ZEROSUMEVAL: SCALING LLM EVALUATION WITH INTER-MODEL COMPETITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating the capabilities of Foundation Models has traditionally relied on static benchmark datasets, human assessments, or model-based evaluations — methods that often suffer from overfitting, high costs, and biases. We introduce ZeroSumEval, a novel competition-based evaluation protocol that leverages zero-sum games to assess LLMs with dynamic benchmarks that resist saturation. ZeroSumEval encompasses a diverse suite of games, including security challenges (Capture the Flag), classic board games (chess), and knowledge tests (MathQuiz). These games are designed to evaluate a range of AI capabilities such as strategic reasoning, planning, knowledge application, safety, and adaptability. A key novelty is integrating automatic prompt optimization to ensure fair comparisons by eliminating biases from human prompt engineering and support arbitrary prompting strategies. Furthermore, ZeroSumEval measures AI models' abilities to self-improve from limited observations and assesses their robustness against adversarial or misleading examples during prompt optimization. Building upon recent studies that highlight the effectiveness of game-based evaluations for LLMs, ZeroSumEval enhances these approaches by providing a standardized and extensible framework for rigorous assessment. We find ZeroSumEval correlates strongly with expensive human evaluations (Chatbot Arena) and disagrees with benchmarks with known overfitting and saturation issues. Inspecting match traces reveals models that allocate more tokens to thought processes perform strongly in games involving planning capabilities.

1 INTRODUCTION

Large Language Models (LLMs) are being developed at an unprecedented pace (Zhao et al., 2024), requiring significant investment for their training and refinement (Kevin Lee, 2024; Miller, 2022; Kimball, 2024). As the performance and complexity of these models continue to grow (Chen et al., 2024b), selecting the most appropriate model for a specific application has become an increasingly challenging and costly decision(Kaplan et al., 2020; Hoffmann et al., 2022). Benchmarking emerges as a critical tool in this context (Laskar et al., 2023; Qin et al., 2023), providing standardized metrics and evaluations to guide these choices.

With the rapid growth of generative technologies built on top of Large Language Models (OpenAI, 2022; Google, 2024; Anthropic, 2024b; Ormazabal et al., 2024; Mistral, 2024; Dubey et al., 2024a; Yang et al., 2024), it has been increasingly difficult to evaluate these models comprehensively (Guo et al., 2023). Current benchmarking practices face several significant issues. Many benchmarks suffer from data contamination (Yang et al., 2023), where models inadvertently train on portions of the test data (Dubey et al., 2024a; Groeneveld et al., 2024), leading to inflated performance metrics. Sensitivity to prompt variations (Alzahrani et al., 2024b) and a lack of diversity in evaluation tasks (Laskar et al., 2024) further undermine the reliability and robustness of these benchmarks. Additionally, the high cost and effort required to develop new benchmarks often result in outdated evaluation methods that do not keep pace with the rapid development of LLMs (Kiela et al., 2021; Vu et al., 2023).

An observed disparity exists between the computational resources measured in floating-point operations per second, or FLOPs used to train LLMs and those allocated for their evaluation. Training these models involves massive computational efforts (Hoffmann et al., 2022), yet the evaluation

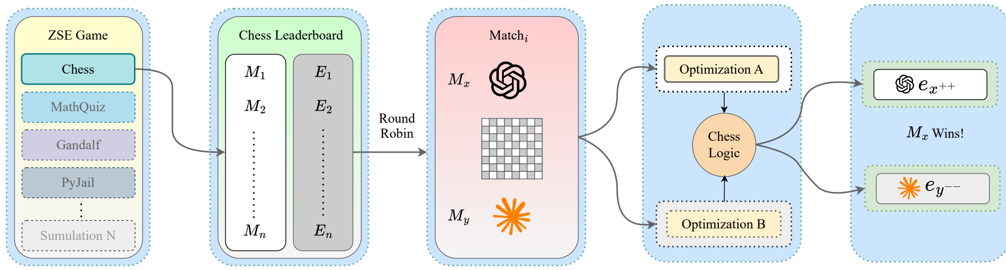

Figure 1: The ZeroSumEval suite of benchmarks provides dynamic simulations with head to head model competition to create robust and scalable model evaluations and leaderboards. Integrated automatic prompt optimization minimizes biases introduced by prompting and hand-engineering.

phase typically utilizes a negligible fraction of this capacity (Laskar et al., 2024). Scaling up evaluation by increasing the number of evaluation tokens is essential for a more thorough understanding of model capabilities. Traditionally, this scaling involves incorporating human-crafted independent and identically distributed (i.i.d.) data (Holland et al., 2018), which is resource-intensive (Hutchinson et al., 2021) and may not adequately capture the complexities of language (Mehrabi et al., 2021) and reasoning required to challenge advanced LLMs (Gudibande et al., 2023) or even LLM generated (Karpinska et al., 2021).

Previous work has proposed the use of games as benchmarks (Topsakal et al., 2024), offering a promising avenue for evaluating complex reasoning (Wong et al., 2023) and decision-making abilities of LLMs (Warstadt et al., 2023; Park et al., 2023; Wang et al., 2023). Games provide interactive and dynamic environments that can test models beyond static datasets. However, existing game-based benchmarks are often *(i)* inflexible and limited in scope, *(ii)* not easily extendable, *(iii)* restricted in their effectiveness for comprehensive model evaluation, and *(iv)* depend on predefined prompts.

