Adaptive Bi-SADA: Bidirectional Structure-Aware Data Augmentation for Robust Aspect Sentiment Quadruplet Prediction

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved state-of-the-art performance in Aspect Sentiment Quadruplet Prediction (ASQP). However, we argue that this success often relies on superficial positional heuristics rather than robust structural reasoning. In this paper, we propose a Probe-and-Cure framework to scrutinize and enhance LLM robustness. First, we introduce a Multi-Surgical Adversarial Attack protocol as a diagnostic tool. Our study reveals that SOTA models are "Fragile Giants": they suffer severe performance degradation when facing logical distractors and deep syntactic embeddings. To address this, we propose Adaptive Bi-SADA (Bidirectional Structure-Aware Data Augmentation). Unlike uniform augmentation strategies, our method constructs a length-aware curriculum: it applies natural structural hardening to short sentences to prevent overfitting, and syntactic normalization to long sentences to distill core dependencies. We implement a strict generation-time verification protocol to ensure semantic invariance. Experiments on ASQP and ACOS tasks demonstrate that our method not only achieves new SOTA F1 scores but also effectively transforms superficial pattern matching into robust structural reasoning, significantly closing the performance gap under adversarial stress tests.

1 Introduction

Aspect-Based Sentiment Analysis (ABSA) (Liu et al., 2020; Nazir et al., 2020; D'Aniello et al., 2022; Scaria et al., 2023) is a fine-grained sentiment analysis task that aims to identify and understand opinions at the aspect level within a given text. It addresses a key limitation of traditional document-level or sentence-level sentiment analysis (Meena and Prabhakar, 2007) by recognizing that a single sentence or review may contain multiple opinions targeting different entities or attributes.

Aspect Sentiment Quadruplet Prediction (ASQP) (Zhang et al., 2021a,b; Zhou et al., 2023; Kim et al., 2024) represents the most granular and challenging task in ABSA. It aims to extract all sentiment elements—aspect terms, categories, opinion terms, and sentiment polarities—as structured quadruplets from user reviews. Recently, Generative Large Language Models (LLMs) have revolutionized this field, achieving State-of-the-Art (SOTA) performance by formulating ASQP as a sequence-to-sequence generation task (Jian et al., 2025; Hellwig et al., 2025; Jiajian Li and Wang, 2025).

However, the impressive results on standard benchmarks often mask underlying fragilities. Recent studies in NLP robustness suggest that deep models frequently rely on spurious correlations rather than robust semantic reasoning (McCoy et al., 2019a). We hypothesize that current SOTA models effectively act as "Fragile Giants"—powerful in standard settings but prone to collapse under slight structural shifts. They tend to overfit to canonical syntactic structures and superficial positional proximity, lacking the ability to handle linguistic variation.

To validate this hypothesis, we introduce a Multi-Surgical Adversarial Attack framework. Unlike black-box attacks that inject random noise (Hofer et al., 2021), our approach employs linguistically motivated perturbations—such as cleft construction and recursive embedding—to stress-test the model. Our diagnostic study reveals a startling reality: despite high standard performance, baseline models suffer significant performance degradation when subjected to these structural shifts. We identify two primary failure modes: syntactic rigidity and positional bias. First, models exhibit syntactic rigidity when processing informal or fragmented reviews (e.g., "*Great food, bad vibes*"), which lack explicit syntactic dependencies. Second, models rely heavily on positional proximity between aspects and

opinions. Their performance drops notably when the two are separated by embedded clauses or when the sentence structure is inverted, leading to extraction failures.

To address these issues, we propose a probe-and-cure framework featuring Adaptive Bidirectional Structure-Aware Data Augmentation (Bi-SADA). Unlike previous methods that apply uniform augmentation, Bi-SADA constructs a Length-Aware Curriculum that adapts strategies to sentence length. For Short Sentences, which are prone to overfitting, we apply Structural Hardening. This involves rewriting them into complex but natural forms, such as using cleft sentences or passive voice, to force the model to learn deep dependency parsing rather than surface heuristics. For Long Sentences, which suffer from noise, we apply Syntactic Normalization, which simplifies them into standard SVO structures to provide clear attention anchors. Crucially, we implement a strict Generation-Time Verification Protocol to ensure semantic invariance without relying on external filtering models.

Our contributions are summarized as follows:

- We conduct a systematic structural robustness analysis using a novel Multi-Surgical Attack protocol, revealing that standard ASQP models are highly vulnerable to logical distractors and syntactic inversion.
- We propose Adaptive Bi-SADA, a training framework that simultaneously repairs syntactic deficiencies and immunizes models against structural noise through a length-adaptive, bidirectional rewriting mechanism.
- Extensive experiments on ASQP and ACOS tasks demonstrate that our method achieves new SOTA performance. More importantly, it demonstrates superior robustness, maintaining stability under adversarial stress tests where baseline models collapse.

