Can Large Language Models Address Open-Target Stance Detection?

Abu Ubaida Akash **Ahmed Fahmy** **Amine Trabelsi**
Department of Computer Science, Université de Sherbrooke
{akaa2803,mosa2801,amine.trabelsi}@usherbrooke.ca

Abstract

Stance detection (SD) identifies a text's position towards a target, typically labeled as *favor*, *against*, or *none*. We introduce Open-Target Stance Detection (OTSD), the most realistic task where targets are neither seen during training nor provided as input. We evaluate Large Language Models (LLMs) GPT-4o, GPT-3.5, Llama-3, and Mistral, comparing their performance to the only existing work, Target-Stance Extraction (TSE), which benefits from predefined targets. Unlike TSE, OTSD removes the dependency of a predefined list, making target generation and evaluation more challenging. We also provide a metric for evaluating target quality that correlates well with human judgment. Our experiments reveal that LLMs outperform TSE in target generation when the real target is explicitly and not explicitly mentioned in the text. Likewise, for stance detection, LLMs excel in explicit cases with comparable performance in non-explicit in general.[1]

1 Introduction

Stance detection (SD) aims to determine the position of a text or person towards a certain *target*, typically categorized as "favor", "against", or "none". The *target* can be mentioned explicitly in the text, or sometimes the idea of the *target* can be conveyed indirectly (Küçük and Can, 2020). In Zero-shot Stance Detection (ZSSD), a model predicts stances for targets it has not seen during training, which is crucial since collecting training data for every potential target is impractical (Allaway et al., 2021). While recent research has focused on ZSSD (Zhang et al., 2023; Li et al., 2023b; Wen and Hauptmann, 2023; Liang et al., 2022; Zhu et al., 2022; Luo et al., 2022; Xu et al., 2022), most studies assume that the target is known or manually identified and given as input, a rare scenario in real-world applications where the target is often unknown or not explicitly mentioned. Additionally, target annotation is an expensive task in SD (Küçük and Can, 2020).

In this paper, we focus on a different yet challenging task we refer to as **`Open-Target Stance Detection (OTSD)`**. In this task, the target is neither seen during training (zero-shot) nor provided as input to the model. In OTSD, one important challenge lies in identifying the target from the text rather than having it provided as input. The target may be mentioned explicitly or not and the stance is predicted with respect to the produced target.

There has been one notable attempt to partially address this real-world setting of SD. Li et al. (2023a) generate targets from the input text and later detect stance based on these targets. However, in target generation, they map the generated targets to a predefined list of golden targets (details in §2) to ensure exact wording and facilitate evaluation. They call their approach **TSE**, which does not fully align with our OTSD setting. In OTSD we assume no input target information is given during the whole process, making it more practical than TSE, which requires a comprehensive list of all possible targets. Moreover, our experiment reveals a performance gap in Li et al. (2023a)'s approach when used on text with explicit or non-explicit target mentions, a factor not addressed by the authors that warrants further investigation in the context of OTSD. In this work, we present and explore the OTSD task, focusing on its two main steps: Target Generation (TG) and Stance Detection (SD) (Eq. 2). As a zero-shot task for both target and stance, Large Language Models (LLMs) appear well-suited for this challenge. Therefore, we examine the performance of LLMs and compare it to the primary existing work, TSE, which benefits from using a predefined list of targets during the process (Eq. 1). Our empirical study aims to address the following research questions: [**RQ#1**] How do proprietary and open LLMs—specifically

[1] Dataset and code are available here: `anonymous.4open.science/r/opentarget-5521`

GPT-4o (Achiam et al., 2023), GPT-3.5 (Brown et al., 2020), Llama 3 (Touvron et al., 2023), and Mistral (Jiang et al., 2023)—perform in open TG compared to the TSE when the real target is explicitly or non-explicitly mentioned in the text for stance classification, and how do they compare to each other? [**RQ#2**] How do the same LLMs perform compared to TSE in SD (in the context of OTSD) for both explicit and non-explicit cases, and how do they compare to each other? Our contributions consist of introducing the task of OTSD, providing a target evaluation metric, and conducting experiments answering the research questions.

2 Target-Stance Extraction (TSE)

In TSE, Li et al. (2023a) demonstrate target extraction and stance detection tasks in both in-target and zero-shot settings. For our comparison, we only consider their zero-shot setting. In TSE zero-shot, they first generate targets using keyphrase generation models, then map these targets to a predefined list of golden targets to find the closest match, and finally use this match to detect the stance of the text (Eq. 1). The best match from the keyphrase model, marked as `TSE-BestGen`, and the best mapped target (`TSE-mapped`) is considered as the predicted target. For example, given the text *"Embracing different faiths teaches ... diverse beliefs and foster unity."*, TSE primarily generates targets such as *Peace*, *Religious diversity*, *Respect*, *etc.*. From a predefined list of possible targets (*e.g.*, *Face Mask*, *Atheism*, *Donald Trump*, *etc.*), *Atheism* matches most closely with *Religious diversity* and is considered as the final predicted target. In OTSD, we aim to generate *Religious diversity* or a closely related phrase directly, bypassing the predefined list. The stance is then detected toward this generated target.

3 Open-Target Stance Detection (OTSD)

3.1 Task Definition

Given an input text x, previous work (Eq. 1) generates a target t' and maps it to a predefined list of targets k, resulting in a mapped target t. Finally, they detect the stance y given x and t. OTSD task (Eq. 2) is to detect the stance y from the input text x and the generated target t', eliminating the need for a predefined list k.

$$x \xrightarrow{\text{generate}} t' \xrightarrow{\text{map}} t \in k, \quad x + t \to y \tag{1}$$

$$x \xrightarrow{\text{generate}} t', \quad x + t' \to y \tag{2}$$

3.2 Approach

To address the research questions outlined in §1, our approach to OTSD leverages the zero-shot learning capabilities of LLMs. We adopt the "Task Definition" prompting strategy, as it has been shown by Cruickshank and Ng (2023) to outperform other strategies—such as "Task-only", "Context Analyze", "Context Question", "Zero-shot CoT", and "CoDA"—particularly for smaller models (7B-8B parameters) across most datasets. Since this is an open-target task, we use a straightforward prompting approach to ensure fairness and generalizability, avoiding tailored prompts for specific targets, or in-context examples, aligning with the zero-shot nature of this task. We employ two different prompting approaches for TG and SD. The prompts and their design justifications can be found in Appendix A. **TG+SD (Two-Step Approach)** In this approach, we first generate the target and then sequentially detect the stance of the text toward the generated target. By directing the LLM to focus on one task at a time, this method may reduce the cognitive load on the model. **TG&SD (Single-Step Joint Approach)** This approach combines TG and SD into a single step using a unified prompt. By jointly generating the target and stance, the model can develop a more holistic understanding of the relationship between the text, target, and stance.

4 Experiments

In OTSD task, we compare our approaches with TSE in both TG and SD. To ensure sensitivity and robustness, we run our experiments three times independently and report the average results [2]. Since OTSD is a generative task, we primarily compare our results with `TSE-BestGen`, despite its reliance on a predefined list of golden targets. Additionally, we consider `TSE-mapped` as the current best score, regardless of task type. Details about our hardware settings can be found in Appendix D.

