

对比不同大语言模型在NLP下游任务上的性能

刘峻池

2252298@qq.com

摘要

本实验通过旨在对比不同大语言模型在NLP下游任务上的性能，实验中对于LLaMA-13b、RWKV-4-Raven、Vicuna-13b、ChatGLM-6b四个大语言模型的原理进行了解释，并检验了其分别对于三种下游任务：机器翻译、情感检测、自动问答的能力，对于各模型在不同下游任务上的能力进行了对比与分析。

一、介绍

大语言模型（Large Language Model, LLM）是指具有大规模参数的语言模型，通常基于深度学习技术，如循环神经网络（RNN）或者转换器（Transformer）。这些模型被广泛应用于自然语言处理领域，能够生成与理解自然语言文本，具备理解、生成和推理等语言处理能力。

LLM的优势包括：

- 语言生成能力：**LLM能够生成自然流畅的文本，包括文章、对话、翻译等。通过训练大规模数据集，LLM能够学习到语言的模式、语法结构和上下文信息，从而生成具有逻辑连贯性和语言风格的文本。
- 多领域适应性：**由于LLM的训练数据可以涵盖多个领域的文本，因此它在不同领域的应用中都能够取得不错的效果。无论是新闻报道、科技论文还是社交媒体上的对话，LLM都可以适应并生成相关内容。
- 知识储备广泛：**LLM在训练过程中接触到大量的文本数据，因此具备广泛的知识储备。它可以回答各种领域的问题，提供背景信息、定义术语、解释概念等。
- 个性化应用：**LLM可以通过根据用户的输入和上下文生成个性化的回复。它可以学习到用户的偏好和习惯，从而为用户提供个性化的建议、推荐和对话。

然而，LLM也存在一些劣势：

- 缺乏常识推理能力：**尽管LLM可以从大规模数据集中学习到丰富的知识，但它在常识推理方面的能力还相对有限。LLM在处理一些需要常识推理的问题时可能会出现错误或生成不合适的回答。
- 受限于训练数据和偏差：**LLM的性能很大程度上依赖于训练数据的质量和数量。如果数据集存在偏差或错误，模型可能会学习到错误的知识或产生偏倚的回复。此外，如果LLM没有遇到某个特定领域或主题的训练数据，它可能会在该领域或主题上表现较差。
- 对抗攻击和滥用风险：**LLM容易受到对抗性攻击，即通过有意设计的输入来引导

其生成误导性或有害的输出。此外，LLM也可能被滥用，用于生成虚假信息、网络钓鱼等不良用途。

4. 能源和计算成本高：由于LLM拥有大规模的参数和复杂的结构，其训练和推理过程需要大量的计算资源和时间。这导致了高昂的能源和计算成本，限制了LLM在某些环境和设备上的应用。

综上所述，大语言模型具有生成流畅文本、多领域适应性和广泛知识储备等优势，但在常识推理能力、数据偏差、安全风险和资源消耗等方面仍存在一些劣势。未来的研究和发​​展将致力于进一步提高LLM的性能和解决其劣势，以更好地应对实际应用需求。

二、方法原理

1、LLaMA-13b[1]

LLaMA-13b模型的重点是训练一系列语言模型，通过训练比通常使用的更多的标记，在各种推理预算下实现最佳性能。参数范围从 7B 到 65B，与现有最好的 LLM 相比具有竞争力的性能。例如，尽管 LLaMA-13B 小了 10 倍，但在大多数基准测试中都优于 GPT-3。该模型将有助于使 LLM 的访问和研究民主化，因为它可以在单个 GPU 上运行。在规模的高端，LLaMA 模型的 65B 参数模型也可以与最好的大型语言模型（如 hinchilla 或 PaLM-540B）竞争。

LLaMA只使用了公开可用的数据，使得LLaMA模型与开源兼容，而大多数现有模型依赖于非公开可用或未记录的数据（例如“Books-2TB”）。

2、RWKV-4-Raven[2]

RWKV-4-Raven是一种基于RWKV-4架构的模型，它使用了一种新颖的循环权重键值（RWKV）机制来实现长期记忆。RWKV-4-Raven在Pile数据集上进行了训练和微调，还在Alpaca, CodeAlpaca, Guanaco, GPT4All, ShareGPT等数据集上进行了进一步的微调。它有不同大小和语言的版本，从1.5B到14B参数。它是一个强大的聊天模型，也可以执行各种指令。

