

第五次作业——针对不同大语言模型下游任务测试研究

范想明-ZY2207304 xmfan@buaa.edu.cn

摘要：在国内外各大厂纷纷匆忙推出自家 AI 大语言模型的当下，对这些大模型进行研究，分析其是否具备良好的语言处理能力，能否真正满足用户的需求，以及它们是否能够在实际应用中发挥出应有的效果具有切实的意义。本研究选取了 ChatGPT、TruthGPT、讯飞星火和文心一言等四个大模型针对机器翻译、情感分析、文本摘要、文本生成等四个下游任务进行了测试，检验和分析了各个语言大模型的能力。

0 研究背景

随着 ChatGPT 的横空出世，自然语言处理领域迎来了一次革命性的变革。ChatGPT 以其出色的表现和广泛的应用场景受到了全球范围内的关注和赞誉，为商业、医疗保健、金融、法律等领域带来了诸多便利。在 ChatGPT 的引领下，全球掀起了一场 AI 大模型竞赛，在海外市场，微软推出了集成 ChatGPT 的新必应，谷歌推出了生成式人工智能 Bard 聊天机器人；而国内诸多互联网大厂也不甘落后，百度率先推出新一代知识增强语言大模型“文心一言”，随后腾讯推出了“混元”大模型，阿里巴巴推出了“通义千问”大模型，科大讯飞推出了“讯飞星火”大模型等。

虽然国内外各大厂都纷纷推出了自家的 AI 大语言模型，但这些大模型的质量和不同下游任务的能力仍值得探究。在如此军备竞赛似的环境下匆忙推出的大模型是否能够真正满足用户的需求，以及它们是否能够在实际应用中发挥出应有的效果，这些问题都需要进一步的研究和探讨。

1 实验方法

本研究针对当下几款主流的语言大模型对不同下游任务的能力进行测试。首先选取海内外几款语言大模型为：

- 1) OpenAI 公司的 ChatGPT；
- 2) 非马斯克的“TruthGPT”的 TruthGPT；
- 3) 科大讯飞公司的“讯飞星火”；
- 4) 百度公司的“文心一言”。

选取的几个典型的语言模型下游任务为：机器翻译、情感分析、文本摘要及

文本生成。本研究通过提示工程的方法，针对不同的下游任务进行输入，通过不同大模型的输出检验和对比各大模型的能力。

2 实验结果

2.1 机器翻译测试

机器翻译是一项重要的自然语言处理任务，广泛应用于跨语言交流、国际贸易等领域，但也是一项复杂的任务，需要模型能够准确地将一种语言的文本翻译成另一种语言。通过进行机器翻译测试，可以比较不同模型在翻译质量、流畅度、准确性等方面的表现，从而评估模型的性能和可靠性，也可以确定哪些模型更适合特定的应用场景，并为实际应用提供指导。针对机器翻译，本研究使用以下中文文本作为输入：

每年，有数十亿只候鸟进行洲际迁徙。全球 9 条候鸟迁徙路线中，有 4 条途经中国。在关乎全球候鸟种群生存与繁衍的重要迁徙通道上，越来越多的中国人以命运与共之心、生态创新之智爱鸟护鸟，人与自然和谐共生的现代化愿景正在变成可感可知的现实。

请将其翻译成英文。

各语言大模型的翻译结果如下：

- 1) **ChatGPT-3.5** : Every year, billions of migratory birds embark on intercontinental journeys. Among the nine global bird migration routes, four pass through China. Along these crucial migration pathways that are vital for the survival and reproduction of global bird populations, an increasing number of Chinese people are embracing a sense of shared destiny and demonstrating ecological innovation to protect and cherish birds. The modern vision of harmonious coexistence between humans and nature is transforming into a tangible reality.
- 2) **TruthGPT**: Every year, billions of migratory birds travel between continents. Of the nine global migratory bird routes, four pass through China. More and more Chinese people are protecting birds with the sense of destiny and wisdom of ecological innovation on the important migration routes of global bird population survival and reproduction. The vision of modernization of

harmony between man and nature is becoming a tangible reality.