Scaling evaluation is fundamental not only for assessing performance but also for uncovering hidden dynamics within LLMs, such as potential backdoors or biases (Schuster et al., 2020), and for evaluating their emerging reasoning capabilities (Brown et al., 2020; Sanh et al., 2022; Wei et al., 2023b;a). Implementing environments for simulations or games offers a scalable solution to these challenges (OpenAI et al., 2019; OpenAI, 2019; Silver et al., 2016; 2017; Zheng et al., 2021).

Existing evaluation protocols possess several key issues:

(i) **Prompt Sensitivity**: Previous work (Zheng et al., 2024; Pezeshkpour & Hruschka, 2023; Lu et al., 2022; Alzahrani et al., 2024a; Wang et al., 2024a) has shown that models are sensitive to benchmark formats. By sheer chance, a model could be presented with a prompting method that's either favorable or detrimental. These prompt modifications are shown to result in substantially different relative performance between models (Alzahrani et al., 2024a). By testing models in varied scenarios within a controlled environment, we can assess and improve their robustness to different prompts. Crucially, different models are not optimized for the same prompts due to variations in data mixtures and algorithmic implementations. Using identical prompts across all models may therefore lead to unfair comparisons.

(ii) **Limited Diversity**: Traditional evaluation methods often rely on static datasets, which are inherently limited by their dependency on human curation and annotation. This makes it challenging to continuously introduce new, diverse test data. An extensible simulated environment, however, allows for a wide array of dynamically generated games and scenarios, enhancing the diversity and scalability of evaluation tasks.

(iii) **Extensibility**: Once established, the environment can be easily expanded to include new games, rules, and scenarios, facilitating continuous evaluation improvements.

(iv) **Crowd and Annotator Bias**: LLM evaluations conducted by large crowds often tend to be susceptible to social hacking, and it can depend on geographic, temporal, and narrative factors Gururan-

gan et al. (2018). Controlled and interpretable environments can mitigate these biases by providing consistent, objective evaluation criteria.

(v) **Saturation**: With the rapid improvement of LLMs, evaluation benchmarks quickly become obsolete and saturated, with frontier models achieving almost perfect scores, which necessitates the development of new benchmarks. On the opposite extreme, benchmarks that are too difficult would result in almost random scores. Both extremes result in a lack of granularity to distinguish models. Therefore, benchmarks posing moderate difficulty to frontier models will need to be continuously developed as models improve. For instance, GSM8K (Cobbe et al., 2021) tests models on grade school-level math, and most state-of-the-art models achieve scores above 90% (Dubey et al., 2024b; Anthropic, 2024a). Thus, the more difficult MATH (Hendrycks et al., 2021b) dataset, which consists of math competition questions, was developed and is now commonly used[1]. A similar trend is observed in academic examination benchmarks with the migration from MMLU (Hendrycks et al., 2020) to MMLU-Pro (Wang et al., 2024b) and GPQA. (Rein et al., 2023)[2].

To address these challenges, we introduce ZEROSUMEVAL, a flexible and extensible open-source framework designed to scale LLM evaluation through the simulation of two-player zero-sum games. Our framework allows for comprehensive and robust assessment by providing models with multiple opportunities to make legal moves, thereby accommodating occasional errors and offering a more nuanced understanding of their capabilities.

1. **Scaling Evaluation by Simulation**: We demonstrate how simulation environments can effectively scale the evaluation process.
2. **Flexible and Extensible Framework**: ZEROSUMEVAL is designed to be adaptable, allowing researchers and practitioners to customize and extend the evaluation environment to suit diverse needs.
3. **Robustness to Prompt Sensitivity**: By incorporating automatic prompt optimization, our framework mitigates issues related to prompt sensitivity, leading to more reliable evaluation outcomes.
4. **Enhanced Interpretability**: The structured environment facilitates easier interpretation of model behaviors, aiding in the identification of strengths and weaknesses.
5. **Error Accommodation**: Models are given multiple chances to make legal moves, ensuring that occasional missteps due to inherent stochasticity do not disproportionately affect the overall evaluation.

2 RELATED WORK

2.1 STATIC LLM BENCHMARKS

Until recently, LLMs were evaluated on Natural Language Understanding (NLU) tasks from benchmark collections like GLUE (Wang, 2018) and SuperGLUE (Wang et al., 2019), which included tasks like paraphrase classification and sentiment analysis. As LLMs developed, they acquired emergent capabilities beyond generating plausible text, such as reasoning, generating code, and instruction following (Brown et al., 2020; Wei et al., 2022). With these newly found capabilities, new benchmarks were developed to quantify these abilities. As models improve, more difficult benchmarks are created. For example:

- **Reasoning:** undergraduate level academic questions are tested via MMLU (Hendrycks et al., 2020), while GPQA (Rein et al., 2023) tests models with graduate level questions. All aforementioned benchmarks score models based on the likelihood of specific tokens for the answer keys in a multiple-choice setting.
- **Mathematics:** GSM8K (Cobbe et al., 2021) evaluates models on elementary level arithmetic, while MATH (Hendrycks et al., 2021b) tests on competition level mathematics. Both benchmarks evaluate the model in a few-shot setting by encouraging models to output chains of thought followed by the numeric answer in a specific format.

