Bilingual Zero-Shot Stance Detection

Chenye Zhao **Cornelia Caragea**
Computer Science
University of Illinois Chicago
czhao43@uic.edu cornelia@uic.edu

Abstract

Zero-shot stance detection (ZSSD) aims to determine whether the author of a text is in support, against, or neutral toward a target that is unseen during training. In this paper, we investigate ZSSD within a bilingual framework and compare it with cross-lingual and monolingual scenarios, in settings that have not previously been explored. Our study focuses on both noun-phrase and claim targets within in-domain and out-of-domain bilingual ZSSD scenarios. To support this research, we assemble Bi-STANCE, a comprehensive bilingual ZSSD dataset consisting of over 100,000 annotated text-target pairs in both Chinese and English, sourced from existing datasets. Additionally, we examine a more challenging aspect of bilingual ZSSD by focusing on claim targets with a low occurrence of shared words with their corresponding texts. As part of Bi-STANCE, we created an extended dataset that emphasizes this challenging scenario. To the best of our knowledge, we are the first to explore this difficult ZSSD setting. We investigate these tasks using state-of-the-art pre-trained language models (PLMs) and large language models (LLMs). We release our dataset and code at https://github.com/chenyez/BiSTANCE.

1 Introduction

Stance detection aims to automatically detect whether the author of a text is in support, against, or neutral toward a specific target such as entities (e.g., "free college") or claims (e.g., "We should subsidize student loans.") (Mohammad et al., 2016; Küçük and Can, 2020; ALDayel and Magdy, 2021). Identifying these stances offers crucial insights for events such as market analysis (Küçük and Can, 2020) and rumor detection (Wei et al., 2019).

Stance detection has been widely studied as in-target stance detection, cross-target stance detection, and zero-shot stance detection (ZSSD). In-target stance detection focuses on training and testing models on identical target(s), such as "clean energy" (Hasan and Ng, 2014; Mohammad et al., 2016; Graells-Garrido et al., 2020). For cross-target stance detection, models are trained with targets that are closely related but different from test targets, for example, using data related to "clean energy" for training and data related to "solar energy" for testing (Augenstein et al., 2016; Wei and Mao, 2019). Zero-shot stance detection (ZSSD), on the other hand, tests models on a large number of targets that are unseen (and unrelated) during training. Since it is hard to include every possible or related target in the training of in-target and cross-target approaches, ZSSD has emerged as a promising approach that more closely reflects real-world scenarios (Allaway and McKeown, 2020; Liu et al., 2021; Luo et al., 2022; Liang et al., 2022). To date, most of the existing works on stance detection study the task in a single language such as English (Mohammad et al., 2016; Conforti et al., 2020; Allaway and McKeown, 2020), Chinese (Xu et al., 2016; Zhao et al., 2023), Italian (Cignarella et al., 2020), Czech (Hercig et al., 2017), and Arabic (Al Hariri and Abu Farha, 2024). However, global topics such as the Ukraine War, NATO, and COVID-19 are widely discussed across various languages, making it essential for models to accurately identify stances in multiple linguistic contexts.

Only recently, several studies have started to explore stance detection from a multilingual perspective (Zotova et al., 2020; Vamvas and Sennrich, 2020; Lai et al., 2020), focusing primarily on in-target or cross-target tasks within a small set of targets. These works typically perform training and testing on targets within a specific domain—in-domain, such as the elections in Switzerland (Vamvas and Sennrich, 2020). Moreover, existing multilingual stance detection research tends to focus on targets as either noun phrases or claims. This narrow focus hinders the applicability of stance detection models in real-world scenarios, where stance classifiers are expected to handle

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29900–29919

Name	Language	Domain	# Target(s)	N/C	Task	Size
VaxxStance	Basque, Spanish	Vaccine	1	N	In-target, in-domain	4,081
TW-1O	Catalan, Spanish	Politics	1	N	In-target, in-domain	20,125
R-ita	English,Spanish, Catalonia, French, Italian	Politics	6	N	In-target, in-domain	14,440
X-STANCE	English, French, Germany, Italian	Politics	194	C	Cross-target, in-domain	67,000
Bi-STANCE	Chinese, English	Sports, Rights, Education, Politics, etc	86,082	N, C	ZSSD, in-domain, out-of-domain	100,782

Table 1: Comparison of our Bi-STANCE dataset with previous multilingual stance detection datasets: VaxxStance (Agerri et al., 2021), TW-10 (Taulé et al., 2017), R-ita (Lai et al., 2020), and X-STANCE (Vamvas and Sennrich, 2020). N and C represent noun-phrase and claim targets, respectively.

diverse zero-shot stance predictions toward unseen targets—including both noun phrases and claims—from a wide range of domains. The absence of studies on multilingual ZSSD restricts the robustness and adaptability of existing ZSSD models when encountering multilingual data. Therefore, it is crucial to develop multilingual ZSSD models that are not confined to specific target types or limited domains, enabling effective stance detection across diverse and multilingual contexts.

To address the above limitations, in this paper, we break the new ground by exploring bilingual zero-shot stance detection, conducting comprehensive comparisons with monolingual and cross-lingual settings to provide a thorough analysis of the performance differences. For each setting, we investigate targets as both noun phrases and claims. This exploration is essential as it addresses the unique challenges posed by bilingual contexts for ZSSD, which, to the best of our knowledge, have not been thoroughly studied before. We also delve into both in-domain and out-of-domain ZSSD, where classifiers trained on targets from diverse domains are evaluated on a large number of unseen targets within the same diverse domains and from an entirely new domain, respectively. To explore these tasks, we compile the first bilingual ZSSD dataset, Bi-STANCE from two existing large datasets for ZSSD: the Chinese C-STANCE dataset (Zhao et al., 2023) and the English EZ-STANCE dataset (Zhao and Caragea, 2024). Additionally, as part of our Bi-STANCE dataset, we construct a more challenging bilingual ZSSD scenario for claim targets with human annotations in which there is a very low word overlap between texts and claim targets. By doing so, we aim to enhance the versatility and effectiveness of stance detection models, making them more applicable to global, multilingual environments.

Our contributions can be summarized as follows:

- We investigate the task of bilingual zero-shot stance detection, and compare it with cross-lingual and monolingual ZSSD tasks. We uniquely address both noun phrase and claim targets within the contexts of in-domain and out-of-domain bilingual ZSSD, areas previously unexplored in the literature. We further explore a more challenging bilingual ZSSD scenario where claim targets exhibit low word overlap with their corresponding texts.
- To explore these tasks, we developed Bi-STANCE, a comprehensive bilingual dataset that includes over 100,000 annotated instances across a comprehensive set of domains, covering both Chinese and English data.
- We carry out extensive experiments to establish baseline results using both pre-trained language models and large language models.

2 Related Work

Multilingual Stance Detection Stance detection has received considerable attention in recent years (Liang et al., 2022; Liu et al., 2023; Li et al., 2023a,b). However, most previous studies have been limited to a single language (Hardalov et al., 2021; Schiller et al.; He et al., 2022; Li and Yuan, 2022; Wen and Hauptmann, 2023; Arakelyan et al., 2023). Despite substantial interest in multilingual tasks within the NLP domain, for stance detection, only a handful of studies have approached the subject from a multilingual perspective (Taulé et al., 2017; Vamvas and Sennrich, 2020; Lai et al., 2020; Agerri et al., 2021; Hardalov et al., 2022). Current multilingual stance detection efforts are largely focused on highly specific domains, featuring a limited variety of targets and types. These studies predominantly address in-target and cross-target stance detection tasks. Vamvas and Sennrich (2020) study multilingual stance detection for claim targets in the domain of Swiss independence. Other

Text:	Where is Morrison? Freedenberg and the general doing a media conference indicates he is not at work. Come on media flush him out. Maybe he is preparing his kids for home schooling.
Stance:	Support
Ori Claim:	The absence of Morrison and the media conference led by Freedenberg raises questions about Morrison's whereabouts, prompting speculation about him possibly preparing his kids for homeschooling.
New Claim1:	It's disheartening when our supposed representatives fail to truly address the issues that matter. They should be accountable and forthcoming with their activities to maintain public trust.
New Claim2:	The media's ability to raise questions and initiate discussions serves as a valuable mechanism for ensuring that politicians remain answerable to the public.
Text:	捷报！英格兰女足联赛杯：曼城女足凭借着福勒、布拉克斯塔的梅开二度，以及拉索和洛萨达的进球6比0击败布莱克本！ Great news! In the England Women's Football League Cup: Manchester City Women's team, thanks to Fowler and Blaxsta scoring twice, along with goals from Lasso and Losada, defeated Blackburn with 6-0 scores!
Stance:	Against
Ori Claim:	布莱克本英格兰女足比赛获得胜利 Blackburn wins the England Women's Football match.
New Claim1:	对于团体竞技运动的胜利，每个队员的贡献都是严格均等的。 In team sports victories, the contribution of every player is strictly equal.
New Claim2:	体育比赛的胜负并不重要，不应该被过分关注。 The outcome of sports competitions is not important and should not receive too much attention.

Table 2: Examples of generated challenging claims with corresponding texts, original claims, and stance labels.

research focus on noun-phrase targets within a singular domain. For example, Taulé et al. (2017) examined stance detection towards the independence of Catalan. Lai et al. (2020) concentrated on noun-phrase targets in the political domain (e.g., Hillary Clinton and Marine LePen), while Agerri et al. (2021) focused on stance detection toward vaccines. Since the training and testing targets in these studies are confined to the same domain, their scope is restricted to in-domain stance detection.

Contrasting with previous studies, we focus on zero-shot stance detection (ZSSD) from a bilingual perspective and develop the first dataset for bilingual ZSSD, encompassing a large number of noun-phrase and claim targets across a wide array of domains. We tackle two demanding multilingual ZSSD scenarios: in-domain and out-of-domain. In Table 1, we compare our work with existing multilingual stance detection works. Compared to previous efforts, our work encompasses a significantly larger number of targets and a more extensive dataset size, spanning a wider variety of domains. We conduct comprehensive investigations into both in-domain and out-of-domain ZSSD.