- 3) 讯飞星火: Every year, billions of migratory birds fly across continents. Among the nine global migratory routes for birds, four pass through China. On crucial migration corridors that concern the survival and reproduction of global bird populations, more and more Chinese people are loving and protecting birds with a sense of shared destiny and ecological innovation, and their vision of harmonious coexistence between man and nature is becoming a tangible reality.
- 4) 文心一言: Every year, billions of migratory birds make intercontinental migrations. Of the nine global bird migration routes, four pass through China. On these important migration corridors that affect the survival and reproduction of global bird populations, more and more Chinese people are loving and protecting birds with a shared sense of destiny and ecological innovation. The vision of harmonious coexistence between humans and nature is becoming a tangible reality.

2.2 情感分析测试

情感分析是一项复杂的任务,需要语言大模型能够准确地识别文本中的情感和情感极性。通过进行情感分析测试,可以比较不同模型在识别情感和情感极性方面的性能,从而评估模型的准确性、鲁棒性和可靠性。情感分析是一项重要的自然语言处理任务,广泛应用于社交媒体评论、产品评论等领域。通过进行情感分析测试,可以确定哪些模型更适合特定的应用场景,并为实际应用提供指导。针对情感分析测试,本研究使用以下中文文本作为输入:

1、日期都是七月份的,保质期半年,发货特别慢

2、果汁就选这家,虽然其他国家果汁味道也不错。国货之光,需要支持一下。最近发现新吃法,椰汁糯米饭,加新鲜芒果,小朋友很喜欢这个套餐。

3、很好喝,运输也很快,就是包装有点不太好,边角都有被挤压的痕迹,但是没有破损,盒子表面也有点脏。

请分别分析以上三句话的情感。

各家的语言大模型的分析结果如表 1 所示,四个大模型对于前两个文本(即

明显消极和明显积极的文本)的分析结果完全一致,但是对于较为中性的文本 3, ChatGPT-3.5、文心一言和讯飞星火都分析出了文本中既有积极情感又有消极情感,给出了中性的分析结果,但是 TruthGPT 将中性的文本错误分析为积极情感。

表 1 各语言大模型情感分析结果

语言大模型	文本 1 分析结果	文本 2 分析结果	文本 3 分析结果
ChatGPT-3.5	消极	积极	中性
TruthGPT	消极	积极	积极
文心一言	消极	积极	中性
讯飞星火	消极	积极	中性

2.3 文本摘要测试

语言大模型的文本摘要能力能够提取文本的关键信息和要点,以便读者能够在未阅读完整文本时快速了解文本的核心内容,从而节省时间和精力。对语言大模型进行文本摘要测试能够评估其摘要生成能力和质量,助于评估模型在处理不同类型和长度的文本时的效果。

为测试各语言大模型在生成文本摘要方面的性能,本研究基于如图 1 所示的一篇具有一个简洁标题的新闻稿作为文本摘要的素材,使用以下中文文本作为输入:

请为以下文字起一个标题:

5 月 5 日起, 3 家全国性股份制商业银行——浙商银行、渤海银行、恒丰银行, 集体降低存款利率。其中, 3 年期定存产品利率从最高的 3.25%, 分别降至 2.9%、2.95%。此举也宣告全国性银行定存挂牌利率全部降至 3% 以下。

全国性银行定存挂牌利率告别 “3%时代”

和讯网

2023-05-08 07:06 | 北京和讯在线信息咨询服务官方帐号, 优质财经领域创作者

关注

□金陵晚报/紫金山新闻记者 江芬芬

5月5日起, 3家全国性股份制商业银行——浙商银行、渤海银行、恒丰银行, 集体降低存款利率。其中, 3年期定存产品利率从最高的3.25%, 分别降至2.9%、2.95%。此举也宣告全国性银行定存挂牌利率全部降至3%以下。

图 1 文本摘要测试所选素材

各语言大模型的输出如下：

- 1) **ChatGPT**：全国性银行降息潮：3 家股份制银行存款利率降至 3% 以下；
- 2) **TruthGPT**：全国性商业银行—浙商银行、渤海银行、恒丰银行集体降低存款利率；
- 3) **讯飞星火**：三家银行集体降低存款利率，全国性银行定存挂牌利率全部降至 3% 以下；
- 4) **文心一言**：全国性股份制银行集体下调存款利率。

