLE:
{puzzle}

GROUND TRUTH:
{ground_truth}

MODEL ANSWER:
{model_answer}

MODEL EXPLANATION:
{model_explanation}
```

C. Dataset Examples

This section provides specific examples of the logic puzzles used in the VERDICT benchmark across three difficulty levels.

C.1. Level 1: Basic Knights and Knaves (2 people)

Level 1 Example (ID: L1_00000)

```
Rules:
On an island, each person is exactly one of the following types:
- Knights always tell the truth
- Knaves always lie

Each person is a different type (no two people are the same type).

Puzzle:
You meet Kevin, and Laura.

Kevin says, "Laura and I are different types."
Laura says, "Kevin is a Knave."

Can you determine who is what?
```

C.2. Level 2: Extended Types (4 people) Level 2 introduces intermediate complexity with 4 agents. We utilize two distinct variants to test different types of logical constraints.

Level 2A: Sane/Insane Variant (ID: L2_00000)

```
Rules:
On an island, each person is exactly one of the following types:
- Knights always tell the truth
- Knaves always lie
- Sane people always tell the truth
- Insane people believe only false things and say what they believe

Each person is a different type (no two people are the same type).

Puzzle:
You meet Kevin, Laura, Marcus, and Natalie.

Natalie says, "If Marcus is an Insane, then Kevin is a Sane."
Kevin says, "Marcus is a Sane if and only if Laura is a Sane."
Marcus says, "Natalie is a Knave if and only if Laura is a Sane."
Laura says, "If Natalie is an Insane, then Marcus is a Knave."
Natalie says, "Kevin is a Knight if and only if Marcus is an Insane."

Can you determine who is what?
```

Level 2B: Random/Alternator Variant (ID: L2_00001)

```
Rules:
On an island, each person is exactly one of the following types:
- Knights always tell the truth
- Knaves always lie
- Random people can randomly tell the truth or lie
- Alternators alternate between telling the truth and lying

Each person is a different type (no two people are the same type).

This round, the Alternator is telling the TRUTH.

Puzzle:
You meet Kevin, Laura, Marcus, and Natalie.

Laura says, "Marcus is a Knave if and only if Kevin is an Alternator."
Kevin says, "Natalie is a Knave if and only if Laura is a Knave."
Natalie says, "If Laura is a Random, then Kevin is a Random."
Kevin says, "Laura is an Alternator if and only if Natalie is a Knight."

Can you determine who is what?
```

C.3. Level 3: Full Complexity (6 people) Level 3 represents the highest difficulty, requiring the model to disentangle the states of 6 distinct character types simultaneously.

Level 3 Example (ID: L3_00000)

```
Rules:
On an island, each person is exactly one of the following types:
- Knights always tell the truth
- Knaves always lie
- Sane people always tell the truth
- Insane people believe only false things and say what they believe
- Random people can randomly tell the truth or lie
- Alternators alternate between telling the truth and lying

Each person is a different type (no two people are the same type).

This round, the Alternator is telling the TRUTH.

Puzzle:
You meet Kevin, Laura, Marcus, Natalie, Oliver, and Penelope.

Marcus says, "If Kevin is an Insane, then Laura is an Insane."
Kevin says, "If Laura is a Random, then Natalie is a Knave."
Oliver says, "If Natalie is an Alternator, then Penelope is a Random."
Oliver says, "Laura is a Knave if and only if Kevin is an Insane."
Kevin says, "If Laura is a Knave, then Natalie is a Knave."
Penelope says, "Natalie is a Knave if and only if Kevin is an Insane."
Penelope says, "If Marcus is a Sane, then Natalie is a Knave."
Natalie says, "Penelope is an Alternator if and only if Laura is a Random."

