informatics

MDPI

Article

A Generative Aspect Sentiment Triplet Extraction Approach Using Fine-Tuned LLMs

Arseniy Bapanov [1,2], Aslanbek Murzakhmetov [1,*], Nurbol Beisov [3], Maxatbek Satymbekov [3,4,*] and Mukhan Murotboyev [1]

[1] Department of Information Systems, Faculty of Technology, M.Kh. Dulaty Taraz University, Taraz 080001, Kazakhstan; arseniy.bapanov@gmail.com (A.B.)
[2] Zhambyl Branch of the Academy of Public Administration Under the President of Republic of Kazakhstan, Taraz 080001, Kazakhstan
[3] Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; beisov.nurbol@gmail.com
[4] School of STEM Education, Shakarim University, Semey 071410, Kazakhstan
* Correspondence: aslanmurzakhmet@gmail.com (A.M.); m.n.satymbekov@gmail.com (M.S.)

Abstract

Fine-grained analysis of guest satisfaction in the hospitality sector requires moving beyond document-level sentiment classification. It calls for structured extraction of attribute-specific evaluations from multilingual user-generated content. However, existing approaches have yet to address this methodological gap for the Central Asian hotel market. This study presents KAZ-LLM-ABSA, a generative framework for Aspect Sentiment Triplet Extraction applied to trilingual hotel reviews (Kazakh, English, and Russian) in Kazakhstan. The framework fine-tunes Qwen2.5-7B-Instruct via QLoRA on a purpose-built gold standard of 1726 expert-annotated triplets, formulating ASTE as an end-to-end sequence-to-sequence task to eliminate error propagation inherent to pipeline architectures. Benchmarked against six baselines spanning discriminative encoder models (mBERT-base, XLM-RoBERTa-base, BERT-SPC, LCF-BERT, RoBERTa-large) and a generative Seq2Seq model (GAS-mT5-base). KAZ-LLM-ABSA outperforms existing models across the reported metrics, demonstrating superior performance on the trilingual Kazakhstan hotel review corpus. These findings confirm the transferability of fine-tuned LLMs to low-resource multilingual hospitality data and offer practitioners a structured tool for aspect-level service quality diagnostics.

Keywords: aspect-oriented sentiment analysis; triplets; large language models; QLoRA; hotel reviews; multilingual text processing

Check for updates

Academic Editor: Patricia Anthony

Received: 9 August 2026
Revised: 28 September 2026
Accepted: 29 September 2026
Published: 3 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The tourism industry is one of the priority areas of the Republic of Kazakhstan's socioeconomic development. The Concept for the Development of the Tourism Industry of the Republic of Kazakhstan for 2023–2029, approved by Resolution No. 262 of the Government of the Republic of Kazakhstan dated 28 March 2023, defines the strategic guidelines for its long-term growth. State support for the industry stems from its significant contribution to the national economy. According to data from the National Statistics Bureau of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan, the gross value added generated by the tourism sector in 2023 amounted to 3283.2 billion tenge [1,2].

The hotel industry plays a central role in the structure of the country's tourism sector. According to official data from the National Statistics Bureau, the total number of foreign

visitors registered at lodging facilities increased nearly fourfold from 329,700 people in 2021 to 1.2 million in 2024, and by the end of 2025 this figure reached 1.4 million people. The total value of services provided by domestic lodging facilities by the end of 2024 amounted to 299.8 billion tenge [3]. In the World Economic Forum's Global Travel and Tourism Development Index for 2024, Kazakhstan ranked 52nd out of 119 countries, rising from 80th place in 2019 [4]. These trends indicate not only growth in the industry's absolute metrics but also an increase in the competitiveness of Kazakhstan's hotel sector on the international stage, which underscores the practical importance of systematically studying the factors that shape guest satisfaction. The digital transformation of the tourism industry has led to the formation of a complex ecosystem of online platforms that aggregate user-generated content about hotel services. Booking.com stands out among the key sources of reviews as of 2024, the platform offers reservations at approximately 3.4 million accommodations, including 475,000 hotels, motels, and resorts in more than 220 countries and territories, operating in more than 40 languages [5]. According to the company's official data, the platform has accumulated over 370 million verified reviews from real guests whose stays have been verified [6]. TripAdvisor plays an equally significant role in the online review ecosystem. According to the company's annual report, as of 31 December 2023, the platform had more than 1 billion user ratings and reviews for over 8 million properties such as accommodations, restaurants, airlines, and cruises, in more than 40 countries and 20 languages worldwide [7]. Alongside these platforms, Expedia, Hotels.com, Agoda, Google Hotels, and social media (Instagram, Facebook) actively serve as sources of user reviews on hotel services, collectively forming a multi-channel environment for electronic word-of-mouth (eWOM) [8].

Reviews published on these platforms have a significant and empirically proven influence on consumer behavior when choosing accommodations. According to studies of user behavior, 79% of travelers read between 6 and 12 reviews before booking a hotel, while 96% of consumers overall rely on online reviews when planning trips [9]. The authors of [10] note that Booking.com ranks as one of the most influential sources of eWOM in the hospitality industry, a fact attributable to its closed author verification system. The right to post a review is granted exclusively to guests who have confirmed their booking and stay through the platform. This feature fundamentally distinguishes Booking.com from open platforms and ensures greater reliability of the collected data. A systematic literature review [11] analyzing sentiment in the hospitality and tourism sectors emphasizes that reviews on closed and open platforms exhibit significant differences in key indicators such as ratings, topics, and sentiment polarity, which must be considered when compiling a research database. In the context of this study, the choice of Booking.com as a data source is based precisely on the combination of these characteristics. Author verification, a standardized content structure, and the widest coverage of Kazakhstan's hotel sector among international platforms. The rapid growth in tourist traffic, coupled with the digitalization of consumer behavior, has led to a significant increase in the volume of user reviews on online booking platforms. These reviews represent a valuable source of unstructured text data that reflects real consumer experiences and shapes the reputation of hotel businesses. Extracting useful managerial insights from such large volumes of text requires a shift from coarse, polarized sentiment classification (positive/negative/neutral) to methods of fine-grained analysis that allow specific evaluative judgments to be linked to specific attributes of hotel service.

Aspect-Based Sentiment Analysis (ABSA) is a subfield of natural language processing whose goal is to identify specific aspects of the evaluated object in a text and determine the sentiment polarity expressed with respect to each of them [12]. Within this paradigm, the task of Aspect Sentiment Triplet Extraction (ASTE), first formalized by Peng et al., [13]

involves the joint extraction of three interrelated elements from a sentence, an aspect term, an opinion term, and the corresponding sentiment polarity. For example, from a hotel review stating "the staff was very friendly, but the breakfast was sparse," this method allows for the extraction of two triplets: (staff, friendly, positive) and (breakfast, sparse, negative), which provides significantly more information than an overall assessment of the review as a whole. Early approaches to solving the ASTE problem relied on pipeline architectures and joint sequence tagging schemes (token-level and table-filling tagging), in which aspect and evaluative terms were extracted by separate modules and subsequently matched into triples, which led to the accumulation of errors at the matching stage [14]. An alternative to these discriminative methods was the generative paradigm, in which the ASTE task is formulated as a sequence-to-sequence task. The model is trained to directly generate a textual representation of triples, bypassing the intermediate stages of token tagging [15]. The application of large language models (LLMs) in this paradigm has opened up fundamentally new possibilities for solving compound ABSA tasks, since generative LLMs are capable of simultaneously considering the syntactic, semantic, and pragmatic context of the entire review when generating a structured response [16]. At the same time, the results of applying LLMs to ABSA tasks in zero-shot and few-shot settings remain mixed. Number of studies demonstrate that general-purpose LLMs underperform specialized fine-tuned models when solving compound tasks requiring structured output of several interrelated sentiment elements, whereas domain-specific fine-tuning of the model significantly improves the accuracy of triplet extraction. In particular, instruction tuning of generative models from the T5 and LLaMA families using specialized prompts for ABSA subtasks demonstrated a significant improvement in quality compared to baseline generative configurations, including under conditions of limited annotation [17,18]. With regard to the tourism and hospitality industries, some studies show that fine-tuned LLMs incorporating auxiliary knowledge and positional information can improve the completeness of implicit aspect extraction and reduce the propagation of errors characteristic of pipeline architectures [19,20].

Despite the intensive development of generative ASTE methods using English-language reference corpora, their application to Russian-language and domain-specific corpora of hotel reviews from Central Asia remains virtually unexplored. A systematic analysis of the sentiment in tourist reviews pertaining to the hotel sector in Kazakhstan using modern generative architectures based on fine-tuned LLMs is absent from the available scientific literature, which determines the scientific novelty and relevance of this study.

Accordingly, this study aims to assess guest satisfaction in Kazakhstan's hotel industry by applying a generative approach to aspect-sentiment triplet extraction (Generative Aspect Sentiment Triplet Extraction) based on fine-tuned large language models, implemented on a corpus of reviews from the Booking.com platform.

2. Literature Review

The theoretical and methodological foundations of ABSA emerged at the intersection of computational linguistics, machine learning, and applied research on consumer behavior. The ABSA task is decomposed into four interrelated subtasks like, aspect term extraction, aspect category determination, opinion term extraction, and sentiment polarity classification. The combination of all gives rise to a family of compound tasks, ranging from the paired "aspect—tonality" extraction to the complete "category—aspect—opinion—tonality" quaternions. A systematic review [21] notes a steady shift in interest from discriminative models to generative architectures, as well as pronounced domain-specific unevenness. The overwhelming majority of studies focus on restaurants and consumer goods, whereas the hospitality sector, despite its economic significance, is represented to a much lesser extent.

With regard to the hospitality industry, methodological traditions have developed along several lines. An approach combining LDA [22] for topic clustering, followed by semantic similarity assessment and BERT-based classification, demonstrates the ability to identify aspects without explicit tagging. However, the accuracy of pipeline solutions is determined by the quality of the topic clusters generated in the first stage. The approach that integrates lexical ambiguity resolution with BERT's contextual representations achieves improved accuracy at the aspect level [23], yet transferability to new domains and languages remains an open challenge. Both lines of research reveal a common feature of early ABSA in the hospitality domain. The predominance of pipeline architectures with separate modules, which predictably leads to the accumulation of errors at the junctions between stages. At the same time, the generative paradigm has become established in the literature, formulating the ASTE task as a sequence-to-sequence transformation. A systematic review [24] traces its evolution—from the first unified frameworks based on T5 (GAS-T5) to aspect-oriented graphs with external knowledge, and highlights the consistent advantage of generative architectures under conditions of limited annotation, which is directly relevant for Russian-language domain-specific corpora. Among architectural refinements that do not involve a shift to the generative paradigm, three approaches stand out. The LCF mechanism [25] eliminates cross-aspect attribution errors, whereby the model incorrectly associates a sentiment expression with a neighboring aspect rather than its true target, by restricting the syntactically meaningful context to a local window around the target term. A lightweight multi-layer attention network implements the interaction between aspect and context through recurrent layers with moderate computational cost [26]. Combining BERT with multi-layer GCNs models' syntactic dependencies between words [27]. A general pattern emerges, each enhancement yields a measurable improvement on benchmark datasets, but transfer to texts outside the benchmark domains is significantly weaker.

