Exploring the Performance of Large Language Models on Subjective Span Identification Tasks

Alphaeus Dmonte[1], Roland Oruche[2], Tharindu Ranasinghe[3]
Marcos Zampieri[1], Prasad Calyam[2]

[1]George Mason University, VA, USA [2]University of Missouri, MO, USA
[3]Lancaster University, UK
admonte@gmu.edu

Abstract

Identifying relevant text spans is important for several downstream tasks in NLP, as it contributes to model explainability. While most span identification approaches rely on relatively smaller pre-trained language models like BERT, a few recent approaches have leveraged the latest generation of Large Language Models (LLMs) for the task. Current work has focused on explicit span identification like Named Entity Recognition (NER), while more subjective span identification with LLMs in tasks like Aspect-based Sentiment Analysis (ABSA) has been underexplored. In this paper, we fill this important gap by presenting an evaluation of the performance of various LLMs on text span identification in three popular tasks, namely sentiment analysis, offensive language identification, and claim verification. We explore several LLM strategies like instruction tuning, in-context learning, and chain of thought. Our results indicate underlying relationships within text aid LLMs in identifying precise text spans.

1 Introduction

Offensive language identification, sentiment analysis, and claim verification are some of the most widely studied tasks at the intersection of social media analysis and NLP (Sandu et al., 2024). Most of the research on these tasks focuses on predicting post-level categorical labels. In the case of sentiment analysis, for example, these are often expressed in terms of *positive*, *neutral*, and *negative* labels or a Likert-scale representing the positive to negative continuum (Birjali et al., 2021).

Various studies have addressed model explainability by developing frameworks, datasets, and models to identify attributes in texts through token span prediction. For example, in the toxic spans detection task, models predict the spans of toxic posts that are indicative of toxic label prediction (Pavlopoulos et al., 2021; Mathew et al., 2021). Going beyond independent token spans, researchers have also proposed more structured formulations to capture relationship between textual elements. One of the most well-established of these formulations is Aspect-Based Sentiment Analysis (ABSA) (Pontiki et al., 2014, 2015), which aims to detect aspects and their associated sentiments within a text. This approach is particularly effective for cases with mixed sentiments, such as *"The food was delicious, but the service was extremely slow"* in a restaurant review. In this example different parts of the text express opposing opinions. In the same vein, in this paper we consider both complex and simple texts and define them as follows:

> **Complex Text** - A text containing more than one type of interrelated spans, and these related spans belong to different categories, such as TARGET and ASPECT in ABSA.
>
> **Simple Text** - A text with only one span category such as a *toxic span* or *claim span* containing a toxic expression and a claim respectively.

LLMs have achieved state-of-the-art performance across various NLP tasks, including generation and prediction (Minaee et al., 2024). Recent studies on evaluating LLMs for sequence labeling tasks such as Named Entity Recognition (NER) (Wang et al., 2023; Pang et al., 2023) suggest that BERT models still outperform LLMs in the in-context learning setting. Li et al. (2023) and Dukić and Šnajder (2024) proposed approaches that transform the objective of LLMs to improve their performance on classification tasks. While LLMs have been explored for NER and sentiment analysis tasks, they have been unexplored for other token classification tasks like offensive spans and claim spans identification, and our work aims to address this gap.

WARNING: This paper contains examples that are offensive in nature.

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 358–371

This paper addresses the following **research questions**:

- **RQ1:** How does the complexity of the text affect the LLM's ability to identify the different types of spans? Do the models identify specific span types more efficiently than others?
- **RQ2:** How do the model size and modeling strategies influence the span identification capabilities of LLMs?
- **RQ3:** Are LLMs efficient in a low-resource setting?

