

基于 Prompt 的 LLM 模型 NLP 下游任务测试

ZY2203810

吴金旺

Abstract

本文将简单介绍 LLM（Large Language Model），以及伴随着 LLM 同步发展的提示工程（Prompt），并基于 Prompt 对 LLM 解决 NLP 下游任务的能力进行测试。

Introduction

LLM 是一种利用大量的语料数据，训练超大型的神经网络得到的深度学习模型。它最大的特点就是体量大，一般认为参数超过十亿（1B）以上才可能被称为 LLM。目前基本所有的 LLM 都是基于 Transformer 架构，这种架构能够降低模型参数过大带来的过拟合风险，伴随着最近硬件计算能力的提升，掀起一股 LLM 的热潮。

得益于大量的参数以及在预训练阶段见过的大量语料，LLM 可以零样本（Zero-Shot）的解决绝大多数自然语言处理（NLP, Natural Language Processing）方面的问题，不需要在训练集上再做训练。而如果给它几个样例，则称为少样本（Few-Shot）学习，这会进一步加强模型解决问题的能力。LLM 还展现出了一种“涌现性”，即参数量大于一定数目时，模型就仿佛解锁了某种能力一般，突然在某些任务上取得了很好的效果，而小于这一数目，则效果会非常差，这是一种量变到质变的过程。

LLM 与传统模型相比，最大的不同在于增加上下文学习（in-context learning）这一项能力，不再需要像传统的预训练模型一样再对模型参数 finetune。具体来说，利用一些提示语（Prompt）就可以让 LLM 认识到当前要执行什么任务，这样它就能在不同任务之间转换，解决多个任务。

自从 ChatGPT 问世以来，大众们第一次看到 LLM 的颠覆性效果，这也兴起了一波热潮。接下来将简单介绍几种 LLM：

- 1) ChatGPT：毫无疑问，ChatGPT 是目前最火热的 LLM，作为 OpenAI 的一个聊天用大模型，几乎成为了一类产品的代名词。ChatGPT 基于 GPT3.5

改进而来，具有 175B 参数量，是 OpenAI 从 GPT1 开始迭代的结果，它选取了 Transformer 的 Docoder 结构用于文本生成，通过大量数据的训练最终达到了划时代的效果。

2) New Bing: 也可以直接称为新必应，是微软搭载在它的搜索引擎 Bing 上的聊天 LLM。早期的 New Bing 基于 ChatGPT，后来随着 GPT4.0 的开发，最新的 New Bing 已经基于 GPT4.0。它的功能基本与 ChatGPT 类似，但可以执行上网搜索的功能。

3) ChatGLM: 这是由清华大学研发的对话机器人，参数最大达到了 130B。但是本次测试只使用了它的在 HuggingFace 上的开源版本 ChatGLM-6B，由于它是由中国人研发的 LLM，因此对中文的支持应该还是相对较好的。

本次实验将使用 LLM 对各个 NLP 的下游任务进行测试，因此选择合适的下游任务也是十分重要的。所以接下来将介绍几个典型的 NLP 下游任务，为后续实验打下基础。

1) 机器翻译: 翻译任务可以认为是最接近我们生活学习的一项任务，简单来说，给定一种语言的文本，模型要将它转换成另一种语言的文本，且表达的意思应该要相同。

2) 文本摘要: 正如字面意思，给定一段文本，模型要提取出文本中的核心内容，并形成一段摘要。

3) 文本生成: 即让模型生成一段符合要求的文本内容。

4) 文本纠错: 模型需要识别并纠正拼写错误、语法错误、标点符号错误等，以提高文本的质量和准确性。

Methodology

本次实验将基于 Prompt，测试三种 LLM 在机器翻译、文本摘要、文本生成和文本分类四个任务上的实验结果。

具体来说，对四种任务，我们分别设计如下 Prompt:

1) 机器翻译: 给定一段中文文本，让模型生成它的英文。

中文: \$ (中文内容)

英文:

2) 文本摘要：给定一段中文文本，让模型生成摘要。

文本内容: \$ (文本内容)

摘要:

3) 文本生成：给定一段中文文本，让模型生成后续内容。

文本内容: \$ (文本内容)

续写:

4) 文本纠错：给定一段中文文本，让模型生成对这段内容的纠错后的结果

错误文本: \$ (文本内容)

正确文本:

Experimental Studies

A. 机器翻译

Prompt:

