MASTERMINDEVAL: A SIMPLE BUT SCALABLE REASONING BENCHMARK

Jonas Golde[†] **Patrick Haller**[†] **Fabio Barth**[‡] **Alan Akbik**[†]
[†]Humboldt-Universität zu Berlin [‡]DFKI Berlin
jonas.max.golde.1@hu-berlin.de

ABSTRACT

Recent advancements in large language models (LLMs) have led to remarkable performance across a wide range of language understanding and mathematical tasks. As a result, increasing attention has been given to assessing the true reasoning capabilities of LLMs, driving research into commonsense, numerical, logical, and qualitative reasoning. However, with the rapid progress of reasoning-focused models such as OpenAI's o1 and DeepSeek's R1, there has been a growing demand for reasoning benchmarks that can keep pace with ongoing model developments. In this paper, we introduce MASTERMINDEVAL, a simple, scalable, and interpretable deductive reasoning benchmark inspired by the board game Mastermind. Our benchmark supports two evaluation paradigms: (1) agentic evaluation, in which the model autonomously plays the game, and (2) deductive reasoning evaluation, in which the model is given a pre-played game state with only one possible valid code to infer. In our experimental results we (1) find that even easy Mastermind instances are difficult for current models and (2) demonstrate that the benchmark is scalable to possibly more advanced models in the future Furthermore, we investigate possible reasons why models cannot deduce the final solution and find that current models are limited in deducing the concealed code as the number of statement to combine information from is increasing.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable performance across various text generation tasks, spanning both text and vision modalities (Grattafiori et al., 2024). These models, characterized by their large parameter counts, have proven effective in a wide range of language understanding tasks (Brown et al., 2020; Zhao et al., 2024). While LLMs have become increasingly proficient at solving complex reasoning tasks, particularly in mathematical domains, concerns persist regarding their actual ability to perform deductive reasoning. This limitation has drawn significant attention from the research community (Sinha et al., 2019; Geva et al., 2021; Dziri et al., 2023).

Several benchmarks have been developed to assess various reasoning capabilities, including LogiQA (Liu et al., 2020b), ReClor (Yu et al., 2020), and FOLIO (Han et al., 2024). However, many of these high-quality, human-generated benchmarks lack an easy and scalable method for extension. On the other hand, existing logic-based game benchmarks, such as LogicBench (Parmar et al., 2024), already incorporate board games to measure reasoning capabilities in language models. Yet, these benchmarks face several limitations. First, they often involve two-player games, where an LLM's decision-making process is heavily influenced by the opponent's moves, making it difficult to isolate and interpret the model's reasoning. Second, these benchmarks are typically not designed for scalability—games like Tic-Tac-Toe do not meaningfully increase in complexity with a larger board, while others, like Mastermind, become disproportionately harder when parameters such as code length are extended. Finally, evaluating models purely based on game outcomes fails to distinguish between genuine strategic reasoning and mere memorization of optimal moves from pretraining data. Furthermore, once these benchmarks are created, they cannot be easily adapted to accommodate future advancements in reasoning models, limiting their long-term applicability.

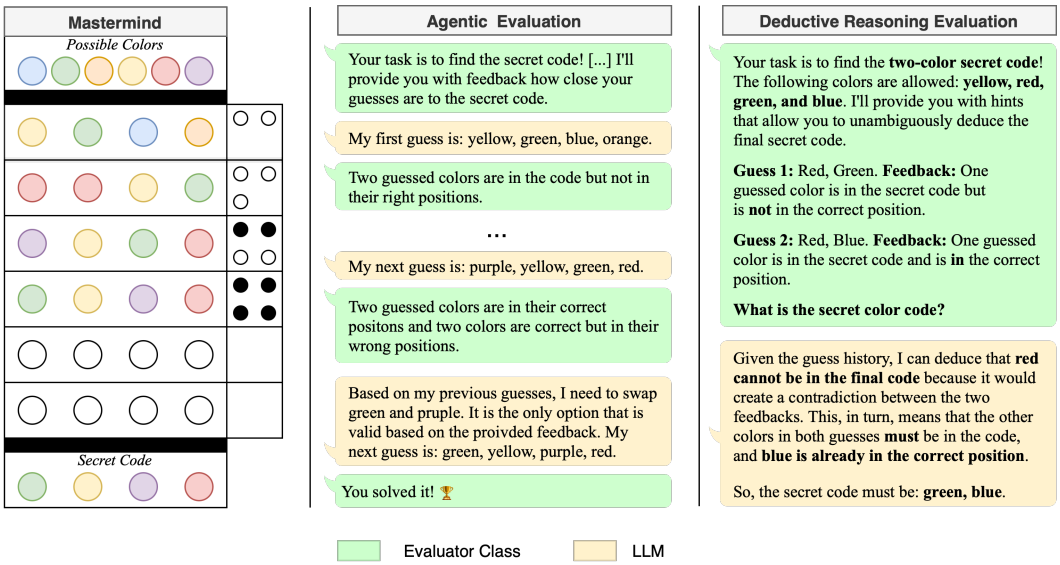

Figure 1: *(Left)* A Mastermind game instance with a secret code of length 4 and 6 possible colors, along with corresponding feedback for each guess (black and white pegs). *(Center)* Our strategy evaluation paradigm in MASTERMINDEVAL, where the LLM acts as the codebreaker in a multi-turn chat environment. *(Right)* Using Knuth's Five-Guess Algorithm, we pre-play games until only one valid code remains, then prompt the model to deduce the final answer based on the provided hints.

