

FACT-SR: Faithfulness-Aware Multi-Task Learning for Aspect-Based Sentiment Reasoning

shirong cheng

National University of Defense Technology

Bei Wu

wubeibetsy@gmail.com

National University of Defense Technology

Xin Zhang

National University of Defense Technology

Yahui Wu

National University of Defense Technology

Qika Lin

National University of Defense Technology

Research Article

Keywords: Aspect-based sentiment analysis, Sentiment reasoning, Rationale faithfulness, Multi-task learning, Adaptive loss weighting

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11080269/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

FACT-SR: Faithfulness-Aware Multi-Task Learning for Aspect-Based Sentiment Reasoning

Abstract

Aspect-based sentiment analysis (ABSA) is a core task in fine-grained opinion mining. Recent studies improve prediction transparency by training student models on teacher-generated rationales. Incoherent, hallucinated, or misattributed rationales can mislead lightweight student models and weaken the reliability of sentiment reasoning. To this end, we propose **FACT-SR**, a scale-invariant multi-task framework for **F**aithful **A**spe**CT**-based **S**entiment **R**easoning. FACT-SR first constructs rationale supervision through label-consistency verification and aspect-conditioned sufficiency filtering. It then jointly trains a student model for rationale generation and polarity prediction, with three faithfulness constraints targeting logical entailment, factual grounding, and aspect exclusivity. To balance these heterogeneous objectives, we introduce Scale-Invariant Automatic Weight Learning (SI-AWL), which combines exponential moving average (EMA) loss normalization with bounded uncertainty mapping. Experiments on the SemEval-2014 Restaurant and Laptop benchmarks show that FACT-SR achieves the strongest classification results among the compared systems at both Flan-T5 model scales, together with improved automatic faithfulness metrics. The 250M model improves macro-F1 over vanilla supervised fine-tuning (SFT) by 6.58 and 2.33 percentage points on Restaurant and Laptop, respectively. A matched Qwen2.5-7B-Instruct comparison further demonstrates the effectiveness of the constraints on a decoder-only backbone.

Keywords: Aspect-based sentiment analysis, Sentiment reasoning, Rationale faithfulness, Multi-task learning, Adaptive loss weighting

1 Introduction

Aspect-based sentiment analysis (ABSA) aims to identify the fine-grained sentiment polarity toward specific aspect terms in text and is a core task in opinion mining (Peng et al. 2020; Zhang et al. 2023). As ABSA is increasingly deployed in decision-sensitive domains, model transparency has become important. Recent studies have shifted from merely predicting sentiment labels to generating reasoning-based explanations. Concurrently, chain-of-thought prompting has enabled large language models (LLMs) to produce structured reasoning for such tasks (Wei et al. 2022).

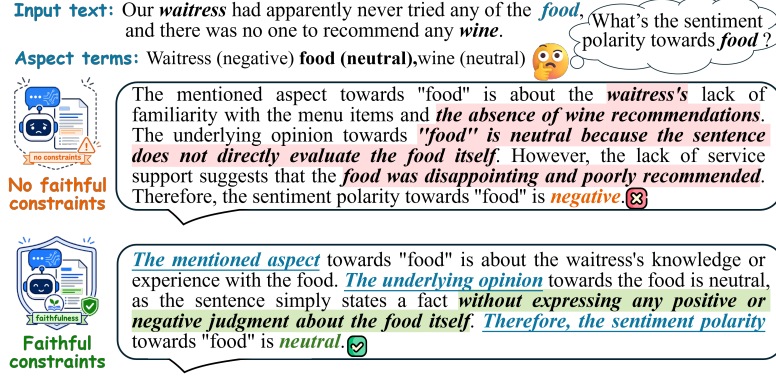


Fig. 1 Illustration of faithful and unfaithful sentiment reasoning. The unfaithful rationale confuses aspect evidence, introduces unsupported information, and does not entail its prediction. The faithful rationale supports the correct polarity with target-specific evidence

However, closed-source models incur substantial costs, and large open-source models demand heavy computation. This motivates distilling structured reasoning supervision from LLM teachers into lightweight models (Ho et al. 2023; Li et al. 2022). Existing reasoning-based methods, such as the three-hop framework, decompose sentiment prediction into aspect identification (Hop-1), opinion inference (Hop-2), and polarity judgment (Hop-3) (Fei et al. 2023). Subsequent verification-based training has shown benefits for classification (Lai et al. 2025). However, a student’s ability to reproduce a rationale does not by itself establish that the rationale is supported by the input.

Meanwhile, these methods mainly optimize final classification accuracy, with limited attention to rationale faithfulness. The benchmark datasets mainly provide aspect-level polarity labels and rarely contain rationales. Although LLMs can generate candidate rationales at scale, plausible explanations are not necessarily faithful to the input or the model decision (Jacovi and Goldberg 2020). Therefore, constructing faithful rationale supervision is a prerequisite for training lightweight models to perform reliable sentiment reasoning.

By examining LLM-generated reasoning chains, we further observe three recurring faithfulness defects. First, *logical disconnection*: the preceding reasoning steps do not logically entail the final polarity (Lyu et al. 2023). Second, *factual hallucination*: the rationale introduces evidence absent from the input (Bang et al. 2025; Zhang et al. 2025b). Third, *aspect confusion*: the model entangles sentiment evidence across different aspects in the same sentence. These defects can mislead student models during training and cause error propagation (Ye and Durrett 2022). Figure 1 illustrates all three defects for the target aspect “food.” The unfaithful rationale acknowledges the absence of direct food evaluation but predicts a negative polarity, introduces unsupported evidence that the food was disappointing, and transfers service-related cues to the food aspect. In contrast, the faithful rationale keeps the evidence target-specific and supports the correct neutral prediction.

To address these defects, we introduce logical entailment, factual grounding, and aspect exclusivity, and jointly optimize them with rationale generation and sentiment classification. However, the five losses differ substantially in numerical scale, gradient magnitude, and convergence dynamics. Fixed weighting requires extensive manual tuning, while uncertainty-based automatic weighting (Kendall et al. 2018; Lai et al. 2026) can suffer from weight degeneration, where large-scale losses dominate the gradient and suppress other objectives. To stably balance such heterogeneous objectives, a scale-invariant weighting strategy is needed.

To this end, we propose **FACT-SR**, a scale-invariant multi-task framework for **Faithful AspeCT-based Sentiment Reasoning**. FACT-SR first constructs a reliable rationale dataset through verification and sufficiency filtering. It then trains a lightweight model with rationale generation, sentiment classification, and three faithfulness constraints. Finally, it incorporates Scale-Invariant Automatic Weight Learning (SI-AWL), which combines EMA-based loss normalization with bounded uncertainty mapping so that each objective continues to contribute throughout training. Our main contributions are as follows:

- We propose a faithful dataset construction pipeline that provides reliable rationale supervision for aspect-based sentiment reasoning.
- We design FACT-SR, a scale-invariant multi-task framework for improving the faithfulness of aspect-based sentiment reasoning.
- We propose SI-AWL to balance heterogeneous losses and improve training stability.
- Experiments show that FACT-SR improves classification and rationale faithfulness across two model scales, and that its constraints remain effective on a decoder-only backbone.

2 Related Work

2.1 Reasoning for Sentiment Analysis

ABSA has been extensively studied, with prior work improving sentiment classification through pre-trained language models, contrastive learning, causal intervention, and sentiment knowledge augmentation (Li et al. 2021; Ouyang et al. 2024; Peng et al. 2020; Rietzler et al. 2020; Wang et al. 2022; Zhang et al. 2023). Recent studies further exploit LLMs to elicit intermediate reasoning steps and transfer this reasoning ability to smaller models (Ho et al. 2023; Li et al. 2022; Magister et al. 2023). For sentiment reasoning, THOR introduces a three-hop chain-of-thought framework for implicit sentiment analysis (Fei et al. 2023). RVISA retains generated rationales and introduces answer-based reasoning verification as an auxiliary learning task (Lai et al. 2025). MT-ISA further combines label-guided rationale refinement with data-level and task-level automatic weighting (Lai et al. 2026). FACT-SR builds on these approaches by integrating three auxiliary constraints with a normalized, bounded weighting formulation, targeting the specific combination of label consistency, source alignment, and aspect attribution.

2.2 Faithfulness-Aware Joint Training

Explanation faithfulness has been widely discussed in NLP, where plausible explanations may not faithfully reflect the input evidence or the model decision process (Jacovi and Goldberg 2020). This issue becomes more critical when LLM-generated rationales are used as supervision, since they may contain hallucinated evidence, unsupported reasoning steps, or conclusions inconsistent with the final prediction (Bang et al. 2025; Zhang et al. 2025b). These studies distinguish explanation plausibility, evidence alignment, and faithfulness to model computation (Parcalabescu and Frank 2024; Tutek et al. 2025). For ABSA, FACT-SR operationalizes complementary output-level properties through polarity-recovery supervision, source-attention regularization, and aspect-level representation learning.

By introducing these constraints into training, sentiment reasoning becomes a multi-objective joint optimization problem. Prior sentiment studies have explored joint learning of related subtasks to improve classification (Schmitt et al. 2018; Wang et al. 2021). To avoid costly manual weighting, uncertainty-based automatic weighting has been proposed for multi-task learning (Kendall et al. 2018; Lai et al. 2026; Liebel and Körner 2018). FACT-SR combines objectives with different units and numerical scales. SI-AWL first normalizes these losses and then learns bounded uncertainty parameters; Section 3.4 specifies the resulting coefficients and their interpretation.

