GPT-3.5 相较于前代模型有多稳健？关于语言理解任务的综合研究

XuantingChen★*, Junjie Ye★*, Can Zu★, Nuo Xu★, Rui Zheng★,Minlong Peng★, Jie Zhou★, TaoGui♦†, Qi Zhang★,♣, Xuanjing Huang★★ 复旦大学计算机科学学院，中国上海♦ 复旦大学现代语言与语言学研究所，中国上海♣ 复旦大学上海智能视觉计算协同创新中心xuantingchen21@m.fudan.edu.cn
{jjye19,tgui,qz,xjhuang}@fudan.edu.cn

摘要

GPT-3.5 模型在各种自然语言处理（NLP）任务中展现出令人印象深刻的性能，体现了其强大的理解和推理能力。然而，其鲁棒性以及应对开放世界中各种复杂情况的能力尚未得到探索，而这对于评估模型的稳定性尤为关键，也是可信人工智能的一个关键方面。在本研究中，我们对 GPT-3.5 进行了全面的实验分析，利用 TextFlint 提供的 66 种文本变换，在涵盖 9 项热门自然语言理解（NLU）任务的 21 个数据集（约 116K 个测试样本）上探索其鲁棒性。我们的发现表明，尽管 GPT-3.5 在某些任务上的表现优于现有微调模型，但它仍然面临显著的鲁棒性退化，例如，其在自然语言推理和情感分析任务中的平均性能分别最多下降 35.74% 和 43.59%。我们还表明，GPT-3.5 面临一些特定的鲁棒性挑战，包括鲁棒性不稳定、提示敏感性和数字敏感性。这些见解有助于理解其局限性，并为未来研究应对这些挑战、提升 GPT-3.5 的整体性能和泛化能力提供指导。

1 引言

大语言模型（LLM），例如 FLAN（Wei 等人，2022年）、GPT-3（Brown 等人，2020年）、OPT（Zhang 等人，2022年）和 PaLM（Chowdhery 等人，2022年），在对话质量、理解和推理方面展现出卓越能力，表现已接近人类水平。只需输入恰当的提示或指令，无需修改任何参数，它们便能执行各种任务（Kaplan 等人， 2020; Thoppilan 等人，2022年）。这些进展正推动自然语言处理领域迈向通用人工智能（AGI）的范式（Goertzel， 2014; Fei 等人，2022年）。在现有的大语言模型中，生成式预训练变换器（GPT）（Brown 等人，2020年）系列模型凭借其卓越能力尤为流行。

大语言模型取得的卓越成就激发了人们对其能力的广泛研究。以往研究主要考察大语言模型在特定能力上的表现，包括常识与逻辑推理（Qin 等人，2023年）、多语言能力与多模态（Bang 等人，2023年）、心智理论（Kosinski，2023年）以及数学能力（Frieder 等人，2023年）。然而，尽管大语言模型在留出测试集上取得高分，并在这些方面展现出显著能力，但它们是否足够鲁棒，能够应对开放世界的复杂性，仍不清楚。评估鲁棒性对于确保模型在现实世界中的可靠性至关重要，尤其是在自动驾驶和医学诊断等安全相关场景中。然而，目前尚未对大语言模型的鲁棒性进行充分的系统评估。

文本鲁棒性评估方法衡量模型在输入存在轻微修改时的稳定性，这些修改发生在字符级、词语级或句子级。其中一类研究

* 贡献相同。
† 通讯作者。

重点关注通过对抗攻击生成语义等价的对抗样本（Jin 等， 2020;Ebrahimi 等，2018年）。然而，这些方法的计算成本高、耗时长，可能不适用于大规模评估。因此，另一类研究设计了通用的鲁棒性评估基准，涉及手动设计或修改数据，包括 CheckList（Ribeiro 等，2020年）、Dynabench（Potts 等，2020年）和 TextFlint（Gui 等，2021年）。在本文中，我们选择了 TextFlint——一个涵盖众多自然语言处理任务和基于语言学的文本变换的统一工具包——来开展全面的鲁棒性评估。此外，我们选取最先进的 GPT-3.5 模型 *text-davinci-003* 作为分析对象。该模型出色的性能以及通过 OpenAI API 使用的便捷性，使我们能够在大规模评估中发现大语言模型的鲁棒性局限，并为未来研究提供有价值的见解。

在本研究中，我们通过实证评估 GPT-3.5 在 21 个数据集上的表现（共 115715 个测试样本），这些数据集涵盖 9 种常见的自然语言理解任务，并包含来自 TextFlint 的 66 种文本变换。我们的研究聚焦于以下方面：(1) GPT-3.5 在零样本和少样本场景下的鲁棒性与性能; (2) 对特定鲁棒性挑战与局限性的探索5; (3) GPT-3.5 与微调模型的比较; (4) GPT-3（*text-davinci-001*）与 GPT-3.5 系列（*text-davinci-002*、*text-davinci-003*）模型之间的差异。

我们的研究得出了以下发现：

- **测试集上的竞争力结果：**与使用任务特定数据进行微调的监督模型相比，GPT-3.5 在某些自然语言理解任务上取得了最先进的结果。具体而言，GPT-3.5 在阅读理解和情感分析任务上表现出色，但在序列标注和关系抽取任务上面临重大挑战（图1）。

- **缺乏鲁棒性：** GPT-3.5 仍然面临显著的鲁棒性退化，例如，在自然语言推理（第4.3.1节）和情感分析（第4.4节）任务中，其平均性能分别下降了最高达35.74%和43.59%。然而，值得注意的是，GPT-3.5在某些任务上表现出显著的鲁棒性，例如阅读理解（第4.1节）和WSC（第4.6节）任务。

- **鲁棒性不稳定：** 在少样本场景下，GPT-3.5在不同任务中的鲁棒性提升幅度差异很大。例如，GPT-3.5在基于方面的情感分析任务中表现出显著提升，而在自然语言推理（第4.3.1节）和语义匹配（第4.3.2节）任务中，鲁棒性反而下降。

- **提示敏感性：** 输入提示的变化[1] 会对结果产生显著影响，而GPT- 3.5对提示变化的鲁棒性仍有待提升。在情感分析（第4.4节）、自然语言推理（第4.3.1节）、语义匹配（第4.3.2节）和阅读理解（第4.1节）任务中，GPT-3.5在不同提示下表现出较高的方差。

- **数字敏感性：** 与预训练微调模型相比，GPT-3.5 对数字输入更加敏感。例如，在*Num Word* 变换中，将句子中的数字词替换为不同的数值，GPT-3.5 表现出显著较高的敏感性（第4.3节）。

- **任务标签敏感性：** 我们推测，指令微调阶段的任务构建可能会对模型的表现产生重大影响。对于 IMDB 二元情感分类数据集，模型会输出大量“中立”回答，而这些回答不属于应用标签空间，因而导致性能下降（第4.4.2节）。

- **零样本/少样本场景下显著提升：**在零样本和少样本场景下，GPT-3.5 在大多数自然语言理解任务中均优于现有大语言模型，尤其是在阅读理解、自然语言推理和语义匹配任务中（表21）。

[1]提示的详细信息见附录B

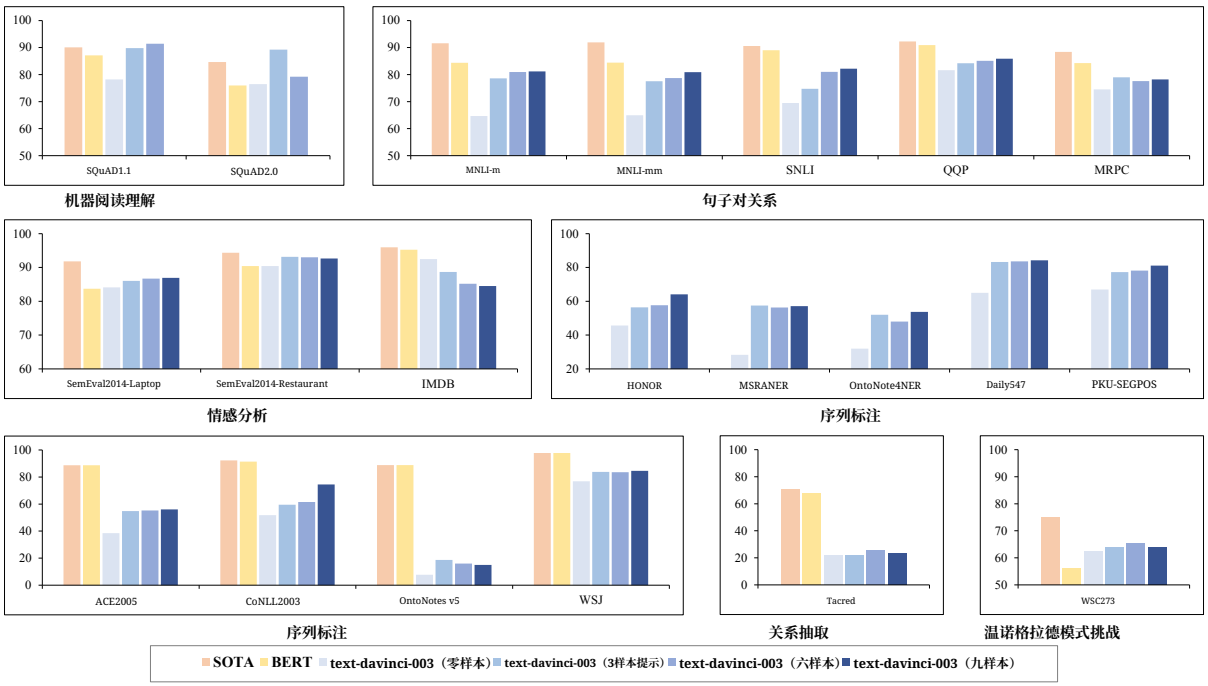

图1：GPT-3.5 在零样本和少样本（单样本、3样本提示、六样本、九样本）场景下、BERT-base模型以及使用任务特定数据微调的模型在 21 个不同数据集上的表现。对于每个数据集，显示所有提示中表现较好的结果。结果详情见附录A。

• **上下文学习能力：**与零样本相比，GPT-3.5 在单样本设置下的大多数任务中表现更好。此外，在大多数任务中，单样本、3样本提示、六样本和九样本设置下的表现差异并不显著。不过，在提示中提供更多示例可能有助于序列标注任务（图1和第4节.5）。

•**GPT模型之间的差异：**在大多数自然语言理解任务中，GPT-3.5模型（*text-davinci-003*、*text-davinci-002*）相比GPT-3模型（*text-davinci-001*）有显著提升，尤其是在阅读理解、自然语言推理和序列标注等需要较高语言理解能力的任务中。

根据上述发现，我们得出以下三个结论：

• **模型理解能力的来源：**预训练、代码训练和指令微调阶段的结合，对增强模型的自然语言理解能力起着决定性作用。我们推测，预训练为模型奠定了语义理解的基础，代码训练增强了模型理解语义依赖的能力，而指令微调则提升了模型对新任务的泛化能力。

• **影响鲁棒性的因素：**在指令微调阶段，对指令和任务采用监督训练，可能是导致出现类似预训练-微调范式所面临的鲁棒性问题的关键因素。因此，除了考察任务的鲁棒性之外，还必须进一步研究指令的鲁棒性。

• **任务标签与标签类型的一致性：**在指令微调阶段和实际应用阶段确保任务标签及其类型保持一致，对于模型的性能和泛化能力至关重要。因此，设计有效的提示或指令，以确保这种一致性并适应不同的标签类型和应用场景，非常重要。

表1：实验中使用的所有数据集信息。

Task	子任务	数据集	样本数	评价指标	语言
机器阅读理解	机器阅读理解	SQuAD1.1	9868	F1值和完全匹配	英语
		SQuAD2.0	11491	F1值与完全匹配	英语
关系抽取	关系抽取	Tacred	15509	F1	英语
句子对关系	自然语言推理	MNLI-m	9815	准确率	英语
		MNLI-mm	9832	准确率	英语
		SNLI	10000	准确率	英语
	语义匹配	MRPC	1724	准确率	英语
		QQP	5000	准确率	英语
情感分析	基于方面的情感分析	SemEval2014-Laptop	331	准确率	英语
		SemEval2014-Restaurant	492	准确率	英语
	情感分类	IMDB	11257	准确率	英语
序列标注	命名实体识别	ACE 2005	1312	F1	英语
		CoNLL 2003	3453	F1	英语
		OntoNotes v5	4019	F1	英语
		HONOR	1120	F1	中文
		MSRANER	4365	F1	中文
		OntoNote4NER	4346	F1	中文
	词性标注	Daily547	546	准确率	英语
		WSJ	5461	准确率	英语
		PKU-SEGPOS	5204	F1	中文
温诺格拉德模式挑战	温诺格拉德模式挑战	WSC273	570	准确率	英语

2 背景与相关工作

2.1 大型语言模型

近年来，大语言模型在自然语言处理研究中日益普及。这些模型的发展得益于多项突破，例如拥有 1750 亿参数的 GPT-3 模型（Brown 等人，2020年）；而 FLAN（Wei 等人，2022年）、T0（Sanh 等人，2021年）、LaMDA（Thoppilan 等人，2022年）和 PaLM（Chowdhery 等人，2022年）等领域进展，也为自然语言处理的发展作出了重要贡献。近期，GPT-3.5 和 ChatGPT[2] 的发布引起了广泛关注，它们能够在多个知识领域给出详尽的回应和流畅的答案（Guo 等人， 2023; Qin 等人，2023年）；这些模型基于 GPT-3.5系列，并通过基于人类反馈的强化学习（RLHF）进行微调（Christiano 等人，2017年）。尽管这些大语言模型潜力巨大，但以往研究（Hendy 等人， 2023; Bang 等人，2023年）主要聚焦于探索它们在各类任务中的表现，很少关注其鲁棒性。鲁棒性能够衡量模型应对开放世界各种复杂情况的能力，也是可信人工智能的关键维度（Liu 等人，2022年）。因此，本文对 GPT-3.5 的鲁棒性进行了全面的实验分析。

2.2 TextFlint

在本研究中，我们利用 TextFlint（Gui 等，2021）这一专为自然语言处理模型设计的鲁棒性评估工具包，开展全面的鲁棒性评估。TextFlint 提供了一系列文本变换，包括通用文本变换、特定任务变换、对抗攻击、子群体及其组合，用于向文本数据引入各种类型的扰动。我们使用 TextFlint 中的 66 种文本变换进行鲁棒性评估。

3 实验设置

3.1 任务和数据集

3.2 任务和数据集

我们使用涵盖 9 个常见自然语言理解任务类别的 21 个数据集对 GPT-3.5 进行评估：机器阅读理解 （**MRC**）（SQuAD1.1（Rajpurkar 等人，2016年）、SQuAD2.0（Rajpurkar 等人，2018年））、关系抽取 （**RE**）（Tacred（Zhang 等人，2017年））、自然语言推理（**NLI**）（MNLI-m（Williams 等人，2017年）、MNLI-mm（Williams 等人，2017年）、SNLI（Williams 等人，2017年））、语义匹配（**SM**）（MRPC（Dolan 和 Brockett，2005年）、QQP（Wang 等人，2017年））、基于方面的情感分析 （**ABSA**）（SemEval2014-Laptop（Pontiki 等人，2014年）、SemEval2014-Restaurant（Pontiki 等人，2014年））、情感分析 （**SA**）（IMDB（Maas 等人，2011年））、命名实体识别（**NER**）（ACE2005[3], CoNLL2003（Sang 和 Meulder，2003年）、OntoNotesv5[4], HONOR[5], MSRANER（Levow，2006年）、OntoNote4NER（Weischedel 等人，2011年））、词性标注（**POS**）（Daily547（Gimpel 等人，2010年）、WSJ（Marcus 等人，1993年）、PKU-SEGPOS[6] ）以及Winograd图式挑战（**WSC**）（WSC273（Levesque 等人，2012年））。为评估模型的鲁棒性和性能，我们在这些任务中共应用了 66 种来自 TextFlint 的文本变换。有关实验数据集的更多细节见表1。

3.3 模型选择

在本研究中，我们的主要目标是评估 GPT-3.5 模型在上述自然语言理解任务中的性能和鲁棒性。为了将结果与早期 GPT 模型进行比较，我们特意选择了 GPT-3.5 系列模型中的*text-davinci-002* 和 GPT-3 系列模型中的*text-davinci-001* 。此外，我们还将结果与现有监督训练模型和零样本/少样本模型的结果进行了比较（第 4.7.2 节）。

3.4 零样本/少样本场景设置

对于每个数据集，我们分别在零样本和少样本场景下评估模型的性能。我们设置了四种不同的场景：1/3/6/9样本。对于单样本和3样本提示，我们使用各种特定任务变换测试整个数据集。对于六样本和九样本，我们仅在原始数据集中评估1000个样本，以评估提示中的示例数量对性能的影响。更多详情请参阅第4节、图1和附录A。

3.5 提示选择

由于大语言模型依赖不同的提示来执行下游任务，我们广泛收集了大量特定任务提示 [7]。我们针对某些任务手动选择并设计新的提示，最终为每个数据集选取表现最佳的三个提示，以获得最客观的结果。通过改变提示中示例的数量，我们将这三个提示 [8] 扩展为五种不同的示例数量级别（0/1/3/6/9样本）。关于这些提示的更多详情请参见附录B。请注意，在第4节中，少样本场景下不同示例数量对应的记录结果，表示由相应的三个提示获得的结果的均值和标准差。

4 实验

本节中，我们展示实验结果，并对每项任务进行分析。

[2]尽管缺乏 ChatGPT 的官方应用程序接口，我们计划在未来的研究中继续探索该模型。

[3]https://catalog.ldc.upenn.edu/LDC2006T06; [4]https://catalog.ldc.upenn.edu/LDC2013T19; [5]我们自行构建了 HONOR 数据集，其重点是与日程安排相关的命名实体识别。

[6]http://cuge.baai.ac.cn/\#/dataset?id=19\&name=PKU-SEGPOS; [7]部分提示来源于 https://github.com/bigscience-workshop/promptsource. [8]在 RE 和 ST任务中，由于原始标注是缩写，我们将其映射为提示中的特定短语，以便模型理解其含义。

4.1 机器阅读理解

表 2：GPT系列模型（零样本和少样本）以及微调模型（即 BERT（Devlin等人，2019）、ALBERT（Lan 等，2019）、DistilBERT（Sanh 等人，2019）、XLNet（Yang等，2019）、LUKE（Yamada 等人，2020）、SpanBERT（Joshi 等人，2020）和 DeBERTa（He 等，2020））在**SQuAD1.1** 数据集上的鲁棒性测试结果，评估指标为 micro-F1分数。