To address these limitations, we propose MASTERMINDEVAL, a benchmark inspired by the game of Mastermind, a logical deduction puzzle in which the player must infer a concealed color code based on feedback from previous guesses. Our benchmark provides two distinct evaluation methodologies: (1) an *agentic evaluation*, where the LLM autonomously plays the game, and (2) a *deductive reasoning evaluation*, where the LLM is tasked to infer the only valid color code in a pre-played game based on prior feedback. We present these variants in Figure 1. Additionally, we introduce a multiple-choice derivation of the deductive reasoning evaluation in which the model needs to select the correct answer based on the log-likelihood, thus performing deductive reasoning only using its pretraining objective without relying on newer paradigms such as test-time compute (Snell et al., 2024; Muennighoff et al., 2025).

We summarize our contributions as follows:

1. We introduce MASTERMINDEVAL, a reasoning benchmark based on the game of Mastermind and propose three distinct evaluation methodologies for assessing language model performance using our framework.
2. We conduct comprehensive experiments on the proposed evaluation methodologies, demonstrating that our benchmark is scalable and highlighting the current limitations of models in combining multiple pieces of information.
3. We open-source our code and datasets, providing the research community with a scalable reasoning benchmark and potential training datasets to improve reasoning capabilities [1].

2 RELATED WORK

Reasoning has long been considered a fundamental capability of language models (Liu et al., 2020a; Cobbe et al., 2021; Suzgun et al., 2023; Srivastava et al., 2023). As model size increases, reasoning abilities naturally emerge, making it a core attribute of LLMs, as demonstrated in prior work (OpenAI et al., 2024). To better utilize this capability, techniques such as chain-of-thought prompting

[1] https://github.com/flairNLP/mastermind

(Wei et al., 2023) and specialized training through reinforcement learning (Lambert et al., 2025) have been widely adopted. Most recently, many works have demonstrated the positive effect of increased test-time compute (generating more tokens at test time) to be better at reasoning tasks (Snell et al., 2024; Muennighoff et al., 2025). Another direction is multi-step reasoning that breaks down complex decision-making and planning tasks into smaller solvable tasks, which are frequently encountered in LLM-based agents (Guo et al., 2024; Xi et al., 2023).

Many benchmarks exist to assess various aspects of reasoning such as inductive reasoning (Sinha et al., 2019), temporal reasoning (Fatemi et al., 2024; Wang & Zhao, 2024), spatial reasoning (Mirzaee et al., 2021), counterfactual reasoning (Tandon et al., 2019), mathematical reasoning (Glazer et al., 2024) or logical reasoning (Han et al., 2024). (Gui et al., 2024). Additionally, knowledge-based reasoning, particularly commonsense reasoning (Talmor et al., 2019), assesses whether models can apply commonsense knowledge to make sound inferences. More advanced reasoning tasks, such as theory-of-mind reasoning, evaluate whether models can infer human thoughts and beliefs (Wagner et al., 2024).

Reasoning can also be evaluated through board games such as Chess (Campbell et al., 2002) or Go (Silver et al., 2016), which require highly complex strategic planning. While specialized models are typically trained for such games, reasoning capabilities in language models are assessed using benchmarks such as LogicGame (Gui et al., 2024), LogicBench (Parmar et al., 2024), BoardgameQA (Kazemi et al., 2023), TruthQuest (Mondorf & Plank, 2024), and GAMEBoT (Lin et al., 2024). These benchmarks introduce simplified yet challenging board game environments. However, they all share the following limitations: *(1)* They are not easily scalable in terms of complexity. For example, increasing the board size in Tic-Tac-Toe does not significantly increase its complexity, whereas Mastermind becomes considerably more difficult when using longer secret codes. *(2)* Evaluating models solely based on game outputs does not determine whether a model is genuinely reasoning and pursuing a strategy to win or merely reproducing good moves memorized from large-scale pretraining data.

3 MASTERMINDEVAL

3.1 GAMEPLAY

In the game of Mastermind, the player (the "codebreaker") attempts to deduce a hidden sequence of symbols by making successive guesses and receiving feedback. The game is defined by three key parameters: the code length c, the number of possible symbols n, and the allowed number of guesses g. A standard game configuration consists of a four-symbol code ($c = 4$) chosen from a set of six possible symbols ($n = 6$), with repetitions allowed, resulting in $6^4 = 1,296$ possible codes.

