

A Reinforcement Learning Framework for Cross-Lingual Stance Detection Using Chain-of-Thought Alignment

Binghui Li, Minghui Zou, Xiaowang Zhang*, Shizhan Chen, Zhiyong Feng

College of Intelligence and Computing, Tianjin University, Tianjin, China

{binghuili, xiaowangzhang}@tju.edu.cn

Abstract

Cross-lingual stance detection identifies users' attitudes toward specific targets in texts by transferring knowledge from source languages to target languages. Previous studies have typically facilitated this transfer by translating and aligning labels or targets. However, these methods cannot effectively perform cross-lingual transfer of the complex reasoning processes in stance detection. To address this challenge, we propose a reinforcement learning framework using cross-lingual Chain-of-Thought (CoT) alignment, referred to as RCCA. Specifically, we adopt a cross-lingual CoT alignment strategy to obtain the high-quality CoTs generated from target language inputs. After that, we leverage reinforcement learning by sampling CoTs and assigning rewards according to pre-defined rules, aiming to enhance the model's generalization capabilities in the target language. Experimental results on four multilingual datasets demonstrate that our approach outperforms competitive methods.

1 Introduction

Stance detection automatically identifies users' attitudes (e.g., "in favor of" or "against") toward specific targets, which is crucial for text mining and social media analytics (Küçük and Can, 2020). Compared to the widely available English corpora used in research (Li et al., 2023; Lan et al., 2024; Weinzierl and Harabagiu, 2024), most other languages often lack sufficient training data. To address the issue of imbalanced data resources among languages and support the application of stance detection in low-resource languages, a cross-lingual stance detection task is proposed to transfer stance detection capabilities from source language to target language (Küçük and Can, 2020). Figure 1 presents an example of cross-lingual stance detection in which the model transfers knowledge from an English dataset to a French dataset.

*Corresponding author

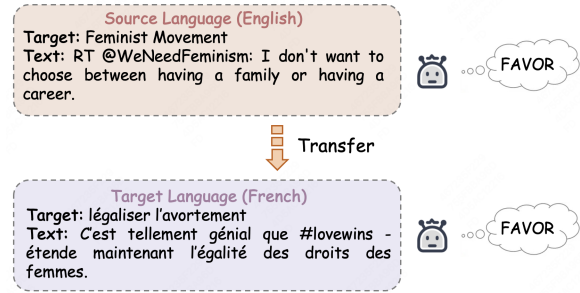


Figure 1: An example of cross-lingual stance detection. English is the source language and French is the target language.

Previous studies have typically facilitated cross-lingual adaptation through several approaches: (1) Label-based Alignment: Mohtarami et al. (2019) introduce a contrastive stance alignment-based adaptation method, aligning class labels across languages. (2) Target-based Alignment: Zhang et al. (2023b) resolve the target inconsistency problem by constructing intra-lingual and cross-lingual target relationship graphs and designing a target alignment strategy. (3) Integrated Label and Target Alignment: Zhang et al. (2023c) proposes a dual distillation framework for cross-lingual and cross-target scenarios. The cross-lingual teacher model is trained using prompt tuning and label alignment methods, while the cross-target teacher model bridge the gap of target inconsistency through target representation learning and contrastive learning.

However, these methods cannot effectively perform cross-lingual transfer of the complex reasoning processes in stance detection. On the one hand, the lack of sufficient training data in target languages makes it difficult for the model to generate accurate reasoning. On the other hand, the complex nature of stance detection reasoning, along with the differences in expression patterns and socio-cultural backgrounds among different languages, prevents the reasoning ability acquired in

the source language from being effectively generalized in the target language.

To overcome this challenge, we propose a reinforcement learning framework using cross-lingual CoT (Wei et al., 2022) alignment with large language models (LLMs). In the cross-lingual CoT alignment phase, we use well-designed prompts for the stance detection task to guide the model in generating clear reasoning processes. We employ a CoT alignment strategy to obtain high-quality CoTs from target language inputs, effectively alleviating the issue of insufficient training data for target languages. In the reinforcement learning phase, we first use the obtained CoTs for initial supervised fine-tuning (SFT) to equip the model with basic stance detection capabilities. Subsequently, through reinforcement learning, we enable the model to better adapt to the unique expressive characteristics of the target language, maintaining reasoning abilities consistent with the source language while enhancing generalization capabilities in the target language.

Our main contributions are summarized as follows: (1) We propose a reinforcement learning framework using cross-lingual CoT alignment, which effectively enables cross-lingual transfer of complex reasoning processes in stance detection tasks. (2) The cross-lingual CoT alignment strategy addresses the issue of scarce high-quality training data in the target language, while reinforcement learning enhances the model’s generalization ability for reasoning stances in cross-lingual scenarios. (3) We conduct experiments on four multilingual datasets, and the results show that RCCA outperforms all other methods in terms of F_{avg} .

2 Method

Figure 2 illustrates the overall structure of the proposed cross-lingual stance detection method, RCCA. In the cross-lingual CoT alignment phase, we designed prompts tailored to the stance detection task, leveraging the CoTs generated by the model in the source language to guide the generation of correct CoTs in the target language. In the reinforcement learning phase, we enhanced the model’s reasoning ability in target language scenarios through CoT sampling and a rule-based reward mechanism.

We denote the set of training data in the source language as $D_{\text{src}} = \{(t_i, x_i), y_i\}_{i=1}^N$, where t_i , x_i and y_i represent the target, text, and stance label

for each instance, respectively, with labels being *FAVOR*, *AGAINST*, or *NEUTRAL*. Similarly, we denote the **translated** training data in the target language as $D_{\text{tgt}} = \{((t'_i, x'_i), y_i)\}_{i=1}^N$, where t'_i and x'_i are the translations of t_i and x_i .

2.1 Cross-Lingual CoT Alignment

CoTs generation from D_{tgt} We first generate the associated CoT for each translated language sample (t'_i, x'_i) . To achieve this, we design a prompt, referred to as *Prompt1*, tailored to the characteristics of the stance detection task. For each target-text pair (t'_i, x'_i) , we implement specific steps to generate a high-quality CoT $c'_i = a'_i \oplus y'_i$, where a'_i represents the analysis component and y'_i denotes the stance. Given that English-centric LLMs often perform better with English prompts (Lai et al., 2023), we choose to use the **English prompt** to generate the CoTs. The *Prompt1* is as follows:

Task: Determine the stance of <text> towards <target>, following a clear and structured process.

[Analysis]:

1. *Read the Comment: Understand the context and sentiment.*
2. *Identify Key Elements: Highlight terms or phrases that suggest support, opposition, or neutrality.*
3. *Analyze Implications: Explain how these elements reflect the author’s viewpoint, considering any relevant references.*
4. *Determine Stance: Decide if the text is "FAVOR," "AGAINST," or "NEUTRAL."*

[Stance]: "FAVOR," "AGAINST," or "NEUTRAL."

