计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202330801
61(5): 1128−1142, 2024

GPT 系列大语言模型在自然语言处理任务中的鲁棒性

陈炫婷[1] 叶俊杰[1] 祖 璨[1] 许 诺[1] 桂 韬[2] 张 奇[1]
[1]（复旦大学计算机科学技术学院 上海 200433）
[2]（复旦大学现代语言学研究院 上海 200433）
（xuantingchen21@m.fudan.edu.cn）

Robustness of GPT Large Language Models on Natural Language Processing Tasks

Chen Xuanting[1], Ye Junjie[1], Zu Can[1], Xu Nuo[1], Gui Tao[2], and Zhang Qi[1]
[1]（*School of Computer Science*, *Fudan University*, *Shanghai* 200433）
[2]（*Institute of Modern Languages and Linguistics*, *Fudan University*, *Shanghai* 200433）

Abstract The GPT models have demonstrated impressive performance in various natural language processing (NLP) tasks. However, their robustness and abilities to handle various complexities of the open world have not yet to be well explored, which is especially crucial in assessing the stability of models and is a key aspect of trustworthy AI. In this study, we perform a comprehensive experimental analysis of GPT-3 and GPT-3.5 series models, exploring their performance and robustness using 15 datasets (about 147 000 original test samples) with 61 robust probing transformations from TextFlint covering 9 popular NLP tasks. Additionally, we analyze the model's robustness across different transformation levels, including character, word, and sentence. Our findings reveal that while GPT models exhibit competitive performance in certain tasks like sentiment analysis, semantic matching, and reading comprehension, they exhibit severe confusion regarding information extraction tasks. For instance, GPT models exhibit severe confusion in relation extraction and even exhibit "hallucination" phenomena. Moreover, they experience significant degradation in robustness in terms of tasks and transformations, especially in classification tasks and sentence-level transformations. Furthermore, we validate the impact of the quantity and the form of demonstrations on performance and robustness. These findings reveal that GPT models are still not fully proficient in handling common NLP tasks, and highlight the difficulty in addressing robustness challenges through enhancing model performance or altering prompt content. By comparing the performance and robustness of the updated version of gpt-3.5-turbo, gpt-4, LLaMA2-7B and LLaMA2-13B, we further validate the experimental findings. Future studies on large language models should strive to enhance their capacities in information extraction and semantic understanding, while simultaneously bolstering overall robustness.

Key words robustness; GPT models; large language models; natural language processing; reliability

摘 要 大语言模型（large language models, LLMs）所展现的处理各种自然语言处理（natural language processing, NLP）任务的能力引发了广泛关注. 然而, 它们在处理现实中各种复杂场景时的鲁棒性尚未得到充分探索, 这对于评估模型的稳定性和可靠性尤为重要. 因此, 使用涵盖了 9 个常见 NLP 任务的 15 个数据集（约 147 000 个原始测试样本）和来自 TextFlint 的 61 种鲁棒的文本变形方法分析 GPT-3 和 GPT-3.5 系列模型在原始数据集上的性能, 以及其在不同任务和文本变形级别（字符、词和句子）上的鲁棒性. 研究结果表明, GPT 模型虽然在情感分析、语义匹配等分类任务和阅读理解任务中表现出良好的性能, 但

收稿日期: 2023-10-10; **修回日期**: 2024-03-12
通信作者: 桂韬（tgui@fudan.edu.cn）

其处理信息抽取任务的能力仍较为欠缺，比如其对关系抽取任务中各种关系类型存在严重混淆，甚至出现“幻觉”现象. 在鲁棒性评估实验中，GPT 模型在任务层面和变形层面的鲁棒性都较弱，其中，在分类任务和句子级别的变形中鲁棒性缺乏更为显著. 此外，探究了模型迭代过程中性能和鲁棒性的变化，以及提示中的演示数量和演示内容对模型性能和鲁棒性的影响. 结果表明，随着模型的迭代以及上下文学习的加入，模型的性能稳步提升，但是鲁棒性依然亟待提升. 这些发现从任务类型、变形种类、提示内容等方面揭示了 GPT 模型还无法完全胜任常见的 NLP 任务，并且模型存在的鲁棒性问题难以通过提升模型性能或改变提示内容等方式解决. 通过对 gpt-3.5-turbo 的更新版本、gpt-4 模型，以及开源模型 LLaMA2-7B 和 LLaMA2-13B 的性能和鲁棒性表现进行对比，进一步验证了实验结论. 鉴于此，未来的大模型研究应当提升模型在信息提取以及语义理解等方面的能力，并且应当在模型训练或微调阶段考虑提升其鲁棒性.

关键词 鲁棒性；GPT 模型；大语言模型；自然语言处理；可靠性

中图法分类号 TP391

大语言模型，如 FLAN[1]，GPT-3[2]，LLaMA[3] 和 PaLM2[4] 等，在对话、理解和推理方面展示了惊人的能力[5]. 在不修改模型参数的情况下，大模型可以仅通过输入合适的提示来执行各种任务. 其中，GPT 系列模型因其出色的能力备受关注.

为定量评估和探究大模型的能力，已有的工作集中于评估大模型在常识和逻辑推理[6]、多语言和多模态[7]、心智理论[8] 和数学[9] 等方面的能力. 尽管这些工作在基准测试集上取得了很好的效果，但大模型是否具备良好的鲁棒性仍然需要进一步研究.

鲁棒性衡量了模型在面对异常情况（如噪音、扰动或故意攻击）时的稳定性，这种能力在现实场景，尤其是在自动驾驶和医学诊断等安全场景下对于大模型至关重要. 鉴于此，现有工作对大模型的鲁棒性展开了探究：Wang 等人[10] 从对抗性和分布外（out of distribution, OOD）的角度出发，使用现有的 AdvGLUE[11] 和 ANLI[12] 对抗基准评估 ChatGPT 等大模型的对抗鲁棒性，使用 DDXPlus[13] 医学诊断数据集等评估分布外鲁棒性；Zhu 等人[14] 则从提示的角度出发，提出了基于对抗性提示的鲁棒性评测基准，并对大模型在对抗提示方面的鲁棒性进行了分析. 然而，已有的研究主要使用对抗攻击策略，这对于大规模评估来说需要消耗大量的算力和时间；并且对抗样本生成的目标是通过对特定模型或数据集的原始输入进行微小的扰动，以误导模型的分类或生成结果，但这些扰动并不总是代表真实世界中的威胁和攻击方式. 此外，现有研究大多针对 ChatGPT 及同时期的其他大模型，对 GPT 系列模型迭代过程中性能和鲁棒性的变化关注较少.

鉴于此，本文选择了图 1 所示的 5 个 GPT-3 和 GPT-3.5 系列模型作为大模型的代表，通过全面的实验分析其性能和鲁棒性，以解决 3 个问题.

问题 1：GPT 模型在自然语言处理 (NLP) 任务的原始数据集上有何性能缺陷？

为给后续的鲁棒性评估提供基础和参考点，本文首先评估模型在原始数据集上的性能. 本文选择 15 个数据集（超过 147 000 个原始测试样本），涵盖了 9 个常见的 NLP 任务，如情感分析、阅读理解和命名实体识别等，评估了 GPT 模型在原始数据集上的性能以及迭代过程中的性能变化. 虽然这些任务没有直接对应具体的对话场景，但它们评估了模型的潜在能力，包括理解上下文、处理不同的语言结构和捕

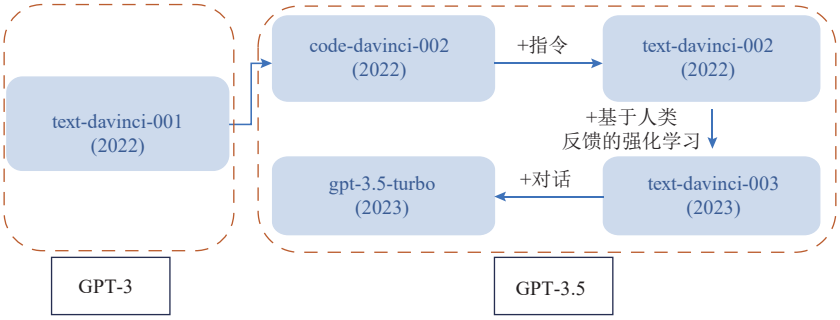


Fig. 1 The evolution of five GPT-3 and GPT-3.5 series models

图 1 5 个 GPT-3 和 GPT-3.5 系列模型的迭代过程

捉微小的信息等，这些能力对于语言理解和生成系统都非常重要.

问题2：GPT模型在NLP任务上面对输入文本扰动时的鲁棒性如何？

本文首先确定评估鲁棒性的方法.为更加真实地模拟现实世界中可能存在的噪音、扰动和攻击，本文选择了TextFlint[15]作为对输入文本进行扰动的工具.TextFlint提供了许多针对NLP任务特定的文本变形，这些变形均基于语言学进行设计，体现了实际使用语言过程中可能发生的情况，保持了变形后文本的语言合理性，能够模拟实际应用中的挑战.本文使用了61种文本变形方法，这些变形按照粒度可以分为句子级、词级和字符级.本文通过实验分析了GPT模型在各种任务和各个变形级别上的鲁棒性，并探究了模型迭代过程中鲁棒性的变化.

