# 基于下游任务的大语言模型性能测试

申海（ZY2203325）

## **摘要：**

本报告旨在对三种大语言模型（ChatGPT3.5、ClueAI/ChatYuan-large-v1和new bing）进行情感分析、机器翻译和文本生成下游任务的测试。我们选择了五个测试问题，分别涵盖了对句子情感色彩的分析、句子的翻译和生成关于生活的诗歌。通过对这些测试问题的实验，我们评估了各个模型在不同任务上的性能表现，并对其进行了比较和总结。

## **引言**

大语言模型在自然语言处理领域取得了显著的成就，其强大的生成能力和语义理解能力为各种下游任务提供了潜力巨大的应用。情感分析、机器翻译和文本生成是自然语言处理中的重要任务，对模型的语义理解、生成能力和语境理解提出了挑战。因此，我们选择了这些任务作为测试的对象，以评估大语言模型在实际应用中的效果和可行性。

## 实验

### 模型选择

我们选择了三种大语言模型进行测试，分别是ChatGPT3.5、ClueAI/ChatYuan-large-v1和new bing。

**ChatGPT3.5:**

ChatGPT3.5是基于GPT-3.5架构的一种大型语言模型，由OpenAI开发。它是在GPT-3的基础上进行改进和扩展的版本。该模型具有1750亿个参数，拥有强大的语言生成和理解能力。

ChatGPT3.5通过预训练和微调的方式进行训练。在预训练阶段，模型通过大量的文本数据学习语言的统计规律和语义信息。在微调阶段，模型通过在特定任务上进行训练，如对话生成、问题回答等，进一步提升模型在具体任务上的性能。

该模型在多个下游任务上展现了强大的表现，包括对话生成、文本摘要、机器翻译等。它能够根据上下文理解和生成连贯的自然语言文本，具有一定的创造性和推理能力。然而，由于模型规模庞大和计算资源要求高，使用ChatGPT3.5可能需要较长的推理时间和较大的计算资源。

**ClueAI/ChatYuan-large-v1:**

ClueAI/ChatYuan-large-v1是一种基于中文的大型语言模型，由ClueAI开发。它是在BERT架构的基础上进行扩展和改进的版本。该模型具有超过1亿个参数，专注于中文自然语言处理任务。

该模型通过预训练和微调的方式进行训练，其中预训练阶段利用了大量的中文文本数据。通过对大规模数据的学习，模型能够理解中文语言的语义和语法规律，并具备一定的语义理解和生成能力。

ClueAI/ChatYuan-large-v1在多个中文自然语言处理任务上表现出色，如情感分析、命名实体识别、文本分类等。它能够根据上下文进行推理和生成自然流畅的中文文本，具有较高的准确性和语义理解能力。

**new bing:**

new bing是一种基于中文的大型语言模型，由一个团队开发。该模型专注于中文自然语言处理任务，并在大规模中文文本数据上进行训练。

该模型采用了类似于GPT的自回归架构，在训练过程中学习中文语言的统计规律和语义信息。它具有较大的模型规模，能够处理较长的文本序列，并具备一定的生成能力。

new bing在中文自然语言处理任务中展现了良好的性能，如中文文本生成、文本摘要、机器翻译等。它能够根据上下文生成连贯的中文文本，具有较高的语义理解和生成能力。

尽管这些模型在各自的任务和语言上都有出色的表现，但它们的使用需要充足的计算资源和推理时间。在选择适合的模型时，需要考虑到具体任务需求、数据规模和可用的计算资源，以获得最佳的性能和效果。

### 3.2下游任务概述

下游任务是指在自然语言处理领域中，基于预训练的大型语言模型（如ChatGPT3.5、ClueAI/ChatYuan-large-v1和new bing）进行微调或使用的具体任务。这些任务通常是基于自然语言文本的处理和生成，旨在解决实际应用中的具体问题。本文选择情感分析、机器翻译和文本生成进行测试。

