Modeling Human-Like Cognition for Stance Detection: Integrating Intuitive Judgment and Analytical Reasoning

Zhaodan Zhang[1,2,3,4], Jin Zhang[2,3,4,*], Jiafeng Guo[2,3,4], Xueqi Cheng[2,3,4]
[1]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences
[2]State Key Laboratory of AI Safety
[3] Institute of Computing Technology, Chinese Academy of Sciences
[4]University of Chinese Academy of Sciences
{zhangzhaodan23s,jinzhang,guojiafeng,cxq}@ict.ac.cn

Abstract

Stance detection aims to identify the attitude expressed in text towards a given target, with applications in public opinion analysis and misinformation mitigation. Despite recent advances in large language models (LLMs), two key challenges remain: (1) spurious correlations between superficial features and stance labels, and (2) lack of cognitive modeling that simulates the transition from intuitive perception to deliberate reasoning. To address these issues, we propose Cognitive-Driven Stance Detection (**CDSD**), inspired by Kahneman's Dual-Process Theory. CDSD integrates fast intuitive judgment (System 1) and analytical reasoning (System 2), enhanced by three key modules: attention-based cognitive alignment to compare system focus, uncertainty-aware belief update using Bayesian inference, and self-doubt-triggered counterfactual reasoning for re-evaluation under low consistency or high uncertainty. Experimental results on SEM16, P-Stance, and VAST show that CDSD outperforms state-of-the-art methods across multiple LLMs. Notably, CDSD exhibits strong robustness against textual perturbations such as emotional word removal and rhetorical restructuring. By integrating cognitive theory with NLP, our work provides a promising path toward more reliable and interpretable stance detection systems.

1 Introduction

Stance detection (Hasan and Ng, 2014; Küçük and Can, 2020), which identifies the textual attitude toward a specific target, plays a vital role in opinion mining (Graells-Garrido et al., 2020), misinformation mitigation (Lai et al., 2020), and public sentiment analysis (Lei et al., 2024). By analyzing structural and linguistic patterns in stance reasoning, researchers can uncover public opinion dynamics, monitor harmful discourse evolution, and foster a more ethical online environment (De Vinco et al., 2024; Graells-Garrido and Baeza-Yates, 2022; Zhang et al., 2023c).

Recent advances in large language models (LLMs) have revolutionized stance detection, enabling complex reasoning strategies such as chain-of-thought prompting (Yao et al., 2024; Zhang et al., 2024c, 2023a; Ding et al., 2024a; Lan et al., 2024), multi-agent collaboration (Wang et al., 2024), and knowledge infusion (Yan et al., 2024). These methods treat stance detection as a reasoning process (rather than a simple classification task) that benefits from contextual understanding, logical inference, and semantic disambiguation. Representative works like COLA (Lan et al., 2024), LC-CoT (Zhang et al., 2023b), and LogiMDF (Zhang et al., 2025b) have improved prediction accuracy, interpretability, and decision consistency.

Despite these gains, most LLM-based stance detection methods suffer from two critical limitations: (1) overreliance on surface-level linguistic cues (e.g., emotionally charged expressions) leads to unstable predictions when such cues are perturbed or missing, undermining model robustness (Li et al., 2025); (2) the lack of an explicit cognitive structure to simulate the human transition from intuitive perception to deliberate analysis, which is essential for reliable judgment.

To address these issues, we propose a novel Cognitive-Driven Stance Detection framework (**CDSD**), inspired by cognitive psychology's Dual-Process Theory (Kahneman, 2011). CDSD simulates human-like reasoning via two complementary systems: System 1 (fast, intuitive judgment for rapid initial predictions) and System 2 (slow, analytical reasoning triggered by high text complexity, ambiguity, or emotional language). We further integrate an attention-based alignment mechanism (to ensure cognitive consistency) and a Bayesian belief update mechanism (to fuse dual-process outputs under uncertainty). When low alignment or high uncertainty is detected, a self-doubt mecha-

* Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5852–5867

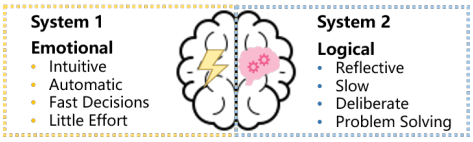

Figure 1: Dual-Process Theory

nism activates counterfactual reasoning for stance reassessment from alternative perspectives.

Our main contributions are as follows:

- We propose a cognitive-inspired dual-process architecture for stance detection, which combines fast intuitive judgment and slow deliberate reasoning to mimic human cognitive patterns.
- We design two core mechanisms (attention-based cognitive alignment and uncertainty-aware Bayesian belief update) and a structured self-doubt with counterfactual reasoning module, jointly enhancing model consistency, interpretability, and robustness against spurious correlations.
- Extensive experiments on three benchmark datasets (SEM16, P-Stance, and VAST) demonstrate that CDSD outperforms state-of-the-art methods and exhibits superior stability under textual perturbations.

2 Related Work

Stance detection has advanced with NLP developments: early statistical models (e.g., BiCond (Augenstein et al., 2016), CrossNet (Du et al., 2017)) leveraged RNNs and attention to capture textual dependencies, while BERT-based methods (e.g., TGA Net (Allaway and McKeown, 2020), BERT-GCN (Liu et al., 2021)) improved performance via contextualized representations.

Recent LLM-based approaches (COLA (Lan et al., 2024), LC-CoT (Zhang et al., 2023b), LogiMDF (Zhang et al., 2025b), KASD-ChatGPT (Li et al., 2023)) adopt multi-agent collaboration, chain-of-thought prompting (Zhang et al., 2025e), or logical frameworks to enhance interpretability (Zhang et al., 2025d,c). However, these methods remain vulnerable to spurious correlations (e.g., emotional expressions) and lack self-correction mechanisms.

Kahneman's Dual-Process Theory (Kahneman, 2011) (Figure 1) underpins our work, positing two complementary cognitive systems: **System 1** (fast, automatic, emotion-driven intuition) and **System 2** (slow, deliberate, logic-based reasoning). System 2 monitors and corrects System 1 biases, balancing cognitive speed and accuracy.

Inspired by this framework, we propose the Cognitive-Driven Stance Detection (**CDSD**) framework. Unlike single-paradigm LLM methods, CDSD simulates dual-process reasoning (System 1: intuitive judgment; System 2: analytical CoT reasoning) integrated via three mechanisms: attention alignment, uncertainty-aware Bayesian belief update, and counterfactual reasoning. This design enhances robustness to superficial linguistic cues, interpretability, and decision consistency - addressing key limitations of existing methods.

3 Methodology

To address the challenges of spurious correlations and lack of cognitive consistency in stance detection, we propose Cognitive-Driven Stance Detection (CDSD) inspired by Kahneman's theory of thinking fast and slow as shown in Figure 2. Our model integrates two complementary reasoning pathways: a fast, intuitive judgment process (System 1) that mimics rapid human perception, and a deliberate, analytical reasoning module (System 2) that engages when uncertainty, ambiguity, or emotional language is detected. System 1 provides an initial stance prediction along with confidence estimation and attention weights, while System 2 performs chain-of-thought reasoning to re-evaluate the stance by dissecting rhetorical strategies and semantic implications. To enhance interpretability and consistency, we further align the attention distributions between the two systems and fuse their outputs through Bayesian belief update. When conflicts arise, a self-doubt mechanism activates counterfactual reasoning, allowing the model to reassess its predictions from alternative perspectives. This cognitive-inspired architecture improves robustness, explainability, and generalization in complex and ambiguous textual scenarios.

