Evaluating LLMs Adversarially with Word Guessing Game

Anonymous ACL submission

Abstract

The increasing significance of evaluating large language models (LLMs) is addressed in this paper. We present a new evaluation framework, Adversarial Guessing Evaluation (AGE), designed for LLMs. AGE employs a systematic set of rules and metrics to evaluate reading comprehension abilities and confusion capabilities of LLMs across different dimensions. Our framework significantly reduces the need for large datasets, requiring only a few pairs of words. The results align with average outcomes from established comprehensive benchmarks and highlight areas for potential improvements in LLMs[1].

1 Introduction

The landscape of natural language processing has undergone huge changes with the advent of large language models (LLMs), starting from foundational models such as BERT and ChatGPT, to more recent advancements like GPT-4 and LLaMA. These models have demonstrated exceptional capabilities in zero-shot generation, complex reasoning tasks, and adherence to nuanced instructions, marking significant progress in the field.

However, as these models evolve, so does the need for effective evaluation frameworks. Traditional benchmarks such as GLUE (Wang et al., 2018a) and MMLU (Hendrycks et al., 2021a) are being replaced by more open-ended evaluations like AGIEval (Zhong et al., 2023) and Chatbot-Arena (Chiang et al., 2024), reflecting a shift towards assessing generalization across broader, more complex scenarios. Current evaluation methods typically fall into two main categories: reference-based, which relies on predefined answers, and preference-based, which involves subjective human judgments or model preferences (Qiao et al., 2023). Each of these approaches has its limitations, ranging from the high costs of annotations to potential biases introduced by human evaluators.

Recent research has pivoted towards using game-based evaluations for LLMs, where models engage in controlled word games (Qiao et al., 2023; Xu et al., 2023; Liang et al., 2023a). This method not only circumvents subjective bias by minimizing direct interaction between researchers and models but also can be evaluated in more dimensions.

Building on these insights, we propose the "Adversarial Guessing Evaluation (AGE)" framework, which leverages simplified rules from the game "Who Is Spy" to evaluate LLMs in a structured yet challenging environment. This framework not only broadens the scope of model evaluation but also provides a direct measure of performance across diverse scenarios.

The primary contributions of our work are as follows:

- Introduction of a robust but light framework for the autonomous evaluation of LLMs using adversarial guessing games, which expands the set of evaluation tools with a methodologically novel approach.
- Comprehensive analysis of LLM performances across ten distinct fields using the AGE framework, thereby identifying potential biases and areas of improvement for model training.

2 Related Works

2.1 Evaluation of LLMs

The rigorous evaluation of LLMs has become a cornerstone in advancing their capabilities and applications. Researchers categorize these evaluations into three primary dimensions: NLP tasks, alignment evaluation, and real-world complex tasks.

[1] Codes, words, conversations, and prompts will be released upon acceptance.

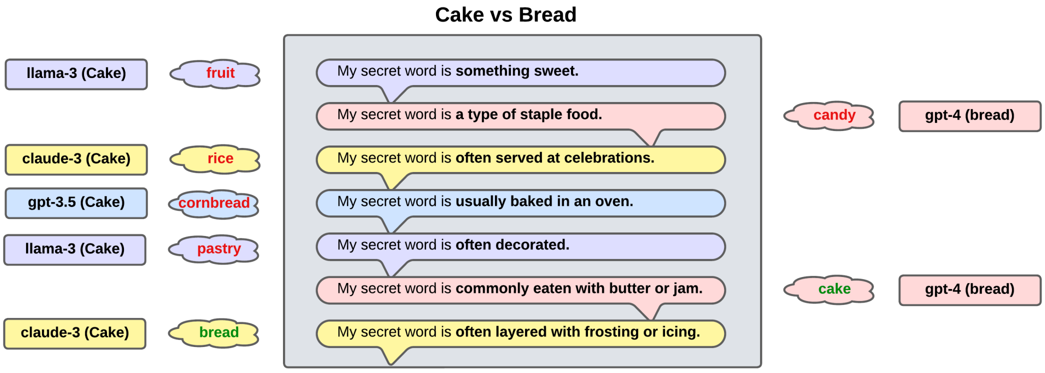

Figure 1: A sample conversation in AGE.

