Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Morphology-Aware Span-First Structured ABSA for Uzbek: Boundary Modeling and a Protocol for Cross-Lingual Turkic Transfer

Botir Elov[1]

elov@navoiy-uni.uz

Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Ruhillo Alaev[2]

Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Sultonbek Normamatov[1]

Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Alisher Abdullayev[3]

Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan

Elov Jamshid[4]

Tashkent University of Information Technology

Muhayyo Xakimova[1]

Alisher Navo'i Tashkent State University of Uzbek Language and Literature

Murotali Latipov[5]

Gulistan State University

Research Article

Keywords: Aspect-based sentiment analysis, span extraction, suffix features, boundary supervision, Turkic transfer

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11692685/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Morphology-Aware Span-First Structured ABSA for Uzbek: Boundary Modeling and a Protocol for Cross-Lingual Turkic Transfer

Botir Elov[1]*, Ruhillo Alaev[2], Sultonbek Normamatov[1], Alisher Abdullayev[3], Elov Jamshid[4], Muhayyo Xakimova[1], Murotali Latipov[5]

[1] Alisher Navoʻi Tashkent State University of Uzbek Language and Literature, Tashkent, Uzbekistan
[2] National University of Uzbekistan, Tashkent, Uzbekistan
[3] Center for Digital Education and Artificial Intelligence Development, Ministry of Higher Education, Science and Innovation of the Republic of Uzbekistan, Tashkent, Uzbekistan
[4] Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Tashkent, Uzbekistan
[5] Gulistan State University, Gulistan, Uzbekistan
*Correspondence: elov@navoiy-uni.uz

Abstract

Recovering aspect mentions is insufficient for structured sentiment analysis when their opinions, occurrences, or attributes are wrong. We study these errors in Uzbek educational texts using UzEDU-ABSA: 7,440 records, 5,683 texts, and 5,622 allocation groups. MorphoSpan-Uz augments XLM-R with a token-level byte convolution, eight transparent suffix features, and auxiliary endpoint supervision with span–endpoint consistency. It maps contiguous word spans to original Unicode offsets and predicts six-field tuples. We reassess eight archived systems across five seeds and independently reproduce their saved final-test metrics. On 744 records from 567 texts and 564 groups, mean candidate aspect exact F1 is 73.11 ± 14.12 and strict tuple F1 is 26.15 ± 8.16, versus 65.61 ± 18.79 and 19.09 ± 7.44 for XLM-R span. Scores are percentages and dispersions are sample standard deviations across seeds. Retrospective paired hierarchical bootstrap intervals for the improvements are [2.14, 14.49] and [4.35, 9.93] percentage points. Byte features and boundary supervision improve observed means; morphology gains are smaller and metric-dependent. Target-only low-rank adaptation (LoRA) produces no candidates, and constrained generators remain syntactically valid despite low tuple accuracy. Offset annotations have mixed review provenance: 525 records have documented human span confirmation. Conclusions are therefore conditional on annotation quality; prediction-level reproducibility does not establish checkpoint reproducibility. Turkish-to-Uzbek transfer is specified as a prospective protocol; no cross-lingual gain is claimed from the completed experiments.

Keywords Aspect-based sentiment analysis · span extraction · suffix features · boundary supervision · Turkic transfer

1 Introduction

Aspect-based sentiment analysis (ABSA) identifies what is evaluated and how it is evaluated. The SemEval tasks formalized aspect term extraction, category identification, and aspect-level sentiment prediction (Pontiki et al., 2014; Pontiki et al., 2016). More recent structured formulations recover related elements together. Their usefulness depends on recovering the correct text occurrence and relation, rather than recognizing a plausible aspect word somewhere in a sentence. A positive sentiment attached to the wrong occurrence, or to an opinion with an incomplete scope, is an incorrect structured prediction even when its individual labels look reasonable.

This distinction matters for Uzbek. Agglutination creates numerous surface forms from a lexical stem, and words may contain possessive, case, derivational, and verbal suffix sequences. Work on Uzbek Universal Dependencies provides linguistic resources for studying these structures (Akhundjanova & Talamo, 2025). In extraction, however, a morphological representation and an output boundary are separate design choices. A model can use suffix information while still producing complete orthographic-word spans. Conversely, a character-indexed output alone does not demonstrate that the model has learned morphology. We retain that distinction throughout the study.

The education domain adds a structured prediction problem. One text can evaluate instruction, workload, assessment, and institutional services together. A review expanded into several aspect records must be presented to a text-only extractor as one source with several target tuples. Supplying a gold aspect term in the input would change the task. Splitting sibling rows across training and evaluation would also change the generalization question, because the same wording could be observed in both partitions. Group-aware evaluation and explicit task definitions are consequently part of the modeling problem, consistent with broader concerns about leakage and random splits (Elangovan et al., 2021; Søgaard et al., 2021).

We investigate a compact implemented model rather than an unrestricted morphological parser. MorphoSpan-Uz combines contextual XLM-R states, a byte convolution inside each orthographic token, and a small suffix feature vector. A span classifier predicts aspect and opinion candidates; a separate relation module attaches four attributes to their linked

pairs. Endpoint supervision regularizes candidate scores. This design makes each component inspectable, but its output remains restricted to contiguous spans on a word and punctuation lattice. It does not predict arbitrary internal character endpoints.

Our empirical analysis uses the versioned UzEDU span release dated 20 October 2026. The fixed Mendeley v1 resource and the mutable Hugging Face distribution are related but distinct data objects (Elov & Alaev, 2026a; Elov & Alaev, 2026b). The experiments contain 7,440 retained records rather than the earlier 7,500-record description. We report the actual partition sizes, reference span provenance, and final predictions. We do not carry forward the historical exact-span value of 0.8585 as a score on this evaluation.

The study addresses four questions. First, do byte and suffix side channels improve exact extraction relative to a common XLM-R span architecture? Second, what is the observed effect of the bundled endpoint and consistency objective? Third, how much of candidate extraction performance survives in the final six-field output? Fourth, what do failed parameter-efficient adaptation and constrained generation reveal about the current operating regime? Turkish-to-Uzbek transfer is an additional research objective for which we specify a compatible follow-up protocol, rather than a completed fifth experiment.

The contribution is a transparent account of an implemented suffix- and byte-augmented extractor, an independently checked five-seed evaluation, and a diagnosis of the gap between span proposals and usable structured tuples. The data audit distinguishes raw texts from allocation groups and distinguishes offset validity from human verification. The transfer protocol specifies what could legitimately be shared across related languages without inventing missing source labels or attributing an unexecuted transfer result to the Uzbek-only LoRA run.

2 Related Work

2.1 Span and boundary models for structured ABSA

Span-ASTE represents complete candidate spans and their interactions, while Boundary-Driven Table-Filling models triplet structure through boundary-oriented table regions (Xu et al., 2021; Zhang et al., 2022). These methods motivate explicit expression boundaries and relation modeling. MorphoSpan-Uz shares that general objective but uses a simpler token-span classifier with four pair-conditioned attribute heads. We do not claim to reproduce either architecture, and neither appears among the eight completed comparison systems.

Generative ABSA provides a common output language for multiple subtasks. Unified sequence generation and paraphrase-based quadruple prediction illustrate this approach (Yan et al., 2021; Zhang et al., 2021). Our generative comparators instead emit a finite JSON grammar whose span values come from the input text. This eliminates many malformed or non-copying outputs by construction, but it does not make the selected offsets, tuple count, or attributes correct. Structural validity and extraction quality therefore require separate measurements.

2.2 Uzbek resources and representation choices

UzABSA studies aspect-level sentiment in Uzbek restaurant reviews (Matlatipov et al., 2024). LASQ is a published Uzbek–Uyghur quadruple extraction resource with a syntax-enriched grid model (Yusufu et al., 2026). These resources establish relevant Uzbek ABSA precedents. They use different domains and output inventories, so published scores cannot be inserted into an UzEDU leaderboard. In particular, a six-field tuple containing ordinal polarity and intensity has a different exactness requirement from a sentiment quadruple.

