TEXT-BASED GAMES AS A CHALLENGING BENCHMARK FOR LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-based games (TBG) are puzzle-solving, interactive dialogue, and generic language tasks that can play a substantial role as a challenging intelligence benchmark for large language models (LLMs.) TBGs are similar to interactive dialogue as they both require the capability for bidirectional communication in natural language while being straightforward to evaluate model performance as a score clearly indicates progress in TBGs. We conduct preliminary experiments on FLAN-T5, Turing, and OPT language models to test their puzzle-solving abilities using an *easy* TBG called "Detective". Our results suggest that LLMs underperform in comparison with state-of-the-art and human performance. We discuss the potential reasons behind the performance gap, such as the complexity of turning TBGs into prompts, not learning from past trials, lack of memory, and statistical prediction instead of goal orientation.

1 THINKING VS. LANGUAGE

Large language models have achieved promising performance across many language tasks (2020; 2022; 2023; 2022; 2022), however, the extent of their human-like thinking capabilities is still unclear. We argue that text-based games (TBGs) are an important benchmark for the development of Artificial General Intelligence (AGI) (2001; 2020), and specifically for improving LLMs. Text-based games, also called interactive fictions, describe the environment and story in natural language and require natural language commands. TBGs can serve as an effective test for measuring the thinking capabilities of LLMs, and high-quality benchmarks such as Jericho (2020) and TextWorld (2019) are publicly available to researchers. These games typically involve both language understanding and generation components and puzzle-solving elements, making them an ideal testing ground for evaluating a model's performance. The language aspect of the TBGs makes them accessible for large language models to interact with and the puzzle aspect adds complexity that the models cannot overcome by memorizing training data.

2 TEXT-BASED GAMES AND INTERACTIVE DIALOGUE

Generic interactive dialogue is a core NLP task as it closely represents and requires human-level language. TBGs are similar to interactive dialogue in two key aspects: (i) The narrative of a TBG is delivered in natural language. Players type text commands to advance in the game, and the game returns textual feedback in response to the player's action. This can be viewed as an interactive dialogue between the player and the game. (ii) Both tasks require language understanding, memory, and reasoning capabilities. However, text-based games make for easier evaluation than dialogue. A TBG agent's interactions with the are pre-defined with clear scoring criteria, while an agent for dialogue generation may need to interact with real users whose inputs are unpredictable and dynamic and ground truth or reliable evaluation metrics are often unavailable.

3 EXPERIMENTS AND DISCUSSION

We formulate text-based games as prompting tasks for large language models in a similar way to how human players interact with the game. At each step, we input the texts from the game, which contain a description of the room that the player is currently in and/or feedback on the player's last

Models	Model Size	Scores
Flan-T5	80M	50
OPT	125M	10
T-NLG	7B	90
RC-DQN	N/A	317
Human Performance	N/A	350
Max Score	N/A	360

Table 1: Scores achieved by LLMs (best of 100 trials), SOTA model (average of 100 trials), and human performance (best of 100 trials) in playing Detective.

action, into the models. Then, we input the texts generated by the models in response to the prompts into the game to perform one step of gameplay.

The LLMs we test are FLAN-T5 (2019; 2023) (Text-To-Text Transfer Transformer), OPT (2022) (Open Pre-Trained Transformer Language Models), and T-NLG (2022) (Turing Natural Language Generation). The text-based game we use is "Detective", an easy game as defined in the Jericho benchmark (2019). The results are shown in table 3, alongside the state-of-the-art model "RC-DQN" (2020) which utilizes a deep reinforcement learning (RL) agent, and human performance (best of 100 trials). According to table 3, the scores achieved by large language models are noticeably lower than the SOTA RC-DQN model in playing Detective. Through analyzing the LLM outputs, we highlight three main reasons why text-based games are difficult for large language models:

