SICI: A Semantic-Pragmatic Complexity Index Reveals Regime Shifts in LLM Stance Detection

Fuqiang Niu[1], **Bowen Zhang**[2]*
[1]School of Cyber Science and Technology,
University of Science and Technology of China, Hefei, China
[2]School of Artificial Intelligence, Shenzhen Technology University, Shenzhen, China

Abstract

Prompt-based LLMs are increasingly used for stance detection, but harder examples are not always repaired by clearer instructions, reasoning prompts, retrieval, or debate. We introduce SICI (Stance Inference Complexity Index), a seven-dimensional diagnostic measure of the semantic-pragmatic burden imposed by a target–text pair. Across SemEval-2016 and VAST, SICI predicts LLM accuracy better than surface proxies and shows substantial cross-scorer reliability ($\alpha = 0.771$). More importantly, LLM errors change regime as SICI increases: low-complexity examples invite over-attribution, especially AGAINST predictions; intermediate examples form an unstable boundary; and high-complexity examples rapidly concentrate on NONE. This phase-transition-like structure persists across GPT-3.5, GPT-4o-mini, DeepSeek-V3, and GPT-4o, although stronger models move the boundaries. A 14-method intervention study further shows that prompting, retrieval, and debate often shift models along the attribution–abstention axis rather than removing the high-complexity bottleneck.

1 Introduction

Stance detection asks whether a text expresses FAVOR, AGAINST, or NONE toward a given target (Mohammad et al., 2016; Aldayel and Magdy, 2021). With the rise of large language models (LLMs), recent work increasingly treats stance detection as prompt-based inference: a model receives the text, target, label definitions, and sometimes demonstrations or reasoning instructions, then outputs one of the stance labels. This paradigm has naturally led to richer elicitation strategies, including zero-shot and few-shot prompting, chain-of-thought reasoning, evidence-oriented prompting, retrieval augmentation, and multi-agent discussion or voting.

These strategies are motivated by a plausible assumption: if an LLM already possesses the relevant linguistic and world knowledge, then clearer instructions, more examples, explicit reasoning, or additional context should elicit better stance judgments. The empirical picture for hard stance examples is less settled. Prior work has explored chain-of-thought, counterfactual, and knowledge-enhanced prompting for stance detection (Weinzierl and Harabagiu, 2024; Taranukhin et al., 2024; Zhang et al., 2024b), while recent work also studies multi-path reasoning, implicit target augmentation, and stereotype-sensitive LLM evaluation (Zhang et al., 2025b; Ji et al., 2025; Dubreuil et al., 2025). These studies suggest that additional reasoning or external information does not automatically yield reliable gains when the stance is implicit or the target-related evidence is weak. This raises a basic question: what limits prompt-based LLM stance detection on hard examples?

We argue that this limitation cannot be understood solely as a problem of how the model is prompted. Stance examples differ substantially in the inferential burden they impose. Some texts explicitly mention the target and state a direct attitude. Others discuss an adjacent issue, rely on pragmatic implication or background knowledge, contain sentiment whose target is ambiguous, or simply lack enough evidence to support a stance judgment (Zhang et al., 2022). Hard examples are therefore not merely harder versions of the same label-selection problem: they may require qualitatively different amounts of semantic-pragmatic inference before a stance label can be licensed.

To study this source of variation, we introduce SICI, the **Stance Inference Complexity Index**. SICI combines seven semantic-pragmatic dimensions: target visibility, scope alignment, pragmatic implicitness, knowledge requirement, context dependence, label ambiguity, and polarity–stance gap. It is not intended to replace standard metrics such

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 39209–39221
October 24-29, 2026 ©2026 Association for Computational Linguistics

as accuracy, macro-F1, or the SemEval F_{avg} score. Instead, it provides an axis for asking a different question: how do LLM errors change as stance inference becomes more complex?

Our main finding is that LLM errors do not degrade smoothly with SICI. Instead, they exhibit a phase-transition-like shift between error regimes. At low complexity, models tend to over-attribute stance, with a pronounced bias toward AGAINST. At intermediate complexity, accuracy drops sharply, forming an unstable boundary region. At high complexity, predictions rapidly concentrate on NONE, producing an abstention-dominated regime. Piecewise regression significantly outperforms a linear model, indicating that this pattern is not well explained as a simple monotonic decline in difficulty.

This regime structure is qualitatively consistent across models and datasets. Stronger models improve aggregate performance and move transition boundaries, but they do not remove the shift itself: GPT-4o reduces low-complexity AGAINST overprediction while becoming more prone to NONE predictions in high-complexity regions. Cross-dataset analysis further shows that NONE is not semantically uniform. In SemEval, high-SICI examples often contain implicit stance, so NONE predictions frequently reflect false abstention. In VAST, many high-SICI examples are genuinely underspecified and labeled NONE, making abstention more often appropriate.

Finally, we return to the prompting bottleneck through a systematic intervention study. We evaluate a broad set of prompting-based interventions, including chain-of-thought, multi-step reasoning, SICI-aware prompting, retrieval augmentation, multi-agent debate, and voting. These interventions do not reliably break the high-complexity bottleneck. Instead, they often move the model along the attribution–abstention axis: prompts that encourage indirect inference can reduce some NONE predictions but increase false attribution, whereas prompts that emphasize caution or scope clarification can reduce over-attribution while pushing the model toward excessive NONE.

Our contributions are:

(1) We identify a prompting bottleneck in LLM stance detection: hard examples are not only underserved by stronger elicitation, but organized by semantic-pragmatic inference complexity.

(2) We introduce SICI, a seven-dimensional diagnostic measure of stance inference complexity, and show that it explains model behavior beyond surface features such as length, lexical diversity, and negation density.