3.1 Task Definition

Let $\mathcal{D} = \{(x_i, t_i, y_i)\}_{i=1}^{N}$ be a dataset of N instances, where x_i denotes the input text, t_i is the target entity, and $y_i \in \{\text{favor}, \text{against}, \text{neutral}\}$ represents the stance of x_i towards t_i. The goal of stance detection is to predict the correct stance label y_i for each (x_i, t_i) pair.

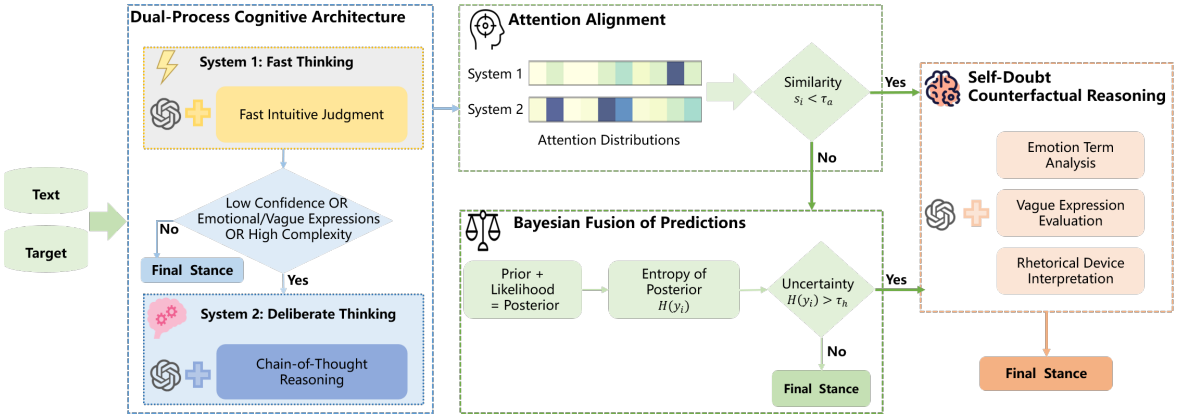

Figure 2: Overview of the Cognitive-Driven Stance Detection (CDSD) architecture. System 1 generates an initial stance prediction with confidence, while System 2 performs Chain-of-Thought (CoT) reasoning when triggered by low confidence, high text complexity, or emotional language. The attention alignment similarity s_i between the two systems is computed and compared against threshold τ_a. If $s_i < \tau_a$, or if the posterior entropy $H(y_i)$ exceeds threshold τ_h, the model enters a self-doubt state and activates counterfactual reasoning to refine the final stance prediction.

3.2 Dual-Process Cognitive Architecture

Inspired by cognitive psychology's Dual-Process Theory (Kahneman, 2011), our framework integrates two complementary reasoning pathways (System 1 and System 2) to address two core challenges of LLM-based stance detection: overreliance on superficial linguistic cues and lack of human-like cognitive transition from intuition to deliberation.

Formally, given input text x_i and target t_i, System 1 generates an initial stance prediction $\hat{y}_i^{(1)}$ and confidence score $c_i^{(1)} \in [0, 1]$ via holistic intuitive judgment (no explicit reasoning). Its prompt explicitly forbids in-depth analysis and requires only a direct stance and confidence output (e.g., judging stance intuitively without reasoning), with full details provided in Appendix A. Mathematically, this is expressed as:

$$(\hat{y}_i^{(1)}, c_i^{(1)}) = f_{\text{S1}}(x_i, t_i), \tag{1}$$

where f_{S1} is System 1's mapping function.

To enhance model robustness against ambiguous or emotionally charged texts (a key challenge of superficial cue reliance), we design a triggering mechanism for System 2 (deliberate Chain-of-Thought (CoT) reasoning), which activates if any of the following conditions hold: - $c_i^{(1)} < \tau_c$ (low System 1 confidence), - $\mathcal{C}(x_i) > \tau_{\text{comp}}$ (high text complexity), - x_i contains vague or emotionally salient expressions.

When triggered, System 2 outputs a refined prediction $\hat{y}_i^{(2)}$ and confidence $c_i^{(2)}$ via step-by-step CoT reasoning. Its prompt follows a structured 4-step protocol (claim identification $\rightarrow$ stance evaluation $\rightarrow$ rhetorical analysis $\rightarrow$ final judgment) to guide in-depth textual analysis, with full details in Appendix B. Mathematically, this is:

$$(\hat{y}_i^{(2)}, c_i^{(2)}) = f_{\text{S2}}(x_i, t_i), \tag{2}$$

where f_{S2} is the CoT-based reasoning function.

3.3 Attention-Based Cognitive Alignment

Grounded in Dual-Process Theory (Kahneman, 2011), we propose an attention-based cognitive alignment mechanism to enforce cognitive consistency between System 1 and System 2, and enhance model interpretability and robustness against ambiguous/emotionally charged texts.

Given an input text x_i (sequence of n tokens), let $\alpha_i^{(1)}$ and $\alpha_i^{(2)}$ denote the attention weight vectors of System 1 and System 2, respectively (each $\alpha_{ij}^{(k)}$ reflects the j-th token's importance for stance prediction). We quantify their focus overlap via cosine similarity s_i:

$$s_i = \frac{\alpha_i^{(1)} \cdot \alpha_i^{(2)}}{\|\alpha_i^{(1)}\| \cdot \|\alpha_i^{(2)}\|}, \tag{3}$$

where $s_i \in [-1, 1]$ (values near 1 = strong alignment; near/below 0 = divergent attention). A threshold τ_a is set: if $s_i < \tau_a$, the focus discrepancy is deemed significant.

This alignment acts as a critical signal to trigger the self-doubt and counterfactual reasoning mechanism, which reconciles conflicting interpretations

between intuitive (System 1) and analytical (System 2) judgments. It also improves model transparency (revealing token-level driving factors) and generalization (discouraging spurious attention correlations), ultimately boosting the reliability and explainability of stance detection decisions.

3.4 Uncertainty-Aware Belief Update

To ensure robust and cognitively consistent stance prediction under conflicting or uncertain evidence from System 1 and System 2, we introduce an uncertainty-aware belief update mechanism grounded in probabilistic reasoning. This module dynamically integrates the outputs of both systems into a unified posterior distribution over stance labels, while also enabling introspective re-evaluation when decision confidence is low.

Let $y_i \in \{\texttt{favor}, \texttt{against}, \texttt{neutral}\}$ denote the true stance label of text x_i towards target t_i. We define a uniform prior to represent initial ignorance about the stance:

$$P(y_i) = \frac{1}{3}. \tag{4}$$

Given the predictions $(\hat{y}_i^{(1)}, c_i^{(1)})$ from System 1 and $(\hat{y}_i^{(2)}, c_i^{(2)})$ from System 2 — where $\hat{y}_i^{(k)}$ denotes the predicted stance and $c_i^{(k)} \in [0, 1]$ represents the system's confidence — we model the likelihood function as follows:

$$P(\hat{y}_i^{(k)} \mid y_i) = \begin{cases} c_i^{(k)} & \text{if } \hat{y}_i^{(k)} = y_i \\ \frac{1-c_i^{(k)}}{2} & \text{otherwise,} \end{cases} \tag{5}$$

for $k = 1, 2$, reflecting that each system assigns high probability to its own prediction and distributes the remaining probability mass equally among the other two classes.