XLM-R supplies the multilingual contextual backbone in the completed discriminative experiments (Conneau et al., 2020). Uzbek-specific models such as BERTbek provide another relevant representation choice (Kuriyozov et al., 2024), but no comparable BERTbek prediction file is present in the final suite. It is discussed as related work rather than represented as an evaluated baseline.

Tokenization-free encoders and byte-to-byte generation reduce dependence on a fixed subword vocabulary. CANINE and ByT5 demonstrate distinct ways of processing lower-level input (Clark et al., 2022; Xue et al., 2022). Our byte side channel is narrower: a shallow convolution encodes the first 64 UTF-8 bytes of each token. It complements XLM-R rather than replacing it with a tokenization-free sentence encoder. The separate ByT5 comparator tests a byte-level sequence-to-sequence model under an offset grammar. mT5 supplies the corresponding multilingual subword generator (Xue et al., 2021).

2.3 Parameter efficiency and reliable comparison

LoRA adapts selected weight matrices through low-rank updates (Hu et al., 2022). Parameter efficiency alone does not imply cross-lingual transfer. A source-trained initialization and a matched target-only comparison are necessary to attribute

a gain to Turkish supervision. The present LoRA system is Uzbek-only; its failed decoding behavior does not resolve whether related-language adaptation would help.

Score distributions across random seeds can change the interpretation of model rankings (Reimers & Gurevych, 2017). Following general guidance on NLP significance testing (Dror et al., 2018), we report uncertainty at the source-family level to preserve dependence among sibling tuples. We report every final seed and identify newly computed intervals as retrospective sensitivity analyses. A positive average difference is not presented as preregistered confirmation or a universal improvement across domains and annotation regimes.

3 Dataset and Task Definition

3.1 Version and corpus lineage

The analyzed snapshot is uzedu-span-v4-20261020, included in the supplied public pipeline archive dated 22 October 2026. It retains 7,440 records and 5,683 distinct raw texts. The release metadata records ten excluded records and fifty pending-review records outside the core, accounting for the difference from the 7,500-record source description. No source-text or original-label changes are recorded for the retained core. This does not imply that every original string annotation already matched a recoverable occurrence; the derived alignment objects define the extraction targets.

Table 1 reports the actual partitions. The 5,622 groups are allocation components: 58 identified copy families join 119 distinct texts, reducing the number of components by 61 relative to the raw-text count. These units are not interchangeable. Predictions are produced for each raw text; group components constrain allocation and uncertainty estimation.

Table 1 Analyzed split and span review coverage

Unit	Training	Development	Test	Total
Records	5,953	743	744	7,440
Distinct raw texts	4,548	568	567	5,683
Allocation groups	4,493	565	564	5,622
Human-span-reviewed records	433	44	48	525
Multi-aspect texts	1,139	146	146	1,431

Record counts and review counts refer to rows. Multi-aspect counts refer to distinct raw texts; the bootstrap unit is the allocation group.

There are 1,431 multi-aspect texts, or 25.18% of the 5,683 sources. Specifically, 4,252 texts have one tuple, 1,105 have two, and 326 have three. Aspect spans average 1.65 regex word or punctuation units, with a maximum of five. Opinion spans average 4.96 units and reach fourteen. Multiword aspects account for 4,338 of 7,440 mentions. These distributions motivate complete-span prediction without implying that the retained collection represents the frequency of educational complaints in a population.

3.2 Reference spans and review provenance

Each aspect and opinion interval is half-open and indexes the original string by Unicode code points. All 14,880 retained intervals are valid, nonempty, and reproduce their stored alignment surface. Every core tuple has an explicit contiguous aspect and opinion; implicit and discontinuous extraction are outside the completed evaluation.

Mechanical validity is not equivalent to independent linguistic adjudication. The release records completed human span review for 525 rows: 513 rows accepted unchanged and twelve rows with corrected opinion boundaries. The remaining 6,915 rows retain the status ai_reviewed_proposal. Both spans are human-confirmed in 433 training rows, 44 development rows, and 48 test rows. No completed re-review of the original attribute labels is recorded in this version. We describe these as reference annotations and preserve their provenance rather than treating the complete offset layer as uniformly expert-adjudicated gold.

Original surface fields and aligned raw slices are also distinct. Across the core, 498 aspect strings and 191 opinion strings differ from their aligned raw surface. Some aspect differences disappear under normalization, but 217 aspect and 191 opinion differences remain. Extraction targets therefore come from the versioned alignment interval, not a later search for the original string. Among aligned surfaces, repeated occurrences are present in 153 aspect rows and two opinion rows; an occurrence is identified by its stored interval.

The fully human-span-reviewed test subset contains 43 complete sources and 45 rows. Three further sources contain only partly reviewed annotations. Restricting an auxiliary analysis to complete sources avoids treating unreviewed sibling tuples

as verified, but the reviewed subset was selected rather than randomly sampled. It cannot estimate the error rate of the entire offset layer or replace a representative independent audit.

3.3 Grouping and allocation checks

The grouping normalizer removes specified invisible formatting characters, applies NFC composition and case folding, maps supported apostrophe variants to U+02BB, and folds whitespace. It does not transliterate between scripts. Raw text remains the extraction coordinate system; normalized strings support grouping and the suffix side channel. Known copy families are allocated together.

Independent checks find zero shared record identifiers, raw-text hashes, normalized texts, or allocation-group identifiers for every partition pair. This verifies the declared exact and family grouping safeguards. It does not establish the absence of all paraphrases, unknown edited copies, author dependence, temporal dependence, or foundation-model pretraining exposure. We consequently use "group-disjoint test" rather than claiming universal leakage elimination.

3.4 Structured output

For a source text x, the target is a set of tuples

$$T(x) = \{(a, c, o, s, p, i)\}, \qquad a = (b_a, e_a), \quad o = (b_o, e_o). \tag{1}$$

Here a and o are explicit aspect and opinion intervals, c is one of 30 categories, s is POS, NEG, NEU, or MIX, p is an ordinal polarity class from −2 to 2, and i is low, medium, or strong intensity. Category, sentiment, numeric polarity, and intensity remain separate labels. We impose no undocumented deterministic mapping between them. Negation, sarcasm, context type, and script are available metadata but are not additional predicted fields in this six-field experiment. No gold aspect, category, or opinion string is supplied to the model input. Table 2 summarizes the attribute and script distributions.

Table 2 Attribute and script distributions in the retained release

Field	Value	Train	Dev	Test
Sentiment	NEG	2,046	255	255
Sentiment	POS	2,006	251	251
Sentiment	NEU	935	117	117
Sentiment	MIX	966	120	121
Polarity	-2	734	91	91
Polarity	-1	1,312	164	164
Polarity	0	1,901	237	238
Polarity	1	1,525	191	191
Polarity	2	481	60	60
Intensity	past	1,294	161	162
Intensity	o‘rta	3,414	427	427
Intensity	kuchli	1,245	155	155
Script	lotin	5,121	640	640
Script	kirill	555	69	69
Script	aralash	277	34	35

All denominators are records. Script is metadata rather than an output attribute. Intensity labels mean low, medium, and strong; script labels mean Latin, Cyrillic, and mixed.

4 MorphoSpan Uz

4.1 Word lattice and contextual representation

The implemented output lattice consists of regex orthographic tokens with internal supported apostrophes, together with standalone punctuation. Let token w_j occupy raw interval $[l_j, r_j)$. XLM-R produces contextual subword states; the token state s_j is the mean of states whose tokenizer intervals overlap that token. Only completely retained tokens enter the lattice.