Language Models Pretrained language models (PLMs) such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have been extensively used in stance detection (Glandt et al., 2021; Allaway and McKeown, 2020; Li et al., 2021). More recently, large language models (LLMs) have been developed, offering the advantage of handling downstream tasks directly through prompting techniques (Le Scao et al., 2023; Touvron et al., 2023; Team et al., 2023; Naveed et al., 2023). While some research on stance detection has utilized LLMs (Gatto et al., 2023; Li et al., 2023a; Fraile-Hernandez and Peñas, 2024), these applications are limited to monolingual scenarios. Existing multilingual stance detection works solely employ PLMs (Vamvas and Sennrich, 2020). In contrast, our study leverages both PLMs and LLMs in the multilingual context, which enables us to explore ZSSD in a more comprehensive setting, enhancing our understanding and capabilities in this area.

3 The Bi-STANCE Benchmark

In this section, we introduce the Bi-STANCE benchmark, a large bilingual ZSSD dataset that includes both Chinese and English, consisting of 100,782 annotated instances spanning a comprehensive range of domains.

3.1 Data Sources

To create Bi-STANCE, we aggregate the C-STANCE dataset (Zhao et al., 2023) and the EZ-STANCE dataset (Zhao and Caragea, 2024). C-STANCE is the first large-scale ZSSD dataset for Chinese, comprising 48,126 annotated microblog-target pairs sourced from Sina Weibo (similar to Twitter) covering 7 domains: "Covid Epidemic" (CoE), "World Events" (WE), "Culture and Education" (CuE), "Entertainment and Consumption" (EC), "Sports" (S), "Rights" (R), and "Environmental Protection" (EP). EZ-STANCE is a recent large English ZSSD dataset. It includes 47,316 annotated tweet-target pairs collected from Twitter. This dataset categorizes data into the same seven domains available in C-STANCE, plus an additional "Politics" domain (P).

Target	Test	Train	LexSim
Noun Phrase	Chinese	English	12.96%
	English	Chinese	9.35%
Original Claim (OC)	Chinese	English	3.88%
	English	Chinese	3.78%
Challenging Claim (CC)	Chinese	English	3.45%
	English	Chinese	3.41%

Table 3: Percentage of *LexSimTopics* across target types within our Chinese and English datasets for in-domain ZSSD.

3.2 Extension to Challenging Claim Targets

We observe that in some cases, there is a high word overlap between the claim targets and texts in the datasets in both languages. Models can exploit such superficial lexical patterns and can infer the stance label without learning the semantic correlation between texts and targets. Examples of this can be found in Table 2 between 'Text' and 'Original Claim' for both Chinese and English datasets. This observation prompts us to investigate a more challenging ZSSD scenario for claim targets—whether models encounter greater difficulty when there is low word overlap between claim targets and texts and stances are expressed implicitly (in a more subtle way). Therefore, we specifically select samples from the Chinese and English claim targets in the Bi-STANCE dataset and instruct human annotators to develop more challenging claim targets that implicitly express the stance of the text, ensuring minimal word overlap with the texts.

3.2.1 Annotation for Bilingual Challenging Claim Dataset

We randomly sample a subset of the Bi-STANCE dataset to manually annotate for challenging claim targets. We performed annotations with two data annotation companies, Cogitotech[1] and Taojinniwo[2], for English and Chinese, respectively. Details about the annotation platform and quality assurance measures can be found in Appendix A. For each language, we select approximately 800 instances from the test set. To evaluate the models' adaptability to challenging claims with limited exposure during training, we randomly choose about 300 instances from the training set and 200 from the validation set, ensuring that these selected instances are evenly distributed across domains and stance categories. In total, we compile 2,670 instances, with 1,323 for Chinese and 1,347 for English.

To annotate challenging claim targets, annotators are provided with texts from the Bi-STANCE dataset, along with their corresponding original claim targets and stance labels. For each instance, we assign an annotator to generate two new claims. These new claims are crafted to ensure that the text's author maintains the same stance as with the original claim, but they are designed to express this stance more subtle, with minimal vocabulary overlap with the corresponding texts. To encourage annotators to generate claim targets with subtle stance, we advise them to: 1) focus on discussing the domain-level context rather than directly mentioning entities from the text, 2) employ logical reasoning to deduce other claims for which the text's author may hold a similar stance as for the original claim, and 3) when it is necessary to mention entities from the text, to try rephrasing them to their synonyms. This methodology encourages both creativity and relevance in formulating claim targets that diverge significantly from their corresponding texts in vocabulary usage. We provide the instructions for annotators in Appendix A.

Examples of new claim targets are shown in Table 2. In the English example (top block), unlike the original claim that directly discusses entities from the text (e.g., Morrison, Freedenberg, home schooling), our challenging claims discuss the topic from fresh perspectives: the necessity for representatives to uphold public trust (New Claim1), and the media's role in monitoring politicians' performance of their duties (New Claim2). In both cases, new claims remain the same stance orientation (support) yet exhibit much lower ratio of vocabulary overlap with the text. For quality assurance of the challenging claim targets, we use a separate group of annotators to determine the stance. This approach achieves a 96% agreement rate, indicating high-quality claim target generation. Finally, our annotated set comprises 1,323 microblogs with 2,646 (2 for each) claim targets in Chinese, and 1,347 tweets with 2,694 claim targets in English.

3.3 LexSimTopics across Languages

To ensure a zero-shot setting for our Bi-STANCE dataset, we analyze the occurrence of *LexSimTopics* (Allaway and McKeown, 2020; Zhao et al., 2023) across training and testing datasets in Chinese and English. *LexSimTopics* (Allaway and McKeown, 2020) is defined as the percentage of target expressions in the test set that achieve a cosine similarity greater than 0.9 with any target in the training dataset, within the word embedding space (Bojanowski et al., 2017). We begin by translating

[1]https://www.cogitotech.com/
[2]http://sjbz.itaojin.cn/

Setting	Language		# Examples			# Unique				Avg. Length			
			N	OC	CC	N	OC	CC	T	N	OC	CC	T
In-domain	Chinese	Train	13,258	20,160	630	6,093	19,694	630	6,740	3.7	15.1	16.3	62.2
		Val	2,865	4,419	420	2,665	4,400	420	1,473	4.6	15.4	18.6	62.5
		Test	2,915	4,509	1,596	2,865	4,487	1,596	1,503	4.7	15.4	15.3	63.7
	English	Train	13,756	18,879	576	7,437	18,861	576	6,293	1.8	19.0	20.4	40.0
		Val	2,354	4,349	420	2,284	4,345	420	1,454	2.4	19.0	20.5	40.0
		Test	2,663	5,135	1,698	2,621	5,130	1,698	1,715	2.4	19.3	19.9	39.7
Out-of-domain (CoE)	Chinese	Train	12,379	18,984	1,662	7,519	18,585	1,662	6,690	4.0	15.2	15.9	61.7
		Val	2,249	3,447	606	2,208	3,436	606	1,087	4.6	15.1	16.3	62.9
		Test	3,474	5,346	378	1,896	5,211	378	1,786	3.7	15.5	16.1	64.7
	English	Train	12,648	19,467	1,544	8,506	19,440	1,544	6,489	2.0	18.9	20.8	39.4
		Val	1,958	3,753	564	1,932	3,749	564	1,251	2.4	19.2	20.9	40.4
		Test	2,639	3,819	302	1,734	3,814	302	1,273	1.9	19.2	20.7	41.8

Table 4: Dataset split statistics for in-domain and out-of-domain ZSSD settings ("Covid Epidemic" (CoE) as the zero-shot domain). N, OC, CC, T represent noun-phrase targets, original claim targets, challenging claim targets, and texts, respectively.

	Noun-phrase targets						Original Claim targets					
	English			Chinese			English			Chinese		
Domain	Con	Pro	Neu	Con	Pro	Neu	Con	Pro	Neu	Con	Pro	Neu
CoE	971	812	853	1,444	1,247	783	1,329	1,328	1,327	1,782	1,782	1,782
WE	856	559	850	870	641	1,616	1,140	1,139	1,140	1,590	1,590	1,590
CuE	615	826	647	734	1,108	554	1,083	1,083	1,083	1,206	1,206	1,206
EC	636	925	1,084	1,355	1,480	1,175	1,405	1,406	1,405	2,051	2,051	2,051
S	179	781	808	435	766	885	941	942	941	1,059	1,059	1,059
R	910	1,015	522	1,020	940	532	1,191	1,192	1,191	1,276	1,276	1,276
EP	515	987	563	264	633	556	979	980	979	732	732	732
P	1,184	846	829	-	-	-	1,386	1,387	1,386	-	-	-
Overall	5,866	6,751	6,156	6,122	6,815	6,101	9,454	9,457	9,452	9,696	9,696	9,696

Table 5: Label distribution for noun-phrase targets and claim targets in each domain from Bi-STANCE. Con, Pro, Neu represent against, favor, and neutral, respectively.

Chinese target expressions into English using a pre-trained machine translation model (Tiedemann and Thottingal, 2020). Next, we calculate the LexSim-Topics ratio: first, between the Chinese test set and the English training set, and then between the English test set and the Chinese training set. Table 3 presents the results. For noun-phrase targets within the English test set, the *LexSimTopics* ratio of 9.35% indicates that this proportion of targets is similar to targets from the Chinese training set, which is lower than ratios observed for previous ZSSD datasets such as C-STANCE (11%) (Zhao et al., 2023), EZ-STANCE (12%) (Zhao and Caragea, 2024), and VAST (16%) (Allaway and McKeown, 2020). For claim targets, the *LexSim-Topics* ratios are notably lower, which can be attributed to the longer and more distinct nature of claim targets, reducing their likelihood of similarity. These findings suggest that our Bi-STANCE dataset maintains a zero-shot setting across the two languages. The details on LexSimTopics for out-of-domain ZSSD are provided in Appendix B.