中文：LLM 是一种利用大量的语料数据，训练超大型的神经网络得到的深度学习模型。它最大的特点就是体量大，一般认为参数超过十亿（1B）以上才可能被称为 LLM。目前基本所有的 LLM 都是基于 Transformer 架构，这种架构能够降低模型参数过大带来的过拟合风险，伴随着最近硬件计算能力的提升，掀起一股 LLM 的热潮。

英文：

1) ChatGPT: LLM, short for Large Language Model, is a deep learning model trained on a massive amount of textual data using a large-scale neural network. Its most prominent feature is its sheer size, typically requiring billions (1B) of parameters to be considered an LLM. Currently, almost all LLMs are based on the Transformer architecture, which helps mitigate the risk of overfitting associated with extremely large model parameters. With the recent advancements in hardware computing capabilities, there has been a surge of interest and enthusiasm surrounding LLMs.

2) New Bing: LLM stands for Large Language Model. It is a type of deep learning model that is trained on a large corpus of data using a very large neural network. The main characteristic of an LLM is its size, with models typically having over 1 billion (1B) parameters. Most LLMs are based on the Transformer architecture, which helps to reduce the risk of overfitting that can arise from having too many model parameters. With recent advances in hardware and computing power, there has been a surge in interest in LLMs.

3) ChatGLM: LLM is a deep learning model obtained by training a very large language model with a large amount of corpus data. Its biggest feature is its size, and it is generally considered that models with parameters above 1 billion (1B) are likely to be called LLM. Currently, almost all LLM are based on the Transformer architecture, which can reduce the risk of overfitting by reducing the model parameters. With the

recent improvement in hardware computing capabilities, a 热潮 for LLM has emerged.

可以看出 ChatGPT 和 New Bing 效果都较好，而 ChatGLM 则效果较差，甚至出现了中文。但以本人薄弱的英语水平，无法分别出谁更地道，于是将 ChatGPT 的翻译给了 New Bing，让它做一个判断，输出结果如下：

Both translations convey the same information and are accurate. It's a matter of personal preference which one is better. Which one do you prefer?

可以看到 New Bing 认为二者都很好，无法做出判断。

B. 文本摘要

Prompt:

文本内容：得益于大量的参数以及在预训练阶段见过的大量语料，LLM 可以零样本（Zero-Shot）的解决绝大多数自然语言处理（NLP, Natural Language Processing）方面的问题，不需要在训练集上再做训练。而如果给它几个样例，则称为少样本（Few-Shot）学习，这会进一步加强模型解决问题的能力。LLM 还展现出了一种“涌现性”，即参数量大于一定数目时，模型就仿佛解锁了某种能力一般，突然在某些任务上取得了很好的效果，而小于这一数目，则效果会非常差，这是一种量变到质变的过程。LLM 与传统模型相比，最大的不同在于增加上下文学习（in-context learning）这一项能力，不再需要像传统的预训练模型一样再对模型参数 finetune。具体来说，利用一些提示语（Prompt）就可以让 LLM 认识到当前要执行什么任务，这样它就能在不同任务之间转换，解决多个任务。

摘要：

1) ChatGPT: 这段文本主要介绍了大型语言模型（LLM）在自然语言处理（NLP）领域的应用。LLM 凭借其庞大的参数和预训练阶段所见的大量语料，能够在零样本情况下解决许多 NLP 问题，无需在训练集上进行额外训练。而如果给予少量样本，就可以进行少样本学习，进一步增强模型的问题解决能力。LLM 还表现出了一种“涌现性”，当参数量超过一定数量时，模型就好像解锁了某种能力，突然在某些任务上表现出良好的效果，而当参数量小于此数量时，效果会非常差，这是一种由数量到质量的转变过程。

与传统模型相比，LLM 最大的不同之处在于引入了上下文学习的能力，不再需要像传统的预训练模型那样对模型参数进行微调。具体来说，通过使用一些提示语（Prompt），LLM 可以了解当前要执行的任务，从而能够在不同任务之间进行转换，解决多个任务。

请问您对这段文本还有其他问题吗？

2) New Bing: LLM 是一种能够零样本或少样本解决自然语言处理问题的模型。它具有“涌现性”，当参数量大于一定数目时，模型能够在某些任务上取得很好的效果。LLM 最大的不同在于增加了上下文学习能力，可以通过提示语来