If the codebreaker's guess is (x_1, x_2, x_3, x_4) and the hidden code is (y_1, y_2, y_3, y_4), the player receives feedback indicating the accuracy of their guess based on two criteria:

1. The number of symbols that are in the correct position, i.e., the number of indices j such that $x_j = y_j$ (referred to as "black hits").
2. The number of symbols that are present in the code but in the wrong position, i.e., the number of indices j where $x_j \neq y_j$ but $x_j = y_k$ for some k, provided that y_k has not already been counted as a black hit (these are called "white hits").

As stated in Bonnet & Viennot (2016), this evaluation process can be formalized as a grading function, which is a symmetric function taking two codes as input and returning a pair of integers (b, w) as output. Given an instance (c, n) of the Mastermind game, a secret code s, and a guess g, the number of black and white hits is computed as:

$$b = \sum_{i=1}^{c} \delta(s_i, g_i), w = \max_{\tilde{g} \in \text{Perm}(g)} \left(\sum_{i=1}^{c} \delta(s_i, \tilde{g}_i) \right) - b,$$

where δ denotes the Kronecker delta function, which returns 1 if $s_i = g_i$ and 0 otherwise, and $Perm(g)$ denotes the set of all permutations of g.

The objective of the codebreaker is to achieve c black hits by iteratively refining guesses based on the feedback. This iterative process allows the player to eliminate incorrect possibilities and progressively narrow down the valid solutions. In Figure 1 (deductive reasoning view, right side), an example illustrates how feedback systematically reduces the set of possible codes until only one valid solution remains. The challenge is to identify this solution within the allotted number of guesses g.

3.2 BENCHMARK CONSTRUCTION

We employ an LLM as the codebreaker and implement a dedicated evaluator class that encapsulates the logic of Mastermind, defined by three parameters: code length c, number of possible symbols n, and the number of allowed guesses g. At the start of each game, a secret code is randomly sampled from the n^c possible combinations. The LLM is then prompted to generate an initial guess.

To ensure correct parsing, the model is instructed to clearly indicate its final guess within a structured response format. The guess is extracted using a regular expression to identify and validate the proposed sequence of symbols. It is then evaluated against the secret code, and feedback is provided based on the rules outlined in Section 3.1 (see also Figure 4 for reference). Each interaction between the model and the game environment is appended to the overall context, which is provided to the model so that it can consider all previous reasoning when generating subsequent guesses. In the following sections, we introduce the two evaluation paradigms in which LLMs are assessed using MASTERMINDEVAL.

3.3 AGENTIC EVALUATION

The first evaluation paradigm presents the classical Mastermind game, where the LLM assumes the role of the codebreaker, tasked with deducing a concealed code through iterative (multi-turn) reasoning. The game continues until the model correctly identifies the secret code or reaches a predefined limit on the number of allowed guesses.

Success in this setting depends on refining strategies to maximize information gain with each move. Algorithmic approaches such as those proposed by Knuth (1977) and Kooi (2005) effectively balance exploration—formulating diverse guesses to extract informative feedback—and exploitation—leveraging prior feedback to make optimal deductions within known constraints. Strong performance in this paradigm demonstrates an LLM's ability to perform logical elimination, engage in adaptive reasoning, and efficiently refine hypotheses, all of which are crucial for structured problem-solving tasks.

3.4 DEDUCTIVE REASONING EVALUATION

The agentic evaluation paradigm assesses two distinct capabilities of language models: (1) deducing the final code based on prior guesses and (2) selecting optimal guesses to reach the solution as efficiently as possible. To isolate and evaluate a model's deductive reasoning ability, we introduce a separate evaluation paradigm in which the guessing strategy is predefined.

To achieve this, we construct a dataset comprising over 30,000 pre-played game states across various game configurations, generated using Knuth's algorithm. Each game progresses until only a single valid code remains based on previous feedback. Unlike the strategy evaluation setting, where models must determine their next move, this paradigm exclusively tests their ability to infer the correct solution given a fixed sequence of prior guesses. The task evaluates whether an LLM can accurately synthesize available information and deduce the unique remaining solution under given constraints.

Knuth's Five-Guess Algorithm. Knuth's algorithm (Knuth, 1977) solves the Mastermind game optimally in at most five guesses using a minimax strategy. Given a secret code s from a set $\mathcal{S}$, the algorithm iteratively selects guesses $g \in \mathcal{S}$ to minimize the worst-case number of remaining possibilities:

$$g^* = \arg\min_{g \in \mathcal{S}} \max_{\substack{s' \in \mathcal{S} \\ f(g,s')=f(g,s)}} |\mathcal{S}'|$$

where $f(g, s)$ denotes the generalized feedback function from Section 3.1, and $\mathcal{S}'$ represents the reduced solution space remaining after incorporating feedback $f(g, s)$. The process begins with an optimal first guess (e.g., $(1, 1, 2, 2)$) and continues until only a single valid code remains ($|\mathcal{S}| = 1$), ensuring a solution within at most five steps.

3.5 Multiple-Choice Evaluation

At last, we derive two multiple-choice versions of the deductive reasoning benchmark, inspired by datasets commonly used in the lm-eval-harness (Gao et al., 2023) project. In this setting, models rely solely on their pretraining objective to determine the concealed code. We implement two variations: (1) a version where incorrect answers are randomly generated codes, and (2) a version where incorrect answers differ from the correct solution by only one symbol. To ensure fairness, no answer option appears in the previous guesses within the prompt.

