# CLAOCS-TX: Cross-Lingual Triplet Extraction with Aspect-Opinion-Aware Code-Switched Prompting and LLM-Guided Contrastive Distillation

**Lipika Dewangan**[1], **Chandresh Kumar Maurya**[1]
[1]Indian Institute of Technology Indore, Madhya Pradesh, India 453552
{phd2101201004, chandresh}@iiti.ac.in

## Abstract

Cross-lingual learning enables the transfer of structured sentiment knowledge from high-resource languages to unlabeled or low-resource languages, but prior work has largely focused on coarse-grained sentiment classification or aspect extraction. In contrast, zero-shot cross-lingual aspect–opinion–sentiment triplet extraction (ASTE), which extracts sentiment triplets of the form (*aspect term*, *opinion term*, *sentiment polarity*), remains underexplored. We propose a unified framework that leverages large language models (LLMs) as both structured pseudo-label generators and semantic teachers for ASTE. Our approach employs stepwise structured prompting over aspect- and opinion-aware code-switched variants to generate reliable pseudo triplets, followed by a multi-variant consistency filter to retain high-confidence supervision. We further introduce a triplet-aware contrastive distillation objective that aligns student triplet representations with LLM-encoded semantic embeddings. During inference, only the student ASTE model is used, without requiring LLM access. Experiments on four non-Indic and four low-resource Indic target languages show consistent improvements over strong cross-lingual and LLM-based baselines. The proposed method yields an absolute micro-F1 improvement of 5.3 points on non-Indic languages and 3.8 points on low-resource Indic languages compared to the best competing approach. Ablation results further validate the complementary roles of aspect- and opinion-aware code-switched prompting and triplet-aware contrastive distillation, with larger relative gains observed in low-resource Indic settings.

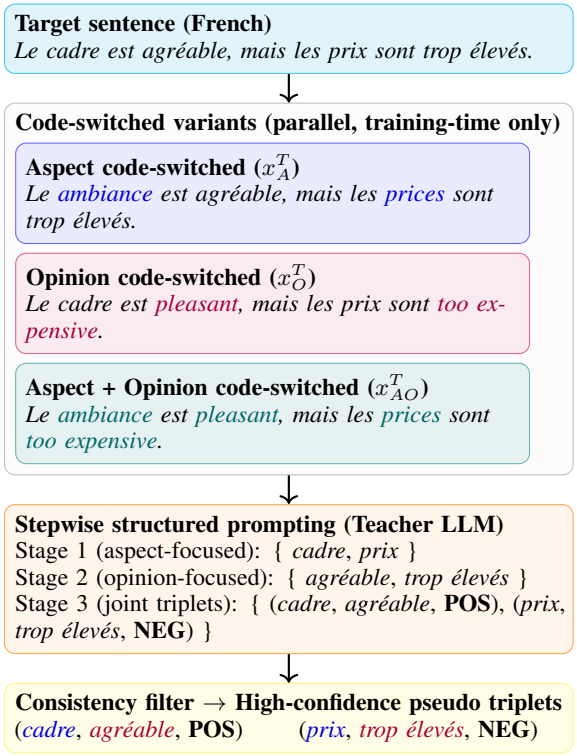


Figure 1: Illustration of training-time pseudo-triplet generation in the restaurant domain. Aspect-only (blue), opinion-only (purple), and joint (teal) code-switched variants are constructed in parallel to support stepwise Teacher LLM prompting and consistency filtering. English translation: "The ambiance is pleasant, but the prices are too expensive."

## 1 Introduction

Aspect-based sentiment analysis (ABSA) aims to identify the sentiment polarity expressed toward specific aspects within a sentence and has been widely studied in natural language processing (Pontiki et al., 2016; Zhang et al., 2018). A more fine-grained formulation is aspect–opinion–sentiment triplet extraction (ASTE), which extracts structured triplets consisting of an aspect term, an opinion term, and their associated sentiment polarity, thereby providing a structured and interpretable representation of sentiment (Peng et al., 2020; Su et al., 2024). For example, in the sentence "The decor feels modern, but the seating is uncomfortable," ASTE



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 44561–44577

extracts the triplets $(decor, modern, \text{POS})$ and $(seating, uncomfortable, \text{NEG})$. Most existing ASTE methods are developed under fully supervised settings and assume the availability of large-scale annotated data. In practice, however, such annotations are available only for a small number of high-resource languages such as English, while most languages lack labeled resources. This limitation has motivated extensive research on cross-lingual aspect-based sentiment analysis (XABSA), where a model trained on a labeled source language is transferred to unlabeled or low-resource target languages (Balamurali et al., 2012; Fei and Li, 2020; Lin et al., 2023). For example, a model trained on English reviews to identify sentiment toward aspects such as *service* or *food* can be applied to French or Spanish reviews without relying on target-language annotations (an illustrative example is shown in Figure 1). However, most cross-lingual ABSA approaches focus on coarse-grained sentiment classification, aspect extraction, or aspect-level polarity transfer rather than full triplet extraction, and therefore do not directly model the structured coupling between aspect spans, opinion spans, and sentiment labels required by ASTE (Jebbara and Cimiano, 2019; Zhang et al., 2021). Extending these approaches to ASTE is non-trivial. Cross-lingual ASTE must not only detect sentiment-bearing spans across languages, but also recover correct aspect–opinion pairings and exact span boundaries, both of which are highly sensitive to lexical, morphological, and syntactic variation (Šmíd and Kral, 2025). Moreover, the absence of target-language annotations forces models to rely on weak or noisy supervision, which can amplify errors in boundary detection, span pairing, and sentiment assignment (Liu et al., 2024). These challenges limit the effectiveness of existing cross-lingual ABSA techniques when applied directly to triplet extraction.

Recent advances in large language models (LLMs) offer new opportunities to address these limitations. LLMs exhibit strong multilingual semantic understanding and can generate structured outputs via prompting, making them attractive for pseudo-label generation in low-resource settings (Brown et al., 2020; Wei et al., 2022; Li et al., 2023). However, directly training on LLM-generated pseudo labels is often unreliable, as the generated triplets may be inconsistent, hallucinated, or misaligned with the input sentence, particularly in low-resource or morphologically rich languages (Bang et al., 2023). In addition, most existing approaches treat LLM outputs as hard supervision and do not fully exploit the semantic representations encoded within LLMs to guide cross-lingual transfer (Hsieh et al., 2023).

In this work, we propose a unified framework for zero-shot cross-lingual ASTE that combines LLM-based pseudo-triplet generation with triplet-aware contrastive distillation in a structured and complementary manner. We introduce *aspect- and opinion-aware code-switched prompting* to anchor key semantic spans in English while preserving the surrounding target-language context, and enforce agreement across multiple prompt variants to improve the reliability of pseudo triplets. We further propose a *triplet-aware contrastive distillation* objective that aligns structured triplet representations between the student model and the LLM, enabling effective cross-lingual transfer without relying on target-language annotations or probability-level distillation.

The key contributions of our study are:

- We propose a unified framework for zero-shot cross-lingual aspect–opinion–sentiment triplet extraction that leverages large language models as both pseudo-label generators and semantic teachers.
- We introduce aspect- and opinion-aware code-switched prompting to improve the reliability of LLM-generated pseudo triplets in low-resource target languages.
- A triplet-aware contrastive distillation objective is employed that aligns student and LLM representations at the structured triplet level for cross-lingual semantic transfer.
- We evaluate the proposed framework on eight target languages (four non-Indic and four low-resource Indic), demonstrating consistent improvements over strong baselines and validating each component through extensive ablation studies. The data and code used in this study are shared for future research.[1]

**Research Questions.** We evaluate the proposed framework by addressing the following questions: **RQ1:** How effective is the proposed framework for zero-shot cross-lingual aspect–opinion–sentiment

[1] https://github.com/Lipika-Dewangan/CLAOCS-TX-Cross-Lingual-Triplet

triplet extraction? **RQ2:** Does aspect- and opinion-aware code-switched prompting improve the quality of LLM-generated pseudo triplets? **RQ3:** What is the contribution of triplet-aware contrastive distillation to cross-lingual semantic alignment? **RQ4:** How well the framework generalizes across both non-Indic and low-resource Indic languages?