3 Methodology

As shown in Figure 2, FACT-SR consists of three tightly connected components: (1) faithful rationale construction for improving LLM-generated supervision, (2) faithfulness-aware multi-task learning for jointly optimizing explanation generation, sentiment prediction, and three faithfulness objectives, and (3) scale-invariant automatic weight learning for stabilizing the joint optimization of heterogeneous losses.

3.1 Task Definition and Framework

Given a sentence x and target aspect a , FACT-SR predicts one of three polarity labels (positive, negative, or neutral) and generates an accompanying rationale r_a . We write y_a for the gold label of aspect a , or y_i for instance i . Explanation generation and direct prediction share parameters but use separate task inputs. The direct prediction stream does not consume the generated rationale.

FACT-SR is designed to address the faithfulness defects of rationales from both the data and model levels. Following prior sentiment reasoning studies (Fei et al. 2023; Lai et al. 2025), we formulate the rationale as a three-hop reasoning chain:

$$r_a = [h_a^{(1)}, h_a^{(2)}, h_a^{(3)}], \quad (1)$$

where $h_a^{(1)}$ describes the facts relevant to the aspect, $h_a^{(2)}$ is the opinion inferred from those facts, and $h_a^{(3)}$ is the final polarity judgment.

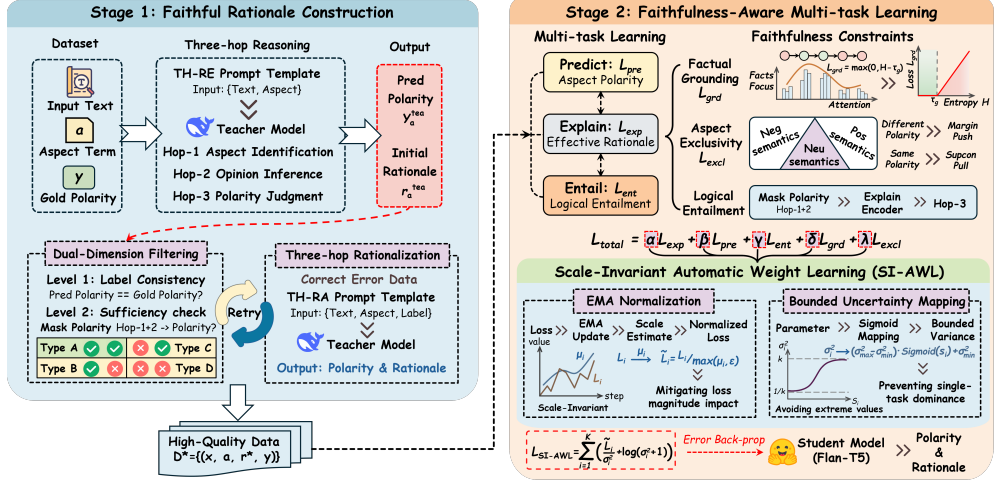


Fig. 2 Overview of FACT-SR: rationale construction and verification, three task streams with five objectives, and scale-invariant automatic weighting. The construction-time sufficiency check uses the sentence, aspect, and masked rationale; the training-time Entail stream uses the aspect and masked intermediate rationale

At the data level, the construction pipeline in Section 3.2 produces a checked supervision set, with N retained aspect instances:

$$\mathcal{D}^* = \{(x_i, a_i, r_{a_i}^*, y_i)\}_{i=1}^N. \quad (2)$$

At the model level, our primary realization trains an encoder-decoder model on $\mathcal{D}^*$ that jointly optimizes explanation generation, polarity prediction, and three faithfulness constraints in §3.3. Since the five training objectives are heterogeneous and hard to balance manually, we further introduce SI-AWL in §3.4. Section 3.5 maps the same supervision to a decoder-only backbone. Sections 4.2 and 4.3 describe model selection and training settings.

During training, the retained teacher rationale supervises the Explain stream, while the gold polarity supervises the Predict and label-recovery streams. At inference, the student receives only the sentence and target aspect. Held-out gold labels are used to score predictions, while retained reference rationales are stored alongside model outputs.

3.2 Rationale Construction and Checking

TH-RE (Three-Hop Reasoning) hides the gold label and requires the LLMs to autonomously complete the full reasoning process “aspect identification → opinion inference → polarity judgment”. **TH-RA** (Three-Hop Rationalization) reveals the gold label and asks the LLMs to produce or repair rationales that support the given conclusion. Appendix A provides the full prompts.

3.2.1 Initial Rationale Generation

For each input (x, a) , we use TH-RE to generate an initial rationale. The LLM outputs a complete reasoning chain r_a^{tea} , from which the predicted polarity $\hat{y}_a^{tea}$ is parsed from Hop-3. For sentences containing multiple aspects, we generate a reasoning chain for each aspect independently, so that subsequent filtering remains strictly aspect-level.

3.2.2 Label Consistency and Rationale Sufficiency

As discussed in Section 1, LLM-generated rationales may exhibit faithfulness defects. To improve supervision quality, we evaluate each candidate using two complementary criteria.

First, **label consistency** requires the polarity inferred by the LLM to match the gold label:

$$\text{LabelMatch}(a) = \mathbb{I}[\hat{y}_a^{tea} = y_a]. \quad (3)$$

Second, **aspect-conditioned rationale sufficiency** evaluates whether the reasoning chain provides evidence supporting the polarity of aspect a , drawing on rationale-sufficiency evaluation (DeYoung et al. 2020). The verifier receives the sentence, aspect, and reasoning chain with the explicit Hop-3 polarity masked, and is instructed to infer the label from the intermediate reasoning. The recovered label is then compared with the gold polarity.

To this end, samples are partitioned into four groups: (i) label-consistent and rationale-sufficient, (ii) label-consistent but rationale-insufficient, (iii) label-inconsistent but rationale-sufficient, and (iv) failing both criteria. Only samples belonging to group (i) are retained. The remaining samples are reconstructed via the TH-RA template or, when necessary, manually corrected, and then re-verified until the full supervision set $\mathcal{D}^*$ is faithful.

3.2.3 Repair and Split Handling

Table 1 shows the four-way sample classification used in the construction pipeline. Type-A samples are directly retained. Type-B samples are repaired by TH-RA because their final labels are correct but the rationales are insufficient. Type-D samples are reconstructed by TH-RA and re-verified. Type-C samples are treated conservatively because they contain a mismatch between the teacher’s final conclusion and the gold label. If TH-RA still fails to produce a reliable correction after rewriting or reconstruction, we manually modify samples with obvious recoverable errors. Ambiguous or irreparable samples are removed from the training dataset. The test dataset undergoes repeated corrections without removing sample data to ensure experimental comparability.

3.3 Faithfulness-Aware Multi-Task Learning

Given the supervision set $\mathcal{D}^*$, our primary realization uses Flan-T5 (Chung et al. 2024) as the backbone. Task-specific prefixes distinguish the input streams, which share the backbone f_θ with parameters θ . The overall loss consists of two main-task losses and three faithfulness-constraint losses.

Table 1 Four-way classification of LLM-generated rationales in the faithful data construction stage.

Type	Label agreement	Label recovery	Action
A	matched	passed	retain
B	matched	failed	rewrite
C	mismatched	passed	rewrite
D	mismatched	failed	rewrite

3.3.1 Task Inputs

As shown in Table 2, each retained training instance produces three examples distinguished by their prefixes. Explain and Predict use the sentence and target aspect as input, with the rationale and polarity label as their respective targets. Entail uses the aspect and masked intermediate rationale to predict the polarity. This design shares parameters across the three tasks while keeping the prediction input independent of the reference rationale.

Table 2 Task-specific input construction for faithfulness-aware multi-task learning.

Task	Input prefix	Target output
Explain	Explain:	three-hop rationale
Predict	Predict:	polarity label
Entail	Entail:	polarity label

For Explain and Predict, the input contains the original sentence and target aspect. For Entail, it contains the target aspect and the first two rationale hops with explicit polarity words removed. Unlike the construction-time recovery check, this training stream does not include the original sentence or Hop-3. Same-sentence aspect metadata supports the exclusivity objective. At test time, the Explain and Predict inputs are built from the sentence and aspect alone; reference rationales and gold labels are retained only for post-hoc evaluation.

3.3.2 Main Tasks

Explanation Generation Conditioned on the prefix “Explain:”, the model generates the complete three-hop rationale $r_{a_i}^*$ from the input text and target aspect. The explanation objective combines sequence generation with additional polarity-token supervision: $\mathcal{L}_{exp} = \mathcal{L}_{gen} + \lambda_{label}\mathcal{L}_{label}$. Let $\ell_{CE}(\cdot, \cdot)$ denote cross-entropy loss. The sequence-generation term is

$$\mathcal{L}_{gen} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(f_{\theta}^{exp}(x_i, a_i), r_{a_i}^*). \quad (4)$$

To further emphasize the polarity decision in Hop-3, we additionally apply a weighted cross-entropy to Hop-3 target tokens corresponding to **positive**, **negative**, or **neutral**. Let $\mathcal{P}_{pol}$ denote these Hop-3 target positions.

$$\mathcal{L}_{label} = \frac{1}{|\mathcal{P}_{pol}|} \sum_{(i,t) \in \mathcal{P}_{pol}} \ell_{CE}(\hat{\mathbf{p}}_{i,t}, r_{a_i,t}^*), \quad (5)$$

where $\hat{\mathbf{p}}_{i,t}$ is the predicted token distribution at position t , and $r_{a_i,t}^*$ is the reference target token at that position. This design aligns token-level supervision with the polarity decision, emphasizing accurate polarity generation alongside rationale fluency.

