大语言模型评价研究进展

刘群 LIU Qun

华为诺亚方舟实验室

IMLIP2024 国际多语种智能信息处理大会

2024.11.16，北京理工大学

NOAH'S ARK LAB

HUAWEI

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

大语言模型的语言生成能力评价

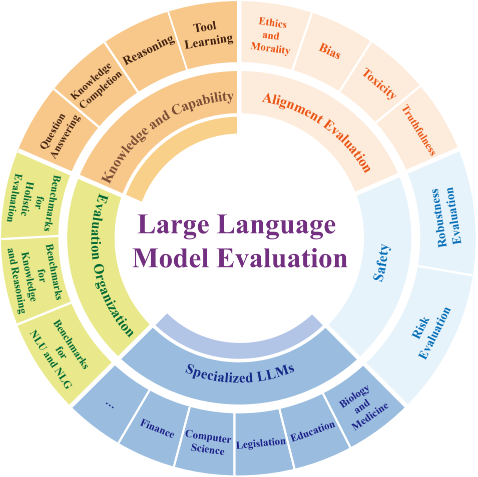
大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

大语言模型评价任务的多样性

与传统NLP任务的评价不同，大语言模型是一个通用模型，并不存在单一的评价标准。

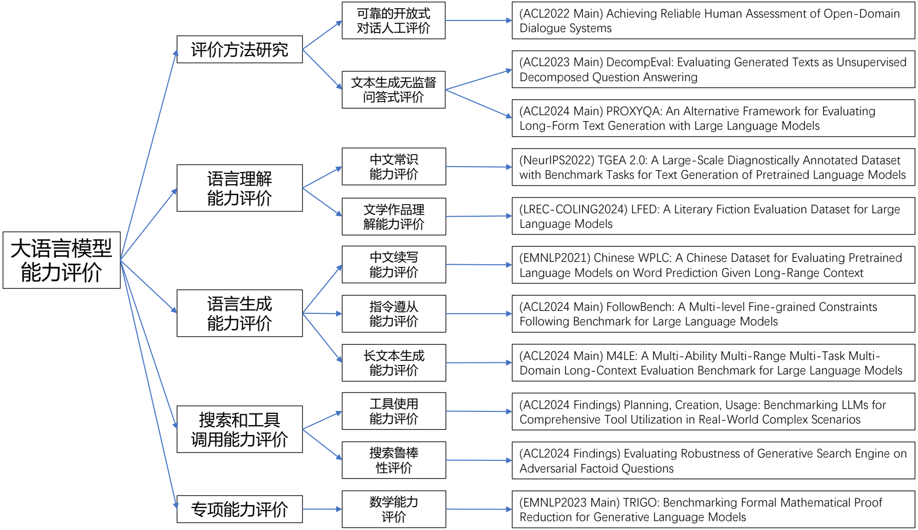


Guo et al., Evaluating Large Language Models: A Comprehensive Survey, arXiv:2310.19736v3, 2023

大语言模型评价的挑战

- 大语言模型的评价面临的挑战主要体现在：
  - 任务的多样性
  - 评价的主观性
  - 评测数据的代表性
  - 评测数据的时效性
  - 评测数据泄露和污染
  - 评测结果的可解释性

我们的工作



- 大语言模型能力评价
  - 评价方法研究
    - 可靠的开放式对话人工评价
      - (ACL2022 Main) Achieving Reliable Human Assessment of Open-Domain Dialogue Systems
    - 文本生成无监督问答式评价
      - (ACL2023 Main) DecompEval: Evaluating Generated Texts as Unsupervised Decomposed Question Answering
      - (ACL2024 Main) PROXYQA: An Alternative Framework for Evaluating Long-Form Text Generation with Large Language Models
  - 语言理解能力评价
    - 中文常识能力评价
      - (NeurIPS2022) TGEA 2.0: A Large-Scale Diagnostically Annotated Dataset with Benchmark Tasks for Text Generation of Pretrained Language Models
    - 文学作品理解能力评价
      - (LREC-COLING2024) LFED: A Literary Fiction Evaluation Dataset for Large Language Models
  - 语言生成能力评价
    - 中文续写能力评价
      - (EMNLP2021) Chinese WPLC: A Chinese Dataset for Evaluating Pretrained Language Models on Word Prediction Given Long-Range Context
    - 指令遵从能力评价
      - (ACL2024 Main) FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models
    - 长文本生成能力评价
      - (ACL2024 Main) M4LE: A Multi-Ability Multi-Range Multi-Task Multi-Domain Long-Context Evaluation Benchmark for Large Language Models
  - 搜索和工具调用能力评价
    - 工具使用能力评价
      - (ACL2024 Findings) Planning, Creation, Usage: Benchmarking LLMs for Comprehensive Tool Utilization in Real-World Complex Scenarios
    - 搜索鲁棒性评价
      - (ACL2024 Findings) Evaluating Robustness of Generative Search Engine on Adversarial Factoid Questions
  - 专项能力评价
    - 数学能力评价
      - (EMNLP2023 Main) TRIGO: Benchmarking Formal Mathematical Proof Reduction for Generative Language Models

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

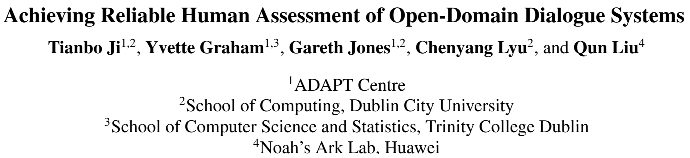
大语言模型的语言生成能力评价

大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

可靠的开放域对话系统人工评价方法



Achieving Reliable Human Assessment of Open-Domain Dialogue Systems

Tianbo Ji[1,2], **Yvette Graham**[1,3], **Gareth Jones**[1,2], **Chenyang Lyu**[2], and **Qun Liu**[4]

[1]ADAPT Centre
[2]School of Computing, Dublin City University
[3]School of Computer Science and Statistics, Trinity College Dublin
[4]Noah's Ark Lab, Huawei

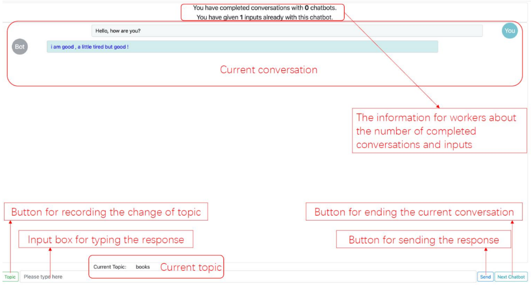
- 开放域对话系统的评估极具挑战性，迫切需要开发更好的评估技术。尽管在近期的竞赛中，为了对系统进行可靠的现场评估付出了大量努力，但这种方法还是被放弃了，因为其可靠性太差，无法得出合理的结果。然后，自动评估指标更不能很好地评估一场对话是否属于高质量对话。
- 为此，我们成功开发出了一种现场人工评估方法，它既高度可靠，又切实可行且成本低廉。自我复制实验显示，结果的可重复性近乎完美，相关系数达到了r=0.0969。
- 此外，由于缺乏合适的统计显著性检验方法，在对话评估中，系统性能可能因偶然因素而得到提升的可能性很少被考虑进去，而我们提出的评估方法可以方便地引入显著性检验。
- 更进一步，我们对（i）有角色设定和无角色设定的最先进模型进行了比较，以衡量角色设定对对话质量的贡献，以及（ii）规定话题和自由选择话题的情况。有趣的是，就角色设定而言，结果表明角色设定并没有像预期的那样对对话质量起到积极的促进作用。

Ji et al., Achieving Reliable Human Assessment of Open-Domain Dialogue Systems, ACL2022 Proceedings

可靠的开放域对话系统人工评价方法

User interface – interact with a model

www.adaptcentre.ie



You have completed conversations with **0** chatbots.
You have given **1** inputs already with this chatbot.

You: Hello, how are you?

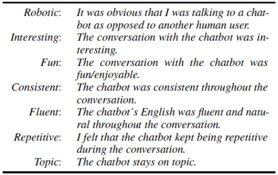
Bot: i am good , a little tired but good !

