What do LLMs know about English language?

Julia May, Krzysztof Sopyła

AI Science Team, Pearson

Introduction

In order to generate high-quality educational content tailored to individual student's needs using LLMs we faced a challenge of efficient evaluation of language capabilities of given LLM before using it in real-life applications. Various available benchmarks for LLM evaluation focus on a broad area of LLM capabilities like mathematical reasoning, academic knowledge and factual questing answering [1] [2]. However, there are significantly less benchmarks and datasets that focus on linguistic capabilities of LLMs and their knowledge of topics related to English language and grammar. We decided to create our custom framework for evaluation of linguistic knowledge and language capabilities of LLMs.

We focused on 2 main categories of language knowledge: language level understanding and language features knowledge. We also ran an experiment where we compared performance of different state-of-the-art LLMs on our custom evaluation framework. Our approach was inspired by popular frameworks like [3], [4] and [5] which aggregate different datasets for extensive few-shot evaluation of LLMs. We implemented our framework using PromptFlow and AzureAIStudio.

Evaluation categories

- **Language level understanding**
  - Check how well LLMs understand the concept of English language difficulty level (coarse-grained level - beginner, intermediate, advanced, GSE level - Pearson Global Scale of English, CEFR level - Common European Framework of Reference for Languages), for example by scoring difficulty level of the text or generating text at a given difficulty level.
- **Language features knowledge**
  - Check how well LLMs understand different language concepts like grammar and language use, for example by generating sentences with given grammar structure or giving dictionary explanations of words.

Tasks and metrics - language level understanding

Task name	Task metric
Scoring general level of the text	Accuracy
Scoring CEFR level of the text	Accuracy
Scoring GSE level of the text	Accuracy
Comparison of two texts with different level	Accuracy
Writing text at general level	Custom GSE leveler
Writing text at GSE level	Custom GSE leveler
Writing text at CEFR level	Custom GSE leveler
Rewriting text to given general level	Custom GSE leveler
Rewriting text to given readability level	Custom readability scorer

Tasks and metrics - language features knowledge

Task name	Task metric
Grammar structure explanation	Custom NLI model
Dictionary words explanation	Custom NLI model
Writing sentences with given structure	Custom syntax matcher
Detecting grammar structures	Custom list matching
Changing grammar tenses	Accuracy
Finding examples of structures in the text	Accuracy

Future work

- Expand datasets for current tasks with more annotated examples
- Implement tasks for evaluating LLM writing capabilites