A candidate token interval $[a, b)$ maps to raw interval $[l_a, r_{b-1})$. All candidates satisfy $1 \leq b - a \leq W$, with $W = 16$. Thus offsets are expressed in characters, while endpoints are constrained to token edges. The present model cannot output a stem-only substring inside a single token. This restriction is compatible with every retained reference span, but it limits conclusions about genuinely internal morphological boundary recovery.

The archived coverage records contain no truncated source and no unreachable reference span in any partition for the discriminative lattice. The test set has a mean of 205.55 candidate spans per text, with a range of 36–840. Candidate coverage is a property of the enumerated lattice; it should not be confused with recall after score thresholding and top-k pruning. Fig. 1 presents the implemented architecture.

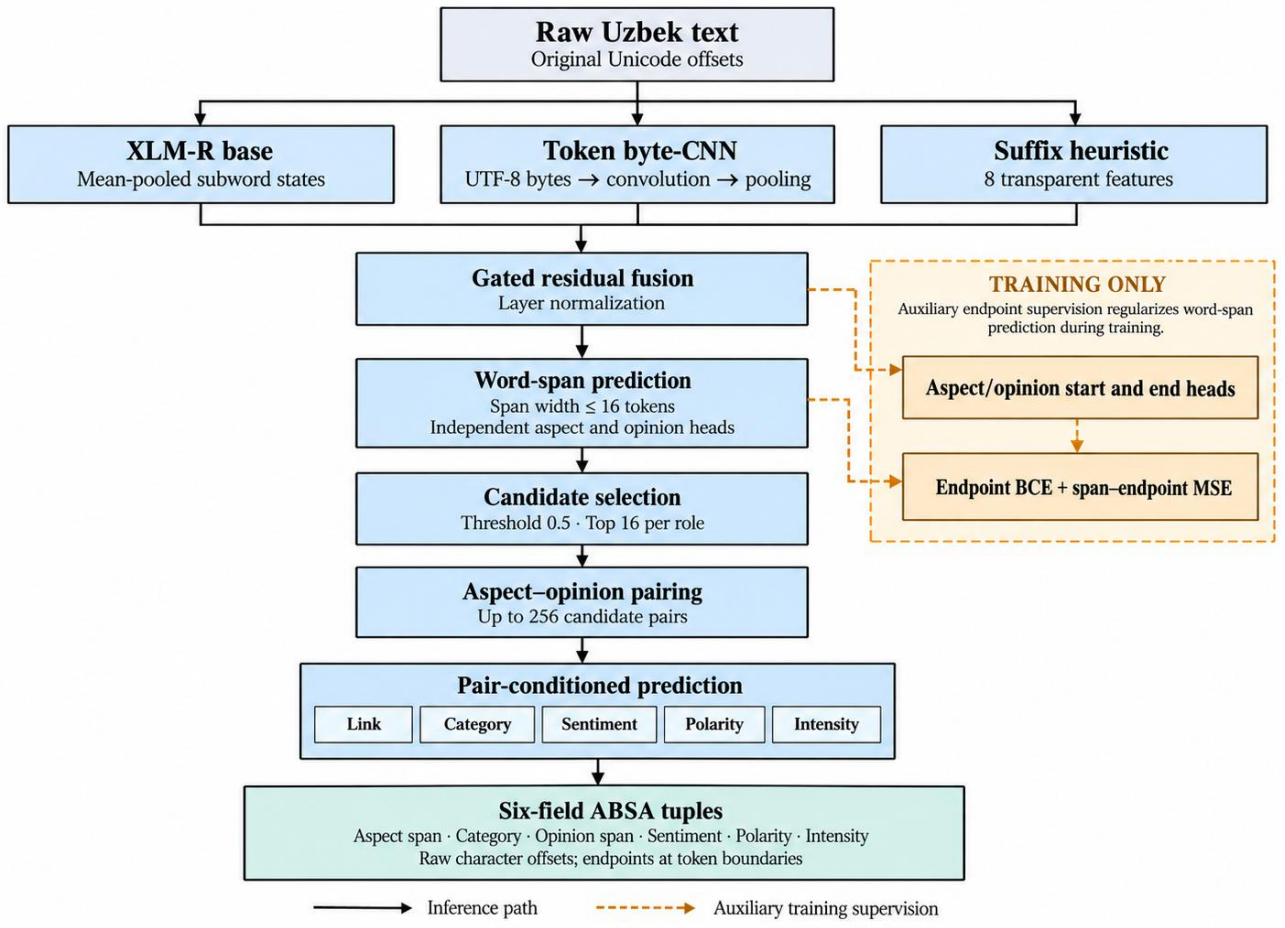


Fig. 1 Implemented MorphoSpan-Uz architecture. Auxiliary endpoint supervision regularizes word-span prediction during training. The model outputs raw character offsets restricted to token boundaries; Turkish adaptation is not part of this executed architecture

4.2 Byte and suffix side channels

The byte encoder embeds up to 64 UTF-8 bytes per token in 32 dimensions, using 257 entries including padding. A width-three convolution produces 64 channels, followed by GELU, max pooling, and projection to the XLM-R hidden size. The result c_j captures token-internal surface patterns. It is neither a full-comment character recurrent network nor an independently pretrained byte encoder.

The morphology channel uses a transparent heuristic rather than a validated morphological analyzer. On a normalized token, it repeatedly strips the longest eligible suffix, at most five times, while preserving a residual string of at least three characters. Six features count suffix classes: case, possessive, plural, derivation, negation, and tense. Two further features are $\min(d_j, 5)/5$ for chain depth and $\min(|w_j|, 30)/30$ for raw token length. The eight values are projected into m_j. The residual string is not asserted to be a correct lemma, and detected suffixes are not treated as expert morphological annotation. Table 3 lists the candidate suffix inventory.

Table 3 Implemented heuristic suffix inventory

Feature class	Candidate suffix strings
Case	ning, dan, ga, ka, qa, ni, da
Possessive	imiz, ingiz, lari, im, ing, si
Plural	lar
Derivation	chilik, lik, chi, siz, li
Negation	maydi, magan, mas, ma
Tense	moqda, yapti, gan, di

Longest matching strings are greedily removed, for at most five passes and with at least three residual characters. These strings are heuristic candidates, not validated morpheme analyses.

The Latin suffix inventory also has a script limitation. A Cyrillic word may receive length information while failing to match these suffix strings. The byte and multilingual channels remain available, but this does not make the suffix feature equally informative across scripts. We do not claim a calibrated morphological reliability gate or a script-robust analyzer.

One learned gate combines the projected views:

$$g_j = \sigma\big(W_g[s_j; c_j; m_j] + b_g\big), \qquad z_j = \mathrm{LayerNorm}\left(s_j + g_j \odot \big(c_j + m_j\big)\right). \tag{2}$$

An ablated side channel is set to zero. This shared gate controls the combined auxiliaries; it does not implement separate analyzer confidence or competing analysis probabilities.

4.3 Candidate spans and endpoint supervision

For a candidate $[a, b)$, the model forms

$$v_{ab} = \mathrm{MLP}\left[z_a; z_{b-1}; \tfrac{1}{b-a}\textstyle\sum_{j=a}^{b-1} z_j\,; \tfrac{b-a}{W}\right]. \tag{3}$$

The MLP is a linear projection, GELU, and dropout of 0.1. Two independent logits classify aspect and opinion membership. Their sigmoid probabilities q_{ab}^r allow overlapping candidates. A token-level head predicts separate start and end probabilities u_a^r and e_{b-1}^r for each role r.