2 Related Work

2.1 Generative Aspect Sentiment Quadruplet Prediction

Aspect Sentiment Quadruplet Prediction (ASQP) aims to extract the comprehensive quadruplet (a, c, o, s) from review texts. Early approaches relied on pipeline methods (Cai et al., 2021) or unified grid tagging schemes (Wu et al., 2020), which often suffered from error propagation or high complexity. Recently, the generative paradigm has become the dominant approach. Zhang et al. (2021a) first reformulated ASQP as a paraphrase generation task using T5 (Raffel et al., 2020), predicting quadruplets as a linearized natural language sequence. Building on this, Hu et al. (2022) proposed Data Augmentation with Layout-Aware Order (DLO) to mitigate the order sensitivity of sequence generation. Gou et al. (2023) further introduced Multi-view Prompting (MvP) to aggregate information from different prompt perspectives.

2.2 Large Language Models for ABSA

With the emergence of Large Language Models (LLMs), recent research has explored their capabilities in fine-grained sentiment analysis. Wang et al. (2024) evaluated ChatGPT's zero-shot performance on ABSA tasks, finding that while it possesses strong reasoning abilities, it still lags behind supervised SOTA models in extracting precise aspect terms. To bridge this gap, Instruction Tuning has been widely adopted. Šmíd et al. (2024) fine-tuned LLaMA (Touvron et al., 2023) on ABSA datasets, demonstrating that smaller, specialized LLMs can outperform larger general-purpose models. Fei et al. (2023) proposed reasoning-enhanced tuning to force LLMs to generate intermediate reasoning steps before predicting sentiment labels. However, most existing works focus on improving standard F1 scores or few-shot capabilities. The *structural robustness* of these instruction-tuned LLMs—specifically their ability to handle complex syntactic variations like long-distance dependencies or passive voice—remains largely underexplored.

2.3 Robustness and Structure-Aware Augmentation

Robustness in NLP is a critical concern (McCoy et al., 2019b). In the context of ABSA, existing augmentation techniques are predominantly lexical-level, such as synonym replacement and mask-filling (Li et al., 2023; Wu et al., 2019). Xing et al. (2020) introduced adversarial training for ABSA, but their perturbations were limited to embedding space noise. While Zhang et al. (2024) introduced adaptivity into data augmentation, their strategy adapts based on model uncertainty or loss magnitudes. In contrast, our Bi-SADA introduces a Linguistically Adaptive curriculum. We define adaptivity based on the syntactic complexity of

System Prompt for Multi-Surgical

You are a linguistics Red Teamer. For each input sentence, generate TWO distinct adversarial versions.

TECHNICAL MENU (Choose one per candidate):

- **Cleft + Contrastive Distractor:** "It is the [Aspect], rather than the [Distractor], that is [Opinion]."
- **Syntactic Inversion:** Change the positional relationship. "[Opinion] is how I found the [Aspect]."
- **Recursive Embedding:** Insert a long parenthetical clause between Aspect and Opinion.

STRICT RULES:

- Meaning must be IDENTICAL.
- Must include ALL 'Keep' terms exactly.
- No 'Yoda-speak' (ungrammatical text).
- Do not change sentiment polarity.

Figure 1: The prompt used to generate the Multi-Surgical Adversarial Test Set. We utilize an LLM agent to enforce linguistic diversity while maintaining strict semantic constraints.

the input itself. By bidirectionally rewriting sentences—hardening short ones to prevent overfitting and normalizing long ones to reduce noise—we explicitly target the model's dependency parsing capability, addressing a blind spot that previous uncertainty-based adaptive methods overlook.

3 Multi-Surgical Adversarial Attack

Standard test sets often fail to expose the structural fragility of ASQP models due to their syntactic homogeneity. To rigorously evaluate whether models have acquired robust semantic reasoning, we propose the Multi-Surgical Adversarial Attack protocol. We employ a Red-Teaming LLM agent to generate linguistically motivated perturbations based on a strict system prompt (see Figure 1). For each sample in the standard test set, we generate distinct adversarial variants that target specific cognitive mechanisms while strictly preserving the original semantic truth conditions.

3.1 Surgical Probe Design

We design three distinct categories of probes to exploit the "shortcut learning" behaviors often observed in neural models.

Contrastive Distractors. First, to challenge the model's Logical Robustness, we introduce Contrastive Distractors. Standard models often rely on bag-of-words co-occurrence, ignoring logical operators. We rewrite simple assertions into cleft constructions that explicitly introduce a negated entity (e.g., "*It is the pizza, rather than the side dishes, that is tasty*"). A robust model must syntactically distinguish the focus from the distractor, whereas a fragile model relying on shallow matching often hallucinates the distractor as a valid aspect.

Recursive Embedding. Second, to test *Working Memory*, we employ Recursive Embedding. We inject contextually relevant but syntactically isolating parenthetical clauses between aspect and opinion terms (e.g., "*The service, which we observed continuously during our visit, remained slow*"). This artificially elongates the dependency path, probing whether the attention mechanism can bridge long-range gaps without suffering from span loss.

Syntactic Inversion. Finally, we target *Positional Bias* via Syntactic Inversion. Models fine-tuned on standard datasets often memorize canonical word orders. We disrupt this using object fronting or passive transformations. A failure in this regime (e.g., polarity flipping) suggests the model relies on absolute positional embeddings rather than true dependency structures.

