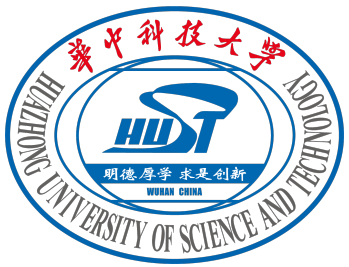
组会
第 1 次

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`

1 Introduction

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

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

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

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.

1 Introduction

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

- 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.

1 Introduction

2 Challenge

3 Methods
- Related Work
- CELLO Benchmark

4 Experiment

5 Conclusion

6 Thoughts

7 References

1 Introduction

2 Challenge

3 Methods
- Related Work
- CELLO Benchmark

4 Experiment

5 Conclusion

6 Thoughts

7 References

Related Works

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

1 Introduction

2 Challenge

3 Methods
- Related Work
- CELLO Benchmark

4 Experiment

5 Conclusion

6 Thoughts

7 References

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

1 Introduction

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

- 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

1 Introduction

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

- 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

1 Introduction

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

- 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

1 Introduction

2 Challenge

3 Methods

4 Experiment

5 Conclusion

6 Thoughts

7 References

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

Thanks!