2 Related Work

Offensive language identification, sentiment analysis, and claim verification are some of the widely studied text classification tasks. Several datasets with post-level annotations have been released for offensive language (Davidson et al., 2017; Zampieri et al., 2019; Ranasinghe and Zampieri, 2021; Mathew et al., 2021), sentiment analysis (Tan et al., 2023), as well as claim verification (Wang, 2017; Thorne et al., 2018; Schlichtkrull et al., 2024). Most of the approaches to these tasks rely on pre-trained transformer-based language models like BERT (Caselli et al., 2020; Sarkar et al., 2021; Tan et al., 2023; Zhang et al., 2024; Dmonte et al., 2024a) while, more recently, LLMs have also been explored (Pan et al., 2023; Zampieri et al., 2023b; Dmonte et al., 2024b).

While most of the work on the aforementioned three tasks addresses post-level analysis, several datasets and approaches for token-level analysis have also been proposed. For offensive language, the TSD (Pavlopoulos et al., 2021) and HateXplain (Mathew et al., 2021) datasets were introduced to identify the token spans containing offensive or toxic content, or specific rationales contributing to the predicted labels while TBO (Zampieri et al., 2023a) was created, to identify the offensive spans and their associated targets. Similarly, ABSA (Pontiki et al., 2014) aims to identify the aspects, which are token spans from the text describing the targets or entities, as well as the sentiment labels associated with these aspects. Wang et al. (2016, 2017) further annotated this dataset to identify the opinion terms. Approaches like Relation-aware Collaborative Learning (RACL) (Chen and Qian, 2020), which considers the relationship between the different types of spans, have showed promising results.

While LLMs have not been used extensively for token span identification tasks, there are some works that have leveraged these models on a few similar tasks. Han et al. (2023) leveraged GPT for four tasks, namely NER, Relation Extraction, Entity Extraction, and ABSA. The authors observe that LLMs achieve lower performance compared to smaller BERT-based models. To improve the LLM performance on token spans tasks like NER, ABSA, etc, approaches that remove the causal mask from the LLM layers have been proposed (Li et al., 2023; Dukić and Šnajder, 2024). While these approaches improve the performance on the token spans identification tasks, removing the causal mask changes the training objective of the models, essentially transforming the model from autoregressive to a masked language model. In this work, we leverage the autoregressive capabilities of the LLMs to evaluate their performance using different approaches.

3 Datasets

We acquire four English datasets for our experiments, two with complex text spans and two with simple text spans. The example instances from each dataset are presented in Table 2, while the data statistics are presented in Table 1.

Span Type	Dataset	Train	Test	Total
Complex Spans	TBO	4,000	673	4,673
	ABSA	3,041	800	3,841
Simple Spans	CSI	3,953	362	4,315
	TSD	8,629	2,000	10,629

Table 1: Number of Train and Test instances in the datasets used for the experiments.

Complex Text Datasets We acquire **Target Based Offensive Language (TBO)** (Zampieri et al., 2023a) and the **Aspect Based Sentiment Analysis (ABSA)** dataset by Pontiki et al. (2014). The instances in the TBO dataset were annotated with *Arguments*, which are offensive phrases in the text, and *Target* representing the subject of the arguments. The ABSA dataset is annotated with *Aspects* and their corresponding *Opinion* spans.

Simple Text Datasets We use the **Claim Spans Identification (CSI)** (Mittal et al., 2023) and **Toxic Span Detection (TSD)** (Pavlopoulos et al., 2021) datasets. CSI is annotated with claim spans from social media posts, while the TSD dataset is annotated for toxic and harmful spans.

Dataset	Type	Example Instance	Annotation
TBO	Complex	@USER Time to stop the voter fraud. These people are evil.	Target 1: None, Argument 1: voter fraud, Target 2: these people, Argument 2: are evil
	Complex	@USER Imma wear my uggs until they turn inside out the hell!	Target 1: @USER, Argument 1: hell
ABSA	Complex	not only was the food outstanding, but the little 'perks' were great	Aspect 1: food, Opinion 1: outstanding, Aspect 2: perks
	Complex	raga's is a romantic, cozy restaurant	Opinion 1: romantic, Opinion 2: cozy
CSI	Simple	It's not Rahul Khan, it	Span: It's not Rahul Khan, it
	Simple	They will try everything to steal it. We will not let them!	Span: They will try everything to steal it.
TSD	Simple	'Another violent and aggressive immigrant killing a innocent and intelligent US Citizen.... Sarcasm'	Argument 1: violent and aggressive immigrant
	Simple	What a knucklehead. How can anyone not know this would be offensive??	Argument 1: knucklehead

Table 2: Example instances from each dataset. The instances, along with their respective text span annotations, are shown. Complex text (TBO and ABSA) have two types of spans, while simple text (CSI and TSD) have only one span type.