4.1 Dataset and Model

To ensure a fair comparison between TSE and LLM models, we use the same dataset as TSE's zero-shot setup. VAST (Allaway and McKeown, 2020) is used to compare LLMs across different settings. Both TSE and VAST datasets contain three stance classes (*Favor*, *Against*, *None*), and contain 3,000 and 5,100 samples, respectively, in our study. VAST, originally a multi-target dataset, is converted

[2] Score variations are negligible, up to four decimal places.

to a single-target format (conversion process in Appendix E) to match the scope of our study. TSE has 500 samples for each of 6 targets (see Appendix F), while VAST includes 2,145 unique targets. We classify samples as explicit or non-explicit by removing stop words and special characters, lemmatizing all words, and checking for target words in the text. This results in 1,804 explicit and 1,196 non-explicit samples from TSE, and 3,120 explicit and 1,980 non-explicit samples from VAST (see Appendix F). Note that, due to the lack of code and details from TSE for reproducing target generation and selection, we cannot use other SD datasets for benchmarking. In our focused contribution, we use both proprietary and non-proprietary LLMs, including GPT-3.5, GPT-4o, Llama-3, and Mistral to address our two research questions (§1).

4.2 Evaluation Methods

BTSD: To measure the generated target quality, we fine-tune the BERTweet model (Nguyen et al., 2020) following the same setup as TSE (fine-tuning details in Appendix G). The SD score (F1-macro) is used as the evaluator for the generated targets. To assess the effectiveness of this metric, we experiment with different levels of target quality by modifying or removing words from the golden targets, substituting them with incorrect targets, and selecting random words from the vocabulary as targets. We find that the BERTweet F1 score consistently and significantly improved with higher-quality targets, in both explicit and non-explicit cases (see Appendix C). More importantly, BTSD score shows strong tau (τ) correlation (KENDALL, 1938) of 0.85 and 0.79 for TSE and VAST, respectively, with human judgments gathered during experimentation (see the section below). Therefore, the BTSD serves as a reliable proxy for evaluating target generation quality.

Human Evaluation: To evaluate the relevance of the generated targets and further validate the BTSD metric, we conduct a small-scale human evaluation using 500 randomly selected samples (300 explicit and 200 non-explicit) from both datasets, maintaining an equal ratio of the stance labels. Three annotators were asked to assess the relevance of the generated targets to the golden ones, classifying them as either 0 (Not Related), 0.5 (Partially Related), or 1 (Completely Related), following the guidelines provided in Appendix K. To measure inter-annotator agreement among the three raters, we compute Krippendorff's α (Krippendorff, 2011), and Fleiss's κ (Fleiss, 1971) for each model, as shown in Appendix Table 6. We observe a solid overall agreement of α=0.76 and κ=0.664.

SemSim: In OTSD, since we do not rely on a predefined list of targets, the generated targets may not precisely match the gold targets but are semantically related. Therefore, we measure the semantic similarity between the generated and gold targets using BERT embeddings. Detailed steps can be found in Appendix H with human judgement correlation of τ=0.57 for TSE, and τ=0.59 for VAST. We use macro-average F1 for the task of SD.

5 Result Analysis

5.1 Target Generation (TG)

As shown in Table 1, both human evaluation (HE) and BTSD indicate that LLMs generate higher-quality targets in explicit cases compared to TSE-BestGen, the best model for generating targets not matched to a predefined list. Although TSE-mapped achieves higher HE score in explicit case, its BTSD score is lower. This is because TSE-mapped produces either a perfect match (score=1) or a completely unrelated target (score=0) favouring higher HE score (see Appendix M). In contrast, LLMs mostly generate targets that are highly or partially related to golden targets, resulting in fewer irrelevant outputs and higher BTSD, despite slightly lower average HE scores.

When the target is not explicitly mentioned, LLMs produce higher-quality targets than both TSE models. However, the quality drops for LLMs (TSE as well) compared to explicit cases (as per Table 1), with the best-performing LLM getting the HE score slightly lower than 0.5 (refers to partially related target) on average. A manual inspection of GPT-3.5's 30 incorrect predictions suggests that the lower score in non-explicit cases is due to insufficient context, often resulting from how the TSE test set is constructed. The lack of implicit hints or surrounding context leads the model to generate targets based solely on the text provided, causing inaccuracies (see example in Appendix J.1).

When comparing tested proprietary and open LLMs using BTSD and HE, there is no clear advantage in TG (*i.e.*, not all proprietary models outperform open ones, or vice versa). We note that GPT-4o consistently produces targets of equal or higher quality than other tested models across the settings, including explicit and non-explicit target mentions, TG+SD, and TG&SD approaches, and across the

Model		TG+SD				TG&SD					TG+SD				TG&SD			
		SS	BTSD	HE	SC	SS	BTSD	HE	SC		SS	BTSD	HE	SC	SS	BTSD	HE	SC
	TSE	Explicit								VAST	Explicit							
TSE-M		0.96	36.63	0.716	38.1	-	-	-	-		-	-	-	-	-	-	-	-
TSE-B		0.86	35.8	0.338	37.81	-	-	-	-		-	-	-	-	-	-	-	-
GPT-3.5		0.87	38.43	0.546	42.68	0.87	**39.6**	**0.663**	**47.61**		0.82	41.67	0.461	47.21	**0.89**	**44.25**	**0.581**	**48.48**
GPT-4o		0.87	41.55	0.566	44.78	**0.88**	**41.92**	**0.69**	**46.83**		0.84	42.69	0.519	40.77	**0.88**	**44.25**	**0.629**	**49.38**
Llama-3		**0.88**	**38.31**	0.54	**43.13**	0.87	37.75	**0.558**	42.1		0.86	41.13	**0.467**	**47.20**	**0.88**	**42.45**	0.458	46.72
Mistral		0.85	38.47	0.586	46.15	**0.87**	**39.02**	**0.65**	**46.32**		0.84	42.13	0.49	45.06	**0.86**	**43.71**	**0.526**	**46.77**
		Non-explicit									Non-explicit							
TSE-M		0.9	30.56	0.391	32	-	-	-	-		-	-	-	-	-	-	-	-
TSE-B		0.82	29.32	0.215	31	-	-	-	-		-	-	-	-	-	-	-	-
GPT-3.5		**0.84**	**33.1**	**0.419**	32.66	0.83	31.32	0.405	**33.94**		0.79	38.1	0.31	45.54	**0.85**	**38.55**	**0.416**	**45.80**
GPT-4o		**0.84**	35.14	0.43	36.39	0.83	**36.12**	**0.513**	**37.5**		0.80	38.55	0.387	39.92	**0.83**	**39.84**	**0.45**	**43.84**
Llama-3		**0.85**	33.06	0.447	**33.36**	0.83	**33.81**	**0.49**	31.9		0.82	38.74	0.33	**47.98**	**0.84**	**39.91**	**0.383**	45.48
Mistral		0.83	30.18	0.386	**33.61**	0.83	**34.12**	**0.495**	30.47		0.80	39.75	0.405	**43.34**	**0.81**	**40.05**	**0.447**	41.87

Table 1: Comparison of TG and SD performance between TSE and LLMs using various metrics: TG is evaluated with SS (SemSim), BTSD (%), and HE (Human Evaluation) scores, while SC (%) measures SD performance. TSE-M and TSE-B represent `TSE-mapped` and `TSE-BestGen`, respectively. Metric-wise best results within a specific setting are underlined, and better results of a model across TG+SD and TG&SD for each dataset are in **bold**.

datasets. Among the tested open LLMs, Mistral generates higher quality targets than Llama-3 in most settings. Comparing between the approaches, most models perform better in TG&SD across both datasets, for explicit and non-explicit cases.