For NLP tasks, benchmarks such as GLUE (Wang et al., 2018b), SuperGLUE (Wang et al., 2019), and MMLU (Hendrycks et al., 2020) are prevalent, testing models on diverse linguistic challenges. Alignment evaluations, such as those conducted using AlpacaEval (Li et al., 2023), focus on the utility and safety of model outputs. Complex task evaluations involve scenarios mimicking real-world interactions, exemplified by Webshop (Yao et al., 2022) and AgentBench (Liu et al., 2023), highlighting the practical implications of deploying LLMs in various environments (Schaeffer, 2023).

2.2 Development of LLM-based Agents

The advent of LLM-based agents marks a significant innovation, particularly in the NLP domain. These agents are crafted to facilitate coherent, multi-turn conversations, simulating human-like interactions (Du et al., 2023; Liang et al., 2023b). Their applications extend beyond communication, contributing to fields such as software development (Qian et al., 2023; Hong et al., 2023), social simulation (Park et al., 2022, 2023), and robotic assistance (Brohan et al., 2023).

2.3 Game Playing with Large Language Models

Integrating LLMs into gaming environments, such as GameEval (Qiao et al., 2023) and Werewolf (Xu et al., 2023), sheds light on their strategic adaptabilities and interaction proficiencies in multi-agent settings. This research area not only examines the gameplay mechanics of LLMs but also their inherent biases and the ways these biases are expressed in complex interaction frameworks.

Expanding on these insights, in our framework. Simplified rules has been adopted from the game "Who Is Spy" (the same game in Qiao et al. (2023) and Liang et al. (2023a)) to assess LLMs in a controlled yet challenging context. AGE not only expands the methodology for evaluating models but also delivers a concrete metric of performance across varied scenarios.

3 Experiments

3.1 Game Setting

In the "Who Is Spy" game, participants are divided into two groups. Each game involves a pair of secret words that share similar attributes but are not identical. At the game's start, each player is assigned one of these secret words, which they must then describe to the others without explicitly revealing it. The game progresses through two pivotal stages: description and guessing.

During the description phase, players provide unique and non-repetitive clues about their assigned words. Creativity is crucial, as overly straightforward descriptions can easily compromise the game. The guessing phase marks the real challenge, where players must interpret the clues from the initial phase and deduce their opponents' words. This stage not only tests the players' vocabulary and creativity but also their ability to deceive their opponents.

3.2 Framework:

Our proposed framework is as follows: LLMs will be given a random order at the beginning. Each LLM is assigned a secret word. In each round, LLMs will review the conversation log, describe their secret word, and attempt to guess the others' secret words. The AGE continues until a predefined condition is met, such as all secret words being guessed or reaching the maximum number of rounds. Figure 1 exemplifies a conversation in AGE, where four LLMs have been assigned the words 'cake' and 'bread'. After several turns, GPT-4 correctly guesses the others' secret words, followed by Claude-3.

3.3 Evaluation:

Different from previous works where the average turn of the First Correct Guessing (FCG) as the score is taken directly (Qiao et al., 2023; Liang et al., 2023a), AGE gets one step further by abstracting two key attributes from the game, which are reading comprehension and confusing capability. Both capabilities can be measured with the locations of FCG. For an agent in AGE, assume the better the reading comprehension, the quicker the FCG will be found. Similarly, the better the confusing capability, the later its opponents' FCG will occur.

AGE incorporates three basic measurements for evaluation. In one AGE scenario, LLMs (Large Language Models) are strategically divided into two groups based on their assigned secret words. Let the first group be denoted as A, which includes LLMs assigned the secret word one, denoted as $\{a_1, a_2, \ldots, a_i\}$. The second group, denoted as B, consists of LLMs assigned the secret word two, represented as $\{b_1, b_2, \ldots, b_j\}$. Each LLM's first successful guess in the game is recorded in the set F, comprising the first correct guessing for both groups: $\{f_{a1}, f_{a2}, \ldots, f_{ai}, f_{b1}, f_{b2}, \ldots, f_{bj}\}$. This structured approach facilitates a systematic analysis of guessing dynamics and strategy efficacy within the AGE framework.

The first metric, comprehension, is defined for each model from sets A and B with the first correct guess f as:

$$\text{comprehension} = \frac{1}{\log_2(f+1)}$$

This metric measures the reading comprehension of the model. As the location of the First Correct Guess (FCG) occurs later, the comprehension decreases.