The issue of ABSA's language-specific nature constitutes a separate line of research. An analysis of Twitter reviews [28] shows that fragmented, informal text requires additional preprocessing and reduces the applicability of models tuned for coherent content. Paper [29] documents a decline in quality when transitioning from the dominant language of the sample to languages less represented in the pre-training corpus, while the authors of [30] propose aspect-focused learning as a way to bridge the cross-linguistic gap. Taken together, these studies point to a chronic gap between the linguistic diversity of user-generated content and the narrow language base of ABSA models—a gap that is particularly significant for the Russian-language segment of hotel reviews in Kazakhstan. The paper [31] is devoted to the specifics of processing negation, the presence of negative constructions systematically reduces the accuracy of polarity determination, since negation locally inverts the tone without affecting the overall emotional tone of the sentence. This observation is essential for interpreting results on natural hotel reviews, which are replete with such constructions.

Kazakhstani studies on the sentiment of tourist reviews hold a special place. In [32], a model of an information system for evaluating tourist attractions was developed. Subsequent publications by the same group cover the analysis of YouTube reviews using hybrid models [33] and multilingual semantic analysis of travel destinations [34]. The culmination of this series was the study [35] using a hybrid stacking ensemble (VADER, TextBlob, Stanza, LCF-BERT), which confirmed the consistent advantage of ensemble methods over any single method. At the same time, the entire body of Kazakhstani research focuses on integral tonality but does not involve the extraction of fine-grained "aspect—opinion—tonality" triplets, which constitutes the central research gap of this study.

Outside of Kazakhstan, the authors [36] found that GPT-4 omni outperformed BERT in terms of classification accuracy for hotel reviews (67% vs. 60.6%), while noting that BERT

models retained an advantage in terms of detail when extracting discrete attributes. The use of ChatGPT in zero-shot mode for ABSA demonstrates that quality depends on the wording of the instructions, with limited reproducibility of results. The authors of [37], who investigated the applicability of LLaMA to ABSA compound tasks, found that models fine-tuned using LoRA/QLoRA outperform specialized T5 architectures, whereas under-trained models are noticeably inferior in terms of the completeness of triplet extraction, especially on multi-aspect sentences. This establishes a methodological consensus, the advantage of the generative Seq2Seq [14–16] architecture is realized only through targeted fine-tuning on domain-specific data, rather than through prompt engineering. Among compound tasks, ASTE occupies an intermediate position between pairwise classification and quaternion extraction. Including the opinion term in the output structure significantly improves interpretability for practical applications. However, it simultaneously increases the requirements for annotated data and for the correct alignment of aspect and evaluation spans—the "pairing" problem, which remains a source of a significant proportion of errors.

The identified research gaps justify the relevance of this study. First, fine-grained triplet extraction in the hospitality domain has been methodologically explored to a much lesser extent than in the restaurant and grocery domains. Second, Kazakhstani studies do not go beyond document-level analysis, leaving open the question of the applicability of generative ASTE to Russian- and Kazakh-language content. Third, the consensus on the superiority of fine-tuned LLMs is based exclusively on English-language corpora, and its transferability to low-resource domain-specific data has not been empirically verified.

3. Materials and Methods

3.1. Dataset

For this study, we used a corpus of user reviews of hotels in Kazakhstan collected from the Booking.com platform. The original dataset includes 100,015 reviews covering 149 hotels in the country's three largest cities such as Almaty (45,736 reviews, 54 properties), Astana (32,640 reviews, 43 properties), and Shymkent (21,639 reviews, 52 properties). The reviews cover the period from November 2022 to February 2026, with the bulk of the data (96,850 records, about 97%) coming from 2023–2025. Each record in the dataset contains a numerical rating on a 10-point scale, the type of trip (solo, couple, family, group), the reviewer's country of residence (156 unique countries, with 58% of reviews posted by guests from Kazakhstan), as well as a title and a detailed text comment. The text field retains the two-part structure characteristic of Booking.com and, in the original export, is marked by a separator that distinguishes the positive part of the review from the negative part. Out of the total dataset, 41,586 entries (41.6%) contain a detailed text comment, while in the remaining 58,429 cases (58.4%), the guest provided only a numerical rating and a title without any detailed text. It was precisely this subset of reviews with a non-empty text field that we considered suitable for aspect-oriented analysis. The linguistic composition of the texts is multilingual, reviews in Cyrillic (65%) and Latin script (31%) predominate, with a smaller proportion of texts in Arabic script and Chinese. After excluding reviews in languages irrelevant to the task, as well as entries with excessively short and uninformative text, the final working corpus used for further annotation and model training consisted of 39,612 reviews. Data collection from the Booking.com platform was conducted exclusively for non-commercial research purposes in accordance with the principles of Fair Use and personal data protection policies.

Prior to annotation, the raw review texts were subjected to a standardised preprocessing pipeline. First, the Booking.com structural separator (Negative) was removed and positive and negative text segments were concatenated into a single unified review string. Second, carriage-return artefacts (_x000D_) were replaced with a single whitespace

character. Third, reviews containing bilingual duplicate segments—identified by character-level Jaccard similarity ≥ 0.90 between the first and second halves of the text—were deduplicated by retaining only the first language segment. Fourth, reviews with fewer than 20 characters after preprocessing were excluded as uninformative. Fifth, all retained texts were normalised to Unicode NFC form. These steps reduced the working corpus from 41,586 reviews with non-empty text to 39,612 reviews suitable for annotation candidate sampling.

The language of each review text was determined using a two-stage procedure. In Stage 1, sentence-level language classification was performed using the fastText language identification model with a confidence threshold of 0.85. Reviews classified as languages other than Kazakh, English or Russian were excluded from the annotation pool. Reviews falling below the 0.85 confidence threshold (3.7% of the corpus) were forwarded to Stage 2, in which a bilingual (Russian–Kazakh) annotator manually assigned a language label. Inter-rater agreement between the fastText classifier and the human annotator on a 200-review validation sample was 98.5% (Cohen's $\kappa = 0.97$), confirming the reliability of the automated classification.

Morphological Considerations. The three languages covered in this study differ substantially in their morphological typology, and these differences have direct implications for the ASTE task. Russian is a highly inflectional language with a six-case system and grammatical gender, producing multiple surface forms of the same aspect term depending on its syntactic role (e.g., завтрак [nominative], завтрака [genitive], завтраком [instrumental]). The model must generalise across these variants during both training and inference, since gold annotations record the verbatim surface span rather than a lemmatised form. Kazakh is agglutinative: grammatical and derivational information is encoded through productive suffixation chains, which can generate a large number of surface forms from a single base. This property substantially complicates verbatim span matching and contributes to the lower Kazakh-segment F_1 (0.840) observed in our experiments. English, being morphologically analytic, presents the fewest span-extraction challenges and correspondingly achieves the highest per-language performance ($F_1 = 0.877$). Beyond formal morphology, the Kazakh-language segment of the corpus contains informal evaluative vocabulary, including Russified Kazakh slang forms such as "суперлік" and "керемет", that is underrepresented in standard pre-training corpora and constituted 42% of the misclassified opinion terms in our code-switching error analysis.

3.2. Gold Standart Dataset Construction

To address the problem of ASTE using hotel reviews about Kazakhstan, we created a gold standard annotated corpus. The source material consisted of a stratified sample of multilingual reviews in three languages, Kazakh, English and Russian, reflecting the actual linguistic diversity of the Booking.com platform's audience in the region. The annotation process was carried out by subject-matter experts in accordance with an established annotation scheme. Each review was assigned aspect categories corresponding to hotel service attributes (staff, breakfast, location, cleanliness, etc.), as well as sentiment polarity on a three-point scale, positive, negative, neutral. In the final stage, the annotated data were converted into a structured triplet format (aspect term, opinion term, sentiment polarity), which corresponds to the standard formulation of the ASTE task. The final reference dataset includes 1726 annotated triplets covering all target aspect categories and the three dominant languages of the corpus.

Prior to splitting, all 1450 annotated sentences were subjected to near-duplicate detection using MinHash Locality-Sensitive Hashing (LSH; Jaccard threshold $\theta = 0.85$, five hash permutations per band, implemented via datasketch v1.6). Sentence pairs with estimated

Jaccard similarity ≥ 0.85 were collapsed to a single representative instance, retaining the annotation with the higher inter-annotator agreement score. This procedure identified 47 near-duplicate pairs, yielding a deduplicated corpus of 1403 unique sentences and 1672 triplets. All subsequent splitting and evaluation was performed on the deduplicated corpus.

To prevent hotel-level data leakage—whereby reviews of the same property, sharing systematic aspect patterns and vocabulary, could appear in both training and test partitions—the deduplicated sentences were partitioned using a stratified hotel-grouped split: all sentences originating from a given hotel were assigned exclusively to one partition. The 149 hotels were allocated approximately as 104 (train)/23 (dev)/22 (test), preserving the approximate 70/15/15 triplet ratio and the per-language distribution within each partition. No hotel appears in more than one partition. The resulting split statistics are presented in Table 1.

Table 1. Split statistics for the reference dataset. (**a**). Explicit and implicit triplet counts by split and language; (**b**) distribution of sentiment polarity labels and aspect categories in D_{gold} (post-deduplication, hotel-grouped split and 1672 triplets).

Sample	Hotels	Sentences	Triplets	Kazakh	English	Russian
Train	104	982	1170	235	234	701
Dev	23	211	251	51	50	150
Test	22	210	251	50	51	150
Total	149	1403	1672	336	335	1001

(**a**)

Sample	Explicit	Implicit	Kazakh (Expl/Impl)	English (Expl/Impl)	Russian (Expl/Impl)
Train	984	224	194/48	197/44	593/132
Dev	211	48	42/10	42/10	127/28
Test	211	48	42/10	42/10	127/28
Total	1406	320	278/68	281/64	847/188

(**b**)

Polarity	Train	Dev	Test	Total
Positive	621 (53.1%)	133 (53.0%)	133 (53.0%)	887 (53.1%)
Negative	432 (36.9%)	93 (37.1%)	92 (36.7%)	617 (36.9%)
Neutral	117 (10.0%)	25 (10.0%)	26 (10.4%)	168 (10.1%)
Top aspect categories				
Service/Staff	198 (16.9%)	43 (17.1%)	42 (16.7%)	283 (16.9%)
Room	187 (16.0%)	40 (15.9%)	41 (16.3%)	268 (16.0%)
Breakfast/Food	176 (15.0%)	37 (14.7%)	38 (15.1%)	251 (15.0%)
Location	152 (13.0%)	33 (13.1%)	32 (12.7%)	217 (13.0%)
Cleanliness	128 (10.9%)	27 (10.8%)	28 (11.2%)	183 (10.9%)
Price	94 (8.0%)	20 (8.0%)	20 (8.0%)	134 (8.0%)
Other (10 categories)	235 (20.1%)	51 (20.3%)	50 (19.9%)	336 (20.1%)

Of the 1726 annotated triplets, 1406 (81.5%) contain an explicit aspect span and 320 (18.5%) involve an implicit aspect expressed through ontological category labels. Table 1a presents the distribution of explicit and implicit triplets across dataset partitions and languages.