After collecting the CoTs c'_i , we divide the data $\{((t'_i, x'_i), c'_i), 1 \leq i \leq N\}$ into two subsets according to whether the model generates the correct stance. More precisely, abbreviating $R = \{1 \leq i \leq N : y'_i = y_i\}$ and $W = \{1 \leq i \leq N : y'_i \neq y_i\}$, we define

$$C_{\text{right}}^{\text{tgt}} = \{((t'_i, x'_i), c'_i), i \in R\},$$

$$C_{\text{wrong}}^{\text{tgt}} = \{((t'_i, x'_i), c'_i), i \in W\}.$$

We refer to these as the correct and incorrect CoT datasets, comprising CoTs generated from target language inputs.

Correct CoTs Generation from D_{src} In this part, for each $i \in W$, we select the corresponding source language pair (t_i, x_i) and generate an associated CoT c_i employing a multi-turn dialogue approach. Initially, we produce a CoT using *Prompt1* with t_i and x_i as input. If the resulting stance does not align with y_i , we proceed to the next dialogue turn using *Prompt2* to incorporate y_i and refine c_i . The

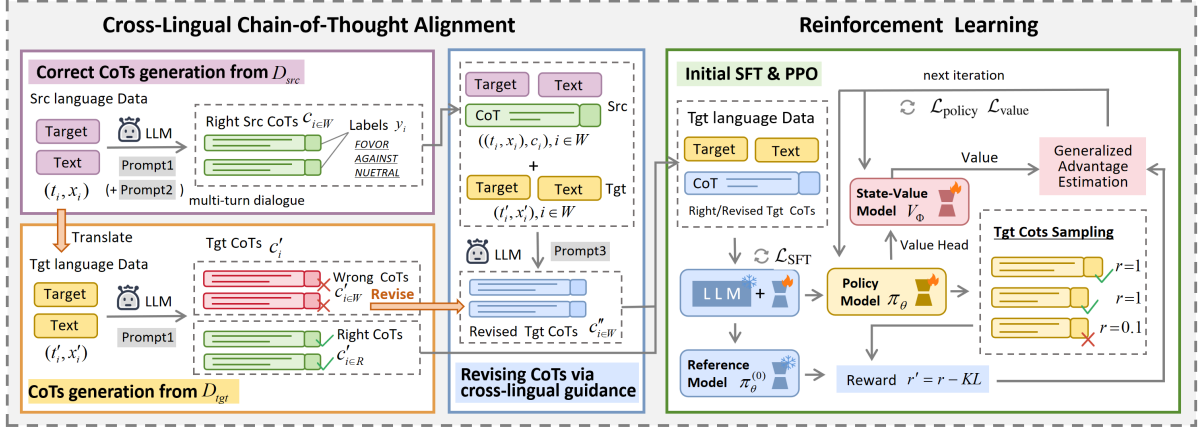


Figure 2: The overall architecture of the cross-lingual stance detection method RCCA. The framework adopts a cross-lingual CoT alignment strategy to generate high-quality CoTs. After the initial SFT, it implements a reinforcement learning approach, sampling CoTs and assigning rewards based on predefined rules.

process results in the refined dataset

$$C_{\text{right}}^{\text{src}} = \{((t_i, x_i), c_i), i \in W\}.$$

The *Prompt2* is as follows:

The correct stance is: $\langle \text{label} \rangle$. Please provide a detailed analysis of why this text is indeed in favor, against, or neutral towards the target, based on the correct stance.

Revising CoTs via cross-lingual guidance Since the model more frequently encounters source language data and because data translated into the target language may contain errors, the model typically performs better with source language data than with target language data. Based on this, we choose to use CoTs generated from source language data as guidance. Specially, for each target language pair (t'_i, x'_i) with associated $((t'_i, x'_i), c'_i) \in C_{i \in W}^{\text{tgt}}$, we guide the model using the source language example $((t_i, x_i), c_i) \in C_{i \in W}^{\text{src}}$ to generate a revised CoT c''_i . The guiding prompt, denoted as *Prompt3*, fills the slots for $t_i, x_i, c_i, t'_i, x'_i$, and the source and target languages, respectively. The revised dataset is denoted as

$$C_{\text{revised}}^{\text{tgt}} = \{((t'_i, x'_i), c''_i), i \in W\}.$$

The *Prompt3* is as follows:

Analyze and judge the stance of a given $\langle \text{tgt language} \rangle$ text by following the example of the $\langle \text{src language} \rangle$ text.
 Target: $\langle \text{tgt target} \rangle$; text: $\langle \text{tgt text} \rangle$
 [Example]:
 Target: $\langle \text{src target} \rangle$; text: $\langle \text{src text} \rangle$;
 Response: $\langle \text{src CoT} \rangle$

2.2 Reinforcement Learning

Initial SFT We begin with a fine-tuning of the policy model π_θ based on the correct dataset $C_{\text{right}}^{\text{tgt}}$ and revised dataset $C_{\text{revised}}^{\text{tgt}}$. Denote q'_i for the *Prompt1* and let L_i be its sequence length. The SFT loss function is taken as follows:

$$\begin{aligned} \mathcal{L}_{\text{SFT}}(\theta) = & -\frac{1}{|R|} \sum_{i \in R} \sum_{t=1}^{L_i} \log \pi_\theta(c'_{i,t} | q'_i, c'_{i,<t}) \\ & -\frac{1}{|W|} \sum_{i \in W} \sum_{t=1}^{L_i} \log \pi_\theta(c''_{i,t} | q'_i, c''_{i,<t}). \end{aligned}$$

Fine-tuning equips the model with basic stance detection skills. After that, we freeze the model's parameters to create a reference model $\pi_\theta^{(0)}$.

Proximal Policy Optimization (PPO) To optimize the reasoning process in the target language, we employ the PPO (Schulman et al., 2017) algorithm to train the policy model π_θ . Specifically, the policy model updates its parameters online by sampling responses, obtaining the corresponding rewards, and adjusting accordingly. Following Ziegler et al. (2019), the value model V_ϕ is constructed by appending a linear value head on top of the policy model π_θ . s_t represents the state at time-step t within the context of the sampled CoT. In the final state, we compare the extracted predicted stance with the true label y and assign rewards based on predefined rules (Luong et al., 2024). If no model answer is extracted, the reward is set to 0; if an answer is extracted and it is correct, the reward is set to 1; if the answer is incorrect, the reward is

set to 0.1. The reward function is as follows:

$$r(s_{t=L}) = \begin{cases} 1, \text{EXTRACT}(\text{stance}) = y \\ 0.1, \text{EXTRACT}(\text{stance}) \neq \text{null}, \neq y \\ 0, \text{EXTRACT}(\text{stance}) = \text{null} \end{cases}$$

Following Lu et al. (2023), the total reward is the sum of the reward function score and the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951) between the probability distributions of π_θ and $\pi_\theta^{(0)}$ weighted by β , the control coefficient. Precisely, we set the reward of each sample as:

$$r_{\text{total}} = r(s_{t=L}) - \beta \text{KL}(\pi_\theta(\cdot | s_t), \pi_\theta^{(0)}(\cdot | s_t)).$$