Polarity Prediction Conditioned on the prefix “**Predict:**”, the model directly generates the sentiment label y_a from the input sentence and aspect. The prediction loss is

$$\mathcal{L}_{pre} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(f_{\theta}^{pre}(x_i, a_i), y_{a_i}). \quad (6)$$

Unlike $\mathcal{L}_{label}$, which supervises Hop-3 polarity tokens within rationale generation, $\mathcal{L}_{pre}$ supervises direct ABSA prediction from the sentence and target aspect.

3.3.3 Faithfulness Constraints

In addition to the main tasks, we introduce three constraints that respectively target the logical disconnection, factual hallucination, and aspect confusion defects identified in §1.

Logical Entailment The logical entailment objective encourages recovery of the gold polarity from Hop-1 and Hop-2. Conditioned on the prefix “**Entail:**”, the model receives the aspect and $\tilde{r}_a^{(1,2)}$, obtained by removing explicit polarity words from the first two hops. Its generation loss is

$$\mathcal{L}_{ent}^{gen} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(f_{\theta}^{ent}(a_i, \tilde{r}_{a_i}^{(1,2)}), y_{a_i}). \quad (7)$$

At the representation level, we further attach a lightweight two-layer multilayer perceptron (MLP) classifier f_{cls} to the mean-pooled encoder hidden state and supervise its class-probability output with the gold polarity:

$$\mathcal{L}_{ent}^{cls} = \frac{1}{N} \sum_{i=1}^N \ell_{CE}(\hat{\mathbf{p}}_i^{ent}, y_{a_i}), \quad (8)$$

where $\hat{\mathbf{p}}_i^{ent}$ denotes the class-probability vector output by the entailment classification head from $(a_i, \tilde{r}_{a_i}^{(1,2)})$.

Combining the two terms supervises polarity recovery at both the sequence and representation levels, giving the logical entailment loss $\mathcal{L}_{ent} = 0.5\mathcal{L}_{ent}^{gen} + 0.5\mathcal{L}_{ent}^{cls}$.

Factual Grounding The factual grounding constraint is intended to reduce unsupported content by encouraging each generated token to concentrate its attention on a

small set of input positions. We average the multi-head cross-attention from the last decoder layer of the Explain task, mask padding positions, and compute the entropy $\mathcal{H}_{b,t}$ at each decoding step. This entropy is an operational proxy for source-attention dispersion rather than proof of factual correctness. We penalize high-entropy positions with a hinge loss:

$$\mathcal{L}_{grd} = \frac{1}{|\mathcal{T}|} \sum_{(b,t) \in \mathcal{T}} \max(0, \mathcal{H}_{b,t} - \tau_g), \quad (9)$$

where $\mathcal{T}$ is the set of valid decoding positions and τ_g is the entropy threshold. This term reuses attention tensors from the Explain forward pass and requires no additional inference-time module.

Aspect Exclusivity The aspect exclusivity constraint mitigates aspect confusion by shaping encoder representations. Same-sentence aspects with *different* polarities should remain distinguishable, whereas aspects sharing the *same* polarity should be closer. Let $\mathbf{z}_i$ denote the ℓ_2 -normalized hidden state at the aspect position, and let $\mathcal{S}^-$ collect within-sentence aspect pairs with different polarities. We apply a margin-based loss that pushes apart different-polarity pairs:

$$\mathcal{L}_{push} = \frac{1}{|\mathcal{S}^-|} \sum_{(i,j) \in \mathcal{S}^-} \max(0, \mathbf{z}_i^\top \mathbf{z}_j - m). \quad (10)$$

We additionally use supervised contrastive learning for same-polarity samples. Let B be the batch size and $\mathcal{P}(i) = \{j \neq i : y_j = y_i\}$ the positive set that shares the polarity of sample i :

$$\mathcal{L}_{pull} = \frac{1}{B} \sum_{i=1}^B \frac{-1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \log \frac{\exp(\mathbf{z}_i^\top \mathbf{z}_j / \tau)}{\sum_{k \neq i} \exp(\mathbf{z}_i^\top \mathbf{z}_k / \tau)}. \quad (11)$$

Here m is a cosine-similarity margin and τ is the contrastive temperature. The final loss is $\mathcal{L}_{excl} = \mathcal{L}_{push} + \mathcal{L}_{pull}$. As with $\mathcal{L}_{grd}$, this term reuses states from the Explain forward pass and requires no additional backbone forward pass.

The five objectives combine into the following weighted loss:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{exp} + \beta \mathcal{L}_{pre} + \gamma \mathcal{L}_{ent} + \delta \mathcal{L}_{grd} + \lambda \mathcal{L}_{excl}. \quad (12)$$

FACT-SR replaces these manually specified weights with coefficients learned through Scale-Invariant Automatic Weight Learning (SI-AWL).

Training–inference separation. In the Flan-T5 implementation, FACT-SR applies the grounding and exclusivity objectives during training by reusing the attention tensors and encoder representations returned by the Explain forward pass. The entailment classifier and SI-AWL parameters are training-time components; inference loads only the fine-tuned student backbone and its low-rank adaptation (LoRA) adapter, without evaluating the auxiliary losses. Thus, the additional objectives strengthen supervision without introducing constraint-specific modules into the deployed predictor.

3.4 SI-AWL

Combining the five objectives above with fixed weights would require searching a five-dimensional space. Even a grid with M candidates per weight contains M^5 combinations. Although uncertainty-based automatic weighting reduces this manual search (Kendall et al. 2018), it can still suffer from weight degeneration when objectives have substantially different numerical scales. SI-AWL builds on automatic uncertainty weighting (Lai et al. 2026; Liebel and Körner 2018) by combining a moving scale estimate for each loss with bounded uncertainty parameters.

3.4.1 EMA-based Loss Normalization

For each objective $\mathcal{L}_i$, we maintain an exponential moving average (EMA) μ_i as its scale estimate. The first mini-batch initializes the buffer directly. Subsequent mini-batches update it before the weighted loss is evaluated:

$$\mu_i^{(t)} = \max(\rho\mu_i^{(t-1)} + (1 - \rho) \text{sg}[\mathcal{L}_i^{(t)}], \varepsilon), \quad (13)$$

where ρ is the momentum coefficient, ε is a numerical floor, and $\text{sg}[\cdot]$ denotes stop-gradient. We normalize each objective as

$$\tilde{\mathcal{L}}_i = \frac{\mathcal{L}_i}{\mu_i}. \quad (14)$$

For a constant rescaling $\mathcal{L}_i \mapsto c_i \mathcal{L}_i$ with $c_i > 0$, the ratio is unchanged if its moving-average buffer is rescaled consistently and the numerical floor is inactive. This normalization makes weight learning less sensitive to the absolute magnitude of each loss and more dependent on its relative training state.

3.4.2 Bounded Uncertainty Mapping

In standard uncertainty-based weighting, variance parameters are typically obtained through an exponential mapping, which can drift toward extreme values during training. We instead map the learnable parameter s_i of each objective into a bounded interval determined by the number of objectives K :

$$\sigma_i^2 = \sigma_{\min}^2 + (\sigma_{\max}^2 - \sigma_{\min}^2) \cdot \text{sigmoid}(s_i), \quad (15)$$

where $\sigma_{\min}^2 = 1/K$, and $\sigma_{\max}^2 = K$. Compared with exponential mapping followed by hard clipping, the sigmoid retains non-zero gradients across its entire domain, and the effective weights $1/\sigma_i^2$ are constrained within $[1/K, K]$, preventing any single objective from dominating the optimization. Therefore, the final SI-AWL objective is

$$\mathcal{L}_{SI-AWL} = \sum_{i=1}^K \left(\frac{\tilde{\mathcal{L}}_i}{\sigma_i^2} + \log(\sigma_i^2 + 1) \right). \quad (16)$$

The logarithmic term regularizes the learned uncertainties and discourages the degenerate solution in which every high-level objective is suppressed (Lai et al. 2026; Liebel and Körner 2018).

3.5 Decoder-Only Realization

We also instantiate FACT-SR with Qwen2.5-7B-Instruct (Yang et al. 2024). The supervision, five objectives, and SI-AWL procedure are retained, while encoder representations and cross-attention are replaced by their decoder-only counterparts. The following adaptations implement the objectives in a decoder-only architecture. Section 5.3 reports the corresponding classification results.

Causal Prediction. Prompt tokens are masked from the language-model labels, and causal cross-entropy is computed only over response tokens, including the end-of-sequence (EOS) token. This provides the decoder-only forms of $\mathcal{L}_{gen}$, $\mathcal{L}_{pre}$, and $\mathcal{L}_{ent}^{gen}$. For $\mathcal{L}_{label}$, $\mathcal{P}_{pol}$ is obtained by matching the complete byte-pair encoding (BPE) sequences of **positive**, **negative**, and **neutral** within Hop-3. The distribution $\hat{\mathbf{p}}_{i,t}$ is obtained from the causal logits at position $t - 1$.