(3) We show that LLM stance errors exhibit a phase-transition-like regime shift: from low-complexity over-attribution, through an unstable boundary region, to high-complexity NONE abstention.

(4) We demonstrate that this structure is qualitatively robust across models and datasets, and that prompting, retrieval, debate, and voting interventions often trade false attribution against false abstention rather than resolving both.

2 Related Work

Stance detection. Stance detection has been formalized in shared tasks and benchmark datasets such as SemEval-2016 Task 6 (Mohammad et al., 2016), VAST (Allaway and McKeown, 2020), P-Stance (Li et al., 2021), and multi-target or conversation-based variants (Wei et al., 2018; Li et al., 2023b,c). Earlier neural approaches use target-specific attention, memory networks, graph models, and transfer learning to model the relation between a text and a target (Du et al., 2017; Wei and Mao, 2019; Liang et al., 2021; Zhang et al., 2020). Pretrained language models and tweet-specific encoders such as BERT and BERTweet further improve in-domain and cross-target performance (Devlin et al., 2019; Nguyen et al., 2020). Recent work also incorporates background knowledge, commonsense, or prompt-based reasoning (Liu et al., 2021; Li et al., 2023a; Ding et al., 2024). Our work is complementary: rather than proposing another stance classifier, we ask which instance properties systematically determine when LLM classifiers fail.

Zero-shot and LLM-based stance inference. Zero-shot stance detection is challenging because test targets may not appear during training (Allaway and McKeown, 2020; Liang et al., 2022; Allaway et al., 2021). LLM prompting provides a natural zero-shot interface, and chain-of-thought or explanation-based variants have been explored for implicit stance and social media reasoning (Wei et al., 2022b; Gatto et al., 2023). Recent LLM-era stance work has explored reasoning over ideological or tree-structured perspectives, retrieval-augmented knowledge, and stance-specific prompting for zero-shot targets (Zhang et al., 2024a; Taranukhin et al., 2024; Zhang et al., 2024b). Other recent studies emphasize multi-path reasoning for

interpretability (Zhang et al., 2025b), LLM-driven implicit target augmentation for target-sparse examples (Ji et al., 2025), and stereotype-sensitive evaluation of zero-shot LLM stance detection (Dubreuil et al., 2025; Zhang et al., 2025a). We differ from these lines by treating difficulty itself as the object of measurement. Instead of adding reasoning paths, targets, or fairness probes, SICI asks which semantic-pragmatic properties predict when such systems fail.

Difficulty, calibration, and regime shifts. Instance difficulty is often estimated from model uncertainty, agreement, or surface features. Such signals are useful but can become circular when the same model both defines and evaluates difficulty. We instead define SICI using semantic-pragmatic attributes of the target–text pair, then validate its agreement across scorers and its relationship to independent prediction behavior. Our notion of a regime shift is inspired by work on emergent abilities and phase-transition-like behavior in machine learning systems (Wei et al., 2022a). Unlike scaling-law studies, which analyze model capability as a function of model or compute scale, we study behavioral transitions as a function of *instance complexity*.

3 The SICI Framework

Seven Dimensions. SICI assigns each target–text pair seven 0–4 scores covering target visibility, scope alignment, pragmatic implicitness, knowledge need, context dependence, label ambiguity, and polarity–stance mismatch (Table 1). Details of the corresponding inference stages are given in Appendix A.

Index Definition. Given seven scores $d_1, \ldots, d_7 \in \{0, 1, 2, 3, 4\}$, we define:

$$\mathrm{SICI}(x,t) = 0.65 \cdot \frac{\mathrm{mean}(d_1,\ldots,d_7)}{4} + 0.35 \cdot \frac{\max(d_1,\ldots,d_7)}{4}. \quad (1)$$

The index ranges from 0 to 1. The mean term captures cumulative load, while the max term captures bottlenecks where one severe ambiguity can dominate the instance. We use equal dimension weights because a pilot comparison against a manually weighted variant produced almost identical scores (Pearson $r = 0.9996$).

Scoring Protocol and Reliability. The main SICI scores are produced by GPT-4o-mini using dimension-specific rubrics and independent 0–4 judgments. To test whether SICI is a model idiosyncrasy, we additionally score a stratified sample of 200 instances with Claude Haiku 4.5 and DeepSeek-V3. Pairwise Spearman correlations are high: 0.829 (GPT vs. Claude), 0.853 (GPT vs. DeepSeek), and 0.884 (Claude vs. DeepSeek). The three-way ordinal Krippendorff's α is 0.771, above the conventional threshold for substantial agreement. Although cross-LLM agreement alone does not establish human validity, it suggests that SICI is not driven by a single model's prediction preferences.

A targeted validation with three blinded annotators on the same 200 instances achieved an ordinal Krippendorff's α of 0.73. Human-consensus SICI correlated strongly with the three LLM scorers ($\rho = 0.79$–0.82) and with human overall-difficulty judgments ($\rho = 0.78$), providing initial human-grounded support beyond cross-LLM agreement.

The scoring prompt is deliberately separated from stance prediction. The scorer is asked to judge the seven attributes of the instance, not to predict FAVOR, AGAINST, or NONE. This separation matters because a difficulty measure based on the same prediction confidence that later enters the evaluation would risk circularity. In our analysis, SICI is computed before any intervention comparison and is held fixed across all downstream models and prompts. The same SICI value is therefore used to analyze GPT-3.5, GPT-4o-mini, DeepSeek-V3, GPT-4o, and all intervention variants.

From Dimensions to Regimes. SICI supports two complementary uses. First, as a continuous score, it orders instances by expected inference complexity. Second, with empirically fitted boundaries, it partitions the data into regimes. We use two transition points, $b_1 = 0.45$ and $b_2 = 0.70$, obtained from segmented-regression analysis on SemEval. Low-complexity instances fall below b_1; intermediate instances occupy the boundary region where over-attribution and abstention compete; high-complexity instances exceed b_2 and are dominated by target invisibility, scope mismatch, and NONE-oriented abstention behavior. The thresholds are not intended as universal constants. Their role is diagnostic: they expose where a model changes its decision strategy as the target–text relation becomes less direct.

