探索大语言模型在主观文本片段识别任务中的性能

Alphaeus Dmonte[1], Roland Oruche[2],Tharindu Ranasinghe[3]Marcos Zampieri[1],Prasad Calyam[2]
[1]乔治·梅森大学，弗吉尼亚州，美国 [2]密苏里大学，密苏里州，美国
[3]兰卡斯特大学，英国 admonte@gmu.edu

arXiv:2601.00736v1 [cs.CL] 2 Jan 2026

摘要

识别相关文本片段对自然语言处理中的多项下游任务至关重要，因为这有助于提高模型可解释性。虽然大多数跨度识别方法依赖于相对较小的预训练语言模型（如 BERT），但近期也有一些方法利用最新一代大语言模型（LLM）来完成这项任务。目前的研究侧重于显式跨度识别，例如命名实体识别（NER），而对于大语言模型在基于方面的情感分析（ABSA）等任务中进行更主观的跨度识别，探讨仍不充分。本文填补了这一重要空白，评估了多种大语言模型在三项热门任务中的文本片段识别性能，即情感分析、冒犯性语言识别和主张验证。我们探讨了多种大语言模型策略，例如指令微调、上下文学习和思维链。结果表明，文本中的内在关联有助于大语言模型精准识别文本片段。

1 引言

冒犯性语言识别、情感分- 析和主张验证是社交媒体分析与自然语言处理交叉领域中研究最广泛的一些任务(Sandu等人，2024年)。这些任务的大多数研究都侧重于预测帖子级别的类别标签。例如，在情感分析中，这些标签通常表示为正面、中性和负面标签，或表示从正面到负面连续变化的李克特量表(Birjali等人，2021)。

各种研究通过开发框架、数据集和模型，利用词元片段预测来识别文本中的属性，从而探讨模型的可解释性。例如，在有毒文本片段检测任务中，模型会预测有毒文本中表明有毒标签预测的片段（Pavlopoulos 等人，2021；Mathew 等人，2021）。除了彼此独立的词元跨度之外，研究人员还提出了更具结构化的表述方式，以捕捉文本元素之间的关系。其中最成熟的表述方式之一是基于方面的情感分析（ABSA）（Pontiki 等人，2014年，2015年），其目标是在文本中识别方面及其相关情感。这种方法尤其适用于包含混合情感的情况，例如餐厅评论中的“食物很美味，但服务非常慢”。在这个例子中，文本的不同部分表达了相互矛盾的观点。同样地，在本文中，我们同时考虑复杂文本和简单文本，并将其定义如下：

> **复杂文本**- 包含多种相互关联文本片段的文本，这些相关片段属于不同类别，例如 ABSA 中的目标和方面。
>
> **简单文本**- 仅包含一种片段类别的文本，例如包含有毒表达的有毒片段或包含主张的主张片段。

大语言模型在自然语言处理的多种任务中都取得了最先进的性能，包括生成和预测（Minaee等人，2024年）。近期关于评估大语言模型在序列标注任务（如命名实体识别（NER））中表现的研究（Wang等人，2023；Pang等人，2023）表明，在上下文学习设置下，BERT 模型的表现仍优于大语言模型。 Li等人 （2023）和Duki´candŠnajder（2024年）提出了将大语言模型的目标转换的方法，以提升其在分类任务中的性能。尽管大语言模型已被用于命名实体识别和情感分析任务，但尚未用于其他令牌分类任务，例如冒犯性片段和主张片段识别；我们的工作旨在弥补这一空白。

警告：本文包含一些具有冒犯性质的示例。

本文探讨以下**研究问题**：

- **研究问题1：** 文本的复杂性如何影响大语言模型识别不同类型文本片段的能力？模型识别某些特定跨度类型是否比其他类型更高效？

- **研究问题2：** 模型规模和建模策略如何影响大语言模型的跨度识别能力？

- **研究问题3：** 大语言模型在低资源设置下是否高效？

2 相关工作

冒犯性语言识别、情感分析和主张验证是受到广泛研究的一些文本分类任务。针对冒犯性语言（Davidson等人，2017年；Zampieri 等人，2019年；Ranasinghe和Zampieri，2021；Mathew等人，2021）、情感分析（Tan等人，2023）以及主张验证（Wang，2017年；Thorne等人，2018年；Schlichtkrull等人，2024年），研究人员已经发布了多个带有帖子级标注的数据集。这些任务的大多数方法都依赖于BERT等预训练的基于Transformer的语言模型（Caselli等人，2020；Sarkar等人，2021；Tan等人，2023；Zhang等人，2024年；Dmonte等人，2024a），而近年来也开始探索大语言模型（Pan等人，2023；Zampieri 等人，2023b；Dmonte等人，2024b）。

尽管前述三项任务的大多数研究都针对篇章级分析，但也有研究提出了用于词元级分析的若干数据集和方法。对于冒犯性语言，研究者推出了 TSD（Pavlopoulos 等人，2021年）和 HateXplain（Mathew 等人，2021年）数据集，用于识别包含冒犯性或有害内容的词元跨度，或对预测标签有贡献的特定理由；而 TBO（Zampieri 等人，2023a）则旨在识别冒犯性片段及其相关目标。类似地，ABSA（Pon- tiki 等人，2014年）旨在识别方面——即文本中描述目标或实体的词元跨度——以及与这些方面相关的情感标签。 Wang等人 (2016年，2017年)进一步对该数据集进行了标注，以识别观点词。关系感知协同学习（RACL）（Chen 和 Qian，2020）等方法考虑了不同类型文本片段之间的关系，并取得了良好的效果。