RWKV-4-Raven的性能非常出色，它可以在不同的任务和领域上生成高质量的文本。RWKV-4-Raven的基本构架是一个L32-D4096的因果语言模型，它使用了一个双向RNN作为编码器和一个单向RNN作为解码器。与基于Transformer的模型相比，RWKV-4-Raven具有更好的长期记忆能力和更低的计算成本。它在多个评测指标上超越了其他模型，并展示了其灵活性和多样性。例如：

1. 在LAMBADA数据集上，RWKV-4-Raven 7B达到了4.38的困惑度（perplexity）和67.18%的准确率（accuracy），超过了GPT-3 13B（5.1和63.24%）和T5 11B（6.8和58.7%）。
2. 在PIQA数据集上，RWKV-4-Raven 7B达到了76.06%的准确率，超过了GPT-3 13B（74.2%）和T5 11B（73.9%）。
3. 在SC2016数据集上，RWKV-4-Raven 7B达到了73.44%的准确率，超过了GPT-3 13B（71.8%）和T5 11B（70.6%）。
4. 在Hellaswag数据集上，RWKV-4-Raven 7B达到了65.51%的归一化准确率（normalized accuracy），超过了GPT-3 13B（64.2%）和T5 11B（63.8%）。

RWKV-4-Raven的特色之一是它可以根据不同的语言和领域进行微调，形成不同的版本。

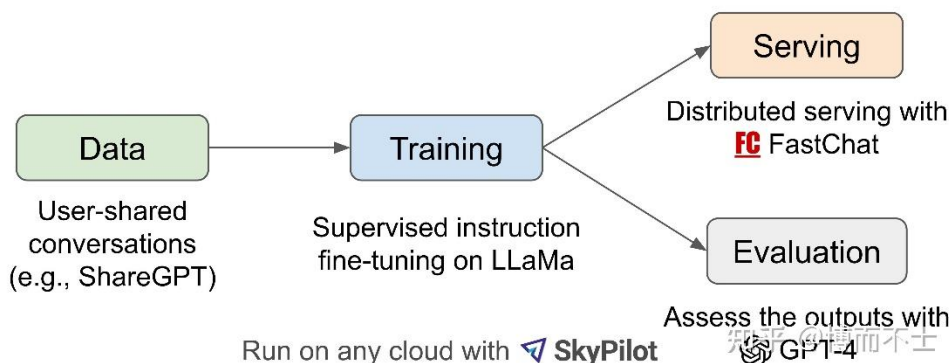
例如：

1. RWKV-4-Raven有英文和中文的版本，可以用于写小说或对话。它们可以在ChatRWKV v2中运行，也可以在Hugging Face Spaces中进行在线测试。
2. RWKV-4-Raven有指令微调的版本，可以用于执行各种指令，如编程、数学、科学等。它们可以使用+i标记来指定指令类型。
3. RWKV-4-Raven有多语言微调的版本，可以用于生成不同语言的文本。它们可以使用+lang标记来指定语言类型。

总之，RWKV-4-Raven是一种基于RNN的大语言模型，它利用了RWKV机制来实现长期记忆，并在多个数据集上进行了微调，以适应不同的任务和领域。它不仅保持了RNN的优点，还借鉴了Transformer的优点，并在多个方面超越了目前主流的大语言模型。RWKV-4-Raven是自然语言处理领域的一个重要贡献，值得我们关注和学习。

3、Vicuna-13b[3]