Dimension	Question measured	Low score	High-score example
V: Target visibility	Is the target explicitly mentioned?	target named directly	stance toward a politician inferred from a policy tweet
S: Scope alignment	Is the text mainly about the target?	text is on target	text discusses a related but shifted topic
P: Pragmatic implicitness	Is stance expressed indirectly?	direct support/opposition	sarcasm, metaphor, rhetorical implication
K: Knowledge requirement	Is background knowledge needed?	self-contained text	requires knowing a law, event, or group relation
C: Context dependency	Is external conversational context needed?	standalone text	reply or quote lacking prior context
A: Label ambiguity	Is the gold label semantically contestable?	clear label boundary	mixed or underspecified stance
G: Polarity–stance gap	Does sentiment polarity align with stance?	sentiment matches stance	positive affect used to oppose via irony

Table 1: The seven semantic-pragmatic dimensions of SICI. Each dimension is scored from 0 (low inference burden) to 4 (high inference burden). Together they track a multi-stage stance inference chain from target identification to polarity–stance bridging.

Dataset	N	Targets	Labels	Mean SICI
SemEval	1,249	5	F/A/N	∼0.35
VAST unseen	1,460	many	F/A/N	∼0.38
MTSD	500	2	F/A/N	∼0.33
P-Stance	777	1	F/A	0.267

Table 2: Datasets used in the analysis. F/A/N denotes FAVOR, AGAINST, and NONE.

4 Experimental Setup

Datasets. We evaluate on two core benchmarks. SemEval-2016 Task 6 contains English tweets labeled FAVOR, AGAINST, or NONE toward social and political targets (Mohammad et al., 2016). Our main SemEval analysis uses 1,249 test instances across five targets. VAST is a zero-shot stance dataset designed for unseen-topic generalization (Allaway and McKeown, 2020); we use the unseen-topic test split with 1,460 instances. We also report supporting cross-dataset model scores on MTSD and P-Stance where available, but the central cross-dataset regime analysis uses SemEval and VAST.

Models and Metrics. We evaluate GPT-3.5-turbo, GPT-4o-mini, DeepSeek-V3, and GPT-4o under zero-shot prompting. SemEval and P-Stance use $F_{\text{avg}} = (F1_{\text{FAVOR}} + F1_{\text{AGAINST}})/2$, the official SemEval-style metric that excludes NONE from the average. MTSD and VAST use macro-F1 or accuracy depending on the analysis; for cross-dataset phase analysis we focus on accuracy within SICI regions and separately inspect NONE label composition to avoid inflated high-SICI scores.

We report both instance-level and binned statistics. Instance-level correlations test whether higher SICI predicts correctness at the sample level. Binned correlations test whether the aggregate trend is monotonic after grouping instances into SICI intervals. This distinction is important for stance detection because individual labels are noisy, while regime-level behavior can still be stable. For SemEval, we also report F_{avg} because it is the official metric and prevents a high NONE prior from artificially improving the score.

We also compute two diagnostic quantities. The first is the systematic-bias rate: the proportion of examples where the model predicts AGAINST when the gold label is not AGAINST, plus the proportion where it predicts NONE when the gold label is not NONE. The second is the bias–SICI interaction index (BII). For each SICI bin b, we define

$$\text{BII}_b = \text{Accuracy}_b - \text{MeanConfidence}_b, \quad (2)$$

where MeanConfidence_b is the average maximum predicted label probability in bin b. Values near zero indicate closer agreement between confidence and correctness; negative values indicate overconfidence, whereas positive values indicate underconfidence.

Surface Baselines. To test whether SICI adds information beyond simple proxies, we compare against text length, target visibility as lexical target coverage, type-token ratio, and negation density. These baselines represent common surface-level difficulty hypotheses: longer texts may contain more evidence, low target mention rate may make inference harder, lexical diversity may increase complexity, and negation may confuse polarity-based decisions.

Metric	Sample r	Binned ρ
SICI	**-0.2405**	**-0.9515**
Target visibility	+0.1328	+0.9245
Text length	+0.0648	+0.6991
Type-token ratio	-0.0508	-0.4877
Negation density	-0.0174	+0.0813

Table 3: Correlation with correctness on merged non-NONE SemEval+VAST instances. Negative values mean that higher difficulty predicts lower accuracy.

Interventions. For the high-complexity SemEval subset (SICI ≥ 0.70, $N = 187$), we evaluate 14 inference-time intervention configurations: self-consistency, few-shot prompting, self-reflection, counterfactual reasoning, target decomposition, generated knowledge, multi-agent debate, evidence chaining, SICI-neighbor few-shot retrieval, dimension-targeted routing, Wikipedia RAG, RAG plus scope routing, cultural-camp debate, and cultural-camp debate with RAG. This set spans prompting, retrieval, routing, and debate-style methods.

The intervention suite is designed to distinguish three possible explanations for high-SICI failure. If the bottleneck is unstable decoding, self-consistency should help. If it is missing reasoning, chain-of-thought, reflection, counterfactual analysis, or target decomposition should help. If it is missing information, few-shot retrieval or Wikipedia RAG should help. Failure across all three families would instead support the stronger interpretation that many Phase-3 examples are underdetermined by the available text-target pair.