## 2 Related Work

Aspect-based sentiment analysis (ABSA) has been widely studied as a fine-grained alternative to sentence-level sentiment classification, with early benchmarks established through the SemEval shared tasks (Pontiki et al., 2016) and subsequent surveys categorizing ABSA into aspect extraction, sentiment classification, and joint modeling paradigms (Zhang et al., 2018). More recently, aspect–opinion–sentiment triplet extraction (ASTE) has emerged as a unified formulation that jointly models aspect terms, opinion terms, and sentiment polarity (Peng et al., 2020). Due to its structured nature, ASTE has been addressed using grid-based tagging (Šmíd and Kral, 2025), span-based models (Liang et al., 2023), and other structured prediction approaches. While effective in supervised settings, these methods rely on extensive labeled data and do not readily generalize to low-resource or cross-lingual scenarios (Jebbara and Cimiano, 2019). To address annotation scarcity, cross-lingual ABSA aims to transfer knowledge from a labeled source language to unlabeled or low-resource target languages. Early approaches relied on machine translation (Lambert, 2015) or cross-lingual word embeddings (Akhtar et al., 2018; Jebbara and Cimiano, 2019),while more recent work leverages multilingual pretrained language models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). Further improvements have been explored through parameter warm-up (Li et al., 2020), alignment-free label projection with distillation (Zhang et al., 2021), contrastive learning for semantic alignment (Lin et al., 2023), and dynamic loss weighting to address class imbalance (Lin et al., 2024). Although large language models (LLMs) generally underperform task-specific models for ABSA (Gou et al., 2023; Zhang et al., 2024), fine-tuned LLaMA-based models show strong monolingual performance (Šmíd and Kral, 2025). As a result, recent work primarily employs LLMs for pseudo-label generation or data augmentation rather than direct inference (Møller et al., 2024). Beyond implicit alignment, prior studies have introduced aspect code-switching to anchor semantic spans across languages (Zhang et al., 2021) and contrastive objectives for representation alignment (Lin et al., 2023). In parallel, LLMs have been explored as weak supervisors for pseudo-label generation (Brown et al., 2020; Wei et al., 2022; Li et al., 2023), though such labels may suffer from hallucination and inconsistency in low-resource settings (Bang et al., 2023). Distillation-based methods further enable knowledge transfer from LLMs to smaller models (Hsieh et al., 2023), with contrastive distillation providing representation-level alignment without relying on probability outputs (Tian et al., 2020). However, existing cross-lingual ABSA and LLM-based approaches largely operate at the token, aspect, or sentence level and do not explicitly model structured triplet semantics. In contrast, our work addresses cross-lingual ASTE by combining aspect- and opinion-aware code-switched prompting with triplet-aware contrastive distillation, enabling structured semantic alignment under a fully zero-shot target-language setting.

## 3 Methodology

We describe our proposed approach in this section. As illustrated in Figure 2, the proposed framework consists of three components: (i) LLM-based pseudo triplet generation using aspect- and opinion-aware code-switched prompting, (ii) a student ASTE model trained with English supervision and filtered pseudo triplets, and (iii) triplet-aware contrastive distillation for structured cross-lingual semantic alignment. The LLM plays a dual role as a pseudo-label generator and a semantic teacher during training, while inference relies solely on the student model. We provide the full step-by-step training and pseudo-label generation algorithm and the prompt template in Appendix A.

### 3.1 Problem Definition

We study cross-lingual aspect–opinion–sentiment triplet extraction (ASTE) under a zero-shot target setting. Let the labeled English source dataset be

$$\mathcal{D}_S = \{(x_i^S, \mathcal{T}_i^S)\}_{i=1}^N, \tag{1}$$

where $x_i^S = (w_1, \ldots, w_n)$ is a sentence and

$$\mathcal{T}_i^S = \{(a, o, y)\} \tag{2}$$

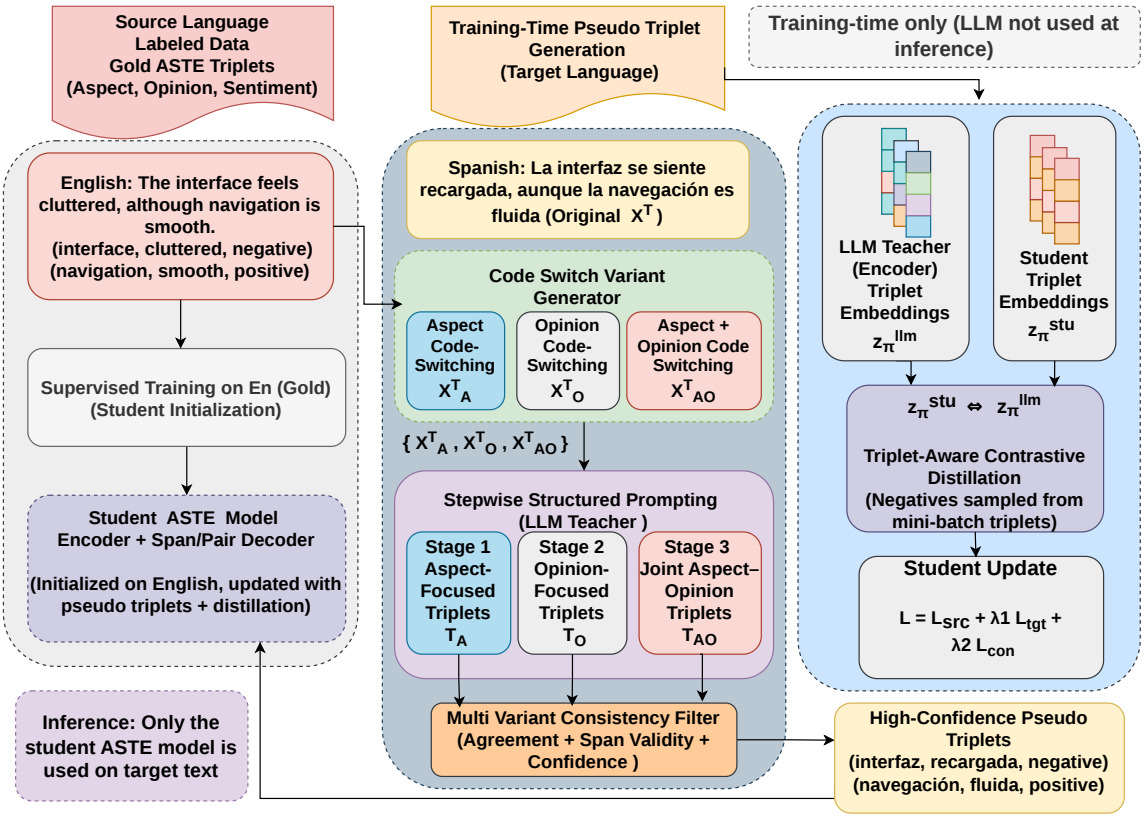


Figure 2: Overview of the proposed cross-lingual ASTE framework. During training, an LLM generates high-confidence pseudo triplets via aspect- and opinion-aware code-switched prompting and guides the student model through triplet-aware contrastive distillation. At inference time, only the student ASTE model is used.

denotes its gold triplets, with aspect span $a$, opinion span $o$, and sentiment polarity $y \in \{\text{POS}, \text{NEG}, \text{NEU}\}$.

Given an unlabeled target-language corpus

$$\mathcal{D}_T = \{x_j^T\}_{j=1}^M, \tag{3}$$

our goal is to learn a student model $M_\theta$ that predicts triplets $\widehat{\mathcal{T}}(x^T)$ without using any target-language annotations.

### 3.2 Student Triplet Extraction Model

**Encoder.** Each sentence $x = (w_1, \ldots, w_n)$ is encoded using XLM-R:

$$\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_n) = \text{Enc}(x), \tag{4}$$

where $\mathbf{h}_i \in \mathbb{R}^d$.

**Aspect and Opinion Tagging.** We adopt a BIOES tagging scheme (Peng et al., 2020; Xu et al., 2020) for aspect and opinion span extraction, as it provides explicit boundary supervision and is widely adopted for improving span precision in structured extraction tasks such as ASTE, especially under noisy or weak supervision (Liang et al., 2023). Aspect and opinion spans are predicted using independent BIOES taggers:

$$\mathbf{p}_i^a = \text{softmax}(\mathbf{W}_a \mathbf{h}_i + \mathbf{b}_a), \tag{5}$$

$$\mathbf{p}_i^o = \text{softmax}(\mathbf{W}_o \mathbf{h}_i + \mathbf{b}_o). \tag{6}$$

The token-level losses are:

$$\mathcal{L}_{\text{tag}}^a = -\sum_i \log p_i^a(y_i^a), \tag{7}$$

$$\mathcal{L}_{\text{tag}}^o = -\sum_i \log p_i^o(y_i^o). \tag{8}$$

**Constrained Decoding.** To ensure valid BIOES sequences, decoding is performed under transition constraints to get well-formed spans and reduce invalid boundary predictions (Šmíd et al., 2025a).

$$\mathbf{y}^* = \arg\max_{\mathbf{y} \in \mathcal{C}} \sum_{i=1}^n \log p_i(y_i) + \sum_{i=2}^n \log A_{y_{i-1}, y_i}, \tag{9}$$

where $\mathcal{C}$ is the set of valid BIOES sequences and $A$ is a transition mask.