尽管大语言模型尚未广泛用于词元跨度识别任务，但已有一些研究将这些模型应用于若干相近任务。 Han等人 (2023)将 GPT 用于四项任务，即命名实体识别、关系抽取、实体抽取和 ABSA。作者发现，大语言模型的性能低于较小的 BERT 模型。为提升大语言模型在命名实体识别、ABSA 等词元跨度任务上的性能，研究者提出了从大语言模型各层移除因果掩码的方法（Li等人，2023；Duki´c 和 Šnajder，2024年）。尽管这些方法能够提升模型在词元跨度识别任务上的性能，但移除因果掩码会改变模型的训练目标，实质上将模型从自回归模型转变为掩码语言模型。在本研究中，我们利用大语言模型的自回归能力，通过不同方法评估其性能。

3 个数据集

我们为实验收集了四个英语数据集，其中两个包含复杂文本片段，另外两个包含简单文本片段。各数据集的示例实例见表 2，数据集统计信息见表1。

跨度类型	数据集	训练	Test	总计
复杂片段	TBO	4,000	673	4,673
	ABSA	3,041	800	3,841
简单片段	CSI	3,953	362	4,315
	TSD	8,629	2,000	10,629

表1：实验所用数据集中的训练和测试实例数量。

复杂文本数据集 我们获取了 **基于目标的冒犯性语言 (TBO)** (Zampieri 等人，2023a) 以及**基于方面的情感分析 (ABSA)** 数据集，该数据集由 Pontiki 等人 (2014年) 提供。TBO 数据集中的实例标注了论据，即文本中的冒犯性短语，以及代表论据主体的 目标 。ABSA 数据集标注了方面 及其对应的观点 文本片段。

简单文本数据集 我们使用 **主张片段识别（CSI）** (Mittal 等人，2023）和 **有毒片段检测（TSD）** (Pavlopoulos 等人，2021）数据集。CSI 使用来自社交媒体帖子的主张片段进行标注，而 TSD 数据集则针对有毒和有害片段进行标注。

数据集	Type	示例实例	标注
TBO	复杂	@USER 是时候制止选民欺诈了。这些人很邪恶。	目标 1：无，论点 1：选民欺诈，目- 论点 2：这些人，论点 2：邪恶
	复杂	@USER 我才不管呢，我要一直穿着我的 UGG靴，直到它们彻底翻过来！	目标 1：@USER，论点 1：地狱
ABSA	复杂	食物不仅味道出众，那些小小的“福利”也很棒	方面 1：食物，观点 1：出色，方- 面 2：福利
	复杂	raga’s 是一家浪漫、舒适的餐厅	观点 1：浪漫，观点 2：舒适
CSI	简单	不是 Rahul Khan，是它	跨度：不是 Rahul Khan，是它
	简单	他们会想尽一切办法把它偷走。我们绝不会让他们得逞！	跨度：他们会想尽一切办法把它偷走。
TSD	简单	’又一个暴力又具有攻击性的移民，杀害了一名无辜且 聪明的美国公民……讽刺’	论点 1：暴力又具有攻击性的移民
	简单	真是个笨蛋。怎么会有人不知道这会 冒犯人？？	论点 1：笨蛋

表2：各数据集中的示例实例。图中展示了这些实例及其对应的文本跨度标注。复杂文本（TBO 和 ABSA）包含两种跨度类型，而简单文本（CSI 和 TSD）仅包含一种跨度类型。

4 实验

我们介绍实验中使用的模型。BERT 模型使用特定任务数据集进行微调，而大语言模型则采用指令调优、上下文学习和思维链。

基线 我们使用 BERT-large（Devlin 等人，2019年）模型作为实验的基线。我们使用特定任务的训练数据集对模型进行微调。在进展测试中，模型使用从训练集随机抽取的子集进行微调。

大语言模型 我们采用 Qwen2.5（Yang 等人，2024年）和 Llama-3.1（Dubey 等人，2024年）模型系列，因为它们提供多种模型规模，能够评估不同规模的模型。具体而言，我们使用 Qwen 的 7B、14B、32B 和 72B 参数版本，以及 Llama 的 8B 和 70B 参数版本。

方法 我们在实验中采用了三种大语言模型方法。所有大语言模型都针对所有任务进行了指令微调（IT）。特定任务的示例提示词见附录 A.1。我们使用上下文学习（ICL）评估开箱即用的模型。具体而言，我们采用 0、3 和 5 样本方法。对于少样本学习，我们使用句子转换器模型为每个测试实例生成嵌入，并从训练集中选取最相似的前k 个实例作为少样本示例。最后，我们采用零样本思维链（CoT）（Kojima 等人，2022）提示策略来识别词元跨度。

评估指标 我们使用以下两种指标评估模型的性能：**词元 F1（TF1）**计算单个词元的 F1 分数。最终 F1 分数为所有实例的平均值。**跨度 F1（SF1）**考察预测是否与金标准标注完全匹配。F1 分数根据所有实例中正确预测的总数计算。

5 结果

表3 展示了所有任务中表现最佳的 Llama 和 Qwen 模型在词元级和跨度级 F1分数（其他所有模型的性能见表 5）。对于 TBO 和 ABSA 任务，少样本学习表现最佳，其次是指令微调模型。思维链在这两项任务上的表现最差；不过，在识别目标和论元跨度方面，其性能与零样本设置相当。然而，对于 ABSA，与思维链相比，零样本设置在识别方面跨度方面效率更高。