5 Results

SICI Predicts Accuracy Better Than Surface Proxies. On the merged non-NONE SemEval+VAST set ($N = 1{,}960$), SICI has the strongest and directionally correct relationship with accuracy. At the sample level, the point-biserial correlation between SICI and correctness is $r = -0.2405$ ($p = 3.4 \times 10^{-27}$). At the binned level, the Spearman correlation is $r = -0.9515$ ($p = 2.3 \times 10^{-5}$). Table 3 shows that surface baselines are either weaker or directionally misleading. Target visibility has a high positive binned correlation because explicit targets are easier, but it captures only one dimension and cannot explain high-SICI collapse caused by scope and pragmatic ambiguity.

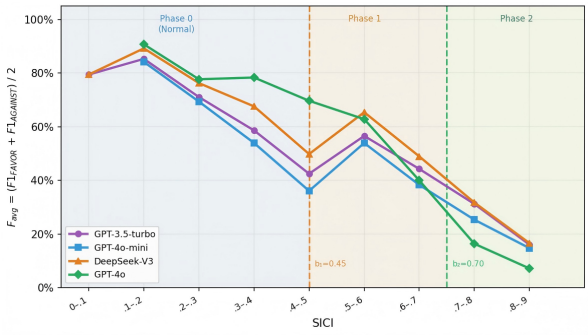
A Dual-Fixation Regime Shift. Figure 1 summarizes the core phenomenon. As SICI increases, model performance decreases, but the more diagnostic signal is the systematic prediction-bias rate. Lower-complexity errors are dominated by AGAINST over-prediction. Around SICI ≈ 0.45, accuracy reaches a trough: AGAINST fixation begins to fail, while NONE escape has not yet become dominant. Above roughly 0.70, NONE predictions rise sharply. A segmented-regression comparison confirms that the three-regime structure fits better than a linear trend ($F(2, 14) = 16.82$, $p = 1.89 \times 10^{-4}$).

The fit improvement is large in absolute terms. On 18 SICI bins, the residual sum of squares drops from 0.2141 for a single linear model to 0.0629 for the two-breakpoint segmented model, a 70.6% reduction. The first breakpoint is stable around 0.45 across GPT-3.5-turbo, GPT-4o-mini, and DeepSeek-V3 (0.425–0.450). The second breakpoint varies more (0.700–0.800), suggesting that models differ mainly in when they begin to rely on NONE escape.

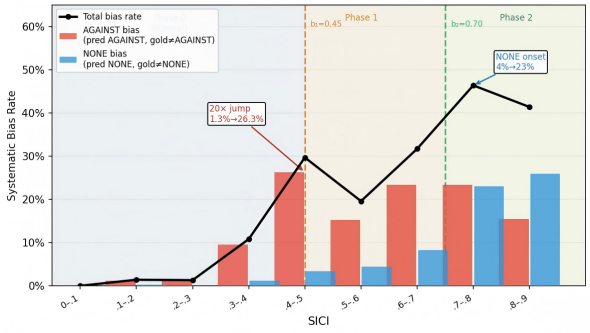
This pattern motivates an attribution–abstention account of the failure. The model is not simply uncertain. It falls back on different heuristics in different regions: first a tendency to treat stance-bearing social media text as opposition, then a tendency to abstain with NONE when the target–text relation becomes indirect.

Table 4 makes the shift concrete. The low-SICI bins contain many AGAINST predictions and high accuracy. The boundary bin at 0.4–0.5 is the weakest point: accuracy drops to 42.3%, while AGAINST remains common and NONE has not yet become the dominant output. After 0.7, the model's output distribution changes qualitatively. NONE predictions rise from 54.5% in the 0.7–0.8 bin to 75.0% above 0.8, while FAVOR nearly disappears. This is why a single monotone "harder means lower accuracy" story is incomplete: the same increasing complexity first produces over-commitment to opposition and then over-abstention.

Figure 2 shows two complementary views. Raw accuracy exhibits a partial rebound after the first boundary because NONE becomes more frequent in some high-SICI bins. This rebound is misleading if interpreted as easier inference: a model can appear more accurate simply because the label prior has shifted toward NONE. The official F_{avg} view and the systematic-bias view avoid this artifact. The former ignores NONE when computing SemEval



(a) Official SemEval F_{avg} across four models.



(b) Systematic prediction-bias rate for GPT-4o-mini.

Figure 1: User-provided main regime-shift figures. As SICI increases, stance prediction quality declines across models, while systematic bias reveals two critical jumps: AGAINST fixation near $b_1 = 0.45$ and NONE escape near $b_2 = 0.70$.

SICI bin	N	%Against	%None	Acc.
0.0–0.1	24	4.2	0.0	87.5
0.1–0.2	599	61.1	1.2	85.8
0.2–0.3	158	61.4	1.9	71.5
0.5–0.6	250	70.0	16.8	57.2
0.6–0.7	337	66.8	29.4	52.2
0.7–0.8	235	43.4	54.5	38.3
0.8–1.0	64	23.4	75.0	37.5

Table 4: SemEval GPT-4o-mini prediction trajectory by SICI bin. Percent columns are model prediction rates, not gold label rates.

Measure	GPT-4o-mini	GPT-4o
Overall macro-F1	∼0.61	0.737
AGAINST peak (0.3–0.4)	81.4	51.4
Accuracy at 0.4–0.5	42.3	70.3
NONE rate at 0.7–0.8	54.5	84.3
First boundary b_1	0.450	0.250
Second boundary b_2	0.700	0.825
SICI–accuracy ρ	-0.243	-0.191

Table 5: Stronger models shift the regime boundaries but preserve a significant relationship between SICI and accuracy. Percent-valued rows omit percent signs for compactness.

quality; the latter directly measures when the model predicts AGAINST or NONE against the gold label. Together, the four views show why the regime-shift claim is not a visual artifact of one metric.

