Can Large Language Models Understand Real-World Complex Instructions?

Qianyu He[@FDU] , Jie Zeng[@FDU] , et al. AAAI, 2024

徐梓航

华中科技大学计算机科学与技术学院

2024 年 12 月 02 日

Can Large Language Models Understand Real-World Complex Instructions?

LLMs 难以处理复杂的指令，这些指令可以是需要多个任务和约束的复杂任务描述，也可以是包含长上下文、噪声、异构信息和多回合格式的复杂输入。

由于这些特性，LLM 常常忽略任务描述中的语义约束，生成错误的格式，违反长度或样本计数约束，并且对输入文本不忠实。

现有的基准测试不足以评估 LLMs 对评估复杂指令的能力，为此，论文提出了 CELLO (ComplEx instruction understanding ability of Large Language MOdels)。

1. Introduction
   - Complex Instructions
   - Existing Benchmarks
2. Challenge
3. Methods
4. Experiment
5. Conclusion
6. Thoughts
7. References

Instruction generally consists of two parts:

- Task description (mandatory)
- Input text (optional)

Two categories of complex instructions:

- complex task descriptions
- complex input

Regarding complex task descriptions, models need to undertake multiple tasks and there can be diverse restrictions describing the task:

- semantics constraints
- format constraints
- quantity constraints

Regarding complex input, the input text generally have:

- long context
- noise
- error accumulation caused by pipeline method
- heterogeneous information (**异构信息**) {e.g. a combination of structured and unstructured data}
- in the form of multi-turn

The complexity of real-world instructions accounts for prevalent errors observed in LLMs. LLMs May:

- Ignore semantic constraints from task description.
- Generate answers in incorrect format.
- Violate the length or sample count constraints, especially when multiple tasks are required to be performed.
- Models can be unfaithful to the input text, especially when it is long, noisy, heterogeneous or in the form of multi-turn.

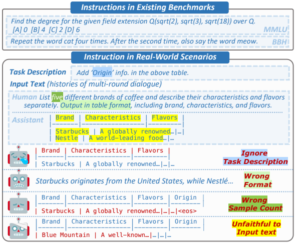

图 1: Real-world applications often involve a diverse range of complex instructions.

1 Introduction
   Complex Instructions
   Existing Benchmarks

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

Existing benchmarks are insufficient for effectively assessing the ability of LLMs to understand complex instructions:

- close-ended (封闭式)
- Contain common and simple instructions, which fail to mirror the complexity of real-world instructions.

They only encompass isolated features:

- count restriction
- semantic restriction
- long text understanding

Real-world instructions comprehensively cover these features.
Overall, none of the existing benchmarks systematically study the complex instructions understanding ability of LLMs.

- Complex instructions in real-world scenarios are open-ended, thus the criteria commonly used for close-ended benchmarks are not suitable in such cases.
- Many studies adopt GPT4 evaluation for automated open-ended assessment, which introduces bias problems.
- The binary pass rate adopted by the benchmarks containing complex instructions is strict and coarsegrained, resulting in universally low scores for smaller LLM without discrimination.

CELLO (ComplEx instruction understanding ability of Large Language MOdels)

- pioneer
- Propose a two-stage framework for constructing the evaluation dataset for LLM's complex instruction understanding.
- Design four evaluation criteria and corresponding automatic metrics for assessing LLMs' ability to understand complex instructions in a comprehensive and discriminative way.
- Tested the benchmark testing framework.

Related Works

- Evaluation for LLMs
- Complex Instruction Following
- Evaluation for Constrained Instructions

Dataset Construction

Diversify the collected complex instructions through In-breadth Evolution and complicate the collected simple instructions through In-breadth Evolution.

Data Source and Selected Tasks

Include common NLP tasks found in existing benchmarks, while incorporating instructions with more complex task descriptions or input beyond those benchmarks.

Dataset Construction

CELLO include nine tasks, classified into six categories:

- Complex NLP Tasks
  - long text summarization
  - long text closed-domain question answering
  - long text keywords extraction
  - complex information extraction
- Meta-prompt
- Planning
- Structured Input
- Well-guided Writing
- Detailed Brainstorming

Dataset Construction

Data Evolution

The collected complex instructions have two limitations:

- For those collected from real-world projects, the human-elaborated task descriptions are complex but alike.
- For those collected from usage logs, many simple instructions are not effectively utilized.