**情感分析：** 情感分析是一种文本分类任务，旨在确定给定文本中的情感色彩，如积极、消极或中性。通过对文本进行情感分析，可以了解用户对特定事物、产品、服务或观点的情感倾向。情感分析在社交媒体分析、品牌声誉管理、用户评论分析等领域具有广泛的应用。

**机器翻译：** 机器翻译是将一种语言的文本自动转换成另一种语言的过程。在机器翻译任务中，模型接受源语言的文本作为输入，并生成目标语言的翻译结果。机器翻译在跨语言交流、多语言文档处理和全球化应用中具有重要作用。通过使用大型语言模型进行机器翻译，可以提高翻译质量和效率。

**文本生成：** 文本生成是指通过模型自动生成符合语法和语义规则的自然语言文本。这种生成可以是根据给定的提示或上下文生成连贯的句子、段落或文章。文本生成在内容创作、智能问答系统、对话生成等领域有广泛的应用。使用大型语言模型进行文本生成可以产生更加自然、流畅和富有创造性的文本。

在这些下游任务中，使用预训练的大型语言模型可以有效地捕捉文本的语义和语法规律，并具备一定的语言理解和生成能力。通过微调或使用这些模型，可以加快任务的开发和部署过程，并取得良好的性能和效果。

### 3.3结果及分析

|  |  |  |
| --- | --- | --- |
| 任务序号 | 任务类型 |  |
| 1 | 情感分析 | "“这部电影太好看了，情节紧凑，演员表演出色。”  分析一下上述语句的感情色彩" |
| 2 | "“我感到非常失望，这家餐厅的服务质量太差了。”  分析一下上述语句的感情色彩" |
| 3 | "“这不可能吧，我很清楚记得电话号码。”  分析一下上述语句的感情色彩" |
| 4 | 机器翻译 | "“飞机若想尽快实现地面机动转弯，则前轮转弯操纵系统操纵速率应尽量大，而为了保证飞机地面机动不会引起前（主）机轮侧向滑动，又要求前轮转弯操纵速率应限制在一定范围内。”将上述语句翻译成英文" |
| 5 | 文本生成 | "写一首诗歌，关于生活" |

**对于情感分析的结果如下：**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Chatgpt3.5 | ClueAI/ChatYuan-large-v1 | New bing |
| 任务1 | 积极 | 中性 | 积极 |
| 任务2 | 消极 | 中性 | 消极 |
| 任务3 | 困惑或怀疑 | 肯定 | 惊讶和困惑 |

分析：可以看到chatgpt3.5和基于chatgpt4.0开发的new bing 均能很好的分析出情感，而ClueAI/ChatYuan-large-v1就比较差，将积极的情感分析成了中性，将消极的也分析成了中性，而比较难分辨的困惑和惊讶分析成了肯定。

**对于机器翻译结果如下：**

|  |  |
| --- | --- |
| 参考翻译（论文） |  |
| Chatgpt3.5 | If an airplane wants to achieve ground maneuvering turns as quickly as possible, the steering rate of the nose wheel control system should be maximized. However, to ensure that ground maneuvers do not cause lateral skidding of the front (main) landing gear, the steering rate of the nose wheel should be limited within a certain range. |
| ClueAI/ChatYuan-large-v1 | Aircraft want to get ground turn quickly, the running rate of forewheel tunnel manipulator should be as big as possible. To ensure that the front wheel slipping cannot cause the first whe |
| New bing | If an airplane wants to achieve ground maneuvering turns as quickly as possible, the steering rate of the front wheel steering system should be as large as possible. However, in order to ensure that the ground maneuvering of the airplane does not cause lateral sliding of the front (main) wheels, the steering rate of the front wheel should be limited within a certain range. |