Using Bayes' theorem, we compute the posterior distribution over stance labels:

$$P(y_i \mid \hat{y}_i^{(1)}, \hat{y}_i^{(2)}) \propto P(\hat{y}_i^{(1)} \mid y_i) \cdot P(\hat{y}_i^{(2)} \mid y_i) \cdot P(y_i). \tag{6}$$

This posterior encapsulates the combined belief of both cognitive systems, weighted by their respective confidences. To quantify the certainty of this belief state, we compute the entropy of the posterior distribution:

$$H(y_i) = -\sum_{y \in \mathcal{Y}} P(y \mid \hat{y}_i^{(1)}, \hat{y}_i^{(2)}) \log P(y \mid \hat{y}_i^{(1)}, \hat{y}_i^{(2)}), \tag{7}$$

where $\mathcal{Y} = \{\texttt{favor}, \texttt{against}, \texttt{neutral}\}$.

If $H(y_i) > \tau_h$, indicating either high uncertainty or conflicting judgments between the two systems, the model enters a self-doubt state and activates counterfactual reasoning to reassess its stance prediction. Otherwise, the most probable stance according to the posterior is selected as the final output:

$$\hat{y}_i = \underset{y \in \mathcal{Y}}{\text{argmax}}\, P(y \mid \hat{y}_i^{(1)}, \hat{y}_i^{(2)}). \tag{8}$$

This uncertainty-aware belief update serves as the central coordination mechanism in our architecture, ensuring that stance predictions are not only accurate but also epistemically justified.

3.5 Self-Doubt and Counterfactual Reasoning

When low cognitive alignment ($s_i < \tau_a$) or high decisional uncertainty ($H(y_i) > \tau_h$) is detected, we activate the counterfactual reasoning module—this step is necessary to simulate human-like introspection, reconcile conflicting judgments between System 1 and System 2, and avoid decisions biased by superficial linguistic cues.

Given input text-target pair (x_i, t_i) and dual-process predictions $\hat{y}_i^{(1)}$, $\hat{y}_i^{(2)}$ (with confidence scores), we define the counterfactual reasoning function $f_{\text{CF}}(\cdot)$ (implemented via structured prompting, full prompt details in Appendix C):

$$(\hat{y}_i^{\text{cf}}, c_i^{\text{cf}}, e_i^{\text{cf}}) = f_{\text{CF}}(x_i, t_i, \hat{y}_i^{(1)}, \hat{y}_i^{(2)}), \tag{9}$$

where $\hat{y}_i^{\text{cf}}$ is the updated stance, $c_i^{\text{cf}} \in [0.0, 1.0]$ is the revised confidence, and e_i^{cf} is the natural language explanation for the update.

$f_{\text{CF}}(\cdot)$ analyzes the text under hypothetical transformations (no explicit re-prediction) by: - Identifying emotional terms and their impact on initial stance; - Evaluating vague expressions and their effect on prediction certainty; - Examining rhetorical devices and alternative interpretations.

This process uses a pre-defined reasoning schema (not free-form prompting), and its output updates the final stance if a more stable interpretation is suggested. This structured analysis is essential for grounding decisions in semantic content (rather than superficial cues), ultimately boosting model robustness and interpretability.

4 Experiments

4.1 Datasets

We conduct experiments on the **VAST**, **SEM16**, and **PStance** datasets to evaluate our proposed method. **VAST** (Allaway and McKeown, 2020) is characterized by its large number of targets across various domains, annotated with *pro*, *con*, or *neutral* stance labels. The **SEM16** dataset (Mohammad et al., 2016) contains six predefined targets, including Donald Trump (DT), Hillary Clinton (HC), Feminist Movement (FM), Legalization of Abortion (LA), Atheism (A), and Climate Change (CC). Each instance is categorized as Favor, Against, or Neutral. We remove two targets A and CC due to data quality issues and apply a leave-one-target-out strategy for zero-shot evaluation (Wei and Mao, 2019). The **PStance** dataset (Li et al., 2021) focuses on the stance of individuals towards three prominent political figures in the United States: Donald Trump (trump), Joe Biden (biden), and Bernie Sanders (sanders). This large-scale dataset includes only two stance labels: favor or against. We exclude inconsistent "none" labeled samples to ensure data quality and avoid noisy sample interference. Detailed dataset statistics are provided in Appendix D.

4.2 Evaluation Metrics

For the **VAST** dataset (Allaway and McKeown, 2020), we calculate the Macro-averaged F1 score across all labels to evaluate the performance of the models on the test set. For the **SEM16** and **PStance** datasets, we report the F_{avg}, which is the average of the F1 scores for the *Favor* and *Against* classes (Mohammad et al., 2016; Li et al., 2021). We compute F_{avg} for each target.

4.3 Baselines

To evaluate the effectiveness of our proposed method, we compare it with a series of strong baselines categorized into three main groups: statistics-based models, BERT-based models, and LLM-based models. For detailed information on these baselines, please refer to the Appendix E; all our baseline performance results are cited from the original papers. For **statistics-based models**, we include: **BiCond** (Augenstein et al., 2016), **CrossNet** (Du et al., 2017), **TPDG** (Liang et al., 2021), and **TOAD** (Allaway et al., 2021). The **BERT-based models** group includes: **TGA Net** (Allaway and McKeown, 2020), **BERT-Joint** (Devlin et al., 2019), **BERT-GCN** (Liu et al., 2021), **JointCL** (Liang et al., 2022b), **TarBK** (Liang et al., 2021), **PT-HCL** (Liang et al., 2022a),**TATA** (Hanley and Durumeric, 2023), **WS-BERT-Dual** (He et al., 2022), **KAI** (Zhang et al., 2024a) and **CKI** (Yan et al., 2024). For **LLM-based models**, we consider: **COLA** (Lan et al., 2024), **GPT-3.5** (Lan et al., 2024), **GPT-EDDA** (Ding et al., 2024b), **LCDA** (Zhang et al., 2025a), **KASD-ChatGPT** (Li et al., 2023), **LC-CoT** (Zhang et al., 2023b), **LogiMDF** (Zhang et al., 2025b) and **FACTUAL** (Li et al., 2025).

4.4 Implementation Details

We utilize open-source models **Qwen2.5-7B-Instruct** (Qwen Team, 2024), **Meta-Llama-3.1-8B-Instruct** (Llama Team, 2024) and **DeepSeek-R1-Distill-Qwen-7B** (DeepSeek-AI, 2025) as they can obtain attention weights. All experiments are conducted on an NVIDIA A800 GPU.

The key parameters and thresholds in our framework were determined empirically: the confidence threshold ($\tau_c = 0.6$) triggers System 2 when System 1's confidence $c_i^{(1)}$ falls below this value; the text complexity threshold ($\tau_{\text{comp}} = 0.7$), based on normalized language model perplexity, activates System 2 if $\mathcal{C}(x_i) > \tau_{\text{comp}}$; the emotional intensity threshold ($\tau_e = 0.3$), derived from VADER sentiment scores (Hutto and Gilbert, 2014), initiates deeper analysis if $|\text{VADER}(x_i)| > \tau_e$; the attention alignment threshold ($\tau_a = 0.5$) triggers self-doubt when the similarity between the attention distributions of the two systems s_i falls below this value; and the posterior entropy threshold ($\tau_h = 0.5$) activates counterfactual reasoning whenever the model's uncertainty $H(y_i)$ exceeds this threshold. Details on hyperparameter sensitivity analysis are provided in Appendix F.

All LLM-based inference is performed using temperature sampling (temperature = 0.7) and nucleus sampling (top-p = 0.9), ensuring both diversity and coherence in generated responses. The computational overhead of System 2 and counterfactual reasoning is modest: only 30–40% of samples trigger these modules (specifically, those with low confidence, high text complexity, or emotionally charged language). To ensure statistical reliability, we report results averaged over 5 repeated runs to mitigate the impact of any variance in model performance. Detailed implementation details are provided in Appendix F.

5 Results and Discussion

We evaluate our proposed **CDSD** through the following four research questions:

RQ1: How does CDSD perform compared to state-of-the-art stance detection models on PStance, SEM16, and VAST datasets?