模型	AddSentDiverse *# 9292* 个样本		ModifyPos *# 9011* 个样本		PerturbAnswer *# 9833* 个样本		PerturbQuestion-BackTranslation *# 9868* 个样本		PerturbQuestion-MLM *# 9867* 个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	66.84±9.79	55.59±9.95	67.55±9.84	67.46±8.94	67.16±9.71	65.97±8.27	67.19±9.66	59.90±8.00	67.18±9.64	56.43±9.15
单样本	88.20±1.30	**70.75±1.98**	88.47±1.24	88.37±1.00	88.13±1.27	85.27±1.11	88.10±1.29	80.31±1.19	88.13±1.26	78.34±1.70
3样本提示	**89.64±0.40**	68.04±0.85	**89.94±0.35**	**89.88±0.38**	**89.60±0.37**	**86.92±0.44**	**89.57±0.38**	**82.10±0.55**	**89.57±0.38**	**79.94±0.90**
text-davinci-002										
零样本	78.56±11.01	55.67±8.01	77.77±11.76	77.10±11.19	78.52±10.86	72.21±10.85	78.44±10.84	69.04±9.52	78.47±10.94	64.47±10.43
单样本	58.82±19.61	47.43±14.23	58.19±19.65	57.48±19.94	58.42±19.04	52.35±15.05	58.62±19.10	52.43±16.37	58.3±18.93	49.55±18.01
3样本提示	82.72±7.87	61.97±2.90	83.37±8.13	82.65±8.01	82.80±7.74	80.98±7.24	82.93±7.66	74.96±7.82	82.93±7.70	71.64±10.81
text-davinci-001										
零样本	73.38±11.70	51.64±14.23	72.45±19.65	72.53±19.94	72.93±19.04	65.89±15.05	73.17±19.10	62.47±16.37	73.09±18.93	61.27±18.01
单样本	85.01±2.11	67.67±0.97	84.83±1.74	84.28±2.68	85.00±2.03	79.28±2.11	85.11±1.90	73.43±1.54	85.02±1.90	73.02±1.36
3样本提示	85.41±0.87	65.86±1.10	85.49±1.31	85.33±2.21	85.73±0.39	80.17±0.61	85.30±0.95	74.28±0.97	85.37±0.85	74.81±0.68
微调										
BERT	87.09	32.47	87.68	87.25	87.13	75.40	87.09	79.37	87.09	79.39
ALBERT	91.27	40.45	91.76	90.82	91.26	80.52	91.23	82.73	91.23	75.59
DistilBERT	87.10	29.60	87.56	86.69	87.04	74.92	87.00	78.50	87.01	78.93
XLNet	89.50	37.48	89.81	88.94	89.45	80.15	89.44	82.08	89.43	78.14
LUKE	**95.15**	**57.46**	**95.57**	**94.93**	**95.14**	**87.43**	**95.13**	**87.93**	**95.13**	76.44
SpanBERT	94.19	50.90	94.63	94.17	94.15	86.56	94.13	86.49	94.13	**84.52**
DeBERTa	93.69	46.44	94.07	93.49	93.66	85.89	93.64	86.30	93.64	75.81

表3：GPT系列模型（零样本和少样本）及微调模型（即 BERT、ALBERT、DistilBERT、XLNet、LUKE、SpanBERT 和 DeBERTa）在**SQuAD1.1** 数据集上的鲁棒性测试结果，使用 EM分数作为评估指标。

模型	AddSentDiverse *# 1312*个样本		ModifyPos *# 1312*个样本		PerturbAnswer *# 1405*个样本		PerturbQuestion-BackTranslation *# 1312*个样本		PerturbQuestion-MLM *# 1312*个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	42.76±15.43	34.01±15.57	43.64±15.58	43.71±14.66	43.37±15.34	42.58±13.32	43.37±15.31	36.73±12.78	43.34±15.31	34.05±13.79
单样本	74.51±3.13	**57.93±3.08**	75.03±3.00	75.26±2.60	74.55±3.07	70.66±2.65	74.47±3.07	65.54±2.80	74.50±3.04	62.86±3.14
3样本提示	**77.85±1.24**	57.53±0.46	**78.41±1.15**	**78.61±1.19**	**77.93±1.18**	74.29±1.19	**77.86±1.15**	**69.04±1.49**	**77.86±1.19**	**66.23±1.75**
text-davinci-002										
零样本	65.33±15.96	43.13±12.76	63.57±17.25	62.80±16.62	65.27±15.61	58.07±15.88	65.33±15.61	55.27±13.57	65.30±15.70	50.50±13.65
单样本	51.20±19.81	40.97±14.12	50.53±19.99	49.87±20.40	50.63±19.27	44.57±14.77	50.9±19.33	44.5±15.96	50.47±19.17	41.27±18.21
3样本提示	76.60±8.77	56.5±3.50	77.23±9.03	76.37±8.61	76.73±8.78	**74.77±7.39**	76.93±8.66	67.77±8.14	76.87±8.70	64.87±11.20
text-davinci-001										
零样本	58.17±16.43	38.40±16.18	56.67±16.95	57.07±15.46	57.63±16.28	50.03±15.24	57.87±16.34	47.87±13.25	57.73±16.09	47.57±13.37
单样本	73.43±3.39	57.43±0.76	73.27±2.85	73.17±4.00	73.43±3.20	66.97±2.52	73.50±3.02	62.07±2.65	73.50±3.06	62.00±2.26
3样本提示	76.05±1.41	57.78±1.23	76.16±2.05	76.23±3.42	76.5±1.37	70.35±1.33	75.97±1.74	65.00±1.35	76.07±1.52	65.86±1.05
微调										
BERT	79.25	27.93	79.95	79.81	79.30	62.48	79.25	70.16	79.24	70.38
ALBERT	84.70	35.87	85.31	84.24	84.63	68.80	84.56	74.44	84.57	61.04
DistilBERT	79.43	25.53	79.96	79.10	79.35	62.21	79.30	69.26	79.31	70.08
XLNet	81.37	32.12	81.79	81.13	81.30	67.15	81.23	72.72	81.23	65.55
LUKE	**90.21**	**52.73**	**90.73**	**90.01**	**90.08**	**77.00**	**90.06**	**81.12**	**90.07**	60.19
SpanBERT	88.64	46.72	89.01	88.50	88.47	75.38	88.42	78.88	88.42	**76.50**
DeBERTa	88.29	42.24	88.76	88.08	88.18	75.10	88.10	79.18	88.10	59.94

表4：GPT系列模型（零样本和少样本）及微调模型（即 BERT、XLNet 和 DeBERTa）在 **斯坦福问答数据集2.0** 上的鲁棒性测试结果，采用 micro-F1分数作为评估指标。

模型	AddSentDiverse *# 5129*个样本		ModifyPos *# 5053*个样本		PerturbAnswer *# 5522*个样本		PerturbQuestion-BackTranslation *# 11492*个样本		PerturbQuestion-MLM *# 11491*个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	65.42±9.22	54.21±9.96	66.31±9.45	66.16±8.65	65.87±9.20	65.19±7.96	65.96±9.17	58.09±7.34	66.00±9.11	54.87±8.62
单样本	87.50±1.26	**69.80±2.07**	88.01±1.26	87.80±1.14	87.34±1.16	84.79±1.07	87.24±1.30	79.16±1.14	87.36±1.24	77.04±1.67
3样本提示	**88.88±0.38**	67.08±0.90	**89.29±0.37**	**89.41±0.47**	**88.83±0.43**	**86.36±0.42**	**88.88±0.36**	**80.89±0.51**	**88.87±0.46**	**78.87±1.03**
text-davinci-002										
零样本	69.31±12.21	48.92±14.03	69.05±12.28	68.67±12.56	68.69±12.14	63.83±11.56	67.36±11.41	56.78±9.97	67.50±11.23	55.52±13.31
单样本	61.16±17.04	48.01±12.27	61.14±16.83	59.61±16.83	61.21±16.64	53.20±15.28	63.73±14.55	55.95±13.82	63.67±14.47	53.01±14.75
3样本提示	84.44±5.89	57.83±2.05	83.64±6.19	81.63±7.04	83.27±5.90	77.73±7.09	84.83±4.34	73.77±5.07	84.67±4.69	71.42±7.39
text-davinci-001										
零样本	65.36±13.26	49.07±15.05	65.22±12.79	65.55±12.61	64.60±13.07	59.56±11.98	62.61±12.37	52.86±8.30	62.67±12.33	50.51±10.25
单样本	80.02±1.33	63.38±4.22	79.96±1.75	77.94±1.67	79.15±1.28	74.07±1.47	76.99±1.03	63.77±0.97	77.00±1.18	62.93±1.50
3样本提示	81.97±1.67	60.88±1.33	81.53±1.65	80.49±2.60	80.63±1.58	75.58±0.85	77.99±1.47	66.04±2.09	77.94±1.52	64.38±0.83
微调										
BERT	80.38	**34.33**	81.40	79.70	80.39	54.74	75.95	67.95	75.95	64.13
SpanBERT	**89.25**	30.85	**89.93**	**88.26**	**89.10**	**67.22**	87.67	**77.68**	87.67	**68.71**
DeBERTa	87.90	28.38	88.52	86.28	87.76	63.19	**87.72**	69.16	**87.72**	56.60

表5：GPT系列模型（零样本和少样本）以及微调模型（即BERT、XLNet和SpanBERT）在 **SQuAD2.0** 数据集上的鲁棒性测试结果，使用EM分数作为评估指标。

模型	AddSentDiverse *# 1312*个样本		ModifyPos *# 1312*个样本		PerturbAnswer *# 1405*个样本		PerturbQuestion-BackTranslation *# 1312*个样本		PerturbQuestion-MLM *# 1312*个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	40.55±14.61	31.78±15.33	41.63±15.09	41.54±14.10	41.34±14.56	41.07±12.80	41.50±14.42	34.33±11.71	41.43±14.52	31.76±13.05
单样本	73.00±3.19	**56.06±3.08**	73.95±3.16	74.08±2.88	72.73±2.98	69.40±2.60	72.83±3.42	63.54±2.93	73.04±3.25	60.72±3.07
3样本提示	**76.2±1.29**	55.7±0.47	**76.96±1.17**	**77.57±1.43**	**76.29±1.25**	**72.95±1.24**	**76.42±1.17**	**67.01±1.46**	**76.43±1.33**	**64.19±1.94**
text-davinci-002										
零样本	50.73±17.72	33.03±17.64	50.13±17.98	49.70±18.20	49.83±17.31	44.47±16.12	49.04±15.81	40.26±14.91	49.11±15.68	38.00±16.24
单样本	49.87±16.57	38.13±12.02	49.67±16.31	48.27±16.34	49.70±16.10	41.50±13.64	51.51±14.46	43.62±13.46	51.51±14.33	40.40±14.94
3样本提示	74.87±6.59	49.6±2.60	73.57±6.83	71.57±7.80	73.03±6.54	66.80±7.30	74.07±5.41	62.83±6.49	73.87±5.64	59.19±7.75
text-davinci-001										
零样本	45.60±18.09	31.87±18.18	45.37±17.68	46.07±17.56	44.67±17.72	39.73±15.84	43.21±16.90	35.46±12.67	43.21±16.90	32.92±12.81
单样本	65.43±2.02	50.03±2.97	64.93±2.26	63.67±2.31	63.93±2.05	58.83±1.81	61.66±1.37	49.66±1.73	61.59±1.66	48.15±1.03
3样本提示	69.75±2.42	49.47±0.67	68.91±2.49	68.27±3.15	67.87±2.22	62.97±1.45	65.16±1.38	54.44±1.99	65.09±1.73	51.45±1.10
微调										
BERT	73.66	**30.55**	74.73	72.93	73.63	45.89	72.67	64.68	72.67	61.21
SpanBERT	**83.04**	28.15	**83.81**	**82.27**	**82.87**	**58.31**	**84.62**	**74.55**	**84.62**	**66.01**
DeBERTa	82.30	26.17	82.90	80.69	82.00	55.00	81.88	63.40	81.90	41.53

对于机器阅读理解任务，我们比较了斯坦福问答数据集SQuAD1.1和SQuAD2.0数据集上，0/1/3次提示学习的GPT系列模型与微调模型在不同鲁棒性测试中的微平均 F1和完全匹配分数（表 2～5）。从表中可以观察到以下特点。**少样本模型在性能和鲁棒性方面与微调模型不相上下**。这一结果表明，少样本学习能够在保持高性能的同时提升模型的鲁棒性。值得注意的是，在SQuAD1.1上针对AddSentDiverse和PerturbAnswer变换进行测试时，GPT系列模型的鲁棒性显著优于微调模型。**提示中的样例数量对鲁棒性和性能的影响并不一致**。0/1/3样本模型的鲁棒性基本保持不变，但与零样本模型相比，1/3样本模型的微平均 F1和完全匹配分数显著提升。这表明，少样本实例可以帮助模型学习当前的阅读理解模式，并增强其在该任务中的阅读理解能力。少样本模型在不同提示之间的表现差异也较小，且受数据和提示变化的影响较小，凸显了它们在机器阅读理解任务中的稳定性。

4.2 关系抽取

表6：GPT系列模型（零样本和少样本）与微调模型（即 BERT-base-uncased（Devlin 等人，2019）、SpanBERT 和 NLL（Zhou和Chen，2021））在**Tacred**数据集上的鲁棒性测试结果。

(a)

模型	插入子句 *# 14897* 个样本		交换实体-低频 *# 15509* 个样本		交换实体-多类型 *# 15509* 个样本		交换实体-相同类型 *# 15509* 个样本	
	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003								
零样本	20.37±1.32	18.76±0.88	20.40±1.33	21.77±1.58	20.36±1.24	20.23±1.02	20.36±1.28	22.01±1.63
单样本	**23.20±1.63**	**22.22±1.21**	**23.24±1.71**	**23.65±1.72**	**23.25±1.68**	**21.91±1.44**	**23.22±1.72**	**24.06±1.70**
3样本提示	22.01±0.11	20.89±0.34	21.98±0.18	21.53±0.29	21.94±0.17	19.24±0.73	21.96±0.20	22.14±0.20
text-davinci-002								
零样本	12.34±0.30	11.50±0.43	12.23±0.09	9.19±0.69	12.18±0.17	9.29±0.81	12.23±0.09	10.70±0.17
单样本	17.25±0.67	14.68±0.68	17.03±0.61	15.14±0.13	17.20±0.61	12.50±0.10	17.20±0.45	15.02±0.87
3样本提示	17.13±1.14	14.48±1.36	17.14±1.14	14.81±0.35	16.85±1.48	12.21±1.13	17.19±1.33	16.11±1.53
text-davinci-001								
零样本	10.29±0.73	9.27±0.75	10.05±0.88	11.08±0.91	10.15±0.73	9.88±1.09	10.18±0.81	10.43±0.35
单样本	10.57±0.61	10.16±0.93	10.55±0.60	11.62±1.55	10.66±0.66	10.25±2.08	10.50±0.56	11.68±1.45
3样本提示	10.10±0.42	9.48±0.19	10.15±0.33	10.48±0.72	9.87±0.19	7.44±0.19	9.99±0.34	10.05±0.66
微调								
BERT-base-uncased	68.01	50.95	68.01	59.16	68.01	48.66	68.01	61.91
SpanBERT	70.81	55.79	70.81	70.81	**70.81**	**70.81**	70.81	70.81
NLL	**70.94**	**57.13**	**70.87**	**70.87**	-	-	**70.87**	**70.87**

(b)

模型	SwapTriplePos-Age *# 28* 个样本		交换三元组位置-出生 *# 48* 个样本		交换三元组位置-员工 *# 251* 个样本	
	ori	转换	ori	转换	ori	转换
text-davinci-003						
零样本	96.43±0.00	**100.00±0.00**	65.28±1.21	71.53±1.21	6.53±2.36	16.37±3.96
单样本	**100.00±0.00**	**100.00±0.00**	62.74±5.92	68.31±2.83	10.88±1.07	16.51±2.26
3样本提示	**100.00±0.00**	**100.00±0.00**	**75.00±2.08**	**81.25±0.00**	19.27±3.72	25.02±4.99
text-davinci-002						
零样本	89.29±6.19	88.10±2.07	56.25±8.33	52.08±4.17	16.47±7.02	17.40±8.52
单样本	96.61±0.31	**100.00±0.00**	62.50±2.08	65.97±3.18	47.54±6.79	52.32±6.68
3样本提示	**100.00±0.00**	**100.00±0.00**	65.97±13.39	63.89±7.89	**60.03±9.42**	**62.02±6.86**
text-davinci-001						
零样本	**100.00±0.00**	**100.00±0.00**	33.90±11.69	36.64±7.31	0.41±0.41	1.22±1.23
单样本	**100.00±0.00**	**100.00±0.00**	32.05±9.63	35.26±10.15	1.44±1.46	1.78±1.47
3样本提示	**100.00±0.00**	**100.00±0.00**	47.91±9.55	48.61±7.89	0.80±1.38	0.66±1.15
微调						
BERT-base-uncased	**100.00**	92.31	82.11	67.42	91.75	**78.28**
SpanBERT	**100.00**	98.18	95.83	**94.74**	**93.14**	77.78
NLL	**100.00**	**100.00**	**96.84**	93.62	-	-

对于关系抽取任务，我们使用Tacred作为测试数据集，并在表6中展示了不同鲁棒性测试下的微平均F1分数。从表中可以得出以下观察结果。**GPT系列模型似乎不足以胜任关系抽取这类复杂的语义理解任务**。具体而言，当数据量较大时，微调模型的微平均F1分数可达到70%，而GPT系列模型的最佳表现仅约为20%，因此几乎不可能将其用于关系抽取任务。**然而，对于某些特定的关系判断，GPT系列模型表现出色，且鲁棒性很强**。尤其是在人物与年龄关系（即SwapTriplePos-Age）上，GPT系列模型的表现非常准确，对变换后的句子的识别效果甚至更好。经过分析，我们发现，数据集中此类特定关系对应句子的句法结构非常固定，而变换后的句子通常对应常见表达，为模型提供了线索，使其能够做出预测。

4.3 句子对关系

我们在两个子任务上测试句子对关系任务：自然语言推理（第4.3.1节）和语义匹配（第4.3.2节）。

4.3.1 自然语言推理

表 7：GPT系列模型（零样本和少样本）与微调模型（即 BERT-base-uncased、RoBERTa-large（Liu 等，2019）、ALBERT-xxlarge-v2（Lan 等，2019）、DeBERTa-v2-xxlarge（He 等，2020））在 **MNLI-m** 数据集上的鲁棒性测试结果。