We use Generalized Advantage Estimation Schulman et al. (2015) to obtain advantages \hat{A}_t and returns \hat{R}_t . The detailed process is presented in the appendix A. The policy loss function and value loss function can be written as follows:

$$\begin{aligned} \mathcal{L}_{\text{policy}}(\theta) &= \mathbb{E}_t \left[\min \left(\rho_t(\theta) \hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t \right) \right], \\ \mathcal{L}_{\text{value}}(\phi) &= \frac{1}{2} \mathbb{E}_t \left[\max \left(\left\| \mathbf{V}_\phi(s_t) - \hat{R}_t \right\|^2, \right. \right. \\ &\quad \left. \left. \left\| \text{clip}(\hat{R}_t - \mathbf{V}_\phi(s_t), \hat{A}_t - \epsilon, \hat{A}_t + \epsilon) \right\|^2 \right) \right], \end{aligned}$$

where $\rho_t(\theta)$ is the probability ratio between the new policy π_θ and the old policy $\pi_{\theta_{\text{old}}}$, and ϵ is the clipping parameter. The overall loss function is a weighted sum of $\mathcal{L}_{\text{policy}}$ and $\mathcal{L}_{\text{value}}$, given by

$$\mathcal{L}_{\text{RL}}(\theta, \phi) = \mathcal{L}_{\text{policy}} + \alpha \mathcal{L}_{\text{value}},$$

with α serving as the weight parameter.

3 Experiments

3.1 Datasets and Baselines

We conduct experiments on four multilingual datasets. **Politics-Czech** and **Society-Czech** datasets, constructed by Zhang et al. (2023c) from X-stance (Vamvas and Sennrich, 2020) and Czech (Hercig et al., 2017), have German as the source language and Czech as the target language. **Independence** (Zotova et al., 2020) contains tweets in Spanish and Catalan about "Independence of Catalonia". Spanish is considered as the source language and Catalan is considered as the target language. **Election**, part of the Multilingual Political Dataset (Lai et al., 2020), contains tweets about US and French elections. We take English as the

source language and French as the target language. See Appendices B.1 for more dataset details.

We conduct experiments with two foundation models: **Qwen2.5-7B-Instruct**¹ and **Llama3-8B-Instruct**². More implementation details are provided in Appendices B.2. We compare RCCA with other methods in cross-lingual stance detection: ADAN (Chen et al., 2018), mBERT-FT (Devlin, 2018), CCSD (Zhang et al., 2023c), TaRA (Zhang et al., 2023b) and ATOM (Zhang et al., 2023a). For comparison, we also include CoT (Wei et al., 2022), which guides the model through the contextual analysis step-by-step, and SFT, which enhances performance through fine-tuning on training data.

3.2 Results

Following Zhang et al. (2023c,b,a), we use accuracy and the average F1 score of "Favor" and "Against" as the evaluation metrics. Table 1 presents the results of RCCA and comparative methods on the four datasets. Specifically, using Qwen2.5-7B-Instruct, RCCA achieves F_{avg} improvements of 2.82%, 3.39%, 12.8% and 5.62% over CoT, and 5.08%, 4.97%, 2.84% and 10.18% over SFT on the four datasets, respectively. Similarly, when using Llama3-8B-Instruct, RCCA also demonstrates remarkable performance. In addition, the performance of our method also outperforming methods such as CCSD, ATOM, and TaRA with F_{avg} . These results suggest that RCCA effectively utilizes source language data for transferring reasoning processes across languages. However, we observed that the accuracy of Llama3-8B-Instruct+RCCA on the Election dataset was not as good as that of SFT. We speculate that this might be due to the relatively small amount of data, which introduces some instability in the reinforcement learning process.

Table 2 presents the results of the ablation study. In the table, "- RL" denotes the experimental setup where only the CoT alignment strategy and initial SFT is retained while the reinforcement learning module is removed. Conversely, "- CoT A" represents the configuration where both reinforcement learning and CoT alignment components are removed, leaving only the baseline CoT method. The comparison reveals that removing either critical component leads to performance degradation across all four datasets. This outcome underscores

¹<https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>

²<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

Method	Politics-Czech (de→cs)		Society-Czech (de→cs)		Independence (es→ca)		Election (en→fr)	
	Acc	F _{avg}	Acc	F _{avg}	Acc	F _{avg}	Acc	F _{avg}
ADAN	52.78	52.70	52.80	50.89	51.34	48.94	63.23	49.82
mBERT-FT	52.78	52.74	51.39	50.36	48.86	47.59	69.06	52.74
CCSD	56.25	55.89	52.78	52.55	-	-	-	-
ATOM	-	-	-	-	52.54	52.99	60.99	56.50
TaRA	-	-	-	-	-	-	75.80	62.30
Qwen2.5-7B-Instruct								
CoT	55.48	55.24	53.47	54.24	43.23	40.38	71.06	71.69
SFT	53.46	52.98	53.02	52.66	48.17	50.34	80.78	66.53
RCCA	56.77	58.06	54.10	57.63	50.89	53.18	83.79	76.71
Llama3-8B-Instruct								
CoT	55.17	55.96	54.26	54.72	47.41	49.17	78.55	68.11
SFT	53.21	52.84	53.61	53.14	46.30	43.52	81.94	67.48
RCCA	56.13	58.71	54.93	56.70	52.96	53.44	79.32	75.40

Table 1: Experimental results of baselines and RCCA on four representative multilingual datasets. The best performances are marked in bold, and the gray shading indicates the results of our method.

	P-C	S-C	Inde	Elec
Qwen2.5-7B-Instruct				
RCCA	58.06	57.63	53.18	76.71
- RL	56.67	55.34	50.50	74.56
- CoT	55.24	54.24	40.38	71.69
Llama3-8B-Instruct				
RCCA	58.71	56.70	53.44	75.40
- RL	56.67	55.01	51.69	71.26
- CoT	55.96	54.72	49.17	68.11

Table 2: Results of ablation study on the four multilingual datasets, using F_{avg} as metrics.

the indispensable roles of both the CoT alignment strategy and the reinforcement learning module in achieving optimal performance within the RCCA framework.

4 Conclusion

In this work, we present a reinforcement learning framework using cross-lingual CoT alignment (RCCA), designed to effectively transfer the complex reasoning processes in cross-lingual stance detection. Experiments show that our method outperforms other methods in cross-lingual stance detection task. Furthermore, RCCA has strong potential for broader application to cross-lingual tasks beyond stance detection and can effectively address challenges in low-resource languages.

5 Limitation

In this study, we combine two mainstream LLMs, Qwen2.5-7B-Instruct and Llama3-8B-Instruct, with the framework we design, achieving promising results in the cross-lingual stance detection task. However, we also recognize that other LLMs or models with different parameter scales

(larger or smaller) have not yet been used for further generalization testing, which somewhat limits the universality of our findings.

During the research process, some datasets have relatively small sample sizes, and fine-tuning with CoTs easily leads to overfitting. Fortunately, LLMs originally perform poorly on target language data, and our proposed method helps improve the model’s performance. In future work, we plan to consider constructing larger training datasets and incorporating more diverse CoTs to mitigate potential overfitting and further optimize the model’s performance.