To ensure the fairness of this diagnostic study, we implement a rigorous filtering pipeline during generation. We discard any adversarial candidate that fails to preserve all ground-truth terms explicitly, introduces new sentiment-bearing adjectives, or violates grammatical fluency constraints. This ensures that any performance drop is attributable solely to the model's structural fragility.

3.2 Pathology Analysis and Motivation

Table 1 presents representative failure cases from our diagnostic study, revealing two critical pathologies in baseline models. As shown in the Contrastive example, the model hallucinates "side dishes" as a positive aspect. This suggests the model fails to parse the logical operator "rather than" and instead relies on superficial lexical matching. In the Inversion case, simply moving the opin-

Attack Type	Original Input (Correctly Predicted)	Adversarial Input	Model Error Type
Contrastive	*The pizza is tasty.*	*It is the pizza, rather than the side dishes, that is tasty.*	**Hallucination**: *Extracts 'side dishes' as aspect.*
Embedding	*The service is slow.*	*The service, if I recall correctly from my visit, is slow.*	**Span Loss**: *Fails to extract 'service'.*
Inversion	*I liked the atmosphere.*	*The atmosphere, to be honest, was what I liked.*	**Polarity Flip**: *Predicts 'Neutral' instead of 'Positive'.*

Table 1: Case studies of successful Multi-Surgical Attacks. The baseline model, while correct on the original samples, fails under structural perturbations. The specific error types reveal the model's reliance on superficial heuristics.

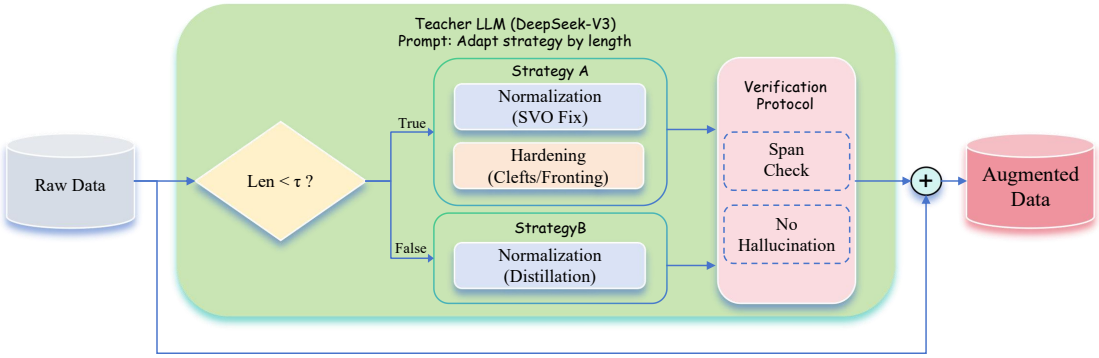

Figure 2: Overview of Adaptive Bi-SADA. Samples flow from left to right. Short sentences (len $< \tau$) undergo bidirectional expansion (Strategy A), while long sentences undergo normalization (Strategy B). A rigorous verification gate ensures semantic integrity before the student model fine-tuning.

ion to the front causes a polarity flip. This indicates that the model's decision boundary is heavily anchored to absolute positions rather than semantic dependencies.

These findings highlight a critical limitation that standard instruction tuning fails to equip models with the structural flexibility needed for complex real-world logic. To address these specific vulnerabilities—namely Syntactic Rigidity and Positional Bias—we propose a targeted curriculum learning framework in the following section.

4 Adaptive Bi-SADA

To address the specific pathologies identified in our diagnostic study—namely positional overfitting in simple structures and attention decay in complex ones—we propose Adaptive Bi-SADA (Bidirectional Structure-Aware Data Augmentation), as shown in Figure 2. Rather than applying a uniform perturbation strategy, Bi-SADA constructs a length-aware curriculum that optimizes the training distribution by treating short and long samples as distinct pedagogical targets.

4.1 Curriculum Formulation

Let D_{train} be the training set. We partition samples using a length threshold τ (empirically $\tau = 15$). As illustrated in Figure 3, the transformation function $\mathcal{T}$ is adaptive:

$$\mathcal{T}(x) = \begin{cases} \{\mathcal{T}_{hard}(x), \mathcal{T}_{norm}(x)\}, & \text{if len}(x) < \tau \\ \{\mathcal{T}_{norm}(x)\}, & \text{if len}(x) \geq \tau \end{cases} \quad (1)$$

For short sentences, we apply a bidirectional expansion to enrich structural diversity. For long sentences, we strictly prohibit complication ($\mathcal{T}_{hard} \to \emptyset$) to prevent generating unreadable text, applying only normalization.

4.2 Direction I: Natural Structural Hardening

We apply $\mathcal{T}_{hard}$ primarily to short samples to counter Positional Bias. Unlike the adversarial probes in Section 3 which aim to break the model, $\mathcal{T}_{hard}$ focuses on teaching Natural Linguistic Generalization. To achieve this, we instruct the Teacher LLM to employ diverse but fluent syntactic structures. Specifically, we implement Focus Shifting

System Prompt for Adaptive Augmentation

You are an expert Data Augmenter. Adapt your output based on the input sentence length ($\tau = 15$).