3.4 Dataset Statistics

We retain the original dataset split for C-STANCE and EZ-STANCE datasets for both in-domain and out-of-domain ZSSD. For in-domain ZSSD, both datasets are split into distinct training, validation, and test sets, ensuring that no texts and targets are shared among them. For out-of-domain ZSSD, each domain in turn is selected as the zero-shot domain for testing, while using the remaining domains for training and validation, again preventing any overlap of texts and targets across all sets. For our newly-annotated challenging claim targets with low word overlap with texts (denoted as CC), our texts sourced from the training, validation, and test sets of the Bi-STANCE dataset were assigned to the corresponding dataset split from which they were sampled. For example, texts originating from the Bi-STANCE test set were allocated to the challenging test set of Bi-STANCE.

Dataset Size The dataset statistics are shown in Table 4, where we observe that the Chinese subset and the English subset includes similar sizes and dataset splits for each target type.

Target Size Table 4 provides the number of unique targets and texts in each set, referred to as # Unique. Overall, Bi-STANCE includes 40,204 distinct Chinese targets, consisting of 11,623 noun phrases (N), 28,581 original claims (OC), and 2,646 challenging claims (CC). For English tar-

Language		Con	Pro	Neu	All
Chinese	OC	34.10	55.33	28.61	39.34
	CC	19.87	23.02	20.26	21.05
English	OC	36.16	45.89	25.48	35.85
	CC	16.81	18.35	18.78	17.98

Table 6: Average token overlap percentages (%) for claim targets: OC and CC denote original and challenging claim targets, respectively.

gets, there are 12,342 noun phrases, 28,336 original claims, and 2,694 challenging claims. We show the analysis on target diversity in Appendix C.

Text/target Length Moreover, we observe that Chinese instances have longer average token lengths for noun-phrase targets and texts, whereas English instances have longer lengths for claim targets (Avg. Length). The full statistics of out-of-domain ZSSD are shown in Appendix D.

Label Distribution Table 5 shows the stance label distribution across domains for original Chinese and English data. We can observe a varied nature of public opinion across different domains. For our newly developed challenging claim targets, we ensure an even distribution across stance categories and domains, as detailed in Appendix E.

Token Overlap We also examine the average token overlap percentage between original and challenging claim targets from Bi-STANCE, defined as the average proportion of words in the claim target that also appear in the corresponding text. Results are shown in Table 6, where challenging claims in both languages demonstrate significantly lower overlap percentage compared with original claims.

4 Models

Here we introduce the multilingual models that we evaluate in our experiments.

Multi-lingual PLM Baselines. We fine-tune the base variants of the following state-of-the art pre-trained language models (PLMs) in our experiments: **mBERT** (Devlin et al., 2019); **XLM-R** (Conneau et al., 2020); and **mT5** (Xue et al., 2021).

Multi-lingual LLM Baselines. We also experiment with multi-lingual large language models (LLMs) on our dataset: **BLOOM** (Le Scao et al., 2023), **LLaMA 2** (Touvron et al., 2023), **LLaMA 3**, **ChatGPT**, and **Gemini** (Team et al., 2023). We construct templates with task description and three in-context examples (one for each class) to prompt the LLMs. Further details on training settings and prompt template design are in Appendix F.

5 Results

In this section, we conduct experiments on bilingual ZSSD and compare the results with cross-lingual and monolingual ZSSD tasks. First, we perform experiments within both the in-domain (§5.1) and out-of-domain (§5.2) settings using the original Bi-STANCE data. Next, we evaluate models using our challenging claim targets and investigate how well models can adjust to the challenging claims if a very small amount of challenging data is available for the models (§5.3). Like prior works (Allaway and McKeown, 2020; Zhao et al., 2023), we employ the macro-averaged F1 score across all classes as our evaluation metrics. Each result is the average of 4 runs with different initializations.

5.1 Bilingual In-domain ZSSD

In-domain ZSSD aims to train a stance classifier on targets from various domains and test it on completely unseen targets from the same domains. We use three training settings: 1) bilingual data (B), 2) Chinese-only data (C), and 3) English-only data (E). Each setting includes a mix of noun-phrase and claim targets. For each setting, models are evaluated on bilingual, Chinese-only, and English-only test sets. Within each test set, evaluations are conducted on: 1) the full test set with both noun-phrase and claim targets, 2) the subset with noun-phrase targets, and 3) the subset with claim targets.

Results are shown in Table 7, where we make the following observations. First, models demonstrate much worse performance in the cross-lingual scenario (C→E or E→C) than in the monolingual scenario (E→E or C→C). For instance, mBERT, when trained on Chinese data, scores only a 49.8% $F1_{macro}$ when evaluated on English *mixed targets* (C→E), yielding a 22.9% decrease compared to its performance when trained with English data (E→E). This indicates that even state-of-the-art multilingual pre-trained language models struggle to generalize across languages. Second, when trained on the full Bi-STANCE dataset, models achieve much better performance when compared with their performance in the cross-lingual scenario. For example, when trained with Bi-STANCE, mBERT demonstrates 23.6% improvement on the $F1_{macro}$ of the English *mixed targets* (B→E) over its performance when trained solely on the Chinese data (C→E), which is indicated by red arrows in Table 7. Third, models trained on the bilingual scenario (B→C, B→E) achieve on par or (in many

Train/Val	Bi-STANCE (B)			Chinese Only (C)			English Only (E)		
Test	B	C	E	B	C	E	B	C	E
Mixed Targets									
mBERT	73.5	73.6*	73.4†	61.7	73.2$_{\downarrow 0.4}$	49.8$_{\downarrow 23.6}$	64.7	55.9$_{\downarrow 17.7}$	72.7$_{\downarrow 0.7}$
XLM-R	76.4	76.9*	76.0†	67.5	76.7$_{\downarrow 0.1}$	58.0$_{\downarrow 18.0}$	69.5	63.1$_{\downarrow 13.8}$	75.6$_{\downarrow 0.4}$
mT5	75.3	76.3*	74.3†	63.6	74.8$_{\downarrow 1.5}$	51.8$_{\downarrow 22.5}$	67.3	62.6$_{\downarrow 13.7}$	71.6$_{\downarrow 2.7}$
BLOOM	33.3	34.2*	32.4	34.0	33.5$_{\downarrow 0.7}$	32.0$_{\downarrow 0.4}$	34.0	32.7$_{\downarrow 1.5}$	34.6$_{\uparrow 2.4}$
LLaMA 2	49.0	51.2*	46.7†	48.0	50.5$_{\downarrow 0.7}$	45.2$_{\downarrow 1.5}$	47.7	46.1$_{\downarrow 5.1}$	46.2$_{\downarrow 0.5}$
ChatGPT	43.9	43.4	45.5†	42.2	42.4$_{\downarrow 1.0}$	42.4$_{\downarrow 3.1}$	43.7	42.9$_{\downarrow 0.5}$	44.6$_{\downarrow 0.9}$
Gemini	59.7	65.3	53.9†	58.9	65.1$_{\downarrow 0.2}$	51.8$_{\downarrow 2.1}$	59.1	63.8$_{\downarrow 1.5}$	53.1$_{\downarrow 0.8}$
Noun-phrase Targets									
mBERT	61.8	62.6*	60.7†	57.7	61.1$_{\downarrow 1.5}$	53.6$_{\downarrow 7.1}$	53.0	48.8$_{\downarrow 13.8}$	57.4$_{\downarrow 3.3}$
XLM-R	65.9	66.6*	65.1†	64.0	65.8$_{\downarrow 0.8}$	61.8$_{\downarrow 3.3}$	61.4	60.0$_{\downarrow 6.6}$	63.1$_{\downarrow 2.1}$
mT5	65.2	66.7*	63.5†	59.9	65.6$_{\downarrow 1.1}$	52.3$_{\downarrow 11.2}$	58.1	56.7$_{\downarrow 10.0}$	59.3$_{\downarrow 4.3}$
BLOOM	33.8	34.3*	33.1	33.3	32.3$_{\downarrow 2.0}$	32.1$_{\downarrow 1.0}$	34.1	33.2$_{\downarrow 1.1}$	34.4$_{\uparrow 1.3}$
LLaMA 2	55.6	53.2*	58.1†	54.9	52.4$_{\downarrow 0.8}$	56.0$_{\downarrow 2.1}$	53.8	46.9$_{\downarrow 6.3}$	58.5$_{\uparrow 0.4}$
ChatGPT	46.7	46.9*	46.3†	44.9	45.9$_{\downarrow 1.0}$	44.0$_{\downarrow 2.4}$	43.6	41.2$_{\downarrow 5.7}$	45.9$_{\downarrow 0.4}$
Gemini	61.6	63.6*	59.2†	60.9	63.1$_{\downarrow 0.5}$	57.2$_{\downarrow 2.0}$	60.8	61.8$_{\downarrow 1.8}$	59.0$_{\downarrow 0.2}$
Original Claim Targets									
mBERT	80.2	80.6*	79.9†	64.1	80.7$_{\uparrow 0.2}$	46.5$_{\downarrow 33.4}$	71.1	60.1$_{\downarrow 20.5}$	80.2$_{\uparrow 0.3}$
XLM-R	82.5	83.4*	81.7†	69.6	83.4$_{\downarrow 0.1}$	55.2$_{\downarrow 26.5}$	74.1	64.9$_{\downarrow 18.5}$	81.8$_{\uparrow 0.1}$
mT5	81.0	82.3*	79.6†	66.0	80.5$_{\downarrow 1.8}$	51.1$_{\downarrow 28.5}$	72.3	66.2$_{\downarrow 16.2}$	77.3$_{\downarrow 2.4}$
BLOOM	33.0	34.1*	32.0	34.4	34.4$_{\uparrow 0.3}$	31.4$_{\downarrow 0.6}$	33.9	32.4$_{\downarrow 1.7}$	33.1$_{\uparrow 1.3}$
LLaMA 2	45.0	49.8*	40.3†	44.1	47.9$_{\downarrow 1.9}$	39.0$_{\downarrow 1.3}$	44.0	45.6$_{\downarrow 4.2}$	39.3$_{\downarrow 1.0}$
ChatGPT	41.1	39.2*	43.1†	36.9	38.4$_{\downarrow 0.8}$	35.4$_{\downarrow 7.6}$	40.2	37.7$_{\downarrow 1.5}$	42.7$_{\downarrow 0.4}$
Gemini	58.6	66.5*	51.0†	57.7	66.2$_{\downarrow 0.3}$	48.9$_{\downarrow 2.1}$	58.2	65.2$_{\downarrow 1.3}$	50.1$_{\downarrow 0.9}$