问题3：提示对GPT模型的性能和鲁棒性有何影响？

在上述2个问题中，本文从测试文本出发，通过将不同的测试样本与任务特定的提示进行拼接，评估了模型的性能和鲁棒性.在这个问题中，本文从提示的角度出发，研究其对性能和鲁棒性的影响.上下文学习[16](in-context learning, ICL)已经成为NLP领域的新范式，语言模型可以仅基于少量示例执行复杂任务.基于此，本文通过改变提示中演示(demonstration)的数量或内容，探究提示对GPT模型的性能和鲁棒性的影响.

本文的定量结果和定性分析表明：

1)GPT模型在情感分析、语义匹配等分类任务和阅读理解任务中表现出较优异的性能，但在信息抽取任务中性能较差.例如，其严重混淆了关系抽取任务中的各种关系类型，甚至出现了“幻觉”现象.

2)在处理被扰动的输入文本时，GPT模型的鲁棒性较弱，它们在分类任务和句子级别变形中鲁棒性缺乏更为显著.

3)随着GPT系列模型的迭代，其在NLP任务上的性能稳步提升，但是鲁棒性并未增强.除情感分析任务外，模型在其余任务上的鲁棒性均未明显提升，甚至出现显著波动.

4)随着提示中演示数量的增加，GPT模型的性能提升，但模型鲁棒性仍然亟待增强；演示内容的改变可以一定程度上增强模型的抗扰动能力，但未能从根本上解决鲁棒性问题.

同时，通过对gpt-3.5-turbo的更新版本、gpt-4、开源模型LLaMA2-7B和LLaMA2-13B的表现进行评估，本文进一步验证了上述实验结论的普适性和可持续性.

1 相关工作

1.1 大模型的性能评测

近期有大量的研究集中于评估大模型在各种任务中的性能.Qin等人[6]对ChatGPT和text-davinci-003等模型在常见NLP任务上的零样本能力进行了评测，结果表明ChatGPT擅长处理推理和对话任务，但是在序列标注任务上表现欠佳；Bang等人[7]评估了ChatGPT在多任务、多语言和多模态方面的能力，发现ChatGPT在大多数任务上优于零样本学习的大模型，甚至在某些任务上优于微调模型；Zhuo等人[17]针对大模型伦理进行了评测工作.此外，大量工作针对大模型在不同领域的能力进行了研究和讨论，包括法律领域[18]、教育领域[19-20]、人机交互领域[21]、医学领域[22]以及写作领域[23]等.然而，这些研究主要集中在大模型的性能上，对鲁棒性的关注有限.模型在固定的测试数据上取得较高准确率，并不能反映出其在现实场景中面对输入的文本噪音、扰动或恶意攻击时的可靠性和稳定性，因此，鲁棒性对于评估模型处理现实世界中的复杂任务的能力至关重要.

1.2 大模型的鲁棒性评测

已有的关于大模型鲁棒性的工作主要集中于2个方面：对抗鲁棒性和分布外鲁棒性.对抗鲁棒性是指模型在对抗样本上的鲁棒性表现，对抗样本[24]的生成方式为：对原始输入施加一个阈值范围内的微小扰动，使得模型的分类或生成结果发生变化.分布外鲁棒性关注于模型的泛化性，即使用与模型训练数据存在分布偏移的数据(包括跨域或跨时间数据)进行鲁棒性评测.Wang等人[10]使用现有的AdvGLUE[11]和ANLI[12]对抗基准评估ChatGPT等大模型的对抗性鲁棒性，使用Flipkart评论和DDXPlus[13]医学诊断数据集评估分布外鲁棒性.结果表明，尽管ChatGPT在大多的分类任务和翻译任务上展现出更优的鲁棒性，但是大模型的对抗性和分布外鲁棒性仍然较弱.Zhu等人[14]针对提示进行对抗攻击，并使用这些对抗性提示对大模型进行鲁棒性测试，结果表明大模型容易受到对抗性提示的影响.然而，对抗样本的数据是以欺骗模型为目的而生成的，与现实场景中产生的噪音和扰动存在明显差异，并且生成对抗样本需要消耗大量算力和时间，不适合进行大规模评测.本文通过考虑更广泛的使用场景，从输入文本的角度出发，利用任务特定的文本变形来评估大模型在每

个任务中的鲁棒性表现，从而进行更全面的分析. 此外，本文关注于 GPT 系列的多个模型的表现，分析了它们在迭代过程中性能和鲁棒性方面的变化.

2 数据集和模型

2.1 数据集

为了全面评估 GPT 模型在各类 NLP 任务上的表现，本文选取了 9 个常见的 NLP 任务，涵盖分类、阅读理解和信息抽取 3 个不同类别，如表 1 所示. 针对每个任务，本文选取了具有代表性的公开数据集进行测试，最终共包含 15 个不同数据集.

2.2 GPT 系列模型

根据图 1 所示，本文主要针对 5 个 GPT-3 和 GPT-3.5 系列模型进行评估和分析，并对 GPT-4 模型在零样本场景下进行抽样测试，所有模型都通过 OpenAI 官方 API[①]进行评估. 根据 OpenAI 官方文档的说明，text-davinci-002 是基于 code-davinci-002 的 InstructGPT[37] 模型，其使用了一种监督式微调策略的方法 FeedME[②]进行训练；text-davinci-003 是 text-davinci-002 的改进版本，其使用近端优化策略（proximal policy optimization, PPO）算法进行训练，该算法被用于基于人类反馈的强化学习[38](reinforcement learning from human feedback, RLHF)；gpt-3.5-turbo 是针对聊天场景进行优化的最强大的 GPT-3.5 模型（本文第 3~5 节所使用的版本均为 gpt-3.5-turbo-0301 版本）.

3 性能评测

性能评测对于评估模型的能力，以及对后续的鲁棒性评估建立基准和参考至关重要. 本节对 GPT 系列模型在 NLP 任务中原始数据集上的性能表现进行了全面的评测，旨在评估它们在不同 NLP 任务中的表现，并分析它们有何缺陷. 同时，本节还探究了 GPT 系列模型在迭代过程中的性能变化.

3.1 方 法

大模型可以通过输入适当的提示或指令来执行各种任务，而无需修改任何参数. 为评估 GPT 模型在 NLP 任务中的性能，本文针对每个具体任务设计了 3 种不同的提示. 如图 2 所示，本文将提示与测试文本拼接起来作为测试样本输入模型，并获得相应的输出，通过对输出结果的定量评估来评测模型的性能.

3.2 实验设定

为定量分析模型的性能，本文使用准确率（accuracy）和 $F1$ 分数（$F1$ score）作为评估指标. 各个数据集对应的评估指标如表 1 所示.

Table 1 Information of 15 Datasets Used in Experiments

表 1 实验使用的 15 个数据集的信息

任务类型	子任务类型	数据集	数据量	评测指标
分类	细粒度情感分析（ABSA）	SemEval2014-Laptop[25]	331	准确率
		SemEval2014-Restaurant[25]	492	准确率
	情感分析（SA）	IMDB[26]	25 000	准确率
	自然语言推理（NLI）	MNLI-m[27]	9 815	准确率
		MNLI-mm[27]	9 832	准确率
		SNLI[27]	10 000	准确率
	语义匹配（SM）	QQP[28]	40 430	准确率
		MRPC[29]	1 725	准确率
	威诺格拉德模式挑战（WSC）	WSC273[30]	570	准确率
阅读理解	机器阅读理解（MRC）	SQuAD 1.1[31]	9 868	$F1$
		SQuAD 2.0[32]	11 491	$F1$
信息抽取	词性标注（POS）	WSJ[33]	5 461	准确率
	命名实体识别（NER）	CoNLL2003[34]	3 453	$F1$
		OntoNotesv5[35]	4 019	$F1$
	关系抽取（RE）	TACRED[36]	15 509	$F1$