The PPO algorithm we employ exhibits some instability, and fine-tuning hyperparameters such as the KL control coefficient and PPO epochs for different datasets and models requires careful adjustment, which incurs additional time and computational costs.

Regarding the datasets, this study covers English, German, French, Spanish, Catalan, and Czech. In subsequent research, we plan to include more low-resource languages from different linguistic systems and explore the challenges and potential solutions in other cross-lingual scenarios. This will broaden the scope and depth of the research, enhancing the applicability and universality of the results.

References

Xilun Chen, Yu Sun, Ben Athiwaratkun, Claire Cardie, and Kilian Weinberger. 2018. Adversarial deep averaging networks for cross-lingual sentiment classification. *Transactions of the Association for Computational Linguistics*, 6:557–570.

- Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. 2017. Detecting stance in czech news commentaries. *ITAT*, 176:180.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Dilek Küçük and Fazli Can. 2020. Stance detection: A survey. *ACM Computing Surveys (CSUR)*, 53(1):1–37.
- Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Mirko Lai, Alessandra Teresa Cignarella, Delia Irazú Hernández Fariás, Cristina Bosco, Viviana Patti, and Paolo Rosso. 2020. Multilingual stance detection in social media political debates. *Computer Speech & Language*, 63:101075.
- Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.
- Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 891–903.
- Yingjie Li, Chenye Zhao, and Cornelia Caragea. 2023. Tts: A target-based teacher-student framework for zero-shot stance detection. In *Proceedings of the ACM Web Conference 2023*, pages 1500–1509.
- Xiuyuan Lu, Benjamin Van Roy, Vikranth Dwaracherla, Morteza Ibrahimi, Ian Osband, Zheng Wen, et al. 2023. Reinforcement learning, bit by bit. *Foundations and Trends® in Machine Learning*, 16(6):733–865.
- Trung Quoc Luong, Xinbo Zhang, Zhanming Jie, Peng Sun, Xiaoran Jin, and Hang Li. 2024. Reft: Reasoning with reinforced fine-tuning. *arXiv preprint arXiv:2401.08967*.
- Mitra Mohtarami, James Glass, and Preslav Nakov. 2019. Contrastive language adaptation for cross-lingual stance detection. *arXiv preprint arXiv:1910.02076*.
- John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. 2015. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
- Jannis Vamvas and Rico Sennrich. 2020. X-stance: A multilingual multi-target dataset for stance detection. *arXiv preprint arXiv:2003.08385*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880.
- Ruike Zhang, Nan Xu, Wenji Mao, and Daniel Zeng. 2023a. Adversarial topic-aware memory network for cross-lingual stance detection. In *2023 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.
- Ruike Zhang, Nan Xu, Hanxuan Yang, Yuan Tian, and Wenji Mao. 2023b. Target-oriented relation alignment for cross-lingual stance detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6391–6404.
- Ruike Zhang, Hanxuan Yang, and Wenji Mao. 2023c. Cross-lingual cross-target stance detection with dual knowledge distillation framework. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10804–10819.
- Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
- 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.

A Generalized Advantage Estimation

In this section we provide details on how we obtain the advantages \hat{A}_t and the returns \hat{R}_t using Generalized Advantage Estimation (GAE).

GAE combines concepts from Temporal Difference (TD) learning and Monte Carlo (MC) estimation to balance the trade-off between bias and variance. The advantage \hat{A}_t at time t , is calculated using the discounted sum of TD residuals as follows:

$$\hat{A}_t = \sum_{l=0}^{L-t} (\gamma\lambda)^l \delta_{t+l},$$

where γ is the discount factor for rewards, and λ is the trace - decay parameter, both typically set between 0 and 1.

The Temporal Difference residual δ_t is computed as:

$$\delta_t = r_t + \gamma V_\phi(s_{t+1}) - V_\phi(s_t),$$

where $V_\phi(s_t)$ is the value function at state s_t parameterized by ϕ , and r_t is the immediate reward received after taking action a_t in state s_t . The estimate of returns \hat{R}_t leverages the λ -return, which is a blend of the GAE and the value estimate:

$$\hat{R}_t = \hat{A}_t + V_\phi(s_t),$$

thus, the returns \hat{R}_t incorporate both the advantages computed via GAE and the current estimate of the state value. By incorporating these calculations, the GAE method provides a more stable foundation for the optimization of policy and value functions in reinforcement learning scenarios.

B Experiment Details

B.1 Datasets

X-stance (Vamvas and Sennrich, 2020) is a multilingual stance dataset on Swiss politics, where German is used as the source language and French is the target language. Each sample consists of a voter’s question and a candidate’s answer, and the stance can be classified into "favor" or "against". Following Zhang et al. (2023c), we construct two datasets from X-stance: (1) **Politics** is comprised of all the data in domains "Foreign Policy" and "Immigration", with 31 different targets in total. (2) **Society** consists of all the data in domains "Society" and "Security", with 32 different targets in total.

Czech (Hercig et al., 2017) contains commons in Czech with targets of "Miloš Zeman" and "Smoking Ban in Restaurants". Following Zhang et al.

(2023c), we construct two subsets from X-stance and Czech datasets: the **Politics-Czech** and the **Society-Czech** dataset, with German as the source language and Czech as the target language.

Independence (Zotova et al., 2020) contains tweets in Spanish and Catalan about "Independence of Catalonia". Spanish is considered as the source language and Catalan is considered as the target language.

Election, part of the Multilingual Political Dataset (Lai et al., 2020), contains tweets about US and French elections. Data about US Election is in English and contains targets of "Hillary Clinton" and "Donald Trump". French Election contains targets of "Macron" and "Le Pen". We take English as the source language and French as the target language. The statistics of the datasets are shown in Table 3.

B.2 Implementation Details

We conduct experiments with two foundation models: Qwen2.5-7B-Instruct and Llama3-8B-Instruct. To save computational resources, we apply LoRA adapter (Hu et al., 2021) to the LLM during both the initial SFT and PPO training stages, with the rank and the scaling factor set to 16 and 32. During the initial SFT, we use AdamW with 10% warm-up ratio, and the learning rate is $2e-5$. We train the model for 5-10 epochs and the maximum length is set to 512. During the PPO training phase, we set the PPO epoch (i.e., the number of updates for each batch of samples) to 2, conduct 5 epochs of training, and then select the optimal result. Following Ziegler et al. (2019), the discount factor γ , the trace-decay parameter λ , the coefficient for the value objective α and the clipping parameter ϵ are set to 0.95, 1, 5 and 2. The number of updates per RL step is set to 2. For Qwen2.5-7B-Instruct, the KL coefficient β on Politics-Czech, Society-Czech, Independence, Election datasets are set to 0.2, 0.2, 0.15 and 0.1. For Llama3-8B-Instruct, the KL coefficient β are set to 0.2, 0.2, 0.1 and 0.25.

When using the CoT baseline, we use the following prompts to guide the model to generate responses:

Please analyze whether the $\langle \text{text} \rangle$ of the comment supports, opposes, or remains neutral towards the $\langle \text{target} \rangle$.

