Linguistic A-Maze-ment: Solving Textual Mazes with Pretrained Language Models

Gabriel Del Castillo*, Nate Webster†
Department of Computer Science
Colorado School of Mines
Golden, CO, USA
*gdelcastillo@mines.edu, †nwebster@mines.edu

***Abstract*—**

I. Introduction

Large Language Models (LLMs) such as GPT-4 and LlaMa 3 are able to produce remarkably accurate results when faced with various natural language processing tasks [1]. As the scale and reach of these models have expanded, so has the focus on developing benchmarks to measure the limits and capabilities of LLMs [2], [3]. However, there has been little research examining how pretrained language models handle maze solving tasks, along with identifying possible common failure patterns. The ability to solve maze-like structures has long served as a way to evaluate the cognitive and algorithmic performance of various subjects, with research ranging from rodent navigation [4], to spatial understanding in infants [5], to path planning for robotic systems [6].

Motivated by this, as well as the prospect of integrating LLMs into physical robotic agents capable of speech, we analyze the responses of four different language models after being prompted with one of several solvable maze configurations. We evaluate the validity and optimality of their answers, and measure if parameter count, model distribution type, or maze representation have an impact on task performance.

The overarching goal of our work was to determine whether publicly available language models have sufficient spatial awareness to correctly execute a given navigational task. In turn, this serves as a foundation for deploying LLMs in contexts where they need to interact with the physical world.

A. Related Work

Popular LLM benchmarks, including MMLU [8], BIG-Bench [2], and HELM [3] provide evaluation frameworks across several language and logic-related tasks, yet fail to test for a model's spatial reasoning or path-finding abilities. Research focused on evaluating language models in embodied settings, such as SayCan [9] and ReAct [10], reflects the growing interest in testing LLM's capacity for interaction beyond pure language. In addition, recent advances in automatic and customizable maze generation have made it feasible to evaluate symbolic spatial reasoning in LLMs. The library presented in [11] enables configurable creation and representation of maze structures, which is well-suited for benchmarking purposes and is utilized in this work.

II. Methodology

III. Results

IV. Conclusion

V. Acknowledgements

References

[1] OpenAI, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774
[2] A. Srivastava *et al.*, "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models," *arXiv preprint arXiv:2206.04615*, 2022.
[3] P. Liang *et al.*, "Holistic Evaluation of Language Models," *arXiv preprint arXiv:2211.09110*, 2022.
[4] E. C. Tolman, "Cognitive Maps in Rats and Men," Psychol. Rev., vol. 55, no. 4, pp. 189-208, Jul. 1948.
[5] J. Piaget and B. Inhelder, "The Child's Conception of Space," Int. J. Psychol., vol. 2, no. 3, pp. 241-242, 1967.
[6] Path-finding Algorithms: A. Stentz, "Optimal and Efficient Path Planning for Partially Known Environments," in IEEE International Conference on Robotics and Automation, 1994, pp. 3310-3317.
[7] "LLM Index — Promptmetheus," Promptmetheus, Mar. 13, 2025. https://promptmetheus.com/resources/llm-index (accessed Mar. 26, 2025).
[8] D. Hendrycks *et al.*, "Measuring Massive Multitask Language Understanding," *arXiv preprint arXiv:2009.03300*, 2021.
[9] M. Ahn *et al.*, "Do As I Can, Not As I Say: Grounding Language in Robotic Affordances," *arXiv preprint arXiv:2204.01691*, 2022.
[10] S. Yao *et al.*, "ReAct: Synergizing Reasoning and Acting in Language Models," *arXiv preprint arXiv:2210.03629*, 2022.
[11] M. I. Ivanitskiy *et al.*, "A configurable library for generating and manipulating maze datasets," *arXiv preprint arXiv:2309.10498*, 2023. [Online]. Available: https://arxiv.org/abs/2309.10498