

大模型下游任务测试

陆旭军 ZY2203320
18376486@buaa.edu.cn

摘要

本文介绍了选定的 3 个前沿语言大模型，GPT-4、GPT-3.5、文心一言，这些模型包括英文和中文模型。然后，本文选择了多个自然语言下游任务，涵盖了不同的应用领域，这些下游任务包括文本分类、命名实体识别、机器翻译、问答系统、文本生成等。我们在这些任务上评估了每个模型的性能，并比较它们之间的差异。为了进行测试，我们采用了提示工程的方法。通过我们的研究，我们得出了一些关于不同前沿语言大模型在不同自然语言下游任务上性能的结论。我们观察到不同模型在不同任务上表现出不同的优势和劣势。有些模型可能在某些任务上表现更好，而在其他任务上表现较差。最后，我们总结了我们的研究结果。

引言

1. 大模型选择

GPT-4 是由 OpenAI 开发的一款自然语言生成模型，于 2023 年 3 月 15 日发布。它是 OpenAI 的第四代自然语言生成模型，也是目前功能最全面的语言模型。GPT-4 在处理任务时展现了非常高的准确性和创造性，例如它能够参加各种职业考试、辅助创作、解答各种问题、进行文学创作、处理各种语言任务等。GPT-4 的优势在于它能够理解人类语言的语法和语义，并进行相应的处理和生成，它还可以根据用户的输入进行推理和总结。

GPT-3.5 是 OpenAI 的第三代自然语言生成模型，于 2020 年 11 月发布。GPT-3.5 已经在各种自然语言处理任务中表现出了非常高的性能，包括文本分类、问答、对话等。GPT-3.5 的优势在于它能够生成连贯、流畅和自然的文本，并且能够理解和应用大量的语言知识。

文心一言是百度研发的知识增强大语言模型，能够与人对话互动，回答问题，协助创作，高效便捷地帮助人们获取信息、知识和灵感。文心一言的优势在于它能够生成高质量的中英文文本，并且能够进行中文语音合成。此外，它还能够应用于多种自然语言处理任务，包括问答、翻译、文本生成等。^[1]

2. 提示工程

提示工程（Prompt Engineering）是一种自然语言处理技术，它通过设计一些有效的提示语言来帮助模型更好地理解 and 生成文本。提示工程已经在自然语言处理领域中得到了广泛的应用，特别是在对话系统、问答系统、文本生成和摘要等任务中。

传统的自然语言处理模型通常需要针对特定的任务和数据集进行训练和微调，例如问答任务、文本分类任务、命名实体识别任务和机器翻译任务等。然而，提示工程提供了一种更加通用的方法，它可以通过设计一些通用的提示语言来帮助模型更好地理解 and 生成文本，而不需要针对特定的任务和数据集进行训练和微调。这种方法使得模型的训练和部署更加容易和高效，并且可以大大降低开发成本。

提示工程的核心思想是利用模型在预训练过程中学到的知识，设计一些有效的提示语言来引导模型生成所需的文本。这些提示语言通常由一些关键字和结构组成，可以告诉模型要生成的内容类型（例如实体、情感、动作等）和文本结构（例如序列、分类、归纳等）。通过使用这些提示语言，模型可以更加准确地理解和生成文本，从而提高模型的性能和鲁棒性。

提示工程的优点之一是它可以帮助模型在特定任务上具有更好的控制性和可解释性。通

过明确指定任务的要求，我们可以更容易地解释模型的输出是如何生成的，并对其进行调整和优化。此外，提示工程还可以帮助解决语言模型中的一些问题，如倾向性、生成不可信信息等。

提示工程已经在各种自然语言处理任务中取得了显著的成果，包括文本分类、命名实体识别、机器翻译、问答系统等。它为改善语言模型的性能和可控性提供了一种有效的方法，并且在实际应用中具有广泛的潜力。

3. 下游任务 LSTM

问答任务：在这个任务中，模型的性能可以通过评估它们在问答数据集上的表现来确定。常用的数据集包括 SQuAD、CoQA 和 TriviaQA 等。

文本分类任务：在这个任务中，模型的性能可以通过评估它们在文本分类数据集上的表现来确定。常用的数据集包括 SST-2、情感分析数据集和 IMDB 等。

命名实体识别任务：在这个任务中，模型的性能可以通过评估它们在命名实体识别数据集上的表现来确定。常用的数据集包括 CoNLL-2003 和 CoNLL-2004 等。