GLOBAL CONSTRAINTS:

- Hard Span Constraint: Keep [Aspect] and [Opinion] exact.
- No Hallucination: Do NOT add new adjectives.

STRATEGY A: For "SHORT" Inputs ($< \tau$)

- **Goal:** Bidirectional Expansion ($1 \rightarrow 2$).
- **Output 1 (Hardening):** Use Clefts/Fronting to disrupt order.
- **Output 2 (Normalization):** Convert fragments to SVO.

STRATEGY B: For "LONG" Inputs ($\geq \tau$)

- **Goal:** Unidirectional Distillation ($1 \rightarrow 1$).
- **Output 1 (Normalization):** Simplify structure; remove fillers.
- **Output 2 (Hardening):** RETURN EMPTY (Avoid complexity explosion).

Figure 3: The adaptive prompt used for Bi-SADA. The curriculum dynamically adjusts the augmentation intensity: short samples undergo bidirectional rewriting (Hardening + Normalization) to prevent overfitting, while long samples undergo only Normalization to reduce noise.

via cleft constructions (e.g., "*It is the service that is slow*"), which forces the model to resolve hierarchical dependencies rather than linear sequences. Furthermore, we utilize Benign Embedding by inserting natural connective phrases into dependency paths; this trains the attention mechanism to bridge gaps, explicitly countering the Short-Range Myopia pathology without compromising sentence fluency.

4.3 Direction II: Syntactic Normalization

We apply $\mathcal{T}_{norm}$ to both short fragments and long run-on sentences to address *Syntactic Rigidity*. Real-world reviews are often colloquial and fragmented. This operator transforms inputs into standard Subject-Verb-Object (SVO) structures (e.g., converting "*Great food, bad service*" to "*The food is great, whereas the service is bad*"). This provides the model with clean dependency anchors, stabilizing the learning of core sentiment semantics before exposing the model to complex variations.

Split	ASQP Task				ACOS Task			
	Rest-15		Rest-16		Laptop		Rest	
	S	Q	S	Q	S	Q	S	Q
Train	834	1,354	1,264	1,264	2,934	4,172	1,530	2,484
Dev	209	347	316	507	326	440	171	261
Test	537	795	544	799	816	1,161	583	916
Train_{aug}	2,074	3,279	3,124	4,737	8,526	12,053	3,760	5,823
Test_{adv}	1,024	1,494	1,019	1,479	1,567	2,216	1,084	1,647

Table 2: Statistics of datasets. S and Q denote the count of sentences and quadruplets. The gray rows indicate the augmented training set and adversarial test set generated in this work.

4.4 Generation-Time Verification

To ensure the integrity of our augmented data without external filtering models, we enforce a rigorous Generation-Time Protocol. As shown in the "Global Constraints" of Figure 3, we enforce Hard Span Preservation (ground-truth terms must persist) and Null-Hallucination Constraints directly within the generation pipeline. Any candidate violating these constraints is discarded immediately. This ensures that $\mathcal{T}(x)$ strictly preserves the semantic manifold of the original sample x, preventing the label noise issues common in previous augmentation works.

5 Experiments

5.1 Experimental Setup

Datasets and Protocols. We conduct comprehensive experiments on four standard benchmarks covering two subtasks of aspect-based sentiment analysis. For the ASQP task, we use Rest-15 and Rest-16 (Zhang et al., 2021a), which focus on restaurant reviews. For the ACOS task, we utilize Laptop and Rest-ACOS (Cai et al., 2021), which present higher linguistic diversity and domain-specific terminology. To rigorously evaluate model robustness, we construct two specific data splits beyond the standard ones. First, we generate an Augmented Training Set ($D_{train\text{-}aug}$) via Adaptive Bi-SADA, expanding the training data by approximately $2.5\times$ to enrich structural diversity, we explicitly retain the original samples to ensure the model maintains proficiency in standard syntactic patterns while learning from the hardened curriculum. Second, we construct an Adversarial Test Set ($D_{test\text{-}adv}$)