4 Experiments

We describe the models used in our experiments. BERT models are fine-tuned with task-specific datasets, while instruction-tuning, in-context learning, and chain-of-thought are used for LLMs.

Baselines We use the BERT-large (Devlin et al., 2019) model as a baseline for our experiments. We fine-tune the model with the task-specific training datasets. For the progress test, the models are fine-tuned with a randomly sampled subset of the training set.

LLMs We utilize the Qwen2.5 (Yang et al., 2024) and Llama-3.1 (Dubey et al., 2024) model families due to the availability of multiple model sizes, enabling evaluation across different model scales. Specifically, we employ the 7B, 14B, 32B, and 72B parameter variants for Qwen, and 8B and 70B for Llama.

Approaches We utilize three LLM approaches in our experiments. All LLMs are instruction-tuned (IT) on all the tasks. The task-specific example prompts are shown in Appendix A.1. In-context Learning (ICL) is used to evaluate off-the-shelf models. More specifically, 0-, 3-, and 5-shot approaches are used. For few-shot learning, an embedding for each test instance is generated using a sentence-transformer model, and top-k similar instances from the training set are used as few-shot exemplars. Finally, we employ the zero-shot chain-of-thought (CoT) (Kojima et al., 2022) prompting strategy for token spans identification.

Evaluation Metrics We evaluate the performance of the models, using the following two metrics: **Token F1 (TF1)** calculates the F1 score, considering the individual tokens. The final F1-score is the average across all the instances. **Span F1 (SF1)** considers the exact match with the gold standard annotation. The F1 score is calculated considering the total correct predictions across all instances.

5 Results

Table 3 show the token-level and span-level F1 scores of the best-performing Llama and Qwen models for all the tasks (the performance of all other models is shown in Table 5). For TBO and ABSA tasks, few-shot learning achieves the best performance, followed by instruction-tuned models. CoT is the least performing on both these tasks, however, the performance is comparable to zero-shot, in identifying the target and argument spans. However, for ABSA, identifying the aspect spans is more efficient in zero-shot setting compared to CoT.

On simple texts, the instruction-tuned models outperform other approaches for the TSD task. However, instruction-tuned Llama models underperform most other models and approaches for CSI. The 5-shot performance of LLMs for this text type is slightly better than the 3-shot evaluation. In the zero-shot setting, the models achieve a comparable performance to the few-shot evaluation for the CSI, whereas there is a significant performance difference for the TSD. CoT achieves a better performance than the zero-shot for the TSD task, while it underperforms for the CSI task.

Model	TBO				ABSA				CSI		TSD	
	Target		Argument		Aspect		Opinion					
	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

Table 3: Combined F1 scores across TBO, ABSA, CSI, and TSD datasets. TF1 = Token-level F1, SF1 = Span-level F1. Here we report the best performing Llama and Qwen model for each approach. Dark Teal cell indicates a higher performance while a lighter tone indicates a lower performance

6 Discussion

In this section, we revisit the four research questions mentioned in Section 1.

RQ1: How does the complexity of the text affect the LLM's ability to identify the different types of spans? Do the models identify specific span types more efficiently than others?