Provide your answer in the following format:

[Analysis]: (detailed explanation of why the comment is in favor, against, or neutral)

[Stance]: ("FAVOR", "AGAINST", or "NEUTRAL")

Dataset	Language	Target	Favor	Against	Neutral	Total
X-stance-Politics	de	31 questions about "Foreign Policy" and "Immigration"	2713	3284	-	5997
X-stance-Society	de	31 questions about "Society" and "Security"	3121	3203	-	6324
Independence	es	"Catalan independence"	419	1807	3174	5400
	ca		3311	163	1926	5400
Election	en	"Hillary Clinton" and "Donald Trump"	462	713	516	1691
	fr	"Emmanuel Macron" and "Marine Le Pen"	156	774	186	1116
Czech	cs	"Miloš Zeman" and "Smoking Ban in Restaurants"	1435	2543	1445	5423

Table 3: The statistics of the five cross-lingual datasets.

For the SFT baseline, we use the source language data for training and the target language data for testing. The LoRA adapter is applied to the LLM with the rank and the scaling factor set to 16 and 32. We use AdamW with 10% warm-up ratio, and the learning rate is $2e-5$. The model is trained for 10 epochs, and the checkpoint with the best performance is chosen.

C Supplementary Experiments

Tables 4 and Figures 3 display the performance variations of the Qwen2.5-7B-Instruct and Llama3-8B-Instruct on the Politics-Czech and Society-Czech datasets as the PPO epoch (i.e., the number of updates per batch of samples) changes. Both models show better performance in terms of Accuracy and F_{avg} when the PPO epoch is set within the range of 1 to 3. However, when the PPO epoch exceeds 3, the performance of both models declines on these datasets, suggesting that too many updates per batch may reduce the models' generalization capabilities.

To provide a more comprehensive comparison, we contrasted RCCA against a direct translation baseline (a strategy that translates target language data into English first and then employs the CoT method for stance judgment). Table 5 shows that the Llama-3-8B+Trans+CoT baseline outperforms the RCCA method in accuracy on the Politics-Czech and Society-Czech datasets. However, the Trans+CoT baseline underperforms RCCA on other datasets or metrics. We hypothesize that for multilingual models proficient in English, translating non-English data into English might amplify translation errors due to the model's heightened sensitivity to English semantics. Conversely, translating English data into non-English languages

for training could lead to more significant improvements in the model's non-English capabilities, thereby mitigating the impact of translation errors.

Method	Politics-Czech		Society-Czech	
	Acc	F_{avg}	Acc	F_{avg}
Qwen2.5-7B-Instruct + RCCA				
PPO epoch = 1	55.47	57.31	54.74	56.22
PPO epoch = 2	56.77	58.06	54.10	57.63
PPO epoch = 3	56.47	58.18	54.01	55.85
PPO epoch = 4	54.64	56.25	53.27	55.64
PPO epoch = 5	53.55	56.20	52.81	55.49
Llama3-8B-Instruct + RCCA				
PPO epoch = 1	56.69	57.60	54.19	56.05
PPO epoch = 2	56.13	58.71	54.93	56.70
PPO epoch = 3	55.20	56.49	54.28	57.01
PPO epoch = 4	53.73	54.38	53.17	54.10
PPO epoch = 5	53.82	53.92	53.08	53.46

Table 4: Experimental results of performance fluctuations for the Qwen2.5-7B-Instruct+RCCA and Llama3-8B-Instruct+RCCA on the Politics-Czech and Society-Czech datasets across various PPO epochs.

D Examples

D.1 Examples of Across-lingual CoT Alignment

Tables 6 and ?? present a complete example of cross-lingual CoT alignment. For incorrect CoTs generated from target language data (shown in the upper half of Table ??), the correct CoTs generated from source language data are used for correction (shown in Table 6, where the target language data is translated from this source language data). Ultimately, the correct CoTs generated from the source language data serve as a model, guiding the model to generate accurate reasoning paths in the corresponding target language (as shown in the lower half of Table ??), thereby improving the quality of

Method	Politics-Czech		Society-Czech		Independence		Election	
	Acc	F _{avg}	Acc	F _{avg}	Acc	F _{avg}	Acc	F _{avg}
Qwen2.5-7B-Instruct								
CoT	55.48	55.24	53.47	54.24	43.23	40.38	71.06	71.69
Trans + CoT	52.46	51.92	51.53	50.73	41.65	40.70	60.53	64.51
RCCA	56.77	58.06	54.10	57.63	50.89	53.18	83.79	76.71
Llama3-8B-Instruct								
CoT	55.17	55.96	54.26	54.72	47.41	49.17	78.55	68.11
Trans + CoT	57.70	58.07	55.17	55.34	43.15	45.08	69.61	63.38
RCCA	56.13	58.71	54.93	56.70	52.96	53.44	79.32	75.40

Table 5: Experimental results of direct translation baseline and RCCA on four multilingual datasets.

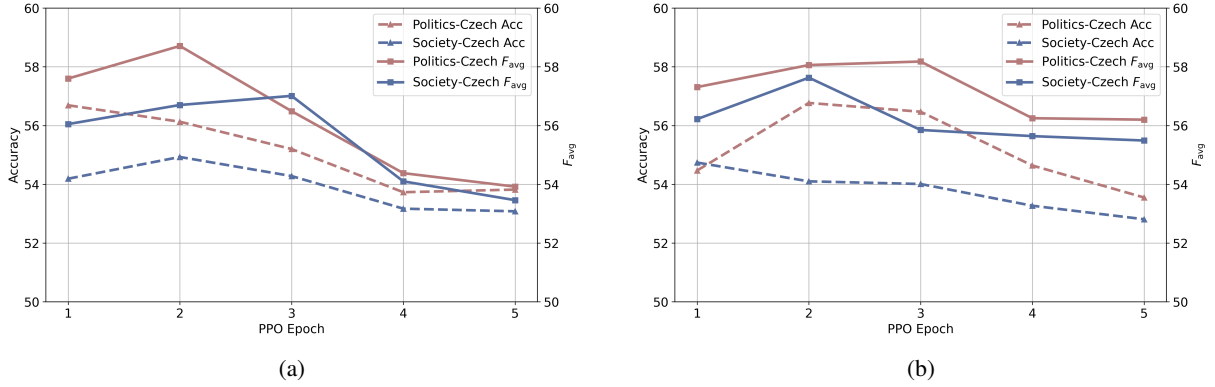


Figure 3: The performance fluctuations of the Qwen2.5-7B-Instruct+RCCA (a) and Llama3-8B-Instruct+RCCA (b) on the Politics-Czech and Society-Czech datasets over various PPO epochs, highlighting trends in both Accuracy and F_{avg} metrics.

the target language CoTs.

Since the model more frequently encounters source language data and because data translated into the target language may contain errors, the model typically performs better with source language data than with target language data. Based on this, we choose to use CoTs generated from source language data as guidance. To ensure the accuracy of the CoTs, if the model fails to generate the correct answer in the first round of dialogue, it will proceed to the next round and use the correct answer as a prompt. This ensures that the model generates accurate CoTs based on source language input, which then guides the input in the target language, as detailed in Table 6.