Can you determine who is what?
```

Character Type Definitions

1. Knight (Truth-teller) **Definition.**

$$\tau_i = \text{Knight} \iff \forall s_{i,j},\ \text{Say}(p_i, s_{i,j}) = s_{i,j}$$

Interpretation. Knights always state true propositions. This follows the standard Knights-and-Knaves semantics.

2. Knave (Liar) **Definition.**

$$\tau_i = \text{Knave} \iff \forall s_{i,j},\ \text{Say}(p_i, s_{i,j}) = \neg s_{i,j}$$

Interpretation. Knaves always state false propositions and are the logical negation of Knights.

3. Sane (Truth-believer) **Definition.**

$$\tau_i = \text{Sane} \iff \begin{cases} \text{Belief}(p_i, s) = s \\ \text{Say}(p_i, s) = \text{Belief}(p_i, s) \end{cases}$$

Interpretation. Sane agents are functionally equivalent to Knights in observable speech. They are distinguished conceptually by their belief model, but not by their truth-telling behavior. This equivalence is made explicit to avoid ambiguity.

4. Insane (False-believer) **Definition.**

$$\tau_i = \text{Insane} \iff \begin{cases} \text{Belief}(p_i, s) = \neg s \\ \text{Say}(p_i, s) = \text{Belief}(p_i, s) \end{cases}$$

Interpretation. Insane agents are functionally equivalent to Knaves in speech behavior.

5. Alternator Each Alternator maintains an internal binary state:

$$\sigma_i \in \{\text{Truth}, \text{Lie}\}$$

Definition.

$$\text{Say}(p_i, s) = \begin{cases} s, & \text{if } \sigma_i = \text{Truth} \\ \neg s, & \text{if } \sigma_i = \text{Lie} \end{cases}$$

After each utterance, the internal state flips:

$$\sigma_i \leftarrow \neg \sigma_i$$

Dataset Constraint. The initial state $\sigma_i^{(0)}$ is explicitly provided in the puzzle. For Example:

This round, the Alternator is telling the truth