Table 7: Comparison of F1$_{macro}$ (%) of multilingual models trained on the mixed targets (mixture of noun-phrase and claim targets) in the bilingual, monolingual, and cross-lingual settings on in-domain ZSSD. B, C, E represent the full Bi-STANCE data, the Chinese subset, and the English subset, respectively. $*$ and †: models trained on B surpass their cross-lingual counterparts at $p < 0.05$ with paired t-test on Chinese test set and English test set, respectively. Blue and red arrows show performance changes for models in monolingual and cross-lingual settings compared to those in bilingual settings, respectively.

cases) even better performance than the monolingual counterparts (C→C, E→E) (denoted as blue arrows in Table 7), suggesting that the combination of ZSSD data from Chinese and English boosts models' stance prediction ability of both languages. Additionally, fine-tuning PLMs results in higher performance than LLMs without additional training. Among the LLMs we compared, Gemini performed the best, followed by LLaMA2 (13B), but both still lag behind the fine-tuned PLMs. Finally, most models yield higher performance on the claim targets than the noun-phrase targets. This could be because claim targets generally provide more context to the models, making stance prediction easier. We show results comparing different LLaMA LLMs in Appendix G. We also train models only on noun-phrase targets and claim targets, as detailed in Appendix H and results integrating VAST into Bi-STANCE in Appendix I.

5.2 Bilingual Out-of-domain ZSSD

Out-of-domain ZSSD aims at evaluating classifiers on unseen targets from new domains. One domain is designated as the left-out domain, with remaining domains serving as source domains. Models are trained using data from source domains and evaluated on data from the left-out domain, resulting in eight out-of-domain settings. Models are trained on the full Bi-STANCE dataset with mixed-targets (noun-phrases and original claims) and evaluated on the mixed-target test set of 1) Bi-STANCE dataset; 2) the Chinese subset; and 3) the English subset, denoted as B, C, and E, respectively.

Table 8 shows F1$_{macro}$ for eight zero-shot domain settings. First, we notice that models show lower performance when compared with the in-domain task (see results in Table 7). This is because the domain shifts between the training and testing stages introduce additional complexity to the task, making out-of-domain ZSSD a more challenging ZSSD task. Second, when fine-tuned on the Bi-STANCE training set, PLM models generally show higher performance when predicting stances for the "Sports" (S) and the "Environmental Protection" (EP) domain. Third, LLMs demonstrate varying stance prediction capabilities across different domains. For instance, Gemini outperforms in the "World Event" (WE) and "Sports" (S) domains, while ChatGPT excels in "Culture and Education" (CuE) and "Entertainment and Consumption" (EC). This variation can be attributed to the different tasks and data distributions on which the

Model		CoE	WE	CuE	EC	S	R	EP	P
mBERT	B	69.3	70.8	71.4	69.1	73.0	72.4	71.6	68.6
	C	70.5	72.0	72.3	70.5	74.4	72.1	73.9	-
	E	67.7	68.9	70.4	67.0	71.3	72.7	69.7	68.6
XLM-R	B	73.8	74.3	74.5	74.7	76.1	75.5	75.5	72.8
	C	75.2	75.3	75.6	75.9	78.1	75.9	78.2	-
	E	71.9	72.8	73.2	72.7	73.8	75.0	73.2	72.8
mT5	B	73.0	73.9	73.1	72.9	75.3	75.1	75.3	69.9
	C	74.5	75.0	74.4	74.8	76.9	75.2	78.2	-
	E	70.9	72.2	71.6	69.6	73.3	74.9	72.9	69.9
BLOOM	B	30.4	30.4	30.6	30.4	28.2	25.7	32.1	24.9
	C	33.4	32.9	33.8	33.2	30.9	27.6	35.9	-
	E	26.3	26.2	26.9	26.2	24.5	23.3	28.2	24.9
LLaMA 2	B	45.4	48.4	48.8	49.2	51.5	47.2	53.3	43.5
	C	46.2	50.0	50.9	50.7	53.0	48.2	56.8	-
	E	43.8	44.9	46.0	46.9	49.4	44.9	49.8	43.5
ChatGPT	B	41.1	41.4	45.0	44.9	43.6	39.2	43.7	36.8
	C	41.0	40.3	43.6	44.6	43.8	36.4	40.1	-
	E	38.4	40.0	46.4	43.7	41.2	41.5	45.1	36.8
Gemini	B	57.7	60.1	56.1	54.8	59.6	57.5	57.7	49.6
	C	61.7	63.3	56.1	56.1	63.6	61.4	63.0	-
	E	52.0	54.2	56.1	53.2	54.0	52.5	52.2	49.6

Table 8: Comparison of $F1_{macro}$ (%) of models on out-of-domain ZSSD. Models are trained and evaluated using datasets for 8 zero-shot domain settings (denoted by each column). Models are trained on the full bilingual training set with mixed targets. Test results are based on the mixed targets of B, C, E (the full Bi-STANCE, the Chinese subset, and the English subset, respectively).

LLMs were pretrained. Last, generally, all models show worse results on the "Covid Epidemic" (CoE) and the "Politics" (P) domain, suggesting that they share less domain knowledge with other domains, making them more difficult zero-shot domains.

5.3 Evaluating on Challenging Claim Targets

In this section, we evaluate bilingual models using our challenging claim targets with low word overlap with corresponding texts (denoted as C_{CC} for Chinese, E_{CC} for English). We train models using 1) original Chinese claim targets (C_{OC}), 2) original English claim targets (E_{OC}), 3) original bilingual claim targets (B_{OC}), and 4) the combination of original and challenging bilingual claim targets (B_{OC}+B_{CC}). At the inference stage, we evaluate models on: 1) Chinese original claim targets (C_{OC}), 2) Chinese challenging claim targets (C_{CC}), 3) English original claim targets (E_{OC}), and 4) English challenging claim targets (E_{CC}). For this experiments, we used PLMs and selected the best performing Gemini as the representative LLM.

Results for in-domain ZSSD are shown in Table 9, where we make the following observations. First, models trained in a bilingual setting outperform those in cross-lingual settings (e.g., B_{OC}→C_{CC} vs. E_{OC}→C_{CC}), reinforcing the advantage of developing a bilingual dataset. Second, models trained on original claim targets perform poorly on challenging claim targets. For instance, when trained on the original Bi-STANCE and evaluated on Chinese challenging claims (B_{OC}→C_{CC}), mT5 shows a 35.4% decrease in $F1_{macro}$ compared to its performance on original claim targets (B_{OC}→C_{OC}). This significant drop indicates that challenging claim targets, with greater vocabulary divergence, introduce additional difficulties. Third, minimal exposure to challenging claim targets during training improves performance, e.g., when trained on a mix of original and challenging bilingual claims (B_{OC}+B_{CC}→E_{CC} vs. B_{OC}→E_{CC}), mT5 shows a 4.8% increase in F1-macro. Results for out-of-domain ZSSD are detailed in Appendix J.

Model	Train	Test			
		C_{OC}	C_{CC}	E_{OC}	E_{CC}
MBERT	C_{OC}	80.4	41.7↓38.7	42.9	37.9↓5.0
	E_{OC}	57.1	32.5↓24.7	79.9	38.6↓41.3
	B_{OC}	80.5	41.1↓39.4	79.7	38.0↓41.7
	B_{OC}+B_{CC}	80.7	43.6↑2.5	80.0	42.2↑4.2
XLM-R	C_{OC}	83.6	46.1↓37.4	53.7	39.6↓14.1
	E_{OC}	60.1	32.2↓27.8	82.3	40.1↓42.2
	B_{OC}	84.2	46.3↓37.9	83.4	40.5↓42.9
	B_{OC}+B_{CC}	84.3	50.3↑4.0	83.5	41.6↑1.1
mT5	C_{OC}	81.3	45.3↓36.0	46.6	39.1↓7.5
	E_{OC}	64.2	34.2↓30.0	78.3	39.0↓39.4
	B_{OC}	81.0	45.7↓35.4	79.3	40.5↓38.8
	B_{OC}+B_{CC}	81.9	49.4↑3.7	79.4	45.3↑4.8
Gemini	C_{OC}	65.8	47.0↓18.8	46.1	43.7↓2.4
	E_{OC}	65.0	47.9↓17.1	49.1	47.6↓1.4
	B_{OC}	66.3	49.9↓16.5	52.0	49.8↓2.2
	B_{OC}+B_{CC}	66.4	54.2↑4.3	52.1	51.4↑1.6

Table 9: Comparison of $F1_{macro}$ (%) between challenging claim tagets and original claim targets for in-domain setting. B_{OC}, C_{OC}, E_{OC} represent Bi-STANCE's bilingual, Chinese, and English data with original claim targets, respectively. B_{CC}, C_{CC}, E_{CC} represent data with challenging targets. Red arrows depict performance shifts from challenging to original claim targets (e.g., B_{OC}→C_{CC} vs. B_{OC}→C_{OC}), and green arrows show the impact of integrating challenging targets for training (e.g., B_{OC}+B_{CC}→C_{CC} vs. B_{OC}→C_{CC}).

6 Conclusion

In this paper, we investigate zero-shot stance detection (ZSSD) in a bilingual scenario, comparing it to monolingual and cross-lingual settings. We explore bilingual ZSSD for both noun-phrase and claim targets within both in- and out-of-domain ZSSD settings. To investigate these tasks, we present Bi-STANCE, a large bilingual ZSSD dataset of over 100,000 annotated Chinese and English instances covering both noun-phrase targets and claim targets from a diverse set of domains. We also explore a more challenging ZSSD scenario where claim targets have low word overlap with their corresponding texts. We hope our work facilitates future research in multilingual stance detection.