Log-Likelihood Prediction. Given a generated statement, we compute the log-likelihood score $\log \hat{p}(a_i \mid t)$, where t represents the prompt and a_i is an answer candidate. Since the template is the same across all answer options, ranking them is sufficient, allowing us to express the log-likelihood as:

$$\log \hat{p}(a_i \mid t) = \log \hat{p}(a_i, t) - \log \hat{p}(t).$$

As causal language models predict token-level log-likelihoods based on prior context, the overall log-likelihood of a sentence is simply the sum of the token-wise log-likelihoods.

4 Experimental Setup

We evaluate a diverse set of open-source models on our benchmark, including Qwen-2/2.5 (Qwen et al., 2025), Llama-3.1/3.2 (Grattafiori et al., 2024), Phi-3.5/4 (Abdin et al., 2024a;b), and distilled Qwen-2.5 models from DeepSeek-R1 (DeepSeek-AI et al., 2025). In our tables, we denote these models as Qwen-2.5 (R1), covering a parameter range from 1.5B to 7B. All experiments are conducted on four Nvidia H100 GPUs with 80GB of memory. Additionally, we report results for proprietary GPT models, including GPT-4o, GPT-4o-mini, and its reasoning variant, o3-mini (OpenAI et al., 2024).

For the agentic and deductive reasoning evaluations, each model plays 100 games, and we report the solve rate, defined as the fraction of games in which the model correctly identifies the concealed code within the allowed number of guesses. Games where a model correctly guesses the concealed code on the first attempt are excluded from the evaluation. For example, in easy game configurations such as $(c = 2, n = 4)$, the probability of guessing the concealed code on the first attempt is approximately 6.3%. We re-play these games to ensure that every model plays the same set of games. For the deductive reasoning evaluation, we use the same setup, except that the model is limited to a single attempt to solve the game. The only exception is the multiple-choice evaluation, where we assess performance across all 30,000+ instances.

5 Results

5.1 Agentic Evaluation

Larger models show better reasoning capabilities. Table 1 presents model performance across different game configurations, highlighting trends in capability scaling and efficiency. GPT-4o, GPT-4o-mini, and o3-mini consistently achieve the highest accuracy, performing near-perfectly in simpler settings ($c = 2, n = 4$) and maintaining strong results even as task complexity increases. Notably, o3-mini, the dedicated reasoning model, achieves high solve rates (over 0.85) even in more challenging settings ($c = 5, n = 7$). Although not shown in the table, we find that o3-mini's performance declines in more difficult settings, solving 60% of games when $c = 6, n = 8$ and only 29% when $c = 7, n = 9$.

Among open-source models, Phi-4 demonstrates the best overall performance, outperforming all other models. However, while it performs well in simpler settings, its accuracy approaches zero

Model	Params	Game Config $(c=2, n=4)$	$(c=3, n=5)$	$(c=4, n=6)$	$(c=5, n=7)$
Qwen-2	1.5B	0.25	0.03	0.01	0.00
Qwen-2.5	3B	0.69	0.12	0.02	0.00
	7B	<u>0.73</u>	0.18	0.04	0.00
Llama-3.2	3B	0.56	0.13	0.04	0.01
Llama-3.1	8B	0.58	<u>0.31</u>	<u>0.07</u>	0.01
Phi-3.5 (mini)	3.8B	0.34	0.16	0.00	0.00
Phi-4	14B	**0.85**	**0.52**	**0.30**	**0.07**
Qwen-2.5 (R1)	1.5B	0.26	0.06	0.00	0.00
	7B	0.52	0.17	0.03	<u>0.03</u>
4o-mini	-	0.99	0.73	0.34	0.10
4o	-	1.00	0.77	0.33	0.02
o3-mini	-	1.00	1.00	0.86	0.92

Table 1: *(Agentic Evaluation)* The performance of LLMs playing and solving various game configurations of Mastermind. We present open-source models in the upper part and results for proprietary GPT models in the lower part. We highlight the best scores in bold and the second-best are underlined.

as game complexity increases. Furthermore, we observe no significant difference between Qwen-2.5 and its distilled version fine-tuned using DeepSeek-R1, suggesting that the reasoning ability of larger models may not yet be effectively distilled into smaller models.

Test-time compute increases with task complexity. Additionally, in Figure 2, we compare the test-time compute of o3-mini and Phi-4. We observe that both models increase their test-time compute as task complexity grows. However, o3-mini exhibits a much stronger adaptive response, allocating significantly more compute to more difficult tasks. This suggests that o3-mini can reflect on task complexity and dynamically adjust its compute usage to achieve better results, even in the most complex settings (see Table 1).