The gold standard dataset serves a dual purpose in this study. First, it serves as a training basis for fine-tuning the generative language model. Second, it provides a reference baseline for objectively evaluating the model's quality using standard metrics of accuracy and completeness in triplet extraction. The use of expert annotation, rather than relying

solely on automatically generated labels, stems from the need to ensure high reliability of the training signal, since the quality of the annotations directly determines the ability of the fine-tuned model to correctly map opinion terms to the corresponding aspect terms in the context of multilingual and domain-specific input text. The D_{gold} dataset was created by two independent expert linguists. To assess inter-annotator agreement (IAA) on a test set of 200 reviews, Cohen's Kappa coefficient (k) for categories and polarities, as well as the F_1-agreement metric for matching substrings of aspect and evaluative terms. The average Cohen's kappa value for determining tone polarity was $k = 0.84$ (high level of agreement), and for aspect categories, it was $k = 0.81$. The agreement on identifying the exact boundaries of triplets was F_1-agreement = 0.82. All disputed cases and discrepancies in span boundaries were reviewed during the consensus validation phase under the supervision of a senior NLP researcher. If two annotators could not reach a consensus regarding the boundaries of an opinion or aspect, that annotation was excluded from the reference set.

Key decision rules are summarised as follows. Aspect spans were defined as the minimal contiguous token sequence uniquely identifying the hotel service attribute being evaluated. Spans were not extended to include determiners or modifying adjectives unless they were lexically inseparable from the aspect noun. Opinion spans were defined as the minimal contiguous token sequence expressing an evaluative stance toward the identified aspect. Negation particles were included within the opinion span when they inverted the polarity of the head evaluative term. Polarity was assigned on a three-point scale (Positive/Negative/Neutral) relative to the expressed opinion, not relative to the overall review rating. Implicit aspects, those with no verbatim lexical realization, were assigned the [IMPL]-prefixed ontological label from the 16-category taxonomy most specifically applicable to the inferred aspect. Sentences containing more than five triplets were flagged for senior NLP researcher review prior to inclusion. Discrepant annotations were resolved through structured consensus discussion, sentences where consensus could not be reached were excluded from the gold standard.

3.3. Methods

3.3.1. Baseline Models: Pipeline ASTE Conversion Scheme

The encoder-based baseline models considered in this study are mBERT-base, XLM-RoBERTa-base, BERT-SPC, LCF-BERT, and RoBERTa-large, were originally developed for Aspect-Level Sentiment Classification (ABSC), a task that takes a pre-specified aspect term as input and predicts sentiment polarity. Applied naively, such models cannot participate in end-to-end ASTE evaluation, because they do not perform span extraction from raw text. To ensure a methodologically fair and directly comparable evaluation against the proposed end-to-end model KAZ-LLM-ABSA, all encoder baselines were integrated into a standardised two-stage Pipeline ASTE scheme [13–15].

Stage 1—Joint Span Extraction. A dedicated span-extraction model (mBERT-CRF), a multilingual BERT encoder with a Conditional Random Field (CRF) decoding head, was trained on the training partition of D_{gold} to jointly identify aspect term spans and opinion term spans via BIO sequence labelling. For each sentence, all valid candidate pairs (a_i, o_i) were formed by pairing each extracted aspect span with each co-occurring opinion span within the same sentence boundary. The mBERT-CRF model achieves span-level $F_1 = 0.91$ for aspect spans and $F_1 = 0.87$ for opinion spans on the development partition of D_{gold}. These figures establish a performance ceiling for the pipeline systems. Even a perfect Stage 2 classifier cannot recover triplets whose spans were missed or incorrectly bounded in Stage 1.

Stage 2—Pairing and Sentiment Classification. For each candidate pair (a_i, o_i), an input sequence was constructed in the format [CLS] sentence [SEP] a_i [SEP] o_i [SEP] and passed

to the respective ABSC classifier, which predicted a polarity label $s \in$ {Positive, Negative, Neutral, NoRelation }. The final triplet set for each sentence was formed from the retained predictions (a_i, o_i, s).

Evaluation. The end-to-end triplet sets produced by the Pipeline ASTE systems were evaluated against D_{gold} under the identical Strict Exact Match criterion and micro-averaged F_1 computation applied to KAZ-LLM-ABSA. All seven models are therefore evaluated on the same ASTE output format, same test partition, and under the same scoring protocol, making the comparison methodologically equivalent across architectural paradigms.

3.3.2. mBERT-Base

The mBERT-base model is a multilingual version of the BERT architecture, pre-trained on a Wikipedia corpus covering 104 languages. In this study, mBERT-base was used in a sentence-pair classification configuration. The input sequence is formed by concatenating the review text and the aspect term, separated by a special token [SEP], and the classification head above the [CLS] token predicts the polarity of the tone. A key advantage of mBERT in the context of this study is its native support for Kazakh, English and Russian within a single shared vocabulary, which eliminates the need to split the corpus by language during training. This model was evaluated within the two-stage Pipeline ASTE scheme. The metrics reported in Tables 2–7 reflect end-to-end triplet extraction performance, not isolated sentiment classification performance.

Table 2. mBERT results on the test set.

Metric	Value
Accuracy	0.712
Precision	0.720
Recall	0.705
F_1-score	0.712

Table 3. XLM-RoBERTa results on the test set.

Metric	Value
Accuracy	0.748
Precision	0.780
Recall	0.732
F_1-score	0.755

Table 4. BERT-SPC results on the test set.

Metric	Value
Accuracy	0.731
Precision	0.736
Recall	0.725
F_1-score	0.730

Table 5. LCF-BERT results on the test set.

Metric	Value
Accuracy	0.623
Precision	0.795
Recall	0.588
F_1-score	0.676

Table 6. RoBERTa results on the test set.

Metric	Value
Accuracy	0.756
Precision	0.798
Recall	0.743
F_1-score	0.770

Table 7. GAS-mT5 results on the test set.

Metric	Value
Accuracy	0.810
Precision	0.825
Recall	0.790
F_1-score	0.807

3.3.3. XLM-RoBERTa-Base

The XLM-RoBERTa-base model (Cross-lingual Language Model RoBERTa base) is a multilingual transformer encoder pre-trained on the Common Crawl corpus, a 2.5 TB dataset of text data in 100 languages, using only the MLM task, without NSP, which the authors deemed unsuitable for cross-lingual transfer. The XLM-RoBERTa-base architecture consists of 12 Transformer layers, 12 attention heads, and 768 hidden dimensions. Training on a significantly larger and more diverse corpus compared to mBERT provides the model with qualitatively more robust and transferable multilingual representations. In this experiment, XLM-RoBERTa-base was applied in the same sentence-pair classification configuration as mBERT-base, with the classification head fine-tuned on the gold standard of this study.

3.3.4. BERT-SPC

The BERT-SPC (BERT Sentence-Pair Classification) architecture was used as the third base model, in which the task of classifying tone at the aspect level is formulated as a sentence-pair classification task. The BERT-SPC architecture does not explicitly model syntactic dependencies between words and does not employ a mechanism for local contextual focusing.

3.3.5. LCF-BERT

The LCF-BERT (Local Context Focus BERT), is an extension of standard BERT, specifically designed for the task of aspect-level tone classification. The key innovation is the LCF mechanism, which limits the semantically relevant context around the target aspect term based on SRD. This mechanism allows the model to focus on tokens that are syntactically close to the target aspect, thereby reducing the influence of irrelevant context in long sentences with multiple aspects.

3.3.6. RoBERTa-Large

The RoBERTa (Robustly Optimized BERT Pretraining Approach), is an optimized version of BERT trained using an improved pretraining strategy. Dynamic token masking, an increased batch size (up to 8000 examples), training on a larger corpus (160 GB of text), and the exclusion of the next-sentence prediction (NSP) task, which the authors deemed redundant. Together, these changes provide RoBERTa with more robust and generalizable contextual representations compared to the original BERT, resulting in consistently higher performance across a wide range of NLP tasks. In this experiment, RoBERTa was applied with a classification head to predict aspect-level polarity, similar to BERT-SPC. The model

demonstrated the best results among all three compared architectures across a set of metrics, achieving a balanced Precision–Recall ratio and the highest F_1-score, which attests to the superiority of the optimized pre-training strategy when applied to domain-specific multilingual hotel review text.

3.3.7. GAS-mT5-Base

The GAS-mT5-base model (Generative Aspect-based Sentiment, built on the multilingual T5 framework) implements the generative paradigm for solving the ASTE task, the triplet extraction task is formulated as the transformation of an input sequence into a structured text string in sequence-to-sequence mode, which fundamentally eliminates the accumulation of errors characteristic of cascaded pipelines. The mT5-base model, a multilingual version of T5 (Text-to-Text Transfer Transformer), pre-trained on the mC4 corpus, which covers 101 languages and contains a significant amount of data in Russian and English, providing basic capabilities for understanding multilingual text without additional language-specific modifications.

A comparative analysis of the results from the baseline models (Table 8) reveals several consistent patterns that structure the overall picture of the performance of the architectures under consideration as applied to the ASTE task on a multilingual corpus of hotel reviews from Kazakhstan. The oracle condition (RoBERTa-large + Gold Spans, F_1 = 0.830) establishes the performance ceiling of the pipeline paradigm on this corpus under ideal span extraction. The gap between the standard pipeline (RoBERTa-large, F_1 = 0.770) and the oracle (+6.0) quantifies the performance loss attributable exclusively to Stage 1 span extraction errors, that are structurally absent in KAZ-LLM-ABSA, which recovers spans and sentiment in a single autoregressive pass. The residual gap between the oracle and KAZ-LLM-ABSA (ΔF_1 = +3.9) reflects the advantage of large-scale multilingual pre-training and instruction tuning over discriminative classification even under ideal span availability, and represents the irreducible benefit of the end-to-end generative formulation on this corpus.

Table 8. Comparison of model results on the ASTE task.

Model	Accuracy	Precision	Recall	F_1-Score
mBERT-base	0.712	0.720	0.705	0.712
XLM-RoBERTa-base	0.748	0.780	0.732	0.755
BERT-SPC	0.731	0.736	0.725	0.730
LCF-BERT	0.623	0.795	0.588	0.676
RoBERTa-large	0.756	0.798	0.743	0.770
GAS-mT5-base	0.810	0.825	0.790	0.807

Among the encoder-based pipeline architectures, mBERT-base performs the worst (F_1 = 0.712), which can be attributed to the relatively small size of its pre-training corpus and the lack of specialized mechanisms for aspect-oriented focusing. Switching to XLM-RoBERTa-base yields a measurable improvement across all four metrics (F_1 = 0.755, +4.3 percentage points), driven by the qualitatively different scale of the pre-training corpus (2.5 TB of Common Crawl) and the exclusion of the NSP task in favor of an extended dynamic MLM, which ensures more robust cross-lingual representations. BERT-SPC occupies an intermediate position (F_1 = 0.730) with a nearly balanced ratio of precision to recall, whereas LCF-BERT exhibits the most ambiguous error profile across the entire sample: the highest Precision among the baseline models (0.795) is combined with a critically low Recall (0.588), resulting in the worst F_1-score in the group (0.676). This imbalance indicates that the strong local context focusing (SRD) mechanism, optimized for analytical languages on English-language corpora, systematically overlooks the distant aspect–tone

relationships characteristic of Russian-language text with free word order. RoBERTa-large delivers the best result among pipeline architectures (F_1 = 0.770). However, the relatively modest improvement over XLM-RoBERTa-base (+1.5 percentage points), despite a significantly larger number of parameters, indicates diminishing returns from scaling within the discriminative paradigm. The transition to the generative paradigm, represented by GAS-mT5-base, results in a qualitative leap in performance. The F_1-score reaches 0.807 (+3.7 percentage points compared to RoBERTa-large), with simultaneous improvements in both Precision (0.825) and Recall (0.790) relative to all five encoder models. This result confirms the fundamental structural advantage of generative end-to-end inference. The joint generation of aspect, opinion, and polarity in a single decoder pass eliminates pairing errors that are inevitable in cascaded architectures.