5.2 Stance Detection (SD)

As shown in Table 1, according to the SC (stance classification), LLMs outperform both TSE models in detecting stances for explicit cases, while performing comparably with TSE models in non-explicit cases (except GPT-4o).

In general, LLMs perform better in SD when the target is explicitly mentioned in the text compared to when it is not. Our manual analysis shows that in non-explicit cases, the input text lacks sufficient context about the target, leading to lower SC scores.

Overall, the experimented proprietary models (GPT-3.5 and GPT-4o) outperform the open ones in TG&SD in both the explicit and non-explicit cases across the datasets. In some cases of TG+SD setting, though, the open LLMs achieve comparatively higher scores. Comparing approaches, LLMs tend to perform better when prompted with the holistic TG&SD, though their TG+SD results remain competitive, specifically in non-explicit cases across both datasets.

Although the quality of the generated targets correlates positively with the SC score in explicit cases (TSE τ=0.71, VAST τ=0.29), this correlation is negative in non-explicit cases (TSE τ=−0.14, VAST τ=−0.29). We manually investigate the negative correlation (lower SC with higher quality target) in non-explicit cases. Indeed, LLMs such GPT-4o may generate highly relevant targets that are closely related to the golden targets but nearly ‘antithetical’ or opposite, leading to a stance reversal in comparison to the golden stance. For example, when the golden target is "permits to carry guns", GPT-4o generates "gun control in universities", while GPT-3.5 generates "guns". Although GPT-4o’s target is more closely related to the golden target, it represents the opposite stance. On the other hand, GPT-3.5’s target, though broader, less accurate and less related, aligns with the stance of golden target. This highlights the need for a coherence measure in OTSD that accounts for the target-stance alignment.

6 Conclusion

We introduce the Open Target Stance Detection (OTSD) task, where targets are neither seen during training nor provided as input, making it more realistic and challenging. We examine proprietary and open LLMs in our OTSD task, and propose a metric to evaluate the quality of Target Generation (TG). We compare to TSE, a method leveraging a predefined list of targets. LLMs outperform it in TG. However, LLMs struggle in generating quality targets when these are not explicitly mentioned. LLMs also surpass TSE in Stance Detection (SD) when the target is explicitly mentioned while performing comparably in non-explicit cases. Finally, a single-step joint prompt approach (TG&SD) of LLMs proves more effective than a two-step approach (TG+SD) for both TG and SD.

Limitation

The limitations of our work can be listed as follows:

- We did not provide a specific coherence measure to evaluate the Open-Target Stance Detection task, which is worth to be investigated in the future work, as the target and the stance toward it are inherently interconnected.

- We conducted a thorough investigation of simple prompting for LLMs, inspired from the stance detection literature (§3.2), which has been shown to outperform other strategies, and provides the answers of our research questions already. However, there might be still room for improvement through a more detailed examination of different prompting techniques or task-specific pre-training. While we present some preliminary results on the chain-of-thought approach in Appendix L, further investigation is needed to fully understand its potential.

- Regarding potential data leakage during the pre-training of LLMs, it is possible that these models were exposed to some of our test data, leading to prior knowledge of certain entities. However, since the LLMs were not explicitly pre-trained on the same task, we continue to refer to our approach as zero-shot stance detection.

- Although we did not claim that the results or findings are true for all LLMs, experimenting with other proprietary LLMs was financially constraining for us.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Emily Allaway and Kathleen McKeown. 2020. Zero-Shot Stance Detection: A Dataset and Model using Generalized Topic Representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931, Online. Association for Computational Linguistics.

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. *arXiv preprint arXiv:2105.06603*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Will-they-won't-they: A very large dataset for stance detection on twitter. *arXiv preprint arXiv:2005.00388*.

Iain J Cruickshank and Lynnette Hui Xian Ng. 2023. Use of large language models for stance classification. *arXiv preprint arXiv:2309.13734*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Akash Gautam, Puneet Mathur, Rakesh Gosangi, Debanjan Mahata, Ramit Sawhney, and Rajiv Ratn Shah. 2020. # metooma: Multi-aspect annotations of tweets related to the metoo movement. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 209–216.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in covid-19 tweets. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (long papers)*, volume 1.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

M. G. KENDALL. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.

Yingjie Li and Cornelia Caragea. 2021. A multi-task learning framework for multi-target stance detection. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2320–2326.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023a. A new direction in stance detection: Target-stance extraction in the wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10071–10085.

Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023b. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, pages 1500–1509.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747.

Yun Luo, Zihan Liu, Yuefeng Shi, Stan Z Li, and Yue Zhang. 2022. Exploiting sentiment and common sense for zero-shot stance detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 7112–7123.

Lin Miao, Mark Last, and Marina Litvak. 2020. Twitter data augmentation for monitoring public opinion on covid-19 intervention measures. In *Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016a. A dataset for detecting stance in tweets. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3945–3952.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016b. Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 31–41.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. Bertweet: A pre-trained language model for english tweets. *arXiv preprint arXiv:2005.10200*.

Amit Singhal. 2001. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24:35–43.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 116–124.

Christian Stab, Tristan Miller, and Iryna Gurevych. 2018. Cross-topic argument mining from heterogeneous sources using attention-based neural networks. *arXiv preprint arXiv:1802.05758*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Haoyang Wen and Alexander G Hauptmann. 2023. Zero-shot and few-shot stance detection on varied topics via conditional generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1491–1499.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. Openstance: Real-world zero-shot stance detection. *arXiv preprint arXiv:2210.14299*.

Jiarui Zhang, Shaojuan Wu, Xiaowang Zhang, and Zhiyong Feng. 2023. Task-specific data augmentation for zero-shot and few-shot stance detection. In *Companion Proceedings of the ACM Web Conference 2023*, pages 160–163.

Chenye Zhao and Cornelia Caragea. 2024. EZ-STANCE: A large dataset for English zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15697–15714, Bangkok, Thailand. Association for Computational Linguistics.

Qinglin Zhu, Bin Liang, Jingyi Sun, Jiachen Du, Lanjun Zhou, and Ruifeng Xu. 2022. Enhancing zero-shot stance detection via targeted background knowledge. In *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval*, pages 2070–2075.