**Span Representation.** Decoded spans are pooled via mean pooling:

$$\mathbf{r}_a = \frac{1}{|a|} \sum_{i \in a} \mathbf{h}_i, \tag{10}$$

$$\mathbf{r}_o = \frac{1}{|o|} \sum_{i \in o} \mathbf{h}_i. \tag{11}$$

We additionally compute a sentence-level context vector:

$$\mathbf{r}_{\text{ctx}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{h}_i. \tag{12}$$

#### 3.2.1 Aspect–Opinion Pairing and Sentiment Classification

Projected span representations are computed as:

$$\tilde{\mathbf{r}}_a = \phi(\mathbf{W}_{pa}\mathbf{r}_a + \mathbf{b}_{pa}), \tag{13}$$

$$\tilde{\mathbf{r}}_o = \phi(\mathbf{W}_{po}\mathbf{r}_o + \mathbf{b}_{po}), \tag{14}$$

where $\phi(\cdot)$ is a non-linear activation.

A biaffine scorer determines valid aspect–opinion pairs:

$$s(a, o) = \tilde{\mathbf{r}}_a^\top \mathbf{U} \tilde{\mathbf{r}}_o + \mathbf{w}^\top [\tilde{\mathbf{r}}_a; \tilde{\mathbf{r}}_o] + b. \tag{15}$$

The pair probability is:

$$p_{\text{pair}}(a, o) = \sigma(s(a, o)). \tag{16}$$

Sentiment polarity is predicted as:

$$\mathbf{g}_{ao} = [\mathbf{r}_a; \mathbf{r}_o; \mathbf{r}_{\text{ctx}}], \tag{17}$$

$$\mathbf{p}^y_{ao} = \text{softmax}(\mathbf{W}_y \mathbf{g}_{ao} + \mathbf{b}_y), \tag{18}$$

Corresponding losses are:

$$\begin{aligned} \mathcal{L}_{\text{pair}} = -\sum_{(a,o)} \Big[ & y_{ao} \log p_{\text{pair}}(a, o) \\ & + (1 - y_{ao}) \log \big(1 - p_{\text{pair}}(a, o)\big) \Big] \end{aligned} \tag{19}$$

### 3.3 Aspect- and Opinion-Aware Code-Switched Prompting

For each target sentence $x^T$, we construct four variants:

$$\mathcal{V}(x^T) = \{x^T, x^T_A, x^T_O, x^T_{AO}\} \tag{20}$$

where $x^T$ denotes the original target-language sentence, $x^T_A$ the aspect-switched variant, $x^T_O$ the opinion-switched variant, and $x^T_{AO}$ the joint aspect-and-opinion-switched variant, as illustrated in Figure 2. Candidate aspect and opinion spans $(C_a, C_o)$ are obtained from the union of (i) low-threshold student predictions and (ii) simple high-recall lexical heuristics. In practice, contiguous noun or noun-phrase spans are treated as aspect candidates, while contiguous adjectival, adverbial, and sentiment-bearing phrase spans are treated as opinion candidates. Duplicate, empty, and punctuation-only spans are removed before prompt construction. Let $\tau(\cdot)$ denote span translation into English. The translation function $\tau(\cdot)$ is applied only to the selected candidate span rather than to the full sentence, so that the surrounding sentence context remains in the original target language during code-switched prompt construction. Code-switching replaces only the selected spans while preserving the surrounding target-language context. These variants are used exclusively for LLM prompting and are never provided to the student model as training input.

### 3.4 LLM-Based Pseudo Triplet Generation and Filtering

For each variant $v \in \mathcal{V}(x^T)$, we sample $K$ structured outputs:

$$\mathcal{P}(x^T) = \bigcup_{v} \bigcup_{k=1}^{K} G(v, k), \tag{21}$$

where $G$ denotes the LLM. All extracted triplets are mapped back to the original sentence using a span alignment function $\Psi$. Let $\pi = (a, o, y)$ be a normalized triplet. Its support count is:

$$c(\pi) = \sum_{v,k} \mathbb{I}[\pi \in \Psi(G(v, k))]. \tag{22}$$

A triplet is retained if:

$$c(\pi) \geq 2 \ \wedge \ |\mathcal{S}(\pi)| \geq 2 \ \wedge \ \text{SpanValid}(\pi), \tag{23}$$

where $\mathcal{S}(\pi)$ is the set of variants supporting $\pi$.

The agreement-based weight is:

$$w(\pi) = \frac{c(\pi)}{\max_{\pi'} c(\pi')}. \tag{24}$$

The exact prompt template used for pseudo-triplet generation is provided in Figure 5 in Appendix A.

### 3.5 Triplet-Aware Contrastive Distillation

For each retained triplet $\pi$, the student representation is:

$$\mathbf{z}^{\text{stu}}_\pi = f([\mathbf{r}_a; \mathbf{r}_o; \mathbf{r}_{\text{ctx}}]), \tag{25}$$

where $f(\cdot)$ is a projection MLP.

The LLM encodes a canonical textual prompt of the triplet:

$$\mathbf{z}^{\text{llm}}_\pi = \text{Pool}(\text{LLMEnc}(\text{Prompt}(\pi))). \tag{26}$$

Using cosine similarity, the InfoNCE loss is:

$$\mathcal{L}_{\text{con}} = -\sum_{\pi} \log \frac{\exp(\text{sim}(\mathbf{z}_{\pi}^{\text{stu}}, \mathbf{z}_{\pi}^{\text{llm}})/\gamma)}{\sum_{\pi'} \exp(\text{sim}(\mathbf{z}_{\pi}^{\text{stu}}, \mathbf{z}_{\pi'}^{\text{llm}})/\gamma)}. \tag{27}$$

**Why triplet-level distillation.** Token-level or probability-based distillation assumes a fixed output space, which is ill-suited for aspect–opinion–sentiment triplet extraction due to the variable number of triplets per sentence. In cross-lingual settings, token-level alignment is further degraded by span boundary noise and morphological variation across languages. We therefore perform distillation directly at the triplet level, aligning structured aspect–opinion–sentiment representations between the student model and a frozen LLM teacher. Specifically, we employ an InfoNCE-based contrastive objective (Sun et al., 2020; Tan et al., 2023) that pulls student triplet embeddings toward their corresponding LLM-encoded embeddings while pushing them away from other triplets in the mini-batch. By operating on structured triplets rather than token-level predictions or probability distributions, this formulation naturally accommodates variable output cardinality and enables more stable cross-lingual semantic alignment. Such contrastive distillation has also proven effective for transferring knowledge from large language model teachers (Katz-Samuels et al., 2024), while preserving triplet structure under noisy pseudo supervision.

### 3.6 Training Objective

At a high level, training proceeds in three stages. First, the student model is initialized on labeled English source data. Second, for each unlabeled target-language sentence, candidate aspect and opinion spans are proposed, code-switched prompt variants are constructed, and the teacher LLM generates multiple candidate triplets, from which only agreement-supported pseudo triplets are retained. Third, the student model is updated using source supervision, filtered pseudo-triplet supervision, and triplet-aware contrastive distillation. This pseudo-label generation and student optimization cycle is repeated for a small number of rounds. The final objective combines source supervision, pseudo-triplet supervision on unlabeled target-language sentences, and triplet-aware contrastive alignment:

$$\mathcal{L} = \mathcal{L}_{\text{src}} + \lambda_1 \mathcal{L}_{\text{tgt}} + \lambda_2 \mathcal{L}_{\text{con}}. \tag{28}$$

Here, $\mathcal{L}_{\text{src}}$ is computed on labeled English source data, while $\mathcal{L}_{\text{tgt}}$ is computed on the filtered pseudo triplets retained from the target-language corpus. The overall training procedure is summarized above and detailed in Algorithm 1.

## 4 Experimental Setup

### 4.1 Dataset

We evaluate the proposed framework on eight target languages in two settings: four non-Indic languages and four low-resource Indic languages. For the non-Indic setting, we use the multilingual restaurant-review benchmark derived from SemEval-2016 Task 5 (Pontiki et al., 2016), where English (En) is the labeled source language and French (Fr), Spanish (Es), Dutch (Nl), and Russian (Ru) are treated as target languages. We adopt the train/dev/test splits introduced by Zhang et al. (2021) and follow the evaluation protocol of Sheng et al. (2025) for comparability with prior cross-lingual ASTE work.