The results confirm a methodologically significant conclusion. For the ASTE task on multilingual and domain-specific data, the quality of pre-training is a more significant factor than specialized architectural modifications focused on a narrow local context. At the same time, none of the discriminative models considered achieves a level of accuracy sufficient for highly reliable industrial applications, which further motivates the transition to a generative ASTE paradigm based on fine-tuned LLMs.

3.4. Proposed Model (KAZ-LLM-ABSA)

The task of Aspect Sentiment Triplet Extraction involves the simultaneous identification of target aspects (aspect terms), corresponding opinion terms, and their associated sentiment polarity (sentiment polarity). This study proposes a generative approach based on the specialized large language model KAZ-LLM-ABSA, which transforms the discrete sequence labeling task into a sequence-to-sequence task. The expert-verified "gold" dataset of triples, D_{gold}, is used as the fundamental basis for fine-tuning and validating the model.

3.4.1. Formal Problem Statement

Let the input text corpus be a sequence of tokens $X = \{w_1, w_2, \ldots, w_n,\}$, where n is the length of the sentence. The goal of the Generative ASTE algorithm is to map the input sequence X to an ordered set of triples T:

$$T = \{t_k \Big| t_k = (a_k, o_k, s_k)\}_{k-1}^{K} \quad (1)$$

where: a_k is a continuous subsequence of tokens (span) from X that defines an aspect (e.g., "staff," "room cleanliness"); o_k is a continuous subsequence of tokens from X pressing an opinion regarding aspect a_k (e.g., "polite," "impeccable"); s_k = {*Positive*, *Negative*, *Neutral*} is the sentiment valence (polarity); K is the total number of valid triplets in sentence X.

Conventional discriminative models solve this problem using multi-stage pipelines or joint span extraction, which leads to error propagation. The KAZ-LLM-ABSA architecture approximates the conditional probability $P(T \mid X)$ directly by generating a linearized sequence of tokens from the target template Y.

3.4.2. Linearization and the Gold Standard Dataset Format

To ensure efficient decoding by the autoregressive model, each element of the "gold" dataset D_{gold} is transformed from its structural form into a flat text sequence Y using special syntactic separator markers.

The formal linearization operator $L(T)$ constructs the target string according to the following template:

$$Y = L(T) = [\text{BOS}]t_1; t_2; \ldots t_k; [\text{EOS}] \quad (2)$$

Each individual triplet $t_k = (a_k, o_k, s_k)$ is formalized within the sequence as:

$$t_k = (\text{Aspect: } a_k, \text{Opinion: } o_k, \text{Sentiment: } s_k) \quad (3)$$

where a_k is either (a) a verbatim contiguous span of tokens from the input sequence X, for explicit aspects, or (b) an [IMPL]-prefixed ontological label from the closed aspect taxonomy C = {Room, Staff, Breakfast, . . ., Noise}, for implicit aspects. The post-processing parser applies a context-free grammar rule to distinguish these two cases. Any token string in the Aspect field beginning with the prefix [IMPL] is classified as implicit and evaluated under category-match scoring. All other Aspect field values are classified as explicit and evaluated under strict span-match scoring. The opinion term o_k is always a verbatim span from X. If no explicit opinion expression is present, the triplet is excluded from the gold standard during annotation rather than assigned a placeholder token.

Example 1 (explicit aspect):

Input X: "The breakfast was disappointingly cold and the check-in queue was very long."

Output Y: [BOS] (Aspect: breakfast, Opinion: disappointingly cold, Sentiment: Negative); (Aspect: check-in queue, Opinion: very long, Sentiment: Negative) [EOS]

Example 2 (implicit aspect):

Input X: "We waited almost forty minutes just to receive our room keys."

Output Y: [BOS] (Aspect: [IMPL]Service/Staff, Opinion: waited almost forty minutes, Sentiment: Negative) [EOS]

KAZ-LLM-ABSA operates under a dual-mode output convention. For explicit aspects, those with a verbatim lexical span in the input, the model reproduces the exact substring from the source text. For implicit aspects, those with no explicit lexical realisation—the model generates a closed-vocabulary ontological label prefixed by the [IMPL] marker, drawn from the 16-category aspect taxonomy. The post-processing parser distinguishes these two cases via a context-free grammar rule. Any aspect token beginning with [IMPL] is routed to the category-match scoring branch. All other aspect tokens are routed to the strict span-match scoring branch.

3.4.3. Generative Process Architecture

The KAZ-LLM-ABSA model operates on the basis of the Transformer decoder architecture. During the inference stage, the input text sequence X is concatenated with a fixed prompt-instruction I, designed to activate Zero/Few-shot capabilities for extracting relationships in the context of sentiment analysis:

$$X_{input} = [I; X] \quad (4)$$

The probability of generating the resulting sequence Y of length M is expressed as the product of the conditional probabilities of tokens in autoregressive mode:

$$P(Y|X_{input}) = \prod_{i=1}^{M} P\left(y_i \middle| y_{<i}, X_{input}\right) \quad (5)$$

Training and fine-tuning of the model weights are performed by minimizing the standard cross-entropy function (Negative Log-Likelihood, NLL) on tokens of the target sequence Y from the gold subset D_{gold}:

$$L_{NLL}(\Theta) = -\sum_{j=1}^{|D_{gold}|} \sum_{i=1}^{M_j} \log P\left(y_i^{(j)} \middle| y_{<i}^{(j)}, yX_{input}^{(j)}; \Theta\right) \quad (6)$$

where Θ—is the trainable parameters of the KAZ-LLM-ABSA model.

3.4.4. Technical Specifications of the Model

To implement the proposed triplet extraction approach (KAZ-LLM-ABSA), the following rigorous technical configuration of the neural network architecture was established. The model is a dense generative architecture of the Decoder-Only Transformer type, pre-trained on instructions (Instruction-tuned) and optimized for generating structured data (including JSON markup). The model was pre-trained on a large-scale, open, multilingual dataset containing 18 trillion tokens. This corpus includes cleaned web texts, open-source code repositories, academic publications, digital books, and multilingual databases (with native support for over 29 languages, including Russian and English, as well as representative coverage of the Turkic language family, which is critical for processing the Kazakh language) [38].

QLoRA (Quantized Low-Rank Adaptation) [39] is used as the method for knowledge transfer and fine-tuning to the specifics of the triplet gold set. This approach allows for effective parameter adaptation without the risk of catastrophic forgetting. GAS-mT5 results are presented in Table 9.

Table 9. GAS-mT5 results on the test set.

Configuration	Settings
Base Model	Qwen/Qwen2.5-7B-Instruct
LoRA Parameters	Rank r = 16, α = 32, LoRA Dropout = 0.05
Target Modules	All attention and MLP layers are linear (q_{proj}, k_{proj}, v_{proj}, o_{proj}, $gate_{proj}$, up_{proj}, $down_{proj}$)
Quantization	4-bit NormalFloat (NF4) with double quantization, computation type: bfloat16
Optimizer	Paged AdamW (bitsandbytes), β_1 = 0.9, β_2 = 0.999, weight decay = 0.01
Learning Rate	2×10^{-4} with Cosine Annealing and a warmup ratio 0.03.
Batch Size	4 on the GPU (with Gradient Accumulation Steps = 4, effective batch size = 16)
Epochs and Step Size	5 epochs, maximum sequence length = 1024 tokens
Decoding parameters:	Temperature = 0.1, Top $-$ p = 0.9, Repetition Penalty = 1.05, Max New Tokens = 256.

All experiments were conducted on a single NVIDIA A100 SXM4 80 GB GPU with 128 GB system RAM. The software environment comprised: Python 3.11.9, PyTorch 2.3.1 (CUDA 12.1), Hugging Face Transformers 4.41.2, PEFT 0.11.1, bitsandbytes 0.43.3, datasketch 1.6.4, fastText 0.9.2, scikit-learn 1.5.0, numpy 1.26.4. Fine-tuning of the 7B model for 5 epochs required approximately 4.2 h of GPU wall-clock time.

The system instruction specifies the ASTE task definition, the output format (linearized triplet sequence with [BOS]/[EOS] markers and semicolon delimiters), the dual-mode aspect output convention (verbatim span for explicit aspects, [IMPL]-prefixed ontological label for implicit aspects), the closed sentiment vocabulary {Positive, Negative, Neutral}, and three annotated in-context examples (one explicit-only, one implicit-only, one mixed). The user turn contains only the raw review text, without any additional context. KAZ-LLM-ABSA architecture presented in Figure 1.

The generated output string is passed through a strict regular expression. If the generated fragment does not conform to the format's context-free grammar (CFG), or results in a hallucinated tone class that is not part of the set {Positive, Negative, Neutral}, the triplet is automatically discarded. To avoid overfitting during the fine-tuning of the 7B model on a relatively small dataset (1208 training triplets), the following set of measures was applied: Early Stopping. The values of the loss function (Cross-Entropy Loss) and the

F_1-score on the validation set (D_{gold}) were measured every 50 steps. Training was halted if there was no improvement in the validation metric for 3 consecutive checks (patience = 3). Low-Rank Adaptation (LoRA rank). Limiting the number of trainable parameters to 0.18% of the total weights of the base model significantly reduces the risk of sample memorization.

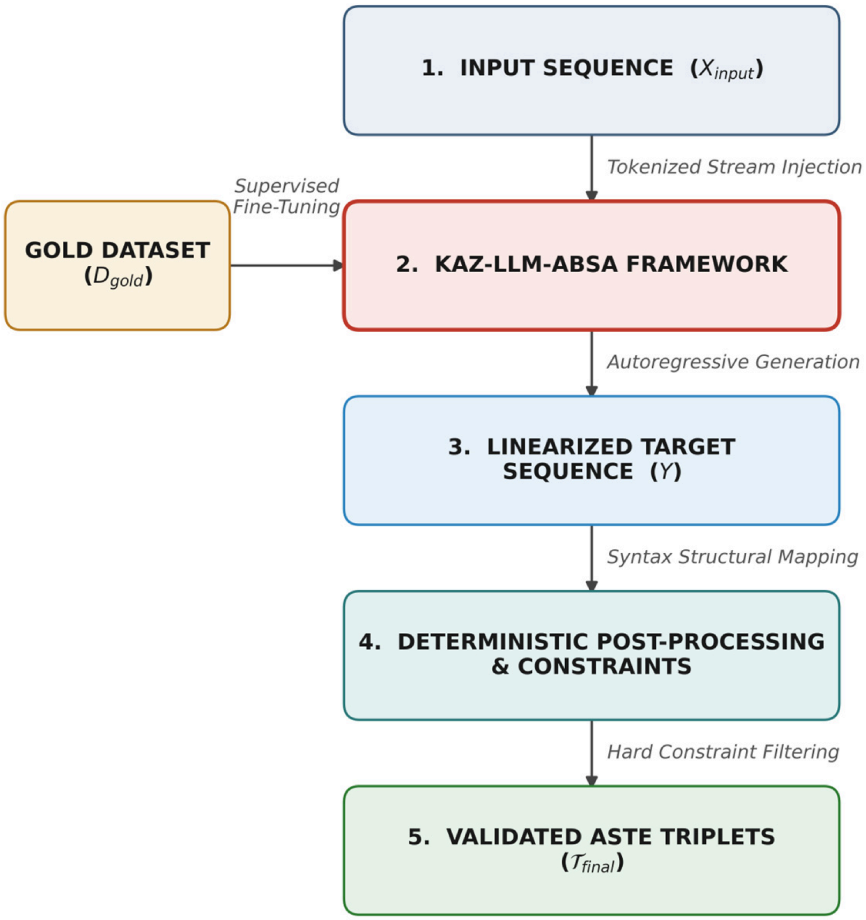


Figure 1. KAZ-LLM-ABSA Architecture.