RQ2: Is each core module in CDSD effective in contributing to final performance?

RQ3: Does CDSD demonstrate improved robustness under textual perturbations such as removal of emotional words or vague expressions?

RQ4: How does the attention mechanism in CDSD highlight key linguistic cues that influence stance prediction?

RQ5: How does the performance of CDSD vary across different models ?

RQ1: Performance Comparison with State-of-the-Art Models Our proposed **Cognitive-Driven Stance Detection (CDSD)** framework demonstrates superior performance across all three datasets compared to state-of-the-art models. As shown in Table 1, CDSD consistently outperforms statistics-based, BERT-based, and LLM-based baselines on the SEM16, P-Stance, and VAST datasets.

On the **SEM16** dataset, CDSD achieves the highest scores on most targets, particularly excelling in complex and emotionally charged texts. For example, using the Llama3.1 backbone, CDSD achieves 80.5% on Donald Trump (DT), 84.8% on Hillary Clinton (HC), and 82.8% on Feminist Movement (FM) - all significantly outperforming the strongest baseline LogiMDF (72.2%, 84.1%, 78.0%, respectively). This indicates that our dual-process architecture effectively handles ambiguity and emotional language by combining intuitive and analytical reasoning.

In the **P-Stance** dataset, CDSD achieves SOTA results across all political figures. Using Qwen2.5, it reaches 88.1%, 88.4%, and 83.8% for Trump, Biden, and Sanders, respectively. These results surpass even strong LLM-based methods like CKI and FACTUAL, highlighting CDSD's ability to maintain high accuracy while reducing reliance on superficial linguistic cues.

On the more diverse and challenging **VAST** dataset, CDSD achieves 84.5% zero-shot accuracy and 84.3% overall F1 score with Llama3.1, significantly outperforming the best baseline LogiMDF (81.6%) and CKI (81.9%). This indicates that our framework improves generalization across unseen topics by leveraging attention alignment and counterfactual reasoning.

These results confirm that CDSD not only addresses spurious correlations but also enhances cognitive consistency, robustness, and interpretability - directly solving the key challenges identified in stance detection.

RQ2: Effectiveness of Each Core Module in CDSD To evaluate the contribution of each core module in CDSD, we perform an ablation study by removing individual modules from the full model. As shown in Table 2, each module plays a distinct and essential role in improving performance and addressing specific challenges in stance detection.

System 2 (w/o S2): Removing System 2 leads to the most significant performance drop across all datasets. This highlights its critical role in handling ambiguous or emotionally charged texts through deliberate reasoning. It enables deeper analysis of rhetorical strategies and semantic implications, directly countering spurious correlations caused by surface-level language cues.

Attention Alignment (w/o AA): Without attention-based cognitive alignment, the model struggles with interpretability and consistency between intuitive and analytical judgments. This results in moderate degradation, especially in complex texts where conflicting interpretations are common.

Belief Update (w/o BU): Removing uncertainty-aware belief update reduces the model's ability to handle conflicting evidence, leading to decreased robustness, particularly noticeable in VAST few-shot settings. This mechanism ensures that decisions are not made based on unreliable or inconsistent signals.

Counterfactual Reasoning (w/o CR): Eliminating this module impairs the model's self-doubt mechanism, making it more susceptible to superficial linguistic cues and reducing its ability to re-evaluate uncertain predictions. Counterfactual reasoning enhances both interpretability and robustness by simulating human-like introspection.

These results confirm that each core module contributes meaningfully to the overall effectiveness of CDSD, directly addressing key challenges such as spurious correlations, cognitive inconsistency, and lack of interpretability.

RQ3: Robustness Under Counterfactual-Inspired Textual Perturbations To evaluate whether our Cognitive-Driven Stance Detection

	Model	SEM16				P-Stance			VAST		
		DT	HC	FM	LA	Trump	Biden	Sanders	Zero-Shot	Few-Shot	Overall
Sta.	Bicond	30.5	32.7	40.6	34.4	73.0	69.4	64.6	42.8	40.0	41.5
	CrossNet	35.6	38.3	41.7	38.5	58.0	65.0	53.0	43.4	47.4	45.5
	TPDG	53.6	46.5	47.3	50.9	60.1	64.4	62.0	51.9	-	-
	TOAD	49.5	51.2	54.1	46.2	53.0	68.4	62.9	41.0	-	-
BERT	TGA Net	40.7	49.3	46.6	45.2	64.4	74.7	69.9	66.6	66.3	66.5
	BERT-Joint	41.0	50.1	42.1	44.8	67.2	75.0	71.1	66.0	64.6	65.3
	BERT-GCN	42.3	50.0	44.3	44.2	70.5	74.9	71.1	68.6	69.7	69.2
	JointCL	50.5	54.8	53.8	49.5	62.0	59.0	73.0	72.3	71.5	-
	TarBK	50.8	55.1	53.8	48.7	65.8	75.5	70.5	73.6	-	-
	PT-HCL	50.1	54.5	54.6	50.9	-	-	-	71.6	-	-
	TATA	63.8	65.4	66.9	62.9	-	-	-	77.1	74.1	76.3
	WS-BERT-Dual	-	53.7	47.1	42.6	69.2	77.9	71.6	75.3	73.6	74.5
	KAI	72.1	76.4	73.7	69.4	75.9	85.7	80.5	76.3	-	-
	CKI	-	-	-	-	86.2	84.1	80.5	81.9	79.6	80.7
LLM	COLA	68.5	81.7	63.4	71.0	86.6	84.0	79.7	73.4	-	-
	GPT-3.5	62.5	68.7	44.7	51.5	79.8	79.7	77.8	65.0	-	-
	GPT-EDDA	69.5	80.1	69.2	62.7	-	-	-	68.5	-	-
	LCDA	79.8	70.0	69.4	70.0	-	-	-	80.3	-	-
	KASD-ChatGPT	-	80.3	70.4	62.7	85.1	84.6	79.9	67.0	-	-
	LC-CoT	71.7	82.9	70.4	63.2	-	-	-	72.5	-	-
	LogiMDF	72.2	84.1	78.0	75.6	-	-	-	81.6	-	-
	FACTUAL	72.8	80.3	75.8	68.8	84.9	86.0	81.6	79.9	-	-
CDSD	Qwen2.5	80.3	84.6	79.9	**80.2***	88.1	88.4	83.8	83.4	83.7	83.5
	Llama3.1	**80.5***	84.8*	**82.8***	78.6	**89.3***	88.2	83.9	**84.5***	**84.2***	**84.3***
	DeepSeek-R1	80.2	**85.0***	81.4*	78.3	89.0	**89.1***	**84.8***	83.7	83.0	83.6

Table 1: Performance Comparison (%) of Stance Detection Methods on SEM16, P-Stance, and VAST Datasets. The best scores are in bold. Results with * denote that CDSD significantly outperforms baselines with the p-value < 0.05.

Model	SEM16				P-Stance			VAST		
	DT	HC	FM	LA	T	B	S	ZS	FS	ALL
CDSD	**80.5**	**84.8**	**82.8**	**78.6**	**89.3**	**88.2**	**83.9**	**84.5**	**84.2**	**84.3**
w/o S2	73.4	77.6	75.2	72.1	83.0	84.1	78.5	78.5	77.1	77.8
w/o AA	77.3	81.0	79.5	75.2	86.0	86.7	81.9	81.2	80.5	80.9
w/o BU	78.1	82.4	80.3	76.0	87.2	87.5	82.6	82.3	81.7	82.0
w/o CR	75.2	80.1	78.4	74.3	85.5	86.6	81.0	80.8	80.0	80.5

Table 2: Ablation study results of CDSD using Llama3.1 as the backbone. Best scores are bolded.