Vicuna是一个由LMSYS Org开发和分享的开源聊天机器人项目，其名称来源于一种生活在南美洲高原上的羊驼科动物。Vicuna共有两个版本，分别是Vicuna-7b-delta-v0和Vicuna-13b-delta-v0，分别拥有70亿和130亿个参数。这两个版本都是在2023年3月至4月之间训练出来的，基于LLaMA模型和ShareGPT数据集。LLaMA模型是一个轻量级、低成本、高效率的PLM，它采用了一种新颖的混合注意力机制（Hybrid Attention Mechanism），结合了自注意力（Self-Attention）和局部注意力（Local Attention），从而降低了计算复杂度和内存消耗，同时保持了较高的表示能力。ShareGPT数据集是一个由用户共享的对话数据集，包含了超过1亿条来自不同领域、主题、风格和情感的对话样本，覆盖了闲聊、问答、故事、诗歌、歌词等多种类型。通过在LLaMA模型上微调ShareGPT数据集，Vicuna可以获得更丰富、更多样、更个性化、更情感化的对话能力。Vicuna的总体训练流程如下图所示。



Vicuna的特色情况主要体现在三个方面，一是其开源性，二是其使用了GPT-4作为评委，三是其训练方法的创新性。首先，Vicuna是一个完全开源的项目，任何人都可以免费地使用、修改、分享和贡献Vicuna的代码和模型。这使得Vicuna具有很强的可扩展性和可定制性，用户可以根据自己的需求和喜好来调整Vicuna的参数、数据、风格等。同时，开源也有利于促进聊天机器人领域的交流和创新，让更多的人参与到这个有趣而又有意义的事业中来。

其次，Vicuna使用了GPT-4作为评委来进行模型评价。GPT-4是目前最先进的PLM之一，它拥有超过1000亿个参数，并且在多个自然语言处理任务上都取得了最佳或接近最佳的结果。

通过让GPT-4对不同聊天机器人生成的对话进行打分和评价，Vicuna可以获得一个更加客观、公正、权威的反馈，从而更好地优化自己的性能。最后，Vicuna采用了一种创新的训练方法，即基于用户共享的对话数据集ShareGPT。ShareGPT是一个由用户主动贡献和分享的对话数据集，它包含了来自不同领域、主题、风格和情感的对话样本，覆盖了闲聊、问答、故事、诗歌、歌词等多种类型。这种数据集具有很高的质量、多样性、个性化和情感化，可以为聊天机器人提供更加丰富和真实的语言知识和语义信息。

4、ChatGLM-6b[4]

ChatGLM-6B 是一个开源的、支持中英双语的对话语言模型，基于 General Language Model (GLM) 架构，具有 62 亿参数。结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存）。ChatGLM-6B 使用了和 ChatGPT 相似的技术，针对中文问答和对话进行了优化。经过约 1T 标识符的中英双语训练，辅以监督微调、反馈自助、人类反馈强化学习等技术的加持，62 亿参数的 ChatGLM-6B 已经能生成相当符合人类偏好的回答。

优点：

1. 较低的部署门槛：FP16 半精度下，ChatGLM-6B 需要至少 13GB 的显存进行推理，结合模型量化技术，一需求可以进一步降低到 10GB（INT8）和 6GB（INT4），使得 ChatGLM-6B 可以部署在消费级显卡上。
2. 更长的序列长度：相比 GLM-10B（序列长度1024），ChatGLM-6B 序列长度达 2048，支持更长对话和应用。
3. 人类意图对齐训练：使用了监督微调（Supervised Fine-Tuning）、反馈自助（Feedback Bootstrap）、人类反馈强化学习（Reinforcement Learning from Human Feedback）等方式，使模型初具理解人类指令意图的能力。输出格式为 markdown，方便展示。目前已开源监督微调方法，

不足：

1. 模型容量较小：6B 的小容量，决定了其相对较弱的模型记忆和语言能力，随着自己训练数据数量和轮次增加，会逐步丧失原来的对话能力。
2. 较弱的多轮对话能力：ChatGLM-6B 的上下文理解能力还不够充分，在面对长答案生成，以及多轮对话的场景时，可能会出现上下文丢失和理解错误的情况。

三、实验结果与分析

本实验中选取的自然语言下游任务来为：机器翻译、情感检测、自动问答，通过以上三类下游任务来对LLaMA-13b、RWKV-4-Raven、Vicuna-13b、ChatGLM-6b四种大语言模型的性能进行评估与对比。

1、机器翻译

提示：

请将以下英语翻译成中文“Looking beyond the 10-year period, the botanists

estimate that some 3,000 native plant species may become extinct in the foreseeable future—more than 10 percent of the approximately 25,000 species of plants in the United States.”