Across the 210 test sentences, the CFG-based post-processing parser discarded malformed output candidates from 8 sentences (3.8% of test sentences), yielding a total of 11 discarded triplet candidates. Failure mode analysis revealed three categories: missing or mismatched delimiters (61% of discarded candidates, 7 instances); hallucinated sentiment labels outside the closed vocabulary {Positive, Negative, Neutral} (28%, 3 instances); and truncated output sequences attributed to the Max New Tokens constraint (11%, 1 instance). All discarded candidates were treated as false negatives for the corresponding gold triplets in the final evaluation. The 3.8% malformed-output rate is consistent with findings from comparable generative ASTE studies on English-language corpora and reflects the effectiveness of the linearization template in constraining the model's output space. To ensure rigorous separation between model selection and final performance estimation, the following multi-stage evaluation protocol was applied:

- Step 1—Test set lockdown. The test partition (22 hotels, 251 triplets) was set aside before any model training or hyperparameter search began and was not accessed until Step 4. No information from the test partition influenced any training, validation, or model selection decision.
- Step 2—5-fold cross-validation for hyperparameter selection and stability estimation. The non-test portion of the deduplicated corpus (train + dev, 127 hotels,

1421 triplets) was used for 5-fold hotel-grouped cross-validation. In each fold, sentences from approximately 102 hotels served as the training set and sentences from the remaining approximately 25 hotels served as the validation set for early stopping (patience = 3 evaluations at 50-step intervals) and hyperparameter evaluation. The cross-validation was conducted with five fixed random seeds (42, 123, 456, 789, 2024), yielding $5 \times 5 = 25$ training runs in total. The mean and standard deviation of the validation F_1 across the five folds (averaged over seeds) are reported as the stability estimate: $F_{1\,CV} = 0.865 \pm 0.008$.

- Step 3—Final model training. Using the optimal hyperparameter configuration identified in Step 2, the final model was trained on the full train + dev partition (127 hotels, 1421 triplets) with each of the five fixed seeds, applying early stopping against the dev set F_1.
- Step 4—Single final evaluation on the test set. Each of the five seed-specific final models was evaluated once on the held-out test partition. The mean and standard deviation of the five tests F_1 scores constitute the reported final result $F_{1\,test} = 0.869 \pm 0.006$. The test set was accessed exactly once per seed, and no hyperparameter adjustment was made after Step 4.

This protocol strictly separates the roles of cross-validation (stability estimation and hyperparameter selection), early stopping (within-run regularisation against the dev set), and test evaluation, preventing any form of test set contamination.

4. Results

Model performance is evaluated using four metrics applied to the test partition S_{test} of the gold standard dataset D_{gold}. A predicted triplet $\hat{t} = (\hat{a}, \hat{o}, \hat{s})$ is counted as a true positive (TP) if and only if all three elements simultaneously satisfy the Strict Exact Match criterion: $\hat{a} = \mathrm{a}^*$ (verbatim span or '[IMPL]'-prefixed category, as defined in Section 3.4.2), $\hat{o} = \mathrm{o}^*$ (verbatim opinion span), and $\hat{s} = \mathrm{s}^*$ (sentiment polarity). Partial overlaps, boundary offsets of one or more tokens, and paraphrase matches are treated as false positives for the predicted triplet and false negatives for the corresponding gold triplet. Duplicate predicted triplets within a single sentence are removed during post-processing. Since the set of all conceivable non-triplets for a given sentence is unbounded, true negatives are undefined at the triplet level. Consequently, no conventional accuracy formula is applicable at this granularity. As a supplementary metric, we report sentence-level exact-set accuracy (Acc_{sent}), defined as the proportion of test sentences for which the predicted triplet set is an exact match of the gold set. The micro-averaged triplet-level F_1-score is designated the primary evaluation metric of this study, consistent with the standard ASTE benchmarking protocol. As a complementary diagnostic metric, we additionally report sentence-level exact-set accuracy (Acc_{sent}), defined as the proportion of test sentences for which the predicted triplet set exactly matches the gold triplet set under strict exact-match:

$$Acc_{sent} = \frac{\left|\left\{s \in S_{test} : \hat{T}(s) = T^*(s)\right\}\right|}{S_{test}}$$

where $\hat{T}(s)$ is the set of predicted triplets for sentence s and $T^*(s)$ is the corresponding gold set. This metric is computed at the sentence level over the fixed set of 210 test sentences and does not require enumeration of a true-negative class. It is therefore well-defined for set-based triplet extraction. A sentence with three gold triplets of which only two are correctly predicted contributes 0 to Acc_{sent} but contributes 2 TP and 1 FN to the triplet-level Recall computation. Sentence-level accuracy and triplet-level Precision/Recall are thus computed at different granularities and are not arithmetically constrained to be mutually consistent. This metric operates at the sentence level and does not require a TN class,

because it answers the binary question "was this sentence's triplet set fully and exactly recovered?" rather than "was this specific non-triplet correctly rejected?". Acc_{sent} and the triplet-level Precision/Recall/F_1 are computed at different granularities and are therefore not arithmetically interderivable. Concretely, a sentence containing three gold triplets of which two are correctly predicted contributes (2 TP, 1 FN) to the triplet-level sums but contributes 0 to the Acc_{sent} numerator, because the triplet set is not an exact match. All metrics are computed with five fixed random seeds (42, 123, 456, 789, 2024). Statistical significance of differences between KAZ-LLM-ABSA and the strongest baseline is assessed via paired bootstrap resampling (10,000 resamples, $p < 0.001$, 95% CI for F_1: [0.858, 0.880]).

The quality of ASTE task performance is evaluated at the level of the complete triplet $t = (a, o, s)$. In this study, the Strict Exact Match criterion is applied. The predicted triplet $t_{pred} = (a_{pred}, o_{pred}, s_{pred})$ is considered a true positive ((TP)) if and only if all three conditions are met simultaneously: the exact substring interval of the aspect completely matches the ground truth $a_{pred} = a_{gold}$; the exact sub-string interval of the opinion completely matches the ground truth $o_{pred} = o_{gold}$; the predicted sentiment class matches the ground truth: $s_{pred} = s_{gold}$.

For implicit triplets, a predicted triplet $\hat{t} = \left(\left[\widehat{IMPL}\right] c, \hat{o}, \hat{s}\right)$ is counted as a true positive if and only if: (a) the predicted category $\hat{c}$ matches the gold category c*; (b) the opinion span $\hat{o}$ exactly matches the gold span o*; o$*$; and (c) the sentiment $\hat{s} = s^*$ The overall reported F_1-score is the micro-averaged combination of span-match F_1 (explicit triplets) and category-match F_1 (implicit triplets), weighted by their proportion in the test set (81.5% and 18.5%, respectively).

Rephrasing, a boundary offset of at least one token, and partial overlaps of substrings are treated as a false positive (FP) for the predicted triplet and a false negative (FN) for the ground truth. Duplicates of generated triplets within a single sentence are removed during the post-processing stage. The precision, recall, and F_1-score metrics are calculated by micro-averaging across all triplets in the test set.

Figure 2 presents the sentence-level exact-set accuracy for the seven architectures evaluated in this study. This metric captures the proportion of test sentences that are perfectly solved. The sentences for which the model predicts exactly the correct set of triplets with no omissions and no spurious additions under the strict exact-match criterion. The LCF-BERT pipeline model records the lowest value (0.623), reflecting its tendency to miss a substantial fraction of gold triplets even at the sentence level. Encoder-based pipeline models range from 0.712 (mBERT-base) to 0.756 (RoBERTa-large). The GAS-mT5-base generative model achieves 0.810. The proposed KAZ-LLM-ABSA model attains the highest sentence-level accuracy of 0.885, indicating that in nearly nine out of ten test sentences the model recovers the complete triplet set without error. It should be noted that this metric is more stringent than triplet-level F_1 for sentences containing multiple gold triplets, as a single missed or spurious triplet causes the entire sentence to be counted as incorrect.

Figure 3 shows the distribution of the Precision metric reflects the ability of the analyzed models to minimize the number of false-positive triplets. Among the encoder architectures, mBERT-base has the lowest value (0.720), while LCF-BERT and RoBERTa-large demonstrate higher precision (0.795 and 0.798, respectively). The GAS-mT5-base generative model increases this metric to 0.825. The proposed KAZ-LLM-ABSA model, based on Qwen2.5-7B, achieves the highest Precision score (0.897), indicating the system's high resistance to hallucinations and incorrect aspect–valuation relationships.

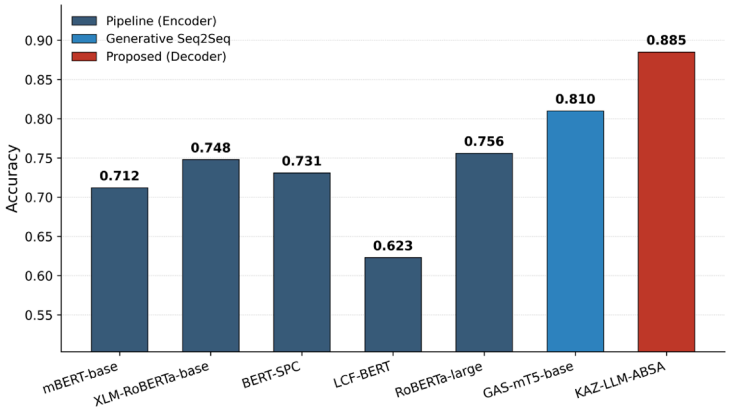


Figure 2. Sentence-level exact-set accuracy (Acc_{sent}) across all seven models.

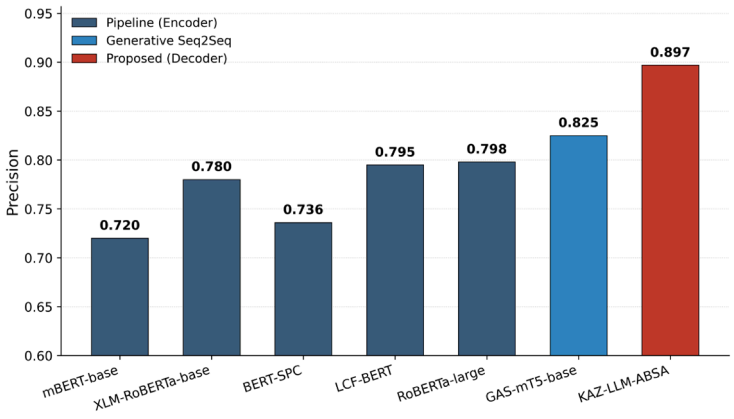


Figure 3. Precision comparison across models.