As seen in Section 5, LLMs generally have a better performance on complex spans compared to simple spans. These models are more efficient at identifying the span types that are explicitly mentioned in the text, like targets in TBO or aspects in ABSA. However, LLMs may struggle to identify subjective spans like offensive arguments or opinion terms, that are context-dependent or indirect expressions. For example, the sentence "You are dead to me" may be perceived as offensive, although it does not contain any profane words. Several factors like ambiguity, interpretability, implicit nature, etc, of the spans can influence the model performance, for example in CSI task. While these factors influence the LLM performance in identifying the token spans, several other factors, like identifying irrelevant tokens, splitting the token spans into multiple distinct spans, etc, can also contribute to a lower performance for the LLMs. For example, in TSD task, LLMs tend to identify not only the toxic spans but also the context words, as seen in Figure 6. Such factors especially contribute to the lower Span F1 scores for certain tasks.

We also aim to identify how text complexity affects the LLMs performance in identifying different types of spans. For the two complex tasks, TBO and ABSA, we assess how well the models identify different span types - individually and combined, using a sample of one hundred instances in a zero-shot setting. As seen in Figure 1, the models, when prompted to identify the two span types together, outperform the models when they are prompted to identify them individually. This indicates that the complexity in the text and the underlying relationships between different span types help the LLMs accurately identify different span types.

RQ2: How do the model size and modeling strategies influence the span identification capabilities of LLMs?

The results indicate that for complex text, LLMs in a few-shot setting outperform all other approaches. In-context examples in the prompts aid the models in identifying the different types of spans. CoT underperforms both zero-shot learning, while IT improves the model performance, especially on subjective spans. Instruction-tuning outperforms all other approaches for both CSI and TSD. The models struggle in zero-shot and CoT for TSD. An analysis of the outputs indicates that some approaches produce irrelevant or extraneous text spans. Furthermore, few-shot learning outperform fine-tuned BERT models on TBO and CSI, while IT has comparable performance to BERT on TSD.

Kaplan et al. (2020) show that with increasing

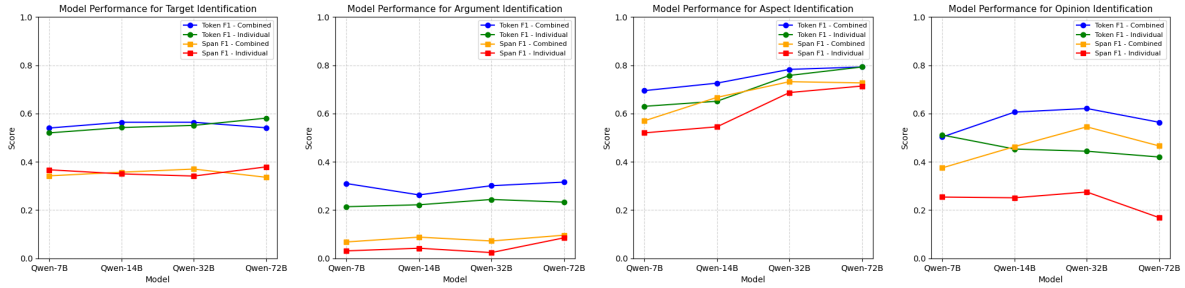

Figure 1: Token F1 and Span F1 scores for the complex text span identification. The plots show the scores for each type of span when extracted individually and combined.

model size, the performance improves. To test this hypothesis, we experiment with different model sizes ranging from 7B to 72B parameters. Our experimental results suggest that while the model size increases, there is only a marginal performance improvement. Overall results show that among the Qwen models, the 7B parameter model is the least performing model, whereas among 14B, 32B, and 72B, either model outperforms the others. While for Llama models, the 70B model consistently outperforms the 8B model. However, the performance difference across models with varying parameter sizes is only marginal. With extensive computational resources required for larger models and only marginal performance improvement, smaller models may represent a more efficient choice for these tasks.

RQ3: Are LLMs efficient in a low-resource setting?

Training language models requires extensive training data. However, some of the token classification tasks may have data scarcity. Hence, to assess how the data scarcity affects the performance of language models, fine-tune both small language models (SLM) and large language models (LLM) with varying training data sizes, ranging from 200 to 1000 samples. For this specific experiment, we compare the performance of BERT-large and Qwen-7B models.