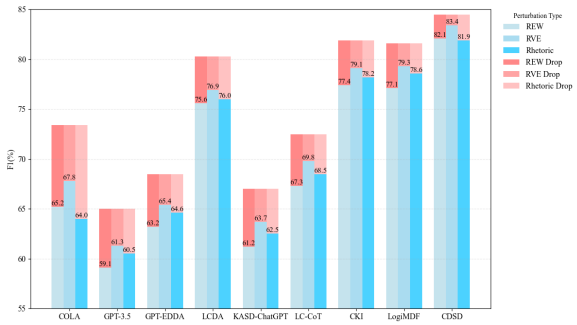

Figure 3: Performance of LLM-based models under textual perturbations on the VAST dataset in zero-shot setting.CDSD demonstrates the smallest performance degradation, indicating superior robustness to emotionally charged language, vague expressions, and rhetorical manipulation.

(CDSD) framework demonstrates improved robustness under linguistic manipulations that challenge intuitive judgment, we design textual perturbations inspired by the counterfactual reasoning module. Specifically, we test model performance under:

- **Removing Emotionally Charged Words (REW)**: Identifying and eliminating emotionally salient terms (e.g., "must", "disaster") to assess reliance on affective language.

- **Replacing Vague Expressions (RVE)**: Substituting uncertain or imprecise phrases (e.g., "possibly", "somewhat") with more definitive alternatives.

- **Altering Rhetorical Structures**: Rewriting texts to change exaggeration, analogy, or other persuasive devices.

As shown in Figure 3, CDSD achieves the highest accuracy across all perturbation types while also exhibiting the smallest performance drop compared to its original zero-shot score. For example, after removing emotional words (REW), most LLM-based models suffer drops ranging from 4.5% to 8.2%, whereas CDSD only experiences a 2.4% decline.

Similarly, under the replacement of vague expressions (RVE) and modification of rhetorical structures (Rhetoric), CDSD consistently maintains the smallest degradation among all models. This demonstrates that CDSD is less susceptible to superficial linguistic cues and instead relies on deeper semantic understanding - a behavior directly enabled by its counterfactual reasoning mechanism.

These results confirm that our cognitive-driven framework significantly improves robust-

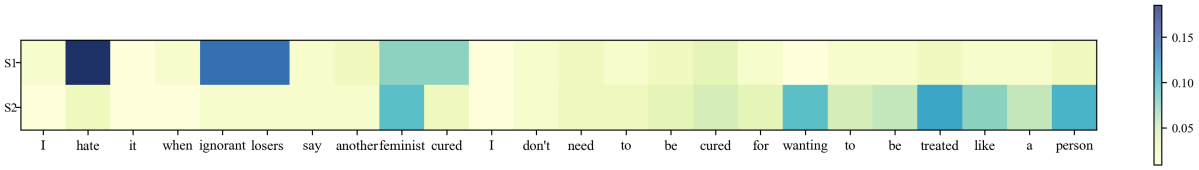

Figure 4: Attention Heatmap Visualization for Stance Prediction. The heatmap compares the attention distributions of System 1 (S1) and System 2 (S2) in processing a statement with complex linguistic cues.

ness against textual manipulations that are common in real-world stance detection scenarios.

RQ4: Attention Mechanism Analysis To investigate how the attention mechanism in our CDSD highlights key linguistic cues influencing stance prediction, we visualize the attention distributions of System 1 (S1) and System 2 (S2) using a heatmap.

Consider the following example: "I hate it when ignorant losers say 'another feminist cured.' I don't need to be cured for wanting to be treated like a person."

Despite the initial negative tone ("I hate it", "ignorant losers"), the true stance is FAVOR towards the Feminist Movement. As shown in Figure 4, System 1 assigns high attention to emotionally charged words such as "hate", "ignorant", and "losers", which may lead to an incorrect prediction. In contrast, System 2 focuses more on semantic core expressions like "wanting to be treated like a person", indicating deeper analysis of rhetorical strategies and emotional language.

The alignment between S1 and S2 is crucial for ensuring cognitive consistency. When discrepancies arise, as highlighted in the heatmap, the model triggers counterfactual reasoning to re-evaluate the stance from alternative perspectives. This process ensures that decisions are not solely based on superficial features but are grounded in stable interpretations supported by both systems.

These visualizations confirm that our attention mechanism effectively captures diverse linguistic cues, enhancing interpretability and robustness in stance detection tasks.

RQ5: Performance Variation Across Different LLMs Our proposed CDSD demonstrates strong generalizability across different large language models (LLMs). We evaluate CDSD using three distinct LLMs, all of which show competitive performance on stance detection tasks.

Among the three, **Llama3.1** achieves the best overall results, particularly excelling in complex reasoning and alignment with subtle linguistic cues. This can be attributed to its superior semantic understanding and logical reasoning capabilities, which are essential for handling ambiguous or emotionally charged texts - a key requirement in stance detection. **Qwen2.5** performs slightly lower than Llama3.1, especially in English-centric stance reasoning. This may reflect its relatively weaker grasp of nuanced expressions and implicit argumentation in English-dominated datasets. **DeepSeek-R1-Distill**, being a distilled version of a larger model, shows somewhat reduced performance, particularly in detecting implicit stances and executing multi-step reasoning. Distillation may lead to some loss of reasoning depth, making it less effective in capturing complex rhetorical patterns.

Despite these differences, all three models benefit significantly from our cognitive-driven framework, confirming that CDSD is not only effective but also model-agnostic. The consistent gains across diverse LLMs validate the robustness and general applicability of our method.

Qualitative Error Analysis We conduct a qualitative error analysis on 50 misclassified samples from the VAST zero-shot test set using Llama3.1. Three primary failure patterns are identified:

Ambiguous Sarcasm. For utterances such as "Oh sure, because listening to scientists has never worked before" (target: Climate Science), System 1 detects negative tone and predicts *against*, while System 2 fails to recognize sarcasm despite chain-of-thought reasoning. Counterfactual reasoning is not triggered due to high attention alignment.

Cultural Reference and Implicit Stance. Statements like "This policy is straight out of 1984" (target: Government Surveillance) require external world knowledge. Both systems lack background retrieval and default to *neutral*.

Intra-sentence Conflicting Signals. In sentences such as "I support free speech, but this tweet crosses the line", System 1 focuses on "support" and predicts *favor*, whereas System 2 focuses on

"crosses the line" and predicts *against*. The fused belief yields high entropy, and counterfactual reasoning still selects an incorrect label due to weak semantic grounding.

These limitations reveal challenges in external knowledge integration and complex rhetorical understanding.

6 Broader Applications: Extending CDSD to Emotional Intelligence Tasks

While this work focuses on stance detection as a core testbed for cognitive-driven modeling, our dual-process framework exhibits strong potential for generalization to broader human-centric NLP tasks, particularly *Emotional Intelligence (EI)* assessment. Stance detection was selected as the initial domain due to its clear alignment with dual-process cognition, well-established benchmarks, and observability of cognitive inconsistencies. Nevertheless, the core mechanisms of CDSD - intuitive System 1 judgment, analytical System 2 reasoning, attention alignment, uncertainty-aware belief update, and counterfactual reasoning - can be naturally adapted to EI-related tasks as follows.

Emotion Perception & Understanding. System 1 supports fast detection of surface-level emotional cues (e.g., valence, arousal), while System 2 performs deliberate reasoning about hidden causes, social context, and mixed emotions. This matches the "perceive $\rightarrow$ understand" pipeline widely used in computational emotional intelligence.