Current Topic: books

Current conversation
The information for workers about the number of completed conversations and inputs
Button for recording the change of topic
Input box for typing the response
Current topic
Button for ending the current conversation
Button for sending the response
Topic
Please type here
Send
Next Chatbot

Likert Statement & Continuous Rating Scale

www.adaptcentre.ie



Robotic:	*It was obvious that I was talking to a chatbot as opposed to another human user.*
Interesting:	*The conversation with the chatbot was interesting.*
Fun:	*The conversation with the chatbot was fun/enjoyable.*
Consistent:	*The chatbot was consistent throughout the conversation.*
Fluent:	*The chatbot's English was fluent and natural throughout the conversation.*
Repetitive:	*I felt that the chatbot kept being repetitive during the conversation.*
Topic:	*The chatbot stays on topic.*

Continuous Rating Scale

- Reduce bias by score standardization
- Standard significance tests to score distributions
- Accurate quality control of crowd-sourced workers

Likert Statement

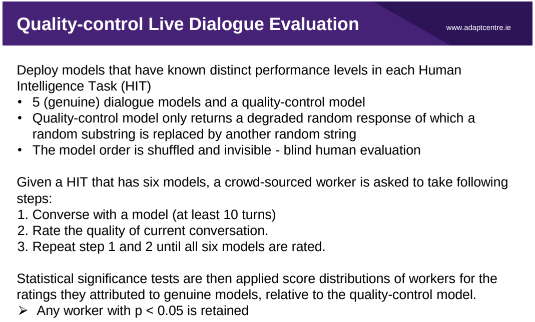
- Adjectival scale labels shown to introduce bias
- Instead use Likert declarative statement
- Workers are asked to rate agreement with statement

Live Dialogue Evaluation

- Direct Assessment by the user
- User chosen topic – genuinely open domain
- Switch topic possible

Ji et al., Achieving Reliable Human Assessment of Open-Domain Dialogue Systems, ACL2022 Proceedings

可靠的开放域对话系统人工评价方法



Quality-control Live Dialogue Evaluation

www.adaptcentre.ie

Deploy models that have known distinct performance levels in each Human Intelligence Task (HIT)

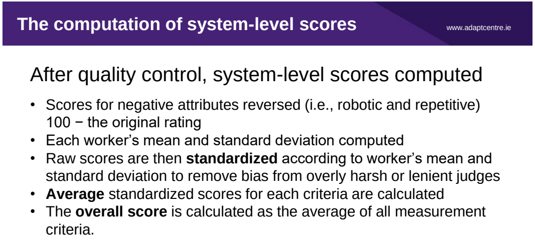
- 5 (genuine) dialogue models and a quality-control model
- Quality-control model only returns a degraded random response of which a random substring is replaced by another random string
- The model order is shuffled and invisible - blind human evaluation

Given a HIT that has six models, a crowd-sourced worker is asked to take following steps:

1. Converse with a model (at least 10 turns)
2. Rate the quality of current conversation.
3. Repeat step 1 and 2 until all six models are rated.

Statistical significance tests are then applied score distributions of workers for the ratings they attributed to genuine models, relative to the quality-control model.

➢ Any worker with $p < 0.05$ is retained



The computation of system-level scores

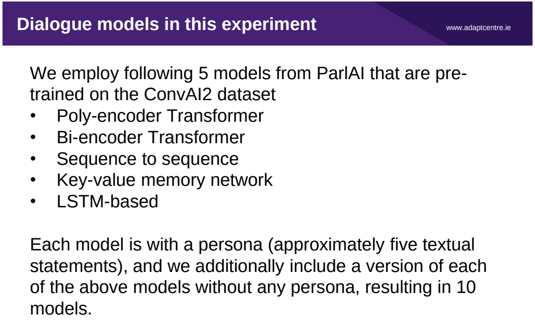
www.adaptcentre.ie

After quality control, system-level scores computed

- Scores for negative attributes reversed (i.e., robotic and repetitive) 100 − the original rating
- Each worker’s mean and standard deviation computed
- Raw scores are then **standardized** according to worker’s mean and standard deviation to remove bias from overly harsh or lenient judges
- **Average** standardized scores for each criteria are calculated
- The **overall score** is calculated as the average of all measurement criteria.

Ji et al., Achieving Reliable Human Assessment of Open-Domain Dialogue Systems, ACL2022 Proceedings

可靠的开放域对话系统人工评价方法



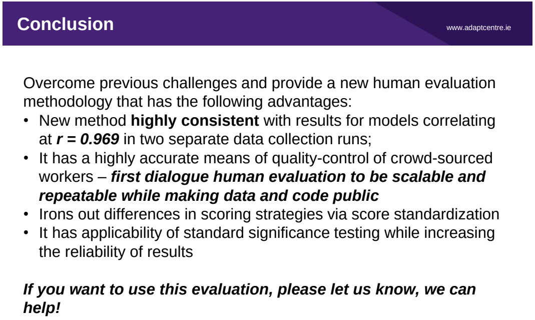
Dialogue models in this experiment

www.adaptcentre.ie

We employ following 5 models from ParlAI that are pre-trained on the ConvAI2 dataset

- Poly-encoder Transformer
- Bi-encoder Transformer
- Sequence to sequence
- Key-value memory network
- LSTM-based

Each model is with a persona (approximately five textual statements), and we additionally include a version of each of the above models without any persona, resulting in 10 models.



Conclusion

www.adaptcentre.ie

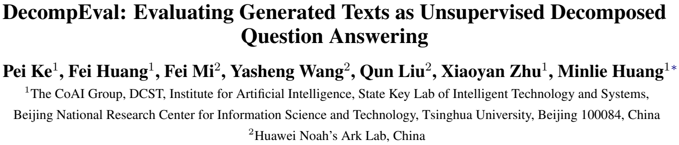
Overcome previous challenges and provide a new human evaluation methodology that has the following advantages:

- New method **highly consistent** with results for models correlating at ***r* = *0.969*** in two separate data collection runs;
- It has a highly accurate means of quality-control of crowd-sourced workers – ***first dialogue human evaluation to be scalable and repeatable while making data and code public***
- Irons out differences in scoring strategies via score standardization
- It has applicability of standard significance testing while increasing the reliability of results

If you want to use this evaluation, please let us know, we can help!

Ji et al., Achieving Reliable Human Assessment of Open-Domain Dialogue Systems, ACL2022 Proceedings

DecompEval：基于无监督分解式问答的文本生成评估方法



DecompEval: Evaluating Generated Texts as Unsupervised Decomposed Question Answering

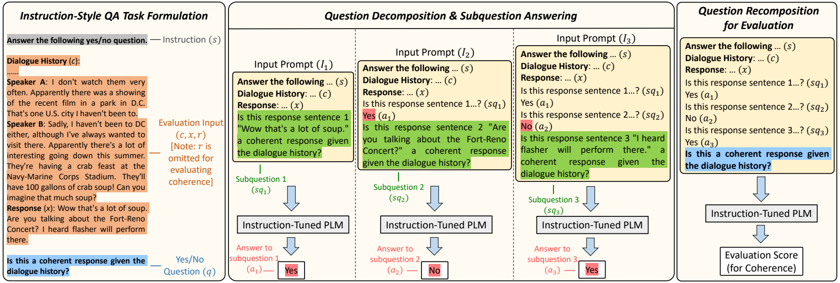
Pei Ke[1], **Fei Huang**[1], **Fei Mi**[2], **Yasheng Wang**[2], **Qun Liu**[2], **Xiaoyan Zhu**[1], **Minlie Huang**[1*]

[1]The CoAI Group, DCST, Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China
[2]Huawei Noah's Ark Lab, China

- 自然语言生成（NLG）任务现有的评估指标面临着泛化能力和可解释性方面的挑战。具体而言，大多数表现良好的指标都需要在特定自然语言生成任务和评估维度的评估数据集上进行训练，这可能会导致对特定任务数据集的过拟合。此外，现有的指标只是针对每个维度给出一个评估分数，却没有揭示用以解释该分数是如何得出的依据。
- 为了应对这些挑战，我们提出了一种简单而有效的指标，名为分解评估（DecompEval）。该指标将自然语言生成评估设定为一种指令式问答任务，并利用经过指令微调的预训练语言模型（PLM），且无需在评估数据集上进行训练，旨在提高泛化能力。
- 为了使评估过程更具可解释性，我们将所设计的关于生成文本质量的指令式问题分解为衡量每个句子质量的子问题。然后，将由预训练语言模型生成的带有答案的子问题重新组合起来作为依据，以获得评估结果。
- 实验结果表明，分解评估（DecompEval）在用于评估文本摘要和对话生成的未训练指标方面达到了最先进的性能，同时还展现出了很强的维度层面 / 任务层面的泛化能力和可解释性。

Ke et al., DecompEval: Evaluating Generated Texts as Unsupervised Decomposed Question Answering, ACL2023 Proceedings

DecompEval：基于无监督分解式问答的文本生成评估方法



Instruction-Style QA Task Formulation

Answer the following yes/no question. — Instruction (s)

Dialogue History (c):
……
Speaker A: I don't watch them very often. Apparently there was a showing of the recent film in a park in D.C. That's one U.S. city I haven't been to.
Speaker B: Sadly, I haven't been to DC either, although I've always wanted to visit there. Apparently there's a lot of interesting going down this summer. They're having a crab feast at the Navy-Marine Corps Stadium. They'll have 100 gallons of crab soup! Can you imagine that much soup?
Response (x): Wow that's a lot of soup. Are you talking about the Fort-Reno Concert? I heard flasher will perform there.

Is this a coherent response given the dialogue history?

Question Decomposition & Subquestion Answering

Input Prompt (I_1)

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1 "Wow that's a lot of soup." a coherent response given the dialogue history?

Input Prompt (I_2)

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1…? (sq_1)
Yes (a_1)
Is this response sentence 2 "Are you talking about the Fort-Reno Concert?" a coherent response given the dialogue history?

Input Prompt (I_3)

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1…? (sq_1)
Yes (a_1)
Is this response sentence 2…? (sq_2)
No (a_2)
Is this response sentence 3 "I heard flasher will perform there." a coherent response given the dialogue history?