机器翻译任务：在这个任务中，模型的性能可以通过评估它们在机器翻译数据集上的表现来确定。常用的数据集包括 WMT14 English-French、WMT14 English-German 和 IWSLT14 German-English 等。

文本生成任务：在这个任务中，模型的性能可以通过评估它们在文本生成数据集上的表现来确定。常用的数据集包括 GLUE 的 Text Generation 任务、SST-2 的生成任务和 SICK-sentence Completion 任务等。

实验过程

1. 选择要比较的模型

确定要比较的模型，如 GPT-3.5、GPT-4 和文心一言等。

2. 选择下游任务

选择要评估的自然语言处理任务，如问答任务、文本分类任务、机器翻译任务等。

3. 准备测试数据

为每个任务准备一组测试数据。测试数据应包含输入文本和相应的期望输出。为每个任务编写明确的提示，引导模型理解任务要求。提示应简洁明了，可包含示例。例如，在问答任务中，可以用以下格式：“问题：... 答案：...”。以下为测试案例：

(1) 问答任务（测试案例）： 文本：量子力学是描述微观粒子行为的一种物理理论，它的发展始于 20 世纪初，主要贡献者包括尼尔斯·玻尔、沃纳·海森堡、埃尔温·薛定谔等。量子力学基本原理包括波动性、不确定性原理和波函数坍缩。 问题：量子力学的发展始于哪个世纪？

(2) 文本分类任务（测试案例）： 文本：这部电影的剧情紧凑，角色塑造生动，让我深感震撼。导演的表现也堪称一绝，我会向朋友们推荐这部作品。

(3) 命名实体识别任务（测试案例）： 文本：乔治·华盛顿是美国的第一任总统，他在 1789 年至 1797 年间任职。他曾领导美国大陆军在独立战争中取得胜利，被誉为美国的国父。

(4) 机器翻译任务（测试案例）： 源文本（英语）：The quick brown fox jumps over the lazy dog. 目标文本（法语）：

(5) 文本生成任务（测试案例）： 输入：科学家们发现了一种新的太阳能电池材料，这种材料可以大幅提高太阳能电池的转换效率。这一发现可能会对可再生能源产业产生重大

影响。

实验结果

1.输出结果

结果按测试案例顺序、按模型输出：

(1) 文心一言

A1:

量子力学的发展始于 20 世纪。

A2:

这段文本可以被分类为“电影评论/推荐”类。根据文本内容，作者表达了对电影的观感和评价，并表示愿意向朋友推荐这部作品。这种文本通常包含对电影的描述、评价和推荐，因此可以将其归类为电影评论或推荐类文本。

A3:

命名实体：乔治·华盛顿（人名）、美国（地名）、1789 年（时间）、1797 年（时间）、独立战争（事件）、美国大陆军（组织）、国父（称号）

A4:

作为一个人工智能语言模型，我还没学习如何回答这个问题，您可以向我问一些其它的问题，我会尽力帮您解决的。

A5:

科学家们发现了一种新的太阳能电池材料，这种材料可以大幅提高太阳能电池的转换效率，使太阳能电池更加高效。这一发现可能会对可再生能源产业产生重大影响，因为它可以为太阳能电池的制造提供更加高效的材料，从而降低成本，提高效率，使可再生能源更加具有竞争力。

(2) GPT-3.5

A1:

对于输入文本：文本：量子力学是描述微观粒子行为的一种物理理论，它的发展始于 20 世纪初，主要贡献者包括尼尔斯·玻尔、沃纳·海森堡、埃尔温·薛定谔等。量子力学基本原理包括波动性、不确定性原理和波函数坍缩。问题：量子力学的发展始于哪个世纪？

模型的期望输出应为："20 世纪初"。

A2:

对于输入文本：文本：这部电影的剧情紧凑，角色塑造生动，让我深感震撼。导演的表现也堪称一绝，我会向朋友们推荐这部作品。

模型的期望输出应为："正面情感"。

A3:

对于输入文本：文本：乔治·华盛顿是美国的第一任总统，他在 1789 年至 1797 年间任职。他曾领导美国大陆军在独立战争中取得胜利，被誉为美国的国父。

模型的期望输出应为："乔治·华盛顿"(人名)。

A4:

对于输入文本：源文本（英语）："The quick brown fox jumps over the lazy dog."