Emotion Regulation Simulation. The counterfactual reasoning module in CDSD can be repurposed to simulate emotion regulation strategies. For instance, it can evaluate hypothetical rephrasings: "If the speaker reframed frustration as concern, how would the emotional trajectory change?" This enables modeling of adaptive emotional expression and regulation.

Empathic Reasoning in Dialogues. In conversational datasets such as EMPATHETICDIALOGUES, CDSD can jointly model a speaker's stance and a listener's inferred emotional state. System 2 can further assess whether responses reflect perspective-taking, a central competency in emotional intelligence.

Multi-dimensional Emotion Labeling. Unlike stance detection with discrete ternary labels, EI tasks often require continuous or fine-grained emotion predictions (e.g., Ekman's six basic emotions, Plutchik's emotion wheel). Our uncertainty-aware Bayesian belief update naturally supports probabilistic modeling over rich emotion taxonomies.

This transferability confirms that the cognitive design of CDSD is not limited to stance detection but provides a generalizable paradigm for reliable, human-like reasoning in complex social NLP tasks.

7 Conclusion

In this work, we propose Cognitive-Driven Stance Detection (CDSD), a novel framework inspired by Kahneman's Dual-Process Theory. Unlike existing methods that rely on superficial cues or lack introspection, CDSD integrates fast intuitive judgment (System 1) and deliberate reasoning (System 2), simulating human-like transitions from intuitive judgment to rational analysis. With attention-based alignment, uncertainty-aware belief update, and counterfactual reasoning, our framework improves decision consistency, interpretability, and robustness against emotionally charged or complex texts. Experimental results on SEM16, P-Stance, and VAST show that CDSD outperforms state-of-the-art models across multiple LLMs. Ablation studies validate the role of each core module in addressing spurious correlations and cognitive inconsistency, while perturbation tests demonstrate superior robustness. By bridging cognitive science and NLP, CDSD offers a promising path toward more reliable stance detection.

Limitations

Despite its strong performance and interpretability, our Cognitive-Driven Stance Detection (CDSD) framework has two key limitations. First, although the computational overhead of System 2 and counterfactual reasoning is modest (only 30–40% of samples trigger these modules), it still introduces non-negligible inference latency, which restricts the framework's deployment in ultra-low-latency applications (e.g., real-time social media monitoring). Second, the effectiveness of our attention-based cognitive alignment mechanism heavily relies on the attention weight quality of backbone large language models (LLMs); models with weak attention fidelity may undermine the reliability of cognitive consistency judgment and further degrade the overall performance. Future work will explore lightweight dual-process architectures and adap-

tive threshold tuning strategies to mitigate these limitations.

Ethical Considerations

When developing and evaluating the CDSD framework for stance detection, we prioritize ethical rigor across all stages of research. First, we exclusively utilize publicly available, de-identified benchmark datasets (SEM16, P-Stance, VAST) with proper citation to avoid unauthorized use of private or sensitive textual data, ensuring compliance with data privacy regulations (e.g., GDPR). Second, we acknowledge that stance detection models may inherit potential biases from backbone LLMs and training data; to mitigate this risk, we report performance across different text genres and emotional tones, and avoid deploying the model for high-stakes applications (e.g., political censorship or discriminatory decision-making) without further bias auditing. Third, we make the core implementation details and prompt templates publicly available (in the supplement) to promote transparent, reproducible research and prevent malicious misuse of the framework. Finally, we emphasize that the CDSD framework is designed for academic research and benign applications (e.g., misinformation mitigation, public opinion analysis) to foster a more ethical and constructive online information environment.

Acknowledgments

This research was supported by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant No.2025ZD0123301, and by funding from the National Natural Science Foundation of China under Grant No.62441229 for the project "High-quality Dataset Construction". This valuable resource significantly enhanced the reliability and robustness of our experimental results. We would like to extend our sincere gratitude to all those who contributed to this work.

References

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

Emily Allaway, Malavika Srikanth, and Kathleen McKeown. 2021. Adversarial learning for zero-shot stance detection on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4756–4767, Online. 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.

Daniele De Vinco, Alessia Antelmi, Carmine Spagnuolo, and Luca Maria Aiello. 2024. Deciphering conversational networks: Stance detection via hypergraphs and llms. In *Companion Publication of the 16th ACM Web Science Conference*, Websci Companion '24, page 3–4, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Daijun Ding, Rong Chen, Liwen Jing, Bowen Zhang, Xu Huang, Li Dong, Xiaowen Zhao, and Ge Song. 2024a. Cross-target stance detection by exploiting target analytical perspectives. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10651–10655.

Daijun Ding, Li Dong, Zhichao Huang, Guangning Xu, Xu Huang, Bo Liu, Liwen Jing, and Bowen Zhang. 2024b. EDDA: An encoder-decoder data augmentation framework for zero-shot stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5484–5494, Torino, Italia. ELRA and ICCL.

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 3988–3994. AAAI Press.

Eduardo Graells-Garrido and Ricardo Baeza-Yates. 2022. Bots don't vote, but they surely bother! a study of anomalous accounts in a national referendum. In

Proceedings of the 14th ACM Web Science Conference 2022, WebSci '22, page 302–306, New York, NY, USA. Association for Computing Machinery.

Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. 2020. Every colour you are: Stance prediction and turnaround in controversial issues. In *Proceedings of the 12th ACM Conference on Web Science*, WebSci '20, page 174–183, New York, NY, USA. Association for Computing Machinery.

Hans Hanley and Zakir Durumeric. 2023. TATA: Stance detection via topic-agnostic and topic-aware embeddings. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11280–11294, Singapore. Association for Computational Linguistics.

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.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*.

Daniel Kahneman. 2011. Thinking, fast and slow.

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. *Comput. Speech Lang.*, 63:101075.

Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance detection with collaborative role-infused llm-based agents. In *Proceedings of the Eighteenth International AAAI Conference on Web and Social Media, ICWSM 2024, Buffalo, New York, USA, June 3-6, 2024*, pages 891–903. AAAI Press.

Yuanyuan Lei, Md Messal Monem Miah, Ayesha Qamar, Sai Ramana Reddy, Jonathan Tong, Haotian Xu, and Ruihong Huang. 2024. EMONA: Event-level moral opinions in news articles. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5239–5251, Mexico City, Mexico. Association for Computational Linguistics.

Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. 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.

Ang Li, Jingqian Zhao, Bin Liang, Lin Gui, Hui Wang, Xi Zeng, Xingwei Liang, Kam-Fai Wong, and Ruifeng Xu. 2025. Mitigating biases of large language models in stance detection with counterfactual augmented calibration. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7075–7092, Albuquerque, New Mexico. 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, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022a. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 2738–2747, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. 2021. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, WWW '21, page 3453–3464, New York, NY, USA. Association for Computing Machinery.

Bin Liang, Qinglin Zhu, Xiang Li, Min Yang, Lin Gui, Yulan He, and Ruifeng Xu. 2022b. 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*, pages 3152–3157, Online. Association for Computational Linguistics.

Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Xiaolong Wang, Yile Wang, Sijie Cheng, Peng Li, and Yang Liu. 2024. DEEM: Dynamic experienced expert modeling for stance detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4530–4541, Torino, Italia. ELRA and ICCL.

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.

Ming Yan, Tianyi Zhou Joey, and W. Tsang Ivor. 2024. Collaborative knowledge infusion for low-resource stance detection. *Big Data Mining and Analytics*, 7(3):682–698.

Zhenyin Yao, Wenzhong Yang, and Fuyuan Wei. 2024. Enhancing zero-shot stance detection with contrastive and prompt learning. *Entropy*, 26(4).