Method	ASQP Task						ACOS Task					
	Rest-15			Rest-16			Laptop			Rest		
	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1	Pre	Rec	F1
LLM Baselines (Instruction-Tuned)												
Qwen-2.5-7B	52.22	53.41	52.81	63.40	64.20	63.80	44.82	42.65	43.71	64.43	60.83	62.58
Llama-3.1-8B	51.92	52.90	52.41	60.27	62.31	61.27	46.13	44.29	45.19	63.77	63.35	63.56
Mistral-v0.3-7B	55.15	52.78	53.94	59.67	62.81	61.20	46.10	44.46	45.27	63.96	63.68	63.82
Generative Baselines												
MvP	-	-	51.04	-	-	60.39	-	-	43.92	-	-	61.54
DOT	-	-	51.91	-	-	61.24	-	-	44.92	-	-	59.25
SimRP	53.12	53.50	53.30	62.74	64.12	63.42	-	-	-	-	-	-
IVLS	54.46	48.53	51.28	62.69	59.75	61.04	43.47	33.45	43.71	60.46	51.14	55.25
ILO	47.78	50.38	49.05	57.58	61.17	59.32	44.14	44.56	44.35	58.43	58.95	58.69
MUL	49.12	50.39	49.75	59.24	61.75	60.47	44.38	43.65	44.01	61.22	59.87	60.53
E4L	54.12	55.35	54.73	63.98	64.46	64.21	45.38	**46.08**	45.73	65.46	64.37	64.91
Ours (Adaptive Bi-SADA)												
+ Qwen-2.5	**56.29**	**56.69**	**56.44**	**64.91**	66.46	**65.67**	47.62	44.98	46.26	**68.02**	**66.08**	**67.04**
+ Llama-3.1	54.62	55.93	55.27	62.12	64.07	63.08	46.45	44.72	45.57	65.74	64.44	65.08
+ Mistral-v0.3	54.56	55.93	55.24	62.95	**67.46**	65.13	**48.05**	45.85	**46.92**	66.26	64.88	65.56

Table 3: Main results on standard benchmarks. Red and pink backgrounds denote the best and second-best F1 scores, respectively. Bi-SADA universally boosts the performance of all three LLM backbones, achieving SOTA across all datasets.

using the Multi-Surgical Attack protocol, applying one-to-many perturbations to create a hardened evaluation suite roughly $2\times$ the size of the original, we intentionally exclude the original test samples to prevent performance dilution. Detailed statistics for all splits, including the number of sentences (S) and quadruplets (Q), are summarized in Table 2.

Backbones and Baselines. To verify the model-agnostic effectiveness of our approach, we verify our method across three representative instruction-tuned LLMs: Qwen-2.5-7B-Instruct (Qwen et al., 2025), Llama-3.1-8B-Instruct (Touvron et al., 2023), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). These models were selected for their superior instruction-following capabilities and diverse architectural traits. For comparative analysis, we include strong generative baselines such as DOT (Jun and Lee, 2025), SimRP (Jian et al., 2025), IVLS (Nie et al., 2024), ILO (Hu et al., 2022), MUL (Hu et al., 2023), E4L (Lai et al., 2025), and MvP (Gou et al., 2023), which represent the previous state-of-the-art in supervised ASQP.

Implementation Details. We implement all experiments using PyTorch on a single NVIDIA A800 (80GB) GPU. Parameter efficiency is achieved via Low-Rank Adaptation (LoRA) applied to all linear projection modules, with rank $r = 64$, $\alpha = 128$, and dropout set to 0.05. We utilize the AdamW optimizer with a learning rate of $2e^{-4}$ and a cosine scheduler with a 0.03 warmup ratio. The global batch size is set to 32, and the maximum sequence length is fixed at 1024 tokens. For the data augmentation phase, DeepSeek-V3 (DeepSeek-AI et al., 2025) serves as the teacher model with a generation temperature of 0.7.

Evaluation Metrics. We report the F1 Score based on exact quadruplet matching. For robustness analysis, we additionally report the F1 Drop Rate (Δ), defined as $\Delta = (F1_{std} - F1_{adv})/F1_{std}$. A lower Δ indicates higher structural resilience.

5.2 Main Results

Table 3 reports the comparative performance of our Adaptive Bi-SADA framework against competitive baselines. We apply our method to three representative backbones: Qwen-2.5, Llama-3.1, and Mistral-v0.3. The results indicate three key findings.

Universal Effectiveness across Backbones. Adaptive Bi-SADA consistently boosts the performance of all three LLMs, demonstrating that our curriculum learning strategy is model-agnostic. Comparing the "Base" models with their "Bi-SADA" enhanced counterparts, we observe significant F1 gains. For instance, Qwen-2.5 sees an improvement of **+3.63%** on Rest-15 and **+4.46%** on Rest (ACOS). Similarly, Mistral-v0.3 gains **+3.93%** on Rest-16. This confirms that explicitly

Method	Rest-15		Rest-16		Laptop		Rest	
	D_{std}	D_{adv}	D_{std}	D_{adv}	D_{std}	D_{adv}	D_{std}	D_{adv}
Llama-3.1-8B-Instruct								
Base	52.41	50.78 (-1.63)	61.27	57.70 (-3.57)	45.19	41.86 (-3.33)	63.56	56.79 (-6.77)
+ Bi-SADA	**55.27**	**54.20** ($\mathbf{-1.07}$)	**63.08**	**60.05** ($\mathbf{-3.03}$)	**45.57**	**43.54** ($\mathbf{-2.03}$)	**65.08**	**59.07** ($\mathbf{-6.01}$)
Qwen-2.5-7B-Instruct								
Base	52.81	46.99 (-5.82)	63.80	55.67 (-8.13)	43.71	41.39 (-2.32)	62.58	54.30 (-8.28)
+ Bi-SADA	**56.44**	**51.86** ($\mathbf{-4.58}$)	**65.67**	**59.30** ($\mathbf{-6.37}$)	**46.26**	**43.69** (-2.57)	**67.04**	**59.43** ($\mathbf{-7.61}$)
Mistral-7B-Instruct-v0.3								
Base	53.94	51.44 (-2.50)	61.20	57.35 (-3.85)	45.27	42.05 (-3.22)	63.82	57.16 (-6.66)
+ Bi-SADA	**55.24**	**55.10** ($\mathbf{-0.14}$)	**65.13**	**63.63** ($\mathbf{-1.50}$)	**46.92**	**44.34** ($\mathbf{-2.58}$)	**65.56**	**59.57** ($\mathbf{-5.99}$)