对于简单文本，指令微调模型在 TSD 任务上的表现优于其他方法。然而，在 CSI 任务上，指令微调的 Llama 模型表现不及大多数其他模型和方法。对于这种文本类型，大语言模型在 5 样本设置下的表现略优于 3 样本评估。在零样本设置下，模型在 CSI 任务上的表现与少样本评估相当，而在 TSD 任务上则存在显著差异。对于 TSD 任务，思维链的表现优于零样本设置；而在 CSI 任务上，其表现则较差。

模型	TBO				ABSA				CSI		TSD	
	目标		论元		方面		观点					
	TF1	SF1	TF1	SF1	TF1	SF1	TF1	SF1	TF1	SF1	TF1	SF1
BERT	0.766	0.611	0.779	0.603	0.924	0.843	0.907	0.840	0.573	0.126	0.794	0.652
Llama-70B-CoT	0.504	0.244	0.350	0.093	0.670	0.531	0.548	0.376	0.287	0.078	0.418	0.032
Qwen-72B-CoT	0.545	0.316	0.404	0.122	0.760	0.682	0.543	0.432	0.444	0.183	0.310	0.028
Llama-70B-0	0.523	0.326	0.412	0.183	0.780	0.695	0.652	0.567	0.475	0.176	0.184	0.076
Qwen-32B-0	0.564	0.351	0.286	0.061	0.795	0.732	0.615	0.516	-	-	-	-
Qwen-72B-0	-	-	-	-	-	-	-	-	0.532	0.269	0.148	0.013
Llama-70B-3	0.904	0.844	0.905	0.833	0.836	0.752	0.759	0.689	0.542	0.252	0.612	0.395
Qwen-32B-3	0.920	0.875	0.889	0.852	0.841	0.767	0.695	0.628	0.598	0.305	0.550	0.330
Llama-70B-5	0.856	0.775	0.892	0.816	0.840	0.763	0.760	0.687	0.539	0.247	0.561	0.445
Qwen-32B-5	0.921	0.879	0.894	0.852	0.850	0.786	0.698	0.635	0.595	0.308	0.593	0.376
Llama-70B-IT	0.737	0.606	0.702	0.550	0.824	0.761	0.762	0.706	0.284	0.108	0.777	0.630
Qwen-72B-IT	0.710	0.570	0.701	0.537	0.771	0.690	0.739	0.675	0.639	0.362	0.775	0.638

表 3：TBO、ABSA、CSI 和 TSD 数据集的综合 F1分数。TF1 = 词元级F1，SF1 = 跨度级F1。此处报告每种方法中性能最佳的 Llama 和 Qwen 模型。深青色单元格表示性能较高，较浅的色调表示性能较低

6 讨论

本节将重新探讨第1节中提到的四个研究问题。

研究问题1：文本的复杂度如何影响大语言模型识别不同类型文本片段的能力？模型识别某些特定跨度类型是否比识别其他类型更高效？

如第5节所示，与简单片段相比，大语言模型通常在复杂片段上的性能更好。这些模型更擅长识别文本中明确提及的跨度类型，例如 TBO 中的目标或 ABSA 中的方面。然而，大语言模型可能难以识别主观跨度，例如冒犯性论点或观点词，因为它们依赖上下文或采用间接表达。例如，“你对我来说已经死了”这句话可能被认为具有冒犯性，尽管其中并未包含任何亵渎词语。跨度的歧义、可解释性、隐含性质等因素可能会影响模型性能，例如在 CSI 任务中。这些因素会影响大语言模型识别词元跨度的性能；此外，识别无关词元、将词元跨度拆分为多个不同跨度等因素，也可能导致大语言模型的性能降低。例如，在 TSD 任务中，大语言模型往往不仅识别有害跨度，还会识别上下文词语，如图6所示。这些因素尤其会导致某些任务的跨度级 F1 分数较低。

我们还旨在确定文本复杂度如何影响大语言模型识别不同跨度类型的性能。对于两个复杂任务 TBO 和 ABSA，我们使用零样本设置下的一百个实例样本，评估模型单独识别和同时识别不同跨度类型的效果。如图1所示，当提示模型同时识别两种跨度类型时，其表现优于分别提示模型识别这两种类型的情况。这表明，文本的复杂性以及不同跨度类型之间的内在关系，有助于大语言模型准确识别不同跨度类型。

研究问题2：模型规模和建模策略如何影响大语言模型的跨度识别能力？

结果表明，对于复杂文本，少样本设置下的大语言模型优于所有其他方法。提示词中的上下文示例有助于模型识别不同类型的文本片段。思维链的表现不如零样本学习，而指令调优提升了模型性能，尤其是在识别主观文本片段方面。对于 CSI 和 TSD，指令微调优于所有其他方法。在 TSD 任务中，模型在零样本和思维链设置下表现不佳。对输出的分析表明，某些方法会生成无关或冗余的文本片段。此外，在 TBO 和 CSI 任务中，少样本学习优于微调后的 BERT 模型；而在 TSD 任务中，指令调优的性能与 BERT 相当。