Interpretation. Alternators behave deterministically given their initial state and support multi-statement temporal reasoning.

6. Random **Definition.**

$$\Pr\left(\text{Say}(p_i, s) = s\right) = 0.5$$

Interpretation. Random agents provide no reliable logical information. Their statements do not constrain the solution space and are ignored during logical deduction.

(a) **One-Shot** Reasoning Metrics

	Level 1					**Level 2**					**Level 3**				
Model	Loc	Glo	Fth	Con	Eff	Loc	Glo	Fth	Con	Eff	Loc	Glo	Fth	Con	Eff
Gemini-2.5-Pro	19.6	19.6	19.8	19.9	19.4	16.6	16.5	18.9	18.5	16.3	13.4	13.0	16.0	15.2	13.2
Gemma-3-27b	15.2	14.3	18.9	14.5	15.8	4.5	3.7	7.2	3.9	4.4	4.6	3.6	6.2	3.6	4.2
Llama-4-Mav	12.6	12.1	17.5	11.7	13.2	4.4	3.5	8.4	6.2	4.5	3.8	3.0	6.2	4.8	3.7
Llama-4-Scout	13.2	12.1	18.4	12.1	13.9	5.3	4.2	9.6	5.5	5.3	5.0	3.9	8.7	7.3	4.9
Phi-4	14.0	13.8	18.3	14.9	14.7	4.5	3.8	8.1	4.4	4.8	3.8	3.1	5.2	3.7	3.9