Table 4: Robustness analysis on Standard (D_{std}) vs. Adversarial (D_{adv}) test sets. Values in parentheses denote the Performance Gap ($D_{std} - D_{adv}$). A smaller gap (closer to 0) indicates higher structural resilience. Bi-SADA consistently reduces this gap across most datasets and backbones.

teaching structural normalization and hardening unlocks the potential of instruction-tuned models better than standard fine-tuning alone.

State-of-the-Art Performance. Our method achieves new SOTA results on all four datasets. As highlighted in Table 3, the top-2 performing models (marked with dark and light gray backgrounds) are consistently Bi-SADA variants. Notably, Bi-SADA (Qwen) dominates the restaurant-domain datasets (Rest-15, Rest-16, Rest-ACOS), outperforming the previous best generative baseline E4L by margins of 1.71%, 1.46%, and 2.13% respectively. This suggests that Qwen's strong multilingual alignment capability, combined with our structural augmentation, yields the most precise sentiment extraction.

Superiority in Complex Domains. On the *Laptop* dataset, which is known for its technical terminology and diverse syntactic structures, Bi-SADA (Mistral) achieves the highest F1 score of **46.92%**, surpassing both Qwen (46.26%) and the strong baseline E4L (45.73%). We hypothesize that Mistral's sliding window attention mechanism may provide an advantage in capturing the local syntactic nuances typical of technical reviews. Nevertheless, both Qwen and Mistral variants of our method outperform all previous baselines, validating that structural hardening is particularly effective for domains with high linguistic variance.

5.3 Robustness Analysis

To verify whether the performance gains stem from genuine structural robustness rather than superficial pattern matching, we subject all models to the Multi-Surgical Adversarial Test Set (D_{adv}). Table 4 compares the performance on standard data (D_{std}) versus adversarial data (D_{adv}), with the Performance Gap ($D_{std} - D_{adv}$) shown in parentheses.

Mitigating Structural Collapse. Standard Instruction-Tuned models exhibit significant fragility when facing structural perturbations. For instance, the Qwen-2.5 baseline suffers a drastic F1 drop of 8.13 points on Rest-16 and 8.28 points on Rest-ACOS. This confirms our diagnostic hypothesis: standard models overfit to canonical structures and struggle with logical distractors. In contrast, Adaptive Bi-SADA acts as an effective immunization strategy. It universally improves absolute adversarial performance (e.g., +3.63 F1 for Qwen on Rest-16) while simultaneously narrowing the performance gap. For Qwen on Rest-16, the gap shrinks from 8.13 to 6.37, indicating that a larger portion of the model's capability is structurally robust.

The Stability of Mistral. A remarkable finding is the resilience of the Mistral-v0.3 + Bi-SADA model. On the Rest-15 dataset, it achieves near-perfect stability with a negligible performance gap of only **0.14 points** (55.24 vs. 55.10). Similarly, on Rest-16, it reduces the gap from 3.85 (Base) to 1.50 (Ours). This suggests that Mistral's architecture, combined with our structural hardening curriculum, learns to decouple semantic sentiment extraction from syntactic surface forms more effectively than other backbones.

Robustness in Diverse Domains. On the ACOS datasets (Laptop and Rest), which feature higher linguistic complexity, Bi-SADA continues to defend against attacks. For Llama-3.1 on Laptop, the performance drop is reduced from 3.33 to 2.03.

Method	Rest-15		Rest-16		Laptop		Rest	
	D_{std}	D_{adv}	D_{std}	D_{adv}	D_{std}	D_{adv}	D_{std}	D_{adv}
Qwen-2.5 (Base)	52.81	46.99 (-5.82)	63.80	55.67 (-8.13)	43.71	41.39 (-2.32)	62.58	54.30 (-8.28)
w/o Normalization (*Hard*)	54.99	51.75 (-3.24)	64.63	58.34 (-6.29)	44.59	42.18 (-2.41)	64.20	**59.88** (-4.32)
w/o Hardening (*Normal*)	53.46	47.71 (-5.75)	64.17	58.46 (-5.71)	44.52	40.94 (-3.58)	64.32	57.06 (-7.26)
Adaptive Bi-SADA	**56.44**	**51.86** (-4.58)	**65.67**	**59.30** (-6.37)	**46.26**	**43.69** (-2.57)	**67.04**	59.43 (-7.61)

Table 5: Ablation study of different augmentation components. "w/o Normalization" refers to using only hardened samples, while "w/o Hardening" uses only normalized samples. The values in parentheses denote the performance gap (Δ). Bi-SADA achieves the best overall performance by synergizing both strategies.