Kaplan等人(2020) 表明，随着

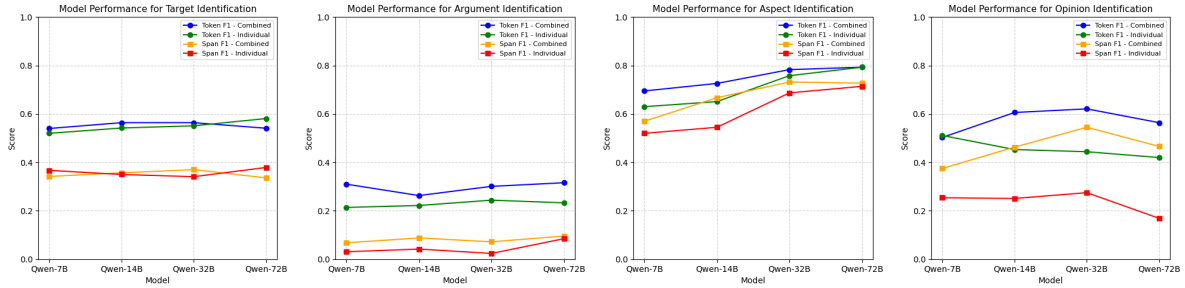


图 1：复杂文本跨度识别任务的词元 F1 和跨度级F1分数。图中展示了分别单独提取和合并提取每种文本片段时的分数。

模型规模增大，性能会提升。为验证这一假设，我们对参数量从 7B 到 72B 的不同模型规模进行了实验。实验结果表明，尽管模型规模不断增大，性能提升却十分有限。总体结果显示，在 Qwen 模型中，参数量为 7B 的模型性能最差；而在 14B、32B 和 72B 模型中，任一模型的性能都优于其他模型。对于 Llama 模型，70B 模型的性能始终优于 8B 模型。然而，不同参数规模模型之间的性能差异仅为小幅差异。考虑到更大模型需要大量计算资源，而性能提升又十分有限，较小的模型或许是执行这些任务时更高效的选择。

研究问题3：大语言模型在低资源设置下是否高效?

训练语言模型需要大量训练数据。然而，某些令牌分类任务可能面临数据稀缺。因此，为了评估数据稀缺对语言模型性能的影响，我们使用数量不等的训练数据（从 200 到 1000 个样本）对小型语言模型（SLM）和大语言模型（LLM）进行微调。在这项特定实验中，我们比较 BERT-large 和 Qwen-7B 模型的性能。

我们的实验表明，在 TBO、ABSA 和 TSD 任务中，对于所有数据规模，BERT 的表现均优于 Qwen- 7B 模型（见图 7）。性能会因跨度类型和所用训练示例的数量而异；对于某些跨度类型，两者表现相当，而对于其他类型，差异则较为显著。然而，在 CSI 任务上，小型语言模型的表现优于大语言模型，尤其是在跨度 F1 分数方面。这表明，大语言模型比更小的模型更能准确识别确切的主张文本片段。研究结果表明，当使用有限的标注训练数据进行微调时，小语言模型通常优于大语言模型。不过，如表3所示，在此类场景中，大语言模型的少样本学习凭借较高的性能也值得采用。

7 结论

在这项工作中，我们评估了多种大语言模型采用不同方法进行主观跨度识别的效果。我们回答了有关文本复杂度和模型规模的重要研究问题，并进一步探究了大语言模型在低资源设置下的能力。我们的研究结果表明，文本的复杂性及其内在关系有助于大语言模型识别精确的文本片段。此外，对于跨度识别这一特定任务，模型规模对性能没有显著影响。尽管 BERT 等小语言模型的表现仍优于大语言模型，但在低资源设置下可以采用少样本学习等方法。大语言模型在显式且上下文无关的跨度识别方面表现出色，但在识别主观跨度时，表现仍不如较小的模型。

在未来的工作中，我们希望探索改进大语言模型对输入内容理解的方法，以实现准确的跨度识别，尤其是同时考虑左侧和右侧的上下文。我们还计划探索其他与主观任务相关的具有挑战性的数据集，包括文本与图像配对的多模态数据（Farabi 等，2024年）。最后，我们计划将这项工作扩展到非英语数据集，旨在评估当前一代大语言模型的多语言能力。

致谢

作者谨此感谢匿名审稿人富有洞见的反馈。

Marcos Zampieri 部分获得弗吉尼亚联邦网络倡议（CCI）项目编号 N-4Q24-009 的资助。

局限性

本文的主要局限性在于，我们仅评估了两个开源模型系列。然而，其他开源模型或专有模型可能达到相当的性能。此外，实验中使用的提示词遵循特定的模板风格。尝试使用不同的提示词模板可能会产生不同的结果。提示词中的任务特定指令可以进行调整，以生成更高效的输出。在少样本实验中，我们在提示词中使用了三个和五个示例。然而，加入更多示例可以提升模型的性能。此外，我们的评估集中于英语数据集。将这项工作扩展到更多语言和任务数据集，可能会进一步揭示大语言模型的词元跨度识别能力。

参考文献

Marouane Birjali、Mohammed Kasri 和 Abderrahim Beni-Hssane。2021。情感分析的全面综述：方法、挑战与趋势。《*Knowledge-Based Systems*》，226。

Tommaso Caselli、Valerio Basile、Jelena Mitrovi´c 和 Michael Granitzer。2020年。HateBERT：为英语辱骂性语言检测重新训练 BERT。*arXiv* 预印本 *arXiv:2010.12472*。