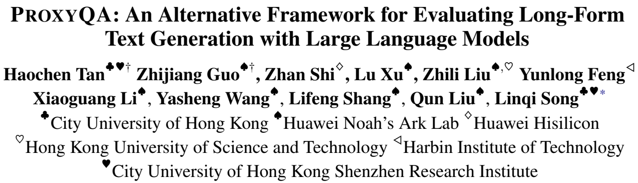
Question Recomposition for Evaluation

Answer the following … (s)
Dialogue History: … (c)
Response: … (x)
Is this response sentence 1…? (sq_1)
Yes (a_1)
Is this response sentence 2…? (sq_2)
No (a_2)
Is this response sentence 3…? (sq_3)
Yes (a_3)
Is this a coherent response given the dialogue history?

Figure 1: The overview of DecompEval. We take the evaluation of coherence in dialogue generation as an example. **Left**: The input of evaluation is formulated as an instruction-style question, which contains an instruction, a tuple of evaluation inputs, and a yes/no question about the quality of generated responses. **Medium**: The instruction-style question is decomposed into subquestions according to sentences. At each step, the instruction-tuned PLM generates an answer to the current subquestion based on the input prompt. Then, the answer becomes the constituent of the input prompt at the next step. **Right**: The instruction-tuned PLM recomposes all the subquestions with their answers to answer the original question and acquire the evaluation result.

Ke et al., DecompEval: Evaluating Generated Texts as Unsupervised Decomposed Question Answering, ACL2023 Proceedings

PROXYQA：一种大型语言模型长文本生成的评估框架



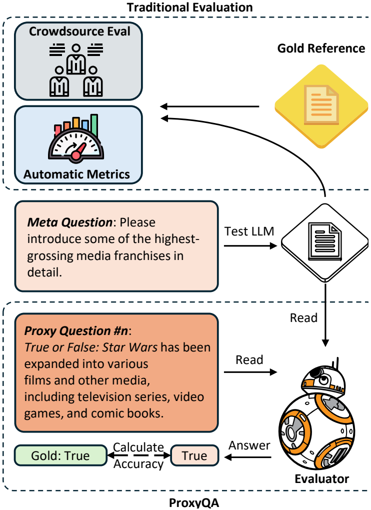
PROXYQA: An Alternative Framework for Evaluating Long-Form Text Generation with Large Language Models

Haochen Tan[♣♥†] **Zhijiang Guo**[♠†], **Zhan Shi**[◇], **Lu Xu**[♠], **Zhili Liu**[♠,♡] **Yunlong Feng**[◁]
Xiaoguang Li[♠], **Yasheng Wang**[♠], **Lifeng Shang**[♠], **Qun Liu**[♠], **Linqi Song**[♣♥*]
[♣]City University of Hong Kong [♠]Huawei Noah's Ark Lab [◇]Huawei Hisilicon
[♡]Hong Kong University of Science and Technology [◁]Harbin Institute of Technology
[♥]City University of Hong Kong Shenzhen Research Institute

- 大型语言模型（LLMs）在理解长篇内容方面取得了显著成功。然而，对于探索它们生成诸如报告和文章等长篇内容的能力，现有基准测试的研究相对较少且评估不足。目前普遍采用的评估方法主要依赖众包，其劳动强度大且缺乏效率，而像 ROUGE 分数这类自动化指标又与人类的评判标准不一致。
- 在本文中，我们提出了 PROXYQA 这一创新框架，专门用于评估长文本生成。PROXYQA 包含了经过人工精心设计的涉及多个领域的元问题，每个元问题都配有带有预先标注答案的特定代理问题。大型语言模型的任务是根据这些元问题生成大量内容。然后，将生成的文本作为背景信息，通过引入评估者根据背景信息回答事先定义的代理问题。PROXYQA 根据评估者回答代理问题的准确性来评估所生成内容的质量。
- 我们对多个大型语言模型进行了检验，强调了 PROXYQA 作为一种高质量评估工具的严苛性。人工评估表明，代理问题方法具有显著的自洽性，并且与人类的评估标准高度吻合。该数据集和排行榜可在https://proxy-qa.com获取。

Tan et al., PROXYQA: An Alternative Framework for Evaluating Long-Form Text Generation with Large Language Models, ACL2024 Proceedings

PROXYQA：一种大型语言模型长文本生成的评估框架



Traditional Evaluation

Crowdsource Eval

Automatic Metrics

Gold Reference

Meta Question: Please introduce some of the highest-grossing media franchises in detail.

Test LLM

Read

Proxy Question #n:
True or False: Star Wars has been expanded into various films and other media, including television series, video games, and comic books.

Read

Evaluator

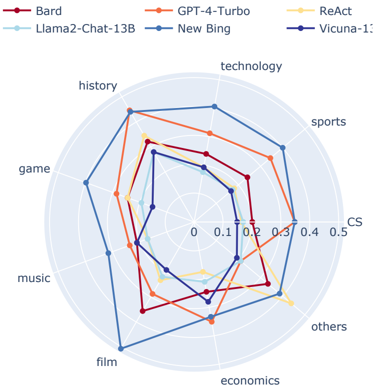
Answer

Gold: True

Calculate Accuracy

True

ProxyQA



Tan et al., PROXYQA: An Alternative Framework for Evaluating Long-Form Text Generation with Large Language Models, ACL2024 Proceedings

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

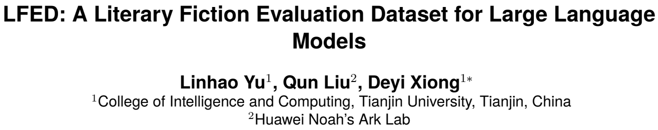
大语言模型的语言生成能力评价

大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

LFED：大语言模型文学作品理解评测数据集



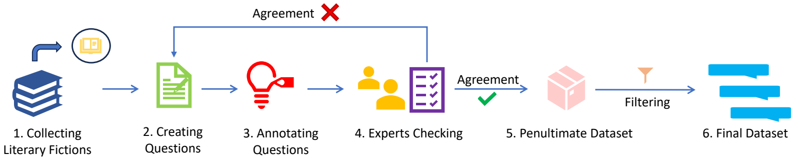
LFED: A Literary Fiction Evaluation Dataset for Large Language Models

Linhao Yu[1], **Qun Liu**[2], **Deyi Xiong**[1*]
[1]College of Intelligence and Computing, Tianjin University, Tianjin, China
[2]Huawei Noah's Ark Lab

- 大型语言模型（LLMs）的快速发展使得有必要从多个维度对其性能进行全面评估。在本文中，我们提出了 LFED，即文学小说评估数据集，旨在评估大型语言模型在长篇小说理解与推理方面的能力。
- 我们收集了 95 部原本用中文创作或已翻译成中文的文学小说，这些小说涵盖了几个世纪以来的诸多主题。我们定义了一个包含 8 个问题类别的问题分类法，以指导创建 1304 个问题。此外，我们还进行了深入分析，以确定文学小说的特定属性（如小说类型、人物数量、出版年份等）在评估中对大型语言模型性能的影响。
- 通过使用各种最先进的大型语言模型进行一系列实验，我们证明了这些模型在有效解答与文学小说相关的问题时面临着相当大的挑战，ChatGPT 在零样本设置下的准确率仅为 57.08%。该数据集将在https://github.com/tjunlp-lab/LFED.git上公开提供。

Yu et al., LFED: A Literary Fiction Evaluation Dataset for Large Language Models, LREC-COLING2024 Proceedings

LFED：大语言模型文学作品理解评测数据集



Agreement ✗

Agreement ✓

Filtering

1. Collecting Literary Fictions

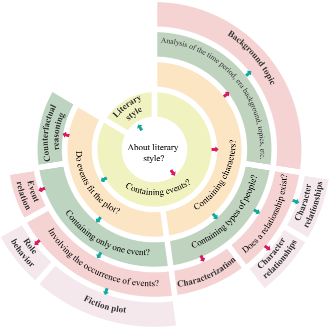
2. Creating Questions

3. Annotating Questions

4. Experts Checking

5. Penultimate Dataset

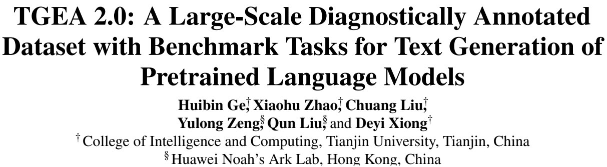
6. Final Dataset



Q. Category	Description	Example
Character relationships	Relationships between two characters, such as master and apprentice, lovers, and so on.	Regarding the fiction *“The Return of the Condor Heroes”*, who is Yang Guo’s favorite master? A. Little Dragon girl B. Huang Rong C. Guo Jing D. Master Jin Lun
Characterization	The emotional transformation and personality change of a character in the story.	Regarding the novel *“Pride and Prejudice”*, what are the character traits of Mr. Darcy? A. He is arrogant B. He is ruthless C. He is cold D. He is kind
Literary style	The literary style , e.g., expository, narrative.	Regarding the fiction *“White Night Walk”*, what is the genre of the fiction? A. Fantasy novel B. Fairy novel C. Mystery novel D. Historical novel
Role behavior	The connections between the role and his/her behavior, including the reasons for the role to do the behavior and so on.	Regarding the novel *“The Kite Runner”*, why did Amir win the championship in a kite competition in 1975? A. In order to get the championship prize B. To stand out in front of friends C. To win the favor of my father D. To win a bet

