The Application and Analysis of Large Language Models in Linguistics

Enrui Mei *

Qingdao No.58 High School of Shandong Province, Qingdao, China

* Corresponding Author Email: alexmei080904@outlook.com

Abstract. The prevail large language models still needs to improve reasoning ability. By studying large language models in linguistics can not only improve large language mode's reasoning ability in complex question especially in ambiguous tasks but also help linguistics and researchers to gain a useful tool to research ancient literacy and ancient cultures. This paper compares different methods to train large language models by linguistics. This paper showcases that some models have the ability of identifying ambiguity in sentences by Linguistic Recursion and Chain-of-Thought. What's more, this paper also shows that sometimes commercial models are not as good as open-source models. Researchers have to carefully choose to use which large language model. There aren't any perfect models. If researchers want high accuracy, they have to give up some efficiency. If users really need high efficiency, the accuracy of large language models may be not good enough. This paper also shows some problems that still need to be solved. The hardware is a significant factor that restrains the improvement reasoning ability of large language models. Linguistics also need to provide more explicit ambiguous data for large language models to train. And nowadays, large language models still have poor performance in systematic reasoning.

Keywords: Large language model, linguistics, reasoning ability, Cohen's κ coefficient.

1. Introduction

In the current era, large language models are used in many fields like statistics and economics. People use large language models for them to search data or information. People use large language models to solve difficult problems. It seems that AI is omniscient. However, AI still needs to learn. This process is called machine learning. The large language models need the reasoning ability to solve problems and complete tasks. For example, Abdulkhaleq Q. A. Hassan, Dr. Shoayee ALotaibi, Wala Bin Subait, Abdullah Al-Dobaian, Hanan Al Sultan, Manar Almanea, Randa Allafi, and Menwa Alshammeri say that sarcasm is a linguistic expression that typically conveys the contrary meaning of what has already been said, making it challenging for machines to find the literal meaning. It depends mainly on context, making it a tedious process for computational analysis [1]. There is an assumption that there must be some relationship between linguistics and large language models. However, the research of applications of linguistics is not thorough enough. There are many papers about the impact of large language models in linguistics. However, they concentrated more on linguistics rather than the improvement of reasoning ability of the large language models. By researching the applications of linguistics, the paper can train the large language models in related aspects. For example, Alexander G. Sboev, Artem V. Gryaznov, Roman B. Rybka, Maxim S. Skorokhodov, Ivan A. Moloshnikov make research by using neural network models to search and replace pronouns in commands, restore missing verbs-actions, decompose a complex command with several actions into simple commands with only one action and classify simple command attribute [2]. Therefore, we can also train the logic and reasoning ability of large language models. There are already some papers about this problem. Lin found that traditional linear aggression has some drawbacks that can be improved by machine learning. She also gives some recommendations of practical programming language and platform [3]. Beguš, Dąbkowski, and Rhodes found that by evaluating the metalinguistic abilities, our understanding of their general capabilities can be improved, and it sheds new light on theoretical models in linguistics. They also speculate that OpenAI o1's unique advantage over other models may result from the model's chain-of-thought mechanism, which

mimics the structure of human reasoning used in complex cognitive tasks, such as linguistic analysis [4]. Ambridge and Blything found that large language models are better than theoretical linguists at theoretical linguistics, at least in the domain of verb argument structure. Large language models succeed where these verbal (i.e., non-computational) linguistic theories fail, precisely because the latter insist – in the service of intuitive interpret ability – on simple yet empirically inadequate (over)generalizations [5]. Vajjala gives the discussion focuses on giving an introduction to the methods and applications of machine learning in applied linguistics [6]. This paper concentrates on improving the reasoning ability of the large language models by studying the application of linguistics. To achieve this goal, different large language models and languages will be used. This paper will compare different methods to train large language models by linguistics.