Causal Representations. The original $\mathcal{L}_{ent}^{cls}$ and $\mathcal{L}_{excl}$ obtain their representations from encoder states. Let $\mathcal{I}_i^k$ denote the task-input positions for task k , excluding the response delimiter, target response, and padding. Let $\mathbf{h}_{i,s}^k$ be the final-layer causal state at position s . For aspect exclusivity, annotated character offsets and boundary markers identify the target-aspect positions $\mathcal{A}_i \subset \mathcal{I}_i^{exp}$. The corresponding decoder-only representations are

$$\begin{aligned}\hat{\mathbf{p}}_i^{ent} &= f_{cls} \left(\frac{1}{|\mathcal{I}_i^{ent}|} \sum_{s \in \mathcal{I}_i^{ent}} \mathbf{h}_{i,s}^{ent} \right), \\ \bar{\mathbf{h}}_i^a &= \frac{1}{|\mathcal{A}_i|} \sum_{s \in \mathcal{A}_i} \mathbf{h}_{i,s}^{exp}, \\ \mathbf{z}_i &= \frac{\bar{\mathbf{h}}_i^a}{\max(\|\bar{\mathbf{h}}_i^a\|_2, \varepsilon_z)}.\end{aligned}\tag{17}$$

The first expression supplies the class-probability vector used by $\mathcal{L}_{ent}^{cls}$. In the normalization denominator, $\varepsilon_z > 0$ is a numerical safeguard. The normalized aspect representation $\mathbf{z}_i$ enters the original $\mathcal{L}_{push}$ and $\mathcal{L}_{pull}$. The classification targets and pair construction remain unchanged.

Self-Attention Grounding. The Flan-T5 grounding constraint obtains $\mathcal{H}_{b,t}$ from decoder-to-encoder cross-attention. In Qwen2.5, we compute the same quantity from last-layer self-attention between each response query and the **Explain:** input. Let $q(t) = t - 1$ denote the causal position that predicts response token t , and let $\bar{A}_{b,q(t),s}$ denote its attention to input position s , averaged across heads. We normalize this

attention over the task input:

$$\begin{aligned}\tilde{A}_{b,t,s} &= \frac{\bar{A}_{b,q(t),s}}{\sum_{j \in \mathcal{I}_b^{exp}} \bar{A}_{b,q(t),j}}, & s \in \mathcal{I}_b^{exp}, \\ \mathcal{H}_{b,t} &= - \sum_{s \in \mathcal{I}_b^{exp}} \tilde{A}_{b,t,s} \log \tilde{A}_{b,t,s}.\end{aligned}\tag{18}$$

The normalized weights form a distribution over the **Explain:** input, and $\mathcal{H}_{b,t}$ measures its concentration. This entropy directly replaces the cross-attention entropy in $\mathcal{L}_{grd}$, while the original hinge and threshold τ_g are retained. The grounding objective therefore preserves its operational role under causal self-attention.

4 Experimental Protocol

4.1 Experimental Setup

Datasets and metrics. We evaluate on the English SemEval-2014 Restaurant and Laptop ABSA benchmarks (Pontiki et al. 2014), using positive, negative, and neutral sentiment labels. Accuracy (Acc) measures the proportion of correct aspect-level predictions. Macro-F1 (F1) gives equal weight to the F1 score of each polarity class. Classification results are expressed as percentages; differences are reported in percentage points. Section 4.2 describes checkpoint selection.

Rationale construction and baselines. DeepSeek-V3.2¹ is the default rationale generator and construction-time verifier. GPT-4o-mini provides a second teacher setting. Comparisons include BERT-based systems (Devlin et al. 2019; Li et al. 2021; Rietzler et al. 2020; Wang et al. 2020), THOR (Fei et al. 2023), RVISA (Lai et al. 2025), and MT-ISA (Lai et al. 2026). Section 5 reports the main comparison, with source markers distinguishing literature-reported results from our experiments.

Student models. Author-trained students and matched controls use LoRA. The Flan-T5 experiments cover 250M and 11B backbones, and the decoder-only comparison uses Qwen2.5-7B-Instruct. Sections 4.3 and 4.4 report the available configuration details. LoRA usage here does not imply that literature baselines used the same adaptation settings.

Inference and evaluation. The Flan-T5 evaluation implementation generates Explain and Predict outputs separately. The final polarity is taken from the Explain output when it can be parsed; otherwise, the Predict output supplies the fallback label. Rationale repair operates on the generated Explain text after this label selection.

4.2 Model Selection

Since the SemEval-2014 ABSA benchmarks do not provide an official validation split, model selection can be sensitive to the chosen train-validation partition. This issue is amplified in multi-aspect sentences, where aspect instances from the same sentence

¹API documentation: <https://api-docs.deepseek.com>.

may otherwise be split across training and validation sets. We therefore adopt sentence-grouped stratified K -fold cross-validation to determine the number of training epochs. All aspect instances from the same sentence are kept in the same fold, preventing train-validation leakage in multi-aspect cases. Each fold is trained independently, and validation performance is recorded at every epoch. For each epoch, we compute the cross-fold mean Macro-F1 and Accuracy on the held-out folds. Instead of relying on a single train-validation split, we use the averaged validation curves to identify a stable training range and select the final training epoch before test evaluation. The model is then retrained on the full training set for the selected number of epochs and evaluated once on the test set.

4.3 Training Configuration

Training Setup. Table 3 summarizes the Flan-T5 configuration, including sequence lengths, batch sizes, optimization, loss hyperparameters, and decoding. Training hyperparameters are tuned separately for each model scale and dataset, and SI-AWL parameters use a separate learning rate. The maximum epochs define the search limits for the model-selection procedure in Section 4.2. The Qwen realization uses response-only causal targets, bfloat16 (BF16), gradient checkpointing, eager attention, and the same loss-specific hyperparameters for both matched variants.

Table 3 Reported training settings. Maximum epochs are search limits, not necessarily the selected training duration.

Setting	Value
Input / target length	256 / 256 tokens
Training / evaluation batch size	16 / 16
Optimizer	AdamW; weight decay 0.01
Learning-rate schedule	Linear decay after 0.1 warmup ratio
Gradient clipping	1.0
SI-AWL learning rate	1×10^{-3}
Loss hyperparameters	$\lambda_{label} = 3.0$, $\tau_g = 2.0$, $\tau = 0.1$
EMA / numerical floor	$\rho = 0.99$, $\varepsilon = 0.01$
Maximum epochs	50 (250M); 20 (11B)
Generation	4 beams; maximum output length 256

Computing Environment. For the 250M Flan-T5 backbone, training and evaluation use 32-bit floating-point precision (FP32) on a single NVIDIA RTX 3090 (24 GB) GPU. For the 11B Flan-T5 backbone, we use BF16 mixed precision and gradient checkpointing on a single NVIDIA RTX PRO 6000 (96 GB) GPU.

4.4 Parameter-Efficient Fine-Tuning

Full-parameter fine-tuning of the backbone under our five-objective formulation is computationally expensive. Low-rank adaptation (LoRA) (Hu et al. 2021) provides

a parameter-efficient alternative by injecting trainable low-rank matrices into the attention and gated feed-forward network (FFN) projection layers while keeping the backbone frozen. This design reduces trainable parameters, optimizer states, and checkpoint storage. We use LoRA for all author-trained students and controls, with the configurations summarized in Table 4.

Table 4 LoRA configuration used for FACT-SR.

Item	Flan-T5	Qwen2.5-7B-Instruct
Rank r	64	64
Scaling factor α	128	128
Dropout rate	0.1	0.1
Backbone	Frozen	Frozen

5 Results and Analysis

5.1 Main Results

Table 5 reports the classification results for discriminative baselines, training-free reasoning methods, fine-tuned systems, and FACT-SR. FACT-SR_D achieves the highest listed scores at both student scales. Source markers distinguish literature-reported results from our experiments.

Results at 250M. FACT-SR_D obtains 89.91/85.12 accuracy/macro-F1 on Restaurant and 83.39/80.17 on Laptop. Relative to the reported MT-ISA_I results, the gains are 1.70/2.67 and 0.48/0.31 points, respectively. Compared with author-run Vanilla SFT, the gains are 3.75/6.58 on Restaurant and 1.41/2.33 on Laptop. The student also exceeds the DeepSeek-V3.2+TH-RE teacher baseline by 2.14/5.38 and 1.42/3.20 points. These results demonstrate the effectiveness of the complete training pipeline in transferring teacher-generated reasoning supervision to a compact student.

Results at 11B. FACT-SR_D reaches 92.77/89.24 on Restaurant and 87.77/85.22 on Laptop. The margins over the strongest listed prior baseline for each metric are 0.09/0.28 and 1.09/1.17 points, respectively. Against author-run Vanilla SFT, the gains are 2.41/4.27 and 3.13/3.84 points. The improvements in both accuracy and macro-F1 show that the benefits extend to equally weighted class-level performance. Notably, FACT-SR (250M) also exceeds the listed THOR (11B) results with approximately 1/44 as many backbone parameters, highlighting the competitiveness of the compact student.