Stronger Models Move Boundaries but Do Not Remove the Regime. GPT-4o substantially improves overall performance compared with GPT-4o-mini, but the SICI relationship remains significant. Table 5 shows that GPT-4o reduces the lower-complexity AGAINST peak and improves the 0.4–0.5 boundary region, but its high-SICI NONE rate is even higher. Thus scale or capability changes the failure profile; it does not erase complexity-conditioned behavior.

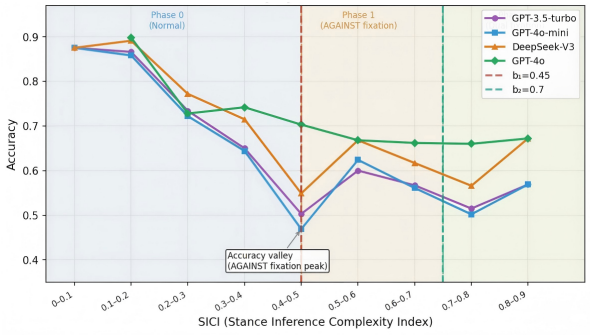
The cross-model pattern also clarifies what "stronger" means in this setting. GPT-4o is better at recovering explicit and moderately implicit stances, but its improvement is not equivalent to solving the inference chain. The first breakpoint moves left, which means the model exits the AGAINST-fixation regime earlier. The second breakpoint moves right, which means the model postpones its strongest NONE escape. Yet the correlation remains negative, and the high-SICI region is still governed by abstention behavior. Capability therefore stretches the phase boundaries rather than flattening the phase diagram.

Phase	N	Accuracy
SICI < 0.45	1,069	0.755
$0.45 \leq$ SICI < 0.70	760	0.605
SICI ≥ 0.70	131	0.323

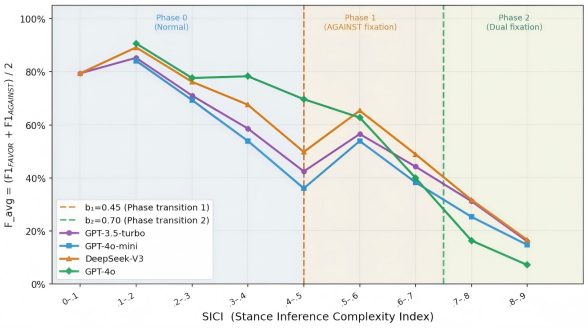
Table 6: Merged SemEval+VAST non-NONE accuracy by SICI phase.

Cross-Dataset Validation Separates Two Kinds of NONE. VAST provides an important contrast. Its high-SICI region has high overall accuracy because the gold labels are mostly NONE: 250 of 288 Phase-3 examples (86.8%). This could appear to contradict the SemEval high-SICI failure pattern. However, after separating non-NONE cases, the difficulty remains: VAST Phase-3 non-NONE accuracy is 0.579, and merged SemEval+VAST non-NONE accuracy drops monotonically from 0.755 to 0.605 to 0.323 across the three SICI phases (Table 6).

This comparison reveals a theoretical distinc-



(a) Accuracy by SICI bin across four models.



(b) Alternative F_{avg} view excluding NONE.

Figure 2: Additional user-provided visualizations of the same regime structure. Accuracy alone can obscure the transition because high-SICI bins often contain more NONE labels; F_{avg} and systematic-bias rate isolate the stance-inference failure more directly.

tion. In SemEval Phase 3, many AGAINST instances are incorrectly mapped to NONE: 68 of 87 AGAINST cases become NONE. In VAST Phase 3, NONE is often a reasonable abstention because the target–text relation is genuinely underspecified. SICI therefore separates two superficially similar outputs: *failed inference fixation* and *information-insufficient abstention*.

Dataset-dependent few-shot behavior. Few-shot degradation is not an intrinsic property of few-shot prompting or a universal consequence of high SICI. On high-SICI SemEval, random few-shot prompting reduces accuracy from 58.3% to 47.1% (-11.2 points). Selecting demonstrations by SICI similarity reaches 49.2%, recovering 2.1 points over random selection but remaining 9.1 points below zero-shot. In contrast, the same random few-shot setting improves VAST accuracy in all three regions, by 1.3, 4.3, and 2.8 points, respectively (Table 7a).

The difference is partly associated with label composition (Table 7b). Only 50.8% of the high-SICI SemEval subset is labeled NONE, and its zero-shot non-NONE accuracy is 21.7%. In VAST, 86.8% of the corresponding subset is labeled NONE, for which abstention is often correct, while non-NONE accuracy is 57.9%. Thus, intervention behavior depends jointly on complexity and dataset label composition. The 2.1-point recovery from SICI-neighbor selection indicates that demonstration choice matters, but similarity in difficulty need not supply missing target-specific or contextual evidence. We therefore interpret the observed degradation as dataset-, demonstration-, and complexity-dependent rather than as an intrinsic failure of few-shot prompting.

Additional dimension-level analysis, reported in Appendix B, shows that target visibility and scope alignment are the strongest individual drivers, while the remaining dimensions help distinguish target sparsity from genuinely complex stance inference.

Interventions under the tested settings. Table 8 reports high-SICI intervention results with 95% Wilson confidence intervals. None of the evaluated interventions exceeds the zero-shot baseline under the tested models, prompts, and single-run settings. Zero-shot and Wikipedia RAG differ by only one correct prediction (109 vs. 108), so their 0.5-point difference should not be interpreted as reliable degradation.

The intervals quantify uncertainty associated with the finite evaluation sample, not sensitivity to prompt wording, decoding, or stochastic model behavior. Results from stochastic prompting and debate are therefore treated as exploratory. Under these conditions, additional reasoning does not help when the missing ingredient is not a reasoning step but an underdetermined pragmatic link between the text and target. Retrieval helps only marginally because the most difficult cases often require conversational, author-specific, or discourse-level context that generic Wikipedia knowledge cannot supply.