Zhuang Chen 和 Tieyun Qian。2020。用于统一方面级情感分析的关系感知协作学习。载于*ACL* 论文集。

Thomas Davidson、Dana Warmsley、Michael Macy 和 Ingmar Weber。2017年。自动仇恨言论检测与冒犯性语言问题。载于*ICWSM* 论文集。

Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019年。BERT：用于语言理解的深度双向变换器预训练。载于*NAACL* 论文集。

Alphaeus Dmonte、Eunmi Ko 和 Marcos Zampieri。2024a。《大语言模型在金融情感分析中的评估》。载于大数据论文集。

Alphaeus Dmonte、Roland Oruche、Marcos Zampieri、Prasad Calyam 和 Isabelle Augenstein。2024b。《大语言模型时代的主张验证：综述》。*arXiv* 预印本 *arXiv:2408.14317*。Abhimanyu Dubey、Abhinav Jauhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Amy Yang、Angela Fan 等。2024。《Llama 3 模型家族》。*arXiv* 预印本 *arXiv:2407.21783*。David Duki´c 和 Jan Šnajder。2024。《看向正确方向有时就是正确的：探究仅解码器大语言模型进行序列标注的能力》。载于*ACL* 论文集（*ACL 2024*）。

Shafkat Farabi、Tharindu Ranasinghe、Diptesh Kanojia、Yu Kong 和 Marcos Zampieri。2024年。多模态讽刺检测综述。载于*AAAI* 论文集。

Ridong Han、Tao Peng、Chaohao Yang、Benyou Wang、Lu Liu 和 Xiang Wan。2023。信息抽取是否已被 ChatGPT 解决？对性能、评估标准、鲁棒性和错误的分析。*arXiv* 预印本 *arXiv:2305.14450*。

Jared Kaplan、Sam McCandlish、Tom Henighan、TomB Brown、Benjamin Chess、Rewon Child、Scott Gray、Alec Radford、Jeffrey Wu 和 Dario Amodei。2020。神经语言模型的缩放定律。*arXiv* 预印本 *arXiv:2001.08361*。

Takeshi Kojima、Shixiang Shane Gu、Machel Reid、Yutaka Matsuo 和 Yusuke Iwasawa。2022。大型语言模型是零样本推理器。载于*NeurIPS* 会议论文集。

Zongxi Li、Xianming Li、Yuzhang Liu、HaoranXie、Jing Li、Fu-lee Wang、Qing Li 和 Xiaoqin Zhong。2023。标签监督的 Llama 微调。*arXiv* 预印本 *arXiv:2310.01208*。

Binny Mathew、Punyajoy Saha、Seid Muhie Yimam、Chris Biemann、Pawan Goyal 和 Animesh Mukher- jee。2021。Hatexplain：用于可解释仇恨言论检测的基准数据集。收录于*AAAI* 论文集。

Shervin Minaee、Tomas Mikolov、Narjes Nikzad、Meysam Chenaghlu、Richard Socher、Xavier Amatriain 和 Jianfeng Gao。2024年。大语言模型：综述。*arXiv* 预印本 *arXiv:2402.06196*。

Shubham Mittal、Megha Sundriyal 和 Preslav Nakov。2023。迷失于翻译，寻获于文本片段：识别多语言社交媒体中的主张。收录于*EMNLP* 论文集。

Liangming Pan、Xiaobao Wu、Xinyuan Lu、Anh Tuan Luu、William YangWang、Min-Yen Kan 和 Preslav Nakov。2023。通过程序引导推理核查复杂主张。收录于*ACL* 论文集。

Chaoxu Pang、Yixuan Cao、Qiang Ding 和 Ping Luo。2023。用于上下文内信息抽取的指南学习。载于*EMNLP* 论文集。

John Pavlopoulos、Jeffrey Sorensen、Léo Laugier 和 Ion Androutsopoulos。2021。SemEval-2021 任务5：有毒文本片段检测。载于*SemEval* 论文集。

Maria Pontiki、Dimitris Galanis、Haris Papageorgiou、Suresh Manandhar 和 Ion Androutsopoulos。2015年。SemEval-2015 任务12：基于方面的情感分- 析。载于 *SemEval* 论文集。

Maria Pontiki、Dimitris Galanis、John Pavlopoulos、Har- ris Papageorgiou、Ion Androutsopoulos 和 Suresh Manandhar。2014年。SemEval-2014 任务4：基于方面的情感分析。载于*Se- mEval* 论文集。