Sensitivity to the construction model. With student backbones and training procedures held fixed, FACT-SR_D and FACT-SR_G differ by approximately 0.14 points on average over the eight dataset-scale-metric combinations. The largest difference is 0.54 points in Laptop accuracy at 11B. FACT-SR_G (250M) also exceeds the reported MT-ISA_I results by 1.61/2.53 on Restaurant and 0.48/0.23 on Laptop. The similar

Table 5 Performance comparison on the benchmarks (Accuracy/Macro-F1, %). Subscripts D and G denote DeepSeek-V3.2 and GPT-4o-mini. All author-trained models use LoRA. Results marked with †, *, and ¶ are from Li et al. (2021), Lai et al. (2025), and Lai et al. (2026). Unmarked rows are author-run. Underlining marks the strongest prior fine-tuned baseline at each scale, and bold marks the FACT-SR result.

Method	Restaurant		Laptop	
	Acc	F1	Acc	F1
• <i>Discriminative ABSA baselines</i>				
BERT+SPC [†] (110M)	83.57	77.16	78.22	73.45
BERT+ADA [†] (110M)	87.14	80.05	78.96	74.18
BERT+RGAT [†] (110M)	86.60	81.35	78.21	74.07
BERT _{Asp} +CEPT [†] (110M)	87.50	82.07	81.66	78.38
BERT+ISAIV [¶] (110M)	87.05	81.40	80.41	77.25
BERT _{Asp} +SCAPT [†] (110M)	89.11	83.79	82.76	79.15
• <i>Training-free reasoning baselines</i>				
GPT-4o-mini (zero-shot) [¶]	84.82	72.52	79.94	73.83
GPT-4o-mini+THOR [¶]	81.25	62.12	72.10	55.42
GPT-4o (zero-shot) [¶]	85.98	73.82	80.72	74.24
GPT-4o+ICL (3-shot) [¶]	87.59	78.82	82.91	79.24
DeepSeek-R1+ICL (3-shot) [¶]	88.82	83.06	81.03	75.13
DeepSeek-V3.2 (zero-shot)	86.07	74.01	80.88	74.33
DeepSeek-V3.2+TH-RE	87.77	79.74	81.97	76.97
• <i>Fine-tuned systems and controls</i>				
Vanilla SFT (250M)	86.16	78.54	81.98	77.84
Vanilla SFT (11B)	90.36	84.97	84.64	81.38
Flan-T5+THOR* (250M)	87.68	81.10	81.66	77.51
Flan-T5+THOR* (11B)	88.57	82.93	82.29	78.78
Flan-T5+RVISA _g * (250M)	86.61	78.92	81.19	77.13
Flan-T5+RVISA _g * (11B)	91.52	86.85	<u>86.68</u>	<u>84.05</u>
Flan-T5+MT-ISA _I [¶] (250M)	<u>88.21</u>	<u>82.45</u>	<u>82.91</u>	<u>79.86</u>
Flan-T5+MT-ISA _I [¶] (11B)	91.34	86.71	85.40	82.73
Flan-T5+MT-ISA _O [¶] (250M)	85.89	77.85	81.98	78.10
Flan-T5+MT-ISA _O [¶] (11B)	<u>92.68</u>	<u>88.96</u>	85.74	83.86
• <i>Our methods</i>				
Flan-T5+FACT-SR _G (250M)	89.82	84.98	83.39	80.09
Flan-T5+FACT-SR _D (250M)	89.91	85.12	83.39	80.17
Flan-T5+FACT-SR _G (11B)	92.61	89.19	87.23	85.14
Flan-T5+FACT-SR _D (11B)	92.77	89.24	87.77	85.22

aggregate performance across these two teacher settings supports the applicability of the training framework to supervision produced by either teacher.

5.2 Ablation Study

To assess the contribution of the three auxiliary constraints, we remove them individually and jointly, as reported in Table 6.

Table 6 Ablation study of the three faithfulness constraints on FACT-SR_D.

	Restaurant		Laptop	
	Acc	F1	Acc	F1
Flan-T5+FACT-SR _D (250M)	89.91	85.12	83.39	80.17
– w/o $\mathcal{L}_{ent}$	87.68	82.07	83.23	78.97
– w/o $\mathcal{L}_{grd}$	88.48	83.11	83.07	78.51
– w/o $\mathcal{L}_{excl}$	88.12	82.26	83.23	79.09
– w/o all constraints	87.23	81.19	82.45	77.88
Flan-T5+FACT-SR _D (11B)	92.77	89.24	87.77	85.22
– w/o $\mathcal{L}_{ent}$	91.87	87.78	85.58	82.71
– w/o $\mathcal{L}_{grd}$	91.61	87.54	86.21	82.97
– w/o $\mathcal{L}_{excl}$	91.43	86.89	85.89	83.65
– w/o all constraints	91.25	86.33	84.95	82.15

The full model has the highest accuracy and macro-F1 in every dataset-scale setting in Table 6. Removing all three constraints produces the largest drop in each reported metric, supporting their combined contribution to classification. Section 5.5 further examines their effects on the faithfulness metrics.

The largest single-removal macro-F1 drop on Restaurant is associated with $\mathcal{L}_{ent}$ at 250M and $\mathcal{L}_{excl}$ at 11B. On Laptop, it is associated with $\mathcal{L}_{grd}$ at 250M and $\mathcal{L}_{ent}$ at 11B. The relative contribution therefore varies by dataset and scale, while retaining all three constraints gives the strongest results throughout. This pattern is consistent with complementary regularization from logical entailment, factual grounding, and aspect exclusivity.

5.3 Decoder-Only Results

We extend the Flan-T5 implementation of FACT-SR_D to the decoder-only Qwen2.5-7B-Instruct backbone. Table 7 compares variants with the same rationale supervision, LoRA configuration, and evaluation protocol. Adding the constraints improves accuracy/macro-F1 by 1.88/3.76 points on Restaurant and 2.04/3.36 on Laptop, supporting the classification utility of the decoder-only realization.

Table 7 Matched decoder-only results. Both variants use Qwen2.5-7B-Instruct with LoRA and the same checked rationale supervision.

	Restaurant		Laptop	
	Acc	F1	Acc	F1
Qwen2.5+FACT-SR _D (7B)	91.52	87.11	86.21	83.99
– w/o all constraints	89.64	83.35	84.17	80.63

5.4 Weighting Strategy Analysis

We compare SI-AWL with the AWL baseline used in prior multi-task sentiment learning (Lai et al. 2026). With the objective set held fixed, this comparison evaluates the combined effect of EMA normalization and bounded uncertainty mapping.

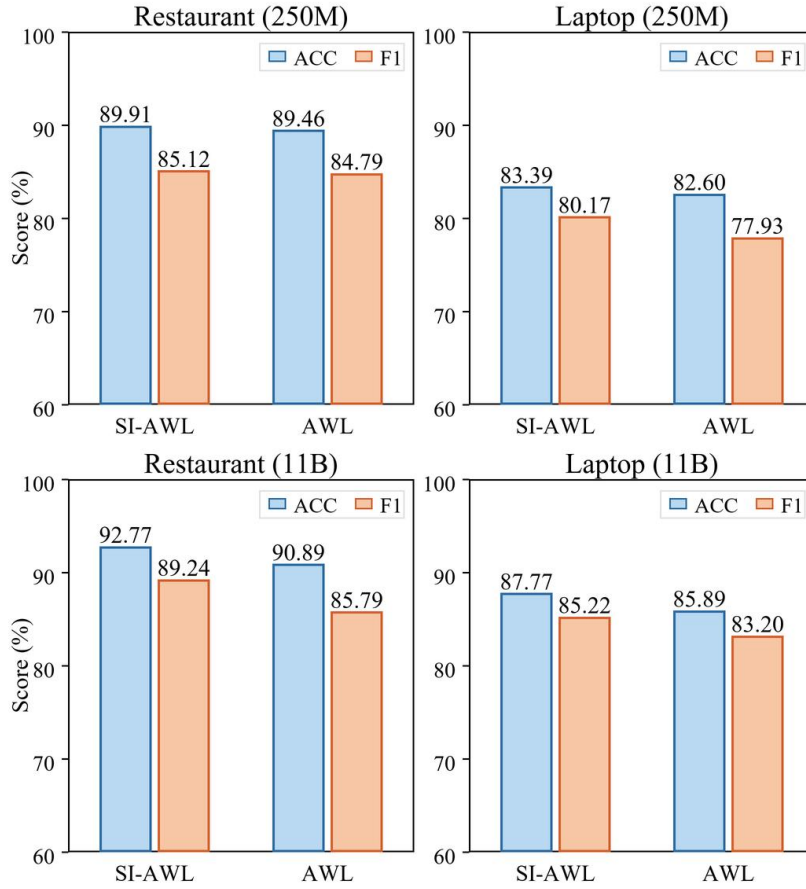


Fig. 3 Comparison of different weighting strategies on the benchmark datasets

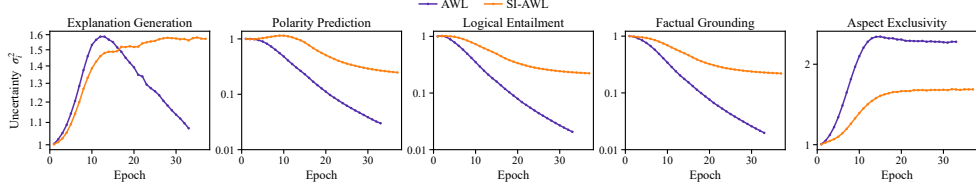


Fig. 4 Training trajectories of the learned uncertainty terms for the five high-level objectives on Restaurant with the 250M backbone. Larger uncertainty corresponds to a smaller effective loss weight

Figure 3 reports higher accuracy and macro-F1 for SI-AWL than AWL in all four evaluated settings. The improvements across both datasets and model scales support the effectiveness of combining loss-scale normalization with bounded uncertainty learning.