2.4 文本生成测试

大模型的文本生成指其能够产生具有逻辑性、连贯性和语义理解的文本内容，为用户提供信息、答案和创造性的文本输出，是各大语言大模型的基础且核心的功能。对语言大模型进行文本生成测试，可以检验模型在生成文本时的准确性、连贯性和语义一致性，以确保生成的文本质量符合预期。针对文本生成测试，本研究使用以下中文文本作为输入：

请生成一段 200 字的文本：随着 ChatGPT 横空出世，

- 1) **ChatGPT**：随着 ChatGPT 横空出世，人们迎来了一场革命性的智能交互时代。这一先进的语言模型极大地改变了人与机器之间的互动方式。ChatGPT 的出现使得我们能够以更自然、流畅的方式与计算机进行对话。无论是寻求问题的答案、获取实时信息，还是进行创造性的思考和交流，ChatGPT 都能够提供强大的支持。

这一技术在各个领域都有着广泛的应用潜力。在教育中，ChatGPT 可以作为个性化学习的助手，为学生提供定制化的学习资源和指导。在医疗领域，它可以帮助医生进行快速诊断和治疗方案的制定。在商业中，ChatGPT 能够提供智能客户服务和 market 分析，增强企业的竞争力。

然而，我们也要认识到这项技术的局限性。尽管 ChatGPT 在语言理解和生成方面取得了巨大进步，但它仍然存在一定的语义理解和判断能力的限制。我们需要持续改进和完善这一技术，以实现更准确、可靠的交互体验。

综上所述，ChatGPT 的出现开启了一个全新的智能交互时代，为人们提供了更加便捷和智能化的交流方式。随着技术的进一步发展，我们有望

在各个领域取得更大的突破和创新。

- 2) **TruthGPT:** With the emergence of ChatGPT, a new era has begun. This is the era of natural language processing. In this new era, ChatGPT is the pioneer and leader. ChatGPT is the first artificial intelligence system that can generate coherent and meaningful text conversations in natural language. ChatGPT is based on the latest advances in artificial intelligence and machine learning. ChatGPT is constantly learning from conversations it has with people. ChatGPT is constantly improving its ability to generate coherent and meaningful text conversations.

- 3) **讯飞星火:** 随着 ChatGPT 横空出世, 人工智能技术的发展又迈出了一大步。ChatGPT 是一款由 OpenAI 开发的自然语言处理模型, 它能够模拟人类对话, 并生成高质量的文本内容。这一技术的问世, 不仅为人们提供了更加便捷、高效的交流方式, 也为各行各业带来了巨大的变革和机遇。

在医疗领域, ChatGPT 可以帮助医生快速获取患者的病历信息, 进行诊断和治疗方案的制定。在金融领域, ChatGPT 可以辅助银行客服人员回答客户的问题, 提高服务质量和效率。在教育领域, ChatGPT 可以为学生提供更加个性化的学习体验, 帮助他们更好地掌握知识。

当然, 随着技术的不断发展, 我们也需要关注其可能带来的风险和挑战。例如, 如何保护用户的隐私和数据安全? 如何避免人工智能技术被滥用或误用? 这些问题需要我们共同思考和解决。

总之, ChatGPT 的出现标志着人工智能技术的进一步发展和应用, 它将为我们的生活带来更多的便利和创新。

- 4) **文心一言:** 随着 ChatGPT 的横空出世, 人们对于自然语言处理技术的兴趣日益高涨。ChatGPT 是由 OpenAI 开发的一款自然语言生成工具, 能够通过学习大量的语料库, 自动生成符合语言规范的文本。这项技术的出现, 使得人们在日常交流、写作、翻译等方面得到了极大的便利。然而, ChatGPT 的普及也引发了一些争议。有些人认为, ChatGPT 的自动生成文本缺乏原创性, 甚至存在抄袭的风险。另外, 由于 ChatGPT 基