Span labels use binary cross-entropy over enumerated candidate-role decisions. Endpoint labels use binary cross-entropy over token-role-boundary decisions. The consistency regularizer is

$$\mathcal{L}_{\mathrm{cons}} = \mathrm{mean}_{(a,b),r}(q_{ab}^r - u_a^r e_{b-1}^r)^2. \tag{4}$$

This is the implemented MSE agreement term. It is a regularizer against a product of endpoint predictions, not a noisy-OR boundary marginal or Jensen–Shannon divergence. Because several spans can share an endpoint, the product is not an exact marginalization of the multilabel span distribution. Removing boundary supervision disables both the endpoint loss and this consistency term; the corresponding ablation cannot isolate their individual effects.

4.4 Linking and attributes

For an aspect candidate a and opinion candidate o, their pair representation is

$$h_{ao} = \mathrm{MLP}[v_a; v_o; v_a \odot v_o]. \tag{5}$$

A sigmoid link head predicts whether the pair is related. Four softmax heads predict category, sentiment, ordinal polarity, and intensity from the same pair representation. Category is therefore pair-conditioned in this implementation. All positive relation pairs come from reference spans during training. Negative relations are bounded uniform samples of nonlinks, targeting at least eight or four times the positive count before deduplication. There is no morphology-conditioned hard-negative margin loss, and the linker is not trained on model-proposed positive spans.

The joint objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{span}} + 0.1(\mathcal{L}_{\mathrm{end}} + \mathcal{L}_{\mathrm{cons}}) + \mathcal{L}_{\mathrm{link}} + \tfrac{1}{4}\textstyle\sum_{t\in\{c,s,p,i\}} \mathcal{L}_t. \tag{6}$$

Each implemented loss uses its active-decision mean. Attribute cross-entropy is computed on reference positive pairs, and conflicting values for an identical pair are masked per attribute. No conflicting pair attributes occur in this retained release.

The unaugmented XLM-R baseline uses the same candidate and pair machinery with byte, suffix, and boundary components disabled.

4.5 Decoding and computational scope

At inference, aspect and opinion span probabilities are independently thresholded at 0.5. Up to sixteen candidates per role are retained by score. Their Cartesian product contains at most 256 pairs; link probabilities of at least 0.5 are emitted with argmax attributes. Endpoint probabilities regularize training but do not rescore candidates at decoding. Exact duplicate tuples are removed by set-based evaluation; overlapping expressions are allowed.

Enumerating width-bounded token spans requires $O(mW)$ candidates for m retained tokens, and linking requires $O(K_A K_O)$ pairs. These bounds describe enumeration and linking, not the complete transformer attention cost. The evaluated implementation has no NULL nodes or discontinuous segment representation.

5 Experimental Setup

5.1 Completed comparison systems

The completed suite contains eight systems and seeds 13, 29, 42, 71, and 101. Table 4 defines the actual conditions. The internal label no-char disables the token byte-CNN; we call it “without bytes” in the paper to identify the representation precisely. All six discriminative conditions share the span and relation formulation. The three component removals change one enabled feature group at a time, but they do not constitute a complete character-by-morphology factorial experiment with every interaction independently identified.

Table 4 Completed systems and enabled components

System	Byte CNN	Suffixes	Boundary loss	Encoder update
MorphoSpan-Uz	Yes	Yes	0.1	Full
XLM-R span	No	No	0	Full
Without bytes	No	Yes	0.1	Full
Without suffix features	Yes	No	0.1	Full
Without boundary objective	Yes	Yes	0	Full
Target-only LoRA	Yes	Yes	0.1	LoRA r = 8
mT5 constrained	—	—	—	Full seq2seq
ByT5 constrained	—	—	—	Full seq2seq

Each system has five final test runs. The LoRA condition is target-only; the generative systems use constrained JSON, with no separate discriminative candidate score.

The LoRA variant inserts rank-eight updates into XLM-R query and value projections, with alpha 16 and dropout 0.05, while freezing the base encoder through PEFT. No Turkish source checkpoint is specified. It is a target-only adaptation condition.

5.2 Constrained generative comparators

mT5-small and ByT5-small generate a JSON object containing a predicted tuple count and a tuple list. The fields are aspect_span, aspect_label, opinion_span, sentiment, polarity_score, and intensity. Span values are drawn from the same width-16 word lattice mapped to raw offsets. Label alternatives come from the training ontology. The maximum tuple count is three, derived from training data rather than supplied per test text.

Training and inference concatenate the same separately tokenized grammar segments. Prefix-constrained greedy decoding uses one beam without sampling. A malformed result is interpreted as an empty prediction; raw outputs are retained. Although ByT5 reads bytes, its allowed output span values remain word-lattice intervals. This comparison therefore evaluates two constrained generation implementations, not arbitrary character-boundary generation or the best attainable performance of either pretrained family.

5.3 Optimization and checkpoint selection

Table 5 describes the executed configuration. Discriminative training processes one source per forward pass and accumulates four steps. The configuration file contains a nominal batch size of eight, but that field is not consumed by the

span trainer; the effective source batch is four, except at the final partial update. Generative training uses batches of two with four-step accumulation. We report the executed behavior rather than multiplying an unused configuration value.

Table 5 Executed optimization and input settings

Setting	Span systems	mT5-small	ByT5-small
Optimizer	AdamW	AdamW	AdamW
Learning rate	0.00002	0.0001	0.0001
Weight decay	0.01	0.01	0.01
Maximum epochs / patience	5 / 2	5 / 2	5 / 2
Source minibatch × accumulation	1 × 4	2 × 4	2 × 4
Input token limit	256 subwords	256 subwords	1,024 byte tokens
Maximum output span width	16 tokens	16 tokens	16 tokens
Selection metric	Mean candidate A/O F1	Strict tuple F1	Strict tuple F1
Gradient clipping	1.0	1.0	1.0
Precision on CUDA	BF16 autocast	BF16 autocast	BF16 autocast
Generation / span decoding	0.5; top 16 per role	Greedy grammar	Greedy grammar

A/O denotes aspect/opinion. Gradient checkpointing is enabled. No learning-rate warm-up or scheduler is implemented in this trainer.

Discriminative checkpoints maximize the mean of development candidate aspect and opinion exact F1. Generative checkpoints maximize development strict tuple F1. The different selection criteria are a cross-family comparison limitation. No executed 20-configuration search or thirty-epoch budget is asserted. The principal full model and span baseline each have five logged epochs for all five seeds; some failed or generative runs stop earlier under patience two.

The final evaluations dated 22 October 2026 reuse the development-selected v4 checkpoints and run no training. The audit underlying this manuscript recomputes scores from saved predictions; it does not regenerate predictions or fit new models. Compatibility records document migration to the public package. Some legacy data_hashes.evaluation fields retain a development hash, so test identity was checked independently: all 40 files contain exactly the 567 final-test source hashes and their expected reference tuple sets. The frozen partition hashes are listed in Appendix C.

5.4 Metrics and denominators

For a source d, let G_d and P_d denote reference and predicted tuple sets. Strict tuple counts are

$$\mathrm{TP} = \sum_d |G_d \cap P_d|, \qquad F_1 = \frac{2\mathrm{TP}}{\sum_d |G_d| + \sum_d |P_d|}. \tag{7}$$

An exact tuple requires both intervals and all four labels to agree. Exact span metrics remove labels and deduplicate intervals within each source before micro-aggregation. We distinguish two quantities. Candidate extraction scores the retained pre-link span lists of discriminative systems. Final-output span scores use only intervals appearing in emitted tuples and are available for all systems. The common cross-family table uses final-output spans and tuple F1; candidate scores appear separately.

Overlap diagnostics use one-to-one interval matching, maximizing match cardinality before summed character intersection-over-union. Mean boundary distance is measured only for overlapping matched pairs, with unmatched counts retained. A relaxed score cannot replace exact extraction. Oracle-span analysis supplies reference candidate spans to the relation module while still predicting links and attributes. It changes the input condition and is a diagnostic rather than the text-only task.