LLaMA-13b:

在 10 年 期间内，生物学家预测 3000 纯米质产品可能在近期时间内变成恢复不可能的。许多抱着这种产品恢复不可能的态度，但也有很多人它们进行保护。

RWKV-4-Raven:

在美国的研究人员们估计，在未来的10年内，可能会出现在美国大约3,000种原生植物中的10%，即超过3千种植物，可能会消失。

Vicuna-13b:

超越10年范围，植物学家预测未来可能会见证3000份美国原生植物种类的灭绝——超过10%的美国植物的约25,000种类。

ChatGLM-6b:

展望未来10年，植物学家估计在未来可预见的未来几年里，约3,000种原生植物物种可能会灭绝——这超过了美国约25,000种植物的10%。

期望答案:

在展望10年后的情况时植物学家们估计，在未来可预见到的时间内，3 000种本地植物——占美国近25 000种植物的10%——将可能灭绝。

本实验主要测试各模型将英文句子翻译为中文的准确性。从上述各模型根据提示所回答的内容可知，LLaMA-13b对于语句从英文单词翻译为中文词汇的准确性较差，同时逻辑性较差，翻译出的译句并不能很好的代表原句；RWKV-4-Raven翻译字词的准确性较高，但是其中忽略了“25000个种类”这一含义而仅仅翻译为“超过3千种”，丢失了原句中原本有的意思；而Vicuna-13b和ChatGLM-6b的翻译较为符合人们的阅读习惯，且字词翻译较准确，整体句子的逻辑性较强，并且保留了原句中的“——”这一标点符号。

整体来说Vicuna-13b和ChatGLM-6b的翻译结果最符合中文的阅读习惯，RWKV-4-Raven模型其次，LLaMA-13b的翻译结果最为不尽人意。

2、情感检测

提示:

请问以下语句对于电影的态度是积极的还是消极的 “我觉得这场电影也不是不好看”

LLaMA-13b:

这句话是消极的。

RWKV-4-Raven:

对于电影的态度，“我觉得这场电影也不是不好看” 是积极的。

Vicuna-13b:

这个语句对于电影的态度是消极的。

ChatGLM-6b:

这句语句对于电影的态度是积极的。尽管电影可能不是传统意义上的好看，但说话者认为它并不是不好看或不堪回首的。表明说话者对电影有一定的鉴赏能力，并且一定程度的欣赏和认可。

期望答案:

该语句对于电影的态度是积极的。

本实验主要测试各模型判断语句所包含的情感准确性，提示中“我觉得这场电影也不是不好看”其中包含了双重否定，“不是不好看”即为“好看”，是积极的情感。从上述各模型根据提示所回答的内容可知，LLaMA-13b和Vicuna-13b将该句子的情感判别为“消极”即代表其并不能很好的判断出句子中双重否定所表达的态度；而RWKV-4-Raven和ChatGLM-6b正确的判断出了该句子含有积极的情感，并且ChatGLM-6b对语句的翻译过程做出了解析，使得输出的结果更具有条理性。

整体来说，RWKV-4-Raven和ChatGLM-6b可以判断出较为复杂的语句其中所包含的情感，而LLaMA-13b和Vicuna-13b由于缺少对于双重否定的判断而并不能对于句子所包含的情感做出正确的判断。

3、自动问答

提示:

请以下面双括号中的语句为参考，回答什么是特殊效应。“一只生活在南美洲亚马孙河流热带雨林中的蝴蝶，偶尔扇动几下翅膀，两周后可能在美国的得克萨斯州引起一场龙卷风，这就是著名的理论假设‘拓扑学连锁反应理论’，也就是所谓的‘特殊效应’，并不是危言耸听，这说明一个什么道理呢，这说明，宇宙间万事万物都是相关联的。那么我们人也是一样，生于宇宙之间，我们谁都不可能脱离其他人事物的影响，也不可能完全不影响其他人和事。我们必须充分尊重这种系统性，只有这样呢，我们才能摆正自己的位置，达到天人合一，内外和谐的境界。这就是我们所说的系统。当一个人越能站在较高层面上理解事物，他就越能照顾全局，越能更好地解决问题。”

LLaMA-13b:

I am sorry, but I do not understand the question. Would you like to rephrase it?