D.2 Examples of CoT Sampling and Reward

Table 8, 9, 10, 11 display the CoT sampling in reinforcement learning for stance detection tasks in Czech, Catalan, and French. The tables include the topics and texts requiring analysis, detailed reasoning processes of the model, and scores awarded for the correctness of the model’s responses. If the reasoning process is correct, a reward of 1.0 is given; if incorrect, a reward of 0.1 is given.

Correct CoTs generation from source language data

Input:

(Prompt1) Task: Determine the stance of the text towards the target, following a clear and structured process.

[Analysis]:

1. Read the Comment: Understand the context and sentiment.
2. Identify Key Elements: Highlight terms or phrases that suggest support, opposition, or neutrality.
3. Analyze Implications: Explain how these elements reflect the author's viewpoint, considering any relevant references.
4. Determine Stance: Decide if the text is "FAVOR," "AGAINST," or "NEUTRAL."

[Stance]: "FAVOR," "AGAINST," or "NEUTRAL."

Target: Donald Trump

Text: @QuiltingMuriel But I do agree about the excess speechifying and posturing and general nonsense.
#SemST

Response:

[Analysis]:

1. Read the Comment: The comment mentions "excess speechifying and posturing and general nonsense," which suggests dissatisfaction with political rhetoric or behavior. However, there is no direct mention of Donald Trump or a specific policy or action related to him.
2. Identify Key Elements: The key elements in the comment include "excess speechifying," "posturing," and "general nonsense." These terms are generally negative and criticize the nature of some political communications or behaviors. The use of these terms indicates a critical view of political discourse, but the lack of direct reference to Trump makes it ambiguous whether this criticism is aimed at him.
3. Analyze Implications: The phrases used imply a critique of political behavior characterized by excessive or meaningless speeches and postures, which could be applied to many political figures. Without explicit reference to Trump or his actions, it's difficult to conclusively link this criticism directly to him, suggesting that the comment could be about political figures more broadly.
4. Determine Stance: Given the absence of direct references to Trump or specific policies/actions of his, and the general nature of the critique, it's most prudent to classify this comment as NEUTRAL concerning Donald Trump specifically.

[Stance]: NEUTRAL ✗

(Next turn)

Input:

(Prompt2) The correct stance is: AGAINST. Please provide a detailed analysis of why this text is indeed in favor, against, or neutral towards the target, based on the correct stance.

Response:

[Analysis]:

1. Read the Comment: The comment appears to be a reply to someone named @QuiltingMuriel, agreeing with them about the excessive speechifying, posturing, and general nonsense.
2. Identify Key Elements: The key elements here are "excess speechifying and posturing and general nonsense." These phrases suggest criticism of political rhetoric and behavior, particularly from politicians who engage in these behaviors.
3. Analyze Implications: The comment indicates agreement with another person regarding the perceived overuse of rhetoric and posturing by politicians. Given the context, it is likely that this criticism is directed at Donald Trump, as he was known for his controversial and often inflammatory speeches and statements during his presidency.
4. Determine Stance: Since Donald Trump was frequently criticized for these aspects of his public persona and political style, it can be inferred that the comment is AGAINST Donald Trump.

[Stance]: AGAINST ✓

Table 6: An example of generating correct CoTs from source language inputs. The model typically performs better with source language data. If the model fails to generate the correct answer in the first round of dialogue, it will proceed to the next round and provide the correct answer as a prompt. This ensures that the model can generate correct CoTs based on the source language data to guide the input in the target language.

Incorrect CoTs generation from target language data

Input:*Prompt1***Target:** Donald Trump**Text:** @QuiltingMuriel Mais je suis d'accord avec les discours excessifs, les postures et les absurdités en général. #SemST (The French text is translated from English text.)**Response:***[Analysis]:*

1. Read the Comment: The comment expresses agreement with excessive rhetoric, postures, and general absurdities.
2. Identify Key Elements: The key elements include "discours excessifs" (excessive rhetoric), "postures" (postures), and "absurdités en général" (general absurdities). The phrase "je suis d'accord" (I agree) indicates that the commenter supports these aspects associated with Donald Trump.
3. Analyze Implications: The commenter agrees with the negative attributes mentioned, suggesting that they view these characteristics positively in the context of Donald Trump.
4. Determine Stance: Given the agreement with the described negative traits of Trump, the stance of the comment is clearly supportive of these aspects, even if not necessarily endorsing all of Trump's actions or policies.

[Stance]: FAVOR ✗

Revising CoTs via cross-lingual guidance

Input:*Prompt1 + (Prompt3) Analyze and judge the stance of a given French comment by following the example of the English comment.***French Target:** Donald Trump**French Text:** @QuiltingMuriel Mais je suis d'accord avec les discours excessifs, les postures et les absurdités en général. #SemST (The French(target language) text is translated from English(source language) text.)**[Example]:** (The example is from source language data, including target, text, and model response.)**French Target:** Donald Trump**English Text:** @QuiltingMuriel But I do agree about the excess speechifying and posturing and general nonsense. #SemST**Response:***[Analysis]:*

1. Read the Comment: The comment appears to be a reply to someone named @QuiltingMuriel, agreeing with them about the excessive speechifying, posturing, and general nonsense.
2. Identify Key Elements: The key elements here are "excess speechifying and posturing and general nonsense." These phrases suggest criticism of political rhetoric and behavior, particularly from politicians who engage in these behaviors.
3. Analyze Implications: The comment indicates agreement with another person regarding the perceived overuse of rhetoric and posturing by politicians. Given the context, it is likely that this criticism is directed at Donald Trump, as he was known for his controversial and often inflammatory speeches and statements during his presidency.
4. Determine Stance: Since Donald Trump was frequently criticized for these aspects of his public persona and political style, it can be inferred that the comment is AGAINST Donald Trump.

[Stance]: AGAINST**Response:***[Analysis]:*

1. Read the Comment: The comment is a reply to someone named @QuiltingMuriel, agreeing with them about the excessive speeches, postures, and general absurdities.
2. Identify Key Elements: Identify Key Elements: The key elements here are "discours excessifs" (excessive speeches), "les postures" (postures), and "absurdités en général" (general absurdities). These phrases suggest criticism of political rhetoric and behavior, particularly from politicians who engage in these behaviors.
3. Analyze Implications: The comment indicates agreement with another person regarding the perceived overuse of rhetoric and posturing by politicians. Given the context, it is likely that this criticism is directed at Donald Trump, as he was known for his controversial and often inflammatory speeches and statements during his presidency.
4. Determine Stance: Since Donald Trump was frequently criticized for these aspects of his public persona and political style, it can be inferred that the comment is AGAINST Donald Trump.

[Stance]: AGAINST ✓ (Guided by the examples from the source language, the CoT generated by the model in the target language data have been effectively corrected.)

Table 7: An example of cross-lingual guidance. The model's performance on the target language data usually falls short compared to the source language data. The cross-lingual CoT alignment strategy uses CoTs from the source language data as guidance, helping the model generate higher quality reasoning processes in the target language.