5.5 Variability and retrospective uncertainty

We report the arithmetic mean and sample standard deviation of five independently scored seed runs. Standard deviation describes seed dispersion; it is not a confidence interval. Differences between the full model and four span comparators additionally receive 10,000 paired hierarchical percentile-bootstrap replicates with random seed 1729. Each replicate samples the five seed indices with replacement and samples all 564 test allocation groups with replacement. All source predictions in a selected group remain together, and the compared models receive identical draws.

For each replicate, micro-F1 is computed separately for each sampled seed and then averaged. The interval is the 2.5th–97.5th percentile of the paired mean differences. This retains the paper's mean-seed estimand rather than pooling

predictions from different fitted models into one artificial system. The intervals were computed after inspection of the supplied results, so they are exploratory sensitivity estimates. We make no preregistered significance claim, and do not infer that five seeds characterize every possible initialization.

6 Results

6.1 Final structured outputs

Table 6 presents the common final-output metrics. MorphoSpan-Uz has the highest observed mean tuple F1, 26.15, compared with 19.09 for the XLM-R span baseline. Its final-output aspect and opinion exact F1 values are 56.08 and 48.55. These are lower than its candidate extraction scores because spans that never appear in an emitted tuple are absent from the final-output inventory. In the archived outputs, all scored pairs pass the link threshold; candidate spans therefore disappear when the opposite-role candidate inventory is empty, rather than through observed link rejection.

Table 6 Common final-output exact F1 across five seeds

System	Aspect F1	Opinion F1	Strict tuple F1
MorphoSpan-Uz	56.08 ± 14.27	48.55 ± 10.19	26.15 ± 8.16
XLM-R span	42.37 ± 13.38	34.68 ± 9.43	19.09 ± 7.44
Without bytes	41.08 ± 21.72	35.39 ± 17.05	19.42 ± 9.71
Without suffix features	47.95 ± 23.21	40.94 ± 17.27	22.74 ± 9.90
Without boundary objective	37.12 ± 15.40	31.58 ± 12.07	16.70 ± 7.29
Target-only LoRA	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
mT5 constrained	18.67 ± 8.23	3.83 ± 0.91	0.99 ± 0.58
ByT5 constrained	22.36 ± 13.70	6.49 ± 3.79	2.16 ± 2.62

Scores are percentages, reported as mean ± sample SD. Aspect and opinion intervals are counted only when they occur in a final emitted tuple. All systems score 567 source texts and 744 reference tuples.

The absolute tuple difference from XLM-R is 7.06 points. Its retrospective paired interval is [4.35, 9.93]. This supports an advantage within these fitted runs and this reference release, while substantial seed variability and annotation uncertainty limit generalization. The full model's tuple F1 ranges from 15.37 to 33.09 across seeds; reporting only the strongest run would materially overstate typical performance. Complete seed values appear in Appendix B. Fig. 2 shows the individual seed results and their means.

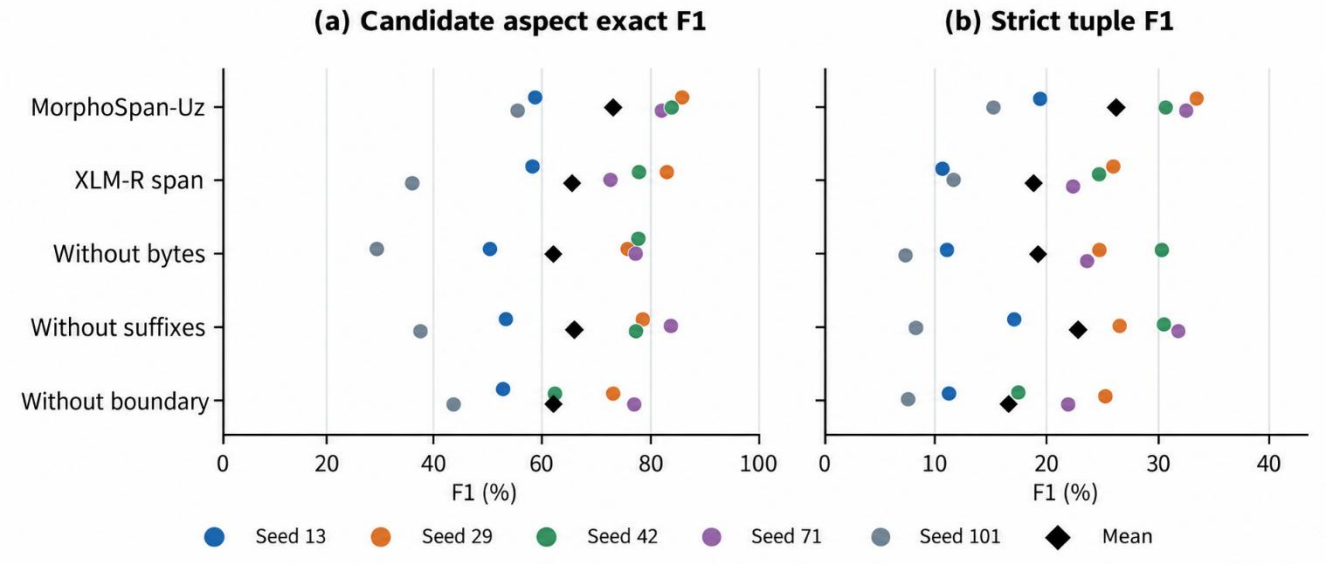


Fig. 2 Five-seed test results. Colored points identify seeds; black diamonds show means. Left: candidate aspect exact F1. Right: strict tuple F1. Horizontal scales differ

6.2 Candidate extraction and component removals

The full model's candidate aspect exact F1 is 73.11, versus 65.61 for XLM-R, a difference of 7.50 points with a retrospective interval of [2.14, 14.49]. Candidate opinion F1 improves from 42.27 to 52.44. Table 7 separates these proposal-stage results from the common final-output table.

Table 7 Discriminative candidate extraction exact F1

System	Aspect F1	Opinion F1
MorphoSpan-Uz	73.11 ± 14.12	52.44 ± 7.34
XLM-R span	65.61 ± 18.79	42.27 ± 7.79
Without bytes	62.27 ± 21.42	44.31 ± 11.64
Without suffix features	66.07 ± 19.57	49.47 ± 11.51
Without boundary objective	61.73 ± 13.84	37.87 ± 8.75
Target-only LoRA	0.00 ± 0.00	0.00 ± 0.00

These metrics use the retained candidates before linking. They are not directly comparable with generative final-output span scores.

Removing bytes reduces mean candidate aspect F1 by 10.84 points and tuple F1 by 6.74 points. Removing suffix features reduces them by 7.04 and 3.41 points. Removing the bundled boundary objective yields the largest tuple reduction among these removals, 9.45 points, together with an 11.38-point candidate aspect reduction. The comparisons support the combined operating design, but do not identify a causal contribution of linguistically correct morphology: the suffix inventory is heuristic, and side-channel capacity and optimization may contribute. Table 8 reports the paired differences and uncertainty intervals.

Table 8 Full-model differences and retrospective 95 percent intervals

Comparator	Candidate aspect Δ [CI]	Strict tuple Δ [CI]
XLM-R span	7.50 [2.14, 14.49]	7.06 [4.35, 9.93]
Without bytes	10.84 [5.39, 19.09]	6.74 [2.96, 9.87]
Without suffix features	7.04 [1.78, 13.30]	3.41 [0.33, 6.62]
Without boundary objective	11.38 [6.40, 17.06]	9.45 [6.94, 12.33]

Differences are percentage points. Intervals use 10,000 paired hierarchical bootstrap replicates over five seeds and 564 complete allocation groups. They are post hoc sensitivity estimates, without confirmatory multiplicity claims.