Bowen Zhang, Daijun Ding, Zhichao Huang, Ang Li, Yangyang Li, Baoquan Zhang, and Hu Huang. 2024a. Knowledge-augmented interpretable network for zero-shot stance detection on social media. *IEEE Transactions on Computational Social Systems*, pages 1–12.

Bowen Zhang, Daijun Ding, Liwen Jing, Genan Dai, and Nan Yin. 2024b. How would stance detection techniques evolve after the launch of chatgpt? *Preprint*, arXiv:2212.14548.

Bowen Zhang, Daijun Ding, Liwen Jing, and Hu Huang. 2023a. A logically consistent chain-of-thought approach for stance detection. *Preprint*, arXiv:2312.16054.

Bowen Zhang, Daijun Ding, Liwen Jing, and Hu Huang. 2023b. A logically consistent chain-of-thought approach for stance detection. *CoRR*, abs/2312.16054.

Bowen Zhang, Xianghua Fu, Daijun Ding, Hu Huang, Genan Dai, Nan Yin, Yangyang Li, and Liwen Jing. 2024c. Investigating chain-of-thought with chatgpt for stance detection on social media. *Preprint*, arXiv:2304.03087.

Bowen Zhang, Xu Li, Jun Ma, Xi Zhang, Genan Dai, and Jianhua Ye. 2025a. Zero-shot stance detection with logically consistent data augmentation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Bowen Zhang, Jun Ma, Xianghua Fu, and Genan Dai. 2025b. Logic augmented multi-decision fusion framework for stance detection on social media. *Information Fusion*, 122:103214.

Hong Zhang, Haewoon Kwak, Wei Gao, and Jisun An. 2023c. Wearing masks implies refuting trump?: Towards target-specific user stance prediction across events in covid-19 and us election 2020. In *Proceedings of the 15th ACM Web Science Conference 2023*, WebSci '23, page 23–32, New York, NY, USA. Association for Computing Machinery.

ZhaoDan Zhang, Jin Zhang, Xueqi Cheng, and Hui Xu. 2025c. T-MAD: Target-driven multimodal alignment for stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 580–595, Suzhou, China. Association for Computational Linguistics.

ZhaoDan Zhang, Jin Zhang, Hui Xu, Jiafeng Guo, and Xueqi Cheng. 2025d. MPRF: Interpretable stance detection through multi-path reasoning framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 454–470, Suzhou, China. Association for Computational Linguistics.

ZhaoDan Zhang, Zhao Zhang, Jin Zhang, Hui Xu, and Xueqi Cheng. 2025e. MPVStance: Mitigating hallucinations in stance detection with multi-perspective verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1053–1067, Vienna, Austria. Association for Computational Linguistics.

A System 1 Intuitive Judgment Prompt

Below is the full prompt used in System 1, which simulates fast, intuitive human judgment for stance detection. This prompt explicitly prohibits detailed reasoning and requests only a direct stance prediction with confidence score:

> *"Please act as an experienced human reader and intuitively judge the stance of the following text towards the target entity. Do not perform in-depth analysis or reasoning. Only output: {*`favor`, `against`, `neutral`*} and your confidence score between 0.0 and 1.0.*
>
> *Output format:*
>
> ```
> {
>   "stance": "favor/against/neutral",
>   "confidence": "0.0 ~ 1.0"
> }
> ```
>
> *Text:* `"{tweet}"`
>
> *Target:* `"{target}"`

This prompt is designed to elicit rapid, holistic evaluations that mimic System 1 thinking in dual-process theory, providing an initial stance prediction $\hat{y}_i^{(1)}$ with confidence $c_i^{(1)}$ without engaging in deliberate reasoning.

B System 2 Reasoning Prompt

Below is the full prompt used in System 2, which guides the model to perform step-by-step Chain-of-Thought (CoT) reasoning for stance detection:

> *"You are an analytical assistant tasked with determining the stance of the following text towards the target entity.*
>
> *Step 1: Identify key claims or opinions in the text regarding the target.*
>
> *Step 2: Evaluate whether these claims express support, opposition, or neutrality.*
>
> *Step 3: Consider if any rhetorical strategies (e.g., sarcasm, exaggeration), emotional language, or ambiguous expressions influence the perceived stance.*
>
> *Step 4: Make your final determination based on the most objective interpretation of the content.*
>
> *Output format:*
>
> ```
> {
>  "reasoning": "Please write down your
>  reasoning process here",
>  "stance": "favor/against/neutral",
>  "confidence": "0.0 ~ 1.0"
> }
> ```
>
> *Text:* `"{tweet}"`
>
> *Target:* `"{target}"` "

C Counterfactual Reasoning Prompt

Below is the full prompt used in the counterfactual reasoning module , which guides the model to reflect on its stance prediction by analyzing linguistic influences:

> *"You previously judged the stance of the following text towards 't_i' as '$\hat{y}_i^{(1)}$', while another interpretation suggests it could be '$\hat{y}_i^{(2)}$'. Please carefully reconsider your judgment by answering the following questions:*
>
> *- If all emotional words (e.g., 'must', 'disaster') were removed, would the stance change?*
>
> *- If vague expressions (e.g., 'possibly', 'somewhat') were omitted, would the stance remain the same?*
>
> *- If rhetorical strategies (e.g., exaggeration, analogy, metaphor) were neutralized, would the stance shift?*
>
> *Based on your analysis, please provide your revised stance, confidence score between 0.0 and 1.0, and explanation."*

D Dataset Statistics

We evaluate our method on three widely used stance detection benchmarks: SemEval-2016 Task 6 (SEM16), P-Stance, and VAST.

The **SEM16** (Mohammad et al., 2016) dataset contains tweets annotated with stance labels towards six predefined targets across multiple domains, including Donald Trump (DT), Hillary Clinton (HC), feminist movement (FM), legalization of abortion (LA), atheism (A), and climate change (CC). Each tweet is labeled as one of three stances: *favor*, *against*, or *neutral*. Following prior work, we remove targets A and CC due to data quality issues.

The **P-Stance** (Li et al., 2021) dataset includes 21,574 tweets related to three prominent political figures: Joe Biden (Biden), Bernie Sanders (Sanders), and Donald Trump (Trump). Each tweet is associated with a stance label indicating whether the author supports, opposes, or remains neutral toward the given politician. As noted in previous studies, samples labeled as "none" exhibit low annotation consistency. Therefore, we follow standard practice and exclude all instances labeled as "none" from our analysis.

The **VAST** (Allaway and McKeown, 2020) provides greater topic diversity and lexical variation. It contains over 3,300 distinct topics spanning various domains such as war, drug policy, natural resources, and education tax. Each instance is annotated with one of three stance categories: *pro*, *con*, or *neutral*.

	Train	**Dev**	**Test**
# Examples	13477	2062	3006
# Unique Comments	1845	682	786
# Zero-shot Topics	4003	383	600
# Few-shot Topics	638	114	159

Table 3: Statistics of **VAST** dataset.

E Baseline Models

This section provides detailed descriptions of the baseline models used in our experiments, grouped

Target	Favor	Against	Neutral
DT	148	299	260
HC	163	565	256
FM	268	511	170
LA	167	544	222
A	124	464	145
CC	335	26	203

Table 4: Statistics of **SEM16** dataset.

		Trump	Biden	Sanders
Train	**Favor**	2,937	2,552	2,858
	Against	3,425	3,254	2,198
Val	**Favor**	365	328	350
	Against	430	417	284
Test	**Favor**	361	337	343
	Against	435	408	292
Total		7,953	7,296	6,325

Table 5: Label distribution across different targets for P-Stance.

by their architectural and learning paradigms.