① https://platform.openai.com

② https://platform.openai.com/docs

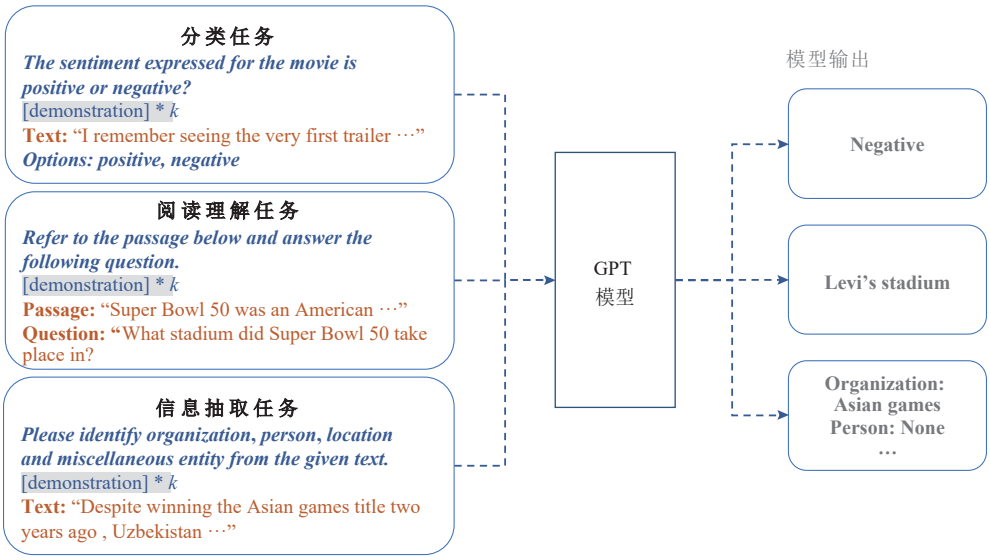


Fig. 2 Overview of experimental evaluating process

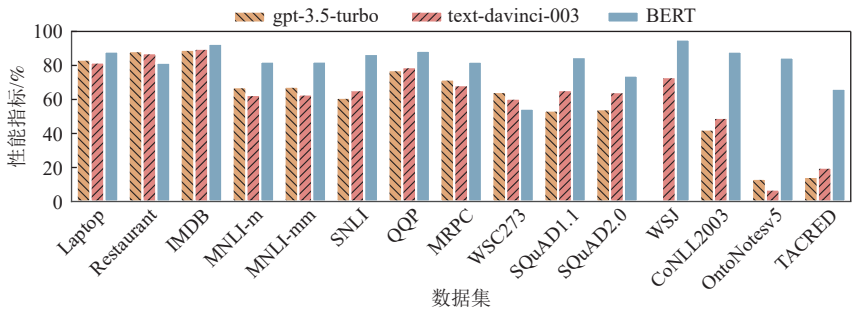
图 2 实验评测流程图

由于本文实验涉及不同模型、数据集、变形类型、提示种类等多个维度，为方便后续从不同维度对结果进行统计、计算和比较，实验选取的基准模型应当在 NLP 研究中具有强大的性能和广泛应用，从而能够适用于本文所有评测数据集. 因此，本文选择 BERT[39] 作为所有数据集的统一基准模型. 对于每个数据集，本文使用在相应数据集上经过有监督微调的 BERT 模型. 具体而言，对于 IMDB 数据集和 WSJ 数据集，本文使用的 BERT 版本分别是 BERT-Large-ITPT 和 BERT-BiLSTM-CRF. 在其他数据集中，本文均使用 BERT-base-uncased 作为基准模型. 此外，本节中 GPT 模型的测试结果均为零样本场景下的结果.

3.3 结果分析

首先分析 2 个最新的 GPT-3.5 模型(即 gpt-3.5-turbo 和 text-davinci-003 模型)的性能表现，其和 BERT 在 15 个数据集上的性能表现如图 3 所示，图中的数据是每个数据集在 3 个提示下的性能均值. 图 3 所示的结果表明，GPT 模型的零样本性能在情感分析、语义匹配、机器阅读理解等分类任务和阅读理解任务中可以与 BERT 相媲美，并且在 SemEval2014-Restaurant 和 WSC273 数据集上的表现均优于 BERT.

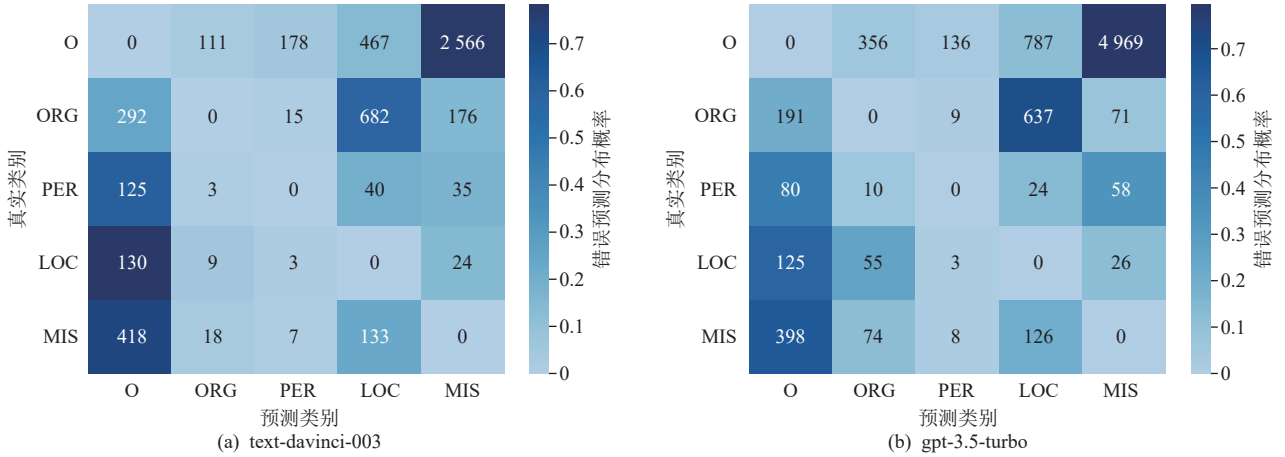
然而，GPT 模型在命名实体识别(NER)和关系抽取(RE)任务上表现不佳. 为深入了解模型错误预测背后的原因，本文选择 CoNLL2003 和 TACRED 数据集作为代表，分析了错误预测的分布情况. 图 4 的 2 个分图的第 1 列表示在 CONLL2003 数据集的预测结果中，实体类型被错误预测为"非实体"类型(即"O")的数量. 结果表明，在 NER 任务中，大多数错误预测来自于"O"标签与特定实体类型的混淆，这表明大模型对实体词缺乏敏感性；在 RE 任务中，如图 5 的 2 个分图的第 1 行所示，GPT 模型倾向于将"无关



注："Laptop"和"Restaurant"分别表示"SemEval2014-Laptop"和"SemEval2014-Restaurant"数据集.

Fig. 3 Performance of GPT-3.5 models and BERT

图 3 GPT-3.5 模型和 BERT 的性能表现



(a) text-davinci-003 (b) gpt-3.5-turbo

Fig. 4 Distribution of prediction errors in CoNLL2003 dateset

图 4 在 CoNLL2003 数据集上的错误预测的分布

系”实例(即“N/A”)错误分类为特定的关系类型.

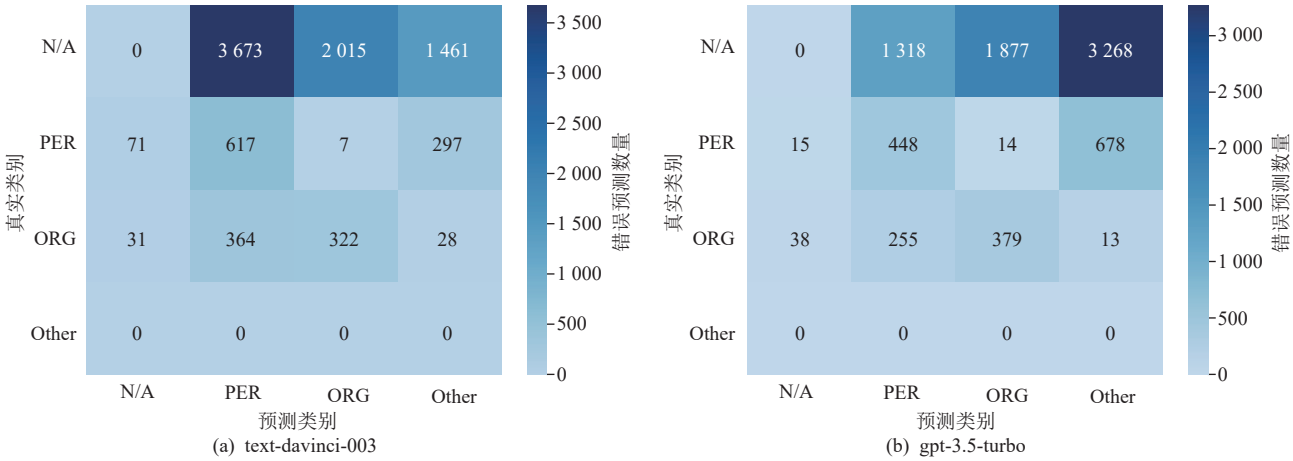
需要注意的是,我们观察到在 RE 任务中模型存在“幻觉”现象,即模型生成了在给定文本和预定义标签空间中不存在的虚构关系. 如图 5 所示,“N/A”表示“无关系”,“PER”和“ORG”分别表示属于“人物”和“组织”关系类别中的关系类型集合,而“Other”表示不属于任何预定义标签的关系集合. 如图 5 的最后 1 列所示,GPT 模型在生成结果中会虚构大量的“Other”关系,而非基于提示中给出的任务特定的关系类型和语义信息. 同时,本文在 IMDB 二分类数据集中也观察到类似的现象,模型将许多句子分类为“中性”标签,而该标签并不属于提示中给定的标签空间.

如图 6 所示,本文按照 OpenAI 官方发布模型的时间顺序和迭代关系(图 1),评测了 GPT-3 和 GPT-3.5 系列模型在迭代过程中性能的变化. 由于测试数据较多,本文按照表 1 所示的子任务类型进行结果展示,每个子任务的数值为其包含数据集的结果的均值. 结果表明,随着模型发布时间的推移,GPT 模型在大多数 NLP 任务上的性能稳步提升. 其中,GPT 模型在情感分析(SA)和细粒度情感分析(ABSA)任务上保持了较高的性能,并在自然语言推理(NLI)、语义匹配(SM)和威诺格拉德模式挑战(WSC273)任务上有显著的性能提升,但在 NER 和 RE 任务上的性能一直处于较低水平.

由于 text-davinci-001 和 gpt-3.5-turbo 在 WSJ 数据集上未能按照提示完成任务,因此图 3、图 6 中未展示该数据集的结果.