The morphology effect is metric-dependent. Candidate opinion F1 differs by only 2.97 points between the full model and the model without suffix features, with an interval of [−1.66, 8.46]. We therefore do not conclude that suffix features reliably improve every output. Likewise, the boundary ablation removes endpoint BCE and consistency together; it cannot show that MSE consistency alone is responsible for the gain.

6.3 Proposal quality and the downstream bottleneck

Supplying reference spans raises the full model's tuple F1 to 59.49 ± 1.49, compared with 26.15 ± 8.16 under predicted candidates. The XLM-R baseline reaches a similar oracle-span tuple score, 58.97 ± 1.20, despite a substantially lower predicted-span score. The auxiliary views therefore appear more consequential for the prediction path leading to usable candidates than for performance under supplied reference boundaries. Table 9 compares predicted and reference candidates.

The oracle experiment still makes attribute errors, and its near-complete final span recovery does not imply correct tuple structure. Furthermore, candidate and linked-span F1 differences are not additive error shares: they use different prediction inventories and denominators. The 33.34-point full-model oracle–predicted tuple gap is a diagnostic contrast, not a causal estimate of the improvement obtained by replacing only one component in deployment.

Table 9 Strict tuple F1 under predicted and reference candidate spans

System	Predicted candidates	Oracle candidates
MorphoSpan-Uz	26.15 ± 8.16	59.49 ± 1.49
XLM-R span	19.09 ± 7.44	58.97 ± 1.20
Without bytes	19.42 ± 9.71	58.79 ± 1.12
Without suffix features	22.74 ± 9.90	59.70 ± 1.31
Without boundary objective	16.70 ± 7.29	58.50 ± 1.26

Oracle candidates supply reference aspect and opinion spans. Relations and attributes are still predicted. The oracle setting is not text-only end-to-end extraction.

In all five full-model runs, every retained aspect–opinion candidate pair is emitted; the same equality holds for the other discriminative conditions. The trained link threshold therefore performs no observed pair pruning on these outputs. Ignoring attributes, the full model has mean pair precision 34.09%, recall 33.63%, and F1 33.67%. With oracle candidate spans, mean pair precision is 64.12% and recall 99.33%, because cross-pair emissions remain. Near-perfect recovery of marginal oracle spans consequently coexists with incorrect relations. This is a direct output-level diagnosis, rather than an inference from the span–tuple F1 gap alone.

6.4 Target only LoRA and constrained generation

All five target-only LoRA runs produce zero aspect candidates, zero opinion candidates, and zero tuples at the fixed decoding threshold. Their development extraction F1 also remains zero, although training loss decreases. This is an unsuccessful operating point under the current loss balance, optimization budget, and threshold. The available artifacts do not distinguish under-training, score calibration, adaptation setup, or another mechanism. Lowering the threshold on the examined test set would be a new exploratory analysis, not a repair of the original result. No conclusion about Turkish transfer or LoRA in general follows from this collapse.

The generative systems produce valid JSON for all 567 test sources in every run. Nevertheless, mT5 and ByT5 reach mean strict tuple F1 of only 0.99 and 2.16. None of their final source outputs is empty, so low accuracy is not explained by parse failure. The finite grammar succeeds at serialization while the model often selects incorrect offsets or labels. Given the short shared epoch budget, different checkpoint criterion, and unoptimized greedy offset serialization, these results should not be generalized to all generative ABSA.

6.5 Review subset and qualitative errors

On the 43 completely human-span-reviewed test sources, the full model has candidate aspect F1 of 54.66 ± 9.90 and strict tuple F1 of 19.46 ± 8.68. Its full-test candidate aspect score is higher, while the candidate opinion score on the reviewed subset is 53.05 ± 7.76, close to the overall mean. Selection of reviewed cases and small support make this a sensitivity result, not evidence that the remaining reference annotations are accurate or inaccurate. Attribute labels in this subset retain the original provenance as well.

Three observed examples clarify the error types. The first contains repeated “ta‘lim sifati” occurrences and a shortened opinion; the second recovers both intervals but confuses positive intensity and ordinal polarity; the third retains the correct aspect but lacks the opinion candidate required to emit a tuple. Table 10 reports the original offset coordinates and selected output. Each illustrated reference tuple has both spans human-confirmed; sibling tuples in the same source may have different review status. The examples are illustrative rather than a prevalence estimate or a new expert error taxonomy.

Table 10 Illustrative errors on sources with human-confirmed spans

Case	Reference	Selected prediction	Error
E1; seed 42	A [72,85); O [72,125)	A [22,35) and [72,85); O [95,125)	Repeated aspect occurrence and shortened opinion
E2; seed 71	A [13,20); O [45,83); POS / 2 / kuchli	Same intervals; POS / 1 / o‘rta	Ordinal polarity and intensity differ
E3; seed 71	A [0,9); O [10,59)	A candidate [0,9); no opinion candidate; no tuple	Missing opinion proposal

A/O denotes aspect/opinion; intervals are half-open raw code-point offsets. Span confirmation does not imply a new independent review of attribute labels.

E1 source. “Bugungi darsdan keyin ta‘lim sifati bilan bog‘liq tartib tushuntirildi: ta‘lim sifati bo‘yicha tahlil ishlari olib borilmoqda.” Source ID prefix c4a9f206809e. The reference opinion includes the second occurrence of “ta‘lim sifati” and its predicate; the selected opinion begins only at “tahlil ishlari”.

E2 source. “online guruh mezonni yuqori baholadi, chunki natijalar shaffof tarzda e‘lon qilindi. Men uchun bu katta yordam bo‘ldi.” Source ID prefix fcf6cbe8724d. Both intervals are correct; the displayed attribute mismatch still prevents exact tuple credit.

E3 source. “vazifalar mavzuga mos; shunga qaramay, murakkabligi notekis.” Source ID prefix e978697a230a. The missing opinion candidate makes tuple emission impossible in this pipeline.

6.6 Training behavior and reproducibility

The five full-model development trajectories are heterogeneous. Three reach a mean aspect/opinion candidate F1 near 0.69–0.70 at epoch five, while two end near 0.48–0.51. XLM-R exhibits a related seed pattern. This suggests that initialization and short-budget optimization are material to the comparison. Five epochs do not establish convergence, and declining aggregate loss is insufficient evidence of useful decoded extraction, as the LoRA condition demonstrates. Fig. 3 displays all five learning trajectories.

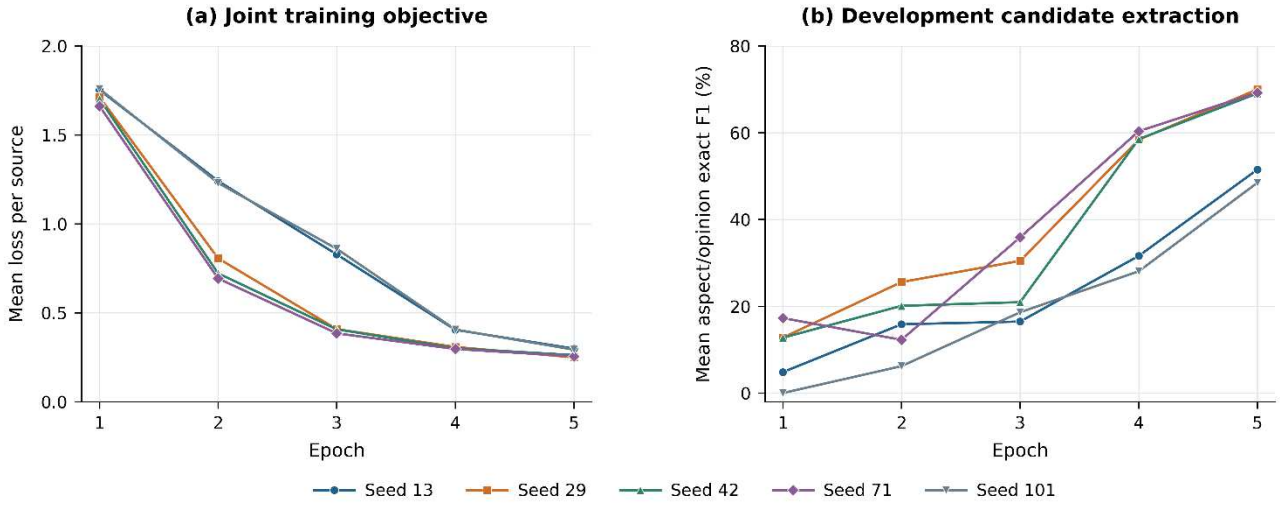


Fig. 3 Archived MorphoSpan-Uz learning curves for all five seeds. Left: mean joint training loss per source. Right: mean development aspect/opinion candidate exact F1