A Prompting Details

Since LLMs are sensitive to the length of generated text, we limit the target length to ensure a fair comparison with TSE in terms of target generation quality. To determine the optimal length, we test various target length instructions across all LLMs (§4.1) on 100 randomly selected samples from our dataset (§4.1), aiming to achieve an average length and standard deviation that closely match those of the golden targets. Based on our empirical study, we set the maximum target length to 5 for GPT-3.5 and 4 for Llama and Mistral to obtain an average length and standard deviations close to those of gold targets (explicit: 3.78 ± 1.5; non-explicit 2.66 ± 1.76).

TG+SD (Two-Step Approach)

Prompt (TG): *You will be provided with a tweet, and your task is to generate a target for this tweet. A target should be the topic on which the tweet is talking. The target can be a single word or a phrase, but its maximum length MUST be 4 words. The output should only be the target, no other words. Do not provide any explanation but you MUST give an output, do not leave any output blank.*

Prompt (SD): *Stance classification is the task of determining the expressed or implied opinion, or stance, of a statement toward a certain, specified target. Analyze the following tweet and determine its stance towards the provided target. If the stance is in favor of the target, write FAVOR, if it is against the target write AGAINST and if it is ambiguous, write NONE. Do not provide any explanation but you MUST give an output, do not leave any output blank. Only return the stance as a single word, and no other text.*

Prompt Justification:

- **Clarity of Task:** The task description explicitly states what the LLM needs to do—generate a target for a given tweet in TG prompt and for SD, it provides a clear and concise definition of stance classification. This ensures that the LLM understands the primary objective without ambiguity.
- **Focus on Relevance:** By instructing the model that the target should be the topic of the tweet, it emphasizes the need for relevance, ensuring that the generated target accurately represents the tweet's content.
- **Conciseness Requirement:** Limiting the target to a maximum of 4 words ensures that the output remains concise and focused, which is crucial for the generated target to be compared fairly with the golden target.
- **Stance Options Provided:** By specifying the possible stances (*Favor*, *Against*, *None*), the prompt sets clear categories for the model to choose from. This simplifies the models decision process and ensures the outputs are standardized.
- **Output Format Specification:** Clearly specifying that the output should only be the target, with no additional words or explanations, helps to standardize the outputs, making them easier to evaluate.
- **Mandating Output Generation:** By instructing the model not to leave any output blank and always to provide an answer, the prompt ensures robustness in responses, reducing the chance of missing data.

TG&SD (Single-Step Joint Approach)

Prompt (TG, SD): *Stance classification is the task of determining the expressed or implied opinion, or stance, of a statement toward a certain, specified target. Analyze the following tweet, generate the target for this tweet, and determine its stance towards the generated target. A target should be the topic on which the tweet is talking. The target can be a single word or a phrase, but its maximum length MUST be 4 words. If the stance is in favor of the target, write FAVOR, if it is against the target write AGAINST and if it is ambiguous, write NONE. If the stance is in favor of the generated target, write FAVOR, if it is against the target write AGAINST and if it is ambiguous, write NONE. The answer only has to be one of these three words: FAVOR, AGAINST, or NONE. Do not provide any explanation but you MUST give an output, do not leave any output blank. The output format should be: “‘Target: <target>, Stance: <stance>“‘.*

Prompt Justification:

- **Integrated Task Definition:** The prompt combines the definitions of both target generation and stance classification, ensuring the LLM understands the combined objective of generating a target and determining the stance towards it.
- **Clear and Concise Instructions:** By clearly outlining the steps—first generating a target and then determining the stance towards it—the prompt guides the model through a logical sequence, reducing ambiguity and improving task performance.
- **Conciseness Requirement:** Limiting the target to a maximum of 4 words ensures that the output remains concise and focused, which is crucial for the generated target to be compared fairly with the golden target.
- **Target Generation Constraints:** Limiting the target to a maximum of 4 words ensures that the output remains concise and focused, which is crucial for the generated target to be compared fairly with the golden target.
- **Stance Classification Options:** Providing explicit stance options (*Favor*, *Against*, *None*) simplifies the decision process for the model and ensures consistency in the outputs.
- **Mandatory Output Generation:** By instructing the model to always provide an output and

not leave any fields blank, the prompt ensures robustness and completeness in the responses, reducing the chance of missing data.

- **Output Format Specification:** Specifying the output format as ’Target: <target>, Stance: <stance>’ standardizes the responses, making them easier to parse and evaluate.

- **Focus on Simplicity and Relevance:** The prompt emphasizes that only the stance should be returned as a single word and no additional text, which maintains simplicity and relevance in the outputs.

B Model Descriptions

GPT-3.5-turbo-0125 is a variant of the GPT-3.5 series with 175 billion parameters which utilizes a transformer-based architecture. This model is decoder-only, similar to models like GPT-3 (Brown et al., 2020) and GPT-4 (Achiam et al., 2023), but with optimizations for efficiency and performance.

GPT-4o is a decoder-only model in the GPT series with an optimized architecture designed to enhance efficiency and accuracy. It builds upon the foundations of GPT-4 (Achiam et al., 2023), incorporating improvements in training and fine-tuning for better performance across diverse language tasks.

Llama-3-8B Instruct is a decoder-only model in the LLaMA (Touvron et al., 2023) series with 8 billion parameters that follows the architectural principles of its predecessor, Llama-2 (Touvron et al., 2023), but has been instruction-tuned to improve its performance on a variety of tasks.

Mistral 7B Instruct (Jiang et al., 2023) is a decoder-only model like Llama-2 (Touvron et al., 2023), but with different architectural features like Grouped-Query Attention and Sliding Window Attention. The Mistral model is 7 billion parameters and has been instruction tuned.

C BTSD Metric Justification

Table 2 contains the results of the stratified experiments with different measure stance detection F1 score when giving as input the text and a target, varying the quality of this latter. The numbers are significantly different.

Input	F1-macro (in %)	
	Explicit	Non-explicit
Tweet only	31.02	30.37
-w/ Gold Target (GT)	41.67	34.65
-w/ altered GT	35.3	32.54
-w/ incorrect Target	25.58	23.58
-w/ random vocab	18.66	14

Table 2: Macro avg. F1 score of BertTweet in detecting stances with different quality levels of the input target.

D Hardware Setting

All of the experiments were run on a computer with Ubuntu 22.04 Linux with x64 CPU with 4 cores, 32 GB of RAM and one NVIDIA V100 GPU.

E VAST Dataset Processing

Originally, the VAST dataset is a multi-target dataset, with multiple targets assigned to each sample text. However, since our work focuses on single-target analysis, we processed the VAST dataset accordingly to align it with our study (samples in Table 4). In VAST, there are two types of target columns: "ori_topic" which contains targets heuristically extracted, and "new_topic," which contains the final annotated targets determined by annotators. The "ori_topic" often includes duplicate targets for each text, while the "new_topic" provides unique targets with different stances.

To convert the dataset to a single-target format, we first filtered the samples to retain only the most frequent stance within each group of duplicate "ori_topic" entries. Next, we computed the average semantic similarity of each "new_topic" within these duplicate groups using BERT embeddings. For groups containing only a single row, the sample was directly retained. Then, we calculated the average pairwise cosine similarity for each "new_topic." Finally, within each duplicate "ori_topic" group, we kept only the row with the highest average semantic similarity score. After this conversion, we obtained a total of 5,100 samples (as shown in Appendix Table 4), covering both explicit and non-explicit samples (the filtering process for explicit and non-explicit samples is described in Section 4.1) from the original 18,545 samples.