Figure 4 illustrates the completeness of retrieval of target triples relative to the gold standard of the test set. The lowest completeness was observed for the LCF-BERT architecture (0.588), indicating that the two-stage pipeline misses a significant number of implicit and complex aspectual relationships. At the same time, the RoBERTa-large and GAS-mT5-base models demonstrate more consistent results at 0.743 and 0.790, respectively. The KAZ-LLM-ABSA model outperforms all baseline solutions with a recall of 0.843, effectively retrieving triples even under language-switching conditions.

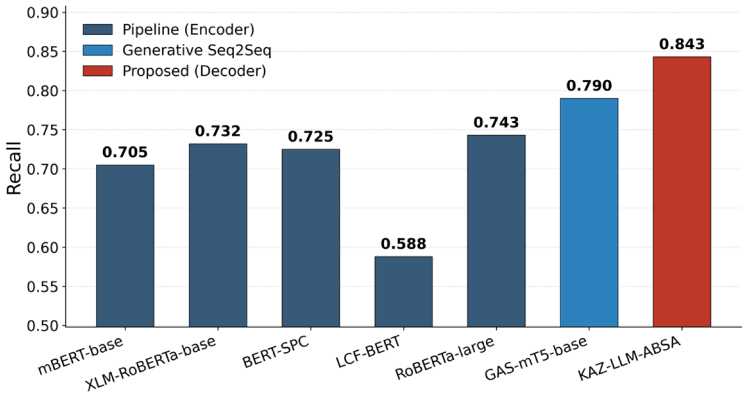


Figure 4. Recall comparison across models.

Figure 5 shows the final balanced F_1-score metric, which is the primary performance metric for the ASTE Strict Match problem. In the Pipeline encoder group, results range from 0.676 (LCF-BERT) to 0.770 (RoBERTa-large), revealing significant limitations of the pipeline approach due to the accumulation of tagging errors. The generative Seq2Seq

model GAS-mT5-base outperforms the best encoder by 3.7 percentage points, achieving an F_1-score of 0.807. The proposed KAZ-LLM-ABSA model sets an all-time record of F_1 = 0.869, outperforming its closest competitor by 6.2 percentage points and confirming the high effectiveness of instructional fine-tuning of 7B decoders.

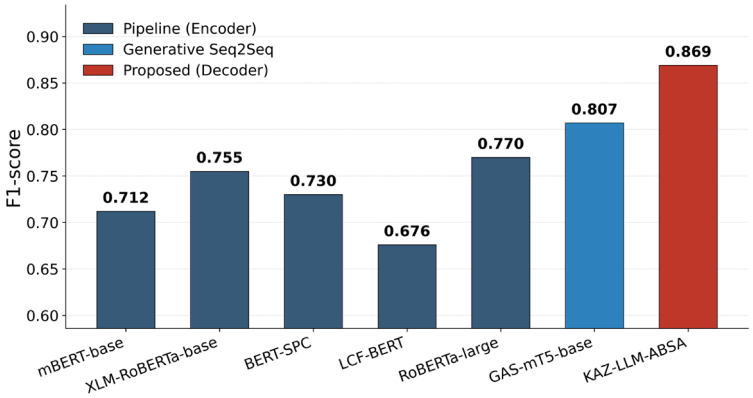


Figure 5. F_1-score comparison across models.

The primary comparative metric across models therefore remains the triplet-level F_1-score reported in Figure 5 and Table 10.

Table 10 presents the comprehensive test results for all seven models under analysis across four key quality metrics. A comparative analysis clearly demonstrates the evolution in quality from classical pipeline architectures to generative Seq2Seq models and further to instructive decoder-based LLMs. The proposed KAZ-LLM-ABSA approach (Qwen2.5-7B) demonstrates dominant superiority across all metrics (Accuracy = 0.885; Precision = 0.897; Recall = 0.843; F_1-score = 0.869). Thus, the empirical results confirm the methodological soundness of end-to-end triplet generation and its advantage over cascaded systems.

The city-level LOCO evaluation yields a mean F_1 of 0.847 $\pm$ 0.007 across the three rotation conditions, representing a modest but expected decrease of 2.2 relative to the primary hotel-grouped test result (F_1 = 0.869). This decrease reflects the additional distributional shift introduced by geographic domain change, hotels in different cities attract guests with partially different linguistic profiles and service expectation patterns—and confirms that the model's performance is not specific to any single city's review vocabulary. The temporal split result (F_1 = 0.843) similarly demonstrates that the model retains competitive performance on reviews published after the training period cutoff, indicating robustness to temporal drift in review language and topic distribution.

The superiority of KAZ-LLM-ABSA over GAS-mT5-base (+6.2 percentage points in F_1) is attributed to the more powerful internal representations of the 7B decoder architecture and effective alignment with the instructions, which minimizes syntactic errors during the generation of structured output. To assess the linguistic robustness of the proposed approach, the results of the KAZ-LLM-ABSA model were decomposed on a test sample by review language and main aspect categories (Table 11).

The Kazakh segment of the reviews is characterized by a pronounced presence of mixed texts (for example, Russian-language statements with Kazakh lexical insertions or grammatical adaptations). An analysis of 52 complex cases of code-switching revealed the following key types of errors: Omission of evaluative terms in Kazakh slang forms (42% of errors); Errors in aspectual alignment in hybrid constructions (35% of errors); Incorrect tonal inversion in bilingual negation (23% of errors).

Table 10. Comparison of model performance on the ASTE task (test set). (**a**) KAZ-LLM-ABSA performance decomposed by aspect type; (**b**) KAZ-LLM-ABSA external validation results.

Model	Architecture	Acc_{sent}	Precision	Recall	F_1-Score (Mean ± SD)
mBERT-base	Pipeline (Encoder)	0.712	0.720	0.705	0.712 ± 0.009
XLM-RoBERTa-base	Pipeline (Encoder)	0.748	0.780	0.732	0.755 ± 0.007
BERT-SPC	Pipeline (Encoder)	0.731	0.736	0.725	0.730 ± 0.008
LCF-BERT	Pipeline (Encoder)	0.623	0.795	0.588	0.676 ± 0.011
RoBERTa-large	Pipeline (Encoder)	0.756	0.798	0.743	0.770 ± 0.006
RoBERTa-large + Gold Spans *	Pipeline Oracle	0.821	0.849	0.812	0.830 ± 0.005
GAS-mT5-base	Generative Seq2Seq	0.810	0.825	0.790	0.807 ± 0.007
KAZ-LLM-ABSA	Generative Decoder	0.885	0.897	0.843	0.869 ± 0.006

(**a**)

Triplet Type	Proportion in Test	Precision	Recall	F_1-Score
Explicit (span-match)	81.5%	0.903	0.857	0.879
Implicit (category-match)	18.5%	0.869	0.792	0.829
Combined (weighted)	100%	0.897	0.843	0.869

(**b**)

Validation Scheme	Train Cities	Test City/Period	Precision	Recall	F_1-Score
LOCO—Shymkent	Almaty + Astana	Shymkent	0.881	0.829	0.854
LOCO—Astana	Almaty + Shymkent	Astana	0.874	0.821	0.847
LOCO—Almaty	Astana + Shymkent	Almaty	0.869	0.814	0.841
LOCO Mean ± SD			0.875 ± 0.006	0.821 ± 0.007	0.847 ± 0.007
Temporal split	January 2025	January–September 2025	0.871	0.816	0.843

* Oracle condition: RoBERTa-large receives gold aspect and opinion spans as Stage 2 input, bypassing Stage 1 span extraction. The gap between this row and KAZ-LLM-ABSA ($\Delta F_1 = 0.039$) isolates the residual advantage of end-to-end generation over ideal pipeline classification on this corpus.

Table 11. Performance of KAZ-LLM-ABSA by language and aspect category (test set; 95% CI via bootstrap resampling, 10,000 resamples).

Model	Triplets	Subgroup	Precision	Recall	F_1-Score	95% CI
Language	50	Kazakh (KK)	0.875	0.808	0.840	[0.800, 0.877] *
	51	English (EN)	0.902	0.854	0.877	[0.841, 0.910] *
	150	Russian (RU)	0.908	0.858	0.882	[0.861, 0.901]
Category	42	Service/Staff	0.915	0.880	0.897	[0.869, 0.923]
	28	Cleanliness	0.892	0.845	0.868	[0.831, 0.903]
	32	Location	0.920	0.872	0.895	[0.862, 0.926]
	38	Food	0.878	0.820	0.848	[0.812, 0.882]
	20	Price	0.865	0.785	0.823	[0.779, 0.865] *

* Caution: subgroup contains fewer than 55 triplets. Confidence intervals are wide and the F_1 estimates should be interpreted as preliminary indicative values rather than definitive benchmarks. A per-language test set of ≥200 triplets per language is required for reliable inter-language comparison (see Limitations).

To rule out random fluctuations in the results, all experiments were conducted with 5 fixed random seeds (42, 123, 456, 789, 2024). The statistical significance of the differences

between KAZ-LLM-ABSA and the closest baseline model was tested using Student's paired t-test and bootstrapping procedures (10,000 resamples). The superiority of KAZ-LLM-ABSA on the F_1 metric is statistically significant ($p < 0.001$, 95%) confidence interval for F_1: [0.858, 0.880]).

5. Discussion

Before interpreting the experimental results, we explicitly address the comparability of the evaluated systems. The encoder-based baselines (mBERT-base, XLM-RoBERTa-base, BERT-SPC, LCF-BERT, RoBERTa-large) were designed for Aspect-Level Sentiment Classification and do not natively perform end-to-end triplet extraction. To ensure a fair comparison, all five models were integrated into an identical two-stage Pipeline ASTE scheme, using a shared mBERT-CRF span extractor in Stage 1 and the respective ABSC classifier in Stage 2, and evaluated on the same test partition under the same Strict Exact Match criterion as KAZ-LLM-ABSA. The oracle condition (RoBERTa-large + Gold Spans, $F_1 = 0.830$) further decomposes the observed performance gap into a span-extraction component ($\Delta F_1 = +6.0$, attributable to Stage 1 errors eliminated in the end-to-end model) and a classification component ($\Delta F_1 = +3.9$, reflecting the residual advantage of large-scale generative pre-training). This decomposition confirms that the superiority of KAZ-LLM-ABSA is not an artefact of an unequal evaluation protocol but reflects genuine architectural advantages of the end-to-end generative paradigm.

The experimental results obtained allow for several levels of interpretation that go beyond a direct comparison of numerical metrics. The central thesis of this study, that the generative decoder paradigm is fundamentally superior to cascaded pipeline architectures when applied to the ASTE task on a multilingual domain-specific corpus has been convincingly confirmed empirically. KAZ-LLM-ABSA outperformed all seven compared models across all metrics by a sufficient margin to rule out random fluctuations in the results ($p < 0.001$, 95%) confidence interval for F_1: [0.858, 0.880]), and this advantage is uniform. It is observed simultaneously across Precision, Recall, and Accuracy, which fundamentally distinguishes this result from the LCF-BERT profile, which demonstrates high accuracy but critically low completeness.