Our experiments indicate that, for the TBO, ABSA, and TSD tasks, BERT outperforms Qwen-7B model for all data sizes (See Figure 7). The performance varies depending on the span type and number of training examples used, where it is comparable for some span types while substantial for the others. However, the SLM outperforms the LLM on CSI, especially on the span F1 score. This indicates that LLM identifies the exact claim spans more precisely than the smaller models. The findings suggest that SLMs generally outperform LLMs when fine-tuned with limited labeled training data. However, few-shot learning with LLMs can be leveraged in such scenarios due to its higher performance, as indicated in Tables 3.

7 Conclusion

In this work, we evaluate several LLMs with different approaches on subjective span identification. We answer important research questions pertaining to text complexity and model size, and further explore the capabilities of LLMs in a low-resource setting. Our findings suggest that the complexity and underlying relationships within text aid LLMs in identifying precise text spans. Furthermore, for the specific task of span identification, the model size does not have a significant impact on the performance. Although SLMs like BERT still outperform LLMs, approaches like few-shot learning can be leveraged in a low-resource setting. While LLMs have shown exceptional ability in explicit and context-independent span identification, they still underperform smaller models in identifying subjective spans.

In future work, we would like to explore approaches to improve the LLMs understanding of the input for accurate span identification, especially considering the context from both left and right. We further plan to explore other challenging datasets on related subjective tasks including multimodal data where text and images are paired (Farabi et al., 2024). Finally, we plan to expand this work to non-English datasets with the goal of evaluating the multilingual capabilities of the current generation of LLMs.

Acknowledgments

The authors would like to thank the anonymous reviewers for their insightful feedback.

Marcos Zampieri is partially supported by the Virginia Commonwealth Cyber Initiative (CCI) award number N-4Q24-009.

Limitations

The main limitation of this paper is that we only evaluate two open-source model families. However, other open-source or proprietary models may achieve comparable performance. Additionally, the prompts used in the experiments follow a specific template style. Experimenting with different prompt templates may generate different results. The task-specific instructions within the prompts can be adjusted to generate more efficient outputs. In the few-shot experiments, we use three and five examples in the prompts. However, including additional examples can enhance the model's performance. Moreover, our evaluations are focused on English datasets. Expanding this work to encompass additional languages and task datasets may offer further insights into the token span identification capabilities of LLMs.

References

Marouane Birjali, Mohammed Kasri, and Abderrahim Beni-Hssane. 2021. A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, 226.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. In *Proceedings of ACL*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of ICWSM*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL*.

Alphaeus Dmonte, Eunmi Ko, and Marcos Zampieri. 2024a. An evaluation of large language models in financial sentiment analysis. In *Proceedings of BigData*.

Alphaeus Dmonte, Roland Oruche, Marcos Zampieri, Prasad Calyam, and Isabelle Augenstein. 2024b. Claim verification in the age of large language models: A survey. *arXiv preprint arXiv:2408.14317*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

David Dukić and Jan Šnajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only llms for sequence labeling. In *Findings of the ACL (ACL 2024)*.

Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A survey of multimodal sarcasm detection. In *Proceedings of AAAI*.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors. *arXiv preprint arXiv:2305.14450*.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of NeurIPS*.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of AAAI*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

Shubham Mittal, Megha Sundriyal, and Preslav Nakov. 2023. Lost in translation, found in spans: Identifying claims in multilingual social media. In *Proceedings of EMNLP*.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of ACL*.

Chaoxu Pang, Yixuan Cao, Qiang Ding, and Ping Luo. 2023. Guideline learning for in-context information extraction. In *Proceedings of EMNLP*.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of SemEval*.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of SemEval*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. SemEval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of SemEval*.

Tharindu Ranasinghe and Marcos Zampieri. 2021. Mudes: Multilingual detection of offensive spans. In *Proceedings of NAACL*.