Limitations

Our Bi-STANCE data originates from social media, which may be perceived as a drawback as it does not cover all aspects of more formal texts found in essays or news articles. In our future work, we plan to broaden the dataset to include various types of texts, such as research articles. However, this limitation is not exclusive to our dataset but is common to all datasets focusing primarily on social media content.

Ethical Statement

Our dataset is derived from two publicly accessible benchmarks. Data for these benchmarks were collected using common keywords based on popular topics from social social media websites, ensuring the dataset does not focus on information from any single user. As a result, our dataset adheres to the privacy policies social networking websites such as Twitter and Sina Weibo, maintaining compliance with established data protection standards.

References

Rodrigo Agerri, Roberto Centeno, Marıa Espinosa, Joseba Fernandez de Landa, and Alvaro Rodrigo. 2021. Vaxxstance: A dataset for cross-lingual stance detection on vaccines.

Youssef Al Hariri and Ibrahim Abu Farha. 2024. SMASH at StanceEval 2024: Prompt engineering LLMs for Arabic stance detection. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 800–806, Bangkok, Thailand. Association for Computational Linguistics.

Abeer ALDayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Inf. Process. Manage.*, 58(4).

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.

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13448–13464, Toronto, Canada. Association for Computational Linguistics.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance detection with bidirectional conditional encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Alessandra Teresa Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, Paolo Rosso, et al. 2020. Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets. In *CEUR WORKSHOP PROCEEDINGS*, pages 1–10. Ceur.

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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1724, Online. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesus M. Fraile-Hernandez and Anselmo Peñas. 2024. HAMiSoN-generative at ClimateActivism 2024: Stance detection using generative large language models. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 79–84, St. Julians, Malta. Association for Computational Linguistics.

Joseph Gatto, Omar Sharif, and Sarah Preum. 2023. Chain-of-thought embeddings for stance detection on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4154–4161, Singapore. Association for Computational Linguistics.

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 (Volume 1: Long Papers), pages 1596–1611, Online. Association for Computational Linguistics.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2020. Representativeness of abortion legislation debate on twitter: A case study in argentina and chile. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 765–774, New York, NY, USA. Association for Computing Machinery.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. Few-shot cross-lingual stance detection with sentiment-based pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10729–10737.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Zihao He, Negar Mokhberian, and Kristina Lerman. 2022. Infusing knowledge from Wikipedia to enhance stance detection. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis*, pages 71–77, Dublin, Ireland. Association for Computational Linguistics.

Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in czech news commentaries. *ITAT*, 176:180.

Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Comput. Surv.*, 53(1).

Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Farías, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech Language*, 63:101075.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023a. Stance detection on social media with background knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15703–15717, Singapore. Association for Computational Linguistics.

Yang Li and Jiawei Yuan. 2022. Generative data augmentation with contrastive learning for zero-shot stance detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6985–6995, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023b. 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, Toronto, Canada. Association for Computational Linguistics.

Yingjie Li, Tiberiu Sosea, Aditya Sawant, Ajith Jayaraman Nair, Diana Inkpen, and Cornelia Caragea. 2021. P-stance: A large dataset for stance detection in political domain. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2355–2365, Online. Association for Computational Linguistics.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022. JointCL: A joint contrastive learning framework for zero-shot stance detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Rui Liu, Zheng Lin, Yutong Tan, and Weiping Wang. 2021. Enhancing zero-shot and few-shot stance detection with commonsense knowledge graph. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu, Yong Keong Yap, Hai Leong Chieu, and Nancy Chen. 2023. Guiding computational stance detection with expanded stance triangle framework. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3987–4001, Toronto, Canada. Association for Computational Linguistics.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016.

SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Nick Barnes, and Ajmal Mian. 2023. A comprehensive overview of large language models. *arXiv preprint arXiv:2307.06435*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. Stance detection benchmark: How robust is your stance detection? *KI-Künstliche Intelligenz*, pages 1–13.

Mariona Taulé, M Antonia Martí, Francisco M Rangel, Paolo Rosso, Cristina Bosco, Viviana Patti, et al. 2017. Overview of the task on stance and gender detection in tweets on catalan independence at ibereval 2017. In *2nd Workshop on Evaluation of Human Language Technologies for Iberian Languages, IberEval 2017*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jannis Vamvas and Rico Sennrich. 2020. X-Stance: A multilingual multi-target dataset for stance detection. In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS)*, Zurich, Switzerland.

Penghui Wei and Wenji Mao. 2019. Modeling transferable topics for cross-target stance detection. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, page 1173–1176, New York, NY, USA. Association for Computing Machinery.

Penghui Wei, Nan Xu, and Wenji Mao. 2019. Modeling conversation structure and temporal dynamics for jointly predicting rumor stance and veracity. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4787–4798, Hong Kong, China. Association for Computational Linguistics.

Haoyang Wen and Alexander 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, Toronto, Canada. Association for Computational Linguistics.

Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. 2016. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications*, pages 907–916, Cham. Springer International Publishing.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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.

Chenye Zhao, Yingjie Li, and Cornelia Caragea. 2023. C-STANCE: A large dataset for Chinese zero-shot stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13369–13385, Toronto, Canada. Association for Computational Linguistics.

Elena Zotova, Rodrigo Agerri, Manuel Nuñez, and German Rigau. 2020. Multilingual stance detection in tweets: The Catalonia independence corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1368–1375, Marseille, France. European Language Resources Association.

Domain	Target	Test	Train	LexSim
CoE	N	Chinese	English	13.13%
	N	English	Chinese	11.94%
	OC	Chinese	English	3.57%
	OC	English	Chinese	4.38%
	CC	Chinese	English	3.43%
	CC	English	Chinese	3.80%
WE	N	Chinese	English	20.49%
	N	English	Chinese	13.14%
	OC	Chinese	English	3.85%
	OC	English	Chinese	4.12%
	CC	Chinese	English	3.88%
	CC	English	Chinese	3.79%
CuE	N	Chinese	English	17.17%
	N	English	Chinese	12.54%
	OC	Chinese	English	3.43%
	OC	English	Chinese	2.70%
	CC	Chinese	English	3.40%
	CC	English	Chinese	2.93%
EC	N	Chinese	English	13.69%
	N	English	Chinese	10.15%
	OC	Chinese	English	3.01%
	OC	English	Chinese	3.49%
	CC	Chinese	English	3.15%
	CC	English	Chinese	3.31%
S	N	Chinese	English	10.35%
	N	English	Chinese	8.19%
	OC	Chinese	English	2.37%
	OC	English	Chinese	3.51%
	CC	Chinese	English	2.43%
	CC	English	Chinese	3.32%
R	N	Chinese	English	14.12%
	N	English	Chinese	10.35%
	OC	Chinese	English	3.06%
	OC	English	Chinese	2.51%
	CC	Chinese	English	3.22%
	CC	English	Chinese	3.19%
EP	N	Chinese	English	14.87%
	N	English	Chinese	13.86%
	OC	Chinese	English	2.82%
	OC	English	Chinese	3.36%
	CC	Chinese	English	2.95%
	CC	English	Chinese	2.96%

Table 10: Percentage of *LexSimTopics* between the two languages for out-of-domain ZSSD. N, OC, CC represent noun-phrase targets, original claim targets, and challenging claim targets, respectively.

A Details on Challenging Claim Annotation

A.1 Annotation Platform

Our English data annotations were obtained through Cogitotech,[3] a premier data annotation company recognized for its work with top AI firms, including OpenAI and AWS. For the Chinese data, annotations were collected from Taojinniwo,[4] a crowd-sourcing platform in China known for providing annotation services to major AI corporations such as Baidu and JD. For both companies, we implemented strict quality standards for annotations by establishing specific requirements. First, all annotators must hold at least a college degree. Second, they must be native speakers of the language they are annotating. Additionally, we conduct quality reviews on a random 10% sample of each annotator's work. Any annotator with an approval rating below 90% is excluded from the project. Annotations that fail these quality checks are reassigned to other annotators for re-labeling.

A.2 Annotation Instructions

We provide the following instructions: "*Based on the message that you learned from the text and the claim target, write two additional claim targets, to which the author of the text would express the same stance as it is toward the original claim. The definition of stance labels are as follows. "Favor": The author is definitely in favor of the point or message of the claim; "Against": The author is definitely against the point or message from the claim; "Neutral": Based solely on the information from the text, we cannot know whether the author definitely supports or opposes the point or message of the claim.*"

To make this task more challenging, we establish an extra requirements: claims labeled with *against* should not merely negate the tweet content (e.g., adding "not" before verbs). Models could easily detect such linguistic patterns and predict stances without learning the content of tweet-claim pairs.

A.3 Quality Check

To ensure the actual stance is indeed the intended one as assigned by the annotator who generated the text, we hide the stance label and present the input text and the generated challenging claim to a set of three annotators (different from those who

