A Blank ACL Paper

Gyeongseo Hwang
Department of Linguistics
2023-10775
rudtj0801@snu.ac.kr

Abstract

table of contents

1 Introduction

The central question of this paper is whether LLMs are merely probabilistic machines based on statistical patterns, or whether they can truly infer linguistic rules. Traditional linguistic theories, particularly generative grammar, have characterized language as a system composed of symbolic elements and explicit rules. In contrast, LLMs operate without explicit rules, instead learning context-dependent probability distributions through the optimization of a vast number of parameters.

This raises the question of whether such models can effectively capture the rule-governed nature of human language—particularly its compositionality and capacity for generalization. Understanding the generalization capabilities and limitations of LLMs can not only help to illuminate the internal mechanisms of these black-box models, but also offer a valuable theoretical framework for exploring the cognitive underpinnings of human linguistic competence(Goldstein et al. 2022).
djWJrh

전문적인 Large Reasoning Model의 추론 능력조차 일정 난도 이상의 task에서는 collapse하기도 했다. (Shojaee*† et al. 2025)

2 Related Work

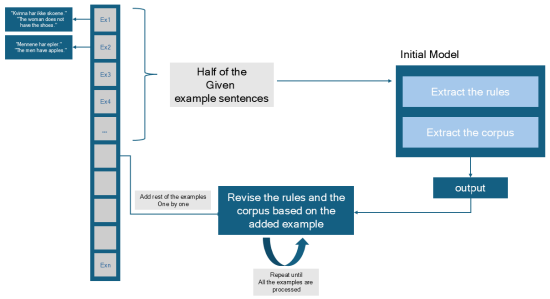
아직까지 언어학 올림피아드를 중점적으로 다룬 AI 논문과 데이터셋은 많지 않다. LINGOLY 데이터셋은 90여 개의 저자원 언어를 포함하는 challenging Linguistic Olympiad puzzle이다 (Bean et al. 2024). PuzzLing이라는, Linguistic Olympiad에 기반한 Rosetta Stone 데이터셋이 존재한다(Şahin et al. 2020).

한편, PuzzLing dataset에서 Tree-of-Thoughts 기법과 Standard I/O의 성능을 비교한 선행연구가 있다(Lin et al. 2023). 이 연구는 Tree-of-Thoughts를 적용했을 때 Standard I/O보다 오히려 성능이 낮아졌음을 보고하였고, 이유를 아래와 같이 추정하였다.

1. 프롬프트의 정확성: 프롬프트가 충분히 정확하지 않아 LLM이 혼동함
2. 평가 방법의 민감도: 현재 평가 방법이 새로운 후보 해결책으로 인해 발생하는 모순을 제대로 감지하지 못했을 수 있음
3. LLM의 능력: GPT-3.5-Turbo를 사용했는데, ToT의 연산량을 감당하기에는 성능이 부족했을 수 있음
4. ToT 구조의 적합성: ToT가 언어학 문제 해결에 적합하지 않을 수 있음. 즉, 언어규칙 추론에는 문제를 부분적으로 해결하는 것보다 전체적인 패턴 분석이 더 중요할 수 있음

따라서, 본 연구에서는 corpus와 rule, 두 가지 요소를 도입하여 또한 GPT-3.5-Turbo와 GPT-4o에서 각각 반복했을 때와 그렇지 않을 때를 비교하여 반복 방식의 성능 부진이 LLM의 capacity(연산량) 때문인지 알아볼 것이다.

3 Method



4 Dataset

LINGOLY 데이터셋으로 분석한다. PuzzLing은 부록으로 수록한다.
기존엔 PuzzLing으로 분석했으나, 정답 공개된 것이 10개뿐이고 나머지의 성능은 대시보드로만 확인할 수 있다. 그 결과도 첨부하긴 할 것이다

5 Experiments

6 Results

필요한 Figures 우선 실험할 모델은 gpt-3.5-turbo, gpt-4o

7 Analysis

7.1 rule 추출보다 **corpus** 추출의 효과가 더 크다. -> **rule** 추출은 어려운 반면 **corpus**는 비교적 잘한다?

7.2 Tree-of thought 방식은 효과가 없다.

7.3 LLM은 자신의 출력을 수정하고 검증하지 못한다.

7.4 번외 실험: 모델 번갈아가며 **test. gpt 3.5-gpt4o-3.5.**그 반대 순서도.

7.5 문제에서 주어진 예문을 결과에 다시 한 번 통합해서 넣어주면 더 자랗ㅁ

8 Conclusion

jjj

9 References

References

Bean, Andrew M., Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. “LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages”. 2024. https://arxiv.org/abs/2406.06196.

Goldstein, A., Z. Zada, E. Buchnik, M. Schain, A. Price, B. Aubrey, S. Nastase, et al. 2022. “Shared Computational Principles for Language Processing in Humans and Deep Language Models”.*Nat Neurosci.*

Lin, Zheng-Lin, Chiao-Han Yen, Jia-Cheng Xu, Deborah Watty, and Shu-Kai Hsieh. 2023. “Solving Linguistic Olympiad Problems with Tree-of-Thought Prompting”. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, edited by Jheng-Long Wu and Ming-Hsiang Su, 262–69. Taipei City, Taiwan: The Association for Computational Linguistics, Chinese Language Processing (ACLCLP). https://aclanthology.org/2023.rocling-1.33/.

Shojaee*†, Parshin, Iman Mirzadeh*, Keivan Alizadeh, Maxwell Horton, Samy Bengio, and Mehrdad Farajtabar. 2025. “The Illusion of Thinking: Understanding the Strengths and Limitations of Reasoning Models via the Lens of Problem Complexity”. 2025. https://ml-site.cdn-apple.com/papers/the-illusion-of-thinking.pdf.

Şahin, Gözde Gül, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. “PuzzLing Machines: A Challenge on Learning From Small Data”. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, edited by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, 1241–54. Online: Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.115.