F Dataset Samples

Following TSE, our targets are Creationism (Somasundaran and Wiebe, 2010), Gay Rights (Somasundaran and Wiebe, 2010), Climate Change is a Concern (Mohammad et al., 2016a), MeToo Movement (Gautam et al., 2020), Merger of Disney and Fox (Conforti et al., 2020), and Lockdown in New York State (Miao et al., 2020). Samples from TSE and VAST in both explicit and non-explicit settings are given in Table 3 and 4, respectively.

G BTSD Fine-tuning

Following TSE, we consider 4 datasets combined i.e. SemEval (Mohammad et al., 2016b), AM (Stab et al., 2018), COVID-19 (Glandt et al., 2021), P-Stance (Li and Caragea, 2021), containing 19 targets, splitting samples by (70%-train, 10%-dev, 20%-test), maintaining equal distribution from each of the targets. To assess the quality of the generated targets, we train a stance classifier model (referred as BTSD) as follows. Given a text x and a target t, we first formulate the input as a sequence $s = [[CLS]t[SEP]x]$ where $[CLS]$ is a token that encodes the sentence and $[SEP]$ is used to separate the sentence x and the target t. Then the representation of $[CLS]$ token is used to predict the stance toward target t. Note that t is the golden target here.

H SemSim Details

Specifically, we employ the bert-base-uncased (Devlin et al., 2018) model to obtain contextual encoded representations of the target words. After performing a mean pooling operation on these representations, we apply cosine similarity (Singhal, 2001) to the encoded sequences to obtain the similarity score.

I OTSD with EZSTANCE

We explore the OTSD task using EZSTANCE (Zhao and Caragea, 2024), the largest zero-shot stance detection (ZSSD) dataset, to address how proprietary and open LLMs compare in both target generation (TG) and stance detection (SD) tasks when the target is explicitly or implicitly mentioned in the text. Given the open-target nature of our task, EZSTANCE was selected for its diverse range of domains and unique target set, reflecting real-world scenarios.

I.1 Dataset Processing

The dataset originally contains 47,316 Twitter samples across three stance categories (*favor*, *against*, and *neutral*). Since EZSTANCE features multiple targets for some texts, we adapted it into a single-target dataset (one target per text) to better align with our research by following a series of steps.

The original dataset is organized into two subtasks (target-based and domain-based), with each subtask containing two target types: claim and noun-phrase. First, we combine the samples from both target types and remove duplicates. For instances where both target types exist, we retain the noun-phrase version, as it poses a greater challenge in generating precise targets. Then, we merge the samples from both subtasks and deduplicate them based on their first occurrence. This process results in 9,462 unique samples with 6,873 distinct targets. We further divide the dataset into explicit and non-explicit cases, as outlined in Section 4.1, yielding 9,313 explicit and 149 non-explicit samples, where the target is mentioned explicitly or implicitly in the text, respectively.

I.2 Result Analysis

As shown in Table 5, LLMs face challenges in target generation for non-explicit samples compared to when the target is explicitly mentioned, based on BTSD and HE scores. When comparing proprietary and open models, consistent with our earlier findings (§5.1), no single LLM consistently outperforms the others. While GPT-4o tends to generate higher-quality targets in the TG+SD approach, Mistral surpasses other models in some TG&SD cases. Across both explicit and non-explicit samples, all models perform noticeably better using the TG&SD approach.

For stance detection, as shown in 5, similar to our findings with TSE and VAST in Section 5.2, LLMs perform better in SC when the target is explicitly mentioned. When comparing proprietary and open models, GPT-4o and Mistral both demonstrate competitive performance across approaches. Overall, LLMs consistently achieve better stance detection in the TG&SD approach.

J Explanation with Samples in TG and SD

J.1 Target Generation

For instance, in the post *"I mean, this reads to me like anyone who's an alcoholic doesn't go to heaven..."*, GPT-3.5 generated "Sinful behavior and heaven's eligibility" as the target instead of the gold target "gay rights".

Target	Text	Stance
Explicit		
creationism	It is not 'appropriate' to teach creationism as a means for upholding the bible. This pre-supposes that the bible should be upheld in a literal sense. Many Christians object to this. More importantly, governments should not be in the business of upholding the Bible.	AGAINST
merger of disney and fox	@ComicBookNOW: FOX reportedly wants to make that deal with DISNEY! X-MEN in the MCU is now closer than ever	FAVOR
gay rights	Most health care organizations support gay parenting as equally capable as heterosexual parenting These organizations are the most capable of determining the capabilities of homosexuals to perform dutifully as parents.	FAVOR
lockdown in new york state	@adamajacoby @RV1026 @CNNPolitics Apparently, better than the "news sources" that keep you informed. You didn't even know America had to lock down. You didn't even know Cuomo put COVID patients in nursing homes. They are keeping you misinformed and looking like a complete moron. Do better!	FAVOR
metoo movement	@KRKBoxOffice Why everyone was silent when she molested Saif Sir ? Why no #MeToo campaign that time??	AGAINST
climate change is a concern	Being an engaged mom, means voting 4 the climate 2. Supporting only candidates who have a plan 2 act on #playin4climate #SemST	FAVOR
Non-explicit		
creationism	My point was not that Genesis contradicted itself but that if you took it literally it contradicted itself. The problem is not with Genesis so much as a literal interpretation of Genesis. You have to get away from literalism to make sense of the two creation stories, with their different orders of creation. You have proved my point by showing how a less than literal interpretation of the passages gets round the problem of a literal interpretation.	AGAINST
gay rights	That's fine. I support that.Here in California, we have a prospective bill that would do just that.I honestly think marriage is a religious thing, but religion is in the eye of the beholder.	FAVOR
lockdown in new york state	RT @Lukewearechange: So NYC announced another lock down coming this Sunday evening, I was suppose to be getting out of here with a friend	FAVOR
climate change is a concern	Considering moving yo Antarctica as thats the only way I could possibly become more #chill #SemST	NONE

Table 3: Samples from TSE dataset with all the 6 targets in both explicit and non-explicit settings.

J.2 Stance Detection

For instance, in the explicit sample text, *"Q. from audience: Malcom X said the most disrespected person in America is the black woman, has #MeToo changed this? Gender has been constructed in the image of the white woman, the idea of a level playing field, white women are often in better positions, says @HeidiMirza #CHevents"*, GPT-3.5 generated the target "Gender roles", which, while related, fails to capture the specific stance-related target "MeToo movement". This mistake reflects the model's need for broader context to establish the correct link.

K Human Annotator Agreement

Dataset Description: The file includes a *Tweet* column with various Twitter posts. The *Gold Target* column represents the true target, indicating the primary topic or issue the tweet addresses, which in turn informs the stance noted in the *Gold Stance* column. For each tweet, there are eight generated targets (in columns *T1* through *T8*) produced by different models, whose names remain anonymous for unbiased annotation. Additionally, there are empty cells (e.g., in the *T1 Score* column) designated for storing human evaluation scores corresponding to each generated target.