To further assess transfer under low-resource conditions, we additionally evaluate on four Indic languages: Hindi (Hi) (Akhtar et al., 2016), Marathi (Mr) (Joshi, 2022), Bengali (Bn) (Shimada, 2023), and Odia (Od) (Dewangan et al., 2025). English is the only source language in all experiments, while target-language gold labels are used strictly for evaluation. Dataset statistics are reported in Table 1. All target-language datasets are evaluated in the restaurant-review setting for consistency, including the Hindi benchmark.

The official SemEval-2016 Task 5 release (Pontiki et al., 2016) was originally developed for aspect-based sentiment analysis and does not provide explicit opinion term span annotations. These annotations became available only in later ASTE benchmark reformulations built on SemEval-style ABSA resources (Peng et al., 2020; Wan et al., 2020; Xu et al., 2020). Accordingly, we use the publicly available ASTE-Data-V1 (Peng et al., 2020) triplet-form benchmark together with the evaluation protocol adopted by Sheng et al. (2025). We do not newly derive, project, or re-annotate opinion terms in this work; instead, we follow the benchmark construction and evaluation practice established in the prior literature. Large language models are used only for pseudo-label generation during training and never to create or modify gold evaluation labels. Further details on dataset provenance are provided in Appendix B.

| Language | Train | Dev | Test (#s) | #a | #o |
|---|---|---|---|---|---|
| English (EN) | 1600 | 400 | 676 | 712 | 684 |
| French (FR) | 1664 | 332 | 668 | 701 | 672 |
| Spanish (ES) | 1656 | 414 | 881 | 938 | 904 |
| Dutch (NL) | 1378 | 344 | 575 | 598 | 563 |
| Russian (RU) | 2924 | 731 | 1209 | 1284 | 1216 |
| Hindi (HI) | 2100 | 520 | 882 | 1025 | 896 |
| Marathi (MR) | 1658 | 408 | 703 | 742 | 708 |
| Bengali (BN) | 1204 | 296 | 506 | 532 | 498 |
| Odia (OD) | 1439 | 356 | 608 | 639 | 602 |

Table 1: Dataset statistics across languages. #s, #a, and #o denote the numbers of sentences, aspects, and opinion terms. English is the source language, while the other languages are treated as targets. Train/dev/test counts are reported for completeness; final evaluation is conducted on the target-language test sets.

### 4.2 Implementation Details

The student ASTE model is based on XLM-R-large ($d = 1024$) (Conneau et al., 2020). Aspect and opinion spans are predicted using independent BIOES taggers with constrained decoding, followed by biaffine scorers for aspect–opinion pairing and sentiment classification. Span representations are obtained via mean pooling over token embeddings. We adopt LLaMA-3-8B-Instruct[2] as a frozen LLM, used only during training for pseudo-label generation and semantic supervision. Generation uses temperature 0.7, top-$p$ 0.9, with $K = 5$ samples per variant, and triplet embeddings are extracted from the final LLM layer via mean pooling. For each target sentence, up to four variants are constructed (original, aspect-switched, opinion-switched, and joint aspect-opinion-switched). Candidate spans are obtained using high-recall heuristics and low-threshold student predictions, translated into English, and used only for LLM prompting. Pseudo triplets are retained if they appear in at least two samples across two variants and align with contiguous spans in the original sentence, and are weighted by agreement frequency during training. Triplet-aware contrastive distillation is implemented using an InfoNCE objective with 256-dimensional projections and temperature 0.1, treating other triplets in the mini-batch as negatives. The student is trained on English data for 10 epochs, followed by two rounds of pseudo labeling and retraining with 5 epochs per round. We use the AdamW optimizer with a learning rate of $2 \times 10^{-5}$, weight decay of 0.01, batch size 16, and loss weights $\lambda_1 = 1.0$ and $\lambda_2 = 0.5$; a sensitivity analysis of these settings is provided in Table 3 in Appendix A. All experiments are conducted on a single NVIDIA A100 GPU, with the LLM loaded in 8-bit precision and triplet embeddings cached. We use a single teacher LLM in order to keep the training setup controlled and computationally tractable while isolating the effect of the proposed pseudo-labeling and triplet-aware distillation framework. The proposed method does not depend on model-specific logits or internal states; rather, the LLM is used as a generic pseudo-label generator and semantic teacher through textual outputs and triplet representations. We report detailed training settings only for the proposed framework and for baselines that were re-run in our experimental setup. Results cited directly from prior work are reported under their original published settings and are included as reference comparisons.

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

### 4.3 Evaluation Metrics

Following prior work (Šmíd et al., 2025c; Lin et al., 2023), we use micro-F1 as the evaluation metric. A prediction is considered correct only when both the extracted boundaries and the associated sentiment polarity are accurately identified. The average F1 score is reported over five runs with different random seeds.

### 4.4 Baseline Methods

We compare our approach with a broad range of cross-lingual ASTE baselines, including zero-shot, translation-based, alignment-free, contrastive learning, and LLM-driven methods. Because dedicated cross-lingual ASTE baselines remain limited, some comparison methods were originally proposed for related cross-lingual ABSA tasks rather than explicit aspect–opinion–sentiment triplet extraction. In such cases, we retain the original training protocol of each baseline as closely as possible and apply only the minimal task-consistent output adaptation required for evaluation under the adopted triplet-form benchmark. The zero-shot setting evaluates direct transfer using only labelled source-language data (Conneau et al., 2020). BILINGUAL-TA (Li et al., 2020) follows the translate-then-align paradigm by fine-tuning on translated data, with the latter additionally incorporating source-language supervision. ACS-DISTILL (Zhang et al., 2021) employs alignment-free projection and aspect code-switching, where distillation further leverages unlabelled target-language data. CL-XABSA (Lin

| Method | Non-English (Non-Indic) | | | | | Low-Resource Indic | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Es | Fr | Nl | Ru | Avg | Hi | Mr | Bn | Od | Avg |
| Bilingual-TA (Li et al., 2020) | 47.53 | 38.04 | 41.37 | 39.66 | 41.65 | 40.27 | 37.69 | 38.81 | 36.44 | 38.30 |
| ACS-Distill (Zhang et al., 2021) | 52.91 | 45.25 | 46.40 | 40.58 | 46.29 | 44.30 | 41.75 | 42.88 | 39.92 | 42.21 |
| CL-XABSA (TL) (Lin et al., 2023) | 53.62 | 48.50 | 50.64 | 42.65 | 48.85 | 46.98 | 43.46 | 45.18 | 42.66 | 44.57 |
| Equi-XABSA (Lin et al., 2024) | 55.08 | 49.08 | 51.85 | 46.59 | 50.65 | 48.16 | 45.52 | 46.73 | 44.19 | 46.15 |
| CAPIT-base (Zhao et al., 2025) | 50.51 | 46.51 | 43.75 | 40.35 | 45.28 | 45.61 | 40.94 | 43.09 | 41.37 | 42.75 |
| Few-Shot (Šmíd et al., 2025b) | 54.82 | 48.78 | 49.17 | 45.61 | 49.60 | 50.13 | 45.44 | 44.28 | 43.99 | 45.96 |
| $\text{LACA}_{\text{LLAMA}_8}$ (Šmíd et al., 2025c) | 56.90 | 50.38 | 51.10 | 47.32 | 51.43 | 49.60 | 46.82 | 47.67 | 44.91 | 47.25 |
| TT-CSW mt5-based (Sheng et al., 2025) | 58.13 | 51.74 | 53.56 | 49.33 | 53.19 | 51.13 | 48.61 | 50.33 | 47.28 | 49.34 |
| **Ours (CLAOCS-TX)** | **63.91** | **57.48** | **58.12** | **54.64** | **58.54** | **55.56** | **52.17** | **53.93** | **50.86** | **53.13** |
| Ours w/o Code-Switching (XT only) | 60.25 | 54.62 | 56.24 | 50.67 | 55.45 | 53.63 | 50.71 | 51.08 | 47.22 | 50.66 |
| Ours w/o Aspect-CS | 62.04 | 57.42 | 57.01 | 53.93 | 57.60 | 54.22 | 51.20 | 52.52 | 49.11 | 51.76 |
| Ours w/o Opinion-CS | 61.90 | 56.75 | 56.51 | 53.48 | 57.16 | 54.01 | 51.33 | 52.02 | 48.75 | 51.53 |
| Ours w/o Triplet-Aware Contrastive Distillation | 60.22 | 55.39 | 56.67 | 52.71 | 56.25 | 52.47 | 49.90 | 50.84 | 47.36 | 50.14 |
| Ours w/o Consistency Filter | 59.26 | 54.80 | 55.42 | 50.49 | 54.99 | 51.79 | 47.08 | 48.33 | 46.58 | 48.45 |

Table 2: Average micro-F1 scores over five runs with different random seeds for zero-shot cross-lingual ASTE on four non-Indic and four low-resource Indic target languages, using English as the source language. Some baselines were originally proposed for related cross-lingual ABSA tasks and are reported here under the triplet-form evaluation setting for comparison with prior cross-lingual ASTE work.

et al., 2023) and EQUI-XABSA (Lin et al., 2024) introduce contrastive objectives and dynamically weighted losses to improve cross-lingual semantic alignment. CAPIT-BASE (Zhao et al., 2025) is an instruction-tuned cross-lingual ABSA baseline that combines contrastive pre-training with instruction tuning to improve transfer across target languages. Among recent LLM-based and code-switching approaches, $\text{LACA}_{\text{LLAMA}_8}$ (Šmíd et al., 2025c) generates high-quality pseudo-labelled target-language data without translation, while TT-CSW (Sheng et al., 2025) applies code-switching for effective test-time augmentation. We further report the FEW-SHOT baseline of Šmíd et al. (2025b) as a reference point for performance under limited target-language supervision, although our primary setting remains strictly zero-shot cross-lingual ASTE.