These consistent improvements across three distinct LLM families (Llama, Qwen, Mistral) and four datasets demonstrate that Adaptive Bi-SADA provides a model-agnostic enhancement to structural reasoning capabilities.

5.4 Ablation Study

To disentangle the contributions of the Syntactic Normalization and Natural Structural Hardening components, we conducted an ablation study across all four datasets using the Qwen-2.5 backbone. We compare the full Adaptive Bi-SADA framework with two variants: w/o Hardening (Normal-only), where the model is trained only on original and normalized (simplified) data; and w/o Normalization (Hard-only), where the model sees only original and hardened (complex) samples. Table 5 summarizes the results, yielding three critical insights.

Hardening Drives Robustness. Removing the hardening component leads to a significant decline in adversarial performance (D_{adv}). For instance, on the Rest-15 dataset, the adversarial F1 drops from 51.86 (Bi-SADA) to 47.71 (Normal-only), and the performance gap widens to 5.75. On Laptop, the Normal-only model (40.94) performs even worse than the Base model (41.39) under attack. This confirms that exposing the model to complex, inverted, and embedded structures during training is non-negotiable for defending against structural perturbations.

Normalization Boosts Standard Accuracy. While the Hard-only variant achieves strong robustness (small gaps), it consistently lags behind Bi-SADA in standard F1 (D_{std}). On the Rest dataset, Bi-SADA achieves 67.04 on D_{std} compared to 64.20 for Hard-only. This suggests that applying normalization to long, fragmented sentences helps the model distill core semantic patterns, preventing it from getting lost in the noise of real-world reviews.

The Synergistic Effect. The full Adaptive Bi-SADA achieves the highest D_{std} scores across all datasets while maintaining excellent robustness. Although the Hard-only variant occasionally yields a slightly higher adversarial score (e.g., 59.88 vs. 59.43 on Rest), it comes at the cost of a significant drop in standard accuracy (-2.84). Bi-SADA strikes the optimal balance, demonstrating that our length-aware curriculum—hardening simple samples while normalizing complex ones—effectively synergizes the benefits of both strategies.

6 Conclusion

In this work, we moved beyond standard performance metrics to scrutinize the structural robustness of Large Language Models in fine-grained sentiment analysis. Our diagnostic study, enabled by the Multi-Surgical Adversarial Attack, exposed a critical vulnerability: even powerful instruction-tuned models often rely on superficial positional heuristics ("Fragile Giants"), crumbling under logical distractors and syntactic shifts. To bridge this gap, we proposed Adaptive Bi-SADA, a curriculum learning framework that bidirectionally normalizes and hardens training samples based on their syntactic complexity. By treating short and long sentences as distinct pedagogical targets, we successfully combined the benefits of structural immunization and syntactic distillation. Extensive experiments across three LLM backbones (Qwen-2.5, Llama-3.1, Mistral-v0.3) and four benchmarks verify that our method achieves a dual victory: setting new State-of-the-Art F1 scores while significantly reducing the performance degradation under adversarial stress. For future work, we plan to extend this adaptive curriculum to cross-domain transfer scenarios and explore whether Chain-of-Thought reasoning can further enhance structural resilience against logical traps.

7 Limitations

Despite the promising results, our work has several limitations that invite future investigation.

Our Adaptive Bi-SADA framework relies on a Teacher LLM (DeepSeek-V3) to generate syntactically diverse samples. While we implemented a strict generation-time verification protocol to filter hallucinations, the quality of the augmented data is ultimately upper-bounded by the linguistic capability and instruction-following accuracy of the teacher model. Applying our method with smaller or less capable teacher models might yield suboptimal results.

The curriculum design relies on the length threshold τ (set to 15) to distinguish between "short" and "long" sentences. This value was determined empirically based on the distribution of the Rest15/16 datasets. In practice, the optimal threshold may vary across different domains where sentence structures are inherently longer or more complex. A dynamic or learnable thresholding mechanism remains to be explored.

Our experiments are currently limited to English datasets in the restaurant and laptop domains. While the ACOS task introduces some complexity, we have not verified the effectiveness of Bi-SADA on low-resource languages or highly specialized domains where syntactic structures differ significantly from standard English. Future work will assess the cross-lingual and cross-domain generalization of our structure-aware curriculum.

References

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 340–350.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Giuseppe D'Aniello, Matteo Gaeta, and Ilaria La Rocca. 2022. Knowmis-absa: an overview and a reference model for applications of sentiment analysis and aspect-based sentiment analysis. *Artificial Intelligence Review*, 55(7):5543–5574.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*.

Zhibin Gou, Qingyan Guo, and Yujiu Yang. 2023. MvP: Multi-view prompting improves aspect sentiment tuple prediction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4380–4397, Toronto, Canada. Association for Computational Linguistics.

Nils Constantin Hellwig, Jakob Fehle, Udo Kruschwitz, and Christian Wolff. 2025. Do we still need human annotators? prompting large language models for aspect sentiment quad prediction. *Preprint*, arXiv:2502.13044.