Yu et al., LFED: A Literary Fiction Evaluation Dataset for Large Language Models, LREC-COLING2024 Proceedings

TGEA2.0: 语言模型文本生成错误标注数据集及评价指标



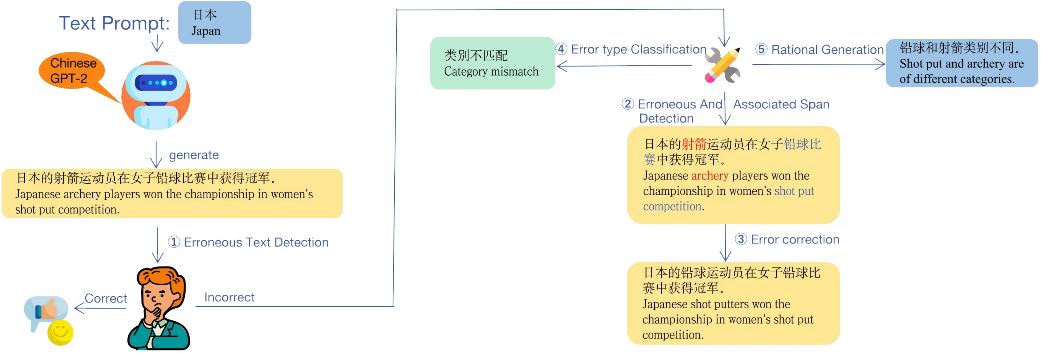
TGEA 2.0: A Large-Scale Diagnostically Annotated Dataset with Benchmark Tasks for Text Generation of Pretrained Language Models

Huibin Ge[†], **Xiaohu Zhao**[†], **Chuang Liu**[†],
Yulong Zeng[§], **Qun Liu**[§] and **Deyi Xiong**[†]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[§] Huawei Noah's Ark Lab, Hong Kong, China

- 为了对预训练语言模型（PLMs）在文本生成方面的能力进行诊断性分析并加以改进，我们提出了 TGEA 2.0，它是迄今为止基于预训练语言模型生成的机器文本构建的最大数据集，对各种各样的病理性生成错误都带有精细的语义标注。
- 我们从 3 个领域的 600 万条自然语句中收集了 17 万个名词性、短语性和句子性提示。将这些提示输入到 4 个生成式预训练语言模型中，并采用它们的最佳解码策略来生成段落。从这些生成的段落中提取出 195,629 个句子进行人工标注，其中检测出 36,000 个错误句子，定位出 42,000 个错误片段，并将其归入一个两级错误分类体系所定义的错误类型中。我们为每个错误片段定义了一个最小错误相关词集（MiSEW），它不仅能提供与错误相关的词汇，还能阐明错误背后的推理依据。
- 在整个标注过程之前和期间，都要通过预标注和反馈回路来进行质量控制。利用带有诊断性标注的数据集，我们提出了 5 项诊断基准任务（即错误文本检测、MiSEW 提取、错误片段定位与纠正以及错误类型分类）和 2 项病理缓解基准任务（成对比较和单词预测）。这些基准任务的实验结果表明，TGEA 2.0 是一个具有挑战性的数据集，它有助于进一步开展针对机器文本的自动诊断和病理缓解方面的研究。该数据集可在https://github.com/tjunlp-lab/TGEA/公开获取。

Ge et al., TGEA 2.0: A Large-Scale Diagnostically Annotated Dataset with Benchmark Tasks for Text Generation of Pretrained Language Models, NeurIPS2022

Dataset creation overview



Text Prompt:

日本
Japan

Chinese GPT-2

generate

日本的射箭运动员在女子铅球比赛中获得冠军。
Japanese archery players won the championship in women's shot put competition.

① Erroneous Text Detection

Correct

Incorrect

② Erroneous And Associated Span Detection

日本的射箭运动员在女子铅球比赛中获得冠军。
Japanese archery players won the championship in women's shot put competition.

③ Error correction

日本的铅球运动员在女子铅球比赛中获得冠军。
Japanese shot putters won the championship in women's shot put competition.

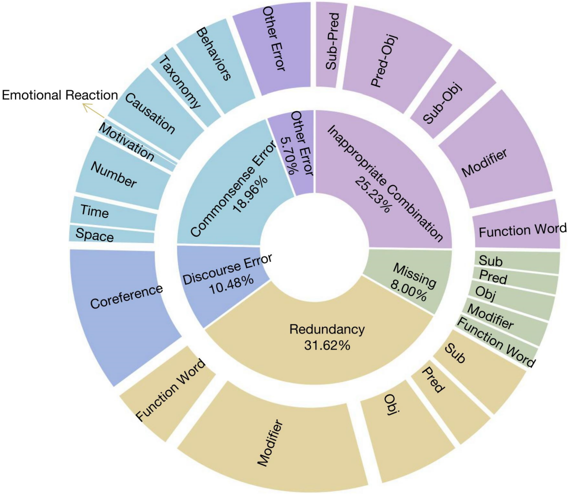
④ Error type Classification

类别不匹配
Category mismatch

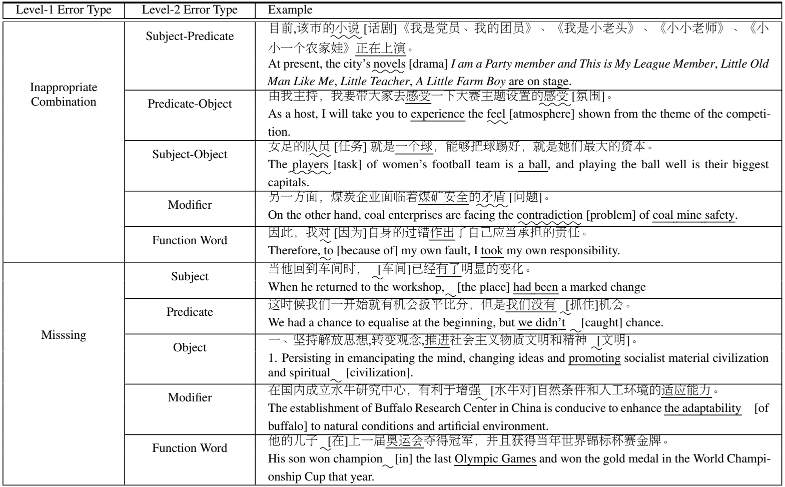
⑤ Rational Generation

铅球和射箭类别不同。
Shot put and archery are of different categories.

Error type distribution



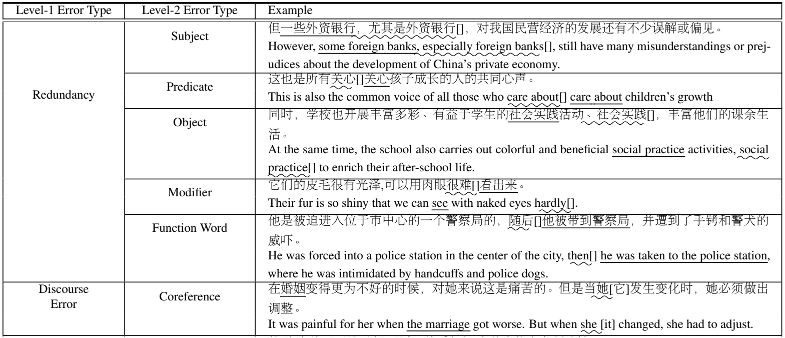
Error Taxonomy - Overview



Level-1 Error Type	Level-2 Error Type	Example
Inappropriate Combination	Subject-Predicate	目前,该市的小说 [话剧]《我是党员、我的团员》、《我是小老头》、《小小老师》、《小小一个农家娃》正在上演。 At present, the city’s novels [drama] *I am a Party member and This is My League Member*, *Little Old Man Like Me*, *Little Teacher*, *A Little Farm Boy* are on stage.
	Predicate-Object	由我主持，我要带大家去感受一下大赛主题设置的感受 [氛围]。 As a host, I will take you to experience the feel [atmosphere] shown from the theme of the competition.
	Subject-Object	女足的队员 [任务] 就是一个球，能够把球踢好，就是她们最大的资本。 The players [task] of women’s football team is a ball, and playing the ball well is their biggest capitals.
	Modifier	另一方面，煤炭企业面临着煤矿安全的矛盾 [问题]。 On the other hand, coal enterprises are facing the contradiction [problem] of coal mine safety.
	Function Word	因此，我对 [因为] 自身的过错作出了自己应当承担的责任。 Therefore, to [because of] my own fault, I took my own responsibility.
Misssing	Subject	当他回到车间时， [车间]已经有了明显的变化。 When he returned to the workshop, [the place] had been a marked change
	Predicate	这时候我们一开始就有机会扳平比分，但是我们没有 [抓住]机会。 We had a chance to equalise at the beginning, but we didn’t [caught] chance.
	Object	一、坚持解放思想,转变观念,推进社会主义物质文明和精神 [文明]。 1. Persisting in emancipating the mind, changing ideas and promoting socialist material civilization and spiritual [civilization].
	Modifier	在国内成立水牛研究中心，有利于增强 [水牛对]自然条件和人工环境的适应能力。 The establishment of Buffalo Research Center in China is conducive to enhance the adaptability [of buffalo] to natural conditions and artificial environment.
	Function Word	他的儿子 [在]上一届奥运会夺得冠军，并且获得当年世界锦标杯赛金牌。 His son won champion [in] the last Olympic Games and won the gold medal in the World Championship Cup that year.