Tharindu Ranasinghe 和 Marcos Zampieri。2021。Mudes：多语言冒犯性文本片段检测。载于*NAACL* 论文集。

Andra Sandu、Liviu-Adrian Cotfas、Aurelia Stˇ anescu 和 Camelia Delcea。2024年。文本挖掘的文献计量分析：探索自然语言处理在社交媒体研究中的应用。*Applied Sciences*，14(8):3144。Diptanu Sarkar、Marcos Zampieri、Tharindu Ranasinghe 和 Alexander Ororbia。2021。f bert：一种用于识别冒犯性内容的神经 Transformer。在*ACL* 论文集（*EMNLP 2021*）。Michael Schlichtkrull、Zhijiang Guo 和 Andreas Vlachos。2024。Averitec：一个基于网络证据、用于真实世界主张验证的数据集。在 *NeurIPS* 论文集。Kian Long Tan、Chin Poo Lee 和 Kian Ming Lim。2023。情感分析综述：方法、数据集与未来研究。*Applied Sciences*，13(7):4550。James Thorne、Andreas Vlachos、Christos Christodoulopoulos 和 Arpit Mittal。2018。Fever：用于事实抽取和验证的大规模数据集。在 *NAACL*。Shuhe Wang、Xiaofei Sun、Xiaoya Li、Rongbin Ouyang、Fei Wu、Tianwei Zhang、Jiwei Li 和 Guoyin Wang。2023。GPT-NER：通过大语言模型进行命名实体识别。*arXiv*预印本 *arXiv:2304.10428*。Wenya Wang、Sinno Jialin Pan、Daniel Dahlmeier 和 Xiaokui Xiao。2016年。用于基于方面的情感分析的递归神经条件随机场。在 *EMNLP*论文集。Wenya Wang、Sinno Jialin Pa n、Daniel Dahlmeier 和 Xiaokui Xiao。2017年。用于方面和观点词联合抽取的耦合多层注意力。在 *AAAI*论文集。William Yang Wang。2017年。“骗子，骗子，裤子着火了”：用于假新闻检测的新基准数据集。在*ACL*。An Yang、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chengyuan Li、Dayiheng Liu、Fei Huang、Haoran Wei 等。2024。Qwen2.5技术报告。*arXiv*预印本 *arXiv:2412.15115*。

Marcos Zampieri、Shervin Malmasi、Preslav Nakov、Sara Rosenthal、Noura Farra 和 Ritesh Kumar。2019年。预测社交媒体中攻击性帖子的类型和目标。在*NAACL*论文集。

Marcos Zampieri、Skye Morgan、Kai North、Tharindu Ranasinghe、Austin Simmons、Paridhi Khandelwal、Sara Rosenthal 和 Preslav Nakov。2023a。基于目标的冒犯性语言识别。载于 *ACL* 论- 文集。

Marcos Zampieri、Sara Rosenthal、Preslav Nakov、Al- phaeus Dmonte 和 Tharindu Ranasinghe。2023b。OffensEval 2023：大语言模型时代的冒犯性语言识别。自然语言工程。

Wenxuan Zhang、Yue Deng、Bing Liu、Sinno Pan 和 Lidong Bing。2024年。大语言模型时代的情感分析：现实检验。载于*NAACL* 论文集。

A 附录

A.1 大语言模型提示

图 2 展示了我们实验中使用的提示模板。表4展示了特定任务提示。每个提示都包含用于描述任务及各类待提取片段的指令。此外，提示还包含输入实例。

<任务特定指令。此处包括对待识别文本片段类型的定义>

输出格式：
<生成的输出格式为“标签：文本片段”。例如，目标：目标文本片段>

示例：
<*n* 个示例>

输入：
<输入文本>

响应：

图 2：我们在实验中使用的提示模板。

A.2 超参数

为了微调模型，我们尝试了多个学习率值，其中 1e-4 可获得最佳性能和最低平均损失。我们使用每设备批次大小为 2、梯度累积大小为 8 的设置对大语言模型进行指令微调。由于可用的计算资源有限，我们选择了较小的批次大小。我们还采用了 Adam 优化器，并对模型进行了十个训练轮次的微调。对于 LoRA，我们使用 alpha 值 16 和 r 值 64，因为这些值能够带来最佳性能。丢弃率设为 0.1。为了使用思维链和上下文学习评估大语言模型，我们将温度设为 0.0001（因为 Llama 等模型不允许将温度设为 0），以获得确定性输出。

A.3 额外结果

在本节中，我们展示实验的额外结果。虽然表 3展示了性能最佳的 Qwen 和 Llama 模型的 F1 分数，但我们在表 5中展示了所有其他模型的结果。

A.4 任务特定输出

我们在图 3-6中展示了所有任务的输出。对于每项任务，我们展示了 Llama 模型采用不同方法生成的输出。对于少样本设置，我们展示了 5 样本实验的输出。

A.5 进展测试

图 7 展示了在不同训练数据规模下训练时，Qwen-7B 与 BERT 的性能对比。我们进行这项实验，以了解大语言模型在低资源设置下的表现。对于 TBO 和 ABSA 任务，BERT 的表现优于 Qwen-7B。类似地，在使用 200 个实例进行微调时，BERT 在 TSD 任务上的表现优于 Qwen-7B；但随着训练实例数量增加，两者的 F1 分数差距逐渐缩小。与其他三项任务不同，Qwen-7B 的表现优于 BERT，两个模型的跨度 F1 差距显著。不过，随着训练数据集规模扩大，该任务的词元 F1 差距逐渐缩小。该实验表明，考虑到任务类型不同，小型语言模型与大语言模型的性能也有所不同。

数据集	提示
TBO	请从给定文本中识别（目标，论据）对。 •目标：论据所针对的个人、群体或组织。 •论元：包含冒犯性、亵渎性或不可接受的语言的短语或句子。 • 论据可能有目标，也可能没有。 • 如果多个论据都指向同一目标，该目标可以出现多次。 • 论元和目标可以相同。 输出格式： 目标 n： <第 n 个目标> 论元 n： <第 n 个论元> 示例：{n examples} 输入： {input text} 响应：
ABSA	请从给定文本中识别方面和观点。 • 方面：与情感相关联的实体。 •观点：表达情感的词语或短语。 输出格式： 方面 n： <第 n 个方面> 观点 n： <第 n 个观点> 示例：{n examples} 输入： {input text} 响应：
CSI	请在给定文本中识别主张跨度。 • 主张片段：明确提及主张、断言或论点的短语或句子。 输出格式： 片段 n： <第 n 个主张片段> 示例：{n examples} 输入： {input text} 响应：
TSD	请识别给定文本中的论据。 •论元：包含冒犯性、亵渎性或不可接受的语言的短语或句子。 输出格式： 论元 n： <第 n 个论元> 示例：{n examples} 输入： {input text} 响应：