GPT-5.1	19.6	19.6	20.0	19.9	19.4	16.9	16.4	18.8	18.8	16.4	14.5	13.6	16.7	16.7	13.9

(b) **Chain-of-Thought (CoT) (CoT)** Reasoning Metrics

	Level 1					**Level 2**					**Level 3**				
Model	Loc	Glo	Fth	Con	Eff	Loc	Glo	Fth	Con	Eff	Loc	Glo	Fth	Con	Eff
Gemini-2.5-Pro	19.3	19.3	19.5	19.8	19.3	15.1	14.4	18.2	18.1	15.2	11.2	10.4	14.2	14.0	11.3
Gemma-3-27b	13.9	13.2	18.2	16.9	14.2	4.3	3.4	10.1	10.9	4.3	3.3	2.5	9.1	10.7	3.4
Llama-4-Mav	11.4	10.3	16.5	15.6	11.4	3.1	2.5	9.6	11.1	3.1	2.4	1.9	8.4	9.4	2.4
Llama-4-Scout	10.2	9.2	16.9	17.6	9.8	2.7	2.2	8.2	10.7	2.5	2.5	2.1	9.2	11.6	2.5
Phi-4	11.4	10.7	17.3	14.6	11.8	4.0	3.2	8.9	9.7	4.0	2.5	2.0	6.9	8.0	2.4
GPT-5.1	19.2	19.3	19.8	19.6	19.2	15.2	14.6	18.0	18.1	15.3	12.9	11.8	16.1	15.7	12.9

(c) **Multi-Agent Debate (MAD)** Reasoning Metrics

	Level 1					**Level 2**					**Level 3**				
Model	Loc	Glo	Fth	Con	Eff	Loc	Glo	Fth	Con	Eff	Loc	Glo	Fth	Con	Eff