2. Overview of Mainstream echnologies

Although many base large language models are powerful enough, there is still great potential for improvement. Beguš, Dąbkowski, Rhodes refute a common idea that language models lack the ability for systematic reasoning [4]. Metalinguistic ability is not just about mere use of language, and it possesses the ability to reflect on and formally analyze the structure and rules of language. This ability can be used for identifying the ambiguity of a sentence and deriving the rules governing phonological changes. The traditional perspective is that the metalinguistic ability is only owned by humans. However, Beguš, Dąbkowski, Rhodes proved that large language models especially the o1 can gain this ability by implicit training data [4]. In that case, it can give a solution to a problem about whether machine can really understand languages. This research has innovative significance. It proves that LLMs possess meta-language analysis capabilities, breaking through the traditional limitation of language models that "can only use language". The chain-like thinking mechanism of o1 provides a reproducible technical path for AI to simulate the complex reasoning of humans. The analytical capabilities of LLMs may provide new empirical tools for theoretical linguistics (such as ambiguity and recursive mechanisms).

2.1. Advantages of the o1 Model

The o1 model of OpenAI significantly outperforms GPT-4 and Llama 3.1 in tasks such as syntactic tree drawing and phonological rule analysis, demonstrating metalinguistic capabilities (that is, not only being able to use language, but also being able to explicitly reason about language structure). The main reason is that the o1 model solves problems by Chain-of-Thought. It will firstly break complicated sentences into simpler parts. Then it will generate intermediate reasoning steps (such as marking the index traces of syntactic movements). Finally, it will iteratively optimize the output based on the intermediate results. There are also some other elements of o1 model’s extraordinary performance. The first is the analysis of ambiguous structures. It requires the model to generate all possible syntactic trees using the forest package of LaTeX. The second is Linguistic Recursion. There are many different types of recursions like adjective recursion. The requirement of this is adding additional layers of the same type of recursion after identifying the recursion. The third point is Syntactic Movement. There are also different types of Syntactic Movement.