[1]HuggingFace's Open LLM Leaderboard (Beeching et al., 2023; Fourrier et al., 2024) migrated from GSM8K in v1 to MATH in v2.

[2]Similar to 1, the leaderboard transitioned from MMLU in v1 to MMLU-Pro and GPQA in v2.

- **Coding:** HumanEval (Chen et al., 2021) test models on basic coding, while APPS (Hendrycks et al., 2021a) uses coding competition questions. These benchmarks generate Python code by prompting LLMs with function docstrings or written specifications, and run input/output test-cases on the generated code.

Critisism of these types of static benchmarks are outlined in Section 1.

2.2 Comparative LLM Benchmarks

LLM Game Evaluations To address the static benchmark issues highlighted in Section 1, the paradigm of evaluating agentic capabilities through simulations has been applied successfully in multiple prior works. Evaluation frameworks comprising multiple games include: *(i) ChatArena* (Wu et al., 2023), which includes Chess, Tic-Tac-Toe, Rock-Paper-Scissors, and others, *(ii) GridGames* (Topsakal et al., 2024), implementing Tic-Tac-Toe, Connect Four, and Gomoku, and *(iii) GameBench* (Costarelli et al., 2024), which is the most diverse, as they developed 9 games, include non-deterministic and imperfect information games.

Limitations of LLM Game Evaluations All the aforementioned benchmark frameworks are implemented with manually written prompts for all models, and sometime suggest a strategy within the prompt, such as ChatArena prompting models to output a random move in Rock-Paper-Scissors. GameBench tries to optimize model results by utilizing two prompting strategies: *(i)* Chain of Though (CoT), and *(ii)* Reasoning via Planning (RAP), but the issue of static prompt still persists. This could explain the poor performances they observed, such as GPT-4 achieving almost random results on some tasks.

Comparative Human Evaluations A popular head-to-head LLM evaluation framework is Chatbot Arena[3] (Chiang et al., 2024), which allows users to prompt two anonymous LLMs with arbitrary prompts and to vote for the better response. This creates a diverse evaluation that effectively ranks all models in a leaderboard. However, it suffers from two issues: *(i)* human evaluations are slow and laborious, and adding new models requires prolonged evaluation periods until sufficient votes are acquired for a confident placement, and *(ii)* human evaluations contain human biases, such as prompt over-representation (Dunlap et al., 2024) and bias to verbose and "pretty" responses (Chen et al., 2024a; Park et al., 2024; Li et al., 2024).

3 Methodology

In this section, we describe the technical details of ZeroSumEval including design choices, the importance of automatic prompt optimization, and game selection/categorization. At its core, ZeroSumEval provides controlled environments to observe models competing against each other to win competitive games. In particular, ZSE controls (i) the role and information each model has access to at any point in the simulation and (ii) the data models can use to optimize/modify their own prompts.

3.1 Capabilities

The games within ZSE are designed to evaluate specific capabilities in a controlled environment:

Reasoning Board games and cybersecurity scenarios require models to perform complex, multi-step reasoning. They test the models' ability to process information, predict outcomes, and formulate strategies in dynamically changing environments.

Planning Board games also involve long-term strategy, requiring models to anticipate the consequences of their actions several moves ahead. This assesses the model's foresight, adaptability, and capacity for nuanced decision-making.

[3] formerly LMSYS, not to be confused with ChatArena.

Knowledge Application Models must recall and apply mathematical knowledge to solve problems in question answering type games. This setup provides a direct assessment of the models' ability to retrieve, interpret, and implement factual information in structured problem-solving.

Creativity Models successful at cybersecurity type games must exhibit creativity to successfully create secure environments and break them.

3.2 GAME DESIGN

ZEROSUMEVAL supports an expanding suite of game types designed to test the aspects of LLM performance described above. The mix we showcase includes both well-known and established games, such as chess, as well as more special-purpose games (e.g. MathQuiz). For completeness and reproducibility, we describe the implementations of MathQuiz and PyJail. The following set of games are selected to encompass a range of cognitive capabilities, including strategic reasoning, planning, knowledge application, and creativity:

Board Games (Chess) Classic board games like chess serve as a benchmark for strategic reasoning and long-term planning. They require models to engage in multi-step thinking, manage trade-offs, and foresee opponent moves. This category is instrumental in evaluating a model's ability to plan several moves ahead, adapt its strategies, and make complex decisions under uncertainty[4].

Question-Answer Games (MathQuiz) These games are constructed to measure models' knowledge recall and logical reasoning abilities. MathQuiz, for instance, challenges models to both create and answer arithmetic and mathematical questions, assessing their understanding of mathematical concepts, computational accuracy, and step-by-step problem-solving skills. Our implementation of MathQuiz tasks a teacher player to create a challenging math problem and prove that the problem is valid and solvable. A student player then attempts to answer the generated math problem. The student wins the game by answering the question correctly or if the teacher fails to create a valid question.