[3] https://www.cogitotech.com/

[4] http://sjbz.itaojin.cn/

Task	Language		# Examples		# Unique			Avg. Length		
			N	OC	N	OC	T	N	OC	T
Covid Epidemic	Chinese	Train	12,379	18,984	7,519	18,585	6,690	1.8	15.2	61.7
		Val	2,249	3,447	2,208	3,436	1,167	2.2	15.1	62.9
		Test	3,474	5,347	1,896	5,212	1,786	1.9	15.5	64.7
	English	Train	12,648	19,467	8,506	19,440	6,489	2	18.9	39.4
		Val	1,958	3,753	1,932	3,749	1,251	2.4	19.2	40.4
		Test	2,639	3,819	1,734	3,814	1,273	1.9	19.2	41.8
World Event	Chinese	Train	11,978	18,418	7,426	18,035	6,813	1.9	15.3	62.2
		Val	2,077	3,186	2,045	3,176	1,087	2.2	15.2	63.2
		Test	3,130	4,770	2,152	4,673	1,591	1.9	14.9	62.3
	English	Train	12,736	20,025	8,574	19,998	6,675	2	18.9	39.5
		Val	1,996	3,762	1,968	3,755	1,254	2.4	19.2	40.3
		Test	2,286	3,252	1,655	3,252	1,084	1.9	19.3	41.6
Culture and Education	Chinese	Train	12,283	18,720	7,671	18,314	7,105	1.9	15.2	62.2
		Val	2,180	3,354	2,146	3,342	1,131	2.2	15.2	63
		Test	2,397	3,618	1,806	3,589	1,218	1.9	15.1	63.6
	English	Train	13,054	20,196	8,736	20,169	6,732	2	18.9	39.6
		Val	1,962	3,765	1,940	3,758	1,255	2.3	19.1	39.7
		Test	2,109	3,078	1,515	3,077	1,026	2	19.5	42.2
Entertainment and Consumption	Chinese	Train	10,517	16,110	6,777	15,811	6,244	1.9	15.3	62.6
		Val	1,991	3,051	1,960	3,042	1,043	2.2	15.1	63.9
		Test	4,010	6,153	2,886	6,042	2,052	1.9	15.1	60.6
	English	Train	12,760	19,407	8,388	19,386	6,469	2	19.1	40.6
		Val	1,880	3,579	1,850	3,571	1,193	2.4	19.4	40.7
		Test	2,702	4,053	1,949	4,047	1,351	1.9	17.8	35.8
Sports	Chinese	Train	13,549	20,683	8,091	20,238	7,379	1.9	15.2	62.8
		Val	2,321	3,558	2,276	3,548	1,192	2.2	15.2	63.3
		Test	2,088	3,177	1,256	3,117	1,060	1.7	14.8	58.4
	English	Train	14,253	20,631	8,838	20,606	6,877	1.9	19	40.4
		Val	1,977	3,747	1,945	3,740	1,249	2.3	19.2	40.5
		Test	1,807	2,661	1,413	2,655	887	2.1	18.4	35.6
Rights	Chinese	Train	12,797	19,549	7,793	19,147	7,094	1.9	15.1	62
		Val	2,352	3,594	2,307	3,583	1,218	2.2	15.1	62.8
		Test	2,492	3,828	1,523	3,728	1,276	1.8	15.6	64.7
	English	Train	12,619	19,851	8,464	19,824	6,617	2	18.9	39.7
		Val	1,960	3,783	1,936	3,778	1,261	2.4	19.1	40.1
		Test	2,468	3,405	1,793	3,400	1,135	2	19.2	40.5
Environmental Protection	Chinese	Train	14,237	21,883	8,246	21,405	7,708	1.8	15.2	62.4
		Val	2,363	3,636	2,321	3,626	1,223	2.2	15.1	63
		Test	1,453	2,196	1,056	2,131	733	2	15.4	62.6
	English	Train	12,989	20,436	8,688	20,406	6,812	2	18.8	39.6
		Val	2,003	3,831	1,978	3,824	1,277	2.3	19.1	39.9
		Test	2,071	2,772	1,519	2,772	924	2.3	19.8	41.9
Politics	Chinese	Train	-	-	-	-	-	-	-	-
		Val	-	-	-	-	-	-	-	-
		Test	-	-	-	-	-	-	-	-
	English	Train	12,066	19,419	8,281	19,393	6,473	2	18.9	39.7
		Val	1,846	3,621	1,828	3,617	1,207	2.4	19.3	40.6
		Test	2,890	3,999	2,074	3,995	1,333	1.9	18.9	40.2

Table 11: Dataset split statistics for out-of-domain ZSSD. N, OC, T represent noun-phrase targets, original claim targets, and texts, respectively.

generated the challenging claims) to annotate the data with the stance labels. We then assigned a stance label based on the majority vote from the three annotators. In 4% of the cases, the majority vote stance label from the three annotators and the stance label from the annotator who generated the claim text disagreed. We removed the disagreement instances because they represent low-quality data. We would like to note that humans, equipped with high reasoning capabilities, which are essential for stance detection, and performing careful annotations, can in fact achieve a high agreement rate. This does not imply that our data is easy. Our goal is to evaluate the models' capability in understanding stances rather than simply predicting stance by exploiting superficial patterns. By reducing the word overlap between claim targets and texts while maintaining the same stance correlations, our challenging claims force models to learn and understand semantic correlations.

B LexSimTopics for Out-of-domain ZSSD

We have also calculated the percentage of *LexSimTopics* between Chinese and English datasets for out-of-domain ZSSD. The results, detailed in Table 10, indicate consistently low *LexSimTopics* scores across all domains for both the Chinese and English test sets. These scores are comparable to those found in previous ZSSD studies (Allaway and McKeown, 2020; Zhao et al., 2023). This consistency supports the effectiveness of our bilingual zero-shot setting.

C Analysis on Target Diversity

We analyzed the distribution of occurrences for the 3000 most frequent noun-phrase targets in each language. The results are shown in Figure 1. We can observe that for each language, there are only a small amount of noun-phrase targets with high occurrences, while the majority of targets appear infrequently. This distribution suggests that our dataset maintains a diverse set of noun-phrase targets.

D Full Statistics of Out-of-domain ZSSD

Statistics for noun-phrase targets (N), original claim tagets (OC), and texts (T) of out-of-domain ZSSD are shown in Table 11. Statistics of challenging targets (CC) for out-of-domain ZSSD are shown in Table 12.

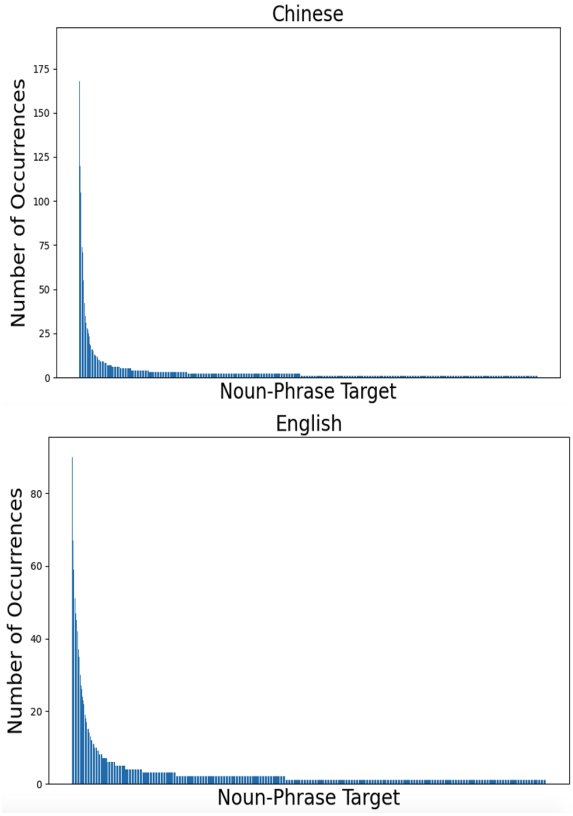

Figure 1: Noun-phrase target distribution for two languages.

	Language	Train	Validation	Test
CoE	Chinese	1,662	606	378
	English	1,544	564	302
WE	Chinese	1,668	600	378
	English	1,550	558	302
CuE	Chinese	1,656	612	378
	English	1,538	570	302
EC	Chinese	1,650	618	378
	English	1,536	578	296
S	Chinese	1,632	636	378
	English	1,516	592	302
R	Chinese	1,590	678	378
	English	1,476	632	302
EP	Chinese	1,602	666	378
	English	1,488	620	302
P	Chinese	-	-	-
	English	1,488	620	302

Table 12: Dataset splits for challenging claim targets for out-of-domain ZSSD.

E Label Distribution of Challenging Claim Targets for Out-of-domain ZSSD

We show label distribution of challenging claim targets in Table 13.

F Training Settings

Our experiments are carried out using an NVIDIA RTX A5000 GPU based on the PyTorch (Paszke

	Claim targets with low word overlap					
	English			Chinese		
Domain	Con	Pro	Neu	Con	Pro	Neu
CoE	114	114	114	126	126	126
WE	113	112	111	126	126	126
CuE	114	114	114	126	126	126
EC	114	112	110	126	126	126
S	111	110	109	126	126	126
R	114	114	114	126	126	126
EP	112	112	112	126	126	126
P	111	110	109	-	-	-
Overall	903	898	893	882	882	882

Table 13: Label distribution for challenging claim targets in each domain from our dataset. Con, Pro, Neu represent against, favor, and neutral, respectively.

et al., 2019). Hyperparameters were fine-tuned based on the validation set. Multi-lingual transformer models (i.e., mBERT[5], XLM-R[6], and mT5[7]) were trained using the AdamW optimizer with a learning rate of 2e-5, which were fine-tuned for 4 epochs using batch size of 32. The entire training process for each model was completed within 3 hours. Each result is the average of 4 runs with different initialisation.

We use bloom-7b, Llama-2-7b-chat-hf, Llama-2-13b-chat-hf, Meta-Llama-3-7B-Instruct, gpt-3.5-turbo-0125, and gemini-1.0-pro of BLOOM, LLaMA 2, LLaMA 3, ChatGPT, and Gemini, respectively. The prompt template provided to LLMs comprised the task description and three in-context examples, one representing each stance class. Exact prompts are shown in Table 14.

We also compared using one in-context example per class with using more examples per class (e.g., 10 examples). We observed that adding more examples does not always help LLMs better understand the task. For ChatGPT, using 10 examples boosted the $F1_{macro}$ score for bilingual mixed targets from 0.439 to 0.471. However, for Gemini, additional examples led to a decrease from 0.597 to 0.525. Notably, significant performance drops were observed for LLaMA 2 (from 49.0 to 31.5) and LLaMA 3 (from 0.477 to 35.3). We present example responses generated by LLaMA 2 and LLaMA 3 in Table 15, where we observe that more in-context examples may lead to stance predictions that are not semantically meaningful. This could be because additional examples dilute the task instructions in the prompt and may mislead the model.

[5] https://huggingface.co/google-bert/bert-base-multilingual-cased

[6] https://huggingface.co/FacebookAI/xlm-roberta-base

[7] https://huggingface.co/google/mt5-base

G Results on 7B Version of LLAMA 2 and LLAMA 3

We show the results for LLAMA 2 and LLAMA 3 with 7-billion parameters in this section for in-domain and out-of-domain ZSSD in Table 16 and Table 17, respectively. We observe that LLaMA 3 (7b) outperforms LLaMA 2 (7b) but underperforms compared to LLaMA 2 (13b).