The second metric, confusion, measures the capability of a model in confusing other participants. For a model belonging to set A, this score is calculated using all of the models in set B and the length of the AGE, l, as:

$$\text{confusion} = \log_{l+1}(\min(f_{b1}, f_{b2}, \ldots, f_{bj}))$$

The confusion of an LLM is calculated based on the timing of correct guesses by its opponents; the later these occur, the higher the confusion score.

Finally, based on the comprehension and confusion metrics, AGE introduces a unified metric called the AGE score, which considers both metrics with equal weight:

$$\text{AGE score} = \frac{2 \times (\text{comprehension} \times \text{confusion})}{(\text{comprehension} + \text{confusion})}$$

3.4 Secret Words

Table 1: A comparison with previous studies (Qiao et al., 2023; Liang et al., 2023a)

Study	Word Pairs	Unified Metrics
GameEval	11	No
SpyGame	50	No
AGE	531	AGE score

The AGE framework features a significantly expanded set of secret word pairs compared to previous studies, encompassing 11 distinct lists. Specifically, List A contains 45 pairs, which were curated by real annotators using web searches. The remaining ten lists were generated with the assistance of ChatGPT and span ten different categories of news, including Business, Entertainment, among others. These lists have between 49 and 60 word pairs each. As indicated in Table 1, AGE offers a substantially larger repository of word pairs (a total of 531 pairs) compared to its predecessors (which provided only 50 and 11 pairs, respectively). Additionally, AGE employs more sophisticated evaluation metrics, thereby enhancing the reliability of LLM assessments in comparison to earlier frameworks.

In the following sections, two experiments based on AGE will be conducted to assess popular LLMs. To ensure rapid performance, the correlation between target words and responses will be evaluated using the Jaro similarity metric (a string metric for measuring the edit distance between two sequences, ranging from 0 to 1, where 1 indicates exact similarity), with a threshold value greater than 0.8, and a dictionary of similar words created by annotators.

4 Results

4.1 AGE with Four LLMs

In this experiment, four prominent LLMs were included in the AGE: GPT-3.5, GPT-4[2], LLama-

[2]https://www.openai.com

Table 2: Performance Metrics for LLMs, ordered by AGE score (macro), R_c for comprehension, C_c for confusion, $Avg.P$ refers to the average of popular metrics (MMLU, HellaSwag, HumanEval, BIG-Bench Hard, GSM-8K and MATH).

Model	R_c	C_c	**AGE**	$Avg.P$
Claude3	0.4011	0.4736	**0.3260**	**84.83%**
LLama3	0.3794	0.4926	**0.3093**	**79.36%**
GPT4	0.3846	0.4668	**0.3089**	**79.45%**
GPT3.5	0.3566	0.4755	**0.2877**	**65.46%**

3-70B-instruct[3], and Claude-3-Opus[4]. The termination condition was set when the conversation reached five rounds.

Out of 972 conversations, 886 successfully extended beyond two rounds. Within this subset, 85.89% (761 conversations) accurately guessed both secret words, demonstrating the effective performance of these LLMs in the game.

To benchmark against common metrics, an average performance score, $Avg.P$, was computed based on data from various sources including the Hugging Face[5], the LLama website[6], and papers by Anthropic (2024) and OpenAI et al. (2024). The results were closely aligned with average scores from renowned metrics such as MMLU (Hendrycks et al., 2021b), HellaSwag (Zellers et al., 2019), HumanEval (Chen et al., 2021), BIG-Bench Hard (Suzgun et al., 2022), GSM-8K (Cobbe et al., 2021), and MATH (Hendrycks et al., 2021b). Claude-3-Opus outperformed other models, with LLama-3-70B and GPT-4 closely behind, and GPT-3.5 showing the least effective performance.

The close correlation between our framework outcomes and established benchmarks underscores the robust capability of these LLMs in conversational games.