Andra Sandu, Liviu-Adrian Cotfas, Aurelia Stănescu, and Camelia Delcea. 2024. A bibliometric analysis of text mining: Exploring the use of natural language processing in social media research. *Applied Sciences*, 14(8):3144.

Diptanu Sarkar, Marcos Zampieri, Tharindu Ranasinghe, and Alexander Ororbia. 2021. fbert: A neural transformer for identifying offensive content. In *Findings of the ACL (EMNLP2021)*.

Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2024. Averitec: A dataset for real-world claim verification with evidence from the web. In *Proceedings of NeurIPS*.

Kian Long Tan, Chin Poo Lee, and Kian Ming Lim. 2023. A survey of sentiment analysis: Approaches, datasets, and future research. *Applied Sciences*, 13(7):4550.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *NAACL*.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023. GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2016. Recursive neural conditional random fields for aspect-based sentiment analysis. In *Proceedings of EMNLP*.

Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. 2017. Coupled multi-layer attentions for co-extraction of aspect and opinion terms. In *Proceedings of AAAI*.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *ACL*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL*.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023a. Target-based offensive language identification. In *Proceedings of ACL*.

Marcos Zampieri, Sara Rosenthal, Preslav Nakov, Alphaeus Dmonte, and Tharindu Ranasinghe. 2023b. OffensEval 2023: Offensive language identification in the age of large language models. *Natural Language Engineering*.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of NAACL*.

A Appendix

A.1 LLM Prompt

Figure 2 shows the prompt template used in our experiments. Table 4 shows the task-specific prompts. Each prompt consists of instructions describing the tasks and each type of span to extract. Additionally, the prompts include the input instances.

<Task-specific instruction. This includes the definitions of the type of spans to be identified>

Output Format:
<The format of the generated output. This will be 'tag: span'. For example, Target: target-span>

Examples:
<n examples>

Input:
<Input Text>

Response*:*

Figure 2: The prompt template used in our experiments.

A.2 Hyper-parameters

To fine-tune the models, we experiment with several learning rate values, with 1e-4 giving an optimal performance and minimum average loss. A per-device batch size of 2 with a gradient accumulation size of 8 was used to instruction-tune the LLMs. We chose a lower batch size to accommodate the limited computational resources available. We further leveraged the Adam optimizer and fine-tuned the models for ten epochs. For LoRA, we use the alpha value of 16 and the r value of 64, as these values provided the best performance. A dropout of 0.1 was used. To evaluate the LLMs with CoT and ICL, we use a temperature value of 0.0001 (as models like Llama do not allow a temperature value of 0) to allow deterministic outputs.

A.3 Additional Results

In this section, we present additional results for our experiments. While Table 3 shows the F1-scores of the best performing Qwen and Llama models, we present the results for all other models in Table 5.

A.4 Task-Specific Outputs

We show outputs for all the tasks in Figures 3- 6. For each task, the outputs generated by the Llama model with different approaches are shown. For the few-shot setting, we show the outputs of 5-shot experiments.

A.5 Progress Test

Figure 7 shows the performance of Qwen-7B compared to BERT when trained with varying training data sizes. We perform this experiment to understand how LLMs perform in a low-resource setting. For the TBO and ABSA tasks, BERT outperform the Qwen-7B model. Similarly, for TSD, BERT outperforms Qwen-7B when fine-tuned with 200 instances, but as we increase the number of training instances, the F1 score difference decreases. Unlike the other three tasks, Qwen-7B outperforms BERT, with a substantial Span F1 difference between the two models. However, as we increase the training dataset, the difference in Token F1 for this task gradually decreases. The experiment indicates that the performance of SLMs and LLMs differs, considering the type of task.