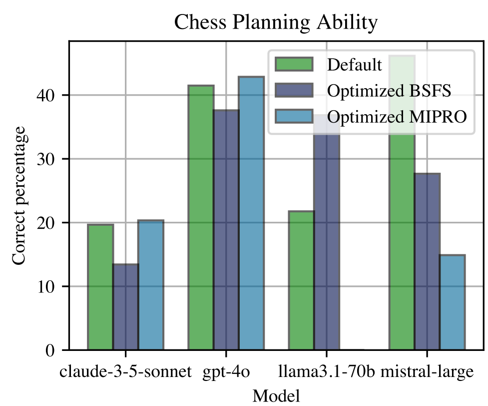

Figure 2: The effect of prompt optimization on the proportion of correct moves. Moves are classified as correct if the evaluation, as determined by Stockfish 17 (The Stockfish Developers, 2024) with depth 15, does not decrease by more than 0.3 points (pawn equivalent). Models react differently to prompts and have varying prompt optimization abilities.

Cybersecurity Games (PyJail) PyJail involves python "capture the flag" cybersecurity challenges, targeting the model's ability to create puzzles and interact with a restricted python environment to strategize solutions. The PyJail game is structured into three stages. The first statically parses a player generated PyJail program to provide feedback on the syntax and semantic structure. Given validity, the challenge code is inserted into the environment, and the same player model must commit a solution that is tested dynamically to prove the challenge's feasibility. A unique flag is stored in the target variable at runtime, which prevents any trivial method to cheat the challenge. The second player will complete the same step, provided a restricted view of the environment and limited context. The game ends if first player is unable to create a valid challenge or the flag is retrieved by the attacker.

[4]Chess has a rich history as a testbed for strategy and planning. See `https://github.com/carlini/chess-llm` and `https://huggingface.co/spaces/mlabonne/chessllm` for examples of LLMs playing chess.

3.3 SCALABLE VERIFICATION

The MathQuiz and PyJail games require competing models to generate complex challenge environments and solutions. Since verification of the knowledge-based challenges by a human in the loop is not scalable, we design a method to verify model output using an automated manager in a two-fold generation and verification process. This is accomplished by defining a target outcome (e.g., the answer to a math question or a CTF flag) as the basis for verifying generated input, and regulating the model context at each stage.

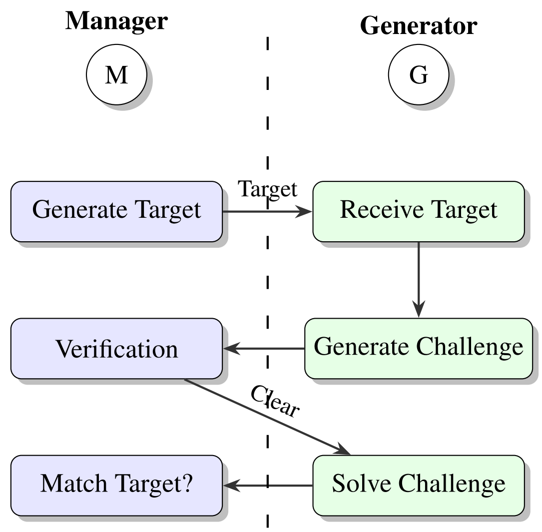

Figure 3: State diagram of the verification process involving the Game Manager and the Generator. Blue boxes indicate deterministic steps and green boxes indicate steps involving the model.

The exact process (illustrated in Figure 3) is outlined as follows:

(i) The generator model receives a target and attempts to output a valid challenge that resolves to the specific target.

(ii) In the verification step, the manager restricts the model's context to ensure no direct access to the target, and asks the generator model to solve the previously generated challenge.

(iii) If the manager determines the verification is successful (by matching the target with the generator's solution), the game proceeds. Otherwise, the generator model is deemed to have failed to generate a valid challenge.

This method ensures the generated challenge environment is valid and a solution is proven possible by the generator. The design also correctly penalizes models that directly generate memorized questions as it is likely to have been memorized by other models, thereby encouraging models to create challenging and novel questions. Finally, the scalability of the evaluation is preserved as the capabilities of models scale.

3.4 AUTOMATIC PROMPTING

Automatic prompting is an essential component of the ZEROSUMEVAL framework for several reasons. First, it allows models to learn to play new games through self-optimization, demonstrating their ability to adapt to different scenarios without human intervention. Second, it removes the human element in prompt engineering, thereby reducing biases and variations introduced by manual prompt construction (Zheng et al., 2024; Pezeshkpour & Hruschka, 2023; Alzahrani et al., 2024a). Third, automatic prompting serves as a measure of a model's ability to self-improve at inference time, providing insight into its adaptability and strategic reasoning skills.

We leverage the DSPy (Khattab et al., 2023) approach to implement automatic prompt optimization in our framework. DSPy allows models to autonomously explore and select optimal prompts based on the current game context, dynamically adjusting strategies to maximize performance. We also make use of DSPy Assertions (Singhvi et al., 2024) to simulate interactivity between the models and the game environment by allowing a number of retries (with feedback from the game) when the model makes an invalid move. Although we find DSPy has the flexibility and generalizability to support various models and games, ZEROSUMEVAL supports alternative automatic prompt optimization techniques if required.

Through prompt optimization, models can develop improved strategies as they encounter diverse game scenarios. For example, in a chess game, models equipped with optimized prompting demonstrated a higher proportion of correct moves compared to their counterparts using default prompts Figure 2. This not only reveals the models' enhanced strategic reasoning but also emphasizes the significance of prompt optimization in robust performance evaluation.