## 5 Results

Table 2 presents the zero-shot cross-lingual ASTE results on four non-Indic and four low-resource Indic target languages, using English as the source language. Our key observations include:

1) Among baseline methods, translation- and alignment-based approaches such as ACS-Distill, Equi-XABSA, and TT-CSW outperform simple zero-shot transfer, confirming the effectiveness of incorporating target-language signals even without gold annotations (RQ1).

2) Recent LLM-assisted methods (e.g., LACA and TT-CSW) further improve performance over earlier baselines, particularly on non-Indic languages, indicating the benefit of LLM-generated supervision for cross-lingual transfer (RQ1).

3) CLAOCS-TX achieves the best average performance across both language groups, with 58.54 micro-F1 on non-Indic languages and 53.13 on Indic languages. This corresponds to gains of approximately 2.5–3.0 absolute F1 points over TT-CSW and around 2.0 points over LACA on Indic languages, demonstrating improved zero-shot transfer capability (RQ1, RQ4).

4) The improvements of our model are more pronounced on Indic languages, where it consistently outperforms the strongest baselines by up to 3–4 absolute F1 points on average, suggesting better robustness under morphological complexity and limited target-language coverage (RQ4).

5) Ablation results indicate that the multi-variant consistency filter is the most critical component, as its removal causes the largest performance drop (about 2.5–3.0 F1 points on Indic and 1.5–2.0 on non-Indic languages), while excluding triplet-aware contrastive distillation leads to a moderate but consistent degradation (around 1.5–2.0 points). In contrast, removing individual code-switched prompting variants results in smaller declines (approximately 0.8–1.2 points), suggesting that structured distillation plays a more central role, with prompting variants providing complementary gains. These trends are further illustrated in Figure 3 (RQ2, RQ3).

### 5.1 Error Analysis

To analyze the remaining challenges of zero-shot cross-lingual ASTE, we conduct a qualitative error analysis across both non-Indic and low-resource Indic languages (Figure 4). We identify four dominant error types. The most frequent errors involve *boundary detection*, where aspect or opinion spans are partially correct but fail to capture exact token boundaries, a problem that is more pronounced in morphologically rich Indic languages such as Marathi and Bengali due to inflection and compounding. We also observe *aspect–opinion pairing errors*, particularly in sentences with multiple aspects or opinions, where correctly identified spans are incorrectly associated, leading to invalid triplets. Another common error type is *sentiment polarity confusion*, especially between neutral and positive classes, which arises when sentiment is expressed implicitly or through subtle lexical cues that are inconsistently captured in LLM-generated pseudo triplets. Finally, *pseudo-label noise* remains a source of error: although the multi-variant consistency filter mitigates hallucinated triplets, occasional spurious or incomplete pseudo labels still propagate to the student model, with higher frequency observed in Marathi, Bengali, and Odia. Representative failure cases are reported in Table 4 (Appendix), and many of these errors correlate with language-specific linguistic properties summarized in Appendix D. Overall, these findings highlight limitations of span-based modeling under weak supervision in morphologically rich languages, motivating future work on morphology-aware span representations or linguistically informed constraints to improve boundary detection and aspect–opinion association in low-resource settings.

**Effect of Pseudo-Labelled Triplets.** Pseudo-labelled triplets enable zero-shot cross-lingual ASTE by providing structured supervision in the absence of target-language annotations. As shown in Table 7 (Appendix), LLM-generated triplets largely preserve the original aspect–opinion–sentiment structure while introducing semantically aligned refinements. When combined with consistency filtering, these high-agreement pseudo triplets offer richer supervision than sentence-level augmentation alone, improving span boundary precision, aspect–opinion association, and sentiment alignment, particularly in languages where direct transfer is challenging.

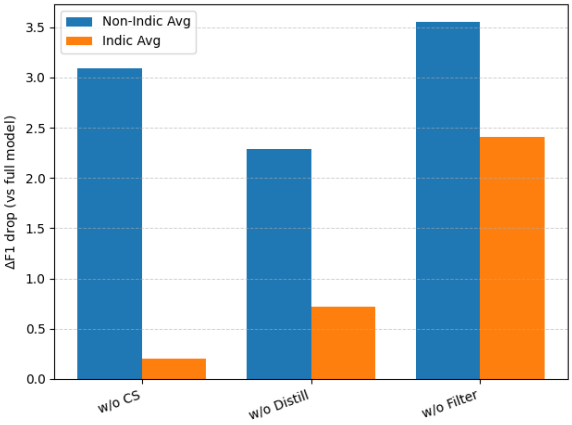


Figure 3: Average micro-F1 drop (ΔF1) from component removal on non-Indic and low-resource Indic languages.

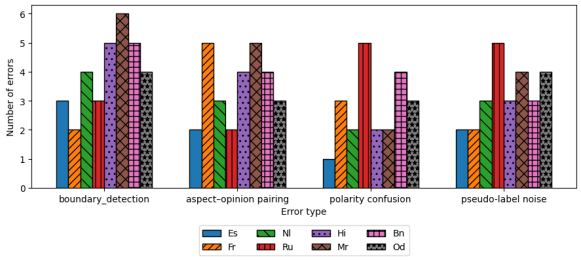


Figure 4: Error distribution across target languages based on 50 sampled predictions per language.

## 6 Conclusion

We proposed CLAOCS-TX, a unified framework for zero-shot cross-lingual aspect–opinion–sentiment triplet extraction that leverages large language models as both structured pseudo-label generators and semantic teachers. Through aspect- and opinion-aware code-switched prompting and multi-variant consistency filtering, the framework generates more reliable pseudo triplets for unlabeled target languages, while triplet-aware contrastive distillation enables structured cross-lingual semantic alignment at the triplet level. Experiments on four non-Indic and four low-resource Indic languages demonstrate consistent improvements over strong cross-lingual and LLM-based baselines. Ablation studies further confirm the complementary roles of code-switched prompting, consistency-based filtering, and triplet-aware contrastive distillation. Future work will explore additional source languages and cross-domain adaptation settings.

## Limitations

Despite the promising results, our work has several limitations. First, the framework relies on large language models to generate pseudo triplets in target languages, which introduces additional computational cost and may limit scalability in resource-constrained settings. While we employ multi-variant consistency filtering to mitigate noise, the quality of pseudo labels remains dependent on the underlying LLM and prompt design. Second, our experiments focus on English as the sole source language and evaluate transfer to a fixed set of non-Indic and Indic target languages. The effectiveness of the proposed approach under different source languages or for more typologically distant language pairs remains an open question. Third, although the proposed triplet-aware contrastive distillation improves cross-lingual alignment, it requires careful hyperparameter tuning and assumes reasonably aligned pseudo triplets. In scenarios where pseudo-label quality is severely degraded, the benefits of structured distillation may be reduced. Furthermore, we evaluate the framework under a zero-shot setting without target-language annotations. While this setting is realistic for low-resource languages, incorporating limited supervised or human-validated data could further improve performance, which we leave for future work. Our experiments use a single teacher LLM, and the quality of pseudo triplets may vary with the underlying model. Although the framework itself is teacher-agnostic, evaluating multiple teacher LLMs would provide a stronger assessment of robustness and is left for future work. Finally, due to the limited availability of annotated datasets across languages, our evaluation is restricted to the restaurant domain. As a result, the generalization of the proposed framework to other domains or cross-domain settings has not been explored. Investigating domain adaptation and cross-domain transfer for cross-lingual triplet extraction remains an important direction for future work.