表 4：各任务使用的提示词。提示词包含一条任务特定指令以及相应的输入实例。

模型	TBO				ABSA				CSI		TSD	
	目标		论元		方面		观点					
	TF1	SF1	TF1	SF1	TF1	SF1	TF1	SF1	TF1	SF1	TF1	SF1
Llama-8B-CoT	0.533	0.171	0.362	0.052	0.569	0.333	0.558	0.236	0.203	0.043	0.515	0.012
Qwen-7B-CoT	0.518	0.282	0.296	0.060	0.649	0.509	0.507	0.340	0.429	0.210	0.262	0.029
Qwen-14B-CoT	0.541	0.151	0.329	0.017	0.607	0.457	0.582	0.410	0.443	0.188	0.323	0.019
Qwen-32B-CoT	0.578	0.329	0.316	0.056	0.576	0.360	0.613	0.302	0.444	0.169	0.425	0.025
Llama-8B-0	0.516	0.310	0.225	0.044	0.707	0.574	0.524	0.401	0.469	0.198	0.135	0.036
Qwen-7B-0	0.531	0.314	0.302	0.078	0.721	0.606	0.526	0.404	0.517	0.275	0.321	0.024
Qwen-14B-0	0.563	0.352	0.255	0.048	0.726	0.648	0.587	0.436	0.516	0.227	0.181	0.010
Qwen-32B-0	-	-	-	-	-	-	-	-	0.528	0.223	0.253	0.008
Qwen-72B-0	0.544	0.332	0.313	0.090	0.776	0.705	0.564	0.462	-	-	-	-
Llama-8B-3	0.880	0.792	0.881	0.786	0.803	0.681	0.700	0.577	0.503	0.248	0.559	0.299
Qwen-7B-3	0.898	0.834	0.881	0.812	0.791	0.709	0.628	0.563	0.539	0.275	0.444	0.244
Qwen-14B-3	0.913	0.864	0.897	0.848	0.816	0.746	0.662	0.586	0.582	0.295	0.542	0.302
Qwen-72B-3	0.852	0.791	0.873	0.812	0.835	0.779	0.686	0.627	0.562	0.281	0.550	0.343
Llama-8B-5	0.836	0.735	0.874	0.759	0.801	0.680	0.716	0.592	0.495	0.228	0.597	0.342
Qwen-7B-5	0.899	0.840	0.889	0.824	0.804	0.728	0.645	0.586	0.530	0.266	0.502	0.290
Qwen-14B-5	0.880	0.816	0.887	0.824	0.830	0.761	0.671	0.603	0.587	0.292	0.582	0.349
Qwen-72B-5	0.838	0.776	0.868	0.801	0.845	0.789	0.708	0.645	0.573	0.277	0.586	0.388
Llama-8B-IT	0.707	0.539	0.692	0.537	0.790	0.717	0.755	0.670	0.178	0.000	0.768	0.629
Qwen-7B-IT	0.645	0.392	0.644	0.441	0.730	0.619	0.571	0.471	0.592	0.316	0.747	0.590
Qwen-14B-IT	0.678	0.482	0.673	0.514	0.776	0.689	0.715	0.643	0.612	0.334	0.753	0.589
Qwen-32B-IT	0.669	0.475	0.654	0.485	0.783	0.712	0.700	0.646	0.590	0.319	0.756	0.600

表 5：0、3、5 样本、思维链及指令调优（IT）模型的 F1分数。报告了各跨度类型的 TF1 和 SF1分数。

输入： 我这会儿要在推特上发一堆彻头彻尾的胡说八道和荒唐话，毕竟都这么晚了

零样本：
目标 1：我
论点 1：彻头彻尾的胡说八道和荒唐话

---

5 样本：
目标 1：nan
论点 1：纯属胡说八道
目标 2：无
论点 2：一派胡言

指令微调：
目标 1：无
论点 1：纯属胡说八道
目标 2：无
论点 2：纯属胡说八道

思维链：
让我们一步一步地思考。文本中包含一个论元："一派胡言，纯属胡说八道"。
这个论元的目标并未明确提及，但可以推断，目标是说话者正在称呼的那个人
或那些人。因此，输出应为：
目标 1：说话者所称呼的一个人或多个人
论点 1：纯属胡说八道，完全是无稽之谈

图 3：TBO 任务的输出。展示了每种方法的输出。对于少样本，仅展示 5 样本输出。

输入： 晚饭后，我听到正在播放的音乐，并发现楼下有一间休息室

零样本：
方面 1：音乐
观点 1：正在播放
方面 2：休息室
观点 2：楼下

---

5 样本：
方面 1：音乐
方面 2：休息室
观点 1：正在播放
方面 3：楼下

指令微调：
方面 1：音乐
观点 1：演奏
方面 2：休息室
观点 2：楼下

思维链：
让我们一步步思考。首先，我们需要识别给定文本中的方面。在本例中，
方面是“正在播放的音乐”和“楼下的休息室”。
接下来，我们需要识别文本中的观点。观点是“我听到了正在播放的音乐”和“我
发现楼下有一间休息室”。
因此，输出格式如下：
方面 1：正在播放的音乐
观点 1：我听到了正在播放的音乐
方面 2：楼下的休息室
观点 2：我发现楼下有一间休息室