**分析：**可以看到chatgpt3.5和new bing 翻译得很像，同时new bing 翻译地稍好一些，而ClueAI/ChatYuan-large-v1也翻译了大致，但并不是很好。

**文本生成结果：**

|  |  |
| --- | --- |
| Chatgpt3.5 | 在日出的瞬间，生活开始绽放，  如同一幅绚烂的画卷展开。  它是一场奇妙的旅程，  将我们带入未知的天地。  生活如潺潺流水，  静静地穿过岁月的河流。  它是一曲悠扬的旋律，  唤醒心灵的深处的琴弦。  时而波澜壮阔，时而平和宁静，  生活的风景千变万化。  它是一次奇幻的冒险，  每一刻都充满了惊喜和感悟。  生活是勇气的燃烧，  面对挑战，我们不畏惧。  它是一座高山的攀登，  脚步跨过困难，迎接辉煌。  在生活的舞台上尽情演绎，  用心感受每个瞬间的呼吸。  它是一幅绘画的画布，  我们是艺术家，创造自己的意义。  生活，如此美丽而深邃，  每一天都是宝贵的礼物。  让我们用爱和希望编织，  一首关于生活的永恒诗篇。 |
| ClueAI/ChatYuan-large-v1 | 生活像一团乱麻， 心烦意乱，  总是找不到出路，  只能去追求梦想， \_  每天重复着这种生活， 但总有一丝幸福，  你总是不知道， 生活就像一场梦， 却总是无法找到出口，  生活就像一匹狼， 只有坚强，  才有未来，  生活中只有我们自己， 才能创造奇迹，  生活的路总是如此漫长， 你永远不知道结果，  只能继续探索， 去找寻答案。 |
| New bing | 生活，如一首歌， 旋律轻快，时而悠长。 忙碌的脚步，匆匆忙忙， 却不忘享受每一天。  清晨的阳光，温暖而明亮， 傍晚的微风，轻柔而舒畅。 每一天都是新的开始， 生活，总是充满希望。  我们在其中奔波劳碌， 却不忘抬头仰望星空。 生活，有苦有甜， 我们在其中成长。  让我们珍惜每一天， 把握每一个机会。 生活，如一首歌， 让我们用心去感受。 |

**分析：**三者均很好地生成了诗歌，但一定要排序的话，我认为chatgpt3.5>new bing> ClueAI/ChatYuan-large-v1。

### 3.4结论

通过对ChatGPT3.5、ClueAI/ChatYuan-large-v1和new bing三种大语言模型在情感分析、机器翻译和文本生成下游任务的测试，我们得出以下结论：

ChatGPT3.5在情感分析任务中表现较好，能够准确判断句子的情感色彩。在情感分析中，ChatGPT3.5=New bing> ClueAI/ChatYuan-large-v1

New bing在机器翻译任务中表现较好，能够准确翻译中文句子为英文。在机器翻译中，我认为New bing >ChatGPT3.5 > ClueAI/ChatYuan-large-v1

文本生成任务中，三个语言模型表现都很好，能够生成创造性的关于生活的诗歌。一定要排序的话，我认为chatgpt3.5>new bing> ClueAI/ChatYuan-large-v1。

然而，这些模型在某些情况下可能存在一些语法、流畅性或语义理解的问题，需要进一步的改进和优化。综合考虑模型的性能和可行性，我们建议根据具体任务的需求和特点选择适合的模型，并进行进一步的调优和训练，以达到更好的性能和效果。

需要注意的是，本报告仅涵盖了对三种大语言模型在情感分析、机器翻译和文本生成任务上的初步测试结果，更全面的评估需要考虑更多因素和更多样本数据。同时，随着技术的不断发展和模型的进一步优化，模型的性能和效果也可能发生变化。因此，在实际应用中，建议根据具体任务和需求进行综合考虑和选择适合的模型。

同时也需要注意的是，本次下游任务测试并没有进行小样本训练，而是直接使用，这可能对数据量不大的语言模型有不公平之处。