识别当前要执行的任务，从而解决多个任务。

3) ChatGLM: LLM(Language Learning Model)是一种能够解决自然语言处理问题的预训练模型。它通过在大量语料上进行预训练，拥有大量的参数，从而能够零样本或少样本解决大多数自然语言处理问题。LLM 还展现出了一种“涌现性”，即参数量大于一定数目时，模型就能够突然在某些任务上取得很好的效果。与传统模型相比，LLM 增加了上下文学习的能力，不再需要对模型参数进行 finetune，从而实现在不同任务之间转换的能力。

显然，New Bing 的效果最好，它完全理解了任务的需求，并很好的回答了内容，ChatGLM 也完成了任务，然而 ChatGPT 反而并没有很好的完成摘要，反而对内容进行了解释。进一步的询问 ChatGPT:“帮我总结一下之前的那段内容”，其结果仍然不符合预期：“这段文本主要介绍了大型语言模型（LLM）在自然语言处理（NLP）领域的应用。LLM 具有庞大的参数和预训练阶段所见的大量语料，使其能够在零样本情况下解决绝大多数 NLP 问题。它还可以进行少样本学习，通过给予少量样本来进一步提升模型的问题解决能力。LLM 展现出一种“涌现性”，当参数数量超过一定阈值时，在某些任务上取得突出的效果，而小于此阈值时效果较差，这是一种量变到质变的过程。

与传统模型相比，LLM 的主要不同之处在于引入了上下文学习能力，无需对模型参数进行微调。通过使用提示语（Prompt），LLM 可以识别当前要执行的任务，并能够在不同任务之间切换和解决多个任务。

总的来说，LLM 利用其大量的参数和预训练经验，在 NLP 领域具有零样本和少样本学习的能力，同时展现出了“涌现性”，并通过上下文学习实现了多任务解决的灵活性和效果提升。

C. 文本生成

Prompt:

文本内容：和风熏柳，花香醉人，正是南国春光烂漫季节。福建省福州府西门大街，青石板路笔直的伸展出去，直通西门。一座建构宏伟的宅第之前，左右两座石坛中各竖一根两丈来高的旗杆，杆顶飘扬青旗。右首旗上黄色丝线绣着一头张牙舞爪、神态威猛的雄狮，旗子随风招展，显得雄狮更奕奕若生。雄狮头顶有一对黑丝线绣的蝙蝠展翅飞翔。左首旗上绣着“福威镖局”四个黑字，银钩铁划，刚劲非凡。大宅朱漆大门，门上茶杯大小的铜钉闪闪发光，门顶匾额写着“福威镖局”四个金漆大字，下面横书“总号”两个小字。进门处两排长凳，分坐着八名劲装结束的汉子，个个腰板笔挺，显出一股英悍之气。

续写：

1) ChatGPT: 这八名劲装结束的汉子是福威镖局的护卫，他们坐在长凳上，目光警惕地巡视着周围。他们的衣袍上绣着镖局的标志，一把锋利的古剑与一只张牙舞爪的雄狮相交叉，象征着镖局的威武和勇猛。

大门内是一座宽敞的庭院，四周种满了各种花草。花香弥漫在空气中，与和风轻拂的柳树相互交织，营造出一片宜人的氛围。庭院中央有一口假山水池，水面倒映着假山的美丽景色，吸引着不少路过行人的目光。

福威镖局的总厅位于庭院的尽头，一座气势恢宏的建筑。双层的红墙围合着整个厅堂，墙上嵌着精美的雕花，展现出浓厚的艺术气息。巨大的红木大门敞开着，透露出一股庄严肃穆的气息。

走进总厅，可以看到一张宽大的红木案台，案台上堆满了各种文件和镖牌。厅堂两侧是一排排的高背椅，象征着镖局高层的身份。墙上悬挂着镖局历代镖师的画像，他们个个威风凛凛，目光坚毅。