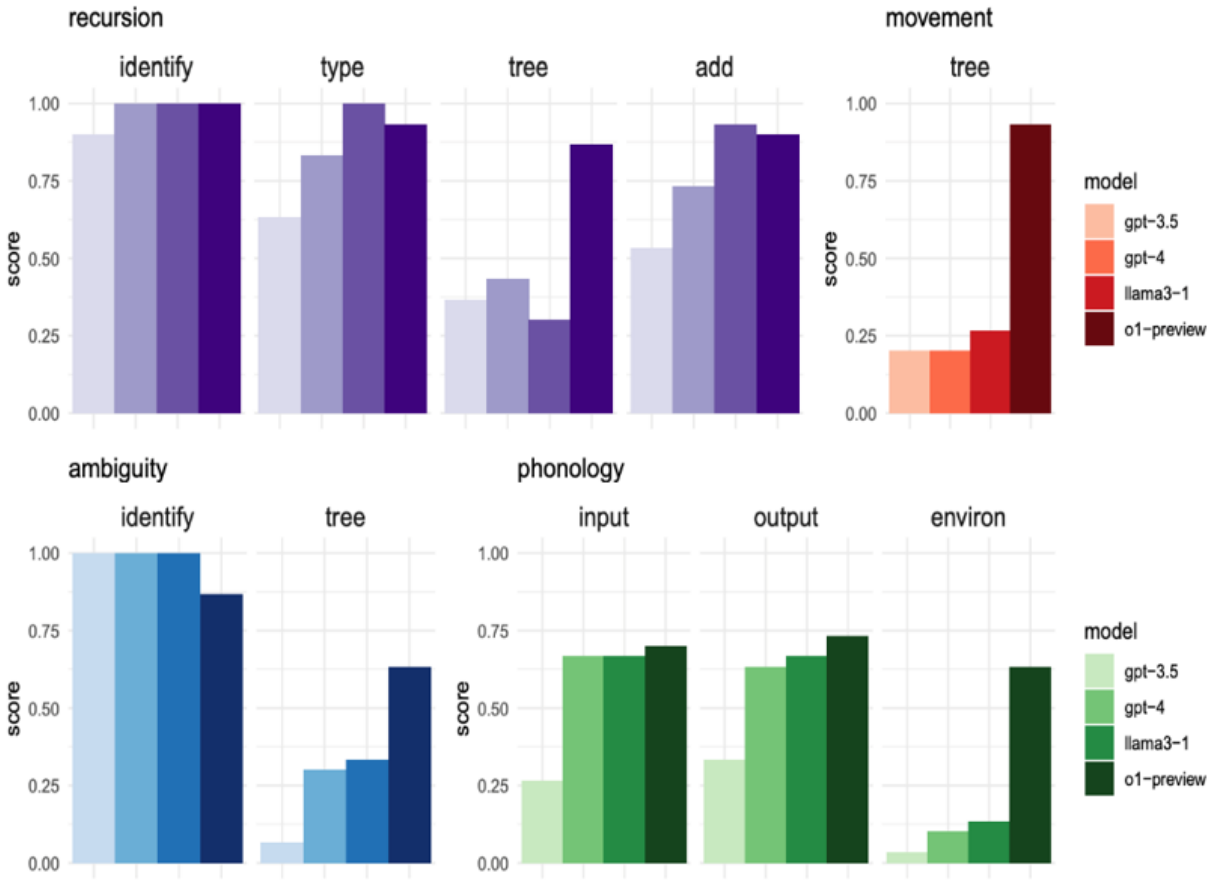


Figure 1. Performance of the four models (gpt-3.5-turbo-0125, gpt-4-0314, meta. llama3-1-405b-instruct-v1, o1-preview-2024-09-12) on four tasks (ten subtasks in total).

As shown in [1, o1-preview performs far better than gpt-3.5, gpt-4, and llama3-1 in movement and environ tasks. The essential part of this point is whether the original positions of the moving components should be marked with index traces or not. The last point is Phonological Rules. The authors made up language datasets. Then they declared the reasoning requirements. For example, clearly definition of the input (in its underlying form), the output (in its surface form), and the environment (such as "the stress is emphasized when the fricative occurs between vowels"). In order to make this research more rigorous, the authors used many ways to achieve this. The research used the Double-blind review. Three linguistics graduate students independently evaluated to ensure the objectivity of the results. What's more, the authors also used statistical methods. They used the logistic regression mixed-effects model for analysis; variables such as sentence length and model version were controlled. They also provided a controlled group. The authors gave analysis of failure cases that Llama 3.1 in the recursive layer expansion task achieved only 23% accuracy, revealing its lack of systematic reasoning ability.

2.2. Cohen's κ Coefficient Evaluation

As contrast to findings of Beguš, Dąbkowski, Rhodes [4]; Mizumoto and Teng said that the performance of open-source models such as DeepSeek-V3 surpasses that of commercial models like GPT-4o, overturning the traditional perception that "commercial models are superior" [7]. The evaluation index is Cohen's κ coefficient (used to measure the consistency of human-machine classification). The authors asked 143 first-year students majoring in English at Japanese universities to describe the procedures of completing English passages. There were two applied linguistics PhDs independently classified them into three categories based on the self-regulated learning cycle model: planning, monitoring, and evaluation. The authors then compared the annotation results of LLM (such as DeepSeek-V3) and two linguistics PhDs. Cohen's κ coefficient measures the consistency between

two evaluators in a classification task. This research chose κ rather than the percentage of accuracy since the accuracy rate only counts cases with 100% consistency, while κ has accounted for random factors (such as uneven distribution of categories). For example, if both evaluators classify 90% of cases as "planning", the accuracy rate could be as high as 81% (even if there is no actual discriminatory power), but the κ value would be close to 0. As for κ, the higher of this, the better of the model. However, even the best DeepSeek-V3 model achieved only moderate consistency (κ = 0.68) when the classification learners provided open responses; other models (such as Claude Sonnet) performed even worse (κ = 0.37 - 0.61). The research gives some possible explanation of this result. The first one is multiple process confusion which the model is difficult to handle the combined response that includes both planning and monitoring. Another possible reason is surface feature dependence which the DeepSeek-V3 software wrongly categorizes all responses containing the word "check" as belonging to the monitoring category, ignoring the context. In order to improve the method, the authors said that it's a useful method to only adapt classification results with high confidence levels (>90%) of the model; the rest will be subject to manual review. This research also has innovative significance. It challenged a consensus that commercial models are better. It proves that open-source models can outperform commercial models in specific tasks and challenge industry perceptions. This research also proposed a hybrid analysis framework of "LLM screening + manual verification" to strike a balance between efficiency and accuracy. Nathan Atox and Mason Clark say that ChatGPT excels in maintaining grammatical accuracy and consistency, making it particularly suitable for applications requiring rigorous language precision, while Google Gemini demonstrates superior contextual comprehension and logical reasoning abilities [8].