Dataset	Prompt
TBO	For the given text, identify the (target, argument) pairs. • Target: The individual, group, or organization towards whom the argument is directed. • Argument: A phrase or sentence containing offensive, profane, or unacceptable language. • An argument may or may not have a target. • Target can appear more than once if referenced by multiple arguments. • The argument and target may be the same. Output Format: Target n: <nth target> Argument n: <nth argument> Examples:{n examples} Input: {input text} Response:
ABSA	For the given text, identify the aspects and opinions. • Aspect: The entities to which sentiments are tied to. • Opinion: The sentiment words or phrases. Output Format: Aspect n: <nth aspect> Opinion n: <nth opinion> Examples:{n examples} Input: {input text} Response:
CSI	For the given text, identify the claim spans. • Claim Span: A phrase or sentence that explicitly mentions a claim, assertion, or argument. Output Format: Span n: <nth claim span> Examples:{n examples} Input: {input text} Response:
TSD	For the given text, identify the arguments. • Argument: A phrase or sentence containing offensive, profane, or unacceptable language. Output Format: Argument n: <nth argument> Examples:{n examples} Input: {input text} Response:

Table 4: Prompts used for each task. The prompt contains a task-specific instruction along with the respective input instance.

Model	TBO				ABSA				CSI		TSD	
	Target		Argument		Aspect		Opinion					
	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

Table 5: F1 scores of 0-, 3-, 5-shot, CoT, and instruction tuned (IT) models. The TF1 and SF1 scores for each span type are reported.

Input: I'll be tweeting utter bullshit and sheer nonsense at this time of the night

---

0-Shot:
Target 1: I
Argument 1: utter bullshit and sheer nonsense

5-Shot:
Target 1: nan
Argument 1: utter bullshit
Target 2: nan
Argument 2: sheer nonsense

Instruction Tuned:
Target 1: None
Argument 1: utter bullshit
Target 2: None
Argument 2: sheer nonsense

Chain of Thought:
Let's think step by step. The text contains one argument: "utter bullshit and sheer nonsense". The target of this argument is not explicitly mentioned, but it can be inferred that the target is the person or people who are being addressed by the speaker. Therefore, the output would be:
Target 1: The person or people being addressed by the speaker
Argument 1: utter bullshit and sheer nonsense

Figure 3: Outputs for the TBO task. The output for each approach is shown. For few-shot, only the 5-shot output is shown.

Input: after dinner i heard music playing and discovered that there is a lounge downstairs

---

0-Shot:
Aspect 1: music
Opinion 1: playing
Aspect 2: lounge
Opinion 2: downstairs

5-Shot:
Aspect 1: music
Aspect 2: lounge
Opinion 1: playing
Aspect 3: downstairs

Instruction Tuned:
Aspect 1: music
Opinion 1: playing
Aspect 2: lounge
Opinion 2: downstairs

Chain of Thought:
Let's think step by step. First, we need to identify the aspects in the given text. In this case, the aspect is "music playing" and "lounge downstairs".
Next, we need to identify the opinions in the text. The opinion is "I heard music playing" and "I discovered that there is a lounge downstairs".
So, the output format would be:
Aspect 1: music playing
Opinion 1: I heard music playing
Aspect 2: lounge downstairs
Opinion 2: I discovered that there is a lounge downstairs

Figure 4: Outputs for the ABSA task. The output for each approach is shown. For few-shot, only the 5-shot output is shown.

Input: Getting the jab protects you from hurricanes? uhh.. let's go Brandon!

0-Shot:
Span 1: Getting the jab protects you from hurricanes?
Span 2: uhh.. let's go Brandon!

5-Shot:
Span 1: Getting the jab protects you from hurricanes?
Span 2: uhh.. let's go Brandon!

Instruction Tuned:
Span 1: Getting the jab protects you from hurricanes?
Span 2: uhh.. let's go Brandon!

Chain of Thought:
Let's think step by step. The text contains one claim span. It is the phrase "Getting the jab protects you from hurricanes?" which is an assertion or argument made by the speaker. Therefore, the output would be:
Span 1: Getting the jab protects you from hurricanes?

Figure 5: Outputs for the CSI task. The output for each approach is shown. For few-shot, only the 5-shot output is shown.

Input: This proposal is so idiotic that its failure is imminent.

0-Shot:
Argument 1: This proposal is so idiotic that its failure is imminent.