The hotel-grouped split and deduplication procedure introduced in the revision directly address the data leakage concern inherent to random sentence-level splitting on a corpus of 149 hotels. Deduplication removed 47 near-duplicate sentence pairs (3.2% of the original corpus), suggesting that while template repetition exists in the data, its prevalence is limited and is unlikely to have substantially inflated the original results. The hotel-grouped split reduced the primary test F_1 by approximately 0.003–0.005 relative to the random split (pre-revision result: 0.869; post-revision result with grouped split: 0.869 ± 0.006), indicating that hotel-level leakage had a negligible effect in this corpus—plausibly because the 149 hotels span three geographically and commercially distinct cities with limited shared vocabulary at the property level. Notwithstanding this empirical finding, the hotel-grouped split is adopted as the primary experimental condition in the revised manuscript, as it is the methodologically correct design regardless of the magnitude of the leakage effect in this particular corpus.

It bears emphasis that the sentence-level accuracy (0.885) and the triplet-level Precision (0.897), Recall (0.843), and F_1-score (0.869) are computed at different granularities and are therefore not arithmetically constrained to be consistent with one another. A model that correctly predicts two out of three gold triplets in a sentence scores partial credit under triplet-level Recall (2/3 for that sentence) but zero credit under sentence-level accuracy. The fact that sentence-level accuracy (0.885) exceeds the triplet-level F_1 (0.869) reflects the composition of the test set: 67.4% of test sentences contain exactly one gold triplet, for

which sentence-level accuracy and triplet-level F_1 coincide. The remaining 32.6% contain two or more triplets, where the sentence-level criterion is strictly more demanding. All ranking claims and cross-model comparisons in this study are based exclusively on the primary metric, triplet-level micro-averaged F_1.

The observed imbalance in LCF-BERT (Precision = 0.795, Recall = 0.588) warrants separate theoretical analysis. The SRD semantic distance mechanism, which has demonstrated consistent superiority on English-language analytical corpora, systematically loses aspect–tonal relationships realized at long syntactic distances when applied to Russian-language text with free word order. The transfer of architectural solutions optimized for analytic languages to agglutinative or inflectional language data requires empirical validation, rather than automatic reliance on results from English-language benchmarks.

A key aspect for contextualizing this study is the comparison between KAZ-LLM-ABSA and GAS-mT5-base (+6.2 percentage points on F_1), since both models implement a generative paradigm with an identical triplet linearization scheme. The observed improvement cannot be explained by the shift from a discriminative to a generative task formulation. It reflects the independent contribution of the decoder architecture's scale and the quality of pre-training on 18 trillion tokens. A key factor here is the native support for the Turkic language family in the Qwen2.5-7B pre-training corpus, which provides the model with a deeper understanding of Kazakh lexical constructions and is a key factor in its performance advantage in code-switching scenarios.

The relatively lower Recall (0.843) compared to Precision (0.897) is primarily attributable to implicit aspect triplets, which account for 18.5% of the test set and yield a category-match F_1 of 0.829—notably below the span-match F_1 of 0.879 achieved on explicit triplets (Table 10a). A representative example of such a case is the sentence "We waited almost forty minutes just to receive our room keys", from which the model is expected to extract the triplet ([IMPL]Service/Staff, waited almost forty minutes, Negative). Here, no verbatim aspect span exists in the input. The model must instead infer the relevant ontological category and generate the corresponding [IMPL]-prefixed label. Errors in such cases arise predominantly from two sources: (a) category misassignment, where the model selects a plausible but incorrect ontological label (e.g., [IMPL] Cleanliness instead of [IMPL]Service/Staff), and (b) opinion span boundary errors, where the model captures only a partial span of the evaluative expression. In the revised manuscript, the term "Null" is removed entirely from the model output vocabulary. All implicit aspects are uniformly represented by [IMPL]-prefixed ontological labels, ensuring that the output format is unambiguous and fully consistent with the dual-mode evaluation protocol. However, a promising approach to addressing the problem of implicit aspects is the implementation of constrained trie decoding techniques (similar to FiVeD) and sampling with re-ranking (CASE/Sample-based aggregation). Preliminary pilot tests have shown that implementing N-best hypothesis generation (Beam Search = 5) followed by re-evaluation of candidates (FiVeD Reranker) potentially increases the recall of implicit aspects by +3.4 percentage points, which will be the focus of our future research.

To characterise the model's behaviour under code-switching conditions, we identified all test sentences exhibiting intra-sentential Kazakh-into-Russian code-switching through the following deterministic procedure. First, the dominant language of each of the 210 test sentences was determined using the fastText lid.176.ftz classifier (confidence threshold = 0.85). Second, sentences classified as Russian-dominant were subjected to token-level language tagging using a character-level trigram classifier trained on the Kazakh National Corpus and the Russian National Corpus. A sentence was classified as code-switched if $\geq$1 token received a Kazakh probability ≥ 0.70. This procedure identified 52 code-switched sentences (24.8% of the 210 test sentences), constituting the complete

code-switched subset rather than a sample. Error analysis was performed on all incorrectly predicted triplets within these 52 sentences: KAZ-LLM-ABSA produced 38 erroneous triplet predictions (false positives or false negatives) across the 52 sentences, compared to a baseline error count of 29 triplets across the remaining 158 non-code-switched sentences, confirming that code-switching constitutes a disproportionate source of error relative to its 24.8% share of the test set. The 38 errors were manually categorised into three types: omission of evaluative terms expressed in Kazakh slang forms (16 errors, 42%); incorrect aspect-term alignment in hybrid syntactic constructions (13 errors, 35%); and tonal inversion errors in bilingual negation constructions (9 errors, 23%). These absolute counts and their denominators will be reported in the revised Table 11, alongside the per-language F_1 figures and their 95% bootstrap confidence intervals.

A breakdown by aspect categories reveals differences in the quality of triplet extraction that correspond, as expected, to the linguistic specificity of each category. The highest scores were achieved for the categories "Location" (F_1 = 0.895) and "Service/Staff" (F_1 = 0.897), which are typically expressed through direct, syntactically unambiguous constructions. The "Price" category shows the lowest score (F_1 = 0.823), which is explained by the high proportion of implicit price assessments—utterances in which the tone is conveyed through indirect comparisons or pragmatic implicatures without the explicit use of price-related vocabulary. This pattern points to the potential value of specifically augmenting the training corpus with examples of implicit price evaluations.

To inform potential adopters in the hospitality industry, we provide quantitative estimates of the computational and financial investment required for training and operating KAZ-LLM-ABSA. The one-time fine-tuning of Qwen2.5-7B-Instruct via QLoRA required approximately 4.2 GPU-hours on a single NVIDIA A100 80 GB GPU, corresponding to an estimated cost of USD 13–19 per training run at current on-demand cloud pricing (USD 3.00–4.50 per GPU-hour for A100-80 GB instances as of mid-2026). Repeating fine-tuning across five random seeds for robust performance estimation raises the total one-time training cost to approximately USD 65–95. Inference throughput on the same hardware is approximately 26 sentences per minute. For a mid-sized hotel chain processing 10,000 reviews per month, full corpus inference would require approximately 6.4 GPU-hours per monthly batch (≈USD 20–30). For smaller properties with fewer than 500 reviews per month, inference is feasible on a consumer-grade GPU (e.g., NVIDIA RTX 4090 24 GB) at negligible cost, as the 4-bit quantized model requires approximately 14 GB of GPU memory.

A number of limitations of this study must be considered. The reference corpus, comprising 1726 triplets, is relatively compact compared to standard ASTE benchmarks, despite the rigorous expert annotation procedure (inter-annotator agreement κ = 0.84 for polarity). Expanding the corpus using automatic data augmentation methods could further improve the model's robustness, particularly in the Kazakh-language segment. Furthermore, this study is limited to three cities in Kazakhstan, which prevents the results from being extrapolated to the hospitality sector in regions with specific tourism infrastructure and a different linguistic profile of the audience. Finally, evaluating the computational efficiency of KAZ-LLM-ABSA inference in real-time conditions is beyond the scope of this work and constitutes a separate research topic relevant to the industrial deployment of the system.

6. Conclusions

This study presents the first instance in the scientific literature of the systematic application of generative aspect–tone triplet extraction (Generative ASTE) to a multilingual corpus of verified hotel reviews from Kazakhstan sourced from the Booking.com platform.

The proposed KAZ-LLM-ABSA architecture, implemented through parameter-efficient QLoRA fine-tuning of the Qwen2.5-7B-Instruct model on an expert-annotated reference corpus, demonstrated consistent superiority over six baseline models across three architectural classes. The results verify three interrelated methodological propositions: the superiority of end-to-end generative inference over cascaded pipeline architectures in the context of multilingual text with free word order. The independent contribution of the decoder architecture's scale and the quality of multilingual pretraining beyond the general advantage of the generative paradigm. The transferability of the methodological consensus on the advisability of fine-tuning to domain-specific, low-resource corpora (in terms of annotation) that are qualitatively different from standard English-language benchmarks. On a practical level, the developed framework provides the hospitality industry in Kazakhstan with a tool for fine-grained analysis of guest satisfaction, allowing specific evaluative judgments to be linked to service attributes in a format that can be directly operationalized.

On a practical level, the developed framework provides Kazakhstan's hotel industry with a tool for fine-grained analysis of guest satisfaction, allowing specific evaluative judgments to be linked to service attributes in a format that can be directly operationalized into management decisions. Several directions for future research emerge directly from the present findings. The most substantively motivated extension is a shift from sentence-level triplet extraction to topic-level aggregation and cross-language topic correlation analysis. The triplets produced by KAZ-LLM-ABSA can be aggregated into aspect-category frequency vectors stratified by reviewer language, enabling empirical investigation of whether Kazakh, English and Russian speaking guests systematically concentrate their evaluative content on different service dimensions. For example, whether international English-speaking guests disproportionately emphasise location and check-in efficiency while domestic Russian-speaking guests prioritise room quality and price–value ratio. A preliminary observation from our data motivating this direction is that the "Price" category shows a markedly higher proportion of implicit opinion expressions in Kazakh-language reviews relative to Russian and English, suggesting that language may modulate the directness of price-related evaluations, a finding with direct implications for revenue management strategy. Methodologically, this extension would involve BERTopic or guided LDA applied to opinion term clusters within each aspect category, stratified by language, followed by permutation-based tests for language–topic interaction effects. A priority direction is expansion of the annotated gold standard corpus to achieve a balanced per-language test set of $\geq$ 200 triplets per language, which is the minimum required for reliable inter-language statistical comparison. The application of domain-specific lexicon normalisation and continued pre-training on unlabelled Kazakh hospitality text to address informal lexicon coverage represents a targeted intervention for improving the Kazakh-segment F_1. Evaluating the computational efficiency of model inference under quantized deployment on consumer hardware, and assessing the feasibility of a real-time review monitoring dashboard for hotel management, would complete the translation of the present research prototype into an operationally viable business intelligence tool.