图 4：O utputs（输出）展示了 ABSA 任务的结果。图中展示了各方法的输出。对于少样本方法，仅展示 5 样本的输出 ut is shown
（已展示）。

输入： 接种疫苗能保护你免受飓风侵袭？呃……布兰登加油！

零样本：
片段 1：接种疫苗能保护你免受飓风侵袭?
片段 2：呃……布兰登加油！

---

5 样本：
片段 1：接种疫苗能保护你免受飓风侵袭?
片段 2：呃……布兰登加油！

指令微调：
片段 1：接种疫苗可以保护你免受飓风侵袭吗?
片段 2：呃……布兰登加油！

思维链：
让我们一步一步地思考。文本中包含一个主张片段，即短语“接种疫苗可以保护
你免受飓风侵袭？”这是说话者提出的断言或论点。因此，输出
将是：
片段 1：接种疫苗能保护你免受飓风侵袭?

图 5：CSI 任务的输出。图中展示了每种方法的输出。对于少样本设置，仅展示 5 样本输出。

输入： 这个提议愚蠢至极，注定会失败。

零样本：
论点 1：这项提案愚蠢至极，注定很快失败。

---

5 样本：
论点 1：愚蠢的

指令微调：
论点 1：这项提案愚蠢至极，注定很快失败。

思维链：
让我们逐步思考。给定的文本是：“这个提案愚蠢至极，失败已不可避免。”
步骤 1：找出包含冒犯性、亵渎性或不可接受的语言的短语或句子。
- 在本例中，“愚蠢的”可被视为冒犯性或不可接受的语言。
步骤 2：根据已识别的短语或句子确定论据。
- 论点 1：“愚蠢的”
输出格式：
论点 1：愚蠢的

图 6：TSD 任务的输出。图中展示了每种方法的输出。对于少样本方法，仅展示五样本输出。

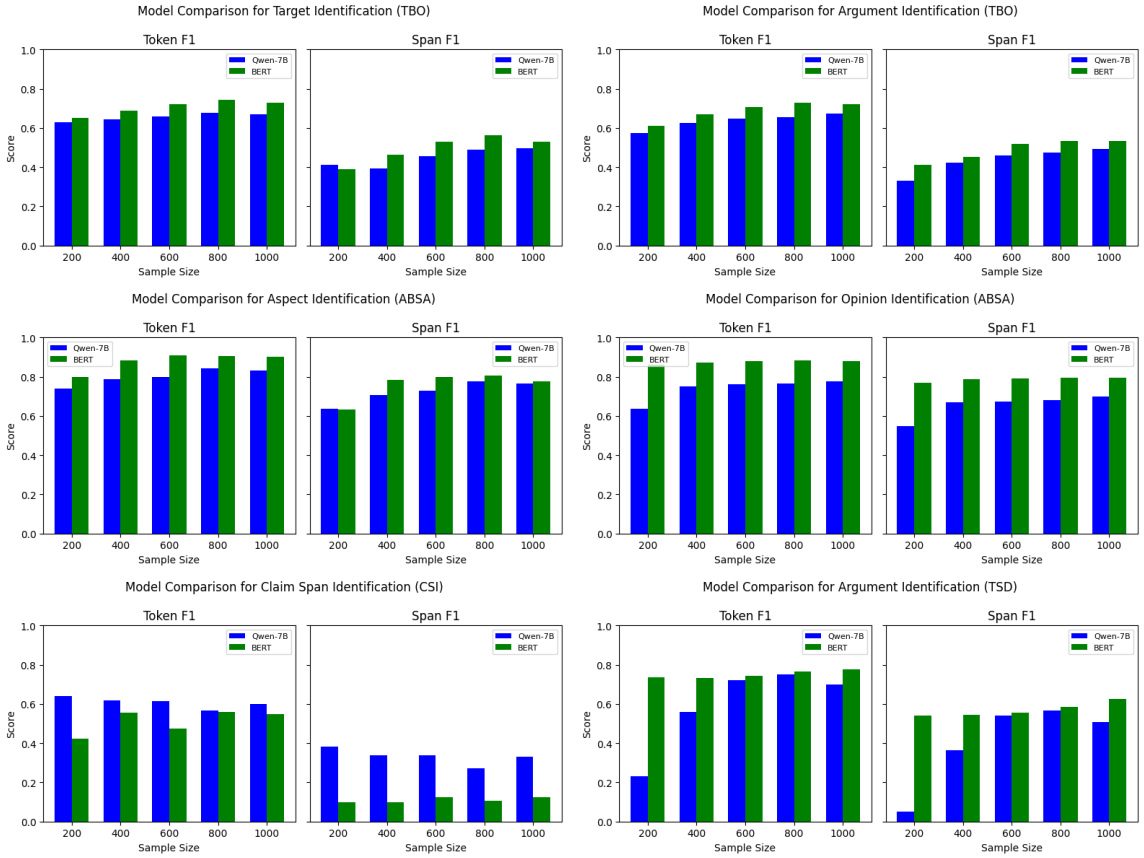


图 7：进展测试结果。报告了每种跨度类型的 TF1 和 SF1分数