5-Shot:
Argument 1: idiotic

Instruction Tuned:
Argument 1: This proposal is so idiotic that its failure is imminent.

Chain of Thought:
Let's think step by step. The given text is: "This proposal is so idiotic that its failure is imminent."
Step 1: Identify the phrases or sentences that contain offensive, profane, or unacceptable language.
- In this case, the phrase "idiotic" can be considered offensive or unacceptable language.
Step 2: Determine the arguments based on the identified phrases or sentences.
- Argument 1: "idiotic"
Output Format:
Argument 1: idiotic

Figure 6: Outputs for the TSD task. The output for each approach is shown. For few-shot, only the 5-shot output is shown.

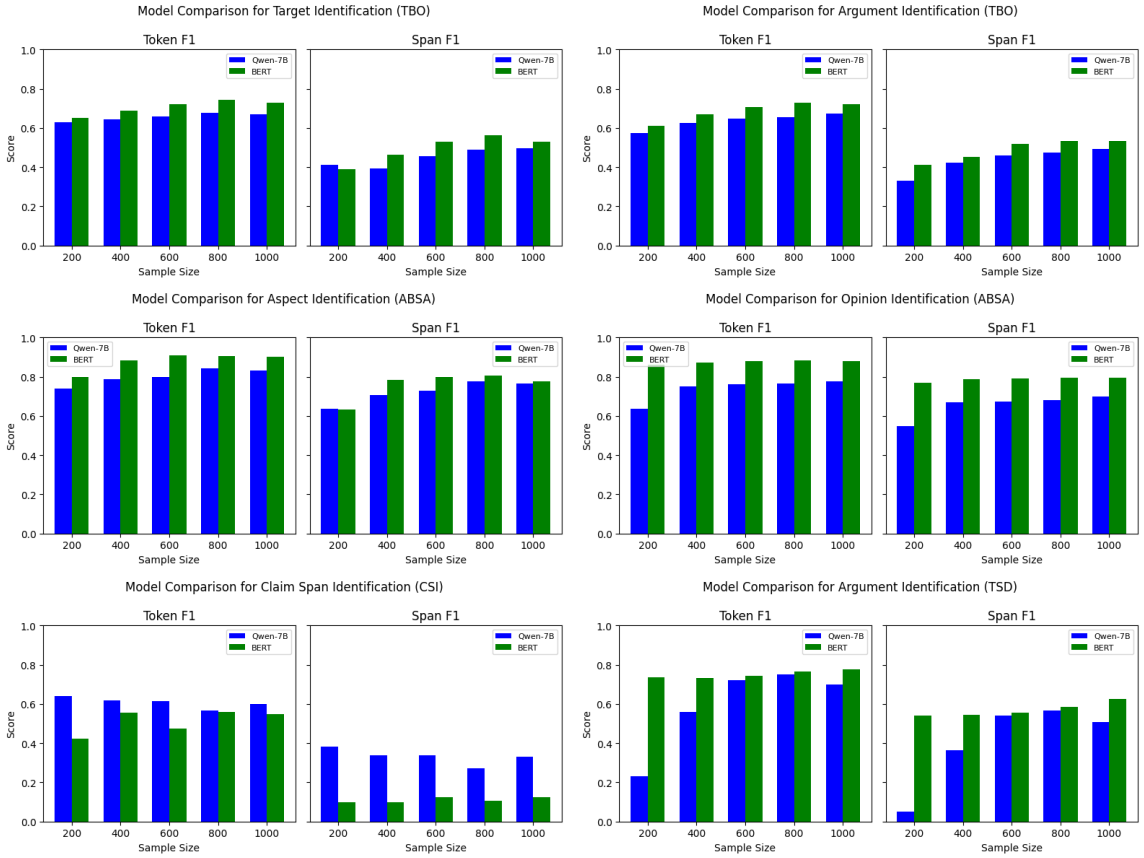

Figure 7: Progress test results. For each span type, the TF1 and SF1 scores are reported