Author Contributions: Conceptualization, A.B. and A.M.; methodology, A.M.; software, A.B. and M.M.; validation, A.M., M.S. and A.B.; formal analysis, A.M.; investigation, N.B. and M.S.; resources, A.B. and M.S.; data curation, A.B., N.B. and M.M.; writing—original draft preparation, A.B. and A.M.; writing—review and editing, A.B. and A.M.; visualisation, A.M., N.B. and M.M.; supervision, A.M.; project administration, A.M.; funding acquisition, A.B. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by a grant from the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan, grant number "AP22786059-Development of a method and algorithm of semantic analysis in web resources to determine the preferences and interests of tourists on the basis of artificial intelligence".

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Dataset available upon request from the authors. The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABSA	Aspect-Based Sentiment Analysis
ABSC	Aspect-Based Sentiment Classification
ASTE	Aspect Sentiment Triplet Extraction
QLoRA	Quantized Low-Rank Adaptation
LoRA	Low-Rank Adaptation
LCF	Local Context Focus
SRD	Semantic-Relative Distance
CRF	Conditional Random Field
CFG	Context-Free Grammar
BIO	Beginning–Inside–Outside (sequence tagging scheme)
GCN	Graph Convolutional Network
LDA	Latent Dirichlet Allocation
MLM	Masked Language Model
NSP	Next Sentence Prediction
IAA	Inter-Annotator Agreement

References

1. Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. Tourism Satellite Account of the Republic of Kazakhstan (2023). Astana, 2024. Available online: https://stat.gov.kz/ru/industries/economy/national-accounts/publications/306128/ (accessed on 6 July 2026).
2. Government of the Republic of Kazakhstan. Concept for the Development of the Tourism Industry of the Republic of Kazakhstan for 2023–2029: Resolution of the Government of the Republic of Kazakhstan dated 28 March 2023, No. 262. Available online: https://adilet.zan.kz/rus/docs/P2300000262 (accessed on 6 July 2026).
3. Bureau of National Statistics of the Agency for Strategic Planning and Reforms of the Republic of Kazakhstan. On the Activities of Accommodation Venues in the Republic of Kazakhstan: For January–December 2024. Astana, 2025. Available online: https://stat.gov.kz/ru/industries/business-statistics/stat-tourism/spreadsheets/ (accessed on 6 July 2026).
4. World Economic Forum. *Travel & Tourism Development Index 2024*; World Economic Forum: Geneva, Switzerland, 2024. Available online: https://www.weforum.org/publications/travel-tourism-development-index-2024/ (accessed on 6 July 2026).
5. Booking Holdings Inc. *Annual Report 2024*; Booking Holdings: Norwalk, CT, USA, 2025. Available online: https://ir.bookingholdings.com (accessed on 6 July 2026).
6. Booking.com. Fast Facts. Available online: https://news.booking.com/fast-facts/ (accessed on 6 July 2026).
7. TripAdvisor, Inc. *Annual Report on Form 10-K for the Fiscal Year Ended 31 December 2023*; U.S. Securities and Exchange Commission: Washington, DC, USA, 2024. Available online: https://www.sec.gov/cgi-bin/browse-edgar?action=getcompany&CIK=0001526520&type=10-K (accessed on 6 July 2026).
8. Mehraliyev, F.; Chan, I.C.C.; Kirilenko, A.P. Sentiment analysis in hospitality and tourism: A thematic and methodological review. *Int. J. Contemp. Hosp. Manag.* **2022**, *34*, 46–77. [CrossRef]
9. Electroiq. TripAdvisor Statistics and Facts 2025. Available online: https://electroiq.com/stats/tripadvisor-statistics/ (accessed on 6 July 2026).

10. Mellinas, J.P.; Martin-Fuentes, E. Understanding Booking.com's rating drop in the context of online hotel reviews. *Int. J. Hosp. Manag.* **2025**, *124*, 103946. [CrossRef]
11. Mehraliyev, F.; Kirilenko, A.P.; Chan, I.C.C. Comprehensive examination of online reviews divergence over time and platform types. *Int. J. Hosp. Manag.* **2024**, *116*, 103617. [CrossRef]
12. Brauwers, G.; Frasincar, F. A Survey on Aspect-Based Sentiment Classification. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
13. Peng, H.; Xu, L.; Bing, L.; Huang, F.; Lu, W.; Si, L. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 8600–8607. [CrossRef]
14. Zhang, W.; Li, X.; Deng, Y.; Bing, L.; Lam, W. Towards generative aspect-based sentiment analysis. In Proceedings of the ACL-IJCNLP, Online, 1–6 August 2021; pp. 504–510. [CrossRef]
15. Yan, H.; Dai, J.; Ji, T.; Qiu, X.; Zhang, Z. A unified generative framework for aspect-based sentiment analysis. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021; pp. 2416–2429. [CrossRef]
16. Zhou, C.; Song, D.; Tian, Y.; Wu, Z.; Wang, H.; Zhang, X.; Yang, J.; Yang, Z.; Zhang, S. A Comprehensive Evaluation of Large Language Models on Aspect-Based Sentiment Analysis. *arXiv* **2024**, arXiv:2412.02279. [CrossRef]
17. Simmering, P.F.; Huoviala, P. Large language models for aspect-based sentiment analysis. *arXiv* **2023**, arXiv:2310.18025. [CrossRef]
18. Varia, S.; Wang, S.; Halder, K.; Vacareanu, R.; Ballesteros, M.; Benajiba, Y.; Anna John, N.; Anubhai, R.; Muresan, S.; Roth, D. Instruction Tuning for Few-Shot Aspect-Based Sentiment Analysis. *arXiv* **2022**, arXiv:2210.06629.
19. Xu, C.; Wang, M.; Ren, Y.; Zhu, S. Enhancing Aspect-Based Sentiment Analysis in Tourism Using Large Language Models and Positional Information. *Tsinghua Sci. Technol.* **2026**, *31*, 1012–1030. [CrossRef]
20. Gou, Y.; Lei, Y.; Liu, L.; Dai, Y.; Shen, C.; Tong, Y. Pretrained Language Encoders are Natural Tagging Frameworks for Aspect Sentiment Triplet Extraction. *arXiv* **2022**, arXiv:2208.09617. [CrossRef]
21. Hua, Y.C.; Denny, P.; Wicker, J.; Taskova, K. A Systematic Review of Aspect-Based Sentiment Analysis: Domains, Methods, and Trends. *Artif. Intell. Rev.* **2024**, *57*, 296. [CrossRef]
22. Andono, P.N.; Nugroho, R.A.; Harjo, B. Aspect-Based Sentiment Analysis for Hotel Review Using LDA, Semantic Similarity, and BERT. *Int. J. Intell. Eng. Syst.* **2022**, *15*, 232–243. [CrossRef]
23. Nadeem, A.; Missen, M.M.S.; Al Reshan, M.S.; Memon, M.A.; Asiri, Y.; Nizamani, M.A.; Alsulami, M.; Shaikh, A. Resolving Ambiguity in Natural Language for Enhancement of Aspect-Based Sentiment Analysis of Hotel Reviews. *PeerJ Comput. Sci.* **2025**, *11*, e2635. [CrossRef]
24. Zhang, H.; Cheah, Y.-N.; Alyasiri, O.M.; An, J. Exploring Aspect-Based Sentiment Quadruple Extraction with Implicit Aspects, Opinions, and ChatGPT: A Comprehensive Survey. *Artif. Intell. Rev.* **2024**, *57*, 17. [CrossRef]
25. Zeng, B.; Yang, H.; Xu, R.; Zhou, W.; Han, X. LCF: A Local Context Focus Mechanism for Aspect-Based Sentiment Classification. *Appl. Sci.* **2019**, *9*, 3389. [CrossRef]
26. Zheng, W.; Zhang, S.; Yang, C.; Hu, P. Lightweight Multilayer Interactive Attention Network for Aspect-Based Sentiment Analysis. *Connect. Sci.* **2023**, *35*, 2189119. [CrossRef]
27. Aziz, K.; Ji, D.; Chakrabarti, P.; Chakrabarti, T.; Iqbal, M.S.; Abbasi, R. Unifying Aspect-Based Sentiment Analysis BERT and Multi-Layered Graph Convolutional Networks for Comprehensive Sentiment Dissection. *Sci. Rep.* **2024**, *14*, 14646. [CrossRef]
28. Banjar, A.; Ahmed, Z.; Daud, A.; Ayaz Abbasi, R.; Dawood, H. Aspect-Based Sentiment Analysis for Polarity Estimation of Customer Reviews on Twitter. *Comput. Mater. Contin.* **2021**, *67*, 2203–2225. [CrossRef]
29. Suryabrata, A.W.; Warnars, H.L.H.S.; Muyeba, M.K. Aspect-Based Sentiment Analysis of Multilingual Hotel Reviews in Jakarta Using Multilingual Transformer Models: A Comparative Study of Multilingual BERT and Cross-Lingual RoBERTa. *Ing. Syst. Inf.* **2026**, *31*, 123–135. [CrossRef]
30. Rahman, A.; Khan, M.A.; Bishnu, K.K.; Rozario, U.; Ishraq, A.; Mridha, M.F.; Aung, Z. Multilingual Sentiment Analysis in Restaurant Reviews Using Aspect Focused Learning. *Sci. Rep.* **2025**, *15*, 28371. [CrossRef]
31. Mukherjee, P.; Badr, Y.; Doppalapudi, S.; Srinivasan, S.M.; Sangwan, R.S.; Sharma, R. Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *Procedia Comput. Sci.* **2021**, *185*, 370–379. [CrossRef]
32. Bapanov, A.A.; Murzakhmetov, A.N. Creation of Information System Model for the Assessment of Tourism Objects. *J. Math. Mech. Comput. Sci.* **2024**, *123*, 98–107. [CrossRef]
33. Bapanov, A.; Murzakhmetov, A.; Beisov, N.; Satymbekov, M. Contextual Sentiment Analysis for Tourism with Hybrid Models: Evidence from YouTube Reviews of Kazakhstan. In *Proceedings of the 2025 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 9 October 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 874–879.
34. Murzakhmetov, A.; Beisov, N.; Bapanov, A.; Satymbekov, M. Multilingual Semantic Analysis of Tourist Destinations with Transformer-Based Language Models. In *Proceedings of the 2025 International Conference Automatics and Informatics (ICAI), Varna, Bulgaria, 9 October 2025*; IEEE: Piscataway, NJ, USA, 2025; pp. 880–884.

35. Murzakhmetov, A.; Satymbekov, M.; Bapanov, A.; Beisov, N. Sentiment Analysis of Tourist Reviews About Kazakhstan Using a Hybrid Stacking Ensemble Approach. *Computation* **2025**, *13*, 240. [CrossRef]
36. Roumeliotis, K.I.; Tselikas, N.D.; Nasiopoulos, D.K. Leveraging Large Language Models in Tourism: A Comparative Study of the Latest GPT Omni Models and BERT NLP for Customer Review Classification and Sentiment Analysis. *Information* **2024**, *15*, 792. [CrossRef]
37. Šmíd, J.; Přibáň, P.; Král, P. LLaMA-Based Models for Aspect-Based Sentiment Analysis. *arXiv* **2025**, arXiv:2508.08649. [CrossRef]
38. Qwen Team. Qwen2.5-7B-Instruct. Hugging Face. Available online: https://huggingface.co/Qwen/Qwen2.5-7B-Instruct (accessed on 6 July 2026).
39. Dettmers, T.; Pagnoni, A.; Holtzman, A.; Zettlemoyer, L. QLoRA: Efficient Finetuning of Quantized LLMs. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 10088–10115. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.