Figure 4 visualizes the learned uncertainty trajectories for Restaurant at 250M. The bounded SI-AWL parameterization restricts their range by construction, while EMA normalization accounts for differences in loss scale. Together, these mechanisms provide a controlled range for the learned coefficients on normalized objectives throughout training.

5.5 Faithfulness Analysis

To complement classification performance, we evaluate rationale quality and aspect-level sentiment behavior along complementary dimensions of reasoning quality (Golovneva et al. 2023). We first define the evaluation metrics and then analyze the full model and its constraint ablations.

5.5.1 Evaluation Metrics

NLI-Align uses natural language inference (NLI) to measure polarity agreement. Explicit polarity words are masked in the generated first two hops, and a DeBERTa-v3-large NLI model (Laurer et al. 2024) selects the most entailed polarity hypothesis. Agreement between the selected label and the gold polarity yields the NLI-Align score.

Sentiment hallucination rate (SHR) is the fraction of gold-neutral instances predicted as positive or negative, isolating cases where hallucinated sentiment is most likely. **Multi-aspect polarity accuracy (MAPA)** measures classification accuracy on multi-aspect sentences whose aspects have different polarities, directly testing whether the model can keep aspect-specific evidence disentangled.

MiniCheck (Tang et al. 2024) and **AlignScore** (Zha et al. 2023) assess source-rationale support using the input sentence and aspect as evidence and the generated rationale as the claim.

5.5.2 Results

As shown in Figure 5, the FACT-SR_D model achieves the strongest overall faithfulness profile across both domains and model scales. Relative to the variant without all constraints, it consistently improves NLI-Align, MiniCheck, AlignScore, and MAPA

while reducing SHR. The largest gains occur in SHR and MAPA. At 250M, SHR decreases by 6.64 and 10.65 points, while MAPA increases by 6.58 and 7.76 points on Restaurant and Laptop, respectively.

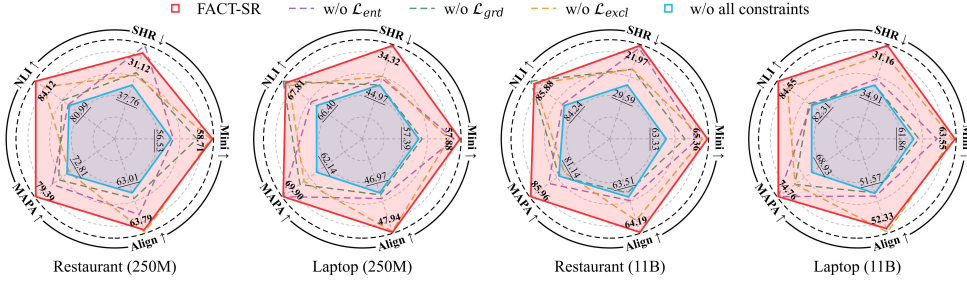


Fig. 5 Faithfulness analysis at the 250M and 11B scales. NLI, Mini, and Align denote NLI-Align, MiniCheck, and AlignScore. SHR is lower-is-better, and the other metrics are higher-is-better

The same pattern holds at 11B. These improvements suggest that FACT-SR_D learns more discriminative aspect-conditioned representations. Polarity-specific evidence is more clearly separated while remaining aligned with the target aspect. The variation in gain magnitude reflects the distinct scope of each diagnostic. AlignScore and MiniCheck primarily assess factual consistency, whereas NLI-Align, SHR, and MAPA measure logical support, polarity discrimination, and aspect specificity, respectively. The single-removal profiles likewise indicate that the three constraints provide complementary supervision. Together with the accuracy and macro-F1 gains in Table 6, these findings demonstrate that FACT-SR_D improves predictive performance while strengthening multiple dimensions of faithfulness.

5.6 Representation Analysis

To better understand the behavior of FACT-SR, we visualize the aspect-based sentiment representations using t-SNE (van der Maaten and Hinton 2008). For each test instance, we extract the hidden states corresponding to the target aspect span and project the resulting representations into two dimensions. Colors denote gold sentiment polarities, and stars indicate multi-aspect conflict samples whose aspects in the same sentence have different gold polarities.

As shown in Figure 6, the base model without faithfulness constraints produces more mixed polarity regions on both domains, indicating limited separation among aspect-based sentiment representations. In contrast, FACT-SR forms clearer polarity-specific clusters, where positive, negative, and neutral samples are more distinguishable. The separation is especially visible for multi-aspect conflict samples, which are placed closer to their corresponding gold-polarity regions after faithfulness-aware multi-task training. These results suggest that FACT-SR improves aspect-specific semantic separation by shaping internal sentiment representations, rather than only improving label prediction.

Across scales, we observe a clear scaling effect. The 11B model produces more compact and better separated clusters, especially on Laptop, where the 250M model

still shows noticeable overlap among polarity regions. This suggests that larger models can more effectively absorb faithfulness constraints and translate them into higher-margin aspect-based sentiment representations.

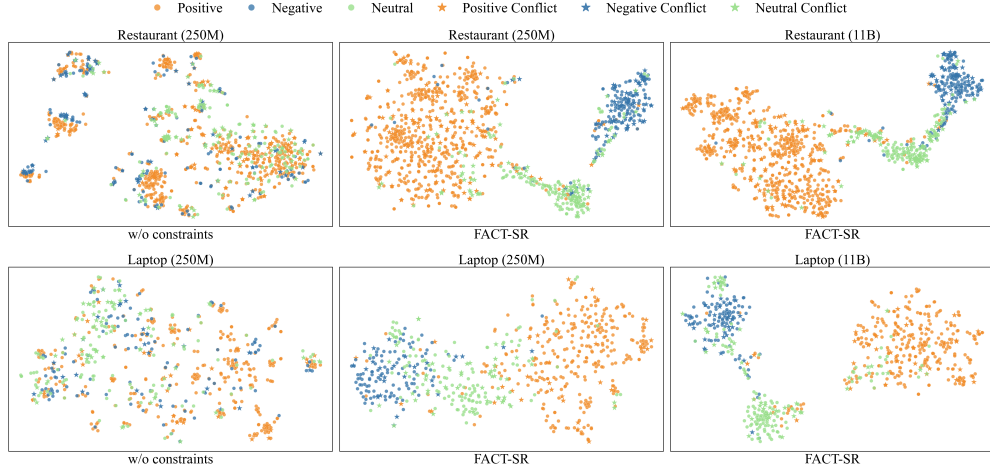


Fig. 6 t-SNE visualization of aspect-based sentiment representations on Restaurant (top) and Laptop (bottom). In each row, the left, middle, and right panels correspond to the 250M model without constraints, FACT-SR (250M), and FACT-SR (11B), respectively. Colors indicate gold sentiment polarity, and stars denote multi-aspect conflict samples whose aspects in the same sentence have different gold polarities. FACT-SR produces more structured polarity-specific regions than the unconstrained model, and the 11B model further yields more compact and better separated clusters.

5.7 Case Study

Tables 8–10 provide a qualitative view of the faithfulness mechanisms introduced in FACT-SR. Each case includes the input text, target aspect, gold label, generated rationale, and sentiment prediction. The examples illustrate how unfaithful rationales can produce incorrect aspect-level decisions, while FACT-SR preserves the evidence needed for the gold polarity.

Case 1 illustrates logical disconnection. For the aspect “Indian food”, DeepSeek-V3.2 identifies the speaker’s past negative attitude but fails to integrate the contrastive cue that Baluchi’s changed the speaker’s mind. As a result, its negative prediction is not entailed by the full reasoning context. In contrast, FACT-SR preserves the temporal contrast between the past attitude and the current opinion shift, leading to the correct positive prediction.

Table 8 Case 1: logical disconnection. Blue highlights mark spans that expose the defect in the DeepSeek-V3.2 rationale; red highlights mark the corresponding evidence in the FACT-SR rationale. Bold aspect terms and gold labels identify the evaluated target.

Input Text	For someone who used to hate Indian food, Baluchi’s has changed my mind.
Aspect Terms	[Indian food]
Gold Polarity	[positive]
DeepSeek-V3.2	Generated Rationale: The mentioned aspect towards “Indian food” is about the speaker’s past dislike for it. The underlying opinion towards “Indian food” is that the speaker used to hate it, indicating a negative past sentiment. Although the sentence also says that Baluchi’s changed the speaker’s mind, this contrastive cue indicates a shift from the speaker’s past negative view. Therefore, the sentiment polarity towards “Indian food” is negative . ✕
FACT-SR	Generated Rationale: The mentioned aspect towards “Indian food” is about the speaker’s previous negative impression of Indian food, which has been changed by Baluchi’s. The underlying opinion towards “Indian food” is that the speaker no longer hates it, indicating a positive change in opinion. Therefore, the sentiment polarity towards “Indian food” is positive . ✓

Case 2 illustrates factual hallucination. The input only says that the reviewer is deciding whether to choose the retina display, but DeepSeek-V3.2 introduces unsupported evaluative evidence such as “higher-quality”, “desirable”, and “worth potentially paying extra for”. These hallucinated cues turn a neutral option into a positive prediction. FACT-SR avoids adding absent sentiment evidence and assigns the neutral label.