Nora Hofer, Pascal Schöttle, Alexander Rietzler, and Sebastian Stabinger. 2021. Adversarial examples against a bert absa model – fooling bert with l33t, misspellign, and punctuation,. In *Proceedings of the 16th International Conference on Availability, Reliability and Security*, ARES '21, New York, NY, USA. Association for Computing Machinery.

Mengting Hu, Yinhao Bai, Yike Wu, Zhen Zhang, Liqi Zhang, Hang Gao, Shiwan Zhao, and Minlie Huang. 2023. Uncertainty-aware unlikelihood learning improves generative aspect sentiment quad prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13481–13494. Association for Computational Linguistics.

Mengting Hu, Yike Wu, Hang Gao, Yinhao Bai, and Shiwan Zhao. 2022. Improving aspect sentiment quad prediction via template-order data augmentation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7889–7900.

Yancheng Wang Qingqiang Wu Jiajian Li, Zhongquan Jian and Meihong Wang. 2025. Representative chain-of-reasoning framework for aspect sentiment quad prediction. In *Proceedings of the 21st International Conference on Intelligent Computing (ICIC 2025)*, pages 1021–1038, Ningbo, China.

Zhongquan Jian, Yanhao Chen, Jiajian Li, Shaopan Wang, Xiangjian Zeng, Junfeng Yao, Xinying An, and Qingqiang Wu. 2025. Simrp: Syntactic and semantic similarity retrieval prompting enhances aspect sentiment quad prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 23, pages 24248–24256.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Yonghyun Jun and Hwanhee Lee. 2025. Dynamic order template prediction for generative aspect-based sentiment analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 614–626.

Jieyong Kim, Ryang Heo, Yongsik Seo, SeongKu Kang, Jinyoung Yeo, and Dongha Lee. 2024. Self-consistent reasoning-based aspect-sentiment quad prediction with extract-then-assign strategy. In *Findings of the Association for Computational Linguistics ACL 2024*, page 7295–7303. Association for Computational Linguistics.

Wenna Lai, Haoran Xie, Guandong Xu, Qing Li, and S. Joe Qin. 2025. Listwise preference optimization with element-wise confusions for aspect sentiment quad prediction. *Preprint*, arXiv:2511.23184.

Guangmin Li, Hui Wang, Yi Ding, Kangan Zhou, and Xiaowei Yan. 2023. Data augmentation for aspect-based sentiment analysis. *International Journal of Machine Learning and Cybernetics*, 14(1):125–133.

Haoyue Liu, Ishani Chatterjee, MengChu Zhou, Xiaoyu Sean Lu, and Abdullah Abusorrah. 2020. Aspect-based sentiment analysis: A survey of deep learning methods. *IEEE Transactions on Computational Social Systems*, 7(6):1358–1375.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019a. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Annual Meeting of the Association for Computational Linguistics*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019b. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. *arXiv preprint arXiv:1902.01007*.

Arun Meena and Tadinada Vankata Prabhakar. 2007. Sentence level sentiment analysis in the presence of conjuncts using linguistic analysis. In *European conference on information retrieval*, pages 573–580. Springer.

Ambreen Nazir, Yuan Rao, Lianwei Wu, and Ling Sun. 2020. Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing*, 13(2):845–863.

Yu Nie, Jianming Fu, Yilai Zhang, and Chao Li. 2024. Modeling implicit variable and latent structure for aspect-based sentiment quadruple extraction. *Neurocomputing*, page 127642.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Kevin Scaria, Himanshu Gupta, Saurabh Arjun Sawant, Swaroop Mishra, and Chitta Baral. 2023. Instructabsa: Instruction learning for aspect based sentiment analysis. *arXiv preprint arXiv:2302.08624*.

Jakub Šmíd, Pavel Přibáň, and Pavel Kral. 2024. Llama-based models for aspect-based sentiment analysis. In *Proceedings of the 14th workshop on computational approaches to subjectivity, sentiment, & social media analysis*, pages 63–70.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is chatgpt a good sentiment analyzer? a preliminary study. *Preprint*, arXiv:2304.04339.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International conference on computational science*, pages 84–95. Springer.

Zhen Wu, Chengcan Ying, Fei Zhao, Zhifang Fan, Xinyu Dai, and Rui Xia. 2020. Grid tagging scheme for aspect-oriented fine-grained opinion extraction. *arXiv preprint arXiv:2010.04640*.

Xiaoyu Xing, Zhijing Jin, Di Jin, Bingning Wang, Qi Zhang, and Xuan-Jing Huang. 2020. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3594–3605.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021a. Aspect sentiment quad prediction as paraphrase generation. *arXiv preprint arXiv:2110.00796*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021b. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510.

Wenyuan Zhang, Xinghua Zhang, Shiyao Cui, Kun Huang, Xuebin Wang, and Tingwen Liu. 2024. Adaptive data augmentation for aspect sentiment quad prediction. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11176–11180.

Junxian Zhou, Haiqin Yang, Yuxuan He, Hao Mou, and Junbo Yang. 2023. A unified one-step solution for aspect sentiment quad prediction. *Preprint*, arXiv:2306.04152.