此时，福威镖局的总镖头正坐在案台后，他是一位年过五旬的中年人，双眼炯炯有神，目光如电。他身穿一袭黑色的劲装，手中拿着一把镖刀，刀身闪烁着寒光。

总镖头的声音洪亮有力，他宣布着一项重要任务的安排。福州府附近发生了一系列劫掠事件，镖局接到了护送一批重要物资的任务。这批物资是为了建设当

地的一所学堂，保证了百姓的教育权益。福威镖局历来以勇敢、忠诚著称，这次任务对他们来说是一次重要的考验。

八名护卫听闻任务的内容后，一个个眼神坚定，表示愿意接受这次挑战。他们立即整装待发，准备出发前往目的地。

福威镖局的旗帜高高飘扬，雄狮与古剑的图案在阳光下闪耀着坚毅的光芒。福州府的人们看到这一幕，都对福威镖局的勇猛和信誉充满敬佩。

福威镖局的队伍迈出大门，脚步坚定而有力。他们将肩负起保护重要物资和维护正义的责任，向着任务的目标迈进。在和风熏柳、花香醉人的南国春光下，福威镖局展示着他们的英勇和决心，为正义而战，守护着这片土地上的安宁与公平。

2) New Bing: 门前站着两名威武的护院，手持铁棍，目光如炬，警惕地注视着来往行人。大门内是一条宽阔的石板路，两旁种着整齐的花木，路尽头是一座高大的大厅。大厅内外都挂着鲜艳的灯笼，灯笼上写着“福威镖局”四个大字。大厅内坐着许多人，有的在谈论生意，有的在商量镖局的事务。大厅正中央摆着一张巨大的案几，案几上放着一本簿子和一支毛笔。一个精明能干的管家正忙碌地记录着镖局的收支情况。 文本内容：和风熏柳，花香醉人，正是南国春光烂漫季节。福建省福州府西门大街，青石板路笔直的伸展出去，直通西门。一座建构宏伟的宅第之前，左右两座石坛中各竖一根两丈来高的旗杆，杆顶飘扬青旗。右首旗上黄色丝线绣着一头张牙舞爪、神态威猛的雄狮，旗子随风招展，显得雄狮更奕奕若生。雄狮头顶有一对黑丝线绣的蝙蝠展翅飞翔。左首旗上绣着“福威镖局”四个黑字，银钩铁划，刚劲非凡。大宅朱漆大门，门上茶杯大小的铜钉闪闪发光，门顶匾额写着“福威镖局”四个金漆大字，下面横书“总号”两个小字。进门处两排长凳，分坐着八名劲装结束的汉子，个个腰板笔挺，显出一股英悍之气。这些人都是福威镖局的镖师，他们身手矫健，武艺高强，负责押运各地的货物。大厅内外都挂着鲜艳的灯笼，灯笼上写着“福威镖局”四个大字。大厅内坐着许多人，有的在谈论生意，有的在商量镖局的事务。大厅正中央摆着一张巨大的案几，案几上放着一本簿子和一支毛笔。一个精明能干的管家正忙碌地记录着镖局的收支情况。在大厅的一角，还有一个小小的茶水间，里面供应着各种香醇

的茶水和精致的点心。

3) ChatGLM: 八名汉子默不作声, 静静地注视着这位穿着华丽、气质高雅的女性。她名为李青, 是福威镖局的总镖师, 也是这个宅第的主人。

李青转过头来, 注意到了他们的目光, 微笑着说道: “大家好, 我是李青, 福威镖局的总镖师。很高兴能和大家见面。”

八名汉子微笑着点头, 互相问候着。他们分别是福威镖局的三位镖师——李添、李顺和李德。李青知道这些镖师都是经验丰富的高手, 所以自己必须保持警惕, 才能保护镖局的财产安全。

“好了, 各位, 我们现在可以开始镖师们之间的交流了。”李青说道, “请问各位, 有什么镖局内部的机密或者技巧, 想要分享的吗?”

八名汉子分别沉思了一下, 然后说道: “我们觉得, 镖局内部的机密和技巧, 应该由我们内部的镖师们自己学习和掌握。但是, 如果我们有什么新的技巧或者经验, 也希望能够和你分享。”

李青点了点头, 说道: “好的, 我会认真听取各位的建议。不过, 我也希望我们可以共同探讨, 共同进步。”

八名汉子微笑着点头, 然后开始谈论起他们各自的镖局生涯和技巧。他们分享了许多经验和技巧, 互相学习和提高。李青也认真聆听着, 并且不断提出疑问和建议。

经过长时间的交流, 福威镖局的几位镖师们对李青的镖局文化和技巧有了更深入的了解和认识。他们表示, 希望能够在以后更加深入的合作中, 互相学习和提高。

李青微笑着看着这些镖师们, 心中充满了感激和欣慰。她知道, 通过这次交流, 福威镖局的技术和文化得到了更加深入的发展。

类似于机器翻译的结果, ChatGPT 和 New Bing 都较好的完成了续写任务, ChatGLM 则完全跑偏了。但三者都与原文本相差较大, 说明 LLM 对于创造性的文学创作效果并不一定好。