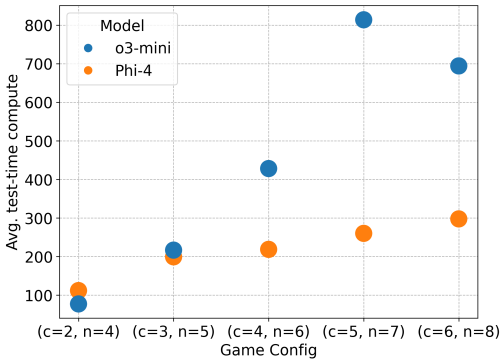

Figure 2: Test-time compute increases as the model complexity increases indicating MASTERMINDEVAL can be easily scaled by increasing the length of the secret code c and the number of possible colors/symbols n.

Models often fail to deduce the secret code, even after eliminating all other possibilities. We further report scores for "perfect games" in Table 5, defined as games in which a model has already reduced the remaining state space to a single possible code and then deduces the correct answer in the next guess. We observe that many models struggle to achieve perfect games, often requiring additional guesses even when only one valid code remains. This suggests that models may rely on local optimization rather than a holistic strategy. However, o3-mini consistently achieves high perfect game rates (over 0.78) across various settings, indicating that models explicitly trained for reasoning can learn to think more systematically.

5.2 DEDUCTIVE REASONING EVALUATION

Among open-source models, Phi-4 achieves the highest success rates across all configurations, with scores of 0.50, 0.12, and 0.03, respectively. Llama-3 (8B) and DeepSeek-R1 (7B) also demonstrate competitive performance, ranking among the top models in certain configurations.

Model	Params	Game Config ($c = 2, n = 4$)	($c = 3, n = 5$)	($c = 4, n = 6$)
Qwen-2.5	3B	0.09	0.02	0.01
	7B	0.25	0.04	<u>0.02</u>
Llama-3.2	3B	0.08	0.04	**0.03**
Llama-3.1	8B	0.13	<u>0.05</u>	0.01
Phi-3.5 (mini)	3.8B	0.14	0.02	0.01
Phi-4	14B	**0.50**	**0.12**	**0.03**
Qwen-2.5 (R1)	1.5B	0.07	0.00	0.00
	7B	<u>0.49</u>	<u>0.05</u>	0.00
GPT-4o-mini	-	0.34	0.09	0.00
GPT-4o	-	0.62	0.14	0.03
o3-mini	-	0.99	1.00	0.99

Table 2: *(Deductive Reasoning Evaluation)* The performance of LLMs deducing the secret code in pre-played games. We present results for open-source models in the upper part and proprietary GPT models in the lower part. We indicate best scores in bold and second-best scores underlined.

Among proprietary models, o3-mini consistently outperforms all others across every configuration, achieving the highest solve rates. GPT-4o-mini also performs well but lags slightly behind GPT-4o.

Notably, solve rates in this setting are lower compared to the agentic evaluation, indicating that models often fail to deduce the secret code even when doing so is theoretically possible - since the dataset is explicitly designed to ensure solvability. One possible explanation is that models require additional guesses to correctly infer the final code.

Models struggle when reasoning over a larger amount of information. Figure 3 illustrates the number of hints required to deduce the secret code. We observe that models with generally strong performance achieve a high solve rate when only a single hint is provided. However, this rate drops significantly and approaches zero when the model must infer the final code using four guesses. While o3-mini successfully solves all states, we conclude that models explicitly trained for reasoning can effectively integrate information across multiple hints, whereas classical instruction-tuned models struggle in more complex settings.

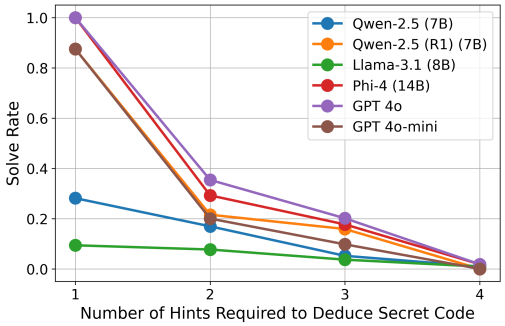

Figure 3: We cluster the solve rate based on the number of hints required to deduce the secret code. We observe that the solve rate approaches 0 when the final code can only be deduced after four chat interactions, even for capable models such as GPT-4o.

5.3 Multiple Choice Evaluation

Table 3 and Table 4 present the performance of various language models in the multiple-choice evaluation setting, where models rank answer options based on log-likelihood. The two tables correspond to different types of incorrect answers: Table 3 features randomly generated codes as distractors, whereas Table 4 presents more challenging distractors that differ from the correct code by only a single pin or color.

Performance Across Evaluation Settings. Across both evaluation settings, larger models generally outperform smaller ones, but certain mid-sized models, such as Phi-3.5 (mini) and Qwen-2.5 (7B), achieve competitive results. In particular, Phi-3.5 (mini) achieves the highest scores in most configurations, while Phi-4 performs best in the $(c = 3, n = 5)$ setting and ranks second in the others.