4 鲁棒性研究

在 NLP 中,鲁棒性通常是指模型在面对噪音、扰动或有意攻击等情况时能够持续可靠地执行任务



(a) text-davinci-003 (b) gpt-3.5-turbo

Fig. 5 Distribution of prediction errors in TACRED dataset

图 5 在 TACRED 数据集上的错误预测的分布

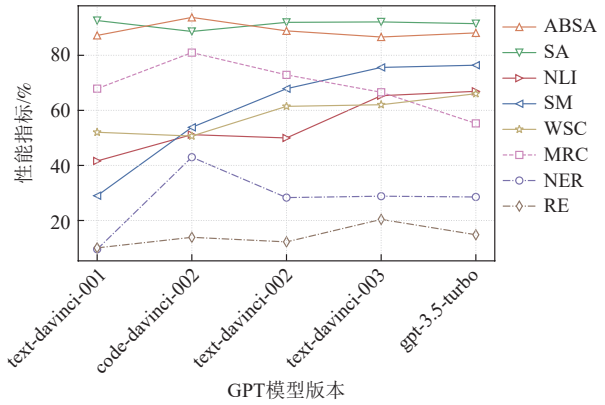


Fig. 6 Performance variations of GPT models

图 6 GPT 模型的性能变化

的能力. 具有较高鲁棒性的模型，在处理不应该对输出造成影响的微小变化的输入时，模型的预测结果不会发生变化. 本节对 GPT 模型面对输入文本扰动时的鲁棒性进行了全面评估，并分析了不同任务和不同变形级别的鲁棒性情况.

4.1 方 法

如表 2 所示，本节使用 TextFlint 提供的 61 种任务特定的变形来评测模型的鲁棒性. 如图 2 所示，每种变形均已通过 TextFlint 提供的变形规则作用于原始数据，从而生成变形数据. 本文通过将提示与变形数据拼接起来，作为测试文本输入模型并获得相应输出.

TextFlint 提供的变形是基于语言学并针对不同的 NLP 任务设计的，在保持变形文本的可接受性的同时，能够更好地代表实际应用中的挑战. 本节中，根据变形的粒度，将变形分为句子级别、词级别和字符级别. 表 3 展示了不同类型的变形样例.

4.2 实验设定

由于在不同任务和变形中使用的评估指标存在差异，本节在鲁棒性评估中引入一个新指标，即性能下降率(performance drop rate, *PDR*). 该指标的计算方式为：

$$PDR(T,P,f_\theta,\mathcal{D})=1-\frac{\sum_{(x,y)\in\mathcal{D}}\mathcal{M}\left[f_\theta([P,T(x)]),y\right]}{\sum_{(x,y)\in\mathcal{D}}\mathcal{M}\left[f_\theta([P,x]),y\right]},\quad(1)$$

其中，$\mathcal{M}$表示不同数据集$\mathcal{D}$使用的评价指标. *PDR* 提供了一种上下文归一化的度量方式，用于量化在处理经过变形 T 的输入 x(使用提示 P)时，模型 f_θ 发生的相对性能下降. 其中，负值的 *PDR* 表示在某些文本变形下会出现性能提升.

本节计算模型在不同数据集和变形中的平均原始性能(ori)、平均变形性能(trans)和平均性能下降率(*APDR*). 此外，使用 BERT 作为基准模型，并且对于每个数据集，GPT 模型和 BERT 都在相同的变形方法和测试数据上进行了评估.

Table 2 Information of 61 Task-Specific Transformations

表 2 61 种任务特定变形的信息

子任务类型	变形类型	变形方式
细粒度情感分析（ABSA）	句子级	AddDiff， RevNon， RevTgt
情感分析（SA）	词级	SwapSpecialEnt-Movie，SwapSpecialEnt-Person
	句子级	AddSum-Movie， AddSum-Person，DoubleDenial
自然语言推理（NLI）	字符级	NumWord
	词级	SwapAnt
	句子级	AddSent
语义匹配（SM）	字符级	NumWord
	词级	SwapAnt
威诺格拉德模式挑战（WSC）	字符级	SwapNames
	词级	SwapGender
	句子级	AddSentences， InsertRelativeClause，SwitchVoice
机器阅读理解（MRC）	句子级	AddSentDiverse， ModifyPos，PerturbAnswer， PerturbQuestion-BackTranslation， PertyrbQuestion-MLM
词性标注（POS）	字符级	SwapPrefix
	词级	SwapMultiPOSJJ， SwapMultiPOSNN，SwapMultiPOSRB， SwapMutliPOSVB
命名实体识别（NER）	字符级	EntTypos， OOV
	词级	CrossCategory， SwapLonger
	句子级	ConcatSent
关系抽取（RE）	词级	SwapEnt-LowFreq， SwapEnt-SamEtype
	句子级	InsertClause， SwapTriplePos-Age，SwapTriplePos-Birth， SwapTriplePos-Employee

Table 3 Examples of Deformations in Different Categories

表 3 不同类型的变形样例

变形类型	变形方式	样例
字符级	SwapPrefix	原始：That is a prefixed string. 变形后：That is a ~~pre~~**un**fixed string.
词级	DoubleDenial	原始：The leading actor is good. 变形后：The leading actor is ~~good~~ **not bad**.
句子级	InsertClause	原始：Shanghai is in the east of China. 变形后：Shanghai **which is a municipality of China** is in the east of China established in Tiananmen.

注：划线单词表示变形后的数据中删掉的部分；黑体单词表示变形后的数据中新增的部分.

4.3 任务层面的鲁棒性

表 4 列出了模型在每个数据集上的平均结果. 具体而言，本文定义 $APDR_D$ 为 *PDR*(式(1))在不同数据集上的平均值：

$$APDR_D(f_\theta,D)=\frac{1}{|\mathcal{T}_D|}\frac{1}{|\mathcal{P}|}\sum_{T\in\mathcal{T}_D}\sum_{P\in\mathcal{P}}PDR(T,P,f_\theta,D),\quad(2)$$

其中，$\mathcal{T}_D$表示特定数据集 D 包含的任务特定变形的集合，$\mathcal{P}$表示 3 个提示的集合.

与第 3 节类似，本节首先分析 gpt-3.5-turbo 和 text-davinci-003 的鲁棒性表现. 表 4 表明，GPT 模型的表现与 BERT 类似，其在分类任务中出现了显著的性能下降. 例如，gpt-3.5-turbo 在 MNLI-mm 数据集上的绝对性能下降了 42.71 个百分点，而 text-davinci-003 在 SemEval2014-Restaurant 数据集上的绝对性能下降了 41.65 个百分点.

此外，GPT 模型在阅读理解（MRC）任务中性能较稳定，其在 SQuAD 1.1 和 SQuAD 2.0 变形前后的数据集上的性能没有出现严重的下降. 但与其他任务不同的是，在 MRC 任务中，text-davinci-003 在性能和鲁棒性方面的表现均优于 gpt-3.5-turbo. 进一步分析发现，如表 4 所示，gpt-3.5-turbo 在该任务上具有较低的精确度（precision），通过抽样分析其生成结果，我们发现原因可能在于 gpt-3.5-turbo 倾向于生成更长的句子. 此外，这 2 个模型的输出均达到 95% 左右的召回率（recall），这表明 GPT 模型在篇章级别的理解任务上具有较强的能力.

同时，GPT 模型对数字和反义词敏感度较高. 在语义匹配任务（包括 QQP 和 MRPC 数据集）中，GPT 模型和 BERT 在变形前后的性能变化上存在显著差距. BERT 在 MRPC 数据集上的变形后性能降至 0，但 GPT 模型在该数据集上的变形后性能甚至有所提升. 通过分析 MRPC 和 QQP 数据集的任务特定变形，即 NumWord 和 SwapAnt，我们发现这 2 种变形通过改变原始数据中的数字或对原始词语进行反义词替换，将原始句子对之间的蕴涵关系转化为矛盾关系. GPT 模型在此类变形上的性能提升表明它们能够较好地捕捉到变形后的文本中数字或反义词所涉及的矛盾关系.

在 NER 和 RE 任务中，GPT 模型性能的下降不明显，有时甚至有提升，尤其是在 OntoNotesv5 和 TACRED 数据集中. 但需要注意的是，模型在这些数据集上的原始性能较低. 因此，在这种情况下，讨论 GPT 模型在这类任务上的鲁棒性缺乏实际意义，提升模型在原始数据上的性能更为紧要.