By incorporating automatic prompting, ZEROSUMEVAL addresses benchmark sensitivity. The prompt optimization integrates game validation mechanisms into the optimization process, allowing models to observe tangible outcomes and refine their prompt strategies. Consequently, this mitigates the variations in performance due to prompt sensitivity, leading to a more consistent and reliable evaluation of model capabilities.

Datasets and Optimizers To perform the automatic prompt optimization process, models require examples of gameplay (inputs and outputs) and prompt optimizers. We create standard datasets manually for each game available to all models for the optimization. The available datasets are described in Table 1. Through DSPy, ZEROSUMEVAL supports multiple types of optimizers. In this work, we focus on (i) BootstrapFewShot (ii) BootstrapFewShotRandomSearch (Khattab et al., 2023) and (iii) MIPROv2 (Opsahl-Ong et al., 2024).

Dataset	Source	Description
chess_stockfish	conacts/stockfish_dataset[5]	stockfish vs stockfish games
chess_puzzles	(Schwarzschild et al., 2021)	chess puzzles.
mathquiz_gsm8k	(Cobbe et al., 2021)	grade school level math QA
mathquiz_hendrycks_math	(Hendrycks et al., 2021b)	advanced math QA
pyjail_ctf_llm	(Shao et al., 2024)	Pyjail style Capture The Flags (CTFs).

Table 1: Overview of datasets used in the evaluation framework.

An interesting direction out of the scope of this work is enabling models to learn games via self-play. This would reduce manual effort needed to create new games for ZEROSUMEVAL and measure a model's ability to effectively explore a space without supervision.

3.5 RATINGS

ZEROSUMEVAL utilizes an easily computable rating system derived from the outcomes of competitive games between models. Each model receives a rating based on its win-loss record over multiple games, allowing for a rapid and scalable oversight of model capabilities. This framework seamlessly incorporates new games, providing continuous and dynamic evaluation as models improve.

Following recent suggestions for LLM rating systems by Boubdir et al. (2023); Chiang et al. (2023), we employ the Bradley-Terry (BT) rating system, an alternative to the Elo system, to rate models. The BT model is permutation-invariant and assumes a fixed win rate for each model pair, maximizing the likelihood of observed outcomes (Bradley & Terry, 1952). This choice is more suitable than the traditional Elo system, which was designed for human chess players with varying skill levels, whereas LLMs have fixed skill levels defined by their weights (Elo, 1967).

ZEROSUMEVAL's rating system facilitates analysis of model behaviors. It allows us to observe not only the relative strategic planning capabilities of models but also their capacity for self-improvement through prompt optimization. For instance, analysis of models' gameplay strategies in chess revealed that prompt-optimized models allocate more reasoning words in their decision-making process, suggesting a deeper level of planning (Figure 4).

4 EXPERIMENTS

In this section, we describe the experiments to demonstrate the effectiveness of the ZEROSUMEVAL as a dynamic leaderboard. We also design experiments to evaluate the effect of prompt optimization on the performance of various large language models (LLMs) under various simulations.

4.1 MODEL SELECTION AND EXPERIMENTAL SETUP

We select four models of varying sizes and capabilities for this study: GPT-4o, Claude 3.5 Sonnet, LLaMA 3.1-70B-Instruct, and Mistral-Large. These models represent a range of architectures and training scales, providing a diverse set for evaluating the generalizability of the ZEROSUMEVAL framework.

The experiments involve running a multiple round-robin tournaments to simulate competitive gameplay among the model (50-100 games per experiment). In addition to measuring model performance on the games in the ZeroSumEval suite, we also examine how the models' performance changes with different prompt optimization techniques. Each tournament round involves all possible match permutations between model variants, after which the models' ratings are calculated using the Bradley-Terry model (Bradley & Terry, 1952). The primary goal of this ablation study is to assess each model's responsiveness to the optimization process and to identify resulting behavioral changes.

For the automatic prompt optimization, we utilize three optimizers commonly used in DSPy: BootstrapFewshot (BSFS), BootstrapFewshotWithRandomSearch (BSFSRS), and MIPROv2, targeting the ChainOfThought module in DSPy.

4.2 Games for Analysis

Although ZeroSumEval supports a range of games for assessing different capabilities, our detailed set of experiments focus primarily on Chess to analyze the models' planning abilities. This decision is motivated by the interpretability of Chess gameplay and its complexity, which provides an ideal testbed for assessing strategic reasoning and decision-making.

5 Results

5.1 Ratings and Performance Trends

Table 2 provides the ratings for each model variant across the games. The results indicate GPT-4o and Claude 3.5 Sonnet typically perform best with GPT-4o slightly ahead. This agrees with leaderboards based on human ratings, such as ChatbotArena (Chiang et al., 2024).