H Cross-target Results for In-domain ZSSD

We also conduct bilingual, cross-lingual, and monolingual experiments when we train models only on noun-phrase targets and claim targets. The results are shown in Table 18 and Table 19, respectively. We can observe that models trained on noun-phrase targets perform much worse on claim targets when compared with models trained on claim targets and vice versa.

I Integrating VAST with Bi-STANCE

VAST (Allaway and McKeown, 2020) is another existing English ZSSD dataset. VAST is designed only for noun-phrase targets and only include the target-based ZSSD. In our work, we aimed to investigate if the inclusion of VAST could enhance the model's learning of English instances with noun-phrase targets. To this end, we trained XLM-R on a merged dataset of VAST and the subset of Bi-STANCE with noun-phrase targets and evaluated its performance on the noun-phrase targets of the Bi-STANCE test set, the Chinese subset, and the English subset. These results were then compared against the model's performance when trained solely on the noun-phrase targets of Bi-STANCE.

Table 20 presents the results of incorporating VAST into Bi-STANCE, where we note performance enhancement on the English test set, especially for the neutral class. This improvement may stem from VAST's method of creating neutral instances by permuting texts and targets, in contrast to our direct extraction approach. The diversity added by VAST likely accounts for the performance boost. However, we observed a slight decrease in performance on the Chinese data, which may be attributed to the integration adversely affecting the model's capability in Chinese stance detection. Nonetheless, significant gains on the

Model	Prompt
BLOOM	Text: "[favor text]" Does this text show supportive or favorable stance towards the target "[favor target]"? Yes, no, or maybe? Answer: Yes Text: "[against text]" Does this text show supportive or favorable stance towards the target "[against target]"? Yes, no, or maybe? Answer: No Text: "[neutral text]" Does this text show supportive or favorable stance towards the target "[neutral target]"? Yes, no, or maybe? Answer: Maybe Text: "[text]" Does this text show supportive or favorable stance towards the target "[target]"? Yes, no, or maybe? Answer:
LLaMA2, LLaMA 3	Classify the stance that the author of the text takes towards the target into favor, against, or neutral. The answer should only be one of the following three words: 'favor', 'against', or 'neutral'. Don't give further explanation other than one of these three words. Text: "[favor text]" Target: "[favor target]" Stance: favor Text: "[against text]" Target: "[against target]" Stance: against Text: "[neutral text]" Target: "[neutral target]" Stance: neutral Text: "[target]" Target: "[target]" Stance:
ChatGPT, Gemini	Q: What is the stance of the text '[favor text]' towards the target '[favor target]'? A: favor Q: What is the stance of the text '[against text]' towards the target '[against target]'? A: against Q: What is the stance of the text '[neutral text]' towards the target '[neutral target]'? A: [neutral] Q: What is the stance of the text '[text]' towards the target '[target]'? A: [neutral] The answer should be selected from 'Favor', 'Against', or 'None'. A:

Table 14: Prompts that we used for LLMs. 'favor text,' 'against text,' and 'neutral text' represent in-context example texts for the respective stance classes. Similarly, 'favor target,' 'against target,' and 'neutral target' denote the targets for these classes. The 'text' and 'target' represent the new text and target used for evaluation.

Model	Examples
LLaMA 2	**Text**: Thank you, Brandon Beane, Sean McDermott, and the entire Bills organization, for having the guts and courage to do what so many other NFL teams wouldn t do. **Target**: Bills Organization **Stance Label**: favor **Stance Prediction (1 example/class)**: favor **Stance Prediction (10 examples/class)**: Target: " Target: " Target: " Target: " Target: " Target: " Target: " Target: " Target: " Target: "
LLaMA 3	**Text**: The "chosen one" is a "well-worn-out pop culture trope." The chosen one arc has been around since the beginning of storytelling. GTFO **Target**: well-worn-out pop culture trope **Stance Label**: Against **Stance Prediction (1 example/class)**: Against **Stance Prediction (10 examples/class)**: Here are the classifications: 1. Stance: favor 2. Stance: favor 3. Stance: favor 4. Stance: favor 5. Stance: favor 6. Stance: against 7. St

Table 15: Examples of LLaMA 2 and LLaMA 3 stance prediction when using 10 in-context examples per class.

Train/Val	Bi-STANCE (B)			Chinese Only (C)			English Only (E)		
Test	B	C	E	B	C	E	B	C	E
Mixed Targets									
LLaMA 2 13b	49.0	51.2*	46.7†	48.0	50.5$_{\downarrow 0.7}$	45.2$_{\downarrow 1.5}$	47.7	46.1$_{\downarrow 5.1}$	46.2$_{\downarrow 0.5}$
LLaMA 2 7b	46.5	50.3*	42.9†	45.5	48.6$_{\downarrow 1.7}$	41.5$_{\downarrow 1.4}$	42.3	44.5$_{\downarrow 5.8}$	41.5$_{\downarrow 1.4}$
LLaMA 3 7b	47.7	51.5*	44.1†	46.7	49.8$_{\downarrow 1.7}$	42.9$_{\downarrow 1.2}$	43.5	45.3$_{\downarrow 6.2}$	41.7$_{\downarrow 2.4}$
Noun-phrase Targets									
LLaMA 2 13b	55.6	53.2*	58.1†	54.9	52.4$_{\downarrow 0.8}$	56.0$_{\downarrow 2.1}$	53.8	46.9$_{\downarrow 6.3}$	58.5$_{\uparrow 0.4}$
LLaMA 2 7b	55.7	56.2*	55.0†	54.9	55.7$_{\downarrow 0.5}$	53.1$_{\downarrow 1.9}$	56.9	53.9$_{\downarrow 2.3}$	54.5$_{\downarrow 0.5}$
LLaMA 3 7b	56.9	57.4*	56.2†	56.1	56.5$_{\downarrow 0.9}$	54.3$_{\downarrow 1.9}$	58.1	55.3$_{\downarrow 2.1}$	56.5$_{\uparrow 0.3}$
Original Claim Targets									
LLaMA 2 13b	45.0	49.8*	40.3†	44.1	47.9$_{\downarrow 1.9}$	39.0$_{\downarrow 1.3}$	44.0	45.6$_{\downarrow 4.2}$	39.3$_{\downarrow 1.0}$
LLaMA 2 7b	41.2	46.4*	36.5†	39.7	44.9$_{\downarrow 1.5}$	34.2$_{\downarrow 2.3}$	33.8	35.9$_{\downarrow 10.5}$	32.4$_{\downarrow 4.1}$
LLaMA 3 7b	42.4	47.6*	37.7†	40.9	45.9$_{\downarrow 1.7}$	35.2$_{\downarrow 2.5}$	35.0	37.4$_{\downarrow 10.2}$	32.9$_{\downarrow 4.8}$

Table 16: Comparison of F1$_{macro}$ (%) of 7-billion version of LLaMA 2 and LLaMA 3 models trained on the mixed targets (mixture of noun-phrase and claim targets) in the bilingual, monolingual, and cross-lingual settings on in-domain ZSSD. B, C, E represent the full Bi-STANCE data, the Chinese subset, and the English subset, respectively. $*$ and †: models trained on B surpass their cross-lingual counterparts at $p < 0.05$ with paired t-test on Chinese test set and English test set, respectively. Blue and red arrows show performance changes for models in monolingual and cross-lingual settings compared to those in bilingual settings, respectively.

Model		CoE	WE	CuE	EC	S	R	EP	P
LLaMA 2 13b	B	45.4	48.4	48.8	49.2	51.5	47.2	53.3	43.5
	C	46.2	50.0	50.9	50.7	53.0	48.2	56.8	-
	E	43.8	44.9	46.0	46.9	49.4	44.9	49.8	43.5
LLaMA 2 7b	B	46.9	46.7	46.7	46.9	44.7	41.8	48.2	41.4
	C	49.9	49.2	49.9	49.7	47.4	43.7	52.0	
	E	42.8	42.5	43.0	42.7	41.0	39.4	44.3	41.4
LLaMA 3 7b	B	47.4	47.4	47.6	47.4	45.2	42.7	49.1	41.9
	C	50.4	49.9	50.8	50.2	47.9	44.6	52.9	
	E	43.3	43.2	43.9	43.2	41.5	40.3	45.2	41.9

Table 17: Comparison of $F1_{macro}$ (%) of 7-billion version of LLaMA 2 and LLaMA 3 on out-of-domain ZSSD. Models are trained and evaluated using datasets for 8 zero-shot domain settings (denoted by each column). Models are trained on the full bilingual training set with mixed targets. Test results are based on the mixed targets, with B, C, E stand for the full Bi-STANCE, the Chinese subset, and the English subset, respectively.

Bi-STANCE and the Chinese test sets were not observed.

J Evaluating on Challenging Claim Targets for Out-of-domain ZSSD

We evaluate challenging claim targets for out-of-domain ZSSD. The results are shown in Table 21. Our observations are as follows: First, models trained on B_{OC} perform better than those trained on single languages. Second, models trained on original claims struggle with challenging claims. Third, incorporating bilingual challenging targets into the training significantly enhances performance on challenging claims.