## Ethics Statement

All experiments are conducted using publicly available benchmark datasets from prior scientific studies, ensuring transparency, reproducibility, and fair evaluation. This research does not involve human participants and poses no risk or harm to individuals. Nevertheless, the pre-trained models employed in this work are trained on large-scale Internet data and may inadvertently reflect societal biases, including those related to gender or race. We acknowledge this limitation and encourage cautious interpretation of the results.

## Acknowledgments

## References


Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2016. Aspect based sentiment analysis in hindi: resource creation and evaluation. In *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)*, pages 2703–2709.

Md Shad Akhtar, Palaash Sawant, Sukanta Sen, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Solving data sparsity for aspect based sentiment analysis using cross-linguality and multi-linguality. In *Proceedings of NAACL-HLT*, pages 572–582.

AR Balamurali, Aditya Joshi, and Pushpak Bhattacharyya. 2012. Cross-lingual sentiment analysis for indian languages using linked wordnets. In *Proceedings of COLING 2012: Posters*, pages 73–82.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, and 1 others. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom B. Brown, Benjamin Mann, Nick Ryder, and 1 others. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Lipika Dewangan and Chandresh Kumar Maurya. 2026. Cmf_hit: Enhancing code-mixed aspect-based sentiment analysis via language-aware gradient-based tokenization and feature fusion. *Expert Systems with Applications*, page 131639.

Lipika Dewangan, Zoyah Afsheen Sayeed, and Chandresh Maurya. 2025. Benchmark creation for aspect-based sentiment analysis in low-resource odia language and evaluation through fine-tuning of multilingual models. In *Proceedings of the 31st Inter-*

*national Conference on Computational Linguistics*, pages 5863–5869.

Hongliang Fei and Ping Li. 2020. Cross-lingual unsupervised sentiment classification with multi-view transfer learning. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5759–5771.

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.

Cho-Jui Hsieh and 1 others. 2023. Distilling step-by-step: Outperforming larger language models with less training data. *arXiv preprint arXiv:2305.02301*.

Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2486–2495.

Raviraj Joshi. 2022. L3cube-hindbert and devbert: Pre-trained bert transformer models for devanagari based hindi and marathi languages. *arXiv preprint arXiv:2211.11418*.

Julian Katz-Samuels, Zheng Li, Hyokun Yun, Priyanka Nigam, Yi Xu, Vaclav Petricek, Bing Yin, and Trishul Chilimbi. 2024. Evolutionary contrastive distillation for language model alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5328–5345.

Patrik Lambert. 2015. Aspect-level cross-lingual sentiment classification with constrained smt. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 781–787.

Xiaoya Li, Yue Zhang, and Jie Zhou. 2023. Generative data augmentation for aspect-based sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Xin Li, Lidong Bing, Wenxuan Zhang, Zheng Li, and Wai Lam. 2020. Unsupervised cross-lingual adaptation for sequence tagging and beyond. *arXiv preprint arXiv:2010.12405*.

Shuo Liang, Wei Wei, Xian-Ling Mao, Yuanyuan Fu, Rui Fang, and Dangyang Chen. 2023. Stage: span tagging and greedy inference scheme for aspect sentiment triplet extraction. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13174–13182.

Nankai Lin, Yingwen Fu, Xiaotian Lin, Dong Zhou, Aimin Yang, and Shengyi Jiang. 2023. Cl-xabsa: Contrastive learning for cross-lingual aspect-based sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2935–2946.

Nankai Lin, Meiyu Zeng, Xingming Liao, Weizhong Liu, Aimin Yang, and Dong Zhou. 2024. Addressing class-imbalance challenges in cross-lingual aspect-based sentiment analysis: Dynamic weighted loss and anti-decoupling. *Expert Systems with Applications*, 257:125059.

Yijiang Liu, Fei Li, and Donghong Ji. 2024. Improving cross-lingual aspect-based sentiment analysis with sememe bridge. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(12):1–22.

Anders Giovanni Møller, Arianna Pera, Jacob Dalsgaard, and Luca Aiello. 2024. The parrot dilemma: Human-labeled vs. llm-augmented data in classification tasks. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 179–192.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2020. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 8600–8607.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphee De Clercq, and 1 others. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 19–30.

Dongming Sheng, Kexin Han, Hao Li, Yan Zhang, Yucheng Huang, Jun Lang, and Wenqiang Liu. 2025. Test-time code-switching for cross-lingual aspect sentiment triplet extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5041–5053.

Kazutaka Shimada. 2023. Dataset construction and evaluation for aspect-opinion extraction in bangla fine-grained sentiment analysis. In *International Conference on Data Science and Applications*, pages 437–449. Springer.

Jakub Šmíd and Pavel Kral. 2025. Cross-lingual aspect-based sentiment analysis: A survey on tasks, approaches, and challenges. *Information Fusion*, page 103073.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025a. Advancing cross-lingual aspect-based sentiment analysis with llms and constrained decoding for sequence-to-sequence models. *arXiv preprint arXiv:2508.10366*.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025b. Few-shot cross-lingual aspect-based sentiment analysis with sequence-to-sequence models. In *International Conference on Text, Speech, and Dialogue*, pages 27–38.

Jakub Šmíd, Pavel Přibáň, and Pavel Král. 2025c. Laca: Improving cross-lingual aspect-based sentiment analysis with llm data augmentation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–853.

Guixin Su, Mingmin Wu, Zhongqiang Huang, Yongcheng Zhang, Tongguan Wang, Yuxue Hu, and Ying Sha. 2024. Refine, align, and aggregate: multi-view linguistic features enhancement for aspect sentiment triplet extraction. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3212–3228.

Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 498–508.

Weiting Tan, Kevin Heffernan, Holger Schwenk, and Philipp Koehn. 2023. Multilingual representation distillation with contrastive learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1477–1490.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive representation distillation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9122–9129.

Jason Wei, Xuezhi Wang, Dale Schuurmans, and 1 others. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2339–2349.

Lei Zhang, Shuai Wang, and Bing Liu. 2018. Aspect-based sentiment analysis in nlp. *ACM Computing Surveys*, 51(2).

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906.

Wenxuan Zhang, Ruidan He, Haiyun Peng, Lidong Bing, and Wai Lam. 2021. Cross-lingual aspect-based sentiment analysis with aspect term code-switching. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 9220–9230.

Wenwen Zhao, Zhisheng Yang, Song Yu, Shiyu Zhu, and Li Li. 2025. Contrastive pre-training and instruction tuning for cross-lingual aspect-based sentiment analysis. *Applied Intelligence*, 55(5):1–22.

## A Additional Experimental Details

As shown in Table 3, performance remains stable across a range of loss-weight settings, with the best results obtained by balancing pseudo-supervised learning and contrastive alignment.

Table 3: Sensitivity analysis of loss-weight hyperparameters. $\lambda_1$ controls the pseudo-supervised target loss $\mathcal{L}_{\text{tgt}}$, and $\lambda_2$ controls the contrastive loss $\mathcal{L}_{\text{con}}$. Results are reported as average micro-F1 on non-Indic and low-resource Indic target languages.

| $\lambda_1$ | $\lambda_2$ | Non-Indic Avg. | Indic Avg. | Overall Avg. |
|---|---|---|---|---|
| 0.5 | 0.5 | 57.02 | 51.84 | 56.02 |
| 1.0 | 0.5 | **58.54** | **53.13** | **58.29** |
| 1.0 | 1.0 | 58.31 | 52.91 | 58.05 |
| 1.5 | 0.5 | 57.88 | 52.37 | 57.81 |
| 1.0 | 0.2 | 57.63 | 52.08 | 57.63 |

## B Clarification on Dataset Version and Triplet Annotations

The official SemEval-2016 Task 5 dataset (Pontiki et al., 2016) was originally released for aspect-based sentiment analysis (ABSA) and does not include explicit opinion term span annotations. Aspect–opinion–sentiment triplet extraction (ASTE) is formalized by Peng et al. (2020) as structured triplets consisting of an aspect term, an opinion term, and a sentiment label, and was made publicly available as ASTE-Data-V1. Later, this ASTE benchmark reformulation was refined by Wan et al. (2020) and released as ASTE-Data-V2 (Xu et al., 2020). In our study, we do not newly derive opinion term annotations from the raw SemEval-2016 release. Instead, we use the publicly available ASTE-Data-V1 triplet-form benchmark (Peng et al., 2020) and follow the benchmark construction and evaluation practice adopted in prior ASTE and cross-lingual ASTE literature (Wan et al., 2020; Xu et al., 2020; Zhang et al., 2021; Sheng et al., 2025). Accordingly, the gold evaluation labels used in our experiments come from existing benchmark resources

**Prompt: Pseudo Triplet Extraction (Target-Language Explicit)**

You will be given a sentence written in a **target language**. Your task is to extract all valid **aspect–opinion–sentiment triplets**. Each triplet must follow the format: (*aspect term*, *opinion term*, *sentiment polarity*).