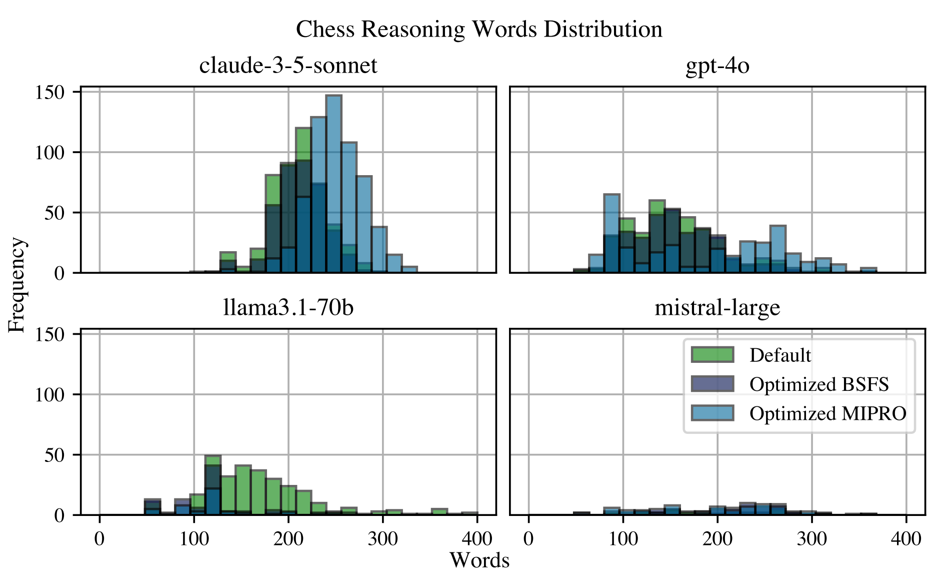

Figure 4: The distribution of CoT words used for each model and prompt optimization technique. In general, prompt optimized models spend more words reasoning than their non-optimized counterparts, especially with MIPROv2 optimization.

5.2 Impact of Prompt Optimization on Performance

The experimental results (Table 3) reveal significant variations in model performance as a result of prompt optimization. Prompt optimization can even flip ranking as is the case with MIPROv2 -

highlighting the significant effect of prompt sensitivity. Prompt-optimized models typically exhibit improved strategic reasoning, as evidenced by an increased number of correct moves and a more favorable distribution of move evaluations (Figure 2). ZEROSUMEVAL provides the capability to compare models across prompt optimization strategies, leading to fairer evaluations and more robust leaderboards.

Figure 4 illustrates the distribution of CoT words for each model with different prompt optimization techniques. Notably, models optimized using MIPROv2 demonstrate a tendency to allocate more words to their reasoning process compared to their default counterparts, suggesting deeper planning and strategic consideration.

Model	Chess (MIPRO)	MathQuiz (Default)	PyJail (Default)
GPT-4o	1202.97	1048.12	1025.58
Claude 3.5 Sonnet	1000.00	962.51	1017.17
Mistral-Large	940.88	982.85	1000.00
LLaMA 3.1 70B	856.15	1006.52	953.15

Table 2: Performance ratings of various models across different tasks. The ratings are computed using the MIPRO-optimized approach for the Chess task and default settings for MathQuiz and PyJail tasks.

Model	Default Rating (CI)	BSFS Rating (CI)	BSFSRS Rating (CI)	MIPRO Rating (CI)
Claude 3.5 Sonnet	**1028** (890-1153)	1000 (871-1126)	1000 (862-1147)	984 (837-1060)
Mistral-Large	942 (889-1005)	1016 (963-1073)	1014 (952-1069)	**1023** (965-1090)
LLaMA 3.1 70B	978 (918-1054)	951 (888-1039)	1030 (962-1089)	**1035** (967-1107)
GPT-4o	962 (880-1034)	1016 (909-1101)	966 (874-1044)	**1055** (987-1133)

Table 3: Results of engaging each model in competition against itself optimized by our choices of optimizers. We can see that for Mistral, LLaMA, and GPT-4o, MIPRO outperforms all other optimizers. It is interestingly not the case with Claude. Ratings are shown with their 95% confidence intervals (CI). The highest rating for each model is in bold.

6 CONCLUSION

The dynamic, competitive nature of ZEROSUMEVAL's evaluation provides a more robust and trustworthy measurement of AI model capabilities, advancing the state of benchmarking in large language models. By leveraging zero-sum games, we ensure that models are consistently challenged with diverse, evolving tasks, minimizing the risk of overfitting and saturation commonly observed in static benchmarks. Additionally, the integration of automatic prompt optimization offers a more holistic evaluation framework that captures not only a model's static performance but also its dynamic capacity for self-improvement.

REFERENCES

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards, 2024a. URL `https://arxiv.org/abs/2402.01781`.

Norah Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yusef Almushaykeh, Faisal Mirza, Nouf Alotaibi, Nora Altwairesh, Areeb Alowisheq, et al. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. *arXiv preprint arXiv:2402.01781*, 2024b.

Anthropic. Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`, 2024a. Accessed: 2024-09-17.

AI Anthropic. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*, 2024b.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open llm leaderboard (2023-2024). https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard, 2023.

Meriem Boubdir, Edward Kim, Beyza Ermis, Sara Hooker, and Marzieh Fadaee. Elo uncovered: Robustness and best practices in language model evaluation, 2023. URL https://arxiv.org/abs/2311.17295.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as the judge? a study on judgement biases, 2024a. URL https://arxiv.org/abs/2402.10669.

Hailin Chen, Fangkai Jiao, Xingxuan Li, Chengwei Qin, Mathieu Ravaut, Ruochen Zhao, Caiming Xiong, and Shafiq Joty. Chatgpt's one-year anniversary: Are open-source large language models catching up?, 2024b. URL https://arxiv.org/abs/2311.16989.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code, 2021. URL https://arxiv.org/abs/2107.03374.