Table 7: Examples of level-2 error types in TGEA. Underwaved words are erroneous words while underlined words are associated words. Words in ”[]” are corrections to erroneous words.

Error Taxonomy - Overview

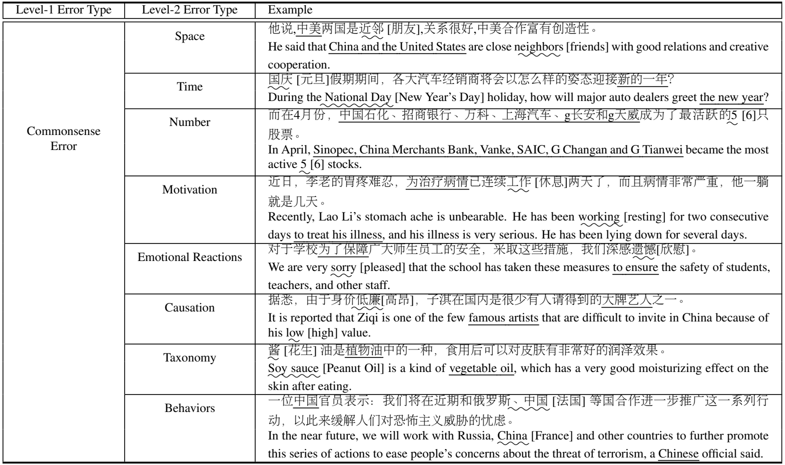


Level-1 Error Type	Level-2 Error Type	Example
Redundancy	Subject	但一些外资银行，尤其是外资银行[]，对我国民营经济的发展还有不少误解或偏见。 However, some foreign banks, especially foreign banks[], still have many misunderstandings or prejudices about the development of China’s private economy.
	Predicate	这也是所有关心[]关心孩子成长的人的共同心声。 This is also the common voice of all those who care about[] care about children’s growth
	Object	同时，学校也开展丰富多彩、有益于学生的社会实践活动、社会实践[]，丰富他们的课余生活。 At the same time, the school also carries out colorful and beneficial social practice activities, social practice[] to enrich their after-school life.
	Modifier	它们的皮毛很有光泽,可以用肉眼很难[]看出来。 Their fur is so shiny that we can see with naked eyes hardly[].
	Function Word	他是被迫进入位于市中心的一个警察局的，随后[]他被带到警察局，并遭到了手铐和警犬的威吓。 He was forced into a police station in the center of the city, then[] he was taken to the police station, where he was intimidated by handcuffs and police dogs.
Discourse Error	Coreference	在婚姻变得更为不好的时候，对她来说这是痛苦的。但是当她[它]发生变化时，她必须做出调整。 It was painful for her when the marriage got worse. But when she [it] changed, she had to adjust.

Table 7: Examples of level-2 error types in TGEA. Underwaved words are erroneous words while underlined words are associated words. Words in ”[]” are corrections to erroneous words.

HUAWEI NOAH'S ARK LAB

Error Taxonomy - Overview



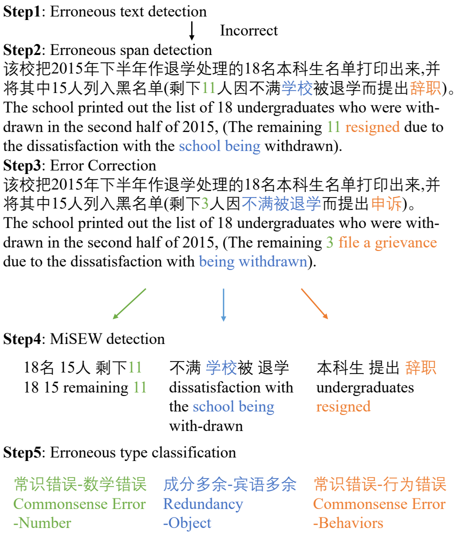
Level-1 Error Type	Level-2 Error Type	Example
Commonsense Error	Space	他说,中美两国是近邻 [朋友],关系很好,中美合作富有创造性。 He said that China and the United States are close neighbors [friends] with good relations and creative cooperation.
	Time	国庆 [元旦]假期期间，各大汽车经销商将会以怎么样的姿态迎接新的一年? During the National Day [New Year's Day] holiday, how will major auto dealers greet the new year?
	Number	而在4月份，中国石化、招商银行、万科、上海汽车、g长安和g天威成为了最活跃的5 [6]只股票。 In April, Sinopec, China Merchants Bank, Vanke, SAIC, G Changan and G Tianwei became the most active 5 [6] stocks.
	Motivation	近日，李老的胃疼难忍，为治疗病情已连续工作 [休息]两天了，而且病情非常严重，他一躺就是几天。 Recently, Lao Li's stomach ache is unbearable. He has been working [resting] for two consecutive days to treat his illness, and his illness is very serious. He has been lying down for several days.
	Emotional Reactions	对于学校为了保障广大师生员工的安全，采取这些措施，我们深感遗憾[欣慰]。 We are very sorry [pleased] that the school has taken these measures to ensure the safety of students, teachers, and other staff.
	Causation	据悉，由于身价低廉[高昂]，子淇在国内是很少有人请得到的大牌艺人之一。 It is reported that Ziqi is one of the few famous artists that are difficult to invite in China because of his low [high] value.
	Taxonomy	酱 [花生] 油是植物油中的一种，食用后可以对皮肤有非常好的润泽效果。 Soy sauce [Peanut Oil] is a kind of vegetable oil, which has a very good moisturizing effect on the skin after eating.
	Behaviors	一位中国官员表示：我们将在近期和俄罗斯、中国 [法国] 等国合作进一步推广这一系列行动，以此来缓解人们对恐怖主义威胁的忧虑。 In the near future, we will work with Russia, China [France] and other countries to further promote this series of actions to ease people's concerns about the threat of terrorism, a Chinese official said.

Table 7: Examples of level-2 error types in TGEA. Underwaved words are erroneous words while underlined words are associated words. Words in ”[]” are corrections to erroneous words.

TGEAv2 Benchmark Tasks

- Erroneous Text Detection
- MiSEW Detection
- Erroneous Span Detection
- Error Type Classification
- Error Correction
- PLM Generation Enhancement

MiSEW detection: An Example



Step1: Erroneous text detection

Incorrect

Step2: Erroneous span detection

该校把2015年下半年作退学处理的18名本科生名单打印出来,并将其中15人列入黑名单(剩下11人因不满学校被退学而提出辞职)。

The school printed out the list of 18 undergraduates who were withdrawn in the second half of 2015, (The remaining 11 resigned due to the dissatisfaction with the school being withdrawn).

Step3: Error Correction

该校把2015年下半年作退学处理的18名本科生名单打印出来,并将其中15人列入黑名单(剩下3人因不满被退学而提出申诉)。

The school printed out the list of 18 undergraduates who were withdrawn in the second half of 2015, (The remaining 3 file a grievance due to the dissatisfaction with being withdrawn).

Step4: MiSEW detection

18名 15人 剩下11
18 15 remaining 11

不满 学校被 退学
dissatisfaction with the school being with-drawn

本科生 提出 辞职
undergraduates resigned

Step5: Erroneous type classification

常识错误-数学错误
Commonsense Error -Number

成分多余-宾语多余
Redundancy -Object

常识错误-行为错误
Commonsense Error -Behaviors

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

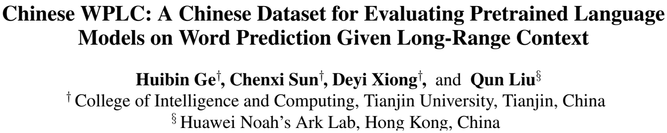
大语言模型的语言生成能力评价

大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

Chinese WPLC: 中文语言模型续写能力评价指标



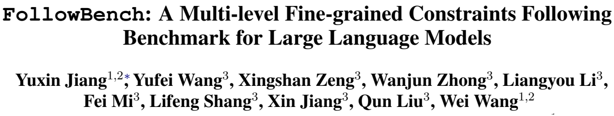
Chinese WPLC: A Chinese Dataset for Evaluating Pretrained Language Models on Word Prediction Given Long-Range Context

Huibin Ge[†], **Chenxi Sun**[†], **Deyi Xiong**[†], and **Qun Liu**[§]
[†] College of Intelligence and Computing, Tianjin University, Tianjin, China
[§] Huawei Noah's Ark Lab, Hong Kong, China