此外，随着迭代的进行，GPT 系列模型在不同任务上平均性能下降率的变化如图 7 所示. 由于不同模型间的结果波动较大，图 7 的纵坐标数值为经过对数变换之后的结果. 平均性能下降率越小，代表模型的鲁棒性越好，但图中的结果没有呈现出一致的趋势. 在 ABSA 和 MRC 任务中，模型间的鲁棒性表现

Table 4 The Robustness Performance of Different Models

表 4 不同模型的鲁棒性表现 %

数据集	gpt-3.5-turbo			text-davinci-003			BERT		
	ori	trans	*APDR*	ori	trans	*APDR*	ori	trans	*APDR*
Restaurant	91.43±1.23	66.00±11.28	27.80±2.74	90.14±1.33	52.59±11.21	41.65±4.26	84.38±1.20	53.49±15.07	36.51±18.43
Laptop	86.67±2.15	59.36±21.97	31.25±23.31	83.30±0.71	54.71±17.75	34.42±19.29	90.48±0.06	49.06±9.03	45.78±9.97
IMDB	91.60±0.20	90.86±0.50	0.80±0.47	91.74±0.68	91.40±0.58	0.37±0.31	95.24±0.12	94.61±0.80	0.66±0.94
MNLI-m	73.03±7.44	41.75±17.05	42.27±21.87	67.49±2.80	54.88±20.93	19.52±24.60	86.31±4.50	52.49±2.97	39.10±4.13
MNLI-mm	72.21±7.69	40.94±19.11	42.71±24.31	66.61±1.57	50.57±20.58	24.46±27.71	84.17±1.09	52.33±5.44	37.87±5.73
SNLI	73.30±12.50	47.80±8.80	32.99±13.66	70.81±9.24	56.44±22.68	18.99±26.16	90.75±1.52	77.61±18.34	14.44±20.25
QQP	79.32±5.97	64.96±20.52	17.17±1.18	70.14±12.03	69.27±13.67	−1.08±9.23	91.75±2.60	52.77±5.93	42.56±4.83
MRPC	80.69±10.28	84.99±10.69	−8.12±22.99	74.87±5.38	74.33±23.12	−0.17±26.51	86.87±6.05	0.00±0.00	100.00±0.00
WSC273	66.05±1.95	64.12±5.82	2.93±5.57	62.05±0.48	61.42±2.41	1.01±3.12	56.00±0.00	53.61±5.31	4.26±9.49
SQuAD 1.1	55.33±8.22	44.55±9.73	19.45±12.39	67.18±8.23	61.07±9.04	9.11±7.13	87.22±0.26	70.78±21.84	18.88±24.95
SQuAD 2.0	55.03±7.39	44.21±9.31	19.62±12.70	65.91±7.81	59.70±8.93	9.45±7.58	78.81±2.65	60.17±16.99	23.48±21.81
WSJ	–	–	–	75.53±2.28	74.63±2.58	1.21±0.90	97.72±0.09	96.23±1.69	1.53±1.79
CoNLL2003	44.61±3.48	37.30±9.29	16.31±20.05	51.54±2.88	42.64±9.24	17.13±17.76	90.57±0.38	72.24±16.75	20.26±18.42
OntoNotesv5	17.74±8.51	18.68±7.00	−12.73±40.09	11.94±9.98	12.30±7.69	−17.51±51.73	79.99±6.54	61.98±20.30	23.47±20.45
TACRED	31.44±31.24	32.64±33.27	0.58±7.88	35.67±30.89	38.67±31.59	−25.69±55.14	77.99±13.47	65.53±15.46	16.54±7.83

注：“±”后的数字表示均值对应的标准差；“Laptop”和“Restaurant”分别表示“SemEval2014-Laptop”和“SemEval2014-Restaurant”数据集；“–”表示模型未完成指定任务.

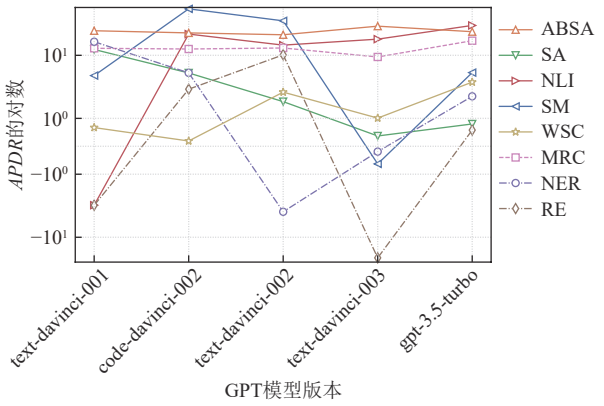


Fig. 7 APDR variations of GPT models

图 7 GPT 模型的平均性能下降率的变化

较为相似; 在 SA 任务上出现了较显著的鲁棒性提升; 但是在其余任务中均呈现出显著的波动, 并且没有出现鲁棒性显著提升的情况. 这可能表明 GPT 模型的迭代过程主要集中于改进模型在一般场景下的性能, 而非解决鲁棒性问题.

4.4 变形层面的鲁棒性

图 8 为 GPT 模型在 3 种变形级别上的性能下降情况. 其中斜杠部分表示模型的变形后性能, 无斜杠部分表示变形后性能与原始性能的差值, 折线表示平均性能下降率(*APDR*). 通过计算每个变形级别下的 *PDR* 的均值得到 $APDR_{\mathcal{T}_t}$:

$$APDR_{\mathcal{T}_t}(f_\theta,\mathcal{T}_t)=\frac{1}{|\mathcal{D}|}\frac{1}{|\mathcal{P}|}\sum_{D\in\mathcal{D}}\sum_{P\in\mathcal{P}}PDR(\mathcal{T}_t,P,f_\theta,D),\quad(3)$$

其中, $\mathcal{T}_t$ 表示某个变形类别 t 的变形集合, $\mathcal{P}$ 表示提示的集合.

根据图 8 所示, GPT 模型的 *APDR* 在句子级、词级、字符级 3 个变形类别上逐级递减, 即处理句子级别的变形文本时, GPT 模型在变形前后的性能下降更为显著. 句子级别的变形通常涉及语义的重新表述或句子整体结构的改变, 这对模型稳定性有更高的要求. 此外, GPT 模型在字符级和词级变形上表现出比 BERT 更好的鲁棒性. GPT 模型的平均性能下降范围为 9.61%~15.22%, 而 BERT 在字符级和词级变形上的性能下降分别为 36.74% 和 37.07%. 可以看出, 与监督微调模型相比, GPT 模型对细粒度扰动表现出更强的稳定性.

5 性能和鲁棒性影响因素

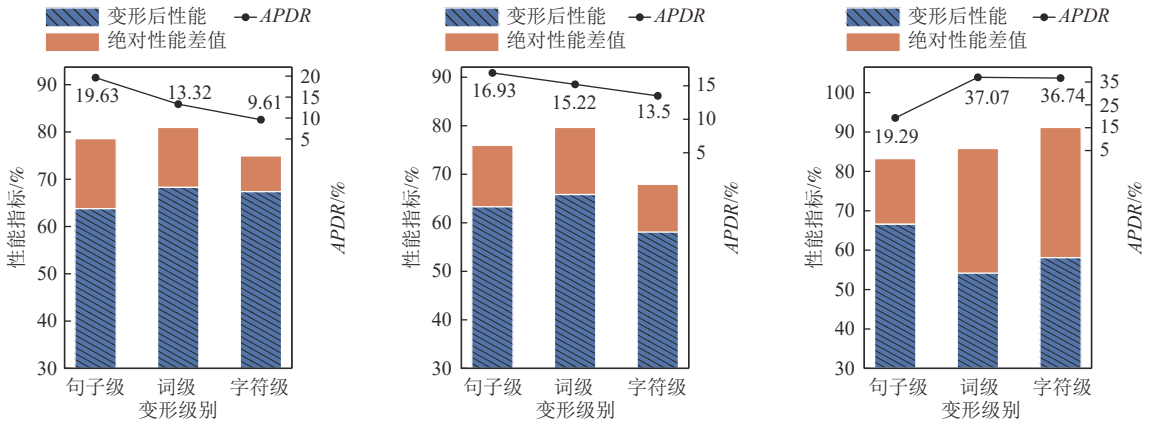
在第 3 节和第 4 节中, 本文使用涵盖了各种任务和文本变形的大量测试数据, 对 GPT 模型的性能和鲁棒性进行了评估. 除测试文本之外, 提示是评测过程中模型输入数据的另一个重要部分, 并且基于提示中少量示例的上下文学习已经成为 NLP 领域的新范式. 基于此, 本节探究提示对 GPT 模型的性能和鲁棒性的影响, 具体关注 2 个方面: 1)提示中演示数量的影响; 2)提示中演示内容的影响. 其中, 演示是指提示中的示例或样本, 通常用来说明我们所期望模型输出的结果.

5.1 演示数量的影响

通过改变演示数量(即图 2 中的“*k*”), 本文研究了在 0、1 和 3 个演示数量下模型的原始性能表现和变形前后性能的变化.

图 9 结果表明, 增加演示数量通常会带来性能的提升. 此外, 从零样本增加为少样本的情况下, 模型性能提升显著, 特别是对于一开始在零样本情景下表现不佳的任务, 如信息抽取任务. 此外, 随着演示数量的增加, 不同 GPT 模型之间的性能差异减小.