Wei-Lin Chiang, Tim Li, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: New models & elo system update, Dec 2023. URL https://lmsys.org/blog/2023-12-07-leaderboard/.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024. URL https://arxiv.org/abs/2403.04132.

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. *arXiv preprint arXiv:2110.14168*, 2021.

Anthony Costarelli, Mat Allen, Roman Hauksson, Grace Sodunke, Suhas Hariharan, Carlson Cheng, Wenjie Li, Joshua Clymer, and Arjun Yadav. Gamebench: Evaluating strategic reasoning abilities of llm agents, 2024. URL https://arxiv.org/abs/2406.06613.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, 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, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, 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, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, 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, 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, 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, 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, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, 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, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, 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, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, 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, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, 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, Karthik Prasad, Kartikay Khandelwal, Katayoun

Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, 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, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, 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, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, 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, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, 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, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, 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, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024a. URL https://arxiv.org/abs/2407.21783.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, 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, 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, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, 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, 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, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, 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, 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, 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, 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, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, 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, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, 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, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, 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, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, 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, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, 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, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, 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, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, 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, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, 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, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, 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, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef

Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models, 2024b. URL https://arxiv.org/abs/2407.21783.

Lisa Dunlap, Evan Frick, Tianle Li, Isaac Ong, Joseph E. Gonzalez, and Wei-Lin Chiang. What's up with llama 3? arena data analysis, May 2024. URL https://lmsys.org/blog/2024-05-08-llama3/.

Arpad E Elo. The proposed uscf rating system, its development, theory, and applications. *Chess life*, 22(8):242–247, 1967.

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard, 2024.

Google. Gemini: A family of highly capable multimodal models, 2024.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. Olmo: Accelerating the science of language models, 2024. URL https://arxiv.org/abs/2402.00838.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. The false promise of imitating proprietary llms, 2023. URL https://arxiv.org/abs/2305.15717.

Zishan Guo, Renren Jin, Chuang Liu, Yufei Huang, Dan Shi, Supryadi, Linhao Yu, Yan Liu, Jiaxuan Li, Bojian Xiong, and Deyi Xiong. Evaluating large language models: A comprehensive survey, 2023. URL https://arxiv.org/abs/2310.19736.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In Marilyn Walker, Heng Ji, and Amanda Stent (eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2017. URL https://aclanthology.org/N18-2017.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Dan Hendrycks, Steven Basart, Saurav Kadavath, Mantas Mazeika, Akul Arora, Ethan Guo, Collin Burns, Samir Puranik, Horace He, Dawn Song, and Jacob Steinhardt. Measuring coding challenge competence with apps. *NeurIPS*, 2021a.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021b.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022. URL https://arxiv.org/abs/2203.15556.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards, 2018. URL https://arxiv.org/abs/1805.03677.

Ben Hutchinson, Andrew Smart, Alex Hanna, Emily Denton, Christina Greer, Oddur Kjartansson, Parker Barnes, and Margaret Mitchell. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure, 2021. URL https://arxiv.org/abs/2010.13561.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation, 2021. URL https://arxiv.org/abs/2109.06835.

Mathew Oldham Kevin Lee, Adi Gangidi. Building meta's genai infrastructure. https://engineering.fb.com/2024/03/12/data-center-engineering/building-metas-genai-infrastructure/, 2024. Accessed: September 28, 2024.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. Dynabench: Rethinking benchmarking in nlp, 2021. URL https://arxiv.org/abs/2104.14337.

Spencer Kimball. Microsoft, brookfield to develop more than 10.5 gigawatts of renewable energy. https://www.cnbc.com/2024/05/01/microsoft-brookfield-to-develop-more-than-10point5-gigawatts-of-renewable-energy.html, 2024. Accessed: September 28, 2024.

Md Tahmid Rahman Laskar, M Saiful Bari, Mizanur Rahman, Md Amran Hossen Bhuiyan, Shafiq Joty, and Jimmy Xiangji Huang. A systematic study and comprehensive evaluation of chatgpt on benchmark datasets, 2023.

Md Tahmid Rahman Laskar, Sawsan Alqahtani, M Saiful Bari, Mizanur Rahman, Mohammad Abdullah Matin Khan, Haidar Khan, Israt Jahan, Amran Bhuiyan, Chee Wei Tan, Md Rizwan Parvez, et al. A systematic survey and critical review on evaluating large language models: Challenges, limitations, and recommendations. *arXiv preprint arXiv:2407.04069*, 2024.

Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. Does style matter? disentangling style and substance in chatbot arena, Aug 2024. URL https://lmsys.org/blog/2024-08-28-style-control/.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Ron Miller. Google launches a 9 exaflop cluster of cloud tpu v4 pods into public preview. https://techcrunch.com/2022/05/11/google-launches-a-9-exaflop-cluster-of-cloud-tpu-v4-pods-into-public-preview/, 2022. Accessed: [Your Access Date].

Mistral. Au large, 2024. URL https://mistral.ai/news/mistral-large/.

OpenAI. Openai five defeats dota 2 world champions. https://openai.com/blog/openai-five-defeats-dota-2-world-champions/, April 2019. Accessed: 2024-09-28.