Table 9 Case 2: factual hallucination. Highlighting and target-marking conventions follow Table 8.

Input Text	It’s been time for a new laptop, and the only debate was which size of the Mac laptops, and whether to spring for the retina display.
Aspect Terms	[size, retina display]
Gold Polarity	[neutral, neutral]
DeepSeek-V3.2	Generated Rationale: The mentioned aspect towards “retina display” is about the consideration of upgrading to a higher-quality screen option for a Mac laptop. The underlying opinion towards “retina display” is that it is presented as a desirable feature worth potentially paying extra for, indicating a positive inclination. Therefore, the sentiment polarity towards “retina display” is positive . ✕
FACT-SR	Generated Rationale: The mentioned aspect towards “retina display” is about its inclusion as an optional feature in a Mac laptop. The underlying opinion towards “retina display” is neutral, as it is simply listed among other options without any evaluative language indicating approval, disapproval, or preference. Therefore, the sentiment polarity towards “retina display” is neutral . ✓

Case 3 illustrates aspect confusion. The positive evidence belongs to “fresh juice concoctions”, while “food” is only the item being paired with the drink. DeepSeek-V3.2 transfers the positive clue to the wrong target aspect. FACT-SR keeps the rationale conditioned on “food” and therefore predicts the neutral label.

Table 10 Case 3: aspect confusion. Highlighting and target-marking conventions follow Table 8.

Input Text	Be sure to accompany your food with one of their fresh juice concoctions.
Aspect Terms	[food , fresh juice concoctions]
Gold Polarity	[neutral , positive]
DeepSeek-V3.2	Generated Rationale: The mentioned aspect towards “food” is about its pairing with fresh juice concoctions. The underlying opinion towards “food” is that it is implicitly recommended or enjoyed, as the suggestion to accompany it with a fresh juice indicates a positive dining experience. Therefore, the sentiment polarity towards “food” is positive . ✕
FACT-SR	Generated Rationale: The mentioned aspect towards “food” is about the recommendation to accompany the food with a fresh juice concoction. The underlying opinion towards the food is neutral, as it is mentioned neutrally without expressing a positive or negative judgment about the food itself. Therefore, the sentiment polarity towards “food” is neutral . ✓

6 Discussion

Implications for sentiment reasoning. The classification results, constraint ablations, and faithfulness analyses jointly support integrating rationale supervision with task-specific constraints. FACT-SR combines a shared backbone with separate Explain and Predict inputs, enabling rationale generation and direct polarity prediction within one training framework. The gains at 250M and 11B, together with the matched decoder-only results, show the utility of this design across the evaluated backbones. SI-AWL provides the accompanying mechanism for balancing the heterogeneous objectives.

Extensions to richer evidence. The present evaluation establishes the framework on the English SemEval-2014 Restaurant and Laptop benchmarks. Recent multimodal aspect-level sentiment studies model fine-grained visual evidence, cognitive cues, and natural-language explanations (Wang et al. 2026; Xiao et al. 2025). Extending FACT-SR to multilingual and multimodal inputs offers a natural next step: the same emphasis on logical support, source grounding, and aspect attribution can guide the design of evidence-specific objectives.

Adaptive reasoning structures. The three-hop format provides a standardized interface for rationale supervision and comparison with THOR and RVISA. Adaptive reasoning offers a complementary direction for extending this interface. AdaptThink (Zhang et al. 2025a) learns when to engage explicit thinking and when to bypass it, while CoLaR (Tan et al. 2025) adapts reasoning length through latent-space

compression. Combining such mechanisms with faithfulness-aware training could allow the number or granularity of reasoning steps to vary with instance difficulty.

7 Conclusion

We introduce FACT-SR, a scale-invariant multi-task framework for faithful aspect-based sentiment reasoning. FACT-SR constructs rationale supervision through label-consistency verification and aspect-conditioned sufficiency filtering, then jointly trains a student for rationale generation, polarity prediction, and three faithfulness constraints. SI-AWL balances the heterogeneous objectives through EMA loss normalization and bounded uncertainty mapping. Experiments on the SemEval-2014 Restaurant and Laptop benchmarks show improved classification and automatic faithfulness metrics at both Flan-T5 model scales. Constraint ablations support the contribution of the auxiliary objectives, and the weighting comparison shows consistent gains over AWL. In the Flan-T5 implementation, these auxiliary objectives are applied during training without adding constraint-specific inference modules. A matched Qwen2.5-7B-Instruct experiment further demonstrates the classification benefits of the decoder-only realization. Together, these results support incorporating faithfulness into the training objectives of aspect-based sentiment reasoning.

A Prompt Templates for Data Construction

We reproduce the construction prompts. In these templates, `{text}` is the input sentence, `{aspect}` is the target aspect, and `{polarity}` is the gold sentiment label supplied in TH-RA. The recovery prompt omits the gold label from its input, although the check subsequently compares its output with that label.

System prompt.

You are an expert in aspect-based sentiment analysis. Your task is to analyze the sentiment polarity (positive, negative, or neutral) towards a specific aspect term in a given sentence. Always follow the given output format strictly.

TH-RE: Three-Hop Reasoning. TH-RE generates an initial rationale and polarity from the sentence and aspect without receiving the gold label.

Given the sentence “`{text}`”, what is the sentiment polarity towards “`{aspect}`”, why? Let’s think step by step.
The mentioned aspect towards “`{aspect}`” is ... The underlying opinion towards “`{aspect}`” is ... Therefore, the sentiment polarity towards “`{aspect}`” is ...

TH-RA: Three-Hop Rationalization. TH-RA provides the gold label and is used to rewrite or reconstruct rationales for low-quality samples identified during the filtering stage.

Given the sentence “`{text}`”, the sentiment polarity towards “`{aspect}`” is `{polarity}`, why? Let’s think step by step.
The mentioned aspect towards “`{aspect}`” is ... The underlying opinion towards “`{aspect}`” is ... Therefore, the sentiment polarity towards “`{aspect}`” is ...

Aspect-conditioned rationale sufficiency prompt. The polarity term in Hop-3 is masked. The construction LLM receives the sentence, aspect, Hop-1, Hop-2, and masked Hop-3, and infers the hidden polarity from the intermediate reasoning. Its output is compared with the gold label to determine whether the candidate passes the sufficiency check.

Sentence: “{text}”

Aspect: “{aspect}”

Reasoning chain (with masked polarity): {masked_rationale}

Based on the reasoning above, the masked sentiment polarity [MASK] should be:

References

- Bang Y, Ji Z, Schelten A, et al (2025) HalluLens: LLM hallucination benchmark. In: Che W, Nabende J, Shutova E, et al (eds) Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Vienna, Austria, pp 24128–24156, <https://doi.org/10.18653/v1/2025.acl-long.1176>, URL <https://aclanthology.org/2025.acl-long.1176/>
- Chung HW, Hou L, Longpre S, et al (2024) Scaling instruction-finetuned language models. Journal of Machine Learning Research 25(70):1–53. URL <https://jmlr.org/papers/v25/23-0870.html>
- Devlin J, Chang MW, Lee K, et al (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) 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). Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186, <https://doi.org/10.18653/v1/N19-1423>, URL <https://aclanthology.org/N19-1423/>
- DeYoung J, Jain S, Rajani NF, et al (2020) ERASER: A benchmark to evaluate rationalized NLP models. In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 4443–4458, <https://doi.org/10.18653/v1/2020.acl-main.408>, URL <https://aclanthology.org/2020.acl-main.408/>
- Fei H, Li B, Liu Q, et al (2023) Reasoning implicit sentiment with chain-of-thought prompting. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada, pp 1171–1182, <https://doi.org/10.18653/v1/2023.acl-short.101>, URL <https://aclanthology.org/2023.acl-short.101/>
- Golovneva O, Chen M, Poff S, et al (2023) ROSCOE: A suite of metrics for scoring step-by-step reasoning. URL <https://arxiv.org/abs/2212.07919>, arXiv:2212.07919

