组会

第 1 次

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

- 工欲善其事，必先利其器

- 工欲善其事，必先利其器
- 我首先改了一个 LaTeX Beamer 模板用于今后的组会 PPT

- 工欲善其事，必先利其器
- 我首先改了一个 LaTeX Beamer 模板用于今后的组会 PPT
- GitHub 项目地址位于 https://github.com/M0rtzz/GroupMeetingSlide

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)。

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

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 l_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 $\{(l_0, a'_0), ..., (l_i-1', a_i-1')\}$。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。

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.

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.

Fine-grained Evaluation

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

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

- 一月：完成文献调研
- 二月：复现并评测各种 Beamer 主题美观程度
- 三、四月：美化 THU Beamer 主题
- 五月：论文撰写

[unk15] unknown.
Thu beamer theme.
In *unknown*, 2015.

Thanks!