大模型（LLMs）评测面

来自： AiGC面试宝典

宁静致远

2023年09月29日 10:23

1 大模型怎么评测?

当前superGLUE, GLUE, 包括中文的CLUE 的benchmark都在不太合适评估大模型。可能评估推理能力、多轮对话能力是核心。

2 大模型的honest原则是如何实现的？模型如何判断回答的知识是训练过的已知的知识，怎么训练这种能力?

大模型需要遵循的helpful，honest， harmless的原则。

可以有意构造如下的训练样本，以提升模型准守honest原则，可以算trick了：

微调时构造知识问答类训练集，给出不知道的不回答，加强honest原则；

阅读理解题，读过的要回答，没读过的不回答，不要胡说八道。

3 如何衡量大模型水平?

要评估一个大型语言模型的水平，可以从以下几个维度提出具有代表性的问题。

- 理解能力：提出一些需要深入理解文本的问题，看模型是否能准确回答。
- 语言生成能力：让模型生成一段有关特定主题的文章或故事，评估其生成的文本在结构、逻辑和语法等方面的质量。
- 知识面广度：请模型回答关于不同主题的问题，以测试其对不同领域的知识掌握程度。这可以是关于科学、历史、文学、体育或其他领域的问题。一个优秀的大语言模型应该可以回答各种领域的问题，并且准确性和深度都很高。
- 适应性：让模型处理各种不同类型的任务，例如：写作、翻译、编程等，看它是否能灵活应对。
- 长文本理解：提出一些需要处理长文本的问题，例如：提供一篇文章，让模型总结出文章的要点，或者请模型创作一个故事或一篇文章，让其有一个完整的情节，并且不要出现明显的逻辑矛盾或故事结构上的错误。一个好的大语言模型应该能够以一个连贯的方式讲述一个故事，让读者沉浸其中。
- 长文本生成：请模型创作一个故事或一篇文章，让其有一个完整的情节，并且不要出现明显的逻辑矛盾或故事结构上的错误。一个好的大语言模型应该能够以一个连贯的方式讲述一个故事，让读者沉浸其中。
- 多样性：提出一个问题，让模型给出多个不同的答案或解决方案，测试模型的创造力和多样性。
- 情感分析和推断：提供一段对话或文本，让模型分析其中的情感和态度，或者推断角色间的关系。
- 情感表达：请模型生成带有情感色彩的文本，如描述某个场景或事件的情感、描述一个人物的情感状态等。一个优秀的大语言模型应该能够准确地捕捉情感，将其表达出来。
- 逻辑推理能力：请模型回答需要进行推理或逻辑分析的问题，如概率或逻辑推理等。这可以帮助判断模型对推理和逻辑思考的能力，以及其在处理逻辑问题方面的准确性。例如：“所有的动物都会呼吸。狗是一种动物。那么狗会呼吸吗？”
- 问题解决能力：提出实际问题，例如：数学题、编程问题等，看模型是否能给出正确的解答。
- 道德和伦理：测试模型在处理有关道德和伦理问题时的表现，例如：“在什么情况下撒谎是可以接受的？”
- 对话和聊天：请模型进行对话，以测试其对自然语言处理的掌握程度和能力。一个优秀的大语言模型应该能够准确地回答问题，并且能够理解人类的语言表达方式。

4 大模型评估方法 有哪些?

- 人工评估：LIMA、Phoenix

- 使用 GPT-4 的反馈进行自动评估：Vicuna、Phoenix、Chimera、BELLE指标评估（BLEU-4、ROUGE分数）：ChatGLM-6B；对于像ROUGE-L分数的指标评估，有些地方称其为非自然指令评估（Unnatural Instruction Evaluation）。
- Chatbot Arena：目前用来衡量一个模型好不好的东西基本都是基于一些学术的benchmark，比如在一个某个NLP任务上构建一个测试数据集，然后看测试数据集上准确率多少。然而，这些学术benchmark（如HELM）在大模型和聊天机器人上就不好用了。其原因在于：
    - 由于评判聊天机器人聊得好不好这件事是非常主观的，因此，现有的方法很难对其进行衡量。
    - 这些大模型在训练的时候就几乎把整个互联网的数据都扫了一个遍，因此，很难保证测试用的数据集没有被看到过。甚至更进一步，用测试集直接对模型进行「特训」，如此一来表现必然更好。
    - 理论上我们可以和聊天机器人聊任何事情，但很多话题或者任务在现存的benchmark里面根本就不存在。

因此，Chatbot Arena 的做法是放弃benchmark，通过对抗，实时聊天，两两比对人工进行打分，采用elo分数进行评测。

5 大模型评估工具 有哪些?

- OpenAI evals：OpenAI的自动化评估脚本，核心思路就是通过写prompt模版来自动化评估。
- PandaLM：其是直接训练了一个自动化打分模型，0,1,2三分制用模型对两个候选模型进行打分。