The failures also reveal a label-prior pendulum. Some interventions, such as indirect inference and scope clarification, reduce AGAINST errors but over-correct toward NONE. In the original local analysis, indirect stance inference predicts NONE for 232 of 299 triggered cases (77.6%), although the gold NONE rate is only 56%. Scope clarification is more extreme: 142 of 154 triggered cases are predicted as NONE (92.2%), while the

SICI region	N	ZS acc.	FS acc.	Δ
SICI < 0.45	634	71.1	72.4	$+1.3$
$0.45 \leq$ SICI < 0.70	538	76.0	80.3	$+4.3$
SICI ≥ 0.70	288	89.9	92.7	$+2.8$

(a) VAST zero-shot and random three-shot accuracy.

Dataset	N	Gold NONE	ZS acc.	Non-NONE acc.
SemEval	187	50.8	58.3	21.7
VAST	288	86.8	89.9	57.9

(b) Composition of the high-SICI subsets.

Table 7: GPT-4o-mini region-wise and cross-dataset results. Accuracies and gold-label rates are percentages; Δ is the few-shot change over zero-shot in percentage points. Both composition rows use SICI ≥ 0.70.

Method	Acc.	95% CI	Δ
Zero-shot	58.3	$[51.1, 65.1]$	–
Self-consistency	58.3	$[51.1, 65.1]$	0.0
Wikipedia RAG	57.8	$[50.6, 64.6]$	-0.5
Target decomposition	55.1	$[47.9, 62.0]$	-3.2
Dimension-targeted routing	52.9	$[45.8, 60.0]$	-5.4
Random few-shot	47.1	$[40.0, 54.2]$	-11.2
Multi-agent debate	47.1	$[40.0, 54.2]$	-11.2
Cultural-camp debate	31.0	$[24.8, 38.0]$	-27.3

Table 8: GPT-4o-mini accuracy on high-SICI SemEval examples ($N = 187$). CIs are 95% Wilson intervals; Δ is the change in percentage points relative to zero-shot.

gold NONE rate is 45.5%. Debate-style prompting moves in the opposite direction in some variants. Cultural-camp debate predicts FAVOR 105 times even though only five gold examples are FAVOR, and its RAG-augmented variant preserves almost the same skew. These shifts can improve global macro-F1 slightly when they counteract a dataset-level bias, but they do not solve the local high-SICI decision. A useful intervention for stance detection must therefore do more than change a model's prior over labels; it must calibrate the strength of the target–text link.

Appendix B further reports routing and confidence analyses: SICI-guided routes can improve global macro-F1 by shifting label priors, but they do not solve the local high-SICI subset, and high complexity is not reducible to low confidence.

6 Discussion

SICI as a diagnostic rather than a leaderboard metric. SICI should not replace task metrics. Instead, it complements them by identifying where a model's aggregate score comes from. Two models with similar macro-F1 may differ in whether they fail through AGAINST fixation, NONE escape, or boundary-region confusion. Reporting performance stratified by SICI would make stance evaluation more informative.

This diagnostic role is especially important for prompt-based evaluation. A low score can arise because the model lacks task competence, because the prompt induces the wrong label prior, or because the instance itself provides too little target-conditioned evidence. SICI helps separate these cases by making the source of difficulty inspectable: the target may be invisible, the text may be off-scope, the stance may be pragmatic, external knowledge may be required, or the gold label may be intrinsically ambiguous. This turns error analysis from a post-hoc list of mistakes into a structured account of where the target–text inference chain breaks.

Why high-complexity prompting can fail. The intervention results suggest a practical warning for LLM-based stance systems. Prompting strategies often shift label priors rather than improve fine-grained calibration. Indirect-inference and scope-clarification prompts reduce one kind of error but overshoot toward NONE; debate prompts can introduce new hallucinated stances. This explains the "pendulum" effect observed in our experiments: interventions swing the model from one fixation mode to another.

The limitation is therefore not simply that the tested prompts are weak. Many interventions add exactly the resources that prompt engineering usually assumes to be helpful: explicit reasoning, multiple samples, retrieved knowledge, or disagreement among agents. Their limited effect suggests that the hard cases often require information that is not recoverable from the visible text-target pair, or require pragmatic commitments that the model cannot license without over-interpreting the author. In such cases, a better prompt may only choose a different trade-off between false attribution and false abstention.

Toward model-level and context-level solutions. If high-SICI failures are caused by missing pragmatic context or ambiguous target–text relations, inference-time prompting may be insufficient. More promising directions include SICI-stratified training, calibrated abstention policies, retrieval

of conversational context rather than encyclopedic background, and human-in-the-loop treatment of intrinsically ambiguous cases.

Implications for benchmark construction. The analysis also suggests that stance benchmarks should report their complexity distribution. A dataset dominated by low-SICI examples primarily tests direct target matching and sentiment-to-stance mapping. A dataset with many high-SICI examples tests pragmatic inference, context recovery, and abstention calibration. Without this distribution, two datasets with the same label set can evaluate very different skills. Reporting SICI histograms and phase-stratified scores would make cross-dataset comparisons less dependent on hidden differences in target visibility and topic alignment.

The per-target results from the original analysis illustrate this point. In VAST, the *election* target is especially difficult: all evaluated models are near 26–34% macro-F1, close to random-level behavior for a three-way task. By contrast, P-Stance removes the NONE label and produces much higher scores for stronger models. These differences are not just dataset names; they reflect distinct inference regimes. A hard benchmark for stance detection should therefore be constructed from high-SICI non-NONE examples rather than from examples that are merely label-balanced.