4.2 GPT-3.5 vs others in 10 fields

In this study, only two different LLMs with varied topics were incorporated into each AGE, specifically including GPT-3.5 to minimize operational costs. The experimental setup spanned ten topics, where each LLM engaged in six conversations per word pair with GPT-3.5, 3 for word A and 3 for word B. The termination criterion for each session

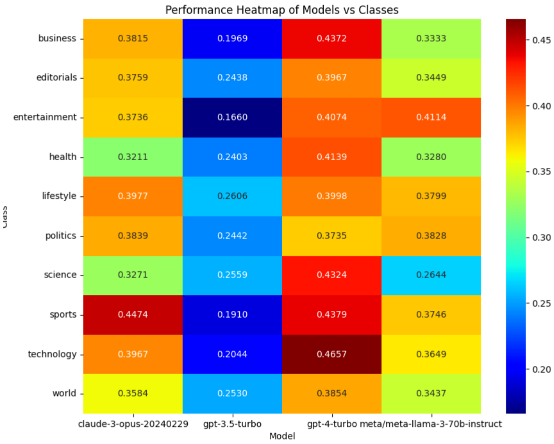

Figure 2: GPT-3.5 vs others in 10 fields

was set at five dialogue rounds. A total of 2,482 successful conversations—defined as those extending beyond two rounds with all intended words correctly identified—were collected and subjected to analysis. Figure 2 illustrates the distribution of the macro-averaged AGE scores comparing each model against GPT-3.5 across the various topics.

The results demonstrate that when using only GPT-3.5 as the adversary, GPT-4 achieves the highest average performance (0.4150), followed by claude-3-opus (0.3763), and then llama-3-70b (0.3528). This indicates that GPT-4 is more familiar with GPT-3.5. The heatmap further reveals that sports (0.3627 average) and technology (0.3579) are the two topics most familiar to LLMs, while science (0.3199) and health (0.3258) rank the lowest. These results may also contribute to the future enhancement of LLMs. Further t-tests were also performed between the models' AGE score. The results, all of p-values were below 0.01, indicate that the models are independent.

5 Conclusion

In this paper, we present a simplified version of the word guessing game rules (Who Is Spy) and propose the AGE framework for evaluating LLMs. We introduce three unified metrics aimed at assessing reading comprehension, confusion capability, and the overall AGE score, which accounts for both aspects. Our findings reveal a close alignment with the average values from multiple well-known benchmarks across four LLMs. Further insights gained from experiments across ten topics suggest avenues for enhancing these models.

[3]https://www.replicate.com
[4]https://www.anthropic.com/claude
[5]https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard
[6]https://ai.meta.com/blog/meta-llama-3/

6 Limitations

Firstly, due to limitations and costs, the comparison between guessed words and target words will include annotations by the authors of this paper (who have agreed to use and publish these annotations). These annotations help align different words with the same meaning to the correct target, but they may introduce biases from the annotators.

Secondly, only four LLMs are considered in this paper. They vary in size and structure. A limitation of the evaluation is that more LLMs, fine-tuned from the same base model, should be tested to control the influence of size and structure. Due to space constraints, this issue will be addressed in future work.

Thirdly, the word pairs used in the second experiment are collected with the assistance of an LLM, which may introduce bias when evaluating such LLMs, as well as the word pairs in the first experiment may also include biases from the creator.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, et al. 2023. Do as i can, not as i say: Grounding language in robotic affordances. In *Conference on Robot Learning*. PMLR.

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. 2021. Evaluating large language models trained on code.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv: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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv*.

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

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, et al. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. 2023. Alpacaeval: An automatic evaluator of instruction-following models.

Tian Liang, Zhiwei He, Jen-tes Huang, Wenxuan Wang, Wenxiang Jiao, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2023a. Leveraging word guessing games to assess the intelligence of large language models. *arXiv preprint arXiv:2310.20499*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023b. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv*.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023. Agentbench: Evaluating llms as agents. *arXiv*.

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. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Joon Sung Park, Joseph C O'Brien, Carrie J Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv*.

Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2022. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*.

Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. 2023. Communicative agents for software development. *arXiv*.

Dan Qiao, Chenfei Wu, Yaobo Liang, Juntao Li, and Nan Duan. 2023. Gameeval: Evaluating llms on conversational games. *arXiv preprint arXiv:2308.10032*.

Rylan Schaeffer. 2023. Pretraining on the test set is all you need. *arXiv*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, , and Jason Wei. 2022. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*.

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

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018a.

GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018b. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations*.

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring large language models for communication games: An empirical study on werewolf. *arXiv preprint arXiv:2309.04658*.

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *Preprint*, arXiv:1905.07830.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *Preprint*, arXiv:2304.06364.

A Example Appendix

This is an appendix.