模型	AddSent *# 9815* 个样本		NumWord *# 745* 个样本		交换反义词 *# 199* 个样本	
	ori	转换	ori	转换	ori	转换
			text-davinci-003			
零样本	64.26±0.53	34.04±1.85	68.12±1.42	**51.62±13.17**	70.11±1.57	78.96±6.20
单样本	**73.14±4.80**	41.07±5.05	**74.43±4.90**	43.80±3.61	93.35±1.66	72.87±5.45
3样本提示	72.07±5.69	41.02±3.14	71.59±6.01	46.49±4.12	**98.66±0.58**	82.41±5.10
			text-davinci-002			
零样本	52.62±6.64	36.27±1.61	52.72±5.83	29.21±14.81	52.27±15.07	50.95±19.87
单样本	68.60±6.20	37.20±1.10	69.44±9.26	45.37±14.51	90.29±1.05	86.26±10.18
3样本提示	70.30±5.47	36.87±2.12	71.18±6.00	44.61±11.68	97.49±1.51	**88.27±12.51**
			text-davinci-001			
零样本	42.20±4.15	36.66±2.06	38.91±3.03	25.76±8.35	42.41±23.67	28.89±23.00
单样本	39.90±8.01	39.40±7.63	40.44±7.19	1.39±2.40	98.16±3.19	4.36±6.28
3样本提示	48.80±3.64	**43.87±3.91**	50.16±6.62	4.34±5.34	91.79±7.33	28.64±16.34
			微调			
BERT-base-uncased	84.39	55.32	83.09	49.40	91.46	52.76
RoBERTa-large	90.61	70.95	88.59	**53.69**	**93.47**	70.35
ALBERT-xxlarge-v2	89.45	**79.74**	86.85	49.53	90.95	68.84
DeBERTa-v2-xxlarge	**91.60**	71.68	**91.14**	48.46	92.46	**78.89**

表 8：GPT系列模型（零样本和少样本）与微调模型（即 BERT-base-uncased、RoBERTa-large、ALBERT-xxlarge-v2、DeBERTa-v2-xxlarge）在**MNLI-mm**数据集上的鲁棒性测试结果。

模型	AddSent *# 9832* 个样本		NumWord *# 775* 个样本		交换反义词 *# 255* 个样本	
	ori	转换	ori	转换	ori	转换
			text-davinci-003			
零样本	64.56±0.32	34.26±2.15	67.50±0.25	39.81±3.04	67.76±0.54	77.65±2.71
单样本	70.28±5.21	37.01±3.34	71.39±4.39	28.45±1.80	93.20±2.94	66.80±6.64
3样本提示	**73.66±3.44**	37.92±1.28	**74.99±2.36**	41.59±4.71	**96.34±1.20**	**89.61±2.90**
			text-davinci-002			
零样本	49.93±6.44	36.05±2.27	52.93±7.57	28.45±14.29	51.40±15.02	58.39±23.61
单样本	51.12±18.88	35.20±6.03	52.96±20.00	16.95±7.03	55.30±36.10	51.29±18.91
3样本提示	61.81±11.09	34.79±0.70	61.40±14.05	**49.71±20.48**	70.67±19.95	83.12±26.82
			text-davinci-001			
零样本	44.07±3.54	35.72±3.46	45.34±6.09	28.11±9.66	33.42±18.59	31.07±18.61
单样本	47.54±2.85	44.87±2.63	45.22±2.84	11.10±5.54	91.76±8.38	36.69±6.32
3样本提示	54.43±3.50	**49.03±10.00**	52.16±5.08	7.87±10.58	89.41±9.90	32.29±23.66
			微调			
BERT-base-uncased	84.43	55.25	82.97	46.06	85.10	55.69
RoBERTa-large	90.12	67.75	88.65	**52.90**	92.16	74.90
ALBERT-xxlarge-v2	89.89	**79.10**	89.03	44.77	91.76	69.80
DeBERTa-v2-xxlarge	**91.90**	72.50	**91.74**	47.23	**94.51**	**76.08**

表9：GPT系列模型（零样本和少样本）与微调模型（即 BERT-base-uncased、BERT-large-uncased（Devlin等人，2019）、XLNet-large-cased（Yang等，2019））在**SNLI** 数据集上的鲁棒性测试结果。

模型	AddSent *# 10000* 个样本		NumWord *# 108* 个样本		交换反义词 *# 523* 个样本	
	ori	转换	ori	转换	ori	转换
			text-davinci-003			
零样本	67.16±3.26	34.11±0.03	63.89±2.45	73.77±22.76	81.39±8.06	61.44±17.44
单样本	71.81±4.12	34.63±0.26	73.30±1.63	54.63±6.68	**97.77±0.67**	49.46±8.36
3样本提示	72.18±2.37	40.00±2.68	**75.00±3.34**	67.29±7.54	94.65±2.20	45.41±7.14
			text-davinci-002			
零样本	47.40±8.24	35.45±2.04	45.98±11.81	30.84±25.56	58.76±41.98	27.89±20.28
单样本	72.20±3.04	**40.70±3.72**	71.60±2.98	46.91±10.65	94.45±4.45	60.48±13.19
3样本提示	**72.43±3.11**	37.83±2.73	70.68±2.97	63.27±15.28	94.39±3.75	66.35±11.72
			text-davinci-001			
零样本	38.61±7.26	33.00±3.80	48.3±26.79	46.42±30.83	24.75±15.46	63.38±37.26
单样本	36.27±0.31	34.50±0.61	41.97±2.33	62.66±11.50	2.10±3.48	38.24±22.93
3样本提示	44.60±4.45	34.10±0.78	47.33±6.43	**87.94±11.21**	41.27±26.07	**89.68±6.84**
			微调			
BERT-base-uncased	88.99	**79.66**	**91.67**	58.33	**91.59**	94.84
BERT-large-uncased	89.37	58.34	**91.67**	67.59	90.63	94.84
XLNet-large-cased	**90.60**	72.97	91.01	**91.67**	62.96	**95.79**

我们在MNLI-m、MNLI-mm和SNLI数据集上测试自然语言推理任务，并得到表7、表8和表9。评估结果显示，**在不同类型的变换中，鲁棒性有所不同，且提示中示例数量的选择也会对结果产生一定影响**。例如，对于MNLI-m和MNLI-mm数据集（表7和表8），与微调模型相比，AddSent变换在0/1/3样本模型中表现出略低的鲁棒性，而NumWord变换的鲁棒性则相当。有趣的是，与微调模型相比，SwapAnt变换在0/1/3样本模型中展现出显著更好的鲁棒性。

4.3.2 语义匹配

表10：GPT系列模型（零样本和少样本）以及微调模型（即 BERT-base-uncased、Funnel-Transformer-medium（Dai等人，2020年）和 DeBERTa-base（He 等，2020））在**MRPC** 数据集上的鲁棒性测试结果。

模型	all *# 1724* 个样本	NumWord *# 402* 个样本		交换反义词 *# 158* 个样本	
	ori	ori	转换	ori	转换
		text-davinci-003			
零样本	70.17±4.51	**74.63±1.97**	94.44±3.90	75.11±8.26	54.22±10.41
单样本	69.50±5.41	69.82±5.31	98.26±1.88	62.02±12.80	69.62±4.56
3样本提示	68.50±10.26	67.99±7.34	97.51±1.63	60.55±17.15	68.78±7.42
		text-davinci-002			
零样本	**72.73±2.55**	68.41±6.24	66.67±35.79	**95.57±5.18**	36.29±18.66
单样本	69.57±8.35	72.31±7.04	**98.59±1.65**	64.14±14.24	78.69±1.93
3样本提示	72.70±3.57	73.14±2.60	96.10±6.53	66.45±5.80	**85.86±9.69**
		text-davinci-001			
零样本	21.60±19.25	17.58±21.19	17.08±17.04	22.79±27.59	13.71±11.91
单样本	70.40±1.87	65.84±3.45	78.44±8.76	89.66±2.86	49.16±7.63
3样本提示	53.80±4.41	56.05±2.91	98.42±1.12	50.00±9.74	75.11±5.74
		微调			
BERT-base-uncased	84.28	**91.14**	0.00	82.59	0.00
Funnel-Transformer-medium	87.07	**91.14**	**1.90**	86.32	0.25
DeBERTa-base	**88.40**	**91.14**	0.00	**88.31**	**0.50**

表11：GPT系列模型（零样本和少样本）及微调模型（即 BERT-base-uncased、ALBERT-xxlarge-v2 和 Electra-base （ Clark 等人，2020 年））在 **QQP** 数据集上的鲁棒性测试结果。

模型	all *# 5000*个样本 ori	NumWord *# 2607*个样本 ori	转换	交换反义词 *# 883*个样本 ori	转换
		text-davinci-003			
零样本	81.03±0.67	79.85±1.37	73.22±19.84	60.44±8.78	65.31±5.15
单样本	80.93±1.91	79.81±1.69	73.98±20.42	63.42±9.92	51.34±7.89
3样本提示	**82.97±1.72**	**83.35±0.14**	71.06±15.50	71.46±9.13	56.21±8.66
		text-davinci-002			
零样本	63.00±5.69	68.07±5.70	25.00±19.56	**85.32±13.47**	28.43±20.73
单样本	77.70±3.05	79.50±7.11	**90.17±24.47**	47.41±12.95	67.87±6.12
3样本提示	80.30±1.15	82.60±6.87	87.13±19.29	57.53±14.28	75.27±6.91
		text-davinci-001			
零样本	36.40±5.82	35.37±2.99	6.07±9.48	75.65±23.45	14.01±19.48
单样本	66.50±3.35	66.17±3.43	86.47±11.10	26.61±23.80	**90.79±11.85**
3样本提示	65.17±5.29	72.60±6.48	72.9±12.68	67.01±6.18	76.48±9.44
		微调			
BERT-base-uncased	90.91	89.92	48.58	93.59	56.96
ALBERT-xxlarge-v2	**92.28**	91.05	68.40	**94.51**	53.47
Electra-base	91.72	**91.39**	**69.99**	93.86	**68.93**

我们在 MRPC 和 QQP 数据集上测试句子匹配任务，得到表10和表11。从表中，我们观察到一些有趣的现象。**与微调模型相比，GPT-3.5的鲁棒性显著提高，NumWord和SwapAnt变换的结果证明了这一点**。零样本、单样本和3样本提示模型的鲁棒性显著优于微调模型，尤其是在 MRPC 数据集上（表10）。这可能是因为不同的提示往往会产生不同的输出。在特定的变换条件下，例如 SwapAnt变换，原始样本全部为蕴含的情况会变成全部矛盾，从而产生这样的结果。**然而，不同变换和实例数量对该数据集的影响并不显著**。单样本和3样本提示模型的鲁棒性优于零样本模型。此外，单样本与3样本提示模型的鲁棒性差异很小，这表明提供示例对可以有效提高模型在该数据集上的鲁棒性，但示例数量对结果影响不大。单样本和3样本提示模型在 MRPC 数据集上经 NumWord变换后的数据上的性能显著提升，准确率达到100%。值得一提的是，**GPT系列模型在 MRPC 数据集上的鲁棒性显著优于微调模型**。微调模型在 MRPC 数据集的 NumWord 和 SwapAnt变体上鲁棒性较差，变换后数据的准确率通常接近0。然而，text-davinci-003、002和001模型均未出现此类问题。这表明 GPT系列模型在该任务中更擅长处理文本变体和异常情况。

4.4 情感分析

我们专门在两个子任务上测试情感分析任务：ABSA（第4.4.1节）和 SC（第4.4.2节）。

4.4.1 基于方面的情感分析

表12：GPT系列模型（零样本和少样本）及微调模型（即 BERT-base-uncased、LCF-BERT（Zhang 等人，2019）、BERT-aspect（Devlin等人，2019）、AEN-BERT（Song 等人，2019）和微软（应用程序接口））在 **SemEval2014-Laptop** 数据集上的鲁棒性测试结果。

模型	AddDiff *# 331* 个样本		ReverseNonTarget *# 104* 个样本		ReverseTarget *# 331* 个样本	
	ori	转换	ori	转换	ori	转换
text-davinci-003						
零样本	83.84±0.33	77.50±2.43	82.43±0.42	39.61±4.83	83.62±0.12	47.04±4.64
单样本	85.77±0.69	87.17±4.62	85.63±1.05	52.22±8.86	85.84±0.57	57.07±7.43
3样本提示	85.91±0.12	**88.73±4.11**	85.08±0.48	**55.59±11.11**	85.98±0.12	**59.14±7.11**
text-davinci-002						
零样本	**86.38±0.11**	81.90±0.35	85.57±0.21	52.97±2.96	**86.4±0.26**	56.68±5.17
单样本	86.05±0.43	82.20±1.98	85.22±0.24	55.18±2.38	86.05±0.43	56.77±3.08
3样本提示	85.41±0.43	81.55±1.86	84.80±0.97	54.01±2.28	85.48±0.33	56.65±2.43
text-davinci-001						
零样本	85.21±1.70	80.10±2.11	**85.89±1.68**	47.35±4.16	85.26±2.17	53.56±0.92
单样本	84.56±2.59	77.55±4.34	84.64±3.34	45.02±1.26	84.76±2.80	51.11±5.13
3样本提示	83.33±0.69	71.90±0.12	83.40±0.24	48.40±1.98	83.44±0.87	50.26±0.77
微调						
BERT-base-uncased	83.69	70.82	85.77	43.51	83.69	46.14
LCF-BERT	81.97	77.90	82.85	56.90	81.97	48.93
BERT-aspect	81.97	72.96	82.85	51.05	81.97	54.08
AEN-BERT	81.55	71.89	83.26	40.17	81.55	50.64
微软（应用程序接口）	**91.79**	**90.63**	**91.79**	**66.06**	**91.79**	**69.70**

表13：GPT系列模型（零样本和少样本）及微调模型（即BERT-base-uncased、LCF-BERT、BERT-aspect、AEN-BERT和微软（应用程序接口））在**SemEval2014-Restaurant** 数据集上的鲁棒性测试结果。

模型	AddDiff *# 492*个样本		ReverseNonTarget *# 227*个样本		ReverseTarget *# 492*个样本	
	ori	转换	ori	转换	ori	转换
text-davinci-003						
零样本	89.45±0.87	55.25±20.03	91.58±0.96	47.93±0.45	89.38±0.85	54.60±7.18
单样本	92.08±0.74	78.30±17.09	92.84±0.52	64.97±11.58	92.05±0.78	66.62±8.78
3样本提示	**92.56±0.83**	**89.47±7.53**	**93.01±0.52**	**70.95±5.76**	**92.52±0.78**	70.68±4.60
text-davinci-002						
零样本	91.86±0.79	74.05±19.15	92.00±0.91	69.50±3.12	91.90±0.79	69.96±5.18
单样本	92.48±0.54	86.46±7.66	92.39±0.48	68.94±2.88	92.41±0.61	**72.39±4.25**
3样本提示	90.36±0.07	74.14±16.42	90.84±0.50	56.82±4.08	90.36±0.07	52.73±13.11
text-davinci-001						
零样本	89.25±0.94	54.56±12.55	90.07±1.24	63.35±1.58	88.89±1.45	63.40±1.94
单样本	88.07±3.09	33.70±4.97	88.49±3.02	61.90±1.22	88.03±3.05	58.03±3.50
3样本提示	89.49±0.21	50.57±3.50	90.43±0.52	62.07±1.04	89.57±0.25	60.73±1.02
微调						
BERT-base-uncased	90.44	55.37	90.55	53.09	90.44	38.72
LCF-BERT	90.32	**85.83**	90.55	61.17	90.32	54.78
BERT-aspect	90.32	81.23	91.41	56.19	90.32	64.70
AEN-BERT	90.20	59.86	90.21	58.08	90.20	34.59
微软（应用程序接口）	**94.36**	82.20	**94.36**	**69.82**	**94.36**	**75.95**

我们在 SemEval2014-Laptop 和 SemEval2014-Restaurant 数据集上测试基于方面的情感分析，结果如表12和表13所示。从表中可以看出：**1）0/1/3样本模型表现良好，与微调模型不相上下**。这些模型在这两个数据集上都表现出稳定的性能和鲁棒性，不同提示之间的性能差异也很小。**2）上下文学习已被证明能显著增强模型的鲁棒性以及处理变换后数据的能力**。与微调模型相比，零样本模型的鲁棒性较低; ，然而，单样本和3样本模型的鲁棒性分别相当和更好。

4.4.2 情感分类

表14：GPT系列模型（零样本和少样本）与微调模型（即 XLNet、ULMFIT（Howard和Ruder，2018年）、BERT-large-ITPT（Sun等人，2019年）和 DistilBERT）在**IMDB** 数据集上的鲁棒性测试结果。

模型	AddSum-Movie *# 11257* 个样本		AddSum-Person *# 12230* 个样本		DoubleDenial *# 22933* 个样本		SwapSpecialEnt-Movie *# 11257* 个样本		SwapSpecialEnt-Person *# 12230* 个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	91.51±1.14	91.56±0.80	91.59±1.04	89.62±0.57	92.02±0.77	91.13±0.69	91.53±1.16	91.17±0.77	91.60±0.97	91.62±0.86
单样本	91.67±0.61	91.4±0.66	91.99±0.53	**89.86±0.83**	92.24±0.50	90.60±0.75	91.68±0.62	91.22±0.46	92.00±0.51	**91.87±0.57**
3样本提示	85.73±3.61	84.57±3.83	86.22±3.45	82.88±3.38	87.38±3.10	85.80±3.79	85.69±3.58	85.08±3.82	86.23±3.44	86.06±3.48
text-davinci-002										
零样本	91.97±1.27	91.17±2.66	92.00±1.59	87.67±1.85	**93.33±0.96**	**92.57±1.11**	91.80±1.41	90.90±1.91	91.97±1.63	91.43±1.97
单样本	89.53±2.20	87.10±3.94	89.33±2.03	80.93±6.54	91.67±1.16	89.87±1.91	89.53±1.78	88.50±2.38	89.27±2.05	88.70±2.75
3样本提示	87.30±0.62	84.77±1.01	88.53±1.08	76.57±2.74	91.33±0.90	90.23±0.40	87.63±0.68	86.53±1.25	88.37±1.11	88.17±0.90
text-davinci-001										
零样本	**92.63±0.54**	**91.59±0.86**	**92.34±0.38**	68.41±18.64	93.20±0.44	76.52±15.39	**92.53±0.47**	83.91±13.67	**92.43±0.42**	81.96±9.59
单样本	91.93±0.06	90.57±0.50	91.53±0.51	86.26±0.78	93.13±0.46	92.43±0.57	91.57±0.15	**91.57±0.42**	91.50±0.35	91.33±0.15
3样本提示	91.13±0.49	89.03±0.32	91.40±0.82	86.08±0.91	93.23±0.49	92.40±0.26	91.00±0.56	90.93±0.45	91.77±0.58	91.83±0.50
微调										
XLNet	**95.97**	**95.38**	**95.92**	93.16	**96.27**	**95.34**	**95.97**	**95.75**	**95.92**	**95.85**
ULMFIT	94.38	94.03	94.76	92.93	94.72	89.55	94.38	94.33	94.76	94.70
BERT-large-ITPT	95.12	94.97	95.27	**94.39**	95.40	93.30	95.12	95.14	95.27	95.23
DistilBERT	92.00	90.65	92.35	86.30	92.80	88.92	92.00	91.91	92.35	92.38