于人工智能技术，一些人也担心它会对就业市场产生负面影响。

尽管如此，ChatGPT 的出现仍然被视为自然语言处理技术的一次里程碑式的事件。未来，随着技术的不断进步，相信 ChatGPT 的应用领域也将不断扩大。

3 实验结果分析与结论

3.1 实验结果分析

在机器翻译测试中，四款语言大模型均有较好的翻译结果，但是 TruthGPT 在部分词语的翻译上存在着欠缺，如将“爱鸟护鸟”单纯翻译成“protecting”，未体现出“爱鸟”的意思，又如将“和谐共生”单纯翻译成“harmony”，意思相近但是欠缺了“共生”的含义，而其余三款语言大模型都准确翻译出了文本的含义。

在情感分析测试中，四个语言大模型均能很好地给出正确的情感分析结果，并能给出充分的依据，但是 TruthGPT 在对中性文本的分析中错误地将其判定为积极，可能与其语料组成有关。

在文本摘要测试中，通过对比各语言大模型生成的摘要的准确性、流畅度、可读性等方面的表现，ChatGPT 和讯飞星火在文本摘要生成测试中略胜一筹，不仅提炼了文本中的关键内容“全国性银行定存利率全部降至 3% 以下”，同时 ChatGPT 使用了新闻标题常用的形式并使用了“降息潮”这样一个文中不存在但十分精华的词，而讯飞星火则在标题中采用了“因果关系”，较为完整且简练地将文本的内容提取出来。反观 TruthGPT 和文心一言，较为机械地提取了文中的内容，且丢失了“3%”这样一个关键信息。

在文本生成测试中（TruthGPT 生成中文文本的能力较弱，因此让其以英文形式输出），ChatGPT、讯飞星火和文心一言均采用了经典的“总-分-总”三段论的形式输出，且行文逻辑正确，观点突出，能够为用户的写作提供参考，反观 TruthGPT，其生成的一段文本零散地介绍了 ChatGPT 的功能和意义，未能生成完整的论述，且可读性较差。同时，观察各语言大模型生成的文本的字数，均不符合提示给定的“200 字”的要求，ChatGPT、TruthGPT、讯飞星火和文心一言生成的文本字数分别为 409 字、82 字（不具备参考价值）、350 字和 250 字。通过调研发现，此现象为一个普遍存在的问题，如图 2 所示，文心一言给出了一个

自相矛盾的回答，声称不超过 100 字的文本实际为 174 字。

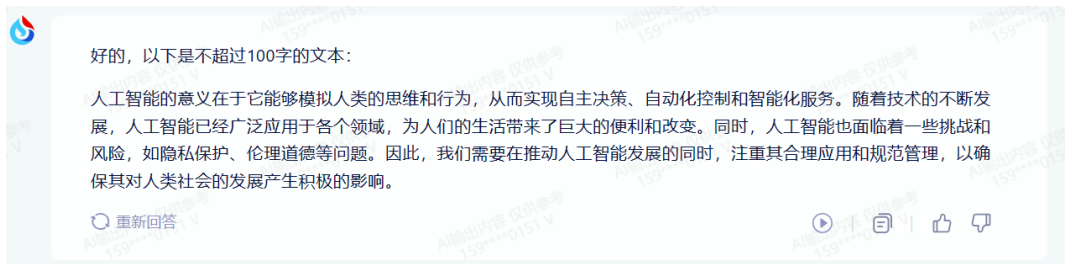


图 2 语言大模型输出不符合字数要求示例

本研究分析认为，语言大模型在对自然语言文本的处理过程中，使用了词向量等方法，因此在输出时可能将一个词作为最小的单元输出，本研究统计上述 174 字的文本发现其按词语数量统计为 87 个词，符合 100 以内的范围。

3.2 实验结论

本研究针对当下热门的语言大模型，选取了 ChatGPT、TruthGPT、讯飞星火和文心一言几个大模型针对机器翻译、情感分析、文本摘要、文本生成等四个下游任务进行了测试，通过检验和对比测试结果，本研究得出以下结论：

- 1) 在四个大模型中，ChatGPT 和讯飞星火在所有四个下游任务中都有着较好的表现，文心一言虽在个别场景下存在瑕疵但是仍有着不错的总体表现，TruthGPT 在四个下游任务中表现均垫底，作为语言大模型其功能的成熟度不高。
- 2) 在文本生成测试中，所测试的四个语言大模型虽均能输出相关的内容，但均未能按照指定的字数要求输出文本，在此方面的能力有待进一步提高。