Implications for LLM evaluation. For LLMs, the most concerning failures are not always low-confidence errors. In our analysis, high-SICI examples often elicit confident but wrong predictions, especially around the boundary where AGAINST fixation and NONE escape compete. This behavior matters for downstream use: a stance system that returns a single label without exposing instance complexity may be least reliable exactly when its output looks decisive. SICI can therefore serve as a triage signal: low-complexity predictions may be used normally, intermediate cases may require calibration checks, and high-complexity cases should trigger abstention, context retrieval, or human review.

7 Conclusion

We introduced SICI, a semantic-pragmatic complexity index for diagnosing LLM stance detection. Across datasets and models, SICI predicts where LLMs fail and reveals an attribution–abstention regime shift from low-complexity over-prediction to high-complexity None escape. Stronger models move the transition boundaries but do not eliminate the phenomenon, and prompting, retrieval, debate, and voting interventions often shift the model's operating point rather than remove the high-complexity bottleneck. The main implication is that stance detection errors should be analyzed as structured failures of target-conditioned inference rather than as undifferentiated classification mistakes. Future stance systems should therefore combine model improvements with complexity-aware evaluation, calibrated abstention, and richer context acquisition.

Limitations

First, the main SICI scores are LLM-generated. Targeted validation with three human annotators on 200 stratified instances provides initial human grounding, but it is not a large-scale human gold standard and cannot rule out all shared model biases. Second, the experiments focus on English social-media datasets; long-form, multilingual, and multimodal stance settings may exhibit different dimension weights. Third, the high-SICI non-NONE subset is relatively small, especially in VAST, so future work should construct larger hard-instance benchmarks. Finally, our intervention study is inference-time only; it does not test whether supervised fine-tuning on SICI-stratified data can reduce the bottleneck observed under the tested inference-time settings.

Ethical Considerations

Stance detection can be used for beneficial analysis of public discourse, but also for political profiling, surveillance, and manipulation. Our work is diagnostic and does not release a new classifier intended for deployment. We use publicly available benchmark datasets and API-accessible LLMs under their respective terms of use, and we do not redistribute dataset contents or model weights. Because high-SICI instances are often ambiguous or context-dependent, automated decisions on such cases should not be treated as ground truth. Systems using stance predictions in sensitive settings should report uncertainty, allow human review, and avoid inferring personal beliefs from sparse or indirect text.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62306184), the Natural Science Foundation for Top Talents of SZTU (Nos. GDRC202518 and GDRC202320), the Shenzhen Science and Technology Program (Nos. RCBS20231211090548077 and JCYJ20240813113218025), the Guangdong Basic and Applied Basic Research Foundation (2026A1515010133), the Project for Improving Scientific Research Capabilities of Key Construction Disciplines in Guangdong Province (No. 2025ZDJS039), and the Innovative Team Project in Guangdong Province under Grant 2025KCXTD040.

References

Abeer Aldayel and Walid Magdy. 2021. Stance detection on social media: State of the art and trends. *Information Processing & Management*, 58(4):102597.

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 2020, Online, November 16-20, 2020*, pages 8913–8931. 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.

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.

Daijun Ding, Xianghua Fu, Xiaojiang Peng, Xiaomao Fan, Hu Huang, and Bowen Zhang. 2024. Leveraging chain-of-thought to enhance stance detection with prompt-tuning. *Mathematics*, 12(4).

Jiachen Du, Ruifeng Xu, Yulan He, and Lin Gui. 2017. Stance classification with target-specific neural attention. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3988–3994.

Anthony Dubreuil, Antoine Gourru, Christine Largeron, and Amine Trabelsi. 2025. Are stereotypes leading LLMs' zero-shot stance detection ? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31517–31530, Suzhou, China. Association for Computational Linguistics.

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

Yanxu Ji, Jinzhong Ning, Yijia Zhang, Zhi Liu, and Hongfei Lin. 2025. LLM-driven implicit target augmentation and fine-grained contextual modeling for zero-shot and few-shot stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 5872–5884, Suzhou, China. Association for Computational Linguistics.

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

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.

Yupeng Li, Haorui He, Shaonan Wang, Francis C. M. Lau, Lau, and Yunya Song. 2023b. Improved target-specific stance detection on social media platforms by delving into conversation threads. *IEEE Transactions on Computational Social Systems*.

Yupeng Li, Dacheng Wen, Haorui He, Jianxiong Guo, Xuan Ning, and Francis C. M. Lau. 2023c. Contextual target-specific stance detection on twitter: Dataset and method. In *2023 IEEE International Conference on Data Mining (ICDM)*, pages 359–367.

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

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*, pages 3453–3464.

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.

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.

Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. 2020. BERTweet: A pre-trained language model for English tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 9–14. Association for Computational Linguistics.

Maksym Taranukhin, Vered Shwartz, and Evangelos Milios. 2024. Stance reasoner: Zero-shot stance detection on social media with explicit reasoning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15257–15272, Torino, Italia. ELRA and ICCL.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022a. Emergent abilities of large language models. *Transactions on Machine Learning Research*. Survey Certification.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022b. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Penghui Wei, Junjie Lin, and Wenji Mao. 2018. Multi-target stance detection via a dynamic memory-augmented network. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1229–1232. ACM.

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*, pages 1173–1176. ACM.

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, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Bowen Zhang, Xu Huang, Zhichao Huang, Hu Huang, Baoquan Zhang, Xianghua Fu, and Liwen Jing. 2022. Sentiment interpretable logic tensor network for aspect-term sentiment analysis. In *Proceedings of the 29th international conference on computational linguistics*, pages 6705–6714.

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

Bowen Zhang, Min Yang, Xutao Li, Yunming Ye, Xiaofei Xu, and Kuai Dai. 2020. Enhancing cross-target stance detection with transferable semantic-emotion knowledge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3188–3197.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024b. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378, Mexico City, Mexico. Association for Computational Linguistics.