All stored exact tuple, final-output span, and candidate exact metrics were independently reproduced from the 40 prediction files with zero discrepancies. Every source identifier and reference tuple set was also matched against the final test partition. This validates prediction-level arithmetic and test identity. The shared archive omits model weights, so it does not by itself permit checkpoint-level regeneration. Final evaluation timings were recorded on shared accelerator resources and contain no training; we do not use them to rank isolated training or inference efficiency.

7 Protocol for Cross Lingual Turkic Transfer

Cross-lingual transfer is prospective. The completed suite contains neither Turkish-trained checkpoints nor matched Turkish-to-Uzbek runs. The existing loader supports Turkish M-ABSA records with aspect–category–sentiment triples (Wu et al., 2025), but the existence of a loader is not evidence of source-language training. M-ABSA’s translated and reviewed provenance also means that such an experiment would measure transfer from that resource, rather than from an independently collected Turkish educational corpus.

Before source optimization, the resource revision, original partitions, permitted uses, text hashes, and conversion output must be frozen. Only official training records may enter source training; the existing directory-wide loader must explicitly filter partitions. Eligible targets are explicit contiguous aspects with an auditable raw occurrence. Repeated strings require occurrence resolution, and missing or implicit expressions cannot be assigned arbitrary intervals. Source duplicate groups and any overlap with Uzbek held-out texts must be checked. A SemEval Turkish source with explicit target offsets is an alternative, but it needs its own parser, licence review, and versioned conversion rather than being described as the currently executed loader.

The implemented aspect-only branch supports a restricted source objective: aspect-span BCE, aspect endpoint BCE, and aspect span–endpoint consistency. Missing opinion, intensity, and ordinal-polarity labels receive no source loss. Source category and sentiment inventories remain private. The Uzbek suffix heuristic is disabled during Turkish adaptation because its inventory is not validated for Turkish. Source representability and truncation must be measured on the same token lattice.

A minimal matched experiment would train rank-eight query/value LoRA updates in the frozen XLM-R encoder on Turkish aspect supervision. The Uzbek stage would initialize only these encoder updates, retain the same adapter parameterization, and continue optimization on Uzbek training groups. Uzbek byte, suffix, span, endpoint, link, and attribute modules would receive identical initializations in transferred and target-only controls. An explicit transferred-tensor manifest is required: unrestricted shape-compatible loading could otherwise copy opinion or relation modules that received no source supervision. This sequential adapter protocol does not assume a learned language gate, a frozen Turkish residual, or two composable adapters.

The primary contrast would compare sequential Turkish-to-Uzbek LoRA with Uzbek-only LoRA at equal target data, seeds, active parameter count, target stopping rule, and selection criterion. A target-only continuation with matched additional updates would control for optimization budget. Given the present failed LoRA operating point, the target-only system must first show nonzero development extraction under an independently specified training protocol. Both beneficial and negative transfer must be reported using exact spans, tuple F1, and paired group uncertainty.

The current test results have already been examined. A later confirmatory transfer study therefore requires new held-out texts or a declared nested grouped evaluation. Reusing this test can support an additional exploratory result, but cannot retrospectively create an untouched transfer benchmark. The protocol remains a reproducible research direction rather than an empirical contribution of the 40 completed evaluations.

8 Discussion

8.1 What the observed improvements establish

Within the shared word-span architecture, the full configuration improves mean candidate extraction and final tuple recovery relative to the XLM-R baseline. Byte augmentation and boundary supervision show larger mean tuple differences than suffix removal. This supports using low-level surface features and explicit boundary supervision in the current Uzbek setting. It does not establish that a morphological analyzer, arbitrary character-boundary model, or cross-lingual component has been validated, because none is the component that produced these scores.

The word lattice also changes the scientific interpretation of "boundary modeling." It learns which complete tokens belong to an aspect or opinion and whether an adjacent modifier is included. It cannot directly evaluate dropping a case or possessive suffix within an orthographic token. An experiment about internal stem–suffix boundaries would need annotations and a model output lattice that permit those distinctions. The present evidence supports word-span selection aided by suffix features.

8.2 Why complete tuples remain difficult

Exact tuple scoring compounds several opportunities for error. Both boundaries must agree, the relation must be emitted, and four attributes must match. A valid-looking opinion may still have a shorter scope than the reference. Correct spans can still fail on medium versus strong intensity or on the ordinal polarity class. An exact six-field score therefore cannot be replaced by an average of attribute F1 values or by a high overlap score.

The relation module is trained with reference positive spans and uniformly sampled nonlinks, whereas inference uses scored candidates with boundary errors. This training–inference difference may reduce its ability to handle plausible wrong candidates. The archive's relation diagnostics should guide experiments with model-proposed positives and difficult near misses, but those changes require new training. Oracle-span outputs identify a conditional performance regime; they do not constitute a corrected end-to-end result.

8.3 Annotation quality and benchmark interpretation

Offset correctness, linguistic correctness, and attribute correctness are distinct. Exact substring validation rules out malformed coordinates. Human review status documents which spans received a recorded check. Neither proves the consistency of all categories or sentiment scales. The large unreviewed portion of the offset layer makes a representative independent annotation audit the most important next data step.

The reviewed test subset illustrates sensitivity to review selection. Its lower aspect score cannot be used to estimate how much the main score is biased, because reviewed and unreviewed examples were not randomly assigned. Future annotation

work should sample independent source groups, use at least two annotators, record occurrence-level decisions and alternative valid boundaries, and evaluate agreement before adjudication. Those new measurements should be reported when performed rather than supplied retrospectively as unsupported agreement coefficients.

8.4 Practical implications

The current system is a research extractor for structured analysis, not a validated basis for individual student or teacher assessment. Aggregated educational insights should retain the source context, express uncertainty, and allow review of span and attribute errors. The low tuple accuracy and seed instability make automated consequential decisions inappropriate. The most actionable engineering targets are stable optimization, representative reference review, opinion-scope recovery, and relation/attribute diagnostics. Cross-lingual expansion becomes informative only after a matched target-only adaptation condition is established.

9 Limitations

The study uses one educational-domain collection and one grouped partition. It does not measure out-of-institution, temporal, or independently collected domain transfer. Known duplicate and copy-family separation is verified, but broader semantic, author, and pretraining overlap remain unmeasured. There is no independent record demonstrating prospective test custody before all earlier model development, so the test should not be described as a sealed confirmatory benchmark.

The reference layer is mixed in review provenance, and attribute labels were not re-reviewed in the v4 correction workflow. The completely reviewed source subset is small and selected. No inter-annotator agreement estimate or fully independent morphology annotation is available. Consequently, extraction improvements are conditional on this reference layer and cannot establish expert-level linguistic correctness.

The suffix inventory is heuristic and Latin-oriented. The model predicts complete token spans rather than arbitrary character endpoints, and the retained corpus contains no implicit or discontinuous targets. All lattice targets are reachable, but performance outside that retained regime is untested. The boundary ablation bundles two losses; component capacity, optimization, and feature information are not fully separated by the executed comparisons.