E.1 Statistics-based Models

These models rely on traditional neural architectures without leveraging pre-trained language representations. **BiCond** (Augenstein et al., 2016) utilizes two BiLSTM models to separately encode the input sentences and their associated targets for stance detection. **CrossNet**(Du et al., 2017) employs bidirectional LSTM to encode both text and topic information, and further introduces a target-specific attention mechanism before classification. **TPDG** (Liang et al., 2021) proposes a method that automatically distinguishes and adjusts the roles of terms in stance expressions based on whether they are target-dependent or independent. **TOAD** (Allaway et al., 2021) applies adversarial learning techniques to improve generalization across diverse topics in zero-shot stance detection.

E.2 BERT-based Models

These models are built upon the BERT architecture with various enhancements tailored for stance detection tasks. **TGA Net** (Allaway and McKeown, 2020) implicitly constructs and leverages associations between training and evaluation topics without requiring supervision. It uses BERT to encode texts and targets, followed by two fully connected layers for classification. **BERT-Joint** (Devlin et al., 2019) refers to the standard BERT framework, where bidirectional encoder representations are learned from large unlabeled corpora to generate dense vector representations for sentences and tokens. **BERT-GCN** (Liu et al., 2021) integrates commonsense knowledge into the model by leveraging both structural and semantic relational information, aiming to enhance generalization under zero- and few-shot settings. **JointCL** (Liang et al., 2022b) enhances feature learning through stance contrastive learning and target-aware prototypical graph contrastive learning, enabling better generalization to unseen targets. **PT-HCL** (Liang et al., 2022a) improves cross-domain transferability by incorporating contrastive learning with both semantic and sentiment knowledge. **TATA** (Hanley and Durumeric, 2023) uses contrastive learning as well as an unlabeled dataset of news articles that cover a variety of different topics to train topic-agnostic/TAG and topic-aware/TAW embeddings for use in downstream stance detection. **TarBK** (Liang et al., 2021) reduces the semantic gap between known and unseen targets by integrating background knowledge from Wikipedia. **WS-BERT-Dual** (He et al., 2022) (Wikipedia Stance Detection BERT) infuses external knowledge into stance encoding through a dual-BERT structure that jointly models content and target information. **KAI** (Zhang et al., 2024a) enhances stance detection performance through a knowledge-aware integration strategy, particularly effective in complex domains. **CKI** (Yan et al., 2024), which is a collaborative knowledge infusion approach for low-resource stance detection tasks, employing a combination of aligned knowledge enhancement and efficient parameter learning techniques.

E.3 LLM-based Models

These approaches utilize large language models with advanced prompting or reasoning strategies. **COLA** (Lan et al., 2024) adopts a collaborative role-infusion framework that involves multiple LLMs to improve stance prediction. **GPT-3.5** (Lan et al., 2024) ,which can be considered zero-shots, implemented in strict accordance with Zhang et al. (2024b). **GPT-EDDA** (Ding et al., 2024b) introduces an encoder-decoder data augmentation framework to enhance the quality and diversity of generated prompts. **LCDA** (Zhang et al., 2025a) focuses on improving data quality by maintaining logical coherence during generation. **KASD-ChatGPT** (Li et al., 2023) enhances stance detection by integrating external knowledge from

Hyperparameter	Low	Mid-Low	Optimal	Mid-High	High
Confidence threshold (τ_c)	0.4 (82.1%)	0.5 (83.7%)	0.6 (**85.2%**)	0.7 (84.3%)	0.8 (82.9%)
Text complexity (τ_{comp})	0.5 (81.8%)	0.6 (83.4%)	0.7 (**85.0%**)	0.8 (83.9%)	0.9 (82.5%)
Emotional intensity (τ_e)	0.1 (82.5%)	0.2 (84.1%)	0.3 (**85.1%**)	0.4 (83.8%)	0.5 (82.7%)
Attention alignment (τ_a)	0.3 (82.3%)	0.4 (83.9%)	0.5 (**85.2%**)	0.6 (84.0%)	0.7 (82.8%)
Posterior entropy (τ_h)	0.3 (82.0%)	0.4 (83.6%)	0.5 (**85.0%**)	0.6 (83.8%)	0.7 (82.6%)

Table 6: Hyperparameter sensitivity analysis is measured by F1-score (%) on VAST validation set. Optimal values are bolded for clarity.

Wikipedia and utilizing retrieval-augmented generation with ChatGPT. **LC-CoT** (Zhang et al., 2023b) leverages a structured chain-of-thought prompting strategy to guide LLMs toward more accurate and interpretable stance predictions. **LogiMDF** (Zhang et al., 2025b), which is , a Logic Augmented Multi-Decision Fusion framework that effectively integrates multiple LLMs' decision processes through a unified logical framework. **FACTUAL** (Li et al., 2025), in which, a trainable calibration network and counterfactual data augmentation are explored to mitigate the biases of LLMs in stance detection.

F Implementation Details

To ensure full reproducibility of our experiments, we provide a comprehensive description of the computational infrastructure, hyperparameter search process, random seed settings, and statistical evaluation methods used throughout this study.

F.1 Computational Infrastructure

All experiments were conducted using the following computing environment:

- **Hardware:** NVIDIA A800 GPU (40GB memory).
- **Software Stack:**
- Python version: 3.9
- PyTorch version: 2.1.0
- CUDA version: 11.8.0
- DeepSeek, Qwen2.5, and Llama3.1 models loaded via HuggingFace Transformers with custom attention extraction hooks.

F.2 Hyperparameter Search and Final Settings

During model development, we systematically explored a wide range of hyperparameters to optimize performance across all datasets. The following parameters were tuned:

- Confidence threshold (τ_c): tested in range [0.4, 0.8], final value set to 0.6 based on validation performance.

- Text complexity threshold (τ_{comp}): tested in range [0.5, 0.9], final value set to 0.7 using perplexity normalization on training data.

- Emotional intensity threshold (τ_e): tested in range [0.1, 0.5], final value set to 0.3 by analyzing VADER sentiment scores on ambiguous samples.

- Attention alignment threshold (τ_a): tested in range [0.3, 0.7], final value set to 0.5 using cosine similarity distribution analysis.

- Posterior entropy threshold (τ_h): tested in range [0.3, 0.7], final value set to 0.5 based on uncertainty calibration curves.

The final values were selected based on best average performance across three validation folds on the VAST dataset. Each parameter was evaluated independently while keeping others fixed to reduce interaction effects. To quantitatively demonstrate the impact of each hyperparameter on model performance, we present a sensitivity analysis table (Table 6) that reports the F1-score on the VAST validation set across different parameter values. The table shows that all parameters exhibit a "first up, then down" performance trend, confirming that the selected optimal values (bolded) achieve the best average performance across three validation folds. Additionally, the independent evaluation of each parameter (with others fixed) effectively reduces the interaction effects between hyperparameters, ensuring the reliability of the selected settings.

F.3 Random Seed and Reproducibility Settings

To ensure consistent and replicable results, we set the random seed for all modules as follows:

- Random seed for all LLM inference runs: 42
- Random seed for preprocessing and postprocessing steps: 42
- Random seed for system-level randomness in attention computation: 42

We followed the standard practice of fixing seeds before each experiment run to ensure deterministic behavior during inference. All reported results are averaged over 5 independent runs with identical

configurations but different prompt executions due to the nature of large language models.

F.4 Statistical Evaluation

To assess the significance of our method's improvements over baseline models, we applied paired t-tests across the five repeated runs for each dataset. Results marked with an asterisk (*) indicate that the improvement is statistically significant at the $p < 0.05$ level.

For example, when comparing CDSD against COLA or GPT-EDDA baselines, we computed t-statistics from the F1 score distributions across the five runs. This approach ensures that observed performance gains are not due to chance variations in model output.