ZhaoDan Zhang, Jin Zhang, Hui Xu, Jiafeng Guo, and Xueqi Cheng. 2025b. 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.

A Additional Experimental Details

Inference-chain interpretation. The seven SICI dimensions can be read as a multi-stage stance inference chain. A model must identify the target (V), decide whether the text's scope bears on that target (S), decode pragmatic cues (P), retrieve needed knowledge (K), integrate missing context (C), resolve label ambiguity (A), and bridge sentiment polarity to stance (G). A severe failure at any stage can derail the final label.

Phase definitions. We use $b_1 = 0.45$ and $b_2 = 0.70$ as empirical transition points, yielding Phase 1 (SICI < 0.45), Phase 2 ($0.45 \leq$ SICI < 0.70), and Phase 3 (SICI ≥ 0.70). These thresholds are selected from the SemEval segmented-regression analysis and then reused for VAST.

Full intervention set. The full Phase-3 intervention suite includes zero-shot, self-consistency, few-shot, self-reflection, counterfactual reasoning, target decomposition, generated knowledge, standard multi-agent debate, evidence chaining, SICI-nearest-neighbor few-shot retrieval, dimension-targeted adaptive prompting, Wikipedia RAG,

RAG plus scope routing, cultural-camp debate, and cultural-camp debate plus RAG. Evidence chaining is excluded from substantive comparisons because a formatting failure caused label parsing to collapse.

B Additional Dimension, Routing, and Confidence Analyses

Dimension-level gradients. Single-dimension analysis on SemEval indicates that target visibility and scope alignment are the strongest drivers. Accuracy drops from 0.909 to 0.607 between $V = 0$ and $V = 4$, and from 0.886 to 0.627 between $S = 0$ and $S = 4$. Their interaction is especially diagnostic: in the SemEval $V = 4, S = 4$ cell ($N = 247$), the model predicts NONE for 85.8% of examples while the gold NONE rate is only 50.2%. In VAST, the same $V = 4, S = 4$ pattern corresponds to a mostly valid NONE abstention: NONE-prediction rate 95.5% and gold NONE rate 96.6%. This suggests that the same inference-chain break can either expose a dataset's legitimate underspecification or reveal a model's erroneous escape behavior.

The remaining dimensions help explain why target visibility alone is insufficient. Pragmatic implicitness captures cases where the target is visible but the stance is conveyed through irony, metaphor, or indirect evaluation. Knowledge requirement captures cases where a stance depends on knowing an event, policy, or social relation. Label ambiguity captures instances where the text may support multiple defensible interpretations. Polarity–stance gap captures the familiar failure of treating sentiment as stance. These dimensions are individually weaker than V and S in our SemEval analysis, but they are essential for distinguishing a merely target-sparse example from a genuinely complex stance inference problem.

Ablation results add a second perspective. Removing S or P changes the binned correlation most strongly, indicating that scope alignment and pragmatic implicitness are central to the monotonic regime signal. Removing G slightly improves the correlation in our current equal-weight setting, which suggests that polarity–stance gap is noisier than the other dimensions on these datasets. We keep G because the construct is theoretically important for stance detection, but the result indicates that future versions of SICI should refine how affective polarity is separated from target-conditioned stance.

Dimension	Low acc.	High acc.	Drop
V: target visibility	0.909	0.607	-0.302
S: scope alignment	0.886	0.627	-0.258
K: knowledge need	0.769	0.549	-0.220
C: context dependence	0.746	0.600	-0.146

Table 9: Representative SemEval single-dimension gradients. "Low" and "high" refer to low- versus high-complexity levels for that dimension; exact sample sizes vary by dimension.

SICI-guided routing. The original experiments also tested staged routing on the full 2,899-instance pool. A simple CoT route triggered on SICI > 0.6 improves global macro-F1 from 0.5916 to 0.6132 (+2.16pp). A more targeted ISI+SC route, which applies indirect stance inference to high-V cases and scope clarification to high-S cases, reaches 0.6209 (+2.93pp). The best full routing policy combines ISI, SC, and CoT on 790 triggered examples (27.2% of the pool), reaching 0.6296 (+3.80pp).

This result prevents an overly pessimistic reading of the intervention study. SICI can guide useful global routing: when the model's dominant error is an overactive AGAINST prior, shifting some examples toward more cautious reasoning can improve macro-F1. However, the same mechanism does not solve the high-SICI subset. Local ISI and SC scores fall below their local baselines because they overshoot toward NONE. Thus SICI is valuable for diagnosis and routing, but routing is not equivalent to resolving the semantic-pragmatic ceiling.

This global-local gap is useful diagnostically. A routing policy can improve the dataset-level score by correcting the most common prior error, but its gains depend on the distribution of phases in the evaluation set. In a deployment setting where high-SICI non-NONE cases are the main concern, the same routing rule could be harmful. We therefore treat routing as an operational use of SICI, not as evidence that the underlying semantic-pragmatic ambiguity has been removed.

Confidence failure. High-SICI examples are not simply low-confidence cases. Using Equation 2, BII is negatively correlated with SICI ($\rho = -0.308$), indicating that the gap between confidence and correctness worsens as complexity rises. The largest concentration of confident errors occurs around SICI $= 0.4$–0.6, where AGAINST fixation has become unreliable but NONE escape

has not yet become a calibrated abstention strategy. In this region, the model can select a wrong label with apparent confidence.

This finding changes how we interpret NONE. A well-calibrated model should reserve NONE for cases where the text does not support a stance toward the target. In the high-SICI SemEval subset, NONE often functions instead as an uncalibrated failure mode: the model gives up on implicit AGAINST examples even when the gold label is not NONE. In VAST, by contrast, high-SICI NONE is often a valid response. Confidence alone cannot distinguish these two cases; the target–text complexity structure is needed.