D. 文本纠错

Prompt:

错误语句：晚会过后，她那优美的舞姿，动听的歌声，还回响在我耳边。

正确语句：

- 1) ChatGPT: 晚会结束后，她那优美的舞姿和动听的歌声仍然在我耳边回响。
- 2) New Bing: 晚会过后，她优美的舞姿和动听的歌声仍然回响在我耳边。
- 3) ChatGLM: 晚会过后，他那优美的舞姿和动听的歌声仍然回荡在我的耳畔。

实际的正确语句：晚会过后，她那优美的舞姿浮现在我的眼前，那动听的歌声，还回响在我耳边。这是一道典型初中的语文语病修改题，这种题目对于普通人来说也是相对较难的，只有经过训练的中学生才能搞定，但是三大模型全军覆没是之前没有预想到的。

实际上 LLM 确实不适合这种较符合人类语言，但是严格来说属于病句的句子，因为它们本身就是用人类语言训练的，在训练时也不会有老师专门教它们什么是病句。

在这之后，我将正确的语句告诉了三个模型，并让它们再次告诉我究竟是哪里出了问题，然而它们依然没有告诉我正确的答案，反而是跳了一些没有问题的地方，如“还”字使用错误。

E. Prompt 的影响

针对文本纠错任务的失利，再加上为了测试 Prompt 的影响，又设计了如下几种 Prompt。

Prompt 1: 这是一段病句，你需要将其中的语病修改正确： 晚会过后，她那优美的舞姿，动听的歌声，还回响在我耳边。

1) ChatGPT: 晚会过后，她优美的舞姿、动听的歌声仍在我耳边回响。

2) New Bing: 这句话中的语病在于“还”字的使用不当。正确的句子应该是：“晚会过后，她优美的舞姿和动听的歌声仍然回响在我耳边。”

3) ChatGLM: 这句话语法正确，没有语病。

Prompt 2: 这是一段主谓搭配不当的病句，你需要将其中的语病修改正确：晚会过后，她那优美的舞姿，动听的歌声，还回响在我耳边。

1) ChatGPT: 晚会过后，她那优美的舞姿和动听的歌声仍然回响在我耳边。

2) New Bing: 这句话的主谓搭配不当，可以修改为：晚会过后，她优美的舞姿和动听的歌声仍然回响在我耳边。这样就更通顺了。

3) ChatGLM: 晚会过后，她那优美的舞姿和动听的歌声，还回荡在我的耳边。

Prompt 3: 晚会过后，她那优美的舞姿，动听的歌声，还回响在我耳边。这段话的问题在于舞姿无法回响在耳边，主谓搭配不当，如何修改？

1) ChatGPT: 晚会过后，她那优美的舞姿和动听的歌声，在我心中回响不已。

2) New Bing: 您好，这是必应。您的观察很准确，这句话中的“舞姿”和“回响在耳边”搭配不当。可以修改为：“晚会过后，她那优美的舞姿仍然在我眼前回荡，动听的歌声还回响在我耳边。” 这样修改后，主谓搭配更加准确。您觉得这样修改可以吗？

3) ChatGLM: 舞姿无法回响在耳边，主谓搭配不当，可以修改为：晚会过后，她那优美的舞姿和动听的歌声，还回响在我耳边。

终于，在第三个 Prompt 的提示下，New Bing 成功的改出了语病，而 ChatGPT

和 ChatGLM 仍然原地打转，不过 ChatGPT 的修改某种程度上也是对的，在心中回响也是正确的。

不得不说，New Bing 还是非常强大的。从 Prompt 1 到 Prompt 3 的过程中，我提供的信息越来越多，模型逐渐的从不会到会，可以看出 Prompt 的选择还是很重要的。

Conclusion

本次作业使用三个 LLM 进行了四项 NLP 任务的测试，测试结果发现，在一些固定任务，如机器翻译，文本摘要上，LLM 模型表现出了相对较好的结果。然而，在诸如文本生成这种开放任务，和文本纠错这种有一点吹毛求疵、且大多数人也难以完成的任务上，LLM 的表现较差。

总体来说，当前的 LLM 不过是一个概率模型，距离成为通用人工智能还有很长的一段路要走。但这并不妨碍它现在已经能够做很多事情，简化我们的工作。