我们在 IMDB 数据集上测试情感分类任务的鲁棒性，并得到表14。**0/1/3样本模型表现出良好的鲁棒性，与微调模型的表现相当**，且不同提示之间的方差较低。然而，GPT-3.5 在这个二分类数据集上表现出明显缺陷：模型给出的答案是“中立”，而不是提示中指定的正面或负面情感。随着提示中示例数量的增加，中性回答的比例也随之上升，导致三样本提示的表现下降。这可能是因为模型在微调阶段接触过三分类数据，从而影响了对二分类数据的评估。**此外，对于 IMDB 数据集而言，提示中的示例越短似乎越有利**。当提示中选取的示例较短时，少样本模型的表现优于使用较长示例时的表现。这是因为输入提示过长可能会对模型的建模能力产生负面影响。

4.5 序列标注

我们评估了 GPT 系列模型在两种常见序列标注任务——命名实体识别和词性标注——上的鲁棒性和性能。在第 4.5.1 节和第 4.5.2 节中，我们将分别介绍这两项任务的实验结果。

4.5.1 命名实体识别

表15：GPT系列模型（零样本和少样本）及微调模型（即BERT-base-cased（Devlin等人，2019年）、BERT-base-uncased、ELMo-BiLSTM-CRF（Peters等人，2018年）、Flair嵌入（Akbik等人，2018年）、CNN-BiLSTM-CRF（Ma和Hovy，2016年）、GRN（Chen等人，2019年））在**ACE2005** 数据集上的鲁棒性测试结果。

模型	句子拼接 *# 1312*个样本		跨类别 *# 1312*个样本		实体拼写错误 *# 1405*个样本		OOV *# 1312*个样本		较长项交换 *# 1312*个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	36.93±1.86	29.43±1.99	37.03±1.69	36.33±1.17	31.31±2.35	26.42±2.31	36.97±1.91	69.86±2.92	36.90±1.85	75.39±2.20
单样本	43.37±1.52	34.68±1.61	43.48±1.48	34.94±0.45	37.63±1.28	30.82±1.32	43.43±1.41	70.35±1.05	43.34±1.49	77.42±0.36
3样本提示	**54.11±0.60**	**47.30±1.01**	**53.99±0.59**	**39.21±0.26**	**45.51±0.44**	**43.10±0.24**	**54.29±0.51**	**71.46±0.22**	**54.17±0.77**	**77.44±0.37**
text-davinci-002										
零样本	35.50±1.64	37.95±11.49	32.65±3.89	30.58±1.64	37.13±0.86	35.22±1.36	35.55±1.57	59.20±1.87	32.60±3.97	65.36±1.23
单样本	45.81±0.78	39.12±0.69	45.71±0.87	35.24±0.69	40.50±0.34	37.27±0.41	45.77±0.73	64.53±0.10	45.54±0.74	73.42±0.95
3样本提示	35.68±3.97	32.92±4.99	35.82±3.89	20.12±4.07	30.70±2.24	35.52±2.08	35.85±4.15	43.46±2.00	35.90±4.11	42.87±4.30
text-davinci-001										
零样本	12.71±5.81	12.75±2.83	12.68±5.77	10.50±3.93	15.93±4.26	14.54±2.22	12.86±5.85	17.18±5.71	12.81±5.87	20.93±8.03
单样本	23.46±3.10	22.58±4.43	23.50±3.25	17.31±0.43	25.61±2.04	21.45±2.66	23.49±3.14	31.72±0.39	23.49±3.01	38.48±0.83
3样本提示	26.56±1.93	26.46±1.27	26.48±2.11	16.50±0.43	25.06±1.56	23.81±1.18	26.57±2.01	30.24±1.99	26.75±2.06	28.47±2.60
微调										
BERT-base-cased	87.27	86.23	87.35	**48.08**	87.53	**83.08**	87.35	78.95	87.35	**82.12**
BERT-base-uncased	**88.75**	**88.65**	**88.68**	46.01	**89.14**	82.95	**88.68**	74.56	**88.68**	78.54
ELMo-BiLSTM-CRF	86.68	86.38	86.06	44.21	86.96	82.24	86.06	84.71	86.06	75.50
Flair嵌入	85.53	85.21	84.57	44.94	86.07	81.54	84.57	81.26	84.57	73.14
CNN-BiLSTM-CRF	82.10	81.24	82.85	43.5	82.46	73.52	82.85	64.20	82.85	67.74
GRN	83.68	83.30	83.97	39.04	84.15	76.86	83.97	**79.61**	83.97	70.89

表16：GPT系列模型（零样本和少样本）以及微调模型（即 BERT-base-cased、BERT-base-uncased、ELMo-BiLSTM-CRF、Flair嵌入、CNN-BiLSTM- CRF、GRN）在 **CoNLL2003** 数据集上的鲁棒性测试结果。

模型	句子拼接 *# 3453* 个样本		跨类别 *# 3453* 个样本		实体拼写错误 *# 2676* 个样本		OOV *# 3453* 个样本		较长项交换 *# 3453* 个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	50.46±2.10	52.97±1.27	50.47±2.15	26.39±1.50	55.93±1.81	43.43±1.31	50.42±2.21	47.13±1.55	50.43±2.14	43.29±1.32
单样本	52.40±1.24	53.96±1.98	52.35±1.29	27.11±1.72	57.54±1.14	47.84±1.11	52.39±1.11	47.95±1.25	52.36±1.13	46.13±1.77
3样本提示	57.73±1.50	56.80±1.58	57.62±1.58	30.93±0.38	61.13±1.55	49.65±1.40	57.70±1.56	56.21±1.86	57.70±1.60	52.34±0.32
text-davinci-002										
零样本	54.69±1.28	58.04±0.93	54.91±1.41	29.19±0.52	57.04±0.75	49.04±0.66	54.79±1.37	55.25±0.51	54.79±1.28	52.14±0.61
单样本	56.70±1.50	55.76±2.09	56.73±1.10	**35.40±1.22**	58.70±1.71	51.53±2.57	56.69±1.44	58.46±1.97	56.82±1.36	59.20±0.82
3样本提示	**61.64±0.87**	**59.15±0.84**	**61.52±0.94**	33.21±0.15	**63.20±0.93**	**54.58±1.83**	**61.50±0.84**	**64.52±0.27**	**61.52±0.92**	**58.78±0.47**
text-davinci-001										
零样本	18.92±1.12	18.48±1.64	18.84±1.48	11.94±0.45	22.00±0.97	18.61±1.24	18.72±0.95	13.83±1.35	19.12±1.17	16.36±1.07
单样本	28.66±1.42	26.06±2.43	28.07±1.50	19.39±0.85	30.85±1.38	27.80±1.69	28.75±1.44	28.04±2.16	28.44±1.17	33.17±2.45
3样本提示	35.69±0.35	30.10±1.73	35.53±0.60	19.32±0.76	37.50±0.47	29.81±1.57	35.54±0.27	28.60±1.81	35.71±0.67	39.36±0.16
微调										
BERT-base-cased	91.43	89.91	91.42	44.42	92.20	85.02	91.42	68.71	91.42	**79.28**
BERT-base-uncased	90.41	90.05	90.40	**47.19**	91.25	81.25	90.40	64.46	90.40	78.26
ELMo-BiLSTM-CRF	91.80	90.67	91.79	44.13	92.48	86.19	91.79	68.10	91.79	61.82
Flair嵌入	**92.25**	**90.73**	**92.24**	45.30	**93.05**	**86.78**	**92.24**	**73.45**	**92.24**	66.13
CNN-BiLSTM-CRF	90.61	87.99	90.59	44.18	91.25	79.10	90.59	58.99	90.59	61.15
GRN	91.57	89.30	91.56	42.90	92.29	82.72	91.56	68.20	91.56	65.38

表17：GPT系列模型（零样本和少样本）及微调模型（即 BERT-base-cased、BERT-base-uncased、ELMo-BiLSTM-CRF、Flair嵌入、CNN-BiLSTM- CRF、GRN）在 **OntoNotesv5** 数据集上的鲁棒性测试结果。

模型	句子拼接 *# 4019*个样本		跨类别 *# 4019*个样本		实体拼写错误 *# 4492*个样本		OOV *# 4019*个样本		较长项交换 *# 4019*个样本	
	ori	转换	ori	转换	ori	转换	ori	转换	ori	转换
text-davinci-003										
零样本	7.19±0.58	7.76±0.73	7.11±0.56	3.30±0.25	31.1±2.69	25.32±2.22	7.13±0.50	13.05±0.19	7.16±0.46	12.05±0.98
单样本	13.54±1.29	15.04±1.53	13.53±1.37	6.32±0.41	**37.40±1.49**	**31.08±1.44**	13.53±1.35	18.76±1.26	13.57±1.32	17.75±1.38
3样本提示	**16.63±1.85**	**19.07±1.02**	**16.61±1.77**	**7.85±1.04**	36.85±1.87	31.04±2.01	**16.64±1.76**	**24.47±2.11**	**16.68±1.85**	**23.13±1.99**
text-davinci-002										
零样本	1.94±0.25	2.17±0.45	2.33±0.64	1.05±0.33	29.78±0.82	25.82±0.67	1.98±0.20	3.95±0.91	1.97±0.31	2.87±0.45
单样本	4.22±0.33	5.06±0.16	4.09±0.11	2.44±0.39	29.29±3.65	27.95±2.47	4.07±0.29	8.93±0.60	4.22±0.38	7.01±0.64
3样本提示	8.00±3.11	9.78±1.14	8.59±2.19	5.43±1.10	33.50±7.37	26.68±5.88	8.57±2.17	17.96±2.36	8.46±2.77	14.9±2.97
text-davinci-001										
零样本	0.17±0.12	0.21±0.04	0.17±0.12	0.22±0.14	15.14±1.98	12.18±0.79	0.17±0.12	0.78±0.07	0.15±0.12	0.37±0.07
单样本	0.87±0.24	1.24±0.24	0.80±0.23	0.50±0.15	10.57±0.59	8.26±0.92	0.83±0.25	1.47±0.09	0.81±0.22	1.20±0.23
3样本提示	2.25±1.61	2.23±1.38	2.25±1.61	2.53±1.88	16.70±3.56	13.60±3.21	2.31±1.71	6.16±4.56	2.23±1.58	4.55±3.28
微调										
BERT-base-cased	**88.85**	**88.31**	**80.11**	**34.96**	**89.46**	**78.31**	**80.11**	**66.39**	**80.11**	**60.83**
BERT-base-uncased	86.85	86.68	75.21	33.47	87.45	75.11	75.21	59.56	75.21	55.09
ELMo-BiLSTM-CRF	86.94	86.59	72.68	27.37	87.55	78.50	72.68	60.04	72.68	57.17
Flair嵌入	86.96	86.67	70.65	25.78	87.66	80.12	70.65	55.76	70.65	55.12
CNN-BiLSTM-CRF	85.95	85.45	69.10	26.49	86.65	73.76	69.10	45.62	69.10	54.43
GRN	87.25	86.80	70.27	24.86	87.94	75.61	70.27	54.35	70.27	52.70

如表1所示，我们评估了GPT系列模型在六个命名实体识别数据集上的性能和鲁棒性。本小节展示了在前三个命名实体识别数据集 ACE2005、CoNLL2003 和 OntoNotesv5 上获得的实验结果。最后三个命名实体识别数据集的结果将在第4.7.1节介绍。我们使用 TextFlint 工具对这些数据集中的文本进行五种不同的变换。所有实验结果均以微平均F1分数给出。有关结果的详细信息，请参阅表 15 ∼ 17。

分析实验结果后，我们可以得出以下结论：**GPT系列模型在命名实体识别任务中面临重大挑战**。 与现有微调模型相比，GPT系列模型的F1分数要低得多。例如，GPT系列模型在ACE2005数据集上的最佳分数徘徊在50左右（表15），而微调模型的分数已接近90。不过，增加提示中的示例数量可以显著提升模型性能。这表明，增加提示中的示例数量对性能有正面影响。**text-davinci-003模型的性能并非总是最佳**。 比较GPT模型时，在ACE2005数据集（表15）和OntoNotesv5数据集（表17）上，text-davinci-003的表现优于text-davinci-002和text-davinci-001，这表明随着模型代际演进，其理解这些命名实体识别任务的能力也在提升。然而，在CoNLL2003数据集（表16）上，text-davinci-002在GPT系列的三个模型中表现最佳。我们推测，在某些命名实体识别任务中，模型的整体能力可能与其在特定语料库上的表现不成正比。**GPT系列模型在命名实体识别任务中的鲁棒性不稳定**。 在命名实体识别任务中，针对不同数据集采用的各种文本变换方法，对鲁棒性的影响程度各不相同。某些类型的数据变换会使鲁棒性略有提升，而另一些则可能导致鲁棒性略有下降，或使其没有显著变化。

4.5.2 词性标注

表18：GPT系列模型（零样本和少样本）及微调模型（即 ELMo-BiLSTM-CRF、Flair-BiLSTM-CRF、CNN-BiLSTM-CRF、BERT-BiLSTM-CRF（Devlin等人，2019））在 **WSJ** 数据集上的鲁棒性测试结果。带有 † 标记的模型表示提示中的示例使用原始标注（例如 NN 等）。由于该任务没有针对全量数据的特定变换，我们单独测试了模型在该数据集上的表现（列 all）。缺失的text-davinci-001结果是因为该模型未按照提示执行此任务。

(a)

模型	SwapMultiPOSJJ *# 3963* 个样本		SwapMultiPOS名词 *# 4952* 个样本		SwapMultiPOSRB *# 2874* 个样本	
	ori	转换	ori	转换	ori	转换
text-davinci-003						
零样本	75.67±2.65	74.92±2.79	75.47±2.63	74.26±2.72	74.81±2.60	72.80±2.81
单样本	71.42±0.69	70.71±0.67	71.35±0.64	70.46±0.71	71.40±0.55	69.87±0.5 4
3样本提示	84.09±0.21	83.57±0.17	83.91±0.20	83.07±0.16	83.55±0.29	81.86±0.33
单样本†	**88.77±0.62**	**88.31±0.57**	**88.18±0.57**	**87.21±0.53**	**87.64±0.48**	**85.69±0.31**
text-davinci-002						
零样本	71.48±0.56	70.78±0.69	71.13±0.67	70.06±0.76	70.62±0.52	69.14±0.77
单样本	68.75±0.71	67.68±0.55	68.29±0.50	67.20±0.38	68.06±0.51	66.26±0.37
3样本提示	79.64±1.00	79.28±1.11	79.53±1.06	78.85±0.95	79.74±1.00	77.99±0.66
text-davinci-001						
0/1/3次提示学习	-	-	-	-	-	-
微调						
ELMo-BiLSTM-CRF	97.73	**96.92**	97.72	96.48	97.55	95.11
Flair-BiLSTM-CRF	97.73	96.58	**97.74**	**97.08**	**97.58**	95.03
CNN-BiLSTM-CRF	97.55	96.46	97.53	96.90	97.35	94.51
BERT-BiLSTM-CRF	**97.78**	**96.92**	97.72	96.48	97.57	**97.75**

(b)

模型	SwapMultiPOSVB *# 2376* 个样本		SwapPrefix *# 4526* 个样本		all *# 5461* 个样本
	ori	转换	ori	转换	ori
text-davinci-003					
零样本	76.21±2.61	76.21±2.61	75.51±2.73	74.94±2.74	75.02±2.59
单样本	72.88±0.69	72.88±0.69	71.46±0.63	71.08±0.64	70.79±0.71
3样本提示	84.61±0.27	84.61±0.27	84.07±0.20	83.67±0.19	**83.69±0.16**
单样本†	**87.75±0.78**	**87.70±0.55**	**88.40±0.63**	**88.10±0.57**	-
text-davinci-002					
零样本	70.73±0.81	70.73±0.81	71.43±0.77	70.66±0.77	71.02±0.62
单样本	68.91±0.47	68.91±0.47	68.70±0.34	68.21±0.50	68.13±0.46
3样本提示	81.09±0.92	81.09±0.92	80.10±1.06	79.73±1.02	79.48±1.11
text-davinci-001					
0/1/3次提示学习	-	-	-	-	-
微调					
CNN-BiLSTM-CRF	97.63	97.30	97.55	96.87	97.54
ELMo-BiLSTM-CRF	97.80	97.39	97.75	96.66	**97.75**
Flair-BiLSTM-CRF	**97.83**	**97.53**	**97.76**	**97.33**	**97.75**
BERT-BiLSTM-CRF	97.80	93.34	97.75	96.66	**97.75**

我们使用三个词性标注数据集评估GPT系列模型的性能和鲁棒性，其中包括两个英语数据集（即WSJ和Daily547）以及一个中文数据集（即PKU-SEGPOS）。我们详细分析了WSJ数据集的结果，并在第4.7.1节中提供了Daily547和PKU-SEGPOS上更详细的结果。

我们还使用TextFlint工具对WSJ数据集进行了五种文本变换，以评估模型的鲁棒性。所有实验结果均以准确率表示，并列于表18。对结果的分析得出以下结论。**首先，GPT系列模型在词性标注任务中也面临挑战**。GPT系列模型在WSJ数据集各种文本变换形式下的准确率仍低于现有微调模型，但与命名实体识别任务相比，差距已显著缩小。与零样本模型相比，单样本模型的性能有所下降，而3样本模型的性能则显著提升。**其次，模型需要具备充分的语义理解能力，才能完成词性标注这类复杂任务**。比较GPT模型可见，同属GPT-3.5系列的text-davinci-003相较于text-davinci-002仅有轻微的性能提升。然而，GPT-3系列模型text-davinci-001无法完成该任务，因为它给出的非标准格式答案过多，超过了阈值。[9] 在其他任务中，从未出现模型完全无法理解所有零样本/少样本提示的情况。这也表明，直到text-davinci-002这一代，GPT模型才开始理解并基本完成相应的词性标注任务。我们推测，这可能是因为从text-davinci-002开始，模型接受了代码训练；这一点仍需进一步研究确认。**第三，GPT-3.5模型在词性标注任务中表现出良好的鲁棒性**。GPT-3.5系列中的text-davinci-003和text-davinci-002模型，以及微调模型，在该任务中面对各种形式的文本变换时均表现出很强的鲁棒性，且鲁棒性大致相当。

4.6 WinogradSchema挑战

表 19：GPT系列模型（零样本和少样本）以及微调模型（即 BERT-base-uncased、BERT-large-uncased、HNN（He 等人，2019年））在**WSC273** 数据集上的鲁棒性测试结果。

模型	all *# 570* 个样本	添加句子 *# 570* 个样本	插入关系从句 *# 566* 个样本	交换姓名 *# 566* 个样本	切换语态 *# 440* 个样本	交换性别 *# 310* 个样本
			text-davinci-003			
零样本	62.05±0.57	65.32±1.77	**59.83±0.51**	60.48±0.74	59.39±0.92	36.99±1.30
单样本	61.40±1.37	62.98±0.63	58.18±0.57	58.42±0.44	57.35±0.13	38.28±0.49
3样本提示	**62.75±1.13**	**65.38±0.53**	57.95±0.47	61.37±0.80	**59.62±0.57**	36.77±0.65
			text-davinci-002			
零样本	61.46±1.57	64.09±1.00	56.89±1.23	59.84±1.77	59.47±1.77	39.03±2.01
单样本	60.94±2.03	65.20±3.62	58.54±2.58	59.54±2.46	57.80±1.26	38.39±1.48
3样本提示	62.22±1.47	64.91±0.88	59.07±2.12	**62.07±0.89**	58.18±1.42	38.93±2.44
			text-davinci-001			
零样本	52.05±1.14	53.22±1.60	50.94±1.37	51.12±1.41	51.14±0.82	47.96±0.38
单样本	50.41±0.56	53.39±2.78	50.18±0.36	49.76±0.10	50.23±0.39	49.35±0.65
3样本提示	51.34±0.97	54.79±2.11	51.47±0.84	51.65±0.80	50.83±1.84	**49.57±1.83**
			微调			
BERT-base-uncased	56.00	-	57.60	52.70	57.70	**46.45**
BERT-large-uncased	61.90	-	59.10	56.50	60.90	41.94
HNN	**75.10**	-	**67.50**	**69.60**	**67.70**	27.10

对于 WSC，我们在 WSC273数据集上进行实验，并在表 19 中报告准确率。分析结果后，我们得出以下发现。**text-davinci-003 在 WSC 任务上的表现可与 BERT-large 相比**。尽管 text-davinci-003 在 WSC273 上的表现仍与当前最佳水平的结果有显著差距，但已接近或超过 BERT-large 的结果，这表明它在应对这一挑战方面具有一定潜力。**然而，GPT 模型能力的增强导致其在某些文本变换中的鲁棒性下降**。模型能力从 text-davinci-001 到 t ext-davinci-003 逐步增强，而 text-davinci-003 确实取得了最佳表现

[9] 当 GPT 模型呈现这种行为时，我们在表格中用符号“-”表示，与下文相同 .