Definition of a Relevant Target: A generated target in the context of the target generation task is considered relevant if it accurately captures or closely aligns with the main topic, issue, or entity identified by the golden target in a given tweet. It should reflect the key subject matter or concern that the golden target represents, maintaining coherence with the overall context and meaning of the tweet.

Task Description: Your task is to assess the quality of each generated target by evaluating its relevance to the golden target in the tweet. You will do this by answering the following question and assigning a score to each generated target. Scoring criteria and examples are provided below.

"How closely does the generated target relate to the golden target in the tweet?"

Scoring Scale: 0 to 1 (Low to High)

- **0 – Not Related:** The generated target is entirely unrelated to the golden target. It does not reference or express anything connected to the intended topic.

- **0.5 – Partially or Indirectly Related:** The generated target has some relevance to the golden target but does not directly match the main concept. It may address a broader or narrower aspect of the topic, such as a parent topic, a subtopic, or a tangentially connected idea that is not fully aligned with the golden target's core meaning.

- **1 – Completely Related:** The generated target is highly relevant to the golden target. It

Target	Text	Stance
	Explicit	
health care law	Congress should delay the law for a year..."? Tell that to my niece whose 17 month old toddler has just been diagnosed with a rare disease that will need years of medical care. Without the provisions in the new health care law, this family would be facing bankruptcy. I prefer living in a nation where we have the decency to realize that health care is a right. Obviously those who seem to be objecting to the new health care law have never faced what 30 million people face in our country every day. For them another year's delay is life or death.	FAVOR
facebook	"...Facebook gets a bad rap; it didn't cause the cheating. It just made it more convenient to do (and perhaps easier to catch)." Hmm, call me crazy, but Facebook shares some of the responsibility – does it not? It's like refusing to assign blame to gun dealers and drug dealers. They increase accessibility to illicit materials, which essentially furthers the end goal of usage. Now, this isn't to suggest that Facebook is entirely at fault, but it does play an active role.	AGAINST
stability of the economy	"...one must ask how much money they must make to demonstrate that they are among the best managed companies on the planet." They must make enough money to insure that they can never fail and threaten the stability of the worlds economy again. That much money.	NONE
palestinian authority	"But let's start with the basics. Any nation wishing to declare independence should meet three essential elements: a strong central government, control of defined territory and security. The Palestinian Authority does not yet meet any of them." Historians would strongly disagree with you. Around sometime late eighteenth century some thirteen odd British colonies did "declare independence" without any resemblance of "strong central government" with its territory controlled by the British. Today a certain French foreign minister called it a "hyper-power" and not a "failed state". Please check upon a historian or read up your history. Trust me! It is true!!	AGAINST
economists	"Economists do certainly over-reach sometimes. We tend to apply the lens of economic efficiency to situations where many people apply the lens of fairness." I get a kick out of economists believing that real live humans are rational economic actors. Really?	AGAINST
	Non-explicit	
restaurant	"...tipping motivates people who work long, busy hours catering to the needs of others. It's the best way to ensure optimal service..." By this logic anyone who works long, busy hours catering to the needs of others should be tipped. Tip the doctor. Tip the grocery clerk. Tip the airline counter agent. Tip the airline pilot. Etc. I fail to see why those whose particular service happens to be delivering plates of food warrant their own method of compensation. One that puts an onus of extra calculation and deliberation on every single customer, every time they sit down to eat and relax. Thankfully, most other services in this world are one-price to the customer. It's left to the employer to do the work of assessing whether the employee is providing good service.	AGAINST
prostitution	"Granted, legalizing the profession might make it attractive for sex traffickers but the benefits outweigh this prospect." !!? The benefits of receiving tax revenue for the state outweighs the negative aspects of having 12 year old girls being sex trafficked into brothels and coerced to work on the streets?! What kind of logic is that, and what kid of person are you Ms. Unigwe?	AGAINST
vaccination	"It is a news media-driven misperception that parents who claim philosophical or religious exemptions are uneducated or misinformed. MOST PARENTS WHO INDIVIDUALIZE THE VACCINE SCHEDULE ARE ACTIVELY EDUCATING THEMSELVES, CONTINUALLY ASSESSING THEIR FAMILY'S SPECIFIC HEALTH NEEDS, and doing everything they can to keep their children safe and healthy." Ms. Margulis offers no data to support her blanket assertion about the industry and motivation of "most parents," a position which, on its face, seems improbable given the breadth and weight of scientific evidence supporting immunization which is not subject to reasonable dispute. Based on anecdotal information, "most parents" are refusing to vaccinate their children based on their gross misunderstandings and unwarranted fears of the alleged risks of immunization, or based on their unique interpretations of religious dictates. If you intend to offer relevant commentary, Ms. Margulis, it should be evidence-based. Your personal opinions are no more interesting (or informed) than mine.	FAVOR
mentally	'...food pornography, musical pornography, mental pornography...' And yet I have never been in a public library and seen a man at a computer masturbating under his coat to videos of food, music, or anything 'mental', so let's no pretend they are all the same beast, okay?	NONE

Table 4: Samples from VAST dataset with few different targets in both explicit and non-explicit settings.

either uses a synonymous term, is semantically similar to the golden one, or conveys the same underlying idea, topic, or issue as the golden target. This means that the generated target captures the essence or main concept of the golden target, even if it uses different wording. For example, if the golden target is "climate change," a conceptually similar term could be "global warming" or "environmental crisis," as both refer to the broader issue of environmental concerns related to climate.

Example Scoring:

1. **Tweet:** Attempts to conceal the creationism-evolution controversy from students are dogmatic promotions of evolution. Not since blasphemy laws has competitive expression of thought been illegalized, and this is what evolutionists want to accomplish. This is evidenced by none other than the title of an evolutionist argument on this very page: "Schools should not teach theories that are completely at odds with each other",
   Gold Target: creationism,
   T1: creationism-evolution controversy, **T1 Score:** 1,
   T2: Education controversy, **T2 Score:** 0.5.

2. **Tweet:** If, on a supernatural level, lust of any sort counts the same consummated or unconsummated, why not just go ahead and consummate. You can't get in any worse trouble, and

Model	TG+SD				TG&SD			
	SS	BTSD	HE	SC	SS	BTSD	HE	SC
				Explicit				
GPT-3.5	0.83	49.27	0.431	39.10	**0.88**	**49.70**	**0.439**	**40.63**
GPT-4o	0.83	49.70	0.449	45.93	**0.87**	**50.69**	**0.469**	**46.22**
Llama-3	0.84	47.07	0.435	40.51	**0.87**	**48.79**	**0.441**	**43.42**
Mistral	0.83	49.53	0.436	44.49	**0.86**	**50.17**	**0.479**	**48.72**
				Non-explicit				
GPT-3.5	0.76	38.78	0.377	36.69	**0.81**	**41.89**	**0.402**	**38.97**
GPT-4o	0.77	42.78	0.41	38.50	**0.80**	**45.70**	**0.436**	**39.26**
Llama-3	0.78	39.68	0.374	**38.35**	**0.82**	**39.75**	**0.386**	37.18
Mistral	0.77	41.36	0.4	38.71	**0.78**	**46.63**	**0.414**	**39.41**

Table 5: Comparison of TG and SD performance on EZSTANCE dataset (Zhao and Caragea, 2024) among the LLMs using various metrics: TG is evaluated with SS (SemSim), BTSD (%), and HE (Human Evaluation) scores, while SC (%) measures SD performance. Metric-wise best results within a specific setting are underlined, and better results of a model across TG+SD and TG&SD for each dataset are in **bold**.