目标文本（法语）：

模型应该生成的期望输出是法语的翻译，这里缺少目标文本的具体内容。请提供您想要的法语目标文本，以便我可以生成对应的翻译输出。

A5:

对于输入文本：输入：科学家们发现了一种新的太阳能电池材料，这种材料可以大幅提

高太阳能电池的转换效率。这一发现可能会对可再生能源产业产生重大影响。

模型的生成输出可以是一种关于该发现的具体描述，例如：

输出：这种新的太阳能电池材料具有独特的光吸收特性，能够更高效地转化太阳能为电能。其独特的材料结构和光电转换机制为太阳能电池技术带来了革命性的突破。这一发现有望推动可再生能源产业的发展，加速可再生能源在能源供应中的比重，从而减少对传统能源的依赖，实现更环保和可持续的能源未来。

(3) GPT-4

A1:

答案：量子力学的发展始于 20 世纪。

A2:

分类：正面评价

A3:

实体：乔治·华盛顿（人名），美国（地名），1789 年（时间），1797 年（时间），美国大陆军（组织），独立战争（事件）

A4:

目标文本（法语）：Le renard brun rapide saute par-dessus le chien paresseux.

A5:

输出：科学家们近日在太阳能电池领域取得了重大突破，他们发现了一种新型材料，有望显著提高太阳能电池的能量转换效率。这一创新性成果有望为可再生能源行业带来革命性的变革，推动全球绿色能源发展。

2.结果分析

本文对所有问题回答的评分如下表 1 所示，每个项目 5 分为满分

	问答系统	文本分类	实体识别	机器翻译	文本生成	总分
GPT-4	5	5	5	5	4	24
GPT-3.5	5	5	5	2	4	21
文心一言	5	5	5	0	3	18

问答系统：

GPT-4、GPT-3.5 和文心一言在问答系统任务上都获得了最高分 5 分，表明它们在回答问题和提供相关信息方面都表现出色。

文本分类：

GPT-4、GPT-3.5 和文心一言在文本分类任务上均获得了 5 分，说明它们在准确地分类文本情感或主题方面都具备很好的能力。

实体识别：

GPT-4、GPT-3.5 和文心一言在实体识别任务上也获得了 5 分，表明它们在准确地识别文本中的命名实体方面具备较高水平。

机器翻译：

GPT-4 在机器翻译任务上获得了 5 分，说明它在准确翻译文本方面具备出色的能力。

GPT-3.5 在机器翻译任务上获得了 2 分，可能意味着它在处理机器翻译任务时相对较弱，可能出现了一些翻译不准确或不流畅的情况。

文心一言在机器翻译任务上获得了 0 分，可能意味着它在这个任务上没有很好的表现或者没有提供相应的功能。

文本生成：

GPT-4 和 GPT-3.5 在文本生成任务上获得了 4 分，表明它们在生成与输入文本相关的连贯和有意义的文本方面表现出良好的能力。

文心一言在文本生成任务上获得了 3 分，说明它也能生成一些相关的文本，但可能在连贯性和内容质量上稍显不足。

总结来说，GPT-4 在所有任务中表现最好，获得总分 24 分。GPT-3.5 紧随其后，总分为 21 分。文心一言在大部分任务上也表现良好，但在机器翻译任务上得分较低，总分为 18 分。

总结

本次作业旨在通过提示工程的方法来检验和对比不同前沿语言大模型在不同下游任务上的性能。我们选择了问答任务、文本分类任务、命名实体识别任务、机器翻译任务和文本生成任务作为测试任务。

根据评分表格的结果，GPT-4 在所有任务中获得了最高的总分 24 分，表现出色。其在问答、文本分类和实体识别任务上均获得满分。GPT-3.5 紧随其后，总分为 21 分，表现也相当不错。文心一言在大部分任务上表现良好，但在机器翻译任务上得分较低，总分为 18 分。

这次作业使我们对这些前沿语言大模型在不同任务上的性能有了更深入的了解。这些模型在处理问答、文本分类和实体识别任务方面表现优秀，但在机器翻译和文本生成任务上还有进一步提升的空间。

参考文献

[1]<https://baike.baidu.com/item/%E6%96%87%E5%BF%83%E4%B8%80%E8%A8%80/62642976?fr=aladdin>