在原始数据集上。然而，对于与名称和性别相关的数据集，text-davinci-002 和 text-davinci-001 分别取得了最佳结果。这表明，我们模型能力的增强使其失去了一些审慎判断力，这一问题仍需进一步研究。

4.7 进一步研究

4.7.1 部分序列标注任务

表20：GPT系列模型在部分序列标注数据集上的性能测试结果（零样本和少样本），Daily547采用准确率作为评估指标，其他数据集采用微平均 F1 分数。表中部分结果缺失，是因为模型未能按照提示完成任务，导致我们无法分析结果。

模型	HONOR *# 1120* 个样本	MSRANER *# 4365* 个样本	OntoNote4NER *# 4346* 个样本	Daily547 *# 546* 个样本	PKU-SEGPOS *# 5204* 个样本
			text-davinci-003		
零样本	47.62±2.00	23.02±4.67	30.66±1.19	64.80±0.18	65.86±1.18
单样本	49.04±0.89	46.89±0.81	49.72±0.81	77.99±1.15	76.43±0.45
3样本提示	**54.01±2.22**	57.14±0.36	**50.98±0.36**	**82.63±0.55**	**76.88±0.36**
			text-davinci-002		
零样本	46.60±8.34	15.13±1.83	-	52.96±4.49	39.11±6.12
单样本	45.14±11.55	35.63±2.14	35.69±4.34	65.25±0.62	56.28±1.55
3样本提示	51.02±13.63	**58.24±1.88**	34.27±11.17	76.03±0.51	51.82±2.00
			text-davinci-001		
零样本	-	-	-	-	-
单样本	32.42±1.05	25.16±2.18	-	-	15.09±0.37
3样本提示	39.09±0.40	42.43±0.68	36.38±0.23	-	24.30±0.81

我们测试了 GPT 系列模型在多个命名实体识别和词性标注数据集的少样本场景下的表现，并得到了表20。**结果表明，它们的得分仍然相当低，这进一步证实了我们的结论：GPT-3.5模型在ST 任务中面临相当大的困难（而 GPT-3 模型的情况甚至更为严重）**。尽管如此，在提示中增加示例数量，仍能显著提高大多数数据集上的得分。与其在所有这些任务中的其他表现相比，text-davinci-003模型在3样本场景下的整体表现最佳。

4.7.2 text-davinci-003 与其他大语言模型的比较

表21：text-davinci-003、LaMDA-PT（Thoppilan 等人，2022年）和 FLAN 的零样本与少样本表现。

子任务	数据集	text-davinci-003		LaMDA-PT		FLAN	
		零样本	3样本提示	零样本	[k]-shot	零样本	[k]样本
机器阅读理解	SQuAD1.1	78.22	**89.80**	22.70	50.20 [3]	80.00	82.70 [4]
	SQuAD2.0	76.46	**89.21**	11.10	34.90 [3]	44.20	43.10 [3]
自然语言推理	MNLI-m	64.67	**78.62**	35.70	43.70 [5]	61.20	63.50 [10]
	MNLI-mm	64.93	**77.51**	37.00	43.80 [5]	62.40	63.50 [10]
	SNLI	69.49	**74.73**	33.30	54.70 [5]	53.40	65.60 [15]
语义匹配	MRPC	74.50	**79.00**	53.70	64.00 [5]	69.10	67.20 [10]
	QQP	81.60	**84.20**	34.90	58.90 [3]	75.90	75.90 [16]
情感分类	IMDB	92.50	88.69	76.90	83.30 [1]	94.30	**95.00** [2]

我们比较了 text-davinci-003 与其他大语言模型（即 LaMDA-PT 和 FLAN）在部分任务上的表现，并在表21中展示了结果。从表中可以看出，在大多数任务上，text-davinci-003在3样本提示的少样本场景下，表现优于原始大语言模型的最佳结果；在某些任务（例如自然语言推理和语义匹配）上，它甚至超过了原始大语言模型在零样本下的最佳表现，这表明其理解能力优于原始大语言模型。

5 结论

本研究全面而深入地分析了 GPT-3.5 模型的鲁棒性。我们的研究涵盖了 9 种不同自然语言理解任务中的 21 个数据集，从而广泛考察了模型的鲁棒性。结果表明，尽管 GPT-3.5 在某些任务上的表现优于现有微调模型，但仍面临显著的鲁棒性下降以及具体的鲁棒性挑战，包括鲁棒性不稳定、提示敏感性、数字敏感性，以及对任务标签和类型的敏感性。我们希望本研究能够启发未来的研究：1）考察鲁棒性的其他方面，例如指令的鲁棒性; 2）探索更有效的提示策略，确保模型的指令微调阶段与实际应用阶段保持一致;以及 3）针对某些任务（如命名实体识别和关系抽取）设计更好的提示，同时考虑到指令构建的复杂性可能是其表现较低的一个影响因素。

参考文献

Alan Akbik、Duncan Blythe 和 Roland Vollgraf。2018年。用于序列标注的上下文字符串嵌入。载于第*27*届国际计算语言学会议论文集，第 1638–1649 页。

Yejin Bang、Samuel Cahyawijaya、Nayeon Lee、Wenliang Dai、Dan Su、Bryan Wilie、Holy Lovenia、Ziwei Ji、Tiezheng Yu、Willy Chung、Quyet V. Do、Yan Xu 和 Pascale Fung。2023年。对 ChatGPT 在推理、幻觉和交互性方面进行的多任务、多语言、多模态评估。*CoRR*，abs/2302.04023。

Tom B. Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel M. Ziegler、Jeffrey Wu、Clemens Winter、Christopher Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于 Hugo Larochelle、Marc'Aurelio Ranzato、Raia Hadsell、Maria-Florina Balcan 和 Hsuan-Tien Lin（编），神经信息处理系统进展*33*：神经信息处理系统进展*2020*年年会，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上举行。

Hui Chen、Zijia Lin、Guiguang Ding、Jianguang Lou、Yusen Zhang 和 Borje Karlsson。2019年。GRN：用于增强命名实体识别卷积神经网络的门控关系网络。载于*AAAI*人工智能会议论文集，第 33 卷，第 6236–6243 页。

Aakanksha Chowdhery、Sharan Narang、Jacob Devlin、Maarten Bosma、Gaurav Mishra、Adam Roberts、Paul Barham、Hyung Won Chung、Charles Sutton、Sebastian Gehrmann、Parker Schuh、Kensen Shi、Sasha Tsvyashchenko、Joshua Maynez、Abhishek Rao、Parker Barnes、Yi Tay、Noam Shazeer、Vinodkumar Prabhakaran、Emily Reif、Nan Du、Ben Hutchinson、Reiner Pope、James Bradbury、Jacob Austin、Michael Isard、Guy Gur-Ari、Pengcheng Yin、Toju Duke、Anselm Levskaya、Sanjay Ghemawat、Sunipa Dev、Henryk Michalewski、Xavier Garcia、Vedant Misra、Kevin Robinson、Liam Fedus、Denny Zhou、Daphne Ippolito、David Luan、Hyeontaek Lim、Barret Zoph、Alexander Spiridonov、Ryan Sepassi、David Dohan、Shivani Agrawal、Mark Omernick、Andrew M. Dai、Thanumalayan Sankaranarayana Pillai、Marie Pellat、Aitor Lewkowycz、Erica Moreira、Rewon Child、Oleksandr Polozov、Katherine Lee、Zongwei Zhou、Xuezhi Wang、Brennan Saeta、Mark Diaz、Orhan Firat、Michele Catasta、Jason Wei、Kathy Meier-Hellstern、Douglas Eck、Jeff Dean、Slav Petrov 和 Noah Fiedel。2022年。PaLM：利用路径扩展语言建模。*CoRR*，abs/2204.02311。

Paul F. Christiano、Jan Leike、Tom B. Brown、Miljan Martic、Shane Legg 和 Dario Amodei。2017年。《基于人类偏好的深度强化学习》。载于 Isabelle Guyon、Ulrike von Luxburg、Samy Bengio、Hanna M. Wallach、Rob Fergus、S. V. N. Vishwanathan 和 Roman Garnett 编，《神经信息处理系统进展 30：2017年神经信息处理系统年会，2017年12月4日至9日，美国加利福尼亚州长滩》，第4299–4307页。

Kevin Clark、Minh-Thang Luong、Quoc V. Le 和 Christopher D. Manning。2020年。ELECTRA：将文本编码器预训练为判别器而非生成器。*CoRR*，abs/2003.10555。

Zihang Dai、Guokun Lai、Yiming Yang 和 Quoc V. Le。2020年。Funnel-transformer：过滤序列冗余，实现高效语言处理。*CoRR*，abs/2006.03236。Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向转换器预训练。载于 Jill Burstein、Christy Doran 和 Thamar Solorio（编），计算语言学协会北美分会*2019*年会议论文集：人类语言技术，*NAACL-HLT 2019*，明尼阿波利斯，明尼苏达州，美国，*2019*年*6*月*2*日至*7*日，第*1*卷（长篇与短篇论文），第4171–4186页。计算语言学协会。Bill Dolan 和 Chris Brockett。2005年。自动构建句子改写语料库。载于第三届国际改写研讨会（*IWP2005*）。Javid Ebrahimi、Anyi Rao、Daniel Lowd 和 Dejing Dou。2018年。Hotflip：用于文本分类的白盒对抗样本。载于 Iryna Gurevych 和 Yusuke Miyao（编），计算语言学协会第*56*届年会论文集，*ACL 2018*，墨尔本，澳大利亚，*2018*年*7*月*15*日至*20*日，第*2*卷：短篇论文，第31–36页。计算语言学协会。Nanyi Fei、Zhiwu Lu、Yizhao Gao、Guoxing Yang、Yuqi Huo、Jingyuan Wen、Haoyu Lu、Ruihua Song、Xin Gao、Tao Xiang 等。2022年。通过多模态基础模型迈向通用人工智能。《自然·通讯》，13(1):3094。

Simon Frieder、Luca Pinchetti、Ryan-Rhys Griffiths、TommasoSalvatori、Thomas Lukasiewicz、Philipp Christian Petersen、Alexis Chevalier 和 Julius Berner。2023年。ChatGPT 的数学能力。*arXiv* 预印本 *arXiv:2301.13867*。

KevinGimpel、Nathan Schneider、Brendan O'Connor、Dipanjan Das、Daniel Mills、Jacob Eisenstein、Michael Heilman、Dani Yogatama、Jeffrey Flanigan 和 Noah A Smith。2010年。Twitter 的词性标注：标注、特征与实验。技术报告，卡内基梅隆大学匹兹堡分校计算机科学学院。

Ben Goertzel。2014年。通用人工智能：概念、现状与未来前景。《通用人工智能杂志》，5(1)：1。

Tao Gui、Xiao Wang、Qi Zhang、Qin Liu、Yicheng Zou、Xin Zhou、Rui Zheng、Chong Zhang、Qinzhuo Wu、Jiacheng Ye 等。2021年。TextFlint：用于自然语言处理的统一多语言鲁棒性评估工具包。*arXiv* 预印本 *arXiv:2103.11441*。

Biyang Guo、Xin Zhang、Ziyuan Wang、Minqi Jiang、Jinran Nie、Yuxuan Ding、Jianwei Yue 和 Yupeng Wu。2023年。ChatGPT 与人类专家的水平有多接近？比较语料库、评估与检测。

Pengcheng He、Xiaodong Liu、Weizhu Chen 和 Jianfeng Gao。2019年。一种用于常识推理的混合神经网络模型。*CoRR*，abs/1907.11983。

Pengcheng He、Xiaodong Liu、Jianfeng Gao 和 Weizhu Chen。2020年。DeBERTa：采用解耦注意力的解码增强型 BERT。*CoRR*，abs/2006.03654。

Amr Hendy、Mohamed Abdelrehim、Amr Sharaf、Vikas Raunak、Mohamed Gabr、Hitokazu Matsushita、Young Jin Kim、Mohamed Afify 和 Hany Hassan Awadalla。2023年。GPT模型的机器翻译水平如何？全面评估。*arXiv* 预印本 *arXiv:2302.09210*。

JeremyHoward 和 Sebastian Ruder。2018年。用于文本分类的微调语言模型。*CoRR*，abs/1801.06146。

Di Jin、Zhijing Jin、Joey Tianyi Zhou 和 Peter Szolovits。2020年。BERT 真的稳健吗？用于文本分类和蕴含任务自然语言攻击的强基线。在第三十四届*AAAI*人工智能大会，*AAAI 2020*，第三十二届人工智能创新应用大会，*IAAI 2020*，第十届*AAAI*人工智能教育进展研讨会，*EAAI 2020*，美国纽约，*2020*年*2*月*7*日至*12*日，第8018–8025页。AAAI出版社。

Mandar Joshi、Danqi Chen、Yinhan Liu、Daniel S. Weld、Luke Zettlemoyer 和 Omer Levy。2020年。SpanBERT：通过表示和预测跨度改进预训练。计算语言学协会汇刊，8：64–77。Jared Kaplan、Sam McCandlish、Tom Henighan、Tom B. Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。2020年。神经语言模型的缩放定律。*CoRR*，abs/2001.08361。

Michal Kosinski。2023年。心智理论可能在大型语言模型中自发涌现。*CoRR*，abs/2302.02083。

Zhenzhong Lan、Mingda Chen、Sebastian Goodman、Kevin Gimpel、Piyush Sharma 和 Radu Soricut。2019年。ALBERT：一种用于语言表示自监督学习的轻量级 BERT。*CoRR*，abs/1909.11942。Hector Levesque、Ernest Davis 和 Leora Morgenstern。2012年。温诺格拉德模式挑战。在第十三届知识表示与推理原理国际会议。Citeseer。Gina-Anne Levow。2006年。第三届国际中文语言处理竞赛：中文分词和命名实体识别。在第五届 *SIGHAN* 中文语言处理研讨会论文集，第 108–117 页。Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019年。RoBERTa：一种经过稳健优化的 BERT预训练方法。*CoRR*，abs/1907.11692。Haochen Liu、Yiqi Wang、Wenqi Fan、Xiaorui Liu、Yaxin Li、Shaili Jain、Yunhao Liu、Anil Jain 和 Jiliang Tang。2022年。可信人工智能：计算视角。*ACM*智能系统与技术汇刊，14(1)：1–59。

Xuezhe Ma 和 Eduard Hovy。2016年。通过双向 LSTM-CNN-CRF 进行端到端序列标注。*arXiv* 预印本 *arXiv:1603.01354*。

Andrew Maas、Raymond E Daly、Peter T Pham、Dan Huang、Andrew Y Ng 和 Christopher Potts。2011年。用于情感分析的词向量学习。载于计算语言学协会第*49*届年会论文集：人类语言技术，第142–150页。

Mitchell P. Marcus、Beatrice Santorini 和 Mary Ann Marcinkiewicz。1993年。构建大型英语标注语料库：宾州树库。《计算语言学》，19(2):313–330。

Matthew E. Peters、Mark Neumann、Mohit Iyyer、Matt Gardner、Christopher Clark、Kenton Lee 和 Luke Zettlemoyer。2018年。深度上下文化词表示。载于 Marilyn A. Walker、Heng Ji 和 Amanda Stent（编），计算语言学协会北美分会*2018*年会议论文集：人类语言技术，*NAACL-HLT 2018*，新奥尔良，路易斯安那州，美国，*2018*年*6*月*1*日至*6*日，第*1*卷（长篇论文），第2227–2237页。计算语言学协会。

Maria Pontiki、Dimitris Galanis、John Pavlopoulos、Harris Papageorgiou、Ion Androutsopoulos 和 Suresh Manandhar。2014 年。SemEval-2014 任务 4：基于方面的情感分析。载于第八届语义评估国际研讨会论文集（*SemEval 2014*），第27–35页，爱尔兰都柏林，8月。计算语言学协会。

Christopher Potts、Zhengxuan Wu、Atticus Geiger 和 Douwe Kiela。2020年。Dynasent：用于情感分析的动态基准。*arXiv*预印本 *arXiv:2012.15349*。

Chengwei Qin、Aston Zhang、Zhuosheng Zhang、Jiaao Chen、Michihiro Yasunaga 和 Diyi Yang。2023年。ChatGPT是通用型自然语言处理任务求解器吗？*arXiv*预印本 *arXiv:2302.06476*。