Model	TSE						VAST					
		KA	FK		KA	FK		KA	FK		KA	FK
TSE-M (TG+SD)	Explicit	1	1	Non-explicit	1	1	Explicit	-	-	Non-explicit	-	-
TSE-B (TG+SD)		0.749	0.633		0.855	0.774		-	-		-	-
GPT-3.5 (TG+SD)		0.726	0.603		0.882	0.722		0.788	0.651		0.762	0.649
GPT-3.5 (TG&SD)		0.82	0.783		0.709	0.616		0.819	0.72		0.714	0.626
GPT-4 (TG+SD)		0.746	0.63		0.717	0.611		0.79	0.668		0.785	0.66
GPT-4 (TG&SD)		0.857	0.75		0.727	0.631		0.852	0.785		0.794	0.681
Lama-3 (TG+SD)		0.823	0.723		0.841	0.762		0.736	0.634		0.764	0.65
Lama-3 (TG&SD)		0.827	0.758		0.708	0.617		0.748	0.639		0.751	0.647
Mistral (TG+SD)		0.884	0.762		0.768	0.709		0.786	0.661		0.724	0.635
Mistral (TG&SD)		0.836	0.786		0.732	0.687		0.81	0.707		0.796	0.694

Table 6: Inter-annotator reliability score based on Krippendorff's α (KA) and Fleiss's κ (FK) coefficient metrics across all the combinations in two datasets. TSE-mapped and TSE-BestGen are reffered by TSE-M and TSE-B, respectively. The reliability score is 1 for TSE-M as it classifies between the golden targets instead of generating.

Metrics	**TG+SD**		**TG&SD**	
	Ex	**N-Ex**	**Ex**	**N-Ex**
BTSD-TSE	38.47	30.73	37.88	32.14
BTSD-VAST	41.10	41.54	41.34	40.38
SC-TSE	47.36	38.11	49.14	36.00
SC-VAST	48.47	46.46	49.00	45.41

Table 7: Caption goes here TG (evaluated by BTSD) and SD (evaluated by SC) performance using COT prompting on both TSE and VAST datasets in both Explicit and Non-explicit cases across the two approaches: TG+SD, TG&SD. Ex and N-Ex represent Explicit and Non-explicit cases, respectively.

if it makes you happy...,
Gold Target: gay rights,
T1: consummation, **T1 Score:** 0.5,
T2: superstition, **T2 Score:** 0.

L COT Initial Result

We initiate an exploration of alternative prompting techniques by employing the chain-of-thought (COT) prompting strategy for both target generation (TG) and stance detection (SD) tasks, across two approaches: TG+SD and TG&SD. As an initial phase, in Table 7, we assess target quality using the BTSD metric and evaluate stance detection performance with the SC metric, applied to both the TSE and VAST datasets for explicit and non-explicit cases.

M Human Annotation Score Distribution

After conducting the human annotation for target quality assessment, we plot the score distribution provided by the annotators, as shown in Figure 1. We observe that LLMs generally produce targets that are either partially or fully related to the golden targets, outperforming the TSE models in this regard. The targets generated by TSE-mapped are either identical to the golden targets or entirely unrelated, as `TSE-mapped` selects targets in its final stage (as described in Section 2) from a pre-defined list. In contrast, the targets generated by `TSE-BestGen` are often either partially or completely unrelated to the golden targets, indicating lower quality compared to LLMs.

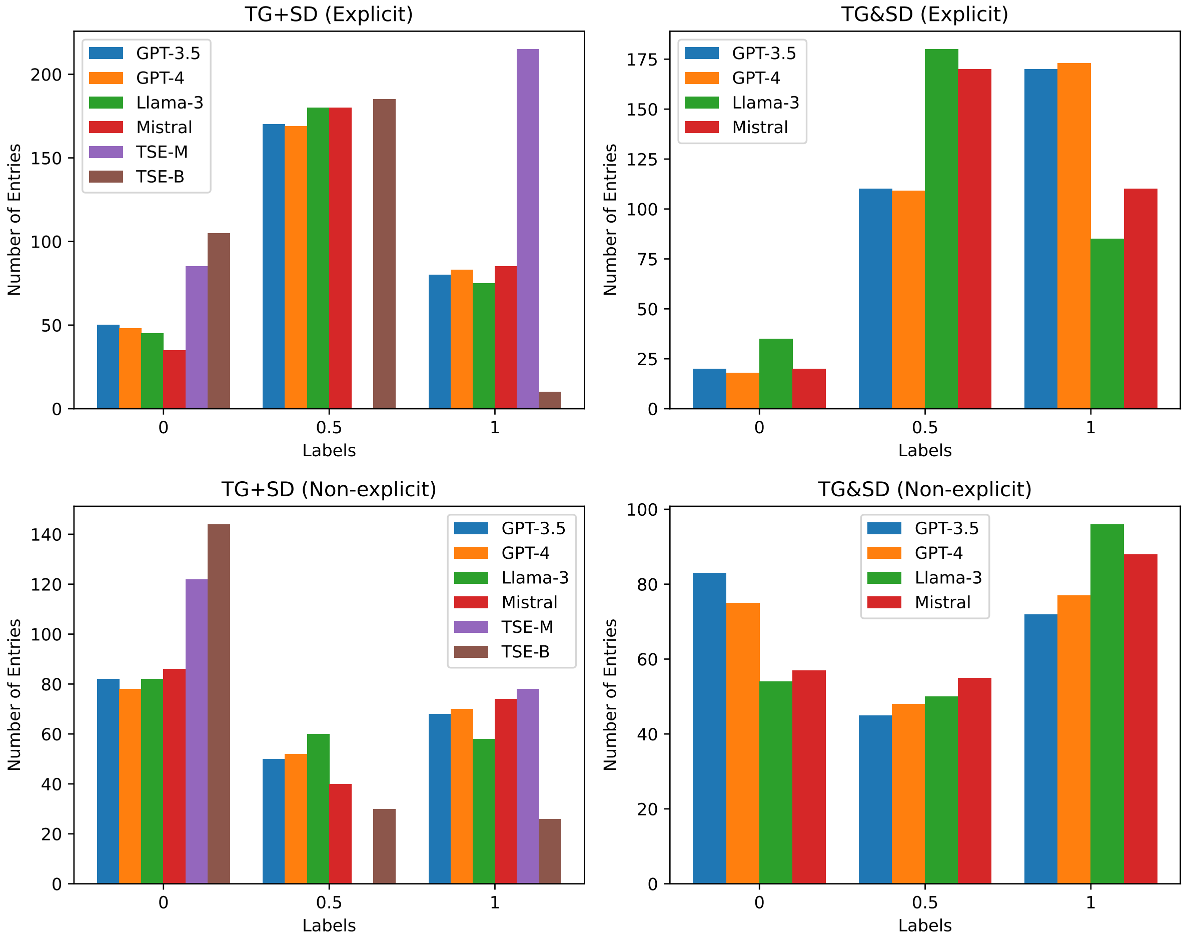

Figure 1: Distribution of human evaluation scores for target relevance assessment (0 – Not Related, 0.5 – Partially or Indirectly Related, 1 – Completely Related). The final score for each sample is determined by the majority vote among the three annotators. In cases of a tie, the average score of 0.5 is assigned. The details on human annotator guideline are provided in Appendix K, and the annotator agreement is in Appendix Table 6.