Introduce two perspectives to evolve data, thereby achieving a more robust and reliable evaluation:

- In-breadth Evolution (Aims to diversify the collected complex instructions)
    - task description relocation
    - task description paraphrasing
    - task emulation

Dataset Construction

- In-depth Evolution (Aims to complicate the simple instructions to increase the data scale)
  - constraints addition
  - multi-round interaction

Evaluation System

Criteria

Encompass common errors made by models:

- count limit
- answer format
- task-prescribed phrases
- input-dependent query

Evaluation System

Evaluation Metrics

每个样本 s_i 由指令 I_i、模型答案 a_i 和给定的历史 h_i 组成，其中 h_i 是多轮对话中的前几轮 $\{(I_0, a_0'), ..., (I_i-1', a_i-1')\}$。对于每个样本 s，其每个标准的分数由多个子分数 C 组成，C 是一个包含 $\{c_1, c_2, ..., c_i\}$ 的集合。

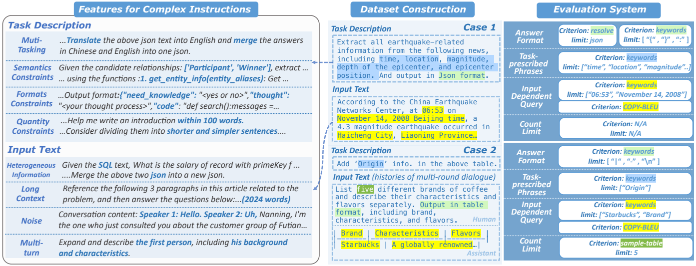

图 2: Eight features for complex instructions, an evaluation dataset covering nine tasks, four evaluation criteria along with their corresponding metrics.

Evaluation System

Count Limit

Four sub-scores:

- word count score
- sentence count score
- sample count score
- revise score

Evaluation System

Answer Format

Two sub-scores:

- parseability (模型输出是否可解析，取 0 或 1)
- keywords (计算模型输出中包含的关键词数量后/总数)

最终两者求均值。

Evaluation System

Input-dependent Query

Two sub-scores:

- keywords($f_{keywords}(a_i, l_q)$, the scoring keywords l_q are extracted from input text)
- COPYBLEU (值随着模型输出与输入文本相似度的增加而减少，即如果模型输出与输入文本过于相似，COPYBLEU 的值会较低，从而对最终得分产生负面影响)

Evaluation System

Task-prescribed Phrases
The more mandatory phrases covered in the answers, the better the model follows complex instructions.

Keywords($f_{keywords}(a_i, l_t)$) is applied where l_t is the scoring keywords extracted from the task description.

Evaluation of the Benchmark

根据四个标准，每个样本由三个 annotators 标记。具体地说，只有当至少两个 annotators 在标准计数限制和输出格式可解析性上达成一致时，我们才保留样本。对于涉及关键字覆盖率的标准，我们只保留至少两个 annotators 一致同意的关键字。

Statistics of the Benchmark

- Dataset has two categories depending on whether the criteria are mainly in the task description or the input text.
- CELLO benchmark is the first to systematically test LLMs' ability to follow complex instructions, which are generally longer and more complex than other benchmarks.
- The tasks we cover are open-ended, which are more realistic and practical.
- Evaluation is also more objective and fine-grained.

Category	Tasks	#Samples	#Format	#Task	#Input	#Count	Avg TD Len.	Avg IP Len.	Avg Ins Len.
Complex Task Description	Extraction	49	49	35	49	N/A	125	169	295
	Planning	52	52	46	48	N/A	1070	534	1606
	Meta.	20	20	15	6	2	765	166	933
	BS(S)	20	20	20	1	15	70	N/A	70
	Writing(S)	23	2	23	2	12	82	25	107
Complex Input	Keywords	15	15	15	15	N/A	546	943	1579
	QA	89	N/A	N/A	89	N/A	25	881	814
	Sum.	108	N/A	N/A	108	N/A	45	514	562
	Struture	38	6	N/A	38	N/A	29	1360	1390
	BS(M)	52	50	50	10	36	31	559	31
	Writing(M)	57	3	35	48	43	30	656	51
Overall		523	217	239	414	108	256	528	676

Evaluated Models

These models are categorized into three groups:

- Chinese-oriented Models (From Scratch, FS) {Trained entirely from scratch using Chinese corpora}
- Chinese-oriented Models (Continue Pretraining, CP) {Continue pretraining on Chinese corpora utilizing an English-oriented base model}
- English-oriented Models

Task-categorized Performance

- General Comparisons
  - Complex instruction comprehension is not language-dependent.
  - There is a strong correlation between the ability to comprehend complex instructions and the instruction.
- Complex Task Description
  - The ability to understand complex task descriptions can transfer across different languages.
  - The supported text context length does not significantly impact the ability to comprehend complex task descriptions.
- Complex Input Text
  - More Chinese training data assists the models in comprehending long and noisy Chinese texts.
  - Within the same model series, larger scales generally improve performance, while longer supported context length can result in performance drops in many cases.