Pranav Rajpurkar、Jian Zhang、Konstantin Lopyrev 和 Percy Liang。2016年。斯坦福问答数据集：用于文本机器阅读理解的 100, 000+ 个问题。*CoRR*，abs/1606.05250。

Pranav Rajpurkar、Robin Jia 和 Percy Liang。2018年。了解自己不知道什么：斯坦福问答数据集中的不可回答的问题。*CoRR*，abs/1806.03822。

Marco Tulio Ribeiro、Tongshuang Wu、Carlos Guestrin 和 Sameer Singh。2020年。超越准确率：使用 CheckList 对自然语言处理模型进行行为测试。*arXiv* 预印本 *arXiv:2005.04118*。

Erik Tjong Kim Sang 和 Fien De Meulder。2003年。CoNLL-2003 共享任务简介：与语言无关的命名实体识别。计算语言学协会北美分会。

Victor Sanh、Lysandre Debut、Julien Chaumond 和 Thomas Wolf。2019年。DistilBERT，一种精简版 BERT：体积更小、速度更快、成本更低、重量更轻。*CoRR*，abs/1910.01108。

Victor Sanh、Albert Webson、Colin Raffel、Stephen H. Bach、Lintang Sutawika、Zaid Alyafeai、Antoine Chaffin、Arnaud Stiegler、Teven Le Scao、Arun Raja、Manan Dey、M Saiful Bari、Canwen Xu、Urmish Thakker、Shanya Sharma、Eliza Szczechla、TaewoonKim、Gunjan Chhablani、Nihal V. Nayak、Debajyoti Datta、Jonathan Chang、Mike Tian-Jian Jiang、Han Wang、Matteo Manica、Sheng Shen、Zheng Xin Yong、Harshit Pandey、Rachel Bawden、Tom Wang、Trishala Neeraj、Jos Rozen、Abheesht Sharma、Andrea Santilli、Thibault Févry、Jason A. Fries、Ryan Teehan、Stella Biderman、Leo Gao、Tali Bers、Thomas Wolf 和 Alexander M. Rush。2021年。多任务提示训练实现零样本任务泛化。

Youwei Song、Jiahai Wang、Tao Jiang、Zhiyue Liu 和 Yanghui Rao。2019年。用于目标情感分类的注意力编码器网络。*CoRR*，abs/1902.09314。

Chi Sun、Xipeng Qiu、Yige Xu 和 Xuanjing Huang。2019年。如何微调 BERT 以进行文本分类？载于 Maosong Sun、Xuanjing Huang、Heng Ji、Zhiyuan Liu 和 YangLiu（编），中文计算语言学——第十八届中国全国会议，*CCL 2019*，中国昆明，*2019*年*10*月*18*日至*20*日，论文集，计算机科学讲义第 11856 卷，第 194–206 页。施普林格。

Romal Thoppilan、Daniel De Freitas、Jamie Hall、Noam Shazeer、Apoorv Kulshreshtha、Heng-Tze Cheng、Alicia Jin、Taylor Bos、Leslie Baker、Yu Du、YaGuang Li、Hongrae Lee、Huaixiu Steven Zheng、Amin Ghafouri、Marcelo Menegali、Yanping Huang、Maxim Krikun、Dmitry Lepikhin、James Qin、Dehao Chen、Yuanzhong Xu、Zhifeng Chen、Adam Roberts、Maarten Bosma、Yanqi Zhou、Chung-Ching Chang、Igor Krivokon、Will Rusch、Marc Pickett、Kathleen S. Meier-Hellstern、Meredith Ringel Morris、Tulsee Doshi、Renelito Delos Santos、Toju Duke、Johnny Soraker、Ben Zevenbergen、Vinodkumar Prabhakaran、Mark Diaz、Ben Hutchinson、Kristen Olson、Alejandra Molina、Erin Hoffman-John、Josh Lee、Lora Aroyo、Ravi Rajakumar、Alena Butryna、Matthew Lamm、Viktoriya Kuzmina、Joe Fenton、Aaron Cohen、Rachel Bernstein、Ray Kurzweil、Blaise Aguera-Arcas、Claire Cui、Marian Croak、Ed H. Chi 和 Quoc Le。2022年。LaMDA：面向对话应用的语言模型。*CoRR*，abs/2201.08239。

Zhiguo Wang、Wael Hamza 和 Radu Florian。2017年。自然语言句子的双边多视角匹配。*CoRR*，abs/1702.03814。

Jason Wei、Maarten Bosma、Vincent Y. Zhao、Kelvin Guu、Adams Wei Yu、Brian Lester、Nan Du、Andrew M. Dai 和 Quoc V. Le。2022年。微调语言模型是零样本学习器。载于第十届国际学习表征会议，*ICLR 2022*，线上活动，*2022*年*4*月*25*日至*29*日。OpenReview.net。

Ralph Weischedel、Sameer Pradhan、Lance Ramshaw、Martha Palmer、Nianwen Xue、Mitchell Marcus、Ann Taylor、Craig Greenberg、Eduard Hovy、Robert Belvin 等。2011。Ontonotes 4.0版。*LDC2011T03*，费城，宾夕法尼亚州：语言数据联盟。

Adina Williams、Nikita Nangia 和 Samuel R. Bowman。2017年。用于句子理解通过推理进行理解。*CoRR*，abs/1704.05426。

Ikuya Yamada、Akari Asai、Hiroyuki Shindo、Hideaki Takeda 和 Yuji Matsumoto。2020年。LUKE：采用实体感知自注意力的深度上下文化实体表示。*CoRR*，abs/2010.01057。

Zhilin Yang、Zihang Dai、Yiming Yang、Jaime G. Carbonell、Ruslan Salakhutdinov 和 Quoc V. Le。2019年。XLNet：用于语言理解的广义自回归预训练。*CoRR*，abs/1906.08237。

Yuhao Zhang、Victor Zhong、Danqi Chen、Gabor Angeli 和 Christopher D. Manning。2017年。位置感知注意力和监督数据改进槽位填充。载于 Martha Palmer、Rebecca Hwa 和 Sebastian Riedel（编），*2017*年自然语言处理实证方法会议论文集，*EMNLP 2017*，哥本哈根，丹麦，*2017*年*9*月*9*日至*11*日，第35–45页。计算语言学协会。

Chen Zhang、Qiuchi Li 和 Dawei Song。2019年。使用方面特定图卷积网络进行基于方面的情感分类。载于*2019*年自然语言处理实证方法会议和第*9*届自然语言处理国际联合会议（*EMNLP-IJCNLP*）论文集，第4568–4578页，中国香港，11月。计算语言学协会。

Susan Zhang、Stephen Roller、Naman Goyal、Mikel Artetxe、Moya Chen、Shuohui Chen、Christopher Dewan、Mona T. Diab、Xian Li、Xi Victoria Lin、Todor Mihaylov、Myle Ott、Sam Shleifer、Kurt Shuster、Daniel Simig、Punit Singh Koura、Anjali Sridhar、Tianlu Wang 和 Luke Zettlemoyer。2022年。OPT：开放预训练Transformer语言模型。*CoRR*，abs/2205.01068。

Wenxuan Zhou 和 Muhao Chen。2021年。通过噪声标签学习进行以实体为中心的信息抽取。载于 Marie-Francine Moens、Xuanjing Huang、Lucia Specia 和 Scott Wen-tau Yih（编），《*2021*年自然语言处理实证方法会议论文集（*EMNLP 2021*），线上活动／多米尼加共和国蓬塔卡纳，*2021*年*11*月7日至*11*日》，第 5381–5392 页。计算语言学协会。

性能详情

表22展示了当前最佳水平模型、BERT、text-davinci-003（0/3/6/九样本）在各数据集上的最佳性能，其可视化结果绘于图1中。

表22：性能详情。标记为 † 的结果来自 BERT-Large-ITPT，标记为 ‡ 的结果来自 BERT-BILSTM-CRF。

Task	子任务	数据集	SOTA	BERT	text-davinci-003			
					零样本	3样本提示	六样本	九样本
机器阅读理解	机器阅读理解	SQuAD1.1	90.07	87.09	78.22	89.80	91.43	-
		SQuAD2.0	84.62	75.95	76.46	89.21	79.23	-
关系抽取	关系抽取	Tacred	70.87	68.01	21.90	22.12	25.77	23.40
句子对关系	自然语言推理	MNLI-m	91.60	84.39	64.67	78.62	80.94	81.20
		MNLI-mm	91.90	84.43	64.93	77.51	78.70	80.90
		SNLI	90.60	88.99	69.49	74.73	81.00	82.20
	语义匹配	QQP	92.28	90.91	81.60	84.20	85.10	85.90
		MRPC	88.40	84.28	74.50	79.00	77.60	78.20
情感分析	基于方面的情感分析	笔记本电脑	91.79	83.69	84.12	86.05	86.70	86.91
		餐厅	94.36	90.44	90.44	93.15	93.03	92.68
	情感分类	IMDB	95.97	95.27 †	92.50	88.69	85.20	84.50
序列标注	命名实体识别	ACE 2005	88.75	88.75	38.51	54.78	55.27	56.03
		CoNLL 2003	92.25	91.43	51.83	59.45	61.51	74.49
		OntoNotes v5	88.85	88.85	7.72	18.61	15.92	15.01
		HONOR	-	-	45.69	56.40	57.63	64.08
		MSRANER	-	-	28.31	57.48	56.34	57.08
		OntoNote4NER	-	-	31.92	52.04	47.99	53.74
	词性标注	WSJ	97.75	97.75 ‡	76.80	83.83	83.56	84.60
		Daily547	-	-	65.00	83.24	83.66	84.22
		PKU-SEGPOS	-	-	66.94	77.29	78.16	81.13
温诺格拉德模式挑战	温诺格拉德模式挑战	WSC273	75.10	56.00	62.46	64.04	65.44	64.04

B 提示

对于每个数据集，我们分别为零样本、单样本、3样本、6样本和9样本场景设计了三种提示。由于与单样本相比，3/6/9样本仅增加了更多上下文示例，因此我们在表23至表44中列出了每个数据集在零样本和单样本场景下使用的提示。

表23：SQuAD1.0数据集的0/1样本提示。“{context}”应替换为段落，“{question}”应替换为问题。

# 样本数	提示
零样本	段落：{context} // 问题：{question} // 根据以上段落，正确的给定问题的答案是
	请参阅以下段落并回答下列问题：// 段落：{context} // 问题：{question} // 答案：
	段落：{context} // 问题：{question} // 答案：
单样本	段落：“从建筑风格来看，这所学校具有天主教特色。主楼的金色穹顶上矗立着圣母玛利亚金像。就在主楼正前方的主楼前，并面向主楼的是一尊双臂高举的基督铜像，上面刻有铭文“Venite Ad Me Omnes”。主楼旁边是圣心大教堂。紧邻大教堂后方的是石窟，一处圣母祈祷与默想之地。这是法国卢尔德石窟的复制品，据说圣母玛利亚曾在那里显现献给1858年圣伯尔纳德·苏比鲁的。在主路尽头（沿一条直线（这条直线经过3座雕像和金色穹顶），矗立着一尊简洁、现代的石制圣母玛利亚雕像。’ // 问题：‘据称圣母玛利亚于1858年在卢尔德（法国）显现给了谁？’ // 根据以上段落，给出问题的正确答案是 // 答案：圣伯尔纳德·苏比鲁 // 段落：‘{context}’ // 问题：‘{question}’ // 根据以上段落，给出问题的正确答案是
	请参考以下段落并回答问题：// 段落：“从建筑风格上看，这所学校具有天主教特色。 主楼的金色穹顶上矗立着一尊圣母玛利亚金像。主楼正前方有一尊面朝主楼、双臂高举的基督铜像，旁边刻有铭文“Venite Ad Me Omnes”。主楼旁边是圣心大教堂。紧接在大教堂内有石窟，这是一处圣母祈祷与默想之地。它是仿照卢尔德的石窟建造的复制品，位于法国卢尔德，据说圣母玛利亚曾在那里显现给圣伯尔纳德·苏比鲁，时间是在1858年。在主车道尽头（沿着一条直线，依次穿过三尊雕像和金色穹顶），立着一尊造型简洁、现代风格的石制圣母玛利亚像。’ // 问题：‘1858年在法国卢尔德，据称圣母玛利亚显现给了谁？’ // 答案：圣伯尔纳德·苏比鲁 // 请参阅以下段落并回答以下问题： // 圣伯尔纳德·苏比鲁 // 请参阅以下段落并回答以下问题： // 段落：‘{context}’ // 问题：‘{question}’ // 答案：
	段落：‘从建筑风格来看，这所学校具有天主教特色。主楼的金色穹顶上矗立着一尊圣母玛利亚金像。主楼正前方有一尊面朝主楼、双臂高举的基督铜像，上面刻着“Venite Ad Me Omnes”。主楼旁边是圣心大教堂。大教堂后方紧邻着石窟，这是一处圣母祈祷与默想之地。它仿照法国卢尔德的石窟建造，据说圣母玛利亚曾于1858年在那里显现给圣伯尔纳德·苏比鲁。主干道尽头（沿着一条直线，途经三尊雕像和金色穹顶）矗立着一尊简朴的现代风格石制圣母像。’ // 问题：‘据称，圣母玛利亚于1858年在法国卢尔德显现给了谁？’ // 答案：圣伯尔纳德·苏比鲁 // 段落：‘{context}’ // 问题：‘{question}’ // 答案：

表 24：SQuAD2.0数据集的 0/1样本提示。“{context}”应替换为段落，“{question}”应替换为问题。

# 样本数	提示

零样本	段落：{context} // 问题：{question} // 根据上述段落，给定问题的正确答案是
	
	请阅读下面的段落并回答以下问题：// 段落：{context} // 问题：{question} // 答案：
	
	段落：{context} // 问题：{question} // 答案：

单样本	段落：‘从建筑风格上看，这所学校具有天主教特色。主楼的金色穹顶上矗立着一尊圣母玛利亚金像。就在主楼正前方楼前矗立着一尊基督铜像，双臂高举，刻有铭文“Venite Ad Me Omnes”。主楼旁边是圣心大教堂。大教堂正后方是石窟，这是一处圣母祈祷与默想之地。它是法国卢尔德石窟的复制品，据说圣母玛利亚曾在那里显现，并于1858年向圣伯尔纳德·苏比鲁显现。沿主车道走到尽头（沿直线前行，这条直线经过三尊雕像和金色穹顶），那里矗立着一尊简朴、现代风格的石像，描绘的是玛利亚。’ // 问题：‘圣母玛利亚据称于1858年在卢尔德向谁显现法国？’ // 根据上述段落，给定问题的正确答案是 // 答案：圣伯尔纳德·苏比鲁 // 段落：‘{context}’ // 问题：‘{question}’ // 根据上述段落，给定问题的正确答案是 参照以下段落回答问题： // 段落：‘从建筑风格来看，这所学校具有天主教特色。主楼金色穹顶顶端有一尊圣母玛利亚金像。主楼正前方有一尊铜像，正对着它，是一尊双臂高举的基督铜像，上面刻着“Venite Ad Me Omnes”。主楼旁边是圣心大教堂。大教堂正后方是石窟，这是一处圣母祈祷与默想之地。它是法国卢尔德石窟的复制品，据说圣母玛利亚曾在那里显现给圣伯尔纳德·苏比鲁在1858年。沿主车道尽头（沿直线穿过3座雕像和金色穹顶），矗立着一尊简洁、现代的石制圣母玛利亚雕像。’ // 问题：‘据称圣母玛利亚向谁于1858年在法国卢尔德显现？’ // 答案：圣伯尔纳德·苏比鲁 // 请参阅下面的段落并回答以下问题： // 段落：‘{context}’ // 问题：‘{question}’ // 答案： 段落：‘从建筑风格上看，这所学校具有天主教特色。主楼顶部的金色穹顶上是圣母玛利亚金像。就在主楼正前方，面向主楼的是一尊双臂高举的基督铜像，旁边刻有铭文“Venite Ad Me Omnes”。主楼旁边是圣心大教堂。大教堂的正后方是石窟，这是一处圣母祈祷与默想之地。它是法国卢尔德石窟的复制品，据称圣母玛利亚曾在那里显现给圣伯尔纳德·苏比鲁，时间是在1858年。沿主干道尽头（并且在一条直线上（它通过3座雕像和金色穹顶相连）是一尊简朴、现代风格的石制玛丽雕像。’ // 问题：‘圣母玛利亚据称于1858年在卢尔德向谁显现法国？’ // 答案：圣伯尔纳德·苏比鲁 // 段落：‘{context}’ // 问题：‘{question}’ // 答案：

表 25：TACRED 数据集的 0/1样本提示。“{subj}”应替换为主语词，“{obj}”应替换为宾语词，“{options}”应替换为“人物与年龄、无关系、人物与头衔、组织与高层成员或雇员、组织与总部所在国家、人物与父母、人物与居住国、人物与子女、组织与别名、人物与指控、人物与居住城市、人物与籍贯、组织与创办者、人物与雇主、人物与兄弟姐妹、人物与别名、组织与网站、人物与宗教、人物与出生州或省、组织与母公司、组织与子公司、人物与其他家庭成员、人物与居住州或省、组织与成员、人物与死因、组织与所属组织、组织与雇员或成员人数、人物与出生国、组织与股东、组织与总部所在州或省、人物与死亡城市、人物与出生日期、人物与配偶、组织与总部所在城市、人物与死亡日期、人物与就读学校、组织与政治或宗教归属、人物与死亡国家、组织与成立时间、人物与出生州或省、人物与出生城市、组织与解散时间”。

# 样本数	提示
零样本	‘{token}’ // 在上面的文本中，‘{subj}’与‘{obj}’之间是什么关系？ // 选项：‘{options}’ // 答案： ‘{token}’ // 限定词‘{subj}’与‘{obj}’在上面的句子中的关系。 // 选项：‘{options}’ // 答案： ‘{token}’ // 根据上面的句子，找出‘{subj}’与‘{obj}’之间的关系。// 选项：‘{options}’ // 答案：

单样本	‘Graham，55 岁，在这起杀人事件中一直坚称自己无罪。’ // 在上述文本中，
	‘Graham’与‘55’之间的关系是什么？ // 答案：人物与年龄 // ‘{token}’ // 在
	根据上面的文本，‘{subj}’和‘{obj}’之间是什么关系？ // 选项：‘{options}’
	// 答案：
	
	‘Graham，55岁，在杀人案中一直坚称自己无罪。’// 限定词关系
	在上面的句子中，‘Graham’和‘55’之间的关系。// 答案：人物与年龄// ‘{token}’//
	确定上面句子中‘{subj}’和‘{obj}’之间的关系。// 选项：
	‘{options}’// 答案：
	
	‘Graham，55 岁，一直坚称自己没有参与杀人。’ // 查找关系
	在上面的句子中，查找 ‘Graham’ 与 ‘55’ 之间的关系。// 答案：人物与年龄 // ‘{token}’
	// 查找上面句子中 ‘{subj}’ 与 ‘{obj}’ 之间的关系。// 选项：
	‘{options}’ // 答案：