Model	Params	Game Config ($c = 2, n = 4$)	($c = 3, n = 5$)	($c = 4, n = 6$)
Qwen-2.5	3B	0.22	0.43	0.60
	7B	0.31	0.48	0.70
Llama-3.2	3B	0.23	0.40	0.56
Llama-3.1	8B	0.30	0.42	0.60
Phi-3.5 (mini)	3.8B	<u>0.38</u>	**0.52**	**0.72**
Phi-4	14B	**0.44**	<u>0.50</u>	<u>0.67</u>
Qwen-2.5 (R1)	1.5B	0.30	0.45	0.65
	7B	0.26	0.39	0.49

Table 3: *(Multiple-Choice Evaluation)* Performance of language models in multiple choice evaluation using log-likelihood-based ranking and **random codes as wrong answer options**. We indicate best scores in bold and second-best scores underlined.

Model	Params	Game Config ($c = 2, n = 4$)	($c = 3, n = 5$)	($c = 4, n = 6$)
Qwen-2.5	3B	0.23	0.42	0.53
	7B	0.37	0.48	0.57
Llama-3.2	3B	0.26	0.42	0.53
Llama-3.1	8B	0.35	0.45	0.56
Phi-3.5 (mini)	3.8B	<u>0.38</u>	<u>0.50</u>	**0.61**
Phi-4	14B	**0.49**	**0.51**	**0.61**
Qwen-2.5 (R1)	1.5B	0.35	0.43	0.58
	7B	0.27	0.38	0.41

Table 4: *(Multiple-Choice Evaluation)* Performance of language models in multiple choice evaluation using log-likelihood-based ranking and **close codes as wrong answer options**. The wrong answers differ only in one pin compared to the concealed code. We indicate best scores in bold and second-best scores underlined.

A key observation is that model performance is consistently higher when incorrect answers are randomly generated codes rather than subtly modified versions of the correct answer. This suggests that eliminating entirely incorrect codes is relatively easy for language models, whereas distinguishing between highly similar codes is more challenging. This aligns with expectations, as small perturbations in answer choices require models to perform fine-grained probabilistic reasoning rather than relying on obvious likelihood differences.

Effect of Game Complexity. Interestingly, we observe an increase in model accuracy as game complexity increases. This trend may seem counterintuitive, as solving more complex instances should generally be harder. However, this phenomenon suggests that in highly constrained settings, language models can leverage statistical regularities from pretraining to make well-informed eliminations. The structured nature of the problem may allow models to reason more effectively using implicit knowledge, even in the absence of explicit step-by-step deductive reasoning.

Nonetheless, it is important to emphasize that this task is inherently different from the agentic evaluation, where models must actively deduce the concealed code by selecting optimal moves. In the multiple-choice setting, the reasoning process is simplified, as models only need to rank pre-given options rather than formulating a search strategy. Thus, while strong performance in this task highlights the effectiveness of pretraining for probabilistic inference, it does not necessarily translate to strategic reasoning capabilities required for interactive gameplay.

Implications for Model Design. These findings suggest that while larger models tend to generalize better, well-optimized mid-sized models can achieve comparable performance in constrained reasoning tasks. Moreover, the challenge of distinguishing between similar answers points to the need for improved fine-grained reasoning mechanisms, possibly through better representation learning or explicit reasoning objectives during training. Future work could explore whether additional

reasoning-specific fine-tuning could bridge the performance gap between purely instruction-tuned models and those explicitly trained for structured problem-solving.

6 CONCLUSION

We introduce a benchmark MASTERMINDEVAL based on the game of Mastermind to evaluate language models in both strategic gameplay and deductive reasoning. Our results show that model performance declines as game complexity increases, revealing fundamental limitations in multi-step reasoning. Even in simple settings, models struggle to deduce the final code and integrate multiple hints effectively.

Despite these challenges, models perform well in eliminating unlikely answers, likely due to their pretraining objective. However, this ability does not translate into robust logical reasoning in structured problem-solving tasks. These findings suggest that while current models exploit statistical patterns efficiently, significant advancements are needed to develop structured and strategic reasoning capabilities.

ACKNOWLEDGMENTS

We thank all reviewers for their valuable comments. Jonas Golde is supported by the Bundesministerium für Bildung und Forschung (BMBF) as part of the project "FewTuRe" (project number 01IS24020). Alan Akbik and Patrick Haller are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Emmy Noether grant "Eidetic Representations of Natural Language" (project number 448414230). Further, Alan Akbik is supported under Germany's Excellence Strategy "Science of Intelligence" (EXC 2002/1, project number 390523135). Fabio Barth is supported by the Bundesministerium für Wirtschaft und Energie (BMWi) as part of the project "OpenGPT-X" (project number 68GX21007D).

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. Phi-3 technical report: A highly capable language model locally on your phone, 2024a. URL https://arxiv.org/abs/2404.14219.

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024b. URL https://arxiv.org/abs/2412.08905.