然而, 就变形前后的性能变化而言, 在大多数情



(a) text-davinci-003 (b) gpt-3.5-turbo (c) BERT

Fig. 8 Performance drop of different models on three transformation categories

图 8 不同模型在 3 种变形类别上的性能下降情况

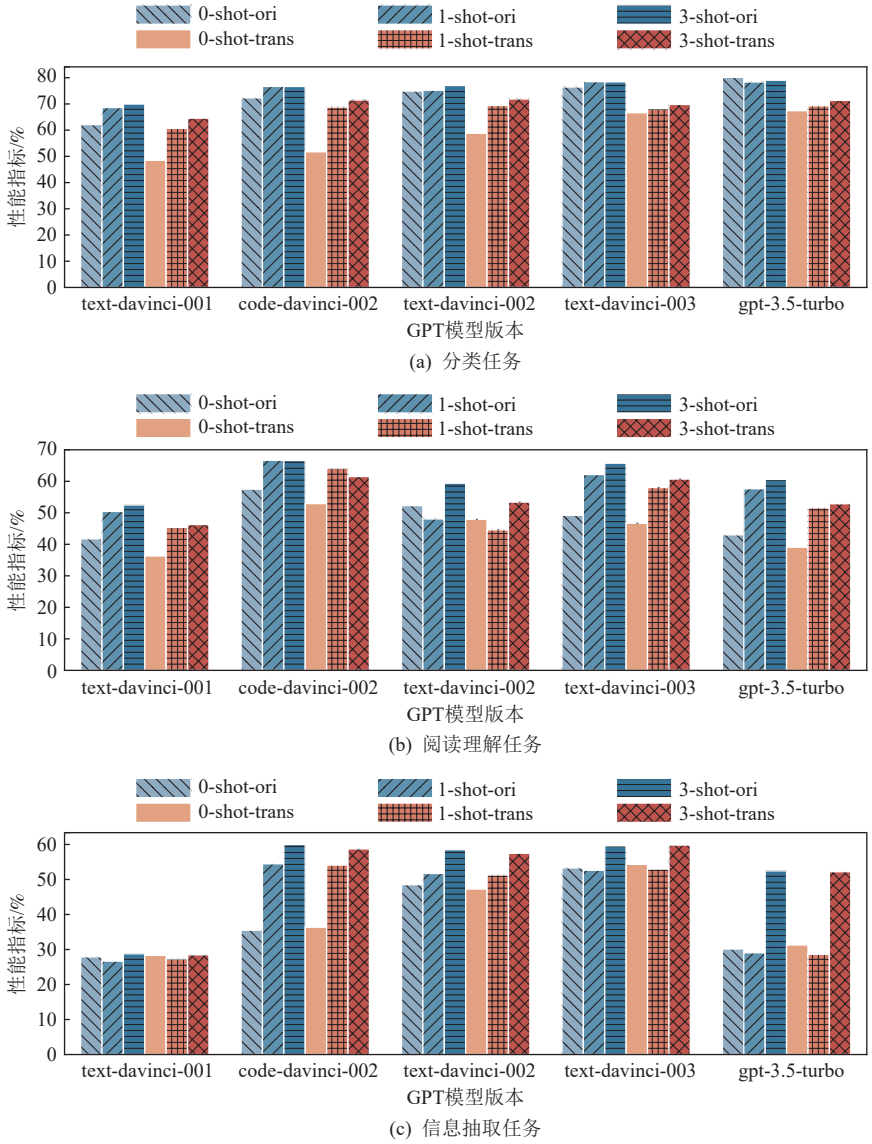


Fig. 9 Original and transformed performance of GPT models on 0-shot, 1-shot, and 3-shot

图 9 GPT 模型在 0-shot、1-shot、3-shot 样本场景下原始性能与变形后的性能表现

况下，增加演示数量没有显著缓解模型的性能下降. 只有在分类任务中，可以观察到 text-davinci-001，code-davinci-002 和 text-davinci-002 的性能下降有所缓解. 这表明增加演示数量虽然可以改善模型在原始任务上的性能，但并不能有效提高模型面对扰动时的鲁棒性.

5.2 演示内容的影响

在 5.1 节中的少样本情景下，原始数据和变形后数据均使用相同的、未经过变形的演示样例来研究变形后测试数据引起的性能变化. 本节研究在提示中使用变形后的演示样例对模型的鲁棒性有何影响. 本文分别从分类、信息抽取和阅读理解三大类任务中选取 SemEval2014-Restaurant (Restaurant), CoNLL2003 和 SQuAD 1.1 数据集作为代表进行实验. 对于每个数据集，演示样例使用该数据集特定的任务变形进行变换，并与变形后的测试数据拼接，用以评估模型变形后的性能. 演示样例的数量为 3.

图 10 展示了变形前后模型的 *APDR*. 结果表明，在演示中使用变形后的样本有助于缓解模型变形后的性能下降，说明演示中包含的扰动信息能够帮助模型更好地处理变形数据. 但是，*APDR* 依然处于较高的数值，这表明这种性能改善是有限的，不足以从根本上解决模型的鲁棒性问题.

6 讨 论

6.1 GPT 更新版本的表现

本文前文主要针对 GPT-3 和 GPT-3.5 系列模型

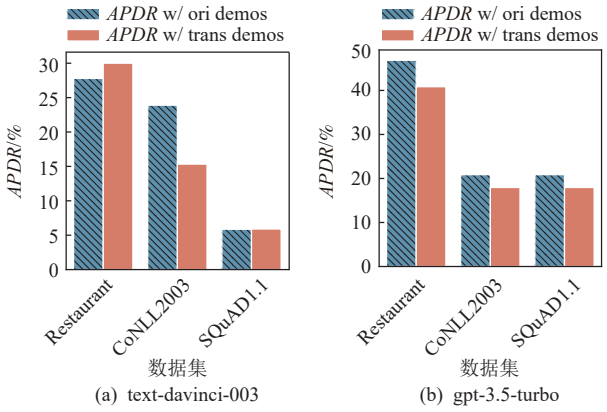


Fig. 10 *APDR* with original and transformed demonstrations date

图 10 模型使用原始和变形后的演示数据的 *APDR*

的性能和鲁棒性表现进行了探究. 随着时间的推进, GPT 系列模型仍然在持续迭代, 并且 Chen 等人[40]、Tu 等人[41] 近期的工作表明模型的表现会随时间发生变化. 为了更好地验证本文实验结果的可持续性, 本节针对 GPT 系列模型的更新版本“gpt-3.5-turbo-0613”(上文中的“gpt-3.5-turbo”为“gpt-3.5-turbo-0301”版本)、“gpt-4”进行性能和鲁棒性评测.

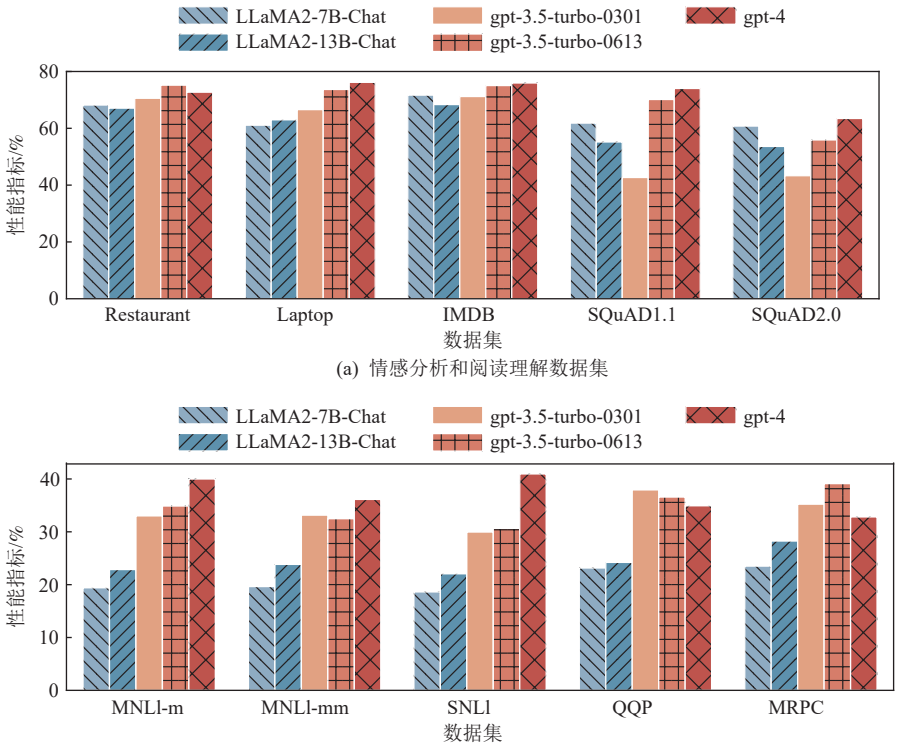
首先是模型的性能表现. 如图 11 所示, 根据模型更新与迭代顺序, gpt-3.5-turbo-0613 和 gpt-4 模型在大部分数据集上的性能表现较为显著的提升. 其中, 在情感分析和阅读理解的数据集中, 这 2 个模型的提升最为显著. 第 3 节中的结果表明 GPT 模型在 NER 和 RE 任务上表现不佳, 图 11 表明 gpt-3.5-turbo-0613 和 gpt-4 模型在 NER 任务的 OntoNotesv5 数据集及 RE 任务的 TACRED 数据集上的表现仍然处于较低水平.

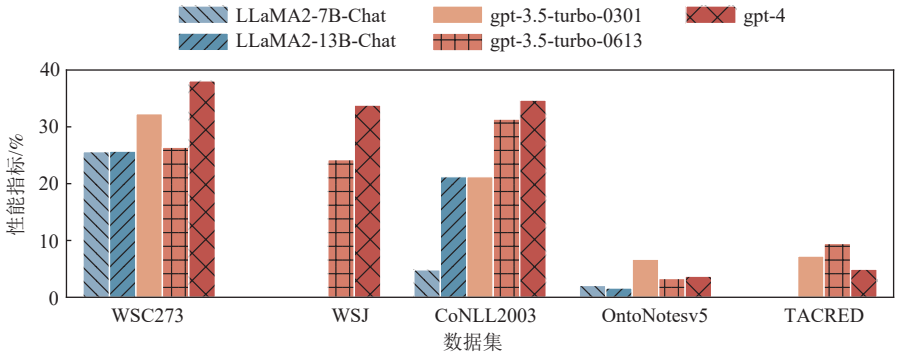
其次是模型的鲁棒性表现. 表 5 展示了 3 个模型的鲁棒性表现. 如表 5 所示, GPT 模型仍然存在 4.3 节中提到的鲁棒性问题, 尤其在分类任务中存在显著的性能下降. 值得注意的是, 在阅读理解任务中 gpt-3.5-turbo-0613 和 gpt-4 的鲁棒性进一步提升, 表现出在该任务上较高的稳定性. 同时, gpt-3.5-turbo 的版本迭代未带来稳定的鲁棒性提升, 而 gpt-4 的鲁棒性在大多任务上都优于 GPT-3.5 系列模型.