表26：MNLI-m数据集的0/1样本提示。“{premise}”应替换为前提，“{hypothesis}”应替换为假设。

# 样本数	提示
零样本	‘{premise}’ 根据前面的段落，它与‘{hypothesis}’之间是蕴含、中立还是矛盾？这‘{假设}’ 假设“{premise}”，我们能否推断出“{hypothesis}”？ 请选择一个答案：蕴含 ，矛盾 ，中立 已知“{premise}”。因此，它必定是蕴含、矛盾或中立：‘{hypothesis}’
单样本	‘他心情矛盾，一方面沉醉于这个村庄的宁静。’根据前面的段落，是蕴含、中立还是矛盾：“他喜欢村庄的宁静的氛围。” // 答案：蕴含 // ‘{premise}’ 根据前面的段落，是蕴含、中立还是矛盾：“{hypothesis}” // 答案： 假设“他心情矛盾，其中一面沉浸于这个村庄的宁静之中。”我们能否推断“他喜欢这个村庄的宁静。”？请选择一个答案：蕴含、矛盾、中立 // 答案：蕴含 // 假设“{premise}”，我们能否推断“{hypothesis}”？请选择一个答案：蕴含、矛盾、中立 // 答案： 鉴于“他心情矛盾，其中一面沉浸于这个村庄的宁静之中。”因此，“他喜欢这个村庄的宁静”必然属于蕴含、矛盾或中立中的一种是。’ // 答案：蕴含 // 鉴于‘{premise}’，因此必定是蕴含，或者矛盾，或者中立，即‘{hypothesis}’ // 答案：

表 27：MNLI-mm 数据集的 0/1 样本提示。应将“{premise}”替换为前提，将“{hypothesis}”替换为假设。

# 样本数	提示
零样本	‘{premise}’ 根据前面的段落，“{hypothesis}”是蕴含、中立还是矛盾这‘{假设}’ 假设“{premise}”，我们能否推断出“{hypothesis}”？ 请选择一个答案：蕴含 ，矛盾 ，中立 鉴于“{premise}”，因此必然是蕴含、矛盾或中立，即‘{hypothesis}’
单样本	‘我会把他扭过来，先生。’ 根据前面的段落，这属于蕴含、中立还是矛盾：“我会让他变直。” // 答案：矛盾 // ‘{premise}’ 根据前面的段落，这属于蕴含、中立还是矛盾：‘{hypothesis}’ // 答案： 假设“我会扭转他，先生。”我们能否推断出“我会让他变直。”？请选择一个答案：蕴含、矛盾、中立 // 答案：矛盾 // 假设“{premise}”我们能否推断出“{hypothesis}”？请选择一个答案：蕴含、矛盾 , 中立 // 答案： 鉴于“我要拧他 , 先生。”因此 , 它必定是蕴含、矛盾或中立即“我要把他弄直。” // 答案：矛盾 // 鉴于“{premise}”因此，它必定是蕴含、矛盾或中立，即“{hypothesis}” // 答案：’

表28：用于SNLI数据集的0/1样本提示。“{premise}”应替换为前提，“{hypothesis}”应替换为假设。

# 样本数	提示
零样本	“{premise}”根据前述段落，这属于蕴含、中立还是矛盾 “{hypothesis}” 假设“{premise}”，我们能否推断出“{hypothesis}”？ 请选择一个答案： 蕴含 ，矛盾 ，中立 鉴于“{premise}”，因此它必然是蕴含、矛盾或中立中的一种： ‘{hypothesis}’

单样本	前提：‘一位骑马的人跳过一架失事飞机。’根据这一前提，
	我们能否得出假设‘一位人物正在训练他的马参加比赛。’为真？
	// 选项：中立、矛盾、蕴含 // 答案：中立 // 前提：‘{premise}’。
	根据这一前提，我们能否得出假设‘{hypothesis}’为真？ // 选项：
	中立 ，矛盾 ，蕴含 // 答案：
	
	假设“骑马的人跳过一架失事飞机。” // 我们能否推断
	“有人正在训练他的马参加比赛。”？ // 选项：中立、矛盾、
	蕴含 // 答案：中立 // 假设“{premise}” // 我们能否推断“{hypothesis}”？
	// 选项：中立、矛盾、蕴含 // 答案：
	
	已知“骑马的人跳过一架失事飞机。” 因此，
	必然可以断定“有人正在训练他的马参加比赛。”吗？ // 选项：中立、
	矛盾、蕴含 // 答案：中立 // 已知“{premise}”因此，必然
	可以断定“{hypothesis}”吗？ // 选项：中立、矛盾、蕴含 // 答案：

表 29：MRPC 数据集的零样本/1样本提示。应将“{sentence1}”替换为第一个句子，将“{sentence2}”替换为第二个句子。

# 样本数	提示
零样本	句子“{sentence1}”是否 释义（也就是说， 意思相同吗）这句子？‘{sentence2}’ // 选项：是，否 我想知道以下两个句子的意思是否相同。‘{sentence1}’ ‘{sentence2}’ // 选项：是，否 以下两个句子的意思相同吗？ // ‘{sentence1}’ // ‘{sentence2}’ // 选项：是，否
单样本	句子“Amrozi 指控他的兄弟 ，他称其为“证人”的人 ，of 故意歪曲他的证据。” 释义（也就是说，意思相同）这个句子吗？“Amrozi 仅称他为‘证人’，指控他的兄弟故意歪曲他的证据。’ // 选项：是，否 // 答案：是 // 是否句子‘{sentence1}’是否是这句话的释义（也就是说，意思相同）？‘{sentence2}’ // 选项：是，否 // 答案： 我想知道以下两个句子的意思是否相同。 ‘Amrozi 指控他的兄弟（他称其为“证人”）故意歪曲他的证据。’ ‘他只称其为“证人”，Amrozi 指控他的兄弟故意歪曲他的证据。’ // 选项：是，否 // 答案：是 // 我想知道以下两个句子的意思是否相同。‘{sentence1}’ ‘{sentence2}’ // 选项：是，否 // 答案： 以下两个句子的意思相同吗？ // ‘Amrozi 指控他的兄弟，他称其为“证人”，故意歪曲他的证据。’ // ‘提到只称他为“这证人”，Amrozi 指责他的兄弟故意歪曲他的证据。’ // 选项：是，否 // 答案：是 // 以下两个句子的意思是否相同？ // ‘{sentence1}’ // ‘{sentence2}’ // 选项：是，否 // 答案：

表30：QQP数据集的0/1样本提示。“{question1}”应替换为第一个问题，“{question}”应替换为第二个问题。

# 样本数	提示
零样本	对“{question1}”的答案也能用来回答“{question2}”吗? 问题“{question1}”和“{question2}”问的是同一件事吗？是或否? 我想知道下面两个问题的意思是否相同。‘{question1}’ ‘{question2}’ 是这样吗？// 选项：是、否
单样本	回答‘在印度股票市场投资的分步指南是什么？’的问题，是否也可以 用来回答‘投资股票市场的分步指南是什么？’？// 答案： 否 // 回答‘{question1}’的问题，是否也可以用来回答‘{question2}’？// 答案： “在印度投资股票市场的分步指南是什么？”和 “投资股票市场的分步指南是什么？”这两个问题问的是同一件事吗？是还 是否？// 答案：否 // “{question1}”和“{question2}”这两个问题问的是同一 件事吗？是还是否？// 答案： 我想知道以下两个问题是否意思相同。“什么是 在印度投资股票市场的分步指南？”“什么是投资股票市场的分步指南 投资股票市场？他们有吗？// 选项：是，否 // 答案：否 // 我想知道 以下两个问题的意思是否相同。‘{question1}’ ‘{question2}’ 他们有吗？// 选项：是，否 // 答案：

表31：SemEval2014 笔记本电脑数据集的0/1样本提示。“{aspect}”应替换为要分析的方面，“{sentence}”应替换为一个句子。

样本数	提示
零样本	分析“{sentence}”中“{aspect}”的情感，并判断其是 正面、负面还是中立。 对于“{sentence}”中的“{aspect}”，情感倾向是什么？它们被认为 是正面、负面还是中立? “{sentence}”请用正面、 负面、或中立。
单样本	分析“但令人遗憾的是，替换的基本输入输出系统（BIOS）在.....时卡死了”中对“基本输入输出系统”的情感倾向 再次更新基本输入输出系统后，电脑关机且无法重新启动。’并限定词它是否 是正面、负面还是中立。// 答案：负面 // 分析这句话的情感 ‘{aspect}’相对于‘{sentence}’的情感，并限定词它是正面、负面还是中立。// 答案: 对于‘但令人遗憾的是，更换后的设备在更新基本输入输出系统时死机了’这句话，情感是什么? 再次更新后关机且无法重新启动。’就‘基本输入输出系统’而言，是正面、负面 还是中立？// 答案：负面 // 这句话的情感是什么 对于‘{sentence}’中的‘{aspect}’，你的看法是怎样的？它们被视为正面、负面，还是 中立？// 答案: ‘但遗憾的是，替换的设备在再次更新基本输入输出系统时死机了，随后关机，而且 无法重新开机。’请表达你对‘基本输入输出系统’这一方面的情感，使用 正面、负面或中立。// 答案：负面 // ‘{sentence}’ 请表达你的情感 对于‘{aspect}’这一方面，使用正面、负面或中立来表达。// 答案:

表32：SemEval2014-Restaurant数据集的0/1样本提示。应将“{aspect}”替换为待分析的方面，并将“{sentence}”替换为句子。

# 样本数	提示
零样本	分析“{sentence}”中“{aspect}”的情感，并限定其为 正面、负面或中立。 对于‘{sentence}’中的‘{aspect}’，情感倾向是什么？它们被认为是 正面、负面还是中立? ‘{sentence}’请用正面、 负面、中立。
单样本	分析对“The food is good, especially their more”中的‘菜肴’的情感倾向 基本的菜肴，而且饮品很美味。’ 并限定其是正面、负面， 还是中立。 // 答案：正面 // 分析对‘{aspect}’的情感 ‘{sentence}’，并限定其是正面、负面还是中立。// 答案： 对‘食物很不错，尤其是他们那些比较简单的菜肴，而且 饮品很美味。’就‘菜肴’而言呢？它们是正面、负面还是 中立的？ // 答案：正面 // 对于‘{sentence}’，就 ‘{aspect}’而言，情感是怎样的？是正面、负面还是中立的？ // 答案： ‘食物很好吃，尤其是他们那些比较简单的菜肴，而且饮品也很美味。’ 请用正面、负面或 中立来表达你对‘菜肴’这一方面的情感。// 答案：正面 // ‘{sentence}’ 请表达你对这一方面的情感 使用正面、负面或中立来标注“{aspect}”。// 答案：

Tab图 33：IMDB 数据集的 0/1样本提示。“{sentence}”应替换为句子 e.

# 样本数	提示
零样本	这部电影表达的情感是正面还是负面? // ‘{sentence}’ // 选项：正面、负面 以下影评表达了什么情感? 正面还是负面? // ‘{sentence}’ // 选项：正面、负面 影评者对这部电影表达了什么情感? 正面还是负面? // ‘{sentence}’ // 选项：正面、负面
单样本	这部电影表达的情感是正面还是负面? //《大独裁者》是一部 超乎卓越的电影。查理·卓别林既极其风趣又机智，表现出色 同时又通过反法西斯讽刺传达了强有力的主张。反 纳粹演讲是卓别林在电影界的伟大时刻之一，其中蕴含着深刻的价值观。 整部电影看下来，我感到其中蕴含着某种更高层次的智慧，超越了 真正聪明的电影制作，正在发挥作用。’ // 选项：正面、负面 // 答案： 正面 // ‘{sentence}’ // 选项：正面、负面 // 答案： 下面这则影评表达了什么情感? 正面还是负面? // ‘这 《大独裁者》是一部优秀至极的电影。 查理·卓别林成功地兼具两者 极其有趣、机智，同时又在他的作品中 对法西斯主义进行了有力的讽刺。卓别林在结尾发表的反纳粹演讲，充满了价值观，是 电影界最伟大的时刻之一。在整部电影中，我感觉到有某种更高层次的 智慧在发挥作用，超越了真正聪明的电影制作。’ // 答案：正面 // ‘{sentence}’ // 选项：正面，负面 // 答案： 评论者对这部电影表达了什么情感? 正面还是负面? // 《大独裁者》是一部超乎寻常的杰作。查理·卓别林成功地兼具 极其滑稽与机智，同时又在他的 反法西斯讽刺中表达了鲜明的立场。卓别林在结尾发表的反纳粹演讲，连同其中传达的价值观，是 电影界最伟大的时刻之一。整部电影看下来，我感觉似乎有某种更高层次的 智慧在发挥作用，超越了真正聪明的电影制作本身。’ // 答案：正面 // ‘{sentence}’ // 选项：正面、负面 // 答案：

Tab表34：ACE 2005 数据集的 0/1样本提示。“{text}”应替换为句子 e.

# 样本数	提示

零样本	请从给定文本中识别组织、人物、地缘政治实体、设施、地点、车辆和武器实体，并按照“实体：组织：”的格式输出
	无|人物：词语1|地缘政治实体：无|设施：词语2|地点：词语3，词语4|车辆：无|武器：无” // 文本：{text} // 实体：
	
	请列出所有组织，人物，地缘政治实体，设施，地点、车辆和武器实体，使用“实体：组织：”的格式输出
	无|人物：单词1|地缘政治实体：无|设施：单词2|地点：单词3，单词4|车辆：无|武器：无” // 文本：{text} // 实体：
	
	从给定文本中提取所有组织、人物、地缘政治实体、设施、地点、车辆和武器实体，并按以下格式输出：“实体：组织：
	无|人物：词语1|地缘政治实体：无|设施：词语2|地点：单词3，单词4|车辆：无|武器：无” // 文本：{text} // 实体：

单样本	请从给定文本中识别组织、人物、地缘政治实体、设施、地点、车辆和武器实体 // 文本：谢谢你，Paula。 // 实体：组-
	织：无|人物：你，Paula|地缘政治实体：无|设施：无|地点：
	无|车辆：无|武器：无 // 文本：{text} // 实体：
	
	请列出所有组织，人物，地缘政治实体，设施，地点、车辆以及给定文本中的武器实体 // 文本：谢谢你，Paula。 // 实体：组织-
	化：None|人物：你，Paula|地缘政治实体：None|设施：None|地点：
	None|车辆：None|武器：None // 文本：{text} // 实体：
	
	从给定文本中提取所有组织、人物、地缘政治实体、设施、地点、车辆和武器实体 // 文本：谢谢你，Paula。 // 实体：组织-
	组织：无|人物：你，Paula|地缘政治实体：无|设施：无|地点：
	无|车辆：无|武器：无 // 文本：{text} // 实体：

Ta表 35：CoNLL 2003 数据集的 0/1样本提示。“{text}”应替换为文本 .

# 样本数	提示	

零样本	请从给定文本中识别组织、人物、地点和其他实体
	，并按以下格式输出：“实体：组织：无|人物：无|地点：
	词1，词2|其他实体：词3” // 文本：{text} // 实体：
	
	请列出给定文本中的所有组织、人物、地点和其他实体，
	并按照以下格式输出：“实体：组织：无|人物：无|地点：词1，
	词2|其他实体：词3” // 文本：{text} // 实体：
	
	从给定文本中提取所有组织、人物、地点和其他实体，
	并按照以下格式输出：“实体：组织：无|人物：无|地点：词1，
	Word2|其他：Word3” // 文本：{text} // 实体：	

单样本	请从给定文本中识别组织、人物、地点和其他实体 文本 // 文本：艾因，阿拉伯联合酋长国 1996-12-06 // 实体： 无|人物：无|地点：艾因，阿拉伯联合酋长国|其他实体：无 // 文本：{text} // 实体：	组织：
	请列出给定文本中的所有组织、人物、地点和其他实体 // 文本：艾因，阿拉伯联合酋长国 1996-12-06 // 实体：组织：无|人物： 无|地点：艾因，阿拉伯联合酋长国|其他实体：无 // 文本：{text} // 实体：	
	从给定文本中提取所有组织、人物、地点和其他实体 // 文本：艾因，阿拉伯联合酋长国 1996-12-06 // 实体：组织：无|人物： 无|地点：艾因，阿拉伯联合酋长国|其他：无 // 文本：{text} // 实体：	

表36：OntoNotes v5数据集的0/1样本提示。“{text}”应替换为句子，“{format}”应替换为“实体：组织：无|人物：单词1|地缘政治实体：无|设施：无|地点：单词2|时间：单词3|基数：无|货币：无|日期：无|百分比：无|语言：无|艺术作品：无|国籍、宗教或政治团体：单词4、单词5|数量：无|序数：无|产品：无|事件：无|法律：无”。

# 样本	提示

零样本	请识别组织，人物，地缘政治实体、设施、地点、时间、基数、货币、日期、百分比、语言、艺术作品、国籍、宗教或政治团体、数量、序数、产品、事件、法律实体。从给定文本中，按照‘{format}’ // 文本：{text} // 实体：的格式输出
	
	请列出所有组织，人物，地缘政治实体，设施，地点，时间，基数、货币、日期、百分比、语言、艺术作品、国籍、宗教或政治团体、数量、序数、产品、事件、给定文本中的法律实体，按以下格式输出：‘{format}’ // 文本：{text} // 实体：
	
	提取所有组织、人物、地缘政治实体、设施、地点、时间、基数、货币、日期、百分比、语言、艺术作品、国籍、宗教或政治团体、数量、序数、产品、事件、给定文本中的法律实体，并按以下格式输出格式为“{format}” // 文本：{text} // 实体：

单样本	请识别组织，人物，地缘政治实体、设施，地点，时间，基数，货币，日期、百分比、语言，艺术作品、国籍、宗教或政治团体、数量，给定文本中的序数、产品、事件、法律实体文本 // 文本：Tsai Chih - 笔绘制的图像 // 实体：组织：无|人物：Tsai Chih - 笔|地缘政治实体：None|设施：None|地点：None|时间：None|基数：None|货币：None|日期：None|百分比：None|语言：None|艺术作品：None|国籍、宗教或政治团体：None|数量：None|序数：None|产品：None|事件：None|法律：None // 文本：{text} // 实体：
	
	请列出所有组织，人物，地缘政治实体，设施，地点，时间，基数，货币，日期、百分比、语言，艺术作品、国籍、宗教或政治团体，数量，序数，产品，事件，在给定文本中的法律实体文本 // 文本：Tsai Chih - 笔绘制的图像 // 实体：组织：None|人物：Tsai Chih - pen|地缘政治实体：None|设施：None|地点：None|时间：无|基数：无|货币：无|日期：无|百分比：无|语言：无|艺术作品：无|国籍、宗教或政治团体：无|数量：无|序数：无|产品：无|事件：无|法律：无 // 文本：{text} // 实体：
	
	从中提取所有组织、人物、地缘政治实体、设施、地点、时间、基数，货币、日期、百分比、语言、艺术作品、国籍、宗教或政治团体、数量、序数、产品、事件、法律实体 // 文本：图像作者 Tsai Chih - pen // 实体：组织：无|人物：Tsai Chih - pen|地缘政治实体：无|设施：无|地点：无|时间：无|基数：None|货币：None|日期：None|百分比：None|语言：None|艺术作品：None|国籍、宗教或政治团体：None|数量：None|序数：None|产品：None|事件：None|法律：None // 文本：{text} // 实体：

Ta表 37：HONOR 数据集的 0/1样本提示。应将“{text}”替换为句子 .