**Definitions**

- **Aspect term**: a specific attribute, feature, or component of a restaurant explicitly mentioned in the sentence.
- **Opinion term**: a word or phrase expressing sentiment toward an aspect.
- **Sentiment polarity**: one of `POS`, `NEG`, or `NEU`.

**Instructions**

1. Aspect and opinion terms must **exactly match spans** appearing in the input sentence. Do **not** infer, paraphrase, or introduce new aspect or opinion terms.
2. Assign sentiment polarity strictly based on the expressed opinion toward the aspect.
3. The output must be written in the **same target language** as the input. If no valid aspect–opinion pairs are present, return an empty list.

Return the output **strictly** in the following JSON format:

```
{
  "triplets": [
    {
      "aspect": "...",
      "opinion": "...",
      "sentiment": "POS | NEG | NEU"
    }
  ]
}
```

---

**Examples (French)**

**Example 1**

**Input:** *Le service a été rapide, mais le personnel était parfois inattentif pendant la soirée.*

**Output:**

```
{ "triplets": [ {"aspect": "service", "opinion":
"rapide", "sentiment": "POS"},
{"aspect": "personnel", "opinion":
"inattentif", "sentiment": "NEG"} ] }
```

**Example 2**

**Input:** *L'emplacement est très pratique, même si le temps d'attente était normal pour un week-end.*

**Output:**

```
{ "triplets": [ {"aspect": "emplacement",
"opinion": "très pratique", "sentiment":
"POS"},
{"aspect": "temps d'attente", "opinion":
"normal", "sentiment": "NEU"} ] }
```

---

**Now process the following input sentence:**

Input:

*Bien que l'emplacement soit pratique, le temps d'attente a été long et le personnel n'était pas toujours attentif.*

Figure 5: LLM prompt used for pseudo aspect–opinion–sentiment triplet generation.

**Algorithm 1:** Training procedure of the proposed framework

**Input:** Source data $\mathcal{D}_S$, target data $\mathcal{D}_T$, teacher LLM $G(\cdot)$, rounds $R$, samples $K$, agreement threshold $\tau_{\text{agree}}$, weights $\lambda_1, \lambda_2$

**Output:** Student parameters $\theta$

Train student $M_\theta$ on $\mathcal{D}_S$ using $\mathcal{L}_{\text{src}}$;

**for** $r \leftarrow 1$ **to** $R$ **do**

  $\hat{\mathcal{D}}_T \leftarrow \emptyset$;

  **foreach** $x^T \in \mathcal{D}_T$ **do**

    $(\mathcal{C}_a, \mathcal{C}_o) \leftarrow \textsc{Propose}(x^T, M_\theta)$;

    $\mathcal{V}(x^T) \leftarrow$ $\textsc{CodeSwitch}(x^T, \mathcal{C}_a, \mathcal{C}_o)$;

    $\mathcal{P} \leftarrow$ $\textsc{CollectTriplets}(G, \mathcal{V}(x^T), K)$;

    $\mathcal{P} \leftarrow \textsc{MapBack}(\mathcal{P}, x^T)$;

    $(\hat{\mathcal{T}}, \mathbf{c}) \leftarrow \textsc{Filter}(\mathcal{P}, \tau_{\text{agree}})$;

    **if** $\hat{\mathcal{T}} \neq \emptyset$ **then**

      $\hat{\mathcal{D}}_T \leftarrow \hat{\mathcal{D}}_T \cup \{(x^T, \hat{\mathcal{T}}, \mathbf{c})\}$;

  **foreach** $\mathcal{B} \subseteq \mathcal{D}_S \cup \hat{\mathcal{D}}_T$ **do**

    $\mathcal{L} \leftarrow \mathcal{L}_{\text{src}} + \lambda_1 \mathcal{L}_{\text{tgt}} + \lambda_2 \mathcal{L}_{\text{con}}$;

    Update $\theta$ by minimizing $\mathcal{L}$;

**return** $\theta$

and evaluation protocols adopted by prior work, rather than from new annotation or reconstruction in this paper. For the non-English target-language setting, we use the benchmark labels and evaluation protocol provided by the adopted cross-lingual ASTE setup, rather than projecting or re-annotating opinion terms within this work.

For the Indic languages, no pre-existing ASTE triplet benchmark was available to our knowledge. We therefore derive triplet-form gold labels from the existing annotations in the corresponding publicly available datasets using a fixed and deterministic conversion procedure defined prior to experimentation (Akhtar et al., 2016; Joshi, 2022; Shimada, 2023; Dewangan et al., 2025; Dewangan and Maurya, 2026). Aspect terms, opinion terms, and sentiment labels are obtained from the corresponding annotation fields in the original resources, and each valid aspect–opinion pair is represented as a triplet. This process does not involve additional manual labeling in the present paper and is entirely independent of the LLM-based pseudo-label generation used during training. The resulting triplets

| Example ($X_{AO}$ Text + Translation) | Error Type | Observed Issue |
|---|---|---|
| **Es:** El [precio] fue [demasiado alto] para una porción pequeña.<br>*(The price was too high for a small portion.)* | Boundary | Opinion span truncated, predicting (precio, alto, NEG) instead of (precio, demasiado alto, NEG). |
| **Fr:** La [nourriture] était [bonne], mais le [service] était [lent].<br>*(The food was good, but the service was slow.)* | Pairing | Incorrect cross-aspect pairing links (service, bonne, POS) and (nourriture, lent, NEG). |
| **Nl:** De [sfeer] was [oké], niets bijzonders.<br>*(The atmosphere was okay, nothing special.)* | Polarity | Neutral opinion misclassified as positive due to weak evaluative cues. |
| **Ru:** *Servis byl [plokhoy], khotya yeda byla normal'noy.*<br>*(The service was bad, although the food was normal.)* | Pairing | Negative opinion incorrectly propagated to an unrelated aspect. |
| **Hi:** *Sevā bahut [dhīmī] thī, lekin khānā ṭhīk thā.*<br>*(The service was very slow, but the food was fine.)* | Boundary | Opinion boundary partially detected, missing intensifier in the extracted triplet. |
| **Mr:** *Sevā khūp [vāiṭ] hotī, tarī vātāvaraṇ śānt hotē.*<br>*(The service was very bad, but the ambience was calm.)* | Pseudo-label noise | Pseudo-label introduces an additional ambience-related triplet not supported by the text. |
| **Bn:** *Khābār chhilo [bhālo], kintu apekkhā dīrgha chhilo.*<br>*(The food was good, but the waiting time was long.)* | Pairing | Aspect–opinion mismatch assigns positive sentiment to the waiting-time aspect. |
| **Od:** *Sebā bahuta [dhīra] thilā, dāma ṭhik thilā.*<br>*(The service was very slow, but the price was reasonable.)* | Polarity | Confusion between neutral and negative sentiment for mildly evaluative opinions. |

Table 4: Failure-case pseudo-labelling examples using joint aspect–opinion code-switched text ($X_{AO}$) across eight target languages. For Spanish, French, and Dutch, original target text is shown. For other languages, transliterated text and English translations are provided. Each example illustrates a dominant error type observed in pseudo-labelled triplets.

are used only as gold evaluation labels, while large language models are used only to generate pseudo-labeled triplets for unlabeled target-language sentences during training.

**Rationale for consistency-based filtering.** Self-consistency across multiple LLM samples reduces stochastic generation noise but is insufficient for structured extraction, as identical errors may persist across samples from the same input. Enforcing agreement across code-switched variants provides a stronger signal by requiring triplets to remain stable under different lexical realizations of the same sentence, thereby filtering hallucinated or context-dependent predictions. In addition, explicit span validity constraints are crucial for ASTE, as incorrect or non-contiguous spans can lead to invalid aspect–opinion pairings and error propagation during student training. Together, cross-variant agreement and span validity ensure that retained pseudo triplets are both semantically consistent and structurally well-formed.

## C Prompt Variants and Failure Modes

This appendix summarizes the prompt variants used for training-time pseudo triplet generation and illustrates representative failure modes. All prompts are applied only during training to support stepwise prompting and consistency-based filtering.

**Aspect-only prompt.** This prompt extracts candidate aspect terms without considering opinions or sentiment.

> **Instruction:** Identify all aspect terms mentioned in the following sentence. Return only the aspect terms as a list. Do not include opinions or sentiment labels.