Gemini-2.5-Pro	18.4	18.3	18.6	19.5	16.9	12.0	11.3	12.4	16.0	11.0	9.2	8.1	8.3	13.6	8.0
Gemma-3-27b	12.4	11.9	11.6	14.4	11.0	3.8	3.1	3.2	6.9	3.3	3.3	2.5	2.3	5.8	2.7
Llama-4-Mav	9.7	9.1	11.9	12.6	9.2	3.8	3.2	7.7	8.9	3.6	2.3	1.8	4.6	6.3	2.1
Llama-4-Scout	14.5	13.8	16.6	17.3	13.8	3.4	2.8	5.0	8.3	3.1	2.7	2.2	4.1	7.5	2.3
Phi-4	15.0	14.9	17.3	15.7	15.1	4.5	3.7	6.2	7.4	4.5	3.3	2.7	3.6	5.6	3.1
GPT-5.1	19.5	19.5	19.9	19.8	19.3	14.4	13.4	15.4	17.0	13.9	11.5	10.4	11.4	14.5	10.4

Table 2: Detailed performance breakdown by prompting strategy. Metrics are scaled 0–20. **Loc**: Local Logic, **Glo**: Global Logic, **Fth**: Faithfulness, **Con**: Consistency, **Eff**: Efficiency. Panel (a) One-Shot, Panel (b) CoT, Panel (c) MAD.

Statement Type	**Natural Language Example**	**Logical Form**
Self-Assertion	X: ”I am a knight.”	P_X
Accusation	X: ”Y is a knight (or knave).”	$P_X \Leftrightarrow \psi_Y \quad (X \neq Y)$