# 样本数	提示
零样本	请识别文本中的所有结束日期、participant、开始日期、开始时间、开始结束时间、发生地、结束时间、开始结束日期，每个词最多出现在一个类别，使用“result:结束日期:无,参与人:无,开始日期:词语 1,开始时间:词语 2,start and end time:无 , location:无,结束时间:词语 4,开始结束日期:无”的格式输出 // 文本:{text} // result:
	请从文本中识别结束日期、参与人、开始日期、开始时间、start and end time、发生地、结束时间、开始结束日期并列举出来,每个词最多出现在一个类别,使用“结果:结束日期:无,参与人:无,开始日期:词语 1,开始时间:词语 2,开始结结束时间:无,location:无,结束时间:词语 4,开始结束日期:无”的格式输出 // 文本:{text} // 结果:
	Please tell me the 结束日期, participant, 开始日期, 开始时间, start and end time in the given text, location、结束时间、开始结束日期分别是什么？每个词最多只能出现在一个类别中，使用“结果:结束日期:无,参与人:无,开始日期:词语 1,开始时间:词语 2,开始结束时间:无,location:无,结束时间:词语 4,开始结束日期:无”的格式输出 // 文本:{text} // 结果:
单样本	请识别文本中的所有结束日期、participant、开始日期、开始时间、start and end time、发生地、结束时间、开始结束日期，每个词最多出现在一个类别 // 文本：让他帮我参加一下每天早上的meeting // result:结束日期:无,participant:无,开始日期:每天,开开始时间:morning,开始结束时间:start and end time,location:无,结束时间:结束时间,开始结束日期:无 // 文本:{text} // result:
	请从文本中识别结束日期、参与人、开始日期、开始时间、start and end time、发生地、结束时间、开始结束日期并列举出来，每个词最多出现在一个类别 // text:让他帮我参加一下每天早上的meeting // 结果:结束日期:无;participant:无;开始日期:每天;开始时间:早上;开始结束时间:无;location:无;结束时间:无;开始结束日期:无 // 文本:{text} // result:
	Please tell me the 结束日期, participant, 开始日期, 开始时间, start and end time in the given text, location、结束时间、开始结束日期分别是什么，每个词最多出现在一个类别 // text:让他帮我参加一下每天早上的meeting // 结果:结束日期:无;participant:无;开始日期:每天;开始时间:早上;开始结束时间:无;location:无;结束时间:无;开始结束日期:无 // 文本:{text} // result:

Tabl图 38：MSRANER数据集的0/1样本提示。“{text}”应替换为句子 e.

# 样本数	提示

零样本	请识别文本中的所有人名、地名、组织名，每个词最多出现在一个类别，使用“结果:人名:词语 1,词语 2; 地名:无; 组织名:词语 3”的格式输出 // text:{text} // 结果:
	
	请从给定文本中识别人名、地名和组织名，并将它们列举出来；每个词最多只能出现在一个类别中别，使用“结果:人名:词语 1,词语 2; 地名:无; 组织名:词语 3”的格式输出 // 文本:{text} // 结果:
	
	请告诉我给定文本中的人名、地名和组织名分别是什么；每个词最多只能出现在一个类别中，使用“result:人名:词语 1,词语 2; 地名:无; 组织名:词语 3”的格式输出 // 文本:{text} // 结果:

单样本	请识别text中的所有人名、地名、组织名，每个词最多出现在一个类别 // text:中共中央致中国致公党十一大的贺词 // result:人名:无; 地名:无; 组织名:中共中央，中国致公党十一大 // 文本:{text} // 结果:
	
	请从给定text中识别人名、地名、组织名并列举出来,每个词最多出现在一个类别 // 文本:中共中央致中国致公党十一大的贺词 // 结果:人名:无; 地名:无; 组织名:中共中央,中国致公党十一大 // 文本:{text} // result:
	
	请告诉我给定文本中的人名、地名、组织名是什么，每个词最多出现在一个类别。// text:中共中央致中国致公党十一大的贺词 // result:人名:无; 地名:无; 组织名:中共中央,中国致公党十一大 // text:{text} // result:

表39：OntoNote4NER 数据集的 0/1样本提示。$ 用作分隔符。“{text}”应替换为句子。

# 样本数	提示
零样本	请识别文本中的所有地缘政治实体、地名、组织机构名、人名，每个词最多出现在一个类别,使用“result:地缘政治实体 $$ 词语 1$ 词语 2$ 词语 3$ 词语 4$$$ 地名 $$ 无 $$$ 组织名 $$ 无 $$$ 人名 $$ 词语 5”的格式输出 // 文本:{text} // result
	请从text中识别地缘政治实体、地名、组织机构名、人名并列举出来，每个词最多出现在一个类别，使用“result:地缘政治实体 $$ 词语 1$ 词语 2$ 词语 3$ 词语 4$$$ 地名 $$ 无 $$$ 组织名 $$ 无 $$$ 人名 $$ 词语 5”的格式输出 // 文本:{text} // result
	请告诉我给定文本中的地缘政治实体、地名、组织机构名、人名是什么，每个词最多出现在一个类别，每个词最多出现在一个类别，使用“result:地缘政治实体 $$ 词语 1$ 词语 2$ 词语 3$ 词语 4$$$ 地名 $$ 无 $$$ 组织名 $$ 无 $$$ 人名 $$ 词语 5”的格式输出 // text:{text} // result
单样本	请识别文本中的所有地缘政治实体、地名、组织机构名、人名，每个词最多出现在一个类别 // text:二次大战日本结束统治后，台湾回归中国，帕劳则成为美国的托管地，并于20世纪80年代开始与我国有了政治上的接触。// result：地缘政治实体 $$ 日本 $ 台湾 $ 中国 $ 帕劳 $$$ 地名 $$ 无 $$$ 组织名 $$ 无 $$$ 人名 $$ 无 // text:{text} // result
	请从text中识别地缘政治实体、地名、组织机构名、人名并列举出来，每个词最多出现在一个类别 // 文本：二次大战日本结束统治后，台湾回归中国，帕劳则成为美国的托管地，并于20世纪80年代开始与我国有了政治上的接触。// result：地缘政治实体 $$ 日本 $ 台湾 $ 中国 $ 帕劳 $$$ 地名 $$ 无 $$$ 组织名 $$ 无 $$$ 人名 $$ 无 // text:{text} // result:
	请告诉我给定文本中的地缘政治实体、地名、组织机构名、人名是什么，每个词最多出现在一个类别 // text:二次大战日本结束统治后,台湾回归中国,帕劳则成为美国的托管地，并于20世纪80年代开始与我国有了政治上的接触。// 结果:地地缘政治实体 $$ 日本 $ 台湾 $ 中国 $ 帕劳 $$$ 地名 $$ 无 $$$ 组织名 $$ 无 $$$ 人名 $$ 无 // 文本:{text} // 结果:

表40：WSJ数据集的0/1样本提示。￥用作分隔符。“{text}”应替换为句子，“{candidate}”应替换为“形容词、复数名词、介词或连词、限定词、单数专有名词、并列连词、过去式动词、单数或物质名词、wh限定词、情态动词、动词原形、wh副词、逗号、动名词或现在分词、to、所有格词尾、句子边界标记、所有格wh代词、非第三人称单数现在时动词、左圆括号、右圆括号、副词、过去分词动词、第三人称单数现在时动词、左双引号、右双引号、比较级副词、货币数值、基数词、比较级形容词、小品词、人称代词、冒号、物主代词、前限定词、最高级副词、wh代词、最高级形容词、外来词、列表标记、感叹词、存在句there、英镑符号、复数专有名词、符号”。

# 样本数	提示

零样本	使用候选列表中的类别，对给定文本执行词性任务，并按以下格式输出
	“词语1¥类别¥¥词语2¥类别¥¥词语3¥类别” //
	候选列表：{candidate} // 文本：{text} // 结果：
	
	使用候选列表中的类别标注给定文本的词性，输出
	格式为“单词1¥类别¥¥单词2¥类别¥¥单词3¥类别” //
	候选列表：{candidate} // 文本：{text} // 结果：
	
	使用候选列表中的类别标注给定文本中的单词，输出格式为
	“单词1¥类别¥¥单词2¥类别¥¥单词3¥类别” // 候选项
	列表：{candidate} // 文本：{text} // 结果：

单样本	Do 词性 task for the 给定 text 使用 the 类别 in 候选项 list // 候选项 列表: {candidate} // 文本: Few 变化 were made in the way the 市场 are 受到监管 . // 结果:
	少数¥形容词¥¥变化¥复数 名词¥¥是¥过去式 动词¥¥made¥过去式
	短语动词¥¥在¥介词或连词¥¥这¥限定词¥¥方式¥单数或
	物质名词¥¥这¥限定词¥¥市场¥复数名词¥¥是¥非第三人称单数
	现在时动词¥¥受到监管¥过去分词动词¥¥。¥句子边界标记 // 文本:
	{text} // 结果：
	
	Tag the 部分 of 语音 in the 给定 text 使用 the 类别 in 候选项 list // 候选项 列表: {candidate} // 文本: Few 变化 were made in the way the 市场 are 受到监管 . // 结果:
	少数¥形容词¥¥变化¥复数 名词¥¥是¥过去式 动词¥¥made¥过去式
	短语动词¥¥在¥介词或连词¥¥这¥限定词¥¥方式¥单数或
	物质名词¥¥这¥限定词¥¥市场¥复数名词¥¥是¥非第三人称单数
	现在时动词¥¥受到监管¥过去分词动词¥¥。¥句子边界标记 // 文本:
	{text} // 结果：
	
	使用候选列表中的类别对给定文本中的单词进行标注 // 候选列表:
	{candidate} // 文本：少数变化改变了市场受到监管的方式。 // 结果:
	少数¥形容词¥¥变化¥复数名词¥¥是¥过去式动词¥¥做出¥过去
	短语动词¥¥在¥介词或连词¥¥这¥限定词¥¥方式¥单数或
	物质名词¥¥这¥限定词¥¥市场¥复数名词¥¥是¥非第三人称单数
	现在时动词¥¥受到监管¥过去分词动词¥¥。¥句子边界标记 // 文本:
	{text} // 结果：

表41：使用原始标签的 WSJ数据集单样本提示。应将“{text}”替换为句子。

# 样本数	提示
单样本	对 WSJ数据集中的给定 X 执行词性任务 // X: “ That ’s baseball . // Y: “”_“”_“” “That”_“限定词” “’s”_“第三人称单数现在时动词” “baseball”_“名词” “.”_“.” // X: {text} // Y:
	给定 X 的标注 // X: “ That ’s baseball . “’s”_“第三人称单数现在时动词” “baseball”_“名词” “.”_“.” // X: {text} // Y:
	X 的语法标注 // X: “ That ’s baseball . “’s”_“第三人称单数现在时动词” “baseball”_“名词” “.”_“.” // X: {text} // Y:

表42：Daily547 数据集的 0/1样本提示。 ¥ 用作分隔符。“{text}”应替换为句子，“{candidate}”应替换为“普通名词、代词、专有名词、名词性 + 所有格、专有名词 + 所有格、动词（包括系动词和助动词）、形容词、副词、感叹词、限定词、前置词或后置词/从属连词、并列连词、动词小品词、存在句 there/前置限定词、话题标签、@提及、话语标记、URL/电子邮件地址、表情符号、数词、标点符号、其他、名词性 + 动词性、专有名词 + 动词性、X + 动词性”。

# 样本数	提示

零样本	使用候选列表中的类别，对给定文本执行词性任务，输出格式为“单词1¥类别¥¥单词2¥类别¥¥单词3¥类别” // 候选列表：{candidate} // 文本：{text} // 结果：
	
	使用候选列表中的类别标注给定文本中的词性，输出格式为“单词1¥类别¥¥单词2¥类别¥¥单词3¥类别” // 候选列表：{candidate} // 文本：{text} // 结果：
	
	使用候选列表中的类别标注给定文本中的单词，输出格式为“单词1¥类别¥¥单词2¥类别¥¥单词3¥类别” // 候选项列表：{candidate} // 文本：{text} // 结果：

单样本	使用候选列表中的类别对给定文本执行词性任务 // 候选列表：{candidate} // 文本：Bridalplasty ! Love 这部 showww . // 结果：Bridalplasty¥专有名词¥¥!¥标点符号¥¥Love¥动词（包括系动词和助动词）系动词和助动词¥¥这个¥限定词¥¥showww¥普通名词¥¥。¥标点符号 // 文本：{text} // 结果：
	
	使用候选列表中的类别标注给定文本的词性 // 候选列表：{candidate} // 文本：Bridalplasty ! Love 这个 showww . // 结果：Bridalplasty¥专有名词¥¥！¥标点符号¥¥Love¥动词（包括系动词和助动词）系动词和助动词¥¥这¥限定词¥¥showww¥普通名词¥¥。¥标点符号 // 文本：{text} // 结果：
	
	使用候选列表中的类别标注给定文本中的单词 // 候选项列表：{candidate} // 文本：Bridalplasty ! Love 这档节目真棒 . // 结果：Bridalplasty¥专有名词¥¥!¥标点符号¥¥Love¥动词，包括。 系动词和助动词-iaries¥¥this¥限定词¥¥showww¥普通名词¥¥.¥标点符号 // 文本：{text} // 结果：

表43： PKU-SEGPOS数据集的0/1样本提示。应将“{text}”替换为句子 .

# 样本数	提示

零样本	请使用候选集的 part of speech，对给定 text 中的每个词语进行标注，使用“词语 1 _ 词性/词语 2_part of speech/词语 3 _ part of speech”的格式输出 // 候选集:noun, 时间词, 处所词, 方位词, 数词, classifier, 区别词, pronoun, verb, 形容词, 状态词, adverb, preposition, conjunction, particle, 语气词, interjection, 拟声词, idiom, 习用语, 简称词, 前接成分, 后接成分, morpheme, 非语素字, 标点符号 // 文本:{text} // result:
	
	请使用候选集标注给定文本中每个词的part of speech, 使用” 词语 1 _ 词性/词语 2 _ 词性/词语 3_ 词性”的格式输出 // 候选集：名词，时间词，处所词，方位词，数词，量词，区别词，代词，动词，形容词，状态词，副词，介词，连词，助词，语气词，叹词，拟声词，成语，习用语，简称词，前接成分，后接成分，语素，非语素字，标点符号 // 文本：{text} // 结果：
	
	对于给定的 text，请使用候选集标注每个词的 part of speech，使用“词语 1 _ 词性/词语 2 _ 词词性/词语 3 _ part of speech”的格式输出 // 候选集:noun, 时间词, 处所词, 方位词, numeral, 量词, 区别词, 代词, 动词, 形容词, 状态词, 副词, preposition, 连词, 助词, 语气词, 叹词, 拟声词, idiom, 习用语, 简称词, 前接成分, 后接成分, morpheme, 非语素字, 标点符号 // 文本:{text} // 结果:

单样本	请使用候选集中的词性，为给定文本中的每个词语进行标注 // 候选集：名词、时间词、处所词、方位词、数词、量词、区别词、代词、动词、形容词、状态词、副词、介词、连词、助词、语气词、叹词、拟声词、成语、习用语、简称词、前接成分、后接成分、语素、非语素字、标点符号 // 文本：Tianjin开发区蒸蒸日上。// result：Tianjin _ 名词/development zone _ 名词/蒸蒸日上 _ 成语/。_ 标点符号 // 文本:{text} // result:
	
	请使用候选集标注给定文本中每个词的part of speech // 候选集:noun, 时间词, 处所词, 方位词, 数词, 量词, 区别词, 代词, 动词, 形容词, 状态词, 副词, preposition, 连词, 助词, 语气词，interjection，拟声词，idiom，习用语，简称词，前接成分，后接成分，morpheme，非语素字，标点符号 // 文本:Tianjin开发区蒸蒸日上。// result:Tianjin _ 名词/development zone _ 名词/蒸蒸日上 _ 成语/。_ 标点符号 // 文本:{text} // 结果:
	
	对于给定的文本，请使用候选集标注每个词的词性 // 候选集：名词、时间词、处所词、方位词、数词、量词、区别词、代词、动词、形容词、状态词、副词、preposition、连词、助词、语气词、叹词、拟声词、成语、习用语、简称词、前接成分、后接成分、语素、非语素字、标点符号 // 文本：Tianjin开发区蒸蒸日上。// 结果：Tianjin _ noun/development zone _ noun/蒸蒸日上 _ 成语/。_ 标点符号 // text:{text} // result:

表44：WSC273数据集的0/1样本提示。应将“{text}”替换为句子，将“{target 1}”替换为第一个目标词，将“{target 2}”替换为第二个目标词。

# 样本数	提示
零样本	‘{text}’ // 在前面的句子中，‘{target 2}’指的是‘{target 1}’吗？ // 选项：是，否 // 答案： ‘{text}’ // 在这里，‘{target 2}’是否指‘{target 1}’？ // 选项：是，否 // 答案： ‘{text}’ // 在上面的段落中，‘{target 2}’可以替换为‘{target 1}’吗？ // 选项：是，否 // 答案：
单样本	‘市议会拒绝向抗议者发放许可，因为他们担心会发生暴力事件。’// 在前面的句子中，‘they’指的是‘市议会’吗？ // 答案：是 // ‘{text}’ // 在前面的句子中，‘{target 2}’是否指‘{target 1}’？ // 选项：是，否 // 答案： ‘市议会拒绝向抗议者发放许可，因为他们担心发生暴力事件。’ // 这里，‘they’是否指‘市议会’？ // 答案：是 // ‘{text}’ // 这里， ‘{target 2}’是否指‘{target 1}’？ // 选项：是，否 // 答案： ‘市议会拒绝向抗议者发放许可，因为他们担心发生暴力事件。’ // 在上面的段落中，能否用“市议会”替换“they”？ // 答案：可以 // “{text}” // 在上面的段落中，能否用“{target 1}”替换“{target 2}”？ // 选项： 可以，不可以 // 答案：