François Bonnet and Simon Viennot. Nash equilibrium in mastermind. In *International Conference on Computers and Games*, pp. 115–128. Springer, 2016.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Murray Campbell, A.Joseph Hoane, and Feng hsiung Hsu. Deep blue. *Artificial Intelligence*, 134(1):57–83, 2002. ISSN 0004-3702. doi: https://doi.org/10.1016/S0004-3702(01)00129-1. URL https://www.sciencedirect.com/science/article/pii/S0004370201001291.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai

Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang (Lorraine) Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena Hwang, Soumya Sanyal, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 70293–70332. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/deb3c28192f979302c157cb653c15e90-Paper-Conference.pdf.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for evaluating llms on temporal reasoning, 2024. URL https://arxiv.org/abs/2406.09170.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023. URL https://zenodo.org/records/10256836.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361, 2021. doi: 10.1162/tacl_a_00370. URL https://aclanthology.org/2021.tacl-1.21/.

Elliot Glazer, Ege Erdil, Tamay Besiroglu, Diego Chicharro, Evan Chen, Alex Gunning, Caroline Falkman Olsson, Jean-Stanislas Denain, Anson Ho, Emily de Oliveira Santos, Olli Järviniemi, Matthew Barnett, Robert Sandler, Matej Vrzala, Jaime Sevilla, Qiuyu Ren, Elizabeth Pratt, Lionel Levine, Grant Barkley, Natalie Stewart, Bogdan Grechuk, Tetiana Grechuk, Shreepranav Varma Enugandla, and Mark Wildon. Frontiermath: A benchmark for evaluating advanced mathematical reasoning in ai, 2024. URL https://arxiv.org/abs/2411.04872.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius,

Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell,

Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Jiayi Gui, Yiming Liu, Jiale Cheng, Xiaotao Gu, Xiao Liu, Hongning Wang, Yuxiao Dong, Jie Tang, and Minlie Huang. Logicgame: Benchmarking rule-based reasoning abilities of large language models, 2024. URL https://arxiv.org/abs/2408.15778.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL https://arxiv.org/abs/2402.01680.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Wenfei Zhou, James Coady, David Peng, Yujie Qiao, Luke Benson, Lucy Sun, Alexander Wardle-Solano, Hannah Szabó, Ekaterina Zubova, Matthew Burtell, Jonathan Fan, Yixin Liu, Brian Wong, Malcolm Sailor, Ansong Ni, Linyong Nan, Jungo Kasai, Tao Yu, Rui Zhang, Alexander Fabbri, Wojciech Maciej Kryscinski, Semih Yavuz, Ye Liu, Xi Victoria Lin, Shafiq Joty, Yingbo Zhou, Caiming Xiong, Rex Ying, Arman Cohan, and Dragomir Radev. FOLIO: Natural language reasoning with first-order logic. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22017–22031, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1229. URL https://aclanthology.org/2024.emnlp-main.1229/.

Mehran Kazemi, Quan Yuan, Deepti Bhatia, Najoung Kim, Xin Xu, Vaiva Imbrasaite, and Deepak Ramachandran. Boardgameqa: A dataset for natural language reasoning with contradictory information. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 39052–39074. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/7adce80e86aa841490e6307109094de5-Paper-Datasets_and_Benchmarks.pdf.

Donald Ervin Knuth. The computer as master mind. *Journal of Recreational Mathematics*, 9:1–6, 1977. URL https://api.semanticscholar.org/CorpusID:201271993.

Barteld Kooi. Yet another mastermind strategy. *ICGA Journal*, 28(1):13–20, 2005.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. Tulu 3: Pushing frontiers in open language model post-training, 2025. URL https://arxiv.org/abs/2411.15124.

Wenye Lin, Jonathan Roberts, Yunhan Yang, Samuel Albanie, Zongqing Lu, and Kai Han. Beyond outcomes: Transparent assessment of llm reasoning in games, 2024. URL https://arxiv.org/abs/2412.13602.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3622–3628. International Joint Conferences on Artificial Intelligence Organization, 7 2020a. doi: 10.24963/ijcai.2020/501. URL https://doi.org/10.24963/ijcai.2020/501. Main track.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning, 2020b. URL https://arxiv.org/abs/2007.08124.

Roshanak Mirzaee, Hossein Rajaby Faghihi, Qiang Ning, and Parisa Kordjamshidi. SPARTQA: A textual question answering benchmark for spatial reasoning. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4582–4598, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.364. URL https://aclanthology.org/2021.naacl-main.364/.

Philipp Mondorf and Barbara Plank. Liar, liar, logical mire: A benchmark for suppositional reasoning in large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7114–7137, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.404. URL https://aclanthology.org/2024.emnlp-main.404/.

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

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook

Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Mihir Parmar, Nisarg Patel, Neeraj Varshney, Mutsumi Nakamura, Man Luo, Santosh Mashetty, Arindam Mitra, and Chitta Baral. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13679–13707, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.739. URL https://aclanthology.org/2024.acl-long.739/.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503, 2016. URL http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 4506–4515, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1458. URL https://aclanthology.org/D19-1458/.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling model parameters, 2024. URL https://arxiv.org/abs/2408.03314.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy

Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=uyTL5Bvosj. Featured Certification.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.824. URL https://aclanthology.org/2023.findings-acl.824/.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge, 2019. URL https://arxiv.org/abs/1811.00937.

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. WIQA: A dataset for "what if..." reasoning over procedural text. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 6076–6085, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1629. URL https://aclanthology.org/D19-1629/.