N Vsiual Comparison of TG and SD Performance

To better illustrate the data in Table 1, we have plotted Figures 2 and 3, which represent the TG and SD performance, respectively. These figures provide a clearer visualization of the conclusions discussed in Sections 5.1 and 5.2.

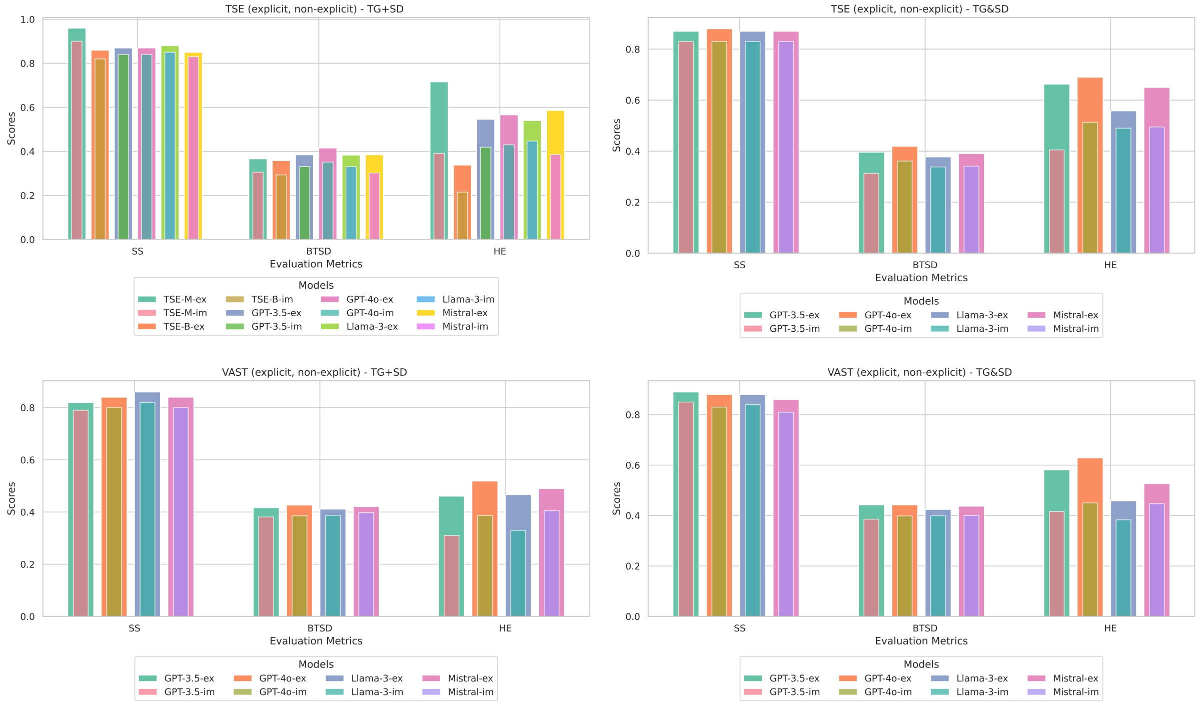

Figure 2: Comparison of TG performance of TSE and LLMs, as well as the performance among LLMs across both the TSE and VAST datasets, using various metrics: SS (Semantic Similarity), BTSD (%), and HE (Human Evaluation). The graphs are generated based on the data in Table 1.

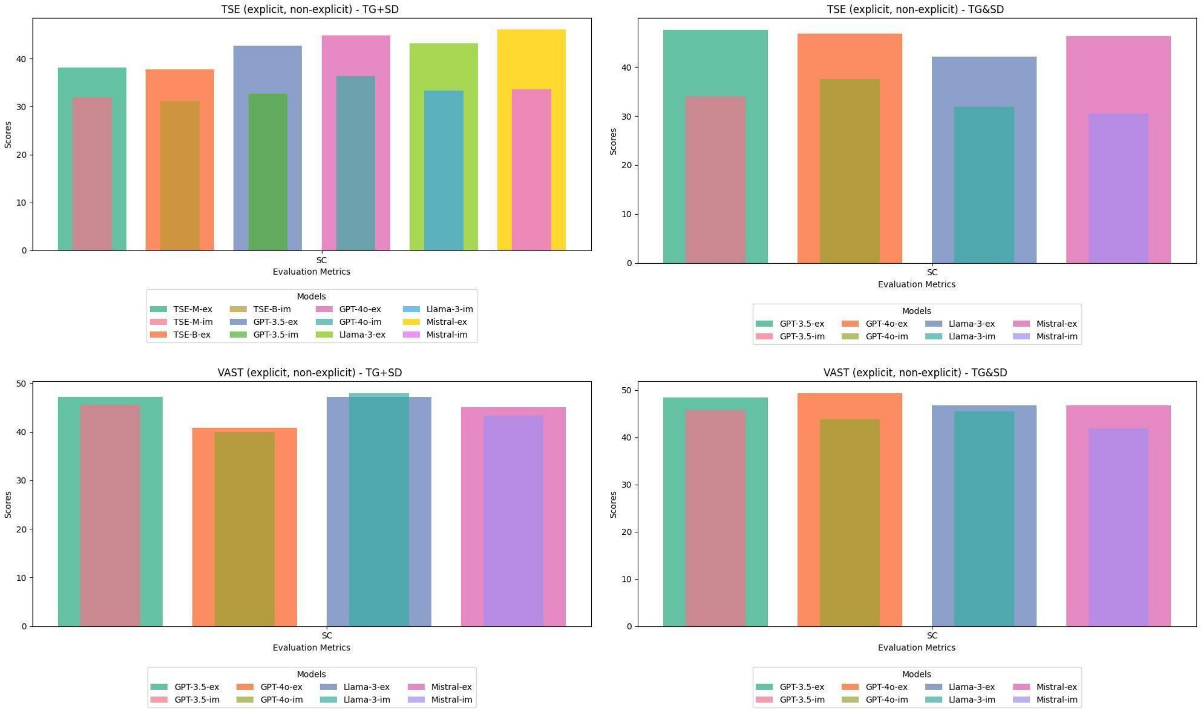

Figure 3: Comparison of SD performance of TSE and LLMs, as well as the performance among LLMs across both the TSE and VAST datasets, using SC (stance classification) score. The graphs are generated based on the data in Table 1.