Conjunctive Claim	X: ”Y and Z are both knights or knaves.”	$P_X \Leftrightarrow (\psi_Y \wedge \psi_Z) \quad (X \neq Y \neq Z)$
Conditional Claim	X: ”If Y is a knight/knave, then Z is as well.”	$P_X \Leftrightarrow (\psi_Y \rightarrow \psi_Z) \quad (X \neq Y \neq Z)$
Biconditional Claim	X: ”Y is a knight/knave exactly when Z is one.”	$P_X \Leftrightarrow (\psi_Y \Leftrightarrow \psi_Z) \quad (X \neq Y \neq Z)$

Table 3: A taxonomy of typical statement patterns used in Knights-and-Knaves puzzles from (Mondorf and Plank 2024)

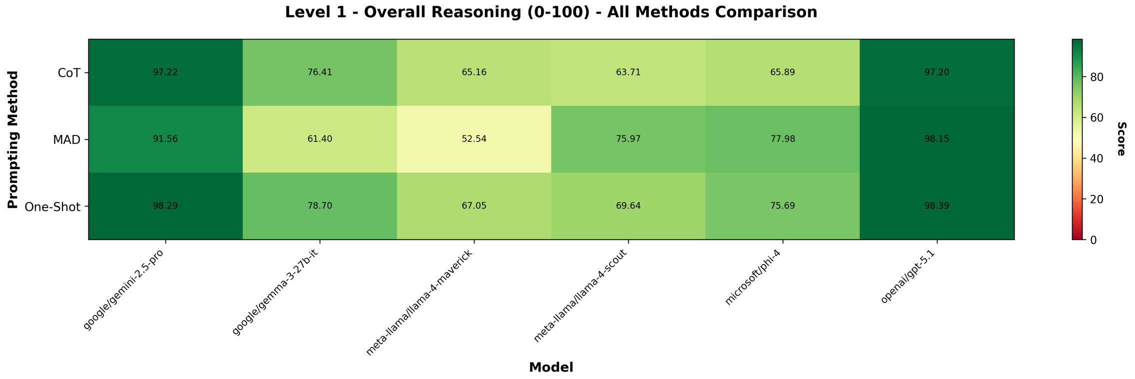

Figure 2: Overall Reasoning – Level 1

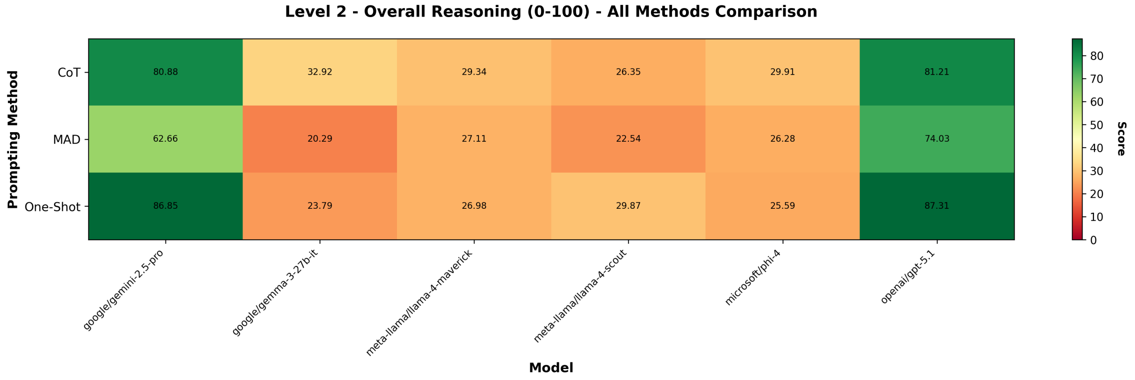

Figure 3: Overall Reasoning – Level 2

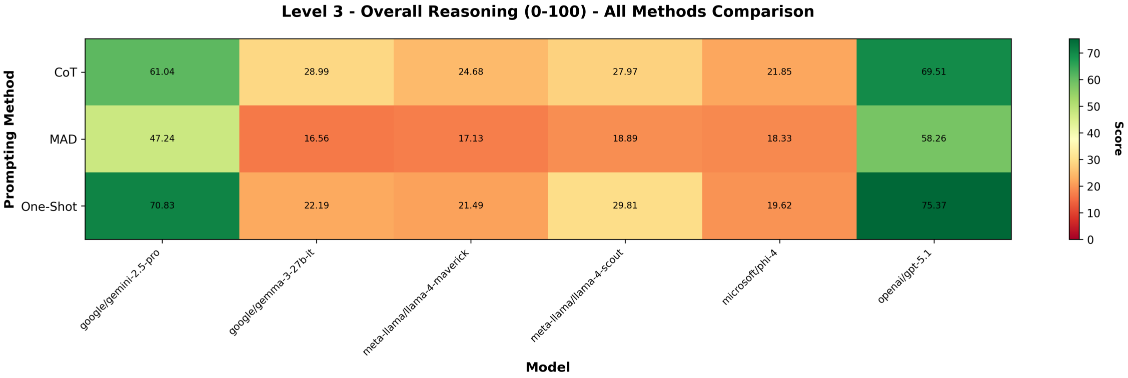

Figure 4: Overall Reasoning – Level 3

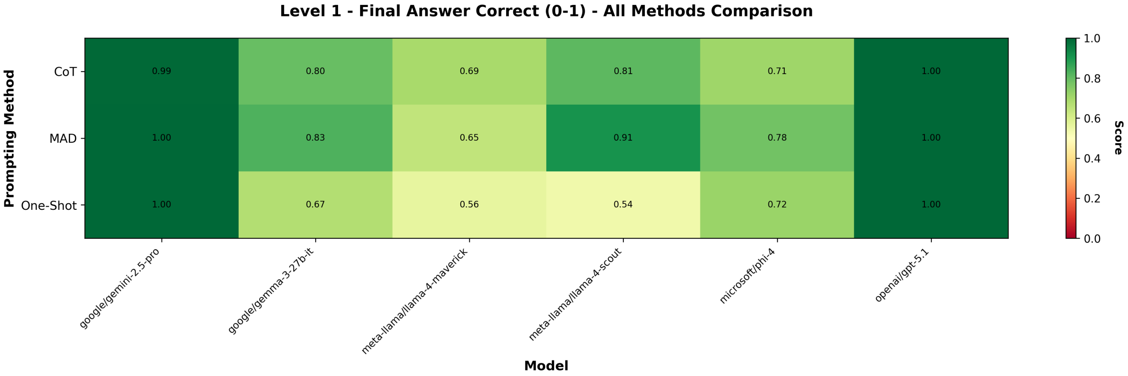

Figure 5: Final Answer Correctness – Level