Eitan Wagner, Nitay Alon, Joseph M. Barnby, and Omri Abend. Mind your theory: Theory of mind goes deeper than reasoning, 2024. URL https://arxiv.org/abs/2412.13631.

Yuqing Wang and Yun Zhao. TRAM: Benchmarking temporal reasoning for large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6389–6415, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.382. URL https://aclanthology.org/2024.findings-acl.382/.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. The rise and potential of large language model based agents: A survey, 2023. URL https://arxiv.org/abs/2309.07864.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. Reclor: A reading comprehension dataset requiring logical reasoning, 2020. URL https://arxiv.org/abs/2002.04326.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models, 2024. URL https://arxiv.org/abs/2303.18223.

A Appendix

A.1 Game Prompt

Your goal is to guess the secret color code like in the game of Mastermind.
The game is defined as follows:

- The code consists out of c colors , and duplicates are allowed.
- The following colors are allowed: n.
- You have a total of 12 guesses.

After each guess, you will receive feedback based on the following rules:

1. You will be told how many pegs are the correct color **and** in the correct position.
2. You will be told how many pegs are the correct color **but** in the wrong position.
3. A color in the guess will only be counted once, prioritizing exact matches before partial ones.

Use the feedback from all your guesses to logically deduce the secret code.
To indicate your final answer, clearly prepend 'FINAL GUESS:' to your last submission.

Example interaction:
<User> What's your next guess?
<Assistant> FINAL GUESS: green, yellow, black, blue.
<User> Feedback: <number> color(s) are in the correct position(s). <number> color(s) are in the concealed code but in wrong position(s).
###

Before giving your next guess, analyze your next guess using all previous feedback step-by-step. What's your next guess?

Figure 4: The initial prompt used to play a game of Mastermind. We append each answer of the LLM to the prompt such that LLM are enabled to reflect on their own thinking in future steps.

Model	Params	Game Config $(c = 2, n = 4)$	$(c = 3, n = 5)$	$(c = 4, n = 6)$	$(c = 5, n = 7)$
Qwen-2.5	1.5B	0.00	0.03	0.00	0.05
Qwen-2.5	3B	0.06	0.01	0.03	0.09
Qwen-2.5	7B	0.12	0.02	0.03	0.15
DeepSeek-R1	1.5B	0.05	0.08	0.00	0.00
DeepSeek-R1	7B	0.09	0.03	0.04	0.06
DeepSeek-R1	14B	0.26	0.00	-	-
Llama-3.1	8B	0.10	0.06	0.04	0.10
Llama-3.2	3B	0.15	0.03	0.09	0.00
Phi-3.5-mini	3.8B	0.12	0.07	0.03	0.05
Phi-4	14B	0.42	0.21	0.10	0.09
GPT-4o	-	0.67	0.17	0.09	0.05
GPT-4o-mini	-	0.33	0.12	0.08	0.02
o3-mini	-	0.96	0.86	0.78	0.89

Table 5: *(Perfect Deduction)* We report the fraction of models that achieve "perfect games" in the agentic evaluation paradigm. A perfect game is defined as a sequence of guesses that systematically reduce the state space until only a single possible code remains, followed by a correct deduction on the next guess. If the model requires additional attempts beyond this step, the game is not considered perfect.

A.2 Perfect Games In Agentic Evaluation

B Limitations and Future Work

While our benchmark provides valuable insights into the reasoning capabilities of language models through the game of Mastermind, our evaluation has several limitations. Due to computational constraints, we were only able to systematically assess smaller open-source models, restricting our analysis to models that could be run efficiently on available hardware. Additionally, we did not train any models; however, methods such as reinforcement learning via feedback (RLVF) (Lambert et al., 2025) could potentially improve both strategic play and deductive reasoning in language models.

Another limitation lies in the scope of qualitative information measurement. While we can track the remaining state space in each game configuration using Knuth's algorithm, contributions from other research areas, such as game theory and psychology, could provide a more fine-grained understanding of how models process structured information. In particular, a more precise quantification of the informational content in each hint could further refine our analysis of LLM reasoning capabilities.

Furthermore, we cannot determine whether models are truly reasoning in MastermindEval or merely relying on pattern-matching learned during pretraining. The paper does not establish whether games of Mastermind are present in the pretraining data, leaving open the possibility that performance is influenced by prior exposure rather than reasoning ability. Verifying this through controlled experiments or data ablation would help clarify the extent to which MastermindEval measures genuine reasoning. Additionally, the study does not analyze how results in MastermindEval correlate with established reasoning benchmarks, making it difficult to assess its validity as a reasoning metric.

Another limitation is the absence of human evaluation. Since performance in MastermindEval appears to correlate with model size, it would be worth investigating whether it also aligns with measures of human intelligence. One straightforward approach would be to test whether people with higher educational attainment perform better in the task. Exploring this connection could provide additional support for MastermindEval as a reasoning benchmark. Future work incorporating human evaluations would help determine whether the task meaningfully reflects reasoning ability beyond model scaling trends.