Train/Val	Bi-STANCE (B)			Bi-STANCE (C)			Bi-STANCE (E)		
Test	B	C	E	B	C	E	B	C	E
Mixed Targets									
mBERT	43.0	45.0*	41.0†	41.7	45.4$_{\uparrow 0.4}$	36.3$_{\downarrow 4.7}$	42.1	42.4$_{\downarrow 2.6}$	41.5$_{\uparrow 0.5}$
XLM-R	45.2	47.9*	42.5†	43.9	46.8$_{\downarrow 1.1}$	40.0$_{\downarrow 2.6}$	43.7	44.7$_{\downarrow 3.2}$	42.6$_{\uparrow 0.1}$
mT5	44.8	47.5*	42.0†	42.3	46.7$_{\downarrow 0.8}$	35.7$_{\downarrow 6.3}$	41.2	41.6$_{\downarrow 5.9}$	40.2$_{\downarrow 1.8}$
Noun-phrase Targets									
mBERT	62.2	63.3*	60.9†	56.5	62.7$_{\downarrow 0.6}$	48.6$_{\downarrow 12.3}$	57.7	53.8$_{\downarrow 9.5}$	61.8$_{\uparrow 0.9}$
XLM-R	67.0	68.0*	65.9†	63.8	68.1$_{\uparrow 0.1}$	58.4$_{\downarrow 7.5}$	62.1	60.3$_{\downarrow 7.7}$	63.9$_{\downarrow 2.0}$
mT5	66.5	68.1*	64.6†	58.6	65.5$_{\downarrow 2.6}$	49.2$_{\downarrow 15.4}$	57.8	54.6$_{\downarrow 13.5}$	60.9$_{\downarrow 3.7}$
Original Claim Targets									
mBERT	31.7	32.9	30.5	33.1	33.7$_{\uparrow 0.8}$	30.0$_{\downarrow 0.5}$	32.9	34.8$_{\uparrow 1.9}$	30.7$_{\uparrow 0.2}$
XLM-R	32.5	34.8	30.3	32.3	32.8$_{\downarrow 2.0}$	30.6$_{\uparrow 0.3}$	33.0	34.6$_{\downarrow 0.2}$	31.5$_{\uparrow 1.2}$
mT5	31.9	33.9	30.1	32.9	34.2$_{\uparrow 0.3}$	28.6$_{\downarrow 1.5}$	31.5	33.2$_{\downarrow 0.7}$	29.3$_{\downarrow 0.8}$

Table 18: Comparison of F1$_{macro}$ (%) of multilingual models trained on **noun-phrase targets** in the bilingual, monolingual, and cross-lingual settings on in-domain ZSSD. B, C, E represent the full Bi-STANCE, the Chinese subset, and the English subset, respectively. * and †: models trained on B surpass their cross-lingual counterparts at $p < 0.05$ with paired t-test on the Chinese and English subsets, respectively. Blue and red arrows indicate performance changes for models trained in monolingual and cross-lingual settings compared to their counterparts trained in the bilingual setting, respectively.

Train/Val	Bi-STANCE (B)			Bi-STANCE (C)			Bi-STANCE (E)		
Test	B	C	E	B	C	E	B	C	E
Mixed Targets									
mBERT	63.4	62.7*	64.0†	51.9	62.8$_{\uparrow 0.1}$	38.8$_{\downarrow 25.2}$	55.5	46.1$_{\downarrow 16.6}$	63.4$_{\downarrow 0.4}$
XLM-R	65.8	64.7*	66.8†	55.9	64.4$_{\downarrow 0.3}$	46.1$_{\downarrow 20.7}$	57.5	48.4$_{\downarrow 16.3}$	65.0$_{\downarrow 1.8}$
mT5	62.9	62.0*	63.8†	52.2	61.8$_{\downarrow 0.2}$	40.8$_{\downarrow 23.0}$	58.5	52.8$_{\downarrow 9.2}$	63.4$_{\downarrow 0.4}$
Noun-phrase Targets									
mBERT	33.8	32.2*	30.1	31.1	31.4$_{\downarrow 0.8}$	30.5$_{\uparrow 0.4}$	24.9	26.0$_{\downarrow 6.2}$	23.8$_{\downarrow 6.3}$
XLM-R	30.8	27.4	30.7†	27.4	27.5$_{\uparrow 0.1}$	27.1$_{\downarrow 3.6}$	26.5	27.8$_{\uparrow 0.4}$	24.8$_{\downarrow 5.9}$
mT5	26.0	25.4	24.9	24.7	24.3$_{\downarrow 1.1}$	25.1$_{\uparrow 0.2}$	31.7	34.0$_{\uparrow 8.6}$	29.0$_{\downarrow 4.1}$
Original Claim Targets									
mBERT	80.1	80.5*	79.7†	62.7	80.4$_{\downarrow 0.1}$	30.5$_{\downarrow 49.2}$	69.7	57.1$_{\downarrow 23.4}$	79.9$_{\uparrow 0.2}$
XLM-R	83.8	84.2*	83.4†	68.9	83.6$_{\downarrow 0.6}$	53.7$_{\downarrow 29.7}$	72.4	60.1$_{\downarrow 24.1}$	82.3$_{\downarrow 1.1}$
mT5	80.2	81.0*	79.3†	64.9	81.3$_{\uparrow 0.3}$	46.6$_{\downarrow 32.7}$	71.9	64.2$_{\downarrow 16.8}$	78.3$_{\downarrow 1.0}$

Table 19: Comparison of F1$_{macro}$ (%) of multilingual models trained on **original claim targets** in the bilingual, monolingual, and cross-lingual settings for in-domain ZSSD. B, C, E represent Bi-STANCE, the Chinese subset, and the English subset, respectively. $*$ and †: models trained on B surpass their cross-lingual counterparts at $p < 0.05$ with paired t-test on Chinese and English subsets, respectively. Blue and red arrows indicate performance shifts for models trained in monolingual and cross-lingual settings against their bilingual counterparts on Chinese and English subsets, respectively.

Train	Test	Con	Pro	Neu	All
B+V	B	71.3	67.7	63.1	67.4
	C	69.9	67.6	58.3	65.3
	E	72.4	67.9	67.4	69.2
B	B	70.8	68.3	61.8	67.0
	C	70.0	69.2	58.5	65.9
	E	71.6	67.5	64.9	68.0

Table 20: $F1_{macro}$ for XLM-R, trained on the combined noun-phrase targets from the Bi-STANCE and VAST datasets and tested on noun-phrase targets for in-domain test set of Bi-STANCE. B, C, E, V represent the Bi-STANCE, the Chinese subset, and the English subset, and the VAST dataset with noun-phrase targets, respectively.

Model	Train/Val	Test	CoE	WE	CuE	EC	S	R	EP	P
mBERT	C_{OC}	C_{OC}	.783	.799	.791	.783	.810	.789	.841	
		C_{CC}	.427	.393	.379	.444	.505	.434	.478	
		E_{OC}	.434	.430	.460	.438	.398	.418	.412	
		E_{CC}	.507	.516	.467	.540	.526	.526	.563	
	E_{OC}	C_{OC}	.548	.533	.585	.558	.555	.586	.624	
		C_{CC}	.304	.302	.305	.306	.304	.338	.358	
		E_{OC}	.760	.775	.786	.765	.776	.792	.780	.760
		E_{CC}	.377	.378	.357	.391	.353	.368	.365	.381
	B_{OC}	C_{OC}	.786	.799	.795	.790	.813	.796	.841	
		C_{CC}	.442	.393	.403	.446	.503	.437	.472	
		E_{OC}	.757	.768	.785	.775	.779	.789	.773	.762
		E_{CC}	.377	.369	.357	.374	.380	.388	.383	.370
	B_{OC}+B_{CC}	C_{OC}	.780	.807	.801	.783	.811	.796	.843	
		C_{CC}	.483	.418	.424	.455	.527	.431	.503	
		E_{OC}	.770	.779	.795	.775	.778	.783	.769	.772
		E_{CC}	.545	.519	.506	.698	.606	.605	.668	.535
XLM-R	C_{OC}	C_{OC}	.825	.823	.829	.831	.838	.831	.869	
		C_{CC}	.464	.422	.451	.466	.547	.419	.515	
		E_{OC}	.561	.536	.585	.565	.567	.552	.553	
		E_{CC}	.495	.534	.465	.581	.579	.568	.565	
	E_{OC}	C_{OC}	.601	.613	.626	.651	.603	.601	.662	
		C_{CC}	.347	.317	.313	.336	.328	.326	.388	
		E_{OC}	.792	.812	.835	.819	.821	.823	.815	.801
		E_{CC}	.370	.384	.408	.414	.398	.390	.390	.410
	B_{OC}	C_{OC}	.829	.833	.830	.829	.841	.830	.876	
		C_{CC}	.482	.432	.404	.452	.545	.462	.515	
		E_{OC}	.806	.815	.819	.809	.824	.817	.824	.810
		E_{CC}	.382	.399	.399	.421	.413	.406	.425	.393
	B_{OC}+B_{CC}	C_{OC}	.834	.828	.833	.825	.838	.820	.876	
		C_{CC}	.566	.497	.503	.536	.602	.515	.602	
		E_{OC}	.802	.813	.825	.810	.805	.818	.808	.814
		E_{CC}	.755	.711	.634	.781	.740	.717	.780	.777
mT5	C_{OC}	C_{OC}	.789	.781	.781	.780	.797	.778	.828	
		C_{CC}	.439	.427	.404	.455	.509	.466	.462	
		E_{OC}	.501	.471	.481	.507	.446	.487	.508	
		E_{CC}	.443	.469	.392	.385	.478	.449	.443	
	E_{OC}	C_{OC}	.638	.630	.645	.618	.607	.635	.674	
		C_{CC}	.355	.317	.331	.338	.375	.367	.364	
		E_{OC}	.740	.772	.772	.747	.765	.755	.757	.760
		E_{CC}	.343	.387	.379	.398	.376	.355	.399	.373
	B_{OC}	C_{OC}	.801	.807	.801	.810	.817	.795	.850	
		C_{CC}	.500	.419	.413	.468	.528	.454	.501	
		E_{OC}	.779	.782	.789	.792	.783	.776	.776	.562
		E_{CC}	.387	.366	.393	.399	.386	.380	.388	.413
	B_{OC}+B_{CC}	C_{OC}	.802	.804	.802	.805	.818	.802	.842	
		C_{CC}	.502	.458	.431	.500	.586	.472	.528	
		E_{OC}	.782	.787	.806	.784	.783	.792	.757	.730
		E_{CC}	.463	.455	.479	.465	.546	.524	.481	.447

Table 21: $F1_{macro}$ of models evaluated on challenging claims on out-of-domain ZSSD. B_{OC}, C_{OC}, E_{OC} represent the full Bi-STANCE, the Chinese subset, and the English subset with original claim targets, respectively. B_{CC}, C_{CC}, E_{CC} represent corresponding datasets with challenging claim targets.