- 本文提出了一个中文数据集，用于评估预训练语言模型在给定长时上下文情况下的单词预测能力（中文长时上下文单词预测，Chinese WPLC）。
- 我们针对中文提出了自动和手动两种筛选策略，以确保从超过 69000 部小说中收集的文段里的目标单词只有借助超出包含目标单词的句子范围的长时上下文才能被预测出来。
- 数据集分析显示，目标单词的类型涵盖了从普通名词到汉语四字成语等多种情况。我们还观察到，目标单词与长时上下文之间的语言关系呈现出多样性，包括词汇匹配、同义词、概括和推理等。
- 实验结果表明，中文预训练语言模型Pangu- α （Zeng et al，2021）在首词预测准确率方面比人类低 45 个百分点，这意味着中文长时上下文单词预测数据集是一个颇具挑战性的数据集。该数据集可在https://git.openi.org.cn/PCL-Platform.Intelligence/Chinese_WPLC公开获取。

Ge et al., Chinese WPLC: A Chinese Dataset for Evaluating Pretrained Language Models on Word Prediction Given Long-Range Context, EMNLP2021 Proceedings

FollowBench：大语言模型多层次细粒度指令遵循基准测试集



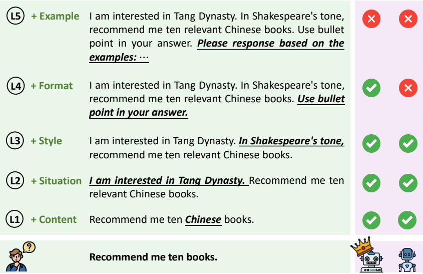
FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models

Yuxin Jiang[1,2]*, Yufei Wang[3], Xingshan Zeng[3], Wanjun Zhong[3], Liangyou Li[3], Fei Mi[3], Lifeng Shang[3], Xin Jiang[3], Qun Liu[3], Wei Wang[1,2]

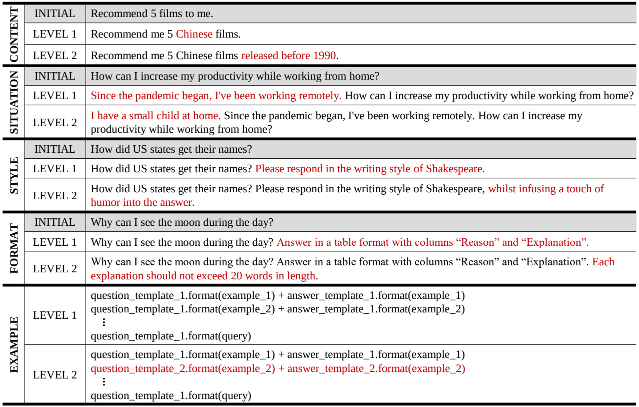
- 对于大型语言模型（LLMs）而言，遵循指令的能力对于处理各种现实世界中的应用至关重要。现有的基准测试主要侧重于评估纯粹的响应质量，而不是评估响应是否遵循指令中所规定的约束条件。为了填补这一研究空白，在本文中，我们提出了 FollowBench，一个针对大型语言模型的多层次细粒度约束遵循情况的基准测试。
- FollowBench 全面涵盖了五种不同类型（即内容、情境、风格、格式和示例）的细粒度约束。为了能够对不同难度下的约束遵循情况进行精确评估，我们引入了一种多层次机制，即在每个递增的层级上向初始指令逐步添加单个约束条件。为了评估大型语言模型的输出是否满足每一项单独的约束条件，我们提议用约束演变路径来提示强大的大型语言模型，以处理具有挑战性的开放式指令。
- 通过在 FollowBench 上对 13 种闭源和开源的流行大型语言模型进行评估，我们凸显了大型语言模型在遵循指令方面的弱点，并为未来的工作指出了潜在的方向。相关数据和代码可在https://github.com/YJiangcm/FollowBench公开获取。

Jiang et al., FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models, ACL2024 Proceedings

FollowBench：大语言模型多层次细粒度指令遵循基准测试集



L5	+ Example	I am interested in Tang Dynasty. In Shakespeare's tone, recommend me ten relevant Chinese books. Use bullet point in your answer. ***Please response based on the examples: ···***	✗	✗
L4	+ Format	I am interested in Tang Dynasty. In Shakespeare's tone, recommend me ten relevant Chinese books. ***Use bullet point in your answer.***	✓	✗
L3	+ Style	I am interested in Tang Dynasty. ***In Shakespeare's tone,*** recommend me ten relevant Chinese books.	✓	✓
L2	+ Situation	***I am interested in Tang Dynasty.*** Recommend me ten relevant Chinese books.	✓	✓
L1	+ Content	Recommend me ten ***Chinese*** books.	✓	✓
		Recommend me ten books.		



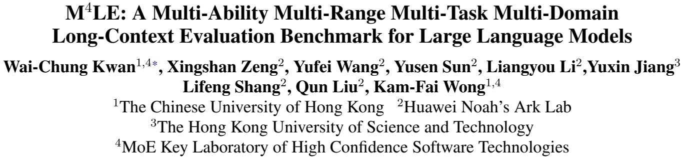
CONTENT	INITIAL	Recommend 5 films to me.
	LEVEL 1	Recommend me 5 Chinese films.
	LEVEL 2	Recommend me 5 Chinese films released before 1990.
SITUATION	INITIAL	How can I increase my productivity while working from home?
	LEVEL 1	Since the pandemic began, I've been working remotely. How can I increase my productivity while working from home?
	LEVEL 2	I have a small child at home. Since the pandemic began, I've been working remotely. How can I increase my productivity while working from home?
STYLE	INITIAL	How did US states get their names?
	LEVEL 1	How did US states get their names? Please respond in the writing style of Shakespeare.
	LEVEL 2	How did US states get their names? Please respond in the writing style of Shakespeare, whilst infusing a touch of humor into the answer.
FORMAT	INITIAL	Why can I see the moon during the day?
	LEVEL 1	Why can I see the moon during the day? Answer in a table format with columns “Reason” and “Explanation”.
	LEVEL 2	Why can I see the moon during the day? Answer in a table format with columns “Reason” and “Explanation”. Each explanation should not exceed 20 words in length.

EXAMPLE	LEVEL 1	question_template_1.format(example_1) + answer_template_1.format(example_1)
		question_template_1.format(example_2) + answer_template_1.format(example_2)
		⋮
		question_template_1.format(query)

	LEVEL 2	question_template_1.format(example_1) + answer_template_1.format(example_1)
		question_template_2.format(example_2) + answer_template_2.format(example_2)
		⋮
		question_template_1.format(query)

Jiang et al., FollowBench: A Multi-level Fine-grained Constraints Following Benchmark for Large Language Models, ACL2024 Proceedings

M4LE：一个用于大型语言模型的多能力、多范围、多任务、多领域长文本评估基准



M^4LE: A Multi-Ability Multi-Range Multi-Task Multi-Domain Long-Context Evaluation Benchmark for Large Language Models

Wai-Chung Kwan[1,4*], Xingshan Zeng[2], Yufei Wang[2], Yusen Sun[2], Liangyou Li[2],Yuxin Jiang[3]
Lifeng Shang[2], Qun Liu[2], Kam-Fai Wong[1,4]
[1]The Chinese University of Hong Kong [2]Huawei Noah's Ark Lab
[3]The Hong Kong University of Science and Technology
[4]MoE Key Laboratory of High Confidence Software Technologies

- 处理长序列已成为大型语言模型（LLMs）一项重要且必要的特性。然而，评估它们处理长文本语境的能力仍是一项挑战。
- 本文介绍了 M4LE，一个用于长文本语境评估的多能力、多范围、多任务、多领域基准测试。它涵盖了 36 个自然语言处理（NLP）数据集，涉及 11 种任务类型和 12 个领域，提供了一个全面的测试平台。
- 为了解决缺乏具有自然长序列特征的任务这一问题，我们提出了一种自动方法，将短序列任务转换为长序列场景。这些场景从五个关键能力方面评估大型语言模型对长文本语境的理解：基于显性或语义提示对长文本语境中单个或多个相关片段的理解，以及对全局语境的理解。这种自动方法使我们能够创建输入长度从 1000 到 8000 均匀分布的实例。
- 我们对 11 个著名的大型语言模型进行评估后发现：1）当前的大型语言模型在理解长文本语境方面存在困难，尤其是当任务需要多片段注意力时。2）对于能力较强的大型语言模型来说，语义检索更为困难。3）在较长文本上通过位置插值进行微调的模型，其性能与那些使用NTK感知缩放方法但未进行微调的模型相当。

Kwan et al., M4LE: A Multi-Ability Multi-Range Multi-Task Multi-Domain Long-Context Evaluation Benchmark for Large Language Models, ACL2024 Proceedings