**Failure case.** Spanish: *El precio fue demasiado alto para la calidad.* Predicted aspects: {precio, calidad}. *Issue: calidad* is spuriously extracted due to implicit comparison.

**Opinion-only prompt.** This prompt focuses on extracting sentiment-bearing expressions independently of aspects.

> **Instruction:** Identify all opinion expressions conveying sentiment in the following sentence. Return only the opinion terms.

**Failure case.** French: *Le service était rapide, mais la salle était bruyante.* Predicted opinions: {rapide, bruyante}. *Issue:* Polarity interpretation remains ambiguous without aspect context.

**Joint aspect–opinion prompt.** This prompt extracts complete aspect–opinion–sentiment triplets in a single step.

> **Instruction:** Extract all valid (aspect, opinion, sentiment) triplets from the sentence. Ensure that extracted spans exactly match the input text.

**Failure case.** Dutch: *Het eten was goed, maar de service was traag.* Predicted triplets: (service, goed, POS), (eten, traag, NEG). *Issue:* Cross-aspect opinion swapping in coordinated structures.

**Discussion.** Different prompt variants exhibit complementary strengths and failure modes. Aspect-only and opinion-only prompts provide high-recall span anchors but lack structural constraints, while joint prompting is more susceptible to pairing errors. Enforcing agreement across these variants therefore filters unstable or structurally inconsistent pseudo triplets, which empirically improves pseudo-label quality and supports the effectiveness of code-switched prompting observed in the ablation results (RQ2).

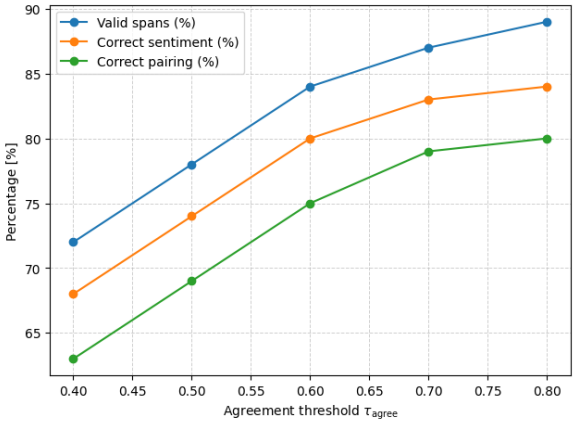


Figure 6: Effect of the agreement threshold $\tau_{\text{agree}}$ on pseudo-label quality metrics, including span validity, sentiment correctness, and aspect–opinion pairing accuracy.

## D Cross-Lingual Linguistic Challenges

This appendix summarizes dominant linguistic characteristics of selected target languages that pose challenges for zero-shot cross-lingual ASTE. These challenges help contextualize the error patterns discussed in Section 5.1 and the limitations of the proposed framework.

These linguistic properties exacerbate span boundary ambiguity and aspect–opinion association errors under weak supervision, particularly when pseudo labels are noisy or incomplete. The larger performance gains of CLAOCS-TX on Indic languages suggest that triplet-aware distillation and consistency-based filtering improve robustness to such language-specific variations, though challenges related to morphology and word order remain.

## E Pseudo-Label Quality Analysis

To better understand the effect of consistency-based filtering, we analyze the quality of LLM-generated pseudo triplets before and after filtering on a random sample of target-language sentences. We evaluate pseudo labels along three dimensions: span validity, sentiment correctness, and aspect–opinion pairing accuracy as shown in Table 6 and Figure 6.

All improvements of the proposed method over the baseline are statistically significant at the 95% confidence level, as shown in Table 10.

| Language | Linguistic Challenges for ASTE |
|---|---|
| French (FR) | Multi-word opinion expressions and frequent use of modifiers, which can lead to partial span extraction and sentiment dilution |
| Spanish (ES) | Rich adjectival modification and degree markers, causing boundary ambiguity for opinion spans and intensity mismatches |
| Dutch (NL) | Separable verb constructions and flexible adjective placement, leading to incorrect aspect–opinion pairing in coordinated clauses |
| Russian (RU) | Case-based inflection and agreement-driven word order variation, often resulting in span boundary shifts and incorrect aspect–opinion association |
| Hindi (HI) | Postpositions and long-distance dependencies, which complicate aspect–opinion alignment under weak supervision |
| Marathi (MR) | Rich morphology with extensive inflection, leading to boundary ambiguity for opinion spans and merged sentiment markers |
| Bengali (BN) | High prevalence of compounding and multi-word expressions, causing fragmented or partially detected aspect–opinion spans |
| Odia (OD) | Relatively free word order and flexible modifier placement, which complicates stable aspect–opinion pairing |

Table 5: Language-specific linguistic challenges affecting zero-shot cross-lingual ASTE performance.

| Pseudo Labels | Valid Spans (%) | Sentiment Correct (%) | Pairing Correct (%) |
|---|---|---|---|
| Before filtering | 71.2 | 68.5 | 64.1 |
| After filtering | 86.7 | 82.3 | 78.9 |

Table 6: Sample-based quality analysis (%) of pseudo triplets before and after consistency-based filtering.

| Lang | Target Snippet | Predicted Triplets | LLM Pseudo Triplets |
|---|---|---|---|
| Es | [servicio] *excelente* | (servicio, excelente, P) | (servicio, excelente, P) |
| Es | [precio] *demasiado alto* | (precio, alto, N) | (precio, demasiado alto, N) |
| Fr | [service] *impeccable* | (service, impeccable, P) | (service, excellent, P) |
| Fr | [nourriture] *pas terrible* | (nourriture, terrible, N) | (nourriture, pas terrible, N) |
| Nl | [service] *erg goed* | (service, goed, POS) | (service, erg goed, P) |
| Nl | [eten] *niet bijzonder* | (eten, bijzonder, NEG) | (eten, niet bijzonder, N) |

Table 7: Pseudo-labelling examples on non-Indic target languages (P-POS, N-NEG).

## F Runtime and Computational Cost

Large language models are used only during training for pseudo-label generation and contrastive supervision. For each target sentence, up to four code-switched variants are constructed, with $K = 5$ LLM samples per variant, resulting in at most 20 LLM calls per sentence. Student training is performed on a single NVIDIA A100 GPU, requiring approximately 10 epochs for English supervision followed by two pseudo-labeling rounds with 5 epochs each. LLM embeddings for retained pseudo triplets are cached and reused across epochs. At inference time, only the student ASTE model is used, with no LLM calls or code-switching, yielding inference efficiency comparable to standard XLM-R-based ASTE models (Table 8).

| Method | GPU Memory | Latency |
|---|---|---|
| XLM-R-based ASTE baselines | 1.8 | 24 |
| Proposed (CLAOCS-TX) | 1.9 | 26 |

Table 8: Inference-time memory usage (GB) and latency (ms) per sentence.

**Choice of LLM Teacher.** We conducted preliminary experiments with multiple instruction-tuned large language models, including LLaMA-3-8B-Instruct[3], Mistral-7B-Instruct[4], Qwen2.5-7B-Instruct[5], and Gemma-2-9B-It[6]. Although downstream performance differences were not large, LLaMA-3-8B-Instruct produced more consistently formatted and structurally coherent aspect–opinion–sentiment triplets, with fewer invalid outputs across code-switched variants. We therefore use LLaMA-3-8B-Instruct as the default teacher model in all experiments. This choice is made for implementation consistency rather than methodological dependence. Importantly, the proposed framework does not rely on a specific LLM, and inference is performed solely using the student ASTE model.

[3] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[4] https://mistral.ai/news/announcing-mistral-7b

[5] https://qwenlm.github.io/blog/qwen2.5-llm/

[6] https://ai.google.dev/gemma/docs/core/model_card_2

| Language | Baseline | $p$-value |
|---|---|---|
| Es | TT-CSW | 0.008 |
| Fr | TT-CSW | 0.012 |
| Nl | TT-CSW | 0.006 |
| Ru | TT-CSW | 0.010 |
| Hi | TT-CSW | 0.009 |
| Mr | TT-CSW | 0.014 |
| Bn | TT-CSW | 0.011 |
| Od | TT-CSW | 0.013 |

Table 9: $p$-values comparing with TT-CSW

| Language Group | Baseline | $p$-value |
|---|---|---|
| Non-Indic (avg) | TT-CSW mT5 | $< 0.01$ |
| Low-resource Indic (avg) | TT-CSW mT5 | $< 0.01$ |

Table 10: Statistical significance analysis (paired two-tailed t-test) comparing the proposed method with the strongest baseline over five random seeds.