Sampling CoTs & Rewards

Input:

(Prompt) Task: Determine the stance of the text towards the target, following a clear and structured process.

[Analysis]:

1. Read the Comment: Understand the context and sentiment.
2. Identify Key Elements: Highlight terms or phrases that suggest support, opposition, or neutrality.
3. Analyze Implications: Explain how these elements reflect the author's viewpoint, considering any relevant references.
4. Determine Stance: Decide if the text is "FAVOR," "AGAINST," or "NEUTRAL."

[Stance]: "FAVOR," "AGAINST," or "NEUTRAL."

Target: Foreign Policy | Švýcarsko již několik let prosazuje aktivnější zahraniční politiku, která je méně orientovaná na přísnou neutralitu. Vítáte to?

Text: Dobré služby Švýcarska a oddanost lidským právům jsou nesmírně důležité! Ty však nesmí být fíkovým listem pro pouhé dosažení ekonomických výhod, například uzavřením nových dohod o volném obchodu s diktátorskými státy a režimy.

Response:

[Analysis]:

1. Read the Comment: The comment discusses Switzerland's push for more active foreign policy that is less focused on strict neutrality. The author is evaluating whether this is a positive development.
2. Identify Key Elements: Key phrases include "Dobré služby Švýcarska a oddanost lidským právům jsou nesmírně důležité" (Switzerland's good services and commitment to human rights are extremely important) and "nesmí být fíkovým listem pro pouhé dosažení ekonomických výhod" (should not be just a green piece of paper for purely economic benefits).
3. Analyze Implications: The author acknowledges the importance of Switzerland's good services and commitment to human rights. However, they express concern that any active foreign policy should not compromise these principles for economic gains, such as entering free trade agreements with authoritarian regimes.
4. Determine Stance: The comment aligns with the idea that Switzerland's active foreign policy should prioritize human rights and not compromise on them for economic benefits, indicating a stance that finds value in active foreign policy but with conditions.

[Stance]: FAVOR

Reward: 1.0 (The answer is correct)

Input:

Prompt

Target: Foreign Policy | question: Švýcarsko již několik let prosazuje aktivnější zahraniční politiku, která je méně orientovaná na přísnou neutralitu. Vítáte to?

Text: Švýcarsko bylo díky své trvalé neutralitě z velké části ušetřeno mezinárodních válek. Tato důležitá skutečnost se dnes stále více odhaluje.

Response:

[Analysis]:

1. Read the Comment: The comment discusses Switzerland's long-standing neutrality and its benefits in avoiding international wars. It also hints at a shift in Switzerland's foreign policy towards a more active approach that is less oriented towards strict neutrality. The author is questioning whether the reader approves of this change.
2. Identify Key Elements: "Švýcarsko již několik let prosazuje aktivnější zahraniční politiku" (Switzerland has been promoting a more active foreign policy for several years) "méně orientovaná na přísnou neutralitu" (less oriented towards strict neutrality) "tato důležitá skutečnost se dnes stále více odhaluje" (this important fact is being revealed more and more today)
3. Analyze Implications: The comment suggests that Switzerland's traditional neutrality has provided the corewarduntry with protection from major international conflicts. It implies that this neutrality is no longer an advantage in today's changing world and that moving towards an active foreign policy is a necessary adaptation. The tone is reflective, suggesting that the shift is seen as inevitable and beneficial.
4. Determine Stance: The author is neutral but leans towards supporting the change, as the argument is presented as a reflection on the evolving nature of international relations and the need for Switzerland to adapt. There is no explicit statement of approval or disapproval, but the implication is that the shift is a positive adaptation.

[Stance]: NEUTRAL

Reward: 0.1 (The answer is not correct)

Table 8: Examples of reinforcement learning sampling by the Qwen2.5-7B-Instruct model in Czech stance detection tasks. Includes Czech text translated from Politics training data, samples of the model's reasoning process, and rewards assessing the correctness of the model's responses.

Sampling CoTs & Rewards

Input:

(Prompt) Task: Determine the stance of the text towards the target, following a clear and structured process.

[Analysis]:

1. Read the Comment: Understand the context and sentiment.
2. Identify Key Elements: Highlight terms or phrases that suggest support, opposition, or neutrality.
3. Analyze Implications: Explain how these elements reflect the author's viewpoint, considering any relevant references.
4. Determine Stance: Decide if the text is "FAVOR," "AGAINST," or "NEUTRAL."

[Stance]: "FAVOR," "AGAINST," or "NEUTRAL."

Target: Society | question: Měly by mít páry stejného pohlaví ve všech oblastech stejná práva jako heterosexuální páry?

Text: Ústava by neměla dělat rozdíly mezi muži a ženami. Všichni jsme lidské bytosti.

Response:

[Analysis]:

1. Read the Comment: The comment argues that the constitution should not differentiate between men and women, emphasizing that all humans are equal.
2. Identify Key Elements: The key elements are "Ústava by neměla dělat rozdíly mezi muži a ženami" (The constitution should not make a difference between men and women) and "Všichni jsme lidské bytosti" (We are all human beings).
3. Analyze Implications: The author is advocating for equality in legal rights, suggesting that both same-sex and heterosexual couples should have the same rights under the law. The phrase "všichni jsme lidské bytosti" (we are all human beings) underscores the core belief that all human beings should be treated equally.
4. Determine Stance: The comment supports the idea that everyone, regardless of sexual orientation, should have equal rights, which implies a stance of support for LGBTQ+ rights and equal treatment under the law.

[Stance]: FAVOR

Reward: 1.0 (The answer is correct)

Input:

(Prompt)

Target: Security | question: Mělo by se upustit od zpřísnění podmínek pro přijetí do civilní služby, které navrhuje Spolková rada?

Text: Obecně prospěšné práce nebyly zamýšleny jako náhradní služba, ale jako poslední možnost v jednotlivých případech. Armáda již nemůže plnit své cílové úrovně. Nejlepším řešením problému by byla povinnost mužů a žen sloužit na stejné úrovni (zbytek vykonávat veřejně prospěšné práce). Kromě toho má být zpřesněna škála vhodnosti, aby motivovaní, okrajově neschopní lidé mohli stále sloužit.

Response:

[Analysis]:

1. Read the Comment: The comment discusses whether the EU Council's proposed tightening of conditions for joining civil service should be abandoned. The author believes that civil service should not be seen as substitute military service but rather as a last resort for individual cases. The author also supports making military service mandatory for both men and women on equal terms, with the option to fulfill civil service if military service is not suitable.
2. Identify Key Elements: The key elements are the criticism of the EU Council's proposal, the clarification of the intended purpose of civil service, the proposal for equal mandatory military service for both genders, and the suggestion to refine the eligibility scale.
3. Analyze Implications: The author is clearly critical of the EU Council's proposed changes as they perceive it to be an unnecessary complication or tightening of conditions. The author suggests that the primary function of civil service is to be a secondary option, and that military and civil service should be integrated, supporting the idea that the main service should be military, but with alternatives for those not suited for it. This shows a preference for maintaining or enhancing traditional military service roles while adding flexibility through civil service.
4. Determine Stance: Based on the analysis, the comment is advocating for maintaining traditional military service and flexibility for those not suited, implying support for the idea of upholding current conditions or slightly modifying them rather than tightening them.