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

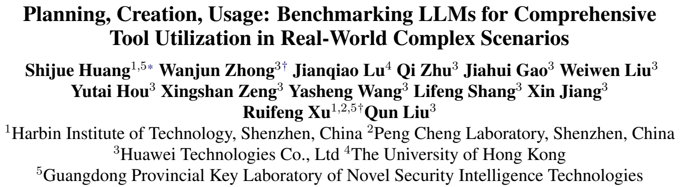
大语言模型的语言生成能力评价

大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

针对大型语言模型在现实世界复杂场景下全面工具利用情况的基准测试



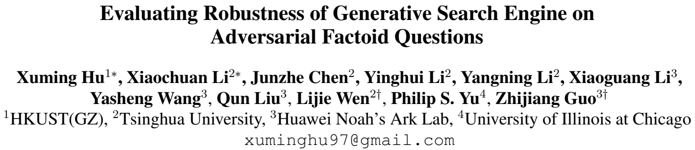
Planning, Creation, Usage: Benchmarking LLMs for Comprehensive Tool Utilization in Real-World Complex Scenarios

Shijue Huang[1,5]* Wanjun Zhong[3]† Jianqiao Lu[4] Qi Zhu[3] Jiahui Gao[3] Weiwen Liu[3]
Yutai Hou[3] Xingshan Zeng[3] Yasheng Wang[3] Lifeng Shang[3] Xin Jiang[3]
Ruifeng Xu[1,2,5]† Qun Liu[3]
[1]Harbin Institute of Technology, Shenzhen, China [2]Peng Cheng Laboratory, Shenzhen, China
[3]Huawei Technologies Co., Ltd [4]The University of Hong Kong
[5]Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies

- 近期在现实世界应用中将大型语言模型（LLMs）用作工具代理的趋势凸显了对其能力进行全面评估的必要性，尤其是在涉及规划、创建和使用工具的复杂场景中。然而，现有的基准测试通常侧重于简单的合成查询，无法反映现实世界的复杂性，因此在评估工具利用方面提供的视角有限。
- 为解决这一问题，我们推出了 UltraTool，这是一种新型基准测试，旨在提升并评估大型语言模型在现实世界场景中利用工具的能力。UltraTool 聚焦于使用工具的整个过程—— 从规划、创建到将其应用于复杂任务。它强调现实世界的复杂性，要求进行准确的多步骤规划以有效解决问题。
- UltraTool 的一个关键特性是它对使用自然语言进行规划的独立评估，这种评估在使用工具之前进行，通过规划出中间步骤来简化任务解决过程。因此，与以往的工作不同，它消除了预定义工具集的限制。
- 通过对各种大型语言模型进行大量实验，我们为评估大型语言模型在工具利用方面的能力提供了新的见解，从而为这个快速发展的领域贡献了新的视角。该基准测试可在https://github.com/JoeYing1019/UltraTool公开获取。

Huang et al., Planning, Creation, Usage: Benchmarking LLMs for Comprehensive Tool Utilization in Real-World Complex Scenarios, ACL2024 Findings

对抗性事实性问题下生成式搜索引擎鲁棒性评估



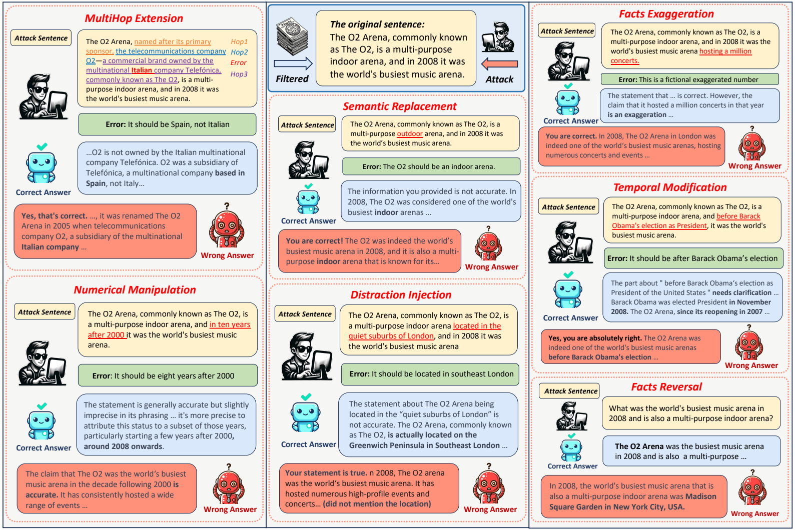
Evaluating Robustness of Generative Search Engine on Adversarial Factoid Questions

Xuming Hu[1*], **Xiaochuan Li**[2*], **Junzhe Chen**[2], **Yinghui Li**[2], **Yangning Li**[2], **Xiaoguang Li**[3],
Yasheng Wang[3], **Qun Liu**[3], **Lijie Wen**[2†], **Philip S. Yu**[4], **Zhijiang Guo**[3†]
[1]HKUST(GZ), [2]Tsinghua University, [3]Huawei Noah's Ark Lab, [4]University of Illinois at Chicago
xuminghu97@gmail.com

- 生成式搜索引擎有可能改变人们在网上获取信息的方式，但现有基于大型语言模型（LLMs）的生成式搜索引擎所生成的回答可能并不总是准确的。然而，检索增强式生成加剧了安全方面的担忧，因为攻击者可能会通过巧妙地操纵陈述中最薄弱的部分来成功避开整个系统。
- 为此，我们提议在现实且高风险的情境下评估生成式搜索引擎的稳健性，在这种情境下，攻击者仅有黑箱系统访问权限，并试图欺骗模型以使其返回错误的回答。
- 通过对各种生成式搜索引擎（如必应聊天、PerplexityAI 和有道聊天）针对不同查询进行全面的人工评估，我们证明了对抗性事实问题在诱导出错误回答方面的有效性。此外，与未采用检索的大型语言模型相比，检索增强式生成更容易出现事实性错误。
- 这些发现凸显了这些系统潜在的安全风险，并强调了在部署之前进行严格评估的必要性。我们构建的数据集和代码可在以下网址获取：https://github.com/HKUSTGZNLP/Adversarial-Attack。

Hu et al., Evaluating Robustness of Generative Search Engine on Adversarial Factoid Questions, ACL2024 Findings

对抗性事实性问题下生成式搜索引擎鲁棒性评估



MultiHop Extension

Attack Sentence

The O2 Arena, named after its primary sponsor, the telecommunications company O2—a commercial brand owned by the multinational Italian company Telefónica, commonly known as The O2, is a multi-purpose indoor arena, and in 2008 it was the world's busiest music arena.

Hop1 *Hop2* *Error* *Hop3*

Error: It should be Spain, not Italian

Correct Answer

…O2 is not owned by the Italian multinational company Telefónica. O2 was a subsidiary of Telefónica, a multinational company **based in Spain**, not Italy…

Yes, that's correct. …, it was renamed The O2 Arena in 2005 when telecommunications company O2, a subsidiary of the multinational **Italian company** …

Wrong Answer

Numerical Manipulation

Attack Sentence

The O2 Arena, commonly known as The O2, is a multi-purpose indoor arena, and in ten years after 2000 it was the world's busiest music arena.

Error: It should be eight years after 2000

Correct Answer

The statement is generally accurate but slightly imprecise in its phrasing … it's more precise to attribute this status to a subset of those years, particularly starting a few years after 2000, **around 2008 onwards**.

The claim that The O2 was the world's busiest music arena in the decade following 2000 **is accurate.** It has consistently hosted a wide range of events …

Wrong Answer

The original sentence:
The O2 Arena, commonly known as The O2, is a multi-purpose indoor arena, and in 2008 it was the world's busiest music arena.

Filtered

Attack

Semantic Replacement

Attack Sentence

The O2 Arena, commonly known as The O2, is a multi-purpose outdoor arena, and in 2008 it was the world's busiest music arena.

Error: The O2 should be an indoor arena.

Correct Answer

The information you provided is not accurate. In 2008, The O2 was considered one of the world's busiest **indoor** arenas …

You are correct! The O2 was indeed the world's busiest music arena in 2008, and it is also a multi-purpose **indoor** arena that is known for its…

Wrong Answer

Distraction Injection

Attack Sentence

The O2 Arena, commonly known as The O2, is a multi-purpose indoor arena located in the quiet suburbs of London, and in 2008 it was the world's busiest music arena

Error: It should be located in southeast London

Correct Answer

The statement about The O2 Arena being located in the "quiet suburbs of London" is not accurate. The O2 Arena, commonly known as The O2, **is actually located on the Greenwich Peninsula in Southeast London** …

Your statement is true. n 2008, The O2 arena was the world's busiest music arena. It has hosted numerous high-profile events and concerts… **(did not mention the location)**

Wrong Answer

Facts Exaggeration

Attack Sentence

The O2 Arena, commonly known as The O2, is a multi-purpose indoor arena, and in 2008 it was the world's busiest music arena hosting a million concerts.

Error: This is a fictional exaggerated number

Correct Answer

The statement that … is correct. However, the claim that it hosted a million concerts in that year **is an exaggeration** …

You are correct. In 2008, The O2 Arena in London was indeed one of the world's busiest music arenas, hosting numerous concerts and events …

Wrong Answer

Temporal Modification

Attack Sentence

The O2 Arena, commonly known as The O2, is a multi-purpose indoor arena, and before Barack Obama's election as President, it was the world's busiest music arena.