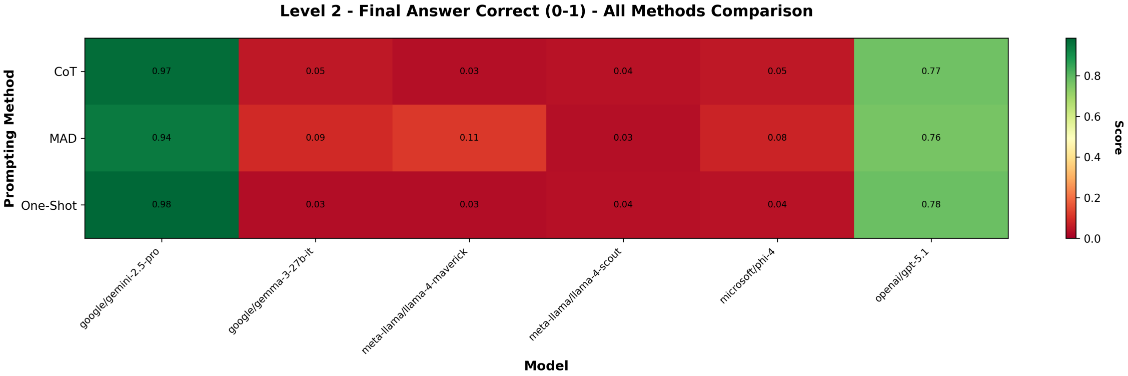

Figure 6: Final Answer Correctness – Level 2

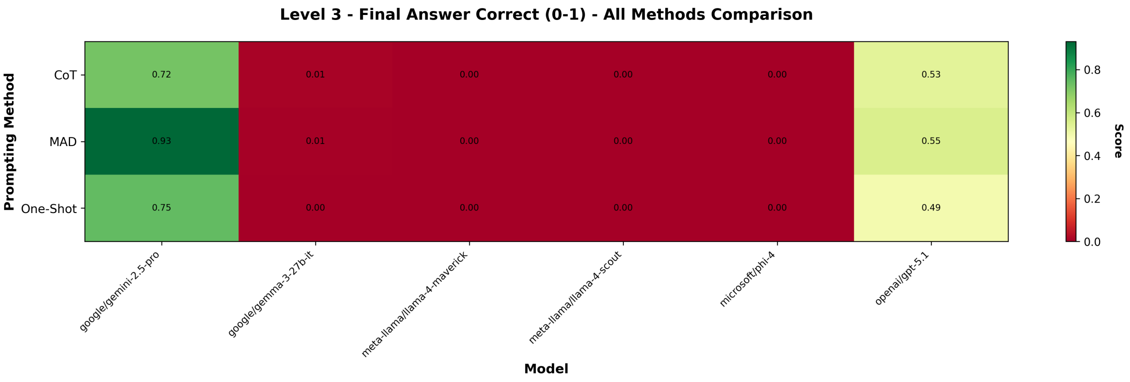

Figure 7: Final Answer Correctness – Level 3

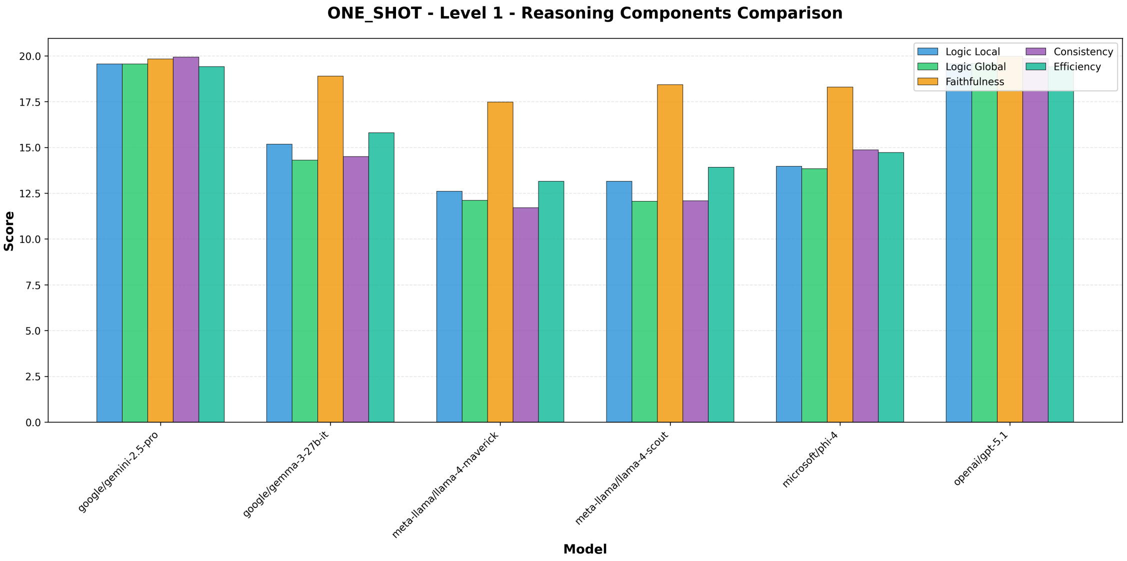

Figure 8: One-Shot – Level 1

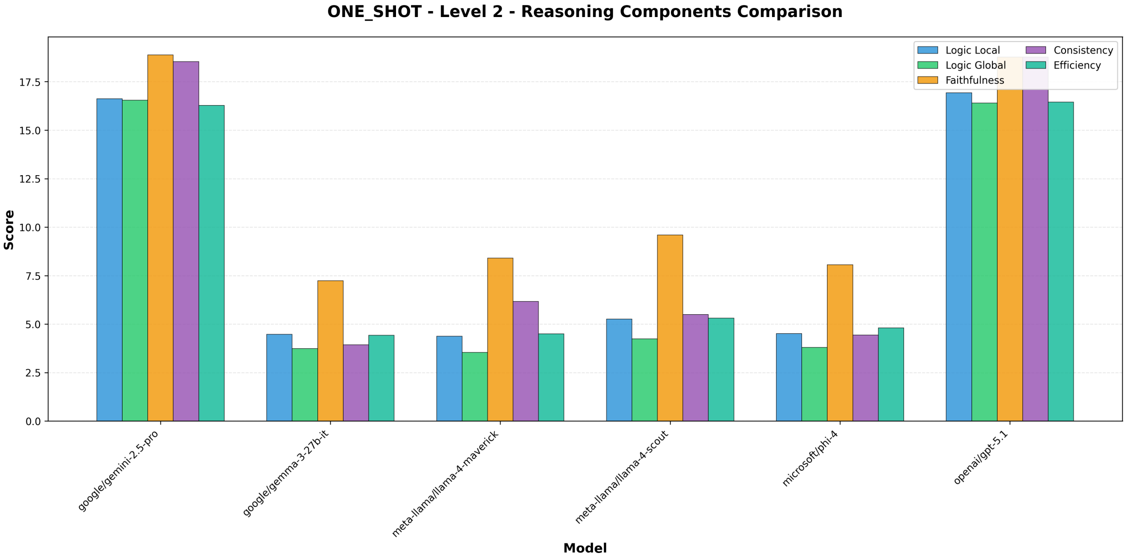

Figure 9: One-Shot – Level 2

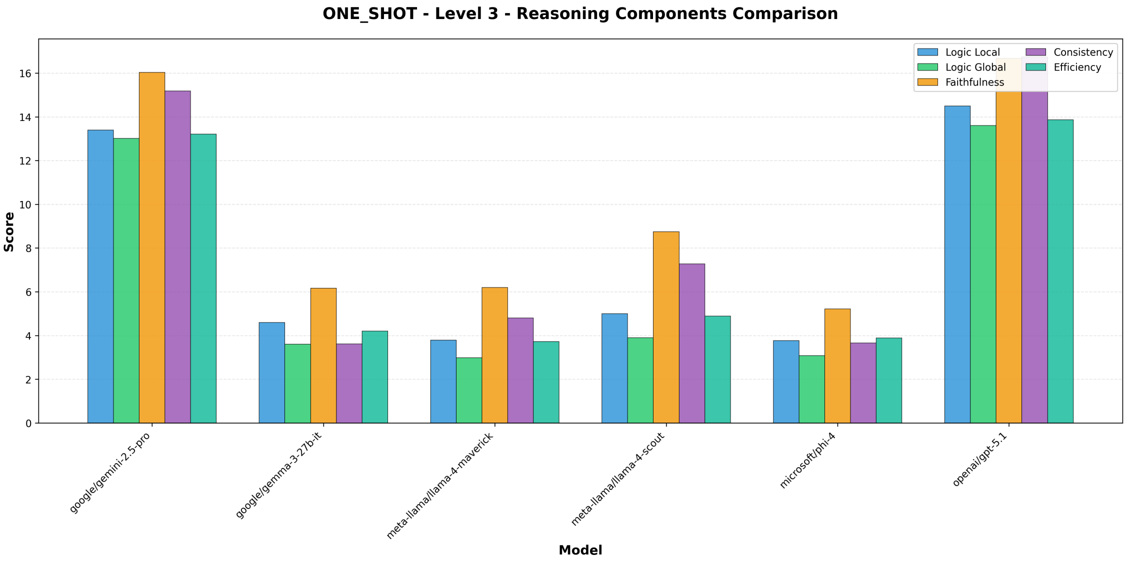

Figure 10: One-Shot – Level 3

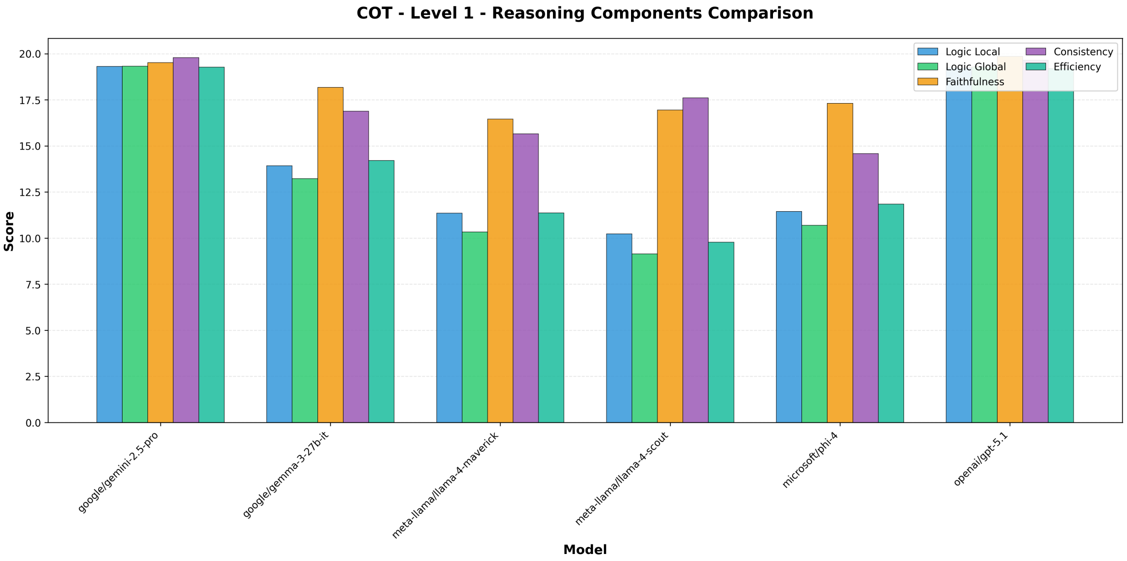

Figure 11: Chain-Of Thought – Level 2

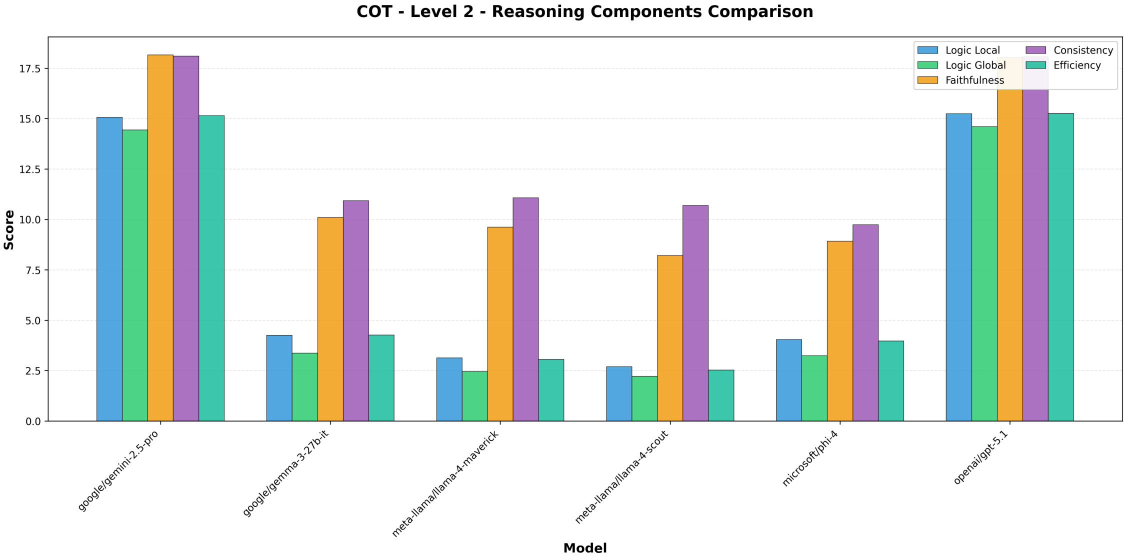

Figure 12: Chain-Of Thought – Level 2

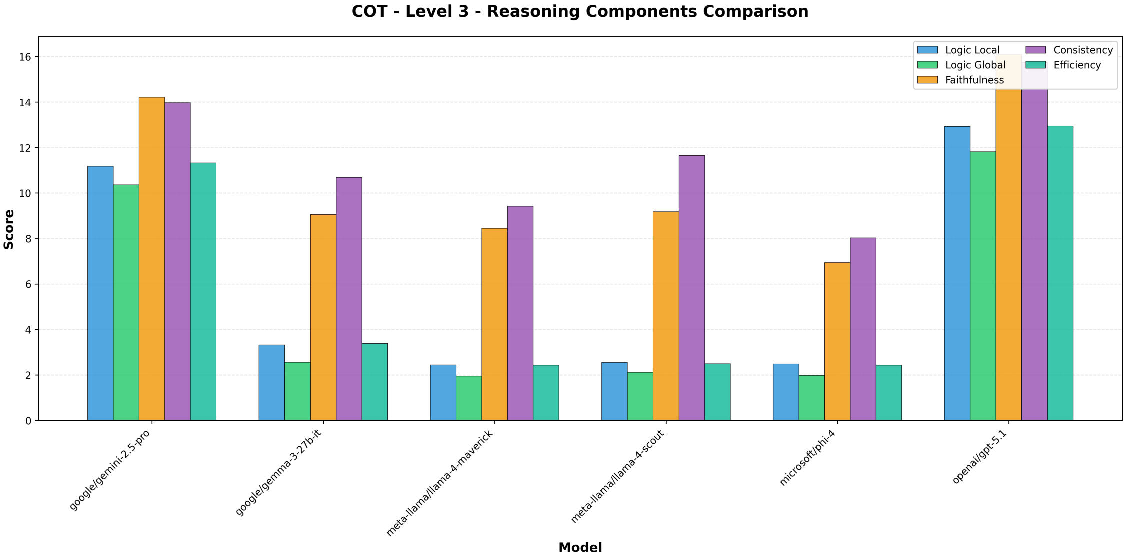

Figure 13: Chain-Of Thought – Level 3

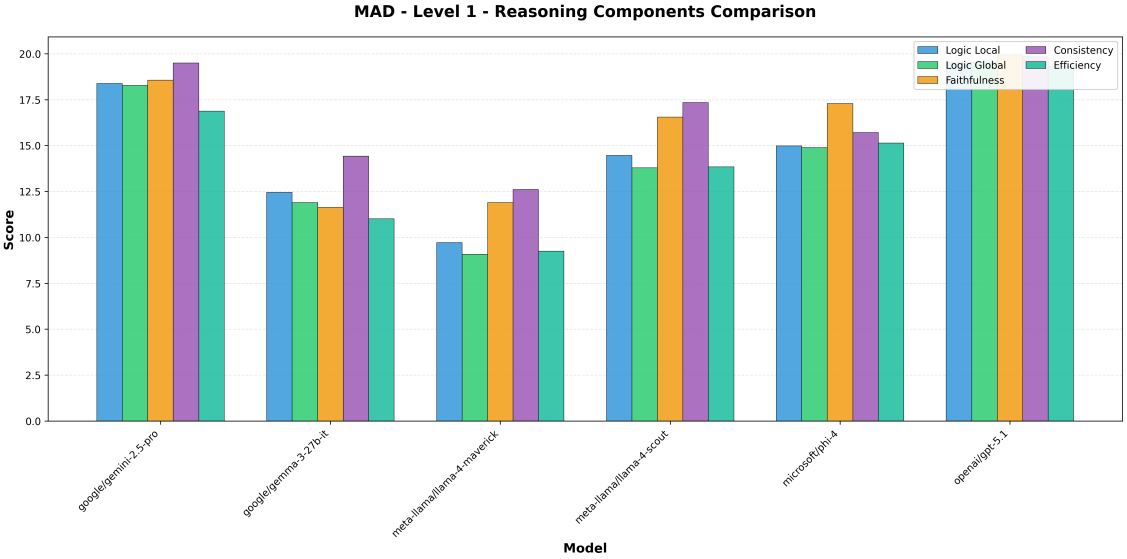

Figure 14: Multi-Agent Debate – Level 1

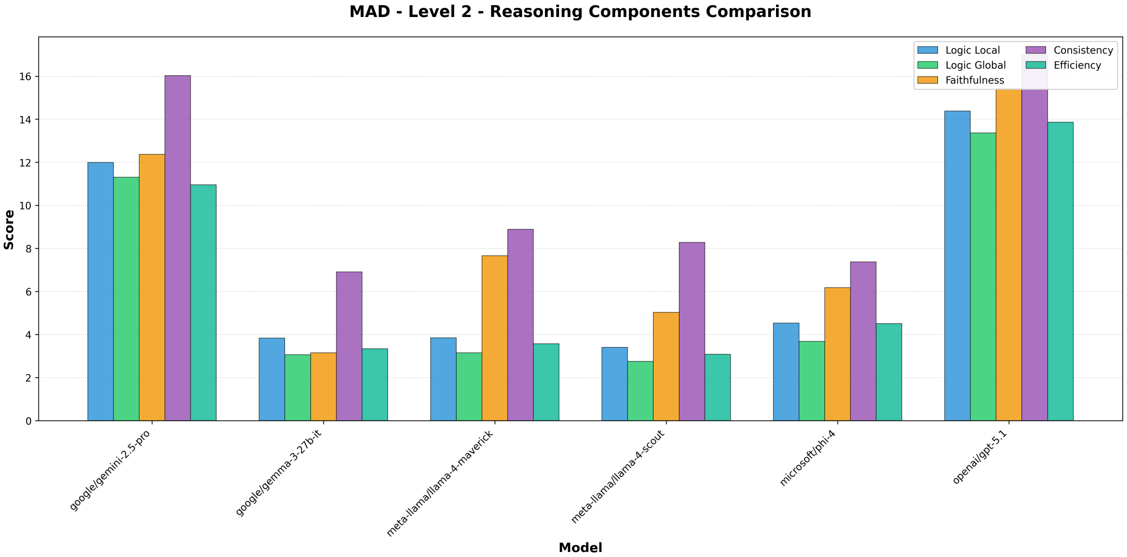

Figure 15: Multi-Agent Debate – Level 2

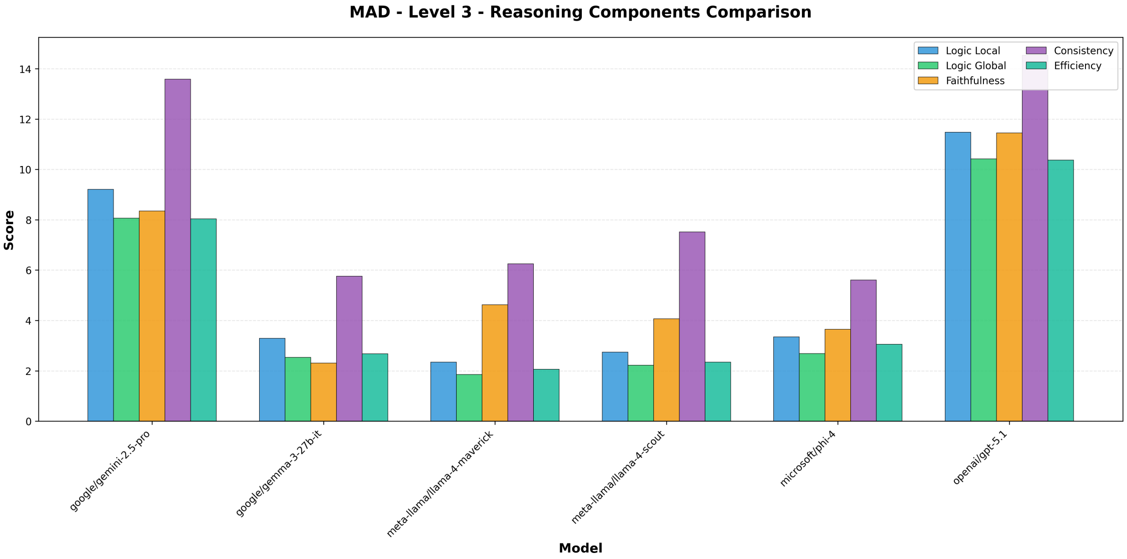

Figure 16: Multi-Agent Debate – Level 3