RWKV-4-Raven:

诸如“一个人的蝴蝶效应是另一个人的龙卷风”之类的陈述是关于特殊性或“特殊效果”概念的陈述。该声明暗示，随着人们对自己和他人的理解和联系的增加，他们会更加意识到自己的行为对他人和周围世界的影响。本声明呼吁注意一个人的行为如何影响他人和环境，并努力建立一个更加联系和富有同情心的世界。

Vicuna-13b:

在这段话中，特殊效应是指一个事件在一个系统中发生的影响，会在其他地方的事件中产生影响。这种效应是系统性的，因为它是由系统内部的相互作用而产生的。这种效应的存在证明了宇宙是一个相互联系的整体，人们也是其中的一部分，必须尊重这种系统性，才能更好地理解和解决问题。

ChatGLM-6b:

双括号中的语句讲述了一种特殊效应，即拓扑学连锁反应理论。这个理论假设，任何一个微小的变化都可能在很短的时间内产生连锁反应，最终导致大规模的改变。这个理论在自然界中得到了广泛的应用，例如蝴蝶翅膀的轻轻扇动就可能导致一场龙卷风的产生。这个理论的意义在于，宇宙间万事万物都是相关联的，我们任何人都可能受到其他人事物的影响，也必须尊重这种系统性。如果我们能够站在更高层面上理解事物，就能够更好地照顾全局，解决问题。

期望答案:

蝴蝶效应是指微小的变化有可能带动整个系统产生长期的巨大的连锁反应。

本实验主要测试各模型从给定文本中提取出所需内容的能力，测试的过程中将文本中的“蝴蝶效应”更改为了“特殊效应”，以防止各模型根据对于“蝴蝶效应”的已有答案进行抄袭，从而更好的测试各模型自动问答的能力。

从上述各模型根据提示所回答的内容可知，LLaMA-13b在该测试中失效，并且回答为英文，可见该模型不仅并不能很好解决自动问答问题，并且在回答时的语言选择上仍存在障碍；RWKV-4-Raven的回答与标准答案相关，但并不完全准确，可见该模型有一定的自动问答能力，但是并不能很好的解决问题；而Vicuna-13b和ChatGLM-6b的都较好的回答了提示中的问题，并且根据全文对于答案进行了拓展解释，可见其均具有较强的自动问答能力。

整体来说，Vicuna-13b和ChatGLM-6b的能较好的从给定文本中提取出所需内容，并根据相关内容加以诠释，RWKV-4-Raven模型可以给出回答但结果不够准确，LLaMA-13b在遇到提示中的问题时失效。

四、结论

本实验通过旨在对比不同大语言模型在NLP下游任务上的性能，实验中对于LLaMA-13b、RWKV-4-Raven、Vicuna-13b、ChatGLM-6b四个大语言模型的原理进行了解释，并检验了其分别对于三种下游任务：机器翻译、情感检测、自动问答的能力，对于各模型在不同下游任务上的能力进行了对比与分析。

实验结果为，在机器翻译问题上，Vicuna-13b和ChatGLM-6b的翻译结果最符合中文的阅读习惯，RWKV-4-Raven模型其次，LLaMA-13b的翻译结果最为不尽人意。在情感检测问题上RWKV-4-Raven和ChatGLM-6b可以判断出较为复杂的语句其中所包含的情感，而LLaMA-13b和Vicuna-13b由于缺少对于双重否定的判断而并不能对于句子所包含的情感做出正确的判

断。在自动问答问题上，Vicuna-13b和ChatGLM-6b的能较好的从给定文本中提取出所需内容，并根据相关内容加以诠释，RWKV-4-Raven模型可以给出回答但结果不够准确，LLaMA-13b在遇到提示中的问题时失效。。

参考资料

- [1] LLaMA: Open and Efficient Foundation Language Models
- [2] RWKV-4-Raven模型介绍 <https://zhuanlan.zhihu.com/p/631693434>
- [3] Vicuna模型介绍 <https://zhuanlan.zhihu.com/p/630785688>
- [4] ChatGLM-6B模型介绍 <https://zhuanlan.zhihu.com/p/630099720>