Performance of different LLMs measured by our framework

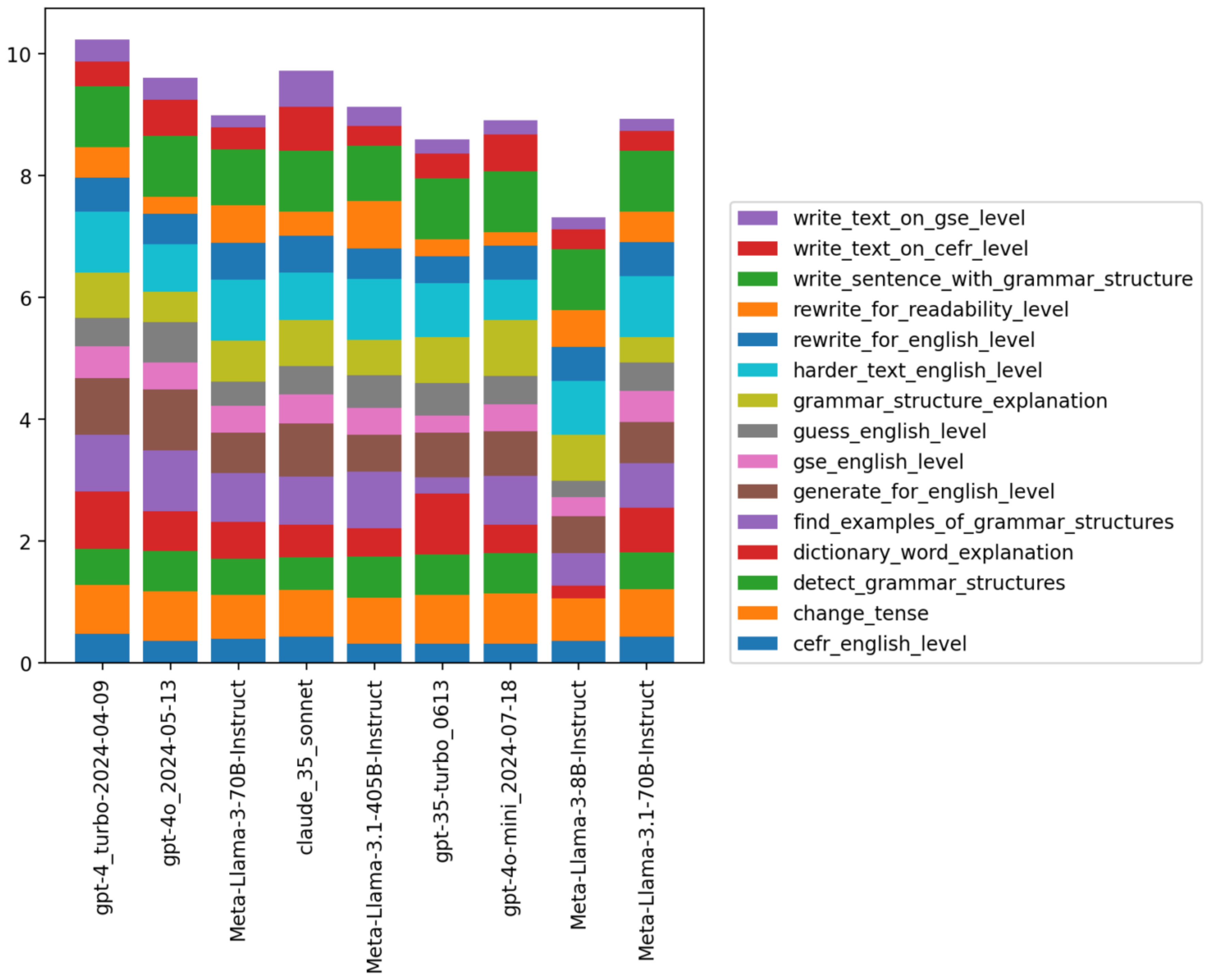

Figure 1. Bar plot showing aggregated performance of LLMs on all tasks

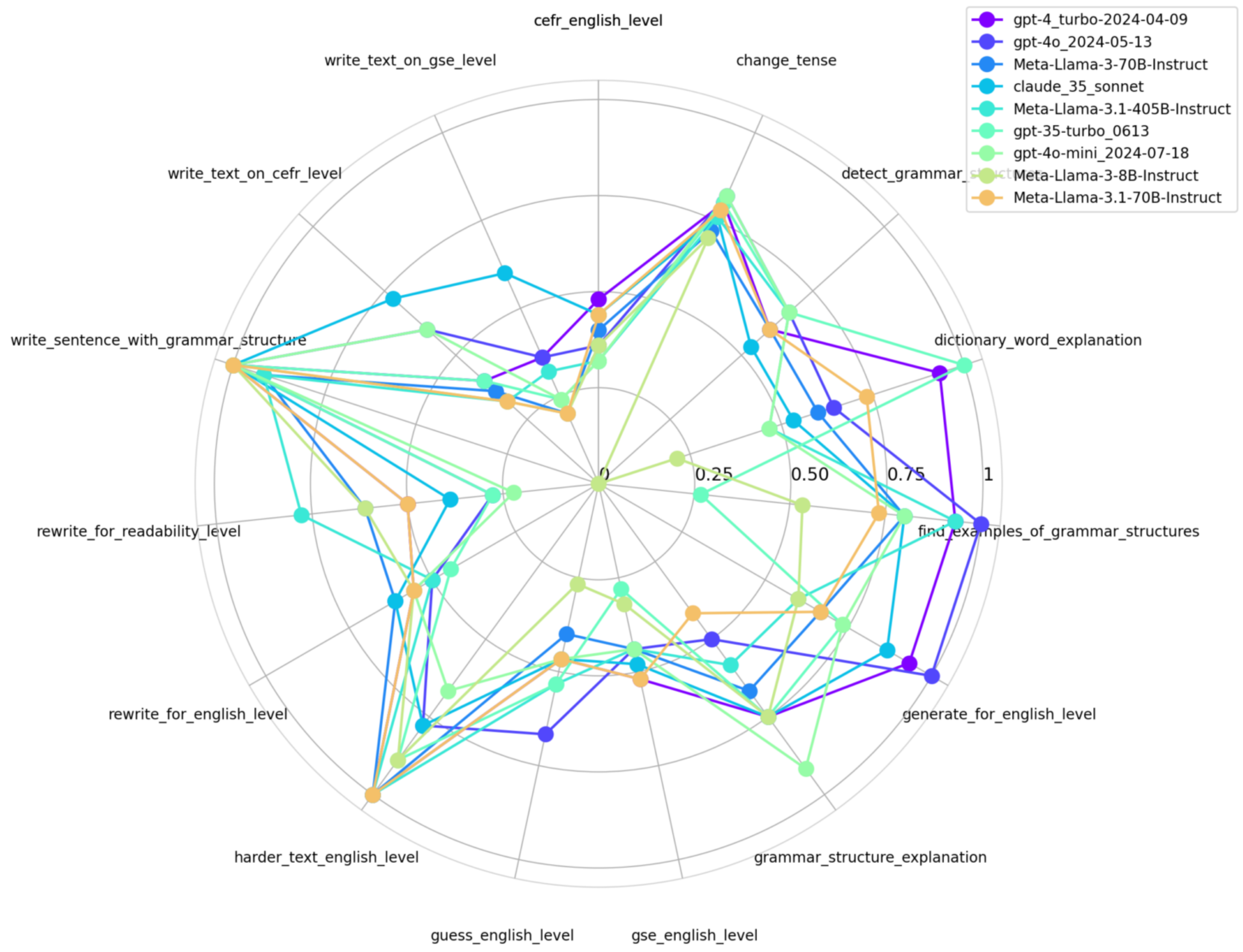

Figure 2. Radar plot showing performance of LLMs on each task

References

[1] Y. Chang, X. Wang, J. Wang, *et al.*, *A survey on evaluation of large language models*, 2023. arXiv: 2307.03109 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2307.03109.

[2] Z. Guo, R. Jin, C. Liu, *et al.*, *Evaluating large language models: A comprehensive survey*, 2023. arXiv: 2310.19736 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2310.19736.

[3] L. Gao, J. Tow, B. Abbasi, *et al.*, *A framework for few-shot language model evaluation*, version v0.4.3, Jul. 2024. DOI: 10.5281/zenodo.12608602. [Online]. Available: https://zenodo.org/records/12608602.

[4] S. Ye, D. Kim, S. Kim, *et al.*, *Flask: Fine-grained language model evaluation based on alignment skill sets*, 2023. arXiv: 2307.10928 [cs.CL].

[5] C. Fourrier, N. Habib, T. Wolf, and L. Tunstall, *Lighteval: A lightweight framework for llm evaluation*, version 0.5.0, 2023. [Online]. Available: https://github.com/huggingface/lighteval.