OpenAI. Chatgpt: Optimizing language models for dialogue, 2022. URL https://openai.com/blog/chatgpt/.

OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, Jonas Schneider, Nikolas Tezak, Jerry Tworek, Peter Welinder, Lilian Weng, Qiming Yuan, Wojciech Zaremba, and Lei Zhang. Solving rubik's cube with a robot hand, 2019. URL https://arxiv.org/abs/1910.07113.

Krista Opsahl-Ong, Michael J Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs, 2024. URL https://arxiv.org/abs/2406.11695.

Aitor Ormazabal, Che Zheng, Cyprien de Masson d'Autume, Dani Yogatama, Deyu Fu, Donovan Ong, Eric Chen, Eugenie Lamprecht, Hai Pham, Isaac Ong, Kaloyan Aleksiev, Lei Li, Matthew Henderson, Max Bain, Mikel Artetxe, Nishant Relan, Piotr Padlewski, Qi Liu, Ren Chen, Samuel Phua, Yazheng Yang, Yi Tay, Yuqi Wang, Zhongkai Zhu, and Zhihui Xie. Reka core, flash, and edge: A series of powerful multimodal language models, 2024.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization, 2024. URL https://arxiv.org/abs/2403.19159.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023. URL https://arxiv.org/abs/2308.11483.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Tali Bers, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M. Rush. Multitask prompted training enables zero-shot task generalization, 2022. URL https://arxiv.org/abs/2110.08207.

Tal Schuster, Roei Schuster, Darsh J. Shah, and Regina Barzilay. The limitations of stylometry for detecting machine-generated fake news. *Computational Linguistics*, 46(2):499–510, June 2020. doi: 10.1162/coli_a_00380. URL https://aclanthology.org/2020.cl-2.8.

Avi Schwarzschild, Eitan Borgnia, Arjun Gupta, Arpit Bansal, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. Datasets for studying generalization from easy to hard examples, 2021.

Minghao Shao, Sofija Jancheska, Meet Udeshi, Brendan Dolan-Gavitt, Haoran Xi, Kimberly Milner, Boyuan Chen, Max Yin, Siddharth Garg, Prashanth Krishnamurthy, Farshad Khorrami, Ramesh Karri, and Muhammad Shafique. Nyu ctf dataset: A scalable open-source benchmark dataset for evaluating llms in offensive security, 2024. URL https://arxiv.org/abs/2406.05590.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL https://arxiv.org/abs/1712.01815.

Arnav Singhvi, Manish Shetty, Shangyin Tan, Christopher Potts, Koushik Sen, Matei Zaharia, and Omar Khattab. Dspy assertions: Computational constraints for self-refining language model pipelines, 2024. URL https://arxiv.org/abs/2312.13382.

The Stockfish Developers. Stockfish, 2024. URL https://github.com/official-stockfish/Stockfish. Version 17.

Oguzhan Topsakal, Colby Jacob Edell, and Jackson Bailey Harper. Evaluating large language models with grid-based game competitions: An extensible llm benchmark and leaderboard, 2024. URL https://arxiv.org/abs/2407.07796.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, et al. Freshllms: Refreshing large language models with search engine augmentation. *arXiv preprint arXiv:2310.03214*, 2023.

Alex Wang. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*, 2018.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32, 2019.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL https://arxiv.org/abs/2305.16291.

Haochun Wang, Sendong Zhao, Zewen Qiang, Nuwa Xi, Bing Qin, and Ting Liu. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models, 2024a. URL https://arxiv.org/abs/2402.01349.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024b.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell (eds.). *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.conll-babylm.0.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022. URL https://arxiv.org/abs/2206.07682.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become u-shaped, 2023a. URL https://arxiv.org/abs/2211.02011.

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, 2023b. URL https://arxiv.org/abs/2201.11903.

Lionel Wong, Gabriel Grand, Alexander K. Lew, Noah D. Goodman, Vikash K. Mansinghka, Jacob Andreas, and Joshua B. Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought, 2023. URL https://arxiv.org/abs/2306.12672.

Yuxiang Wu, Zhengyao Jiang, Akbir Khan, Yao Fu, Laura Ruis, Edward Grefenstette, and Tim Rocktäschel. Chatarena: Multi-agent language game environments for large language models. https://github.com/chatarena/chatarena, 2023.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report, 2024. URL https://arxiv.org/abs/2407.10671.

Shuo Yang, Wei-Lin Chiang, Lianmin Zheng, Joseph E. Gonzalez, and Ion Stoica. Rethinking benchmark and contamination for language models with rephrased samples, 2023.

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.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors, 2024. URL https://arxiv.org/abs/2309.03882.

Stephan Zheng, Alexander Trott, Sunil Srinivasa, David C. Parkes, and Richard Socher. The ai economist: Optimal economic policy design via two-level deep reinforcement learning, 2021. URL https://arxiv.org/abs/2108.02755.

A APPENDIX

Table 4: Exact model versions used in our evaluations.

Model	Version
GPT-4o	gpt-4o-2024-08-06
Claude 3.5 Sonnet	claude-3-5-sonnet-20240620
Mistral-Large	mistralai/Mistral-Large-Instruct-2407
Llama 3.1 70B	meta-llama/Meta-Llama-3.1-70B-Instruct