6.2 开源模型的表现

由于 GPT 系列模型出色的性能和较完善的迭代过程, 对其进行的性能和鲁棒性评测有助于更全面地了解大模型的能力及其发展进程中的变化, 但是由于闭源模型的限制, 后续在 GPT 系列模型上进行优化较为困难. 为此, 本节对开源模型 LLaMA2-7B 和 LLaMA2-13B 进行性能和鲁棒性评测.

如图 11 第 1 个子图所示, LLaMA2-7B 和 LLaMA2-13B 在情感分析和阅读理解类任务上的表现与 GPT-3.5 系列模型相当; 在第 2 个子图中, 其在自然语言推理和语义匹配任务中却与 GPT-3.5 系列





(c) 信息抽取数据集

注：“Laptop”和“Restaurant”分别表示“SemEval2014-Laptop”和“SemEval2014-Restaurant”数据集. 柱状图中 WSJ 和 TACRED 数据集空缺的部分表示模型未完成在该数据集上的指定任务.

Fig. 11 Performance of GPT and LLaMA2 models

图 11 GPT 和 LLaMA2 模型的性能表现

Table 5 The Robustness Performance of Three GPT Models

表 5 3 个 GPT 模型的鲁棒性表现 %

数据集	gpt-3.5-turbo-0301			gpt-3.5-turbo-0613			gpt-4		
	ori	trans	*APDR*	ori	trans	*APDR*	ori	trans	*APDR*
Restaurant	91.43±1.23	66.00±11.28	27.80±2.74	97.05±0.86	59.98±16.37	38.28±16.56	95.81±2.27	71.07±9.15	25.80±9.69
Laptop	86.67±2.15	59.36±21.97	31.25±23.31	93.91±1.45	63.82±19.10	32.16±19.83	98.74±1.88	74.42±16.01	24.75±15.42
IMDB	91.60±0.20	90.86±0.50	0.80±0.47	96.58±1.05	95.99±1.63	0.62±0.90	93.81±3.69	91.91±5.31	2.05±3.83
MNLI-m	73.03±7.44	41.75±17.05	42.27±21.87	71.88±7.99	35.30±16.00	51.85±20.03	84.24±7.00	53.46±10.50	36.81±9.04
MNLI-mm	72.21±7.69	40.94±19.11	42.71±24.31	71.78±7.68	35.59±15.45	50.28±22.50	80.23±8.14	53.88±14.19	33.28±14.43
SNLI	73.30±12.50	47.80±8.80	32.99±13.66	75.67±15.70	38.58±11.11	47.61±16.40	89.10±5.64	70.65±21.60	21.25±21.31
QQP	79.32±5.97	64.96±20.52	17.17±1.18	81.42±8.49	49.71±16.16	38.22±22.66	53.14±19.48	84.91±15.74	−105.86±159.05
MRPC	80.69±10.28	84.99±10.69	−8.12±22.99	85.70±11.16	70.65±16.74	14.29±30.49	60.38±7.06	94.65±4.68	−58.46±18.46
WSC273	66.05±1.95	64.12±5.82	2.93±5.57	53.98±0.75	51.92±3.13	3.80±6.10	77.88±6.12	64.42±23.57	16.91±30.39
SQuAD1.1	55.33±8.22	44.55±9.73	19.45±12.39	90.11±1.09	80.84±8.65	10.27±9.70	95.14±1.74	84.96±13.75	10.69±14.41
SQuAD2.0	55.03±7.39	44.21±9.31	19.62±12.70	73.68±4.61	64.25±10.76	12.85±13.16	81.94±3.17	74.15±7.17	9.50±8.02
WSJ	–	–	–	50.35±5.22	49.31±5.61	2.07±4.52	68.66±3.03	67.88±5.58	1.10±7.39
CoNLL2003	44.61±3.48	37.30±9.29	16.31±20.05	66.78±2.98	49.76±11.69	25.38±17.69	83.23±1.86	65.53±13.86	21.25±16.66
OntoNotesv5	17.74±8.51	18.68±7.00	−12.73±40.09	9.85±6.53	13.50±4.13	−66.86±72.42	7.58±15.72	6.70±10.70	10.87±15.47
TACRED	31.44±31.24	32.64±33.27	0.58±7.88	37.00±35.29	40.23±34.38	−20.07±36.33	14.32±7.57	13.31±9.17	−0.02±74.59

注：“±”后的数字表示均值对应的标准差；“Laptop”和“Restaurant”分别表示“SemEval2014-Laptop”和“SemEval2014-Restaurant”数据集；“–”表示模型未完成指定任务.

模型存在较大差距. 需要注意的是，LLaMA2-7B 和 LLaMA2-13B 在 WSJ 和 TACRED 数据集中均未按照指令完成相应任务，并且在 NER 任务中的表现亟待提升.

如表 6 所示，与 GPT 系列模型的鲁棒性表现类似，LLaMA2-7B 和 LLaMA2-13B 在大多分类任务上的性能下降都较为严重，但在阅读理解任务中的鲁棒性与 gpt-4 相当，且好于 GPT-3.5 系列模型. 同时，LLaMA2-13B 比 LLaMA2-7B 具有更好的鲁棒性.

7 总 结

本文通过评估涵盖 9 个不同 NLP 任务的 15 个数据集，使用 61 种任务特定的变形方法，对 GPT-3 和 GPT-3.5 系列模型的性能和鲁棒性进行了全面分析. 研究结果表明，尽管 GPT 模型在情感分析、语义匹配等分类任务和阅读理解任务表现出色，但在面对输入文本扰动时仍然存在明显的鲁棒性问题. 其

Table 6 The Robustness Performance of LLaMA2 Model

表 6 LLaMA2 模型的鲁棒性表现 %

数据集	LLaMA2-7B			LLaMA2-13B		
	ori	trans	*APDR*	ori	trans	*APDR*
Restaurant	87.85±1.68	52.38±7.01	40.34±8.22	87.10±3.17	35.16±9.07	59.84±9.45
Laptop	79.40±2.93	56.23±12.68	28.96±16.86	81.15±2.82	47.21±18.58	41.87±22.81
IMDB	92.04±1.68	91.06±2.68	1.08±1.43	88.17±2.30	87.40±2.89	0.88±1.21
MNLI-m	46.76±16.03	27.64±13.39	34.77±34.65	54.47±15.15	44.70±18.95	12.52±43.92
MNLI-mm	50.16±17.23	27.92±13.99	39.21±32.29	57.04±15.11	45.47±19.30	15.94±42.02
SNLI	47.77±19.73	30.73±17.44	27.79±41.43	54.79±15.20	43.75±24.22	12.83±53.93
QQP	59.93±16.77	33.18±11.02	40.58±24.61	54.49±12.91	40.17±14.45	21.36±32.47
MRPC	70.66±14.76	66.49±16.68	1.92±33.62	69.59±17.74	33.75±32.70	43.09±63.48
WSC273	52.40±3.60	53.10±1.68	−1.65±7.48	52.57±0.73	56.43±2.77	−7.33±4.58
SQuAD1.1	79.64±0.69	67.85±9.98	14.80±12.51	71.27±1.16	63.67±5.14	10.65±7.12
SQuAD2.0	78.25±0.95	66.30±9.66	15.26±12.36	69.40±1.27	61.77±5.05	10.99±7.20
WSJ	–	–	–	–	–	–
CoNLL2003	20.05±8.92	4.44±5.36	74.37±36.93	45.66±10.22	20.26±10.27	53.47±26.94
OntoNotesv5	4.97±2.57	4.94±2.03	−19.85±76.91	5.87±5.21	5.36±3.34	−8.23±51.59
TACRED	–	–	–	4.26±2.60	5.95±5.45	−16.67±104.08

注: “±”后的数字表示均值对应的标准差; “Laptop”和“Restaurant”分别表示“SemEval2014-Laptop”和“SemEval2014-Restaurant”数据集; “–”表示模型未完成指定任务.

中, 本文分别从任务层面和变形级别层面具体分析了 GPT 模型的鲁棒性表现, 表明其在分类任务和句子级变形中的鲁棒性亟待提升. 同时, 随着 GPT 系列模型的迭代, 其性能在大多数任务上稳步提升, 但鲁棒性依然面临很大的挑战. 此外, 本文探讨了提示对 GPT 模型的性能和鲁棒性的影响, 包括提示中演示数量和演示内容 2 方面. 这些发现从任务类型、变形种类、提示内容等方面揭示了 GPT 模型还无法完全胜任常见的 NLP 任务, 并且模型存在的鲁棒性问题难以通过提升模型性能或改变提示内容等方式解决. 与此同时, 本文通过评估 gpt-3.5-turbo 的更新版本、gpt-4 模型, 以及开源模型 LLaMA2-7B 和 LLaMA2-13B 的性能和鲁棒性表现, 进一步验证了实验结论. 鉴于此, 未来的大模型研究应当提升模型在信息提取和语义理解方面的能力, 并且应当在模型训练或微调阶段考虑提升模型的鲁棒性.