3. Discussion

As the paper from Beguš, Dąbkowski, Rhodes [4], the analytical capabilities of LLMs may provide new empirical tools for theoretical linguistics (such as ambiguity and recursive mechanisms). Svitlana Krasniuk and Svitlana GoncharenkoArtificial say that intelligence is becoming increasingly central to philological research. It opens new possibilities for the analysis of language, literature, and the corresponding culture [9]. However, it's still pretty hard to use large language models to improve theoretical linguistics. It's true that recursive mechanisms can help these models to have the ability of basic reasoning. But when these models face something really ambiguous, the efficiency and accuracy of reasoning are still low. Schneider, Gerold says that the obvious disadvantage of using an automatic tool: that there is always a certain level of errors, which entails that evaluations are essential [10]. Nowadays, the usage of large language models in linguistics is still more about theoretical. Researchers and linguistics are more willing to think of these models as a tool. They use this tool to check and do some basic reasoning. Researchers can't rely on these models to achieve breakthroughs. What's more, large language models have to balance between efficiency and accuracy. In many times, high efficiency means low accuracy, and vice versa. And because different models have different performances, linguistics have to choose which large language models to use. It's hard to identify which model is the best. Some models have high efficiency but low accuracy. In contrast, some other models have low efficiency but high accuracy. In the current era, because of the restriction of hardware. These problems are hard to resolve. Researchers can start to use models that have high accuracy to solve ambiguous problems. Although it will take tons of time, the process may be finally faster. Researchers can also use multiple large language models simultaneously. What's more, developing more advanced hardware is also necessary.

4. Conclusion

This paper contains two perspectives of large language models in linguistics. Some large language models like o1 model of OpenAI can outperform other large language models in basic reasoning and some other aspects. However, it still lacks the ability of systematic reasoning. This paper also shows

the perception that commercial models are not always superior. The comparison of Cohen's κ coefficient shows that open-source models can outperform commercial models in specific tasks and challenge industry perceptions. What's more, different models have their own advantages and disadvantages. Models with better accuracy may have low efficiency. However, as the improvement of large language models and hardware, it's possible that these models can overcome these drawbacks. They can do some pretty complicated reasoning. They may even help linguistics decode ancient literature and ancient culture.

References

[1] Q. A. H. Abdulkhaleq, D. A. Shoayee, B. S. Wala, et al., "Automated Sarcasm Recognition Using Applied Linguistics Driven Deep Learning with Large Language Model." Fractals. (2024).

[2] G. S. Alexander, V. G. Artem, B.R. Roman, et al., "A deep learning method based on language models for processing natural language Russian commands in human robot interaction". Natural Language Processing Journal. (2024).

[3] Z. Lin, "A Methodological Review of Machine Learning in Applied Linguistics". English Language Teaching. (2021).

[4] B. Gašper, D. Maksymilian, R. Ryan, "Large Linguistic Models: Investigating LLMs' metalinguistic abilities". https://arxiv.org/abs/2305.00948. (2025).

[5] A. Ben, B. Liam, "Large language models are better than theoretical linguists at theoretical linguistics". Theoretical Linguistics. (2024).

[6] V. Sowmya, "Machine Learning and Applied Linguistics". https://arxiv.org/abs/1803.09103. (2018).

[7] M. Atsushi, F. Mark, "Large language models fall short in classifying learners' open-ended responses". Research Methods in Applied Linguistics. (2025).

[8] A. Nathan, C. Mason, "Evaluating Large Language Models through the Lens of Linguistic Proficiency and World Knowledge: A Comparative Study". A Comparative Study. (2024).

[9] K. Svitlana, Svitlana, Goncharenko, "ETHICS OF USING LARGE LANGUAGE MODELS IN MACHINE LINGUISTICS". The 5th International Conference on Scientific Practice. (2024).

[10] G. Schneider, "Applying Computational Linguistics and Language Models: From Descriptive Linguistics to Text Mining and Psycholinguistics". University of Zurich. (2014).