- Ho N, Schmid L, Yun SY (2023) Large language models are reasoning teachers. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 14852–14882, <https://doi.org/10.18653/v1/2023.acl-long.830>, URL <https://aclanthology.org/2023.acl-long.830/>
- Hu EJ, Shen Y, Wallis P, et al (2021) LoRA: Low-rank adaptation of large language models. URL <https://arxiv.org/abs/2106.09685>, arXiv:2106.09685
- Jacovi A, Goldberg Y (2020) Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Jurafsky D, Chai J, Schluter N, et al (eds) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 4198–4205, <https://doi.org/10.18653/v1/2020.acl-main.386>, URL <https://aclanthology.org/2020.acl-main.386/>
- Kendall A, Gal Y, Cipolla R (2018) Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7482–7491, <https://doi.org/10.1109/CVPR.2018.00781>, URL https://openaccess.thecvf.com/content_cvpr_2018/html/Kendall_Multi-Task_Learning_Using_CVPR_2018_paper.html
- Lai W, Xie H, Xu G, et al (2025) RVISA: Reasoning and verification for implicit sentiment analysis. IEEE Transactions on Affective Computing 16(3):1760–1771. <https://doi.org/10.1109/TAFFC.2025.3537799>
- Lai W, Xie H, Xu G, et al (2026) Multi-task learning with LLMs for implicit sentiment analysis: Data-level and task-level automatic weight learning. IEEE Transactions on Knowledge and Data Engineering 38(1):506–517. <https://doi.org/10.1109/TKDE.2025.3623941>, URL <https://doi.ieeecomputersociety.org/10.1109/TKDE.2025.3623941>
- Laurer M, van Atteveldt W, Casas A, et al (2024) Building efficient universal classifiers with natural language inference. URL <https://arxiv.org/abs/2312.17543>, arXiv:2312.17543
- Li S, Chen J, Shen Y, et al (2022) Explanations from large language models make small reasoners better. URL <https://arxiv.org/abs/2210.06726>, arXiv:2210.06726
- Li Z, Zou Y, Zhang C, et al (2021) Learning implicit sentiment in aspect-based sentiment analysis with supervised contrastive pre-training. In: Moens MF, Huang X, Specia L, et al (eds) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pp 246–256, <https://doi.org/10.18653/v1/2021.emnlp-main.22>, URL <https://aclanthology.org/2021.emnlp-main.22/>

- Liebel L, Körner M (2018) Auxiliary tasks in multi-task learning. URL <https://arxiv.org/abs/1805.06334>, [arXiv:1805.06334](https://arxiv.org/abs/1805.06334)
- Lyu Q, Havaladar S, Stein A, et al (2023) Faithful chain-of-thought reasoning. In: Park JC, Arase Y, Hu B, et al (eds) Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Nusa Dua, Bali, pp 305–329, <https://doi.org/10.18653/v1/2023.ijcnlp-main.20>, URL <https://aclanthology.org/2023.ijcnlp-main.20/>
- van der Maaten L, Hinton G (2008) Visualizing data using t-SNE. Journal of Machine Learning Research 9(86):2579–2605. URL <https://jmlr.org/papers/v9/vandemaaten08a.html>
- Magister LC, Mallinson J, Adamek J, et al (2023) Teaching small language models to reason. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Association for Computational Linguistics, Toronto, Canada, pp 1773–1781, <https://doi.org/10.18653/v1/2023.acl-short.151>, URL <https://aclanthology.org/2023.acl-short.151/>
- Ouyang J, Yang Z, Liang S, et al (2024) Aspect-based sentiment analysis with explicit sentiment augmentations. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 38. AAAI Press, pp 18842–18850, <https://doi.org/10.1609/aaai.v38i17.29849>, URL <https://doi.org/10.1609/aaai.v38i17.29849>
- Parcalabescu L, Frank A (2024) On measuring faithfulness or self-consistency of natural language explanations. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Bangkok, Thailand, pp 6048–6089, <https://doi.org/10.18653/v1/2024.acl-long.329>, URL <https://aclanthology.org/2024.acl-long.329/>
- Peng H, Xu L, Bing L, et al (2020) Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 8600–8607, <https://doi.org/10.1609/aaai.v34i05.6383>, URL <https://ojs.aaai.org/index.php/AAAI/article/view/6383>
- Pontiki M, Galanis D, Pavlopoulos J, et al (2014) SemEval-2014 task 4: Aspect based sentiment analysis. In: Nakov P, Zesch T (eds) Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014). Association for Computational Linguistics, Dublin, Ireland, pp 27–35, <https://doi.org/10.3115/v1/S14-2004>, URL <https://aclanthology.org/S14-2004/>
- Rietzler A, Stabinger S, Opitz P, et al (2020) Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In: Calzolari N, Béchet F, Blache P, et al (eds) Proceedings of

- the Twelfth Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp 4933–4941, URL <https://aclanthology.org/2020.lrec-1.607/>
- Schmitt M, Steinheber S, Schreiber K, et al (2018) Joint aspect and polarity classification for aspect-based sentiment analysis with end-to-end neural networks. In: Riloff E, Chiang D, Hockenmaier J, et al (eds) Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Brussels, Belgium, pp 1109–1114, <https://doi.org/10.18653/v1/D18-1139>, URL <https://aclanthology.org/D18-1139/>
- Tan W, Li J, Ju J, et al (2025) Think silently, think fast: Dynamic latent compression of LLM reasoning chains. In: Belgrave D, Zhang C, Lin H, et al (eds) Advances in Neural Information Processing Systems, vol 38. Curran Associates, Inc., pp 4646–4668, URL https://proceedings.neurips.cc/paper_files/paper/2025/file/0706261aedab63814a2b73c32564b4c4-Paper-Conference.pdf
- Tang L, Laban P, Durrett G (2024) MiniCheck: Efficient fact-checking of LLMs on grounding documents. In: Al-Onaizan Y, Bansal M, Chen YN (eds) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Miami, Florida, USA, pp 8818–8847, <https://doi.org/10.18653/v1/2024.emnlp-main.499>, URL <https://aclanthology.org/2024.emnlp-main.499/>
- Tutek M, Hashemi Chaleshtori F, Marasovic A, et al (2025) Measuring chain of thought faithfulness by unlearning reasoning steps. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 9935–9960, <https://doi.org/10.18653/v1/2025.emnlp-main.504>, URL <https://aclanthology.org/2025.emnlp-main.504/>
- Wang K, Shen W, Yang Y, et al (2020) Relational graph attention network for aspect-based sentiment analysis. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, pp 3229–3238, <https://doi.org/10.18653/v1/2020.acl-main.295>, URL <https://aclanthology.org/2020.acl-main.295/>
- Wang S, Zhou J, Sun C, et al (2022) Causal intervention improves implicit sentiment analysis. In: Calzolari N, Huang CR, Kim H, et al (eds) Proceedings of the 29th International Conference on Computational Linguistics. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pp 6966–6977, URL <https://aclanthology.org/2022.coling-1.607/>
- Wang Y, Chen Q, Wang W (2021) Multi-task BERT for aspect-based sentiment analysis. In: 2021 IEEE International Conference on Smart Computing (SMARTCOMP), pp 383–385, <https://doi.org/10.1109/SMARTCOMP52413.2021.00077>

- Wang Z, Tian Y, Wang H, et al (2026) Explainable multimodal aspect-based sentiment analysis with dependency-guided large language model. URL <https://arxiv.org/abs/2601.06848>, arXiv:2601.06848
- Wei J, Wang X, Schuurmans D, et al (2022) Chain-of-thought prompting elicits reasoning in large language models. In: Koyejo S, Mohamed S, Agarwal A, et al (eds) Advances in Neural Information Processing Systems, vol 35. Curran Associates, Inc., pp 24824–24837, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf
- Xiao L, Mao R, Zhao S, et al (2025) Exploring cognitive and aesthetic causality for multimodal aspect-based sentiment analysis. IEEE Transactions on Affective Computing 16(4):3248–3265. <https://doi.org/10.1109/TAFFC.2025.3565506>, URL <https://doi.ieeecomputersociety.org/10.1109/TAFFC.2025.3565506>
- Yang A, Yang B, Zhang B, et al (2024) Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 URL <https://arxiv.org/abs/2412.15115>
- Ye X, Durrett G (2022) The unreliability of explanations in few-shot prompting for textual reasoning. In: Koyejo S, Mohamed S, Agarwal A, et al (eds) Advances in Neural Information Processing Systems, vol 35. Curran Associates, Inc., pp 30378–30392, URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c402501846f9fe03e2cac015b3f0e6b1-Paper-Conference.pdf
- Zha Y, Yang Y, Li R, et al (2023) AlignScore: Evaluating factual consistency with a unified alignment function. In: Rogers A, Boyd-Graber J, Okazaki N (eds) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Toronto, Canada, pp 11328–11348, <https://doi.org/10.18653/v1/2023.acl-long.634>, URL <https://aclanthology.org/2023.acl-long.634/>
- Zhang J, Lin N, Hou L, et al (2025a) AdaptThink: Reasoning models can learn when to think. In: Christodoulopoulos C, Chakraborty T, Rose C, et al (eds) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Suzhou, China, pp 3716–3730, <https://doi.org/10.18653/v1/2025.emnlp-main.184>, URL <https://aclanthology.org/2025.emnlp-main.184/>
- Zhang W, Li X, Deng Y, et al (2023) A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. IEEE Transactions on Knowledge and Data Engineering 35(11):11019–11038. <https://doi.org/10.1109/TKDE.2022.3230975>, URL <https://doi.ieeecomputersociety.org/10.1109/TKDE.2022.3230975>
- Zhang Y, Li Y, Cui L, et al (2025b) Siren’s song in the AI ocean: A survey on hallucination in large language models. Computational Linguistics 51(4):1373–1418. <https://doi.org/10.1162/coli.a.16>, URL <https://aclanthology.org/2025.cl-4.9/>

Data and Code Availability

The study uses the SemEval-2014 Restaurant and Laptop benchmarks (Pontiki et al. 2014). The processed rationale data, implementation, and configuration files are available for peer review at <https://anonymous.4open.science/r/FACT2026>.

Use of AI-Assisted Tools

The authors used AI assistants for language editing, coding assistance, and figure preparation. The authors reviewed and approved the final manuscript and take responsibility for its content.