First, TBGs require players to learn from trials which LLMs are not good at. Text-based games are designed for players to try many times to solve puzzles. There usually exist multiple actions that seem to be reasonable given the current and previous game texts. Only by trying these options can the player figure out which action is the best one. The connection between the given game texts and the best actions is often not linear. This characteristic makes it hard for LLMs to play TBGs because they are pre-trained and do not support learning from individual users or prompts in real time. **Second**, TBGs require players to have long-term memory. Players need to memorize the important information given in the previous game texts to better perform actions in later scenes. However, many LLMs before ChatGPT such as FLAN-T5 and OPT do not contain any explicit memory or state. They rely entirely on the current prompts to take in information. Therefore, when prompts only contain the description/feedback from the game at the current step, the models forget about the previous information and have difficulty finding optimal or even reasonable actions. **Third**, TBGs require players to be aware of their goal that they are playing a game and want to achieve high scores in it. With this goal in mind, players will input reasonable commands that they believe could help achieve their goal. However, LLMs generate text by predicting the next token based on statistical patterns learned from the large training dataset. Therefore, the models sometimes will generate outputs that are reasonable for continuing the prompts while not reasonable for playing a game. For example, the OPT-125M model sometimes gets confused that it is going to be a player to play a game. Instead, the model continues or repeats the scenes described in the game descriptions.

4 Conclusion and Future Work

Our results demonstrate that text-based games are difficult for large language models. The three reasons: inability to learn from past trials, memory, and goal, may be summarized into one, that is large language models lack an effective strategy or a way to learn an effective strategy to play a game because they are probability distribution over sequences of words. We also acknowledge that our experiments are preliminary and do not rule out the existence of a "golden prompt" that effectively solves TBGs using LLMs. Additionally, LLMs evolve and new models such as ChatGPT (2022) exhibit a limited memory of recent interactions.

In future work we plan to explore more complex prompting that may simplify the dynamic and stateful aspects of TBGs, making them more approachable for LLMs. Recent work has fused offline reinforcement learning (2021) with language capabilities into long-context language decision transformers (LLDTs.) (2023) Such directions are ripe for developing stateful and more capable LLMs and plan to develop models and benchmarks to solve TBGs.

4.0.1 URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 15084–15097. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/7f489f642a0ddb10272b5c31057f0663-Paper.pdf.

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. Textworld: A learning environment for text-based games. In Tristan Cazenave, Abdallah Saffidine, and Nathan Sturtevant (eds.), *Computer Games*, pp. 41–75, Cham, 2019. Springer International Publishing. ISBN 978-3-030-24337-1.

Avia Efrat, Or Honovich, and Omer Levy. Lmentry: A language model benchmark of elementary language tasks, 2022. URL https://arxiv.org/abs/2211.02069.

Nicolas Gontier, Pau Rodriguez, Issam Laradji, David Vazquez, and Christopher Pal. Long-context language decision transformers and exponential tilt for interactive text environments, 2023. URL https://arxiv.org/abs/2302.05507.

Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of gpt-3, 2022. URL https://arxiv.org/abs/2209.12356.

Xiaoxiao Guo, Mo Yu, Yupeng Gao, Chuang Gan, Murray Campbell, and Shiyu Chang. Interactive fiction game playing as multi-paragraph reading comprehension with reinforcement learning. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pp. 7755–7765. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.624. URL https://doi.org/10.18653/v1/2020.emnlp-main.624.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7903–7910, Apr. 2020. doi: 10.1609/aaai.v34i05.6297. URL https://ojs.aaai.org/index.php/AAAI/article/view/6297.

Matthew J. Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. Interactive fiction games: A colossal adventure. *CoRR*, abs/1909.05398, 2019. URL http://arxiv.org/abs/1909.05398.

John Laird and Michael VanLent. Human-level ai's killer application: Interactive computer games. *AI Magazine*, 22(2):15, Jun. 2001. doi: 10.1609/aimag.v22i2.1558. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1558.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. The flan collection: Designing data and methods for effective instruction tuning, 2023. URL https://arxiv.org/abs/2301.13688.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683, 2019. URL http://arxiv.org/abs/1910.10683.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model, 2022. URL https://arxiv.org/abs/2201.11990.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022. URL https://arxiv.org/abs/2205.01068.

Tianyi Zhang, Faisal Ladhak, Esin Durmus, Percy Liang, Kathleen McKeown, and Tatsunori B. Hashimoto. Benchmarking large language models for news summarization, 2023. URL https://arxiv.org/abs/2301.13848.