Criteria-categorized Performance

- Regarding Answer format, the English-oriented Models significantly perform better than Chinese-oriented Models. This demonstrates the English-oriented Models' ability to follow few-shot examples and generate code, as well as partially explains why their complex instruction-following ability can transfer across languages.
- For Task-prescribed phrases, Chinese data helps the models understand Chinese semantic restrictions.
- Finally, the performance differences between models for Count limit criteria are not big compared to other criteria, which shows that the models have similar comprehension of numerical concepts.

Comparisons between Benchmarks

- On benchmarks focusing on Chinese knowledge (C-eval, CMMLU, and GAOKAO), smaller models achieve similar or even better performance compared to GPT-3.5-turbo.
- On challenging benchmarks like complex reasoning (BBH, GSM8k) and programming ability (HumanEval), there is a lack of distinction between smaller models.

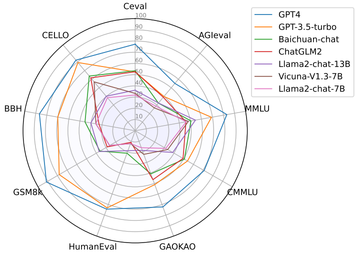

图 3: The performance of models on mainstream benchmarks.

Fine-grained Evaluation

- Different models have different strengths for different criteria.
- Different models also excel in specific tasks.

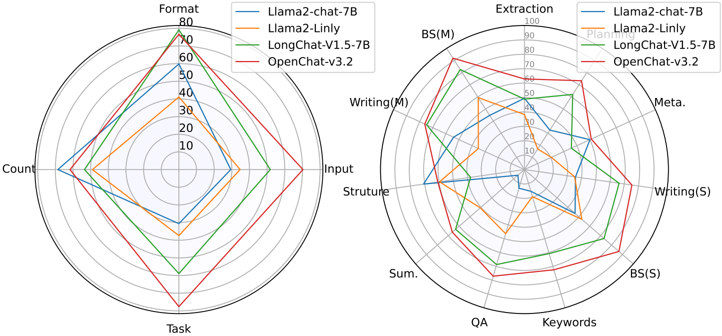

图 4: The performance of LLMs grounded on the same base model regarding different tasks and criteria.

- complex instructions following ability of LLMs
- CELLO Benchmark
- Conduct extensive experiments to compare the performance of representative models.

思考

- **数据集的多样性和代表性**：论文中提出的 CELLO 基准测试涵盖了多种复杂指令的特征，并从真实世界场景中构建了评估数据集。然而，数据集的多样性和代表性始终是一个可以进一步探讨的话题。未来的工作可以探索如何确保数据集覆盖更广泛的语言、地区和文化背景，以及如何平衡不同领域和任务类型的样本。
- **评估标准的细化**：尽管论文提出了四个评估标准（Count limit、Answer format、Task-prescribed phrases、Input-dependent query），但这些标准是否可以进一步细化，以便更精确地捕捉模型在处理复杂指令时的细微差异，是一个值得考虑的问题。
- **模型的可解释性**：论文主要关注模型的性能评估，但对于模型的决策过程和内部机制的可解释性讨论不多。未来的研究可以探索如何提高模型在处理复杂指令时的透明度和可解释性，以便更好地理解模型的行为。

思考

- **模型的适应性和泛化能力**：论文中的实验主要关注模型在特定数据集上的表现。未来的研究可以探讨模型在面对新的、未见过的复杂指令时的适应性和泛化能力，以及如何通过持续学习或迁移学习来提高这些能力。
- **多模态和跨领域指令的理解**：随着多模态学习和跨领域应用的兴起，未来的研究可以考虑如何评估和提高模型在处理包含图像、声音等多种模态信息的复杂指令时的性能。
- **实时性能和资源消耗**：论文中的评估主要关注模型的准确性和完成度。在实际应用中，模型的实时性能和资源消耗（如计算时间、内存使用等）也是非常重要的考量因素。未来的工作可以探讨如何在保证性能的同时优化模型的效率。

[HZH+23] Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Lida Chen, Xintao Wang, Yuncheng Huang, Haoning Ye, Zihan Li, Shisong Chen, Yikai Zhang, Zhouhong Gu, Jiaqing Liang, and Yanghua Xiao.
Can Large Language Models Understand Real-World Complex Instructions?
arXiv e-prints, page arXiv:2309.09150, September 2023.

[lat23] latexstudio.
Hust-beamer-theme, 2023.

Thanks!