作者贡献声明: 陈炫婷提出研究思路和实验方案, 负责部分实验和论文写作; 叶俊杰负责部分实验和完善论文; 祖璨负责部分实验并整理分析实验结果; 许诺协助实验和完善论文; 桂韬提出指导意见并修改论文; 张奇提出指导意见并审阅论文.

参 考 文 献

[1] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners[J]. arXiv preprint, arXiv: 2109.01652, 2021

[2] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C/OL]//Advances in Neural Information Processing Systems. [2023-09-10].https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[3] Touvron H, Lavril T, Izacard G, et al. LLaMA: Open and efficient foundation language models[J]. arXiv preprint, arXiv: 2302.13971, 2023

[4] Anil R, Dai A M, Firat O, et al. PaLM 2 technical report[J]. arXiv preprint, arXiv: 2305.10403, 2023

[5] Kaplan J, McCandlish S, Henighan T, et al. Scaling laws for neural language models[J]. arXiv preprint, arXiv: 2001.08361, 2020

[6] Qin Chengwei, Zhang A, Zhang Zhuosheng, et al. Is ChatGPT a general-purpose natural language processing task solver?[C]//Proc of the 2023 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2023: 1339–1384

[7] Bang Y, Cahyawijaya S, Lee N, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity[J]. arXiv preprint, arXiv: 2302.04023, 2023

[8] Kosinski M. Theory of mind may have spontaneously emerged in large language models[J]. arXiv preprint, arXiv: 2302.02083, 2023

[9] Frieder S, Pinchetti L, Griffiths R R, et al. Mathematical capabilities of ChatGPT[C/OL]//Advances in Neural Information Processing Systems. [2023-09-10]. https://neurips.cc/virtual/2023/poster/73421

[10] Wang Jindong, Hu Xixu, Hou Wenxin, et al. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective[J]. arXiv preprint, arXiv: 2302.12095, 2023

[11] Wang Boxin, Xu Chejian, Wang Shuohang, et al. Adversarial glue: A multi-task benchmark for robustness evaluation of language models[J]. arXiv preprint, arXiv: 2111.02840, 2021

[12] Nie Y, Williams A, Dinan E, et al. Adversarial NLI: A new benchmark for natural language understanding[C]//Proc of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 4885−4901

[13] Fansi T A, Goel R, Wen Zhi, et al. DDXPlus: A new dataset for automatic medical diagnosis[C/OL]//Advances in Neural Information Processing Systems. [2023-09-10].https://proceedings.neurips.cc/paper_files/paper/2022/hash/cae73a974390c0edd95ae7aeae09139c-Abstract-Datasets_and_Benchmarks.html

[14] Zhu Kaijie, Wang Jindong, Zhou Jiaheng, et al. PromptBench: Towards evaluating the robustness of large language models on adversarial prompts[J]. arXiv preprint, arXiv: 2306.04528, 2023

[15] Wang Xiao, Liu Qin, Gui Tao, et al. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing[C]//Proc of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int Joint Conf on Natural Language Processing: System Demonstrations. Stroudsburg, PA: ACL, 2021: 347−355

[16] Dong Qingxiu, Li Lei, Dai Damai, et al. A survey for in-context learning[J]. arXiv preprint, arXiv: 2301.00234, 2022

[17] Zhuo T Y, Huang Yujin, Chen Chunyang, et al. Exploring AI ethics of ChatGPT: A diagnostic analysis[J]. arXiv preprint, arXiv: 2301.12867, 2023

[18] Choi J H, Hickman K E, Monahan A B, et al. ChatGPT goes to law school[J]. Journal of Legal Education, 2021, 71(3): 387

[19] Khalil M, Er E. Will ChatGPT get you caught? Rethinking of plagiarism detection[C]//Proc of Int Conf on Human-Computer Interaction. Berlin: Springer，2023: 475−487

[20] Alshater M. Exploring the role of artificial intelligence in enhancing academic performance: A case study of ChatGPT[J/OL]. [2023-09-12]. http://dx. doi.org/10.2139/ssrn.4312358

[21] Tabone W, De Winter J. Using ChatGPT for human–computer interaction research: A primer[J]. Royal Society Open Science, 2023, 10(9): 21

[22] Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports[J]. European Radiology, 2023: 1−9

[23] Biswas S. ChatGPT and the future of medical writing[J]. Radiology, 2023, 307(2): e223312

[24] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014

[25] Pontiki M, Galanis D, Pavlopoulos J, et al. SemEval-2014 Task 4: Aspect based sentiment analysis[C]//Proc of the 8th Int Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2004: 27−35

[26] Maas A L, Daly R E, Pham P T, et al. Learning word vectors for sentiment analysis[C]// Proc of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: ACL, 2011: 142−150

[27] Williams A, Nangia N, Bowman S R. A broad-coverage challenge corpus for sentence understanding through inference[J]. arXiv preprint, arXiv: 1704.05426, 2017

[28] Dolan W B, Brockett C. Automatically constructing a corpus of sentential paraphrases[C]//Proc of the 3rd Int Workshop on Paraphrasing (IWP2005). Jeju Island: Asia Federation of Natural Language Processing, 2005: 9−16

[29] Wang Zhiguo, Hamza W, Florian R. Bilateral multi-perspective matching for natural language sentences[C]//Proc of the 26th Int Joint Conf on Artificial Intelligence. Australia: IJCAI. org, 2017: 4144−4150

[30] Levesque H, Davis E, Morgenstern L. The winograd schema challenge[C]//Proc of 13th Int Conf on the Principles of Knowledge Representation and Reasoning. Palo Alto, CA: AAAI, 2012: 552−561

[31] Rajpurkar P, Zhang Jian, Lopyrev K, et al. SQuAD: 100, 000+ questions for machine comprehension of text[C]//Proc of the 2016 Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2016: 2383−2392

[32] Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD[C]//Proc of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Stroudsburg, PA: ACL, 2018: 784−789

[33] Marcus M, Santorini B, Marcinkiewicz M A. Building a large annotated corpus of English: The Penn Treebank[J]. Computational Linguistics, 1993, 19(2): 313−330

[34] Sang E T K, De Meulder F. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition[C]//Proc of the 7th Conf on Natural Language Learning at HLT-NAACL 2003. Stroudsburg, PA: ACL, 2003: 142−147

[35] Weischedel R, Palmer M, Marcus M, et al. Ontonotes release 5.0 ldc2013t19[J]. Linguistic Data Consortium, 2013, 23(1): 170

[36] Zhang Yuhao, Zhong V, Chen Danqi, et al. Position-aware attention and supervised data improve slot filling[C]//Proc of Conf on Empirical Methods in Natural Language Processing. Stroudsburg, PA: ACL, 2017: 35−45

[37] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback[C/OL]//Advances in Neural Information Processing Systems. [2023-09-10]. https://proceedings.neurips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html

[38] Christiano P F, Leike J, Brown T, et al. Deep reinforcement learning from human preferences[C/OL]// Advances in Neural Information Processing Systems. [2023-09-10]. https://papers.nips.cc/paper_files/paper/2017/hash/d5e2c0adad503c91f91df240d0cd4e49-Abstract.html

[39] Kenton J D M W C, Toutanova L K. BERT: Pre-training of deep bidirectional Transformers for language understanding[C]//Proc of NAACL-HLT. Stroudsburg, PA: ACL, 2019: 4171−4186

[40] Chen Lingjiao, Zaharia M, Zou J. How is ChatGPT's behavior changing over time?[J]. arXiv preprint, arXiv: 2307.09009, 2023

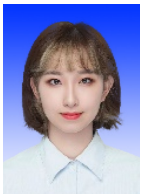
[41] Tu Shangqing, Li Chunyang, Yu Jifan, et al. ChatLog: Recording and analyzing ChatGPT across time[J]. arXiv preprint, arXiv: 2304.14106, 2023

Chen Xuanting, born in 1999. Master candidate. Her main research interests include natural language processing and robust models.

陈炫婷, 1999 年生. 硕士研究生. 主要研究方向为自然语言处理、鲁棒模型.

Ye Junjie, born in 2001. PhD candidate. His main research interest includes natural language processing.

叶俊杰, 2001 年生. 博士研究生. 主要研究方向为自然语言处理.



Zu Can, born in 2000. Master candidate. Her main research interests include large language models and information extraction.

祖　璨, 2000 年生. 硕士研究生. 主要研究方向为大语言模型、信息抽取.

Xu Nuo, born in 1998. Master candidate. Her main research interests include natural language processing and large language models.

许　诺, 1998 年生. 硕士研究生. 主要研究方向为自然语言处理、大语言模型.

Gui Tao, born in 1989. PhD, associate professor, master supervisor. His main research interests include pre-training models, information extraction, and robust models.

桂　韬, 1989 年生. 博士，副研究员，硕士生导师. 主要研究方向为预训练模型、信息抽取、鲁棒模型.

Zhang Qi, born in 1981. PhD, professor, PhD supervisor. His main research interests include natural language processing and information retrieval.

张　奇, 1981 年生. 博士，教授，博士生导师. 主要研究方向为自然语言处理、信息检索.