[Stance]: FAVOR

Reward: 0.1 (The answer is not correct)

Table 9: Examples of reinforcement learning sampling by the Qwen2.5-7B-Instruct model in Czech stance detection tasks. Includes Czech text translated from Society training data, samples of the model's reasoning process, and rewards assessing the correctness of the model's responses.

Sampling CoTs & Rewards

Input:

(Prompt1) Task: Determine the stance of the text towards the target, following a clear and structured process.

[Analysis]:

1. Read the Comment: Understand the context and sentiment.
2. Identify Key Elements: Highlight terms or phrases that suggest support, opposition, or neutrality.
3. Analyze Implications: Explain how these elements reflect the author's viewpoint, considering any relevant references.
4. Determine Stance: Decide if the text is "FAVOR," "AGAINST," or "NEUTRAL."

[Stance]: "FAVOR," "AGAINST," or "NEUTRAL."

Target: Independence of Catalonia

Text: @azukikiyo @agarzon Fotre @agarzon, només t'has assabentat del judici per mitjans internacionals? Que poc cas ens feu als catalans.... Per enderrocar el règim del 78, cal començar per la #republicacatalana, després ja estarem amb vosaltres per enderrocar el monarca.

Response:

[Analysis]:

1. Read the Comment: The comment mentions that large bankers should intensify their propaganda claiming that "Hillary cares about ordinary people." The overall tone is critical and sarcastic.
2. Identify Key Elements: The key elements here include references to the trial ("judici"), international media, and the idea of a Catalan Republic. The phrase "Per enderrocar el règim del 78" implies dissatisfaction with the current Spanish constitutional regime established in 1978.
3. Analyze Implications: The author emphasizes the importance of prioritizing Catalan independence through the establishment of a republic. This indicates a strong support for the independence movement and a belief in taking decisive action towards achieving it.
4. Determine Stance: The comment strongly advocates for Catalan independence and calls for immediate steps towards establishing a Catalan Republic.

[Stance]: FAVOR

Reward: 1.0 (The answer is correct)

Input:

Prompt1

Target: Independence of Catalonia

Text: 100 anys d'honradesa NO!!!! 100 anys de golfs i lladres que no dimiteixen #eleccions2019votodrets

Response:

[Analysis]:

1. Read the Comment: The comment translates to "100 years of honesty NO!!!! 100 years of golfers and criminals who do not resign #eleccions2019votodrets". The author is expressing strong disagreement with something related to the past 100 years, possibly referring to the current political situation in Catalonia.
2. Identify Key Elements: "100 anys d'honradesa" (100 years of honesty) - "100 anys de golfs i lladres" (100 years of golfers and criminals) - "que no dimiteixen" (who do not resign) - "#eleccions2019votodrets" (referring to the 2019 elections and voting rights)
3. Analyze Implications: The author is using sarcasm to criticize the past 100 years, suggesting that instead of being honest, there have been corrupt individuals who have not resigned from their positions. This implies dissatisfaction with the current political leadership and a desire for change, which is often associated with supporting independence movements.
4. Determine Stance: The comment suggests a critical view of the current political establishment and a push for change, which is indicative of support for independence.

[Stance]: FAVOR

Reward: 0.1 (The answer is not correct)

Table 10: Examples of reinforcement learning sampling by the Qwen2.5-7B-Instruct model in Catalan stance detection tasks. Includes Catalan text translated from Independence training data, samples of the model's reasoning process, and rewards assessing the correctness of the model's responses.

Sampling CoTs & Rewards

Input:

(Prompt) Task: Determine the stance of the text towards the target, following a clear and structured process.

[Analysis]:

1. Read the Comment: Understand the context and sentiment.
2. Identify Key Elements: Highlight terms or phrases that suggest support, opposition, or neutrality.
3. Analyze Implications: Explain how these elements reflect the author's viewpoint, considering any relevant references.
4. Determine Stance: Decide if the text is "FAVOR," "AGAINST," or "NEUTRAL."

[Stance]: "FAVOR," "AGAINST," or "NEUTRAL."

Target: Hillary Clinton

Text: @nyinvesting, les grands banquiers devraient intensifier leur propagande selon laquelle « Hillary se soucie des petites gens » #SemST

Response:

[Analysis]:

1. Read the Comment: The comment suggests that the author believes focusing only on international media regarding the trial is insufficient. They argue that this approach disrespects Catalans and propose instead starting with establishing a Catalan Republic before addressing other issues like the monarchy.
2. Identify Key Elements: The key elements here include references to the trial ("judici"), international media, and the idea of a Catalan Republic. The phrase "Per enderrocar el règim del 78" implies dissatisfaction with the current Spanish constitutional regime established in 1978.
3. Analyze Implications: The author emphasizes the importance of prioritizing Catalan independence through the establishment of a republic. This indicates a strong support for the independence movement and a belief in taking decisive action towards achieving it.
4. Determine Stance: The comment strongly advocates for Catalan independence and calls for immediate steps towards establishing a Catalan Republic.

[Stance]: FAVOR

Reward: 1.0 (The answer is correct)

Input:

Prompt

Target: Donald Trump

Text: Osée 4:8 Ils se repaissent des péchés de mon peuple, Et ils mettent leur cœur à leurs iniquités. #pasteurs #chrétiens #repentir #SemST

Response:

[Analysis]:

1. Read the Comment: The comment references a biblical verse from Hosea 4:8, which states, "They feed on the sin of my people, and they are greedy for their iniquities." The comment also includes hashtags related to pastors, Christians, repentance, and a mention of SemST, which likely refers to a specific organization or movement.
2. Identify Key Elements: - The biblical verse is used to criticize those who are benefiting from the sins of the people. - The hashtags suggest a religious context, possibly linking the comment to Christian or religious leaders. - The phrase "they are greedy for their iniquities" can be interpreted as a criticism of people's selfish actions and moral failings.
3. Analyze Implications: - By referencing this biblical verse, the commenter is using religious language to criticize individuals or groups who are profiting from or indulging in sinful behavior. - The hashtags indicate that the commenter might be part of a broader religious or Christian community, possibly expressing moral or ethical concerns about current events or individuals. - The mention of Donald Trump is not explicitly stated but can be inferred from the context, as the comment could be referring to his critics or opponents within the Christian community.
4. Determine Stance: Given the critical tone and the use of biblical language to condemn those who benefit from sin, the comment appears to be against Donald Trump, especially if he is the subject of the criticism within the Christian community.

[Stance]: AGAINST

Reward: 0.1 (The answer is not correct)

Table 11: Examples of reinforcement learning sampling by the Qwen2.5-7B-Instruct model in French stance detection tasks. Includes French text translated from Election training data, samples of the model's reasoning process, and rewards assessing the correctness of the model's responses.