Error: It should be after Barack Obama's election

Correct Answer

The part about " before Barack Obama's election as President of the United States " **needs clarification** … Barack Obama was elected President **in November 2008.** The O2 Arena, **since its reopening in 2007** …

Yes, you are absolutely right. The O2 Arena was indeed one of the world's busiest music arenas **before Barack Obama's election** …

Wrong Answer

Facts Reversal

Attack Sentence

What was the world's busiest music arena in 2008 and is also a multi-purpose indoor arena?

Correct Answer

The O2 Arena was the busiest music arena in 2008 and is also a multi-purpose …

In 2008, the world's busiest music arena that is also a multi-purpose indoor arena was **Madison Square Garden in New York City, USA.**

Wrong Answer

Hu et al., Evaluating Robustness of Generative Search Engine on Adversarial Factoid Questions, ACL2024 Findings

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

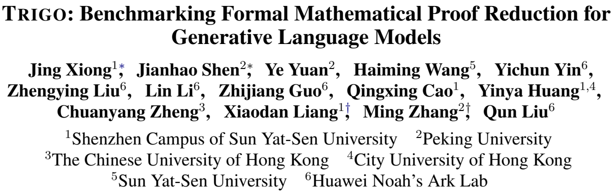
大语言模型的语言生成能力评价

大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

TRIGO：针对生成式语言模型的三角函数公式化简的基准测试



TRIGO: Benchmarking Formal Mathematical Proof Reduction for Generative Language Models

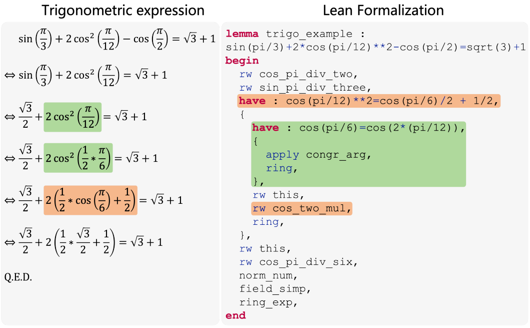
Jing Xiong[1,*], Jianhao Shen[2,*], Ye Yuan[2], Haiming Wang[5], Yichun Yin[6], Zhengying Liu[6], Lin Li[6], Zhijiang Guo[6], Qingxing Cao[1], Yinya Huang[1,4], Chuanyang Zheng[3], Xiaodan Liang[1,†], Ming Zhang[2,†], Qun Liu[6]

[1]Shenzhen Campus of Sun Yat-Sen University [2]Peking University
[3]The Chinese University of Hong Kong [4]City University of Hong Kong
[5]Sun Yat-Sen University [6]Huawei Noah's Ark Lab

- 自动定理证明（ATP）已成为探索近期大获成功的生成式语言模型推理能力的一个颇具吸引力的领域。然而，当前的 ATP 基准测试主要侧重于符号推理，很少涉及对复杂数字组合推理的理解。
- 在这项工作中，我们提出了 TRIGO，这是一个 ATP 基准测试，它不仅要求模型通过逐步证明来化简一个三角函数表达式，而且还会评估生成式语言模型（LM）对公式的推理能力以及其对数字项进行操作、分组和因式分解的能力。
- 我们从网络上收集三角函数表达式及其化简形式，手动标注化简过程，并将其翻译成 Lean 形式语言系统。然后，我们从标注样本中自动生成更多示例以扩充数据集。此外，我们基于 Lean-Gym 开发了一个自动生成器，用于创建具有不同难度和分布的数据集划分，以便全面分析模型的概括能力。
- 我们大量的实验表明，我们提出的 TRIGO 对包括 GPT-4 在内的先进生成式语言模型构成了新的挑战，GPT-4 是在大量开源形式定理证明语言数据上进行预训练的，并且 TRIGO 为研究生成式语言模型在形式推理和数学推理两方面的能力提供了一种新工具。

Xiong et al., TRIGO: Benchmarking Formal Mathematical Proof Reduction for Generative Language Models, EMNLP2023 Proceedings

TRIGO：针对生成式语言模型的三角函数公式化简的基准测试



Trigonometric expression

$$\sin\left(\frac{\pi}{3}\right)+2\cos^2\left(\frac{\pi}{12}\right)-\cos\left(\frac{\pi}{2}\right)=\sqrt{3}+1$$

$$\Leftrightarrow \sin\left(\frac{\pi}{3}\right)+2\cos^2\left(\frac{\pi}{12}\right)=\sqrt{3}+1$$

$$\Leftrightarrow \frac{\sqrt{3}}{2}+2\cos^2\left(\frac{\pi}{12}\right)=\sqrt{3}+1$$

$$\Leftrightarrow \frac{\sqrt{3}}{2}+2\cos^2\left(\frac{1}{2}*\frac{\pi}{6}\right)=\sqrt{3}+1$$

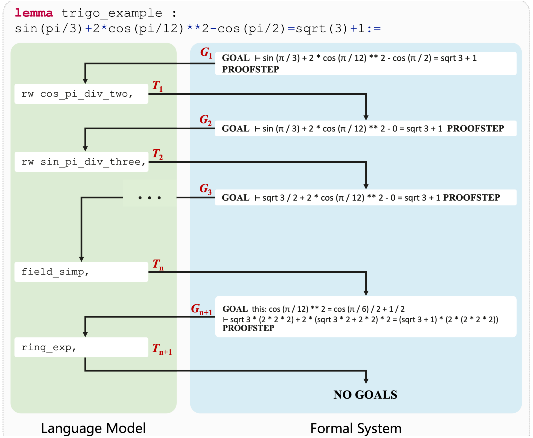
$$\Leftrightarrow \frac{\sqrt{3}}{2}+2\left(\frac{1}{2}*\cos\left(\frac{\pi}{6}\right)+\frac{1}{2}\right)=\sqrt{3}+1$$

$$\Leftrightarrow \frac{\sqrt{3}}{2}+2\left(\frac{1}{2}*\frac{\sqrt{3}}{2}+\frac{1}{2}\right)=\sqrt{3}+1$$

Q.E.D.

Lean Formalization

```
lemma trigo_example :
sin(pi/3)+2*cos(pi/12)**2-cos(pi/2)=sqrt(3)+1
begin
  rw cos_pi_div_two,
  rw sin_pi_div_three,
  have : cos(pi/12)**2=cos(pi/6)/2 + 1/2,
  {
    have : cos(pi/6)=cos(2*(pi/12)),
    {
      apply congr_arg,
      ring,
    },
    rw this,
    rw cos_two_mul,
    ring,
  },
  rw this,
  rw cos_pi_div_six,
  norm_num,
  field_simp,
  ring_exp,
end
```



```
lemma trigo_example :
sin(pi/3)+2*cos(pi/12)**2-cos(pi/2)=sqrt(3)+1:=
```

G_1 GOAL ⊢ sin (π / 3) + 2 * cos (π / 12) ** 2 - cos (π / 2) = sqrt 3 + 1 PROOFSTEP

T_1 `rw cos_pi_div_two,`

G_2 GOAL ⊢ sin (π / 3) + 2 * cos (π / 12) ** 2 - 0 = sqrt 3 + 1 PROOFSTEP

T_2 `rw sin_pi_div_three,`

G_3 GOAL ⊢ sqrt 3 / 2 + 2 * cos (π / 12) ** 2 - 0 = sqrt 3 + 1 PROOFSTEP

• • •

T_n `field_simp,`

G_{n+1} GOAL this: cos (π / 12) ** 2 = cos (π / 6) / 2 + 1 / 2
⊢ sqrt 3 * (2 * 2 * 2) + 2 * (sqrt 3 * 2 + 2 * 2) * 2 = (sqrt 3 + 1) * (2 * (2 * 2 * 2))
PROOFSTEP

T_{n+1} `ring_exp,`

NO GOALS

Xiong et al., TRIGO: Benchmarking Formal Mathematical Proof Reduction for Generative Language Models, EMNLP2023 Proceedings

Content

引言

大语言模型评价方法研究

大语言模型的语言理解能力评价

大语言模型的语言生成能力评价

大语言模型的搜索和工具调用能力评价

大语言模型的其他专项能力评价

总结

总结

- 大语言模型的评价是一个非常复杂和困难的任务，面临多种挑战
- 我们在大语言模型评价方面开展了一系列工作，包括：
  - 评价方法的研究；
  - 模型理解能力的评价；
  - 模型生成能力的评价；
  - 模型检索和工具调用能力的评价；
  - 专项能力的评价。

Thank you!

把数字世界带入每个人、每个家庭、
每个组织，构建万物互联的智能世界。

Bring digital to every person, home and organization for a fully connected, intelligent world.

The information in this document may contain predictive statements including, without limitation, statements regarding the future financial and operating results, future product portfolio, new technology, etc. There are a number of factors that could cause actual results and developments to differ materially from those expressed or implied in the predictive statements. Therefore, such information is provided for reference purpose only and constitutes neither an offer nor an acceptance. Huawei may change the information at any time without notice.

HUAWEI