Training is limited to five epochs, with five seeds and no documented equal-budget search over model families. Discriminative and generative checkpoint criteria differ. Generative results are sensitive to numeric-offset serialization and greedy decoding. The LoRA run fails before useful extraction and provides no general negative result about parameter-efficient learning. No Turkish transfer experiment is completed.

Finally, inference artifacts permit metric recomputation but not weight-based regeneration because checkpoint tensors are absent. Shared-device timings do not establish comparative efficiency. The newly computed uncertainty intervals are retrospective and do not include uncertainty about annotation decisions, corpus collection, or alternative split assignments.

10 Conclusion

MorphoSpan-Uz combines XLM-R, token-internal byte features, suffix heuristics, and auxiliary boundary supervision for Uzbek six-field ABSA. Across five seeds on the group-disjoint v4 test partition, it obtains candidate aspect exact F1 of 73.11 and strict tuple F1 of 26.15, compared with 65.61 and 19.09 for the common XLM-R span baseline. Byte and boundary component removals reduce the observed means, while the suffix contribution is smaller and depends on the metric.

The main remaining difficulty is converting candidate spans into reliable complete tuples. Oracle-span performance, seed dispersion, failed target-only LoRA decoding, and valid but inaccurate constrained generation show why boundary F1 alone is insufficient. The results are reproducible from saved predictions and explicitly conditional on a partially human-reviewed offset layer. A representative annotation audit, more stable adaptation, and a new held-out evaluation are necessary to establish stronger benchmark and transfer claims. The proposed Turkish-to-Uzbek protocol defines that next experiment without presenting it as completed evidence.

Statements and Declarations

Data availability. The source resource is available from Mendeley Data, DOI 10.17632/hwv5wzpvpk.1, and the Hugging Face distribution, DOI 10.57967/hf/9275. The analyzed snapshot is uzedu-span-v4-20261020; partition and manifest checksums identify it independently of a mutable dataset page. The dataset record lists CC BY 4.0. Review flags must accompany reused offset annotations, and redistribution must respect the applicable source terms.

Code and reproducibility. The study's uzedu_public_pipeline_20261022 archive contains implementation code, pinned model revisions, dependency specifications, data manifests, development logs, 40 final prediction files, evaluation configurations, and

metric summaries. It omits trained weights. Appendix C identifies the artifacts supporting prediction-level reproduction; no permanent public code DOI is asserted here. The numerical record accompanying Figs. 2 and 3 is supplied in Online Resource 1.

Ethics and responsible use. This work analyzes an existing educational-domain resource and does not introduce new recruitment or participant intervention. The manuscript reports text-level prediction behavior rather than identifying individual reviewers. Public availability does not establish consent for every downstream use or substitute for an institutional ethics determination. Collection consent, original anonymization procedures, and the applicable review or exemption record require documentary confirmation before submission. Results should not be used for automated individual educational decisions.

Author contributions.

Botir Elov: Conceptualization, Methodology, Supervision, Project administration, Writing – original draft, Writing – review & editing.
Ruhillo Alaev: Data curation, Investigation, Software, Resources, Validation, Software, Visualization, Writing – original draft, Writing – review & editing.

Sultonbek Normamatov: Resources, Validation, Visualization

Alisher Abdullayev: Methodology, Formal analysis, Resources

Elov Jamshid: Methodology, Formal analysis, Software, Resources, Validation

Muhayyo Xakimova: Resources, Validation, Investigation, Writing – review & editing

Murotali Latipov: Methodology, Formal analysis, Resources, Visualization

Funding. This research did not receive any specific grant from public, commercial, or not-for-profit funding agencies.

Competing interests. The authors declare no known competing financial interests or personal relationships that could influence this work.

AI assistance. ChatGPT and Codex assisted with manuscript organization, language revision, code inspection, and computation of retrospective evaluation diagnostics. The offset release separately records AI span proposals and selected human review. Manuscript assistance and annotation provenance are distinct. The authors retain responsibility for verifying the manuscript, references, data governance, and reported analyses.

References

Akhundjanova, A., & Talamo, L. (2025). Universal Dependencies treebank for Uzbek. In *Proceedings of RESOURCEFUL 2025* (pp. 1–6). University of Tartu Library. https://aclanthology.org/2025.resourceful-1.1/

Clark, J. H., Garrette, D., Turc, I., & Wieting, J. (2022). Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, *10*, 73–91. https://doi.org/10.1162/tacl_a_00448

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., et al. (2020). Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL 2020* (pp. 8440–8451). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.747

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of ACL 2018* (Vol. 1, pp. 1383–1392). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1128

Elangovan, A., He, J., & Verspoor, K. (2021). Memorization vs. generalization: Quantifying data leakage in NLP performance evaluation. In *Proceedings of EACL 2021* (pp. 1325–1335). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.113

Elov, B., & Alaev, R. (2026a). *uzEDU-ABSA* [Data set]. Hugging Face. Retrieved October 23, 2026, from https://huggingface.co/datasets/uznlp-uz/uzEDU-ABSA. https://doi.org/10.57967/hf/9275

Elov, B., & Alaev, R. (2026b). *UzEDU-ABSA: An Uzbek educational domain dataset for aspect-based sentiment analysis* (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/hwv5wzpvpk.1

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., et al. (2022). LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

Kuriyozov, E., Vilares, D., & Gómez-Rodríguez, C. (2024). BERTbek: A pretrained language model for Uzbek. In *Proceedings of SIGUL at LREC-COLING 2024* (pp. 33–44). ELRA and ICCL. https://aclanthology.org/2024.sigul-1.5/

Matlatipov, S. G., Rajabov, J., Kuriyozov, E., & Aripov, M. (2024). UzABSA: Aspect-based sentiment analysis for the Uzbek language. In *Proceedings of SIGUL at LREC-COLING 2024* (pp. 394–403). ELRA and ICCL. https://aclanthology.org/2024.sigul-1.47/

Pontiki, M., Galanis, D., Papageorgiou, H., Androutsopoulos, I., Manandhar, S., Al-Smadi, M., et al. (2016). SemEval-2016 Task 5: Aspect based sentiment analysis. In *Proceedings of SemEval 2016* (pp. 19–30). Association for Computational Linguistics. https://doi.org/10.18653/v1/S16-1002

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect based sentiment analysis. In *Proceedings of SemEval 2014* (pp. 27–35). Association for Computational Linguistics. https://doi.org/10.3115/v1/S14-2004

Reimers, N., & Gurevych, I. (2017). Reporting score distributions makes a difference: Performance study of LSTM-networks for sequence tagging. In *Proceedings of EMNLP 2017* (pp. 338–348). Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1035

Søgaard, A., Ebert, S., Bastings, J., & Filippova, K. (2021). We need to talk about random splits. In *Proceedings of EACL 2021* (pp. 1823–1832). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.156

Wu, C. Y., Ma, B., Liu, Y., Zhang, Z., Deng, N., Li, Y., Chen, B., Zhang, Y., Xue, Y., & Plank, B. (2025). M-ABSA: A multilingual dataset for aspect-based sentiment analysis. In *Proceedings of EMNLP 2025* (pp. 2530–2557). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.128

Xu, L., Chia, Y. K., & Bing, L. (2021). Learning span-level interactions for aspect sentiment triplet extraction. In *Proceedings of ACL-IJCNLP 2021* (Vol. 1, pp. 4755–4766). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.367

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., et al. (2022). ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, *10*, 291–306. https://doi.org/10.1162/tacl_a_00461

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., et al. (2021). mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of NAACL-HLT 2021* (pp. 483–498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.41

Yan, H., Dai, J., Ji, T., Qiu, X., & Zhang, Z. (2021). A unified generative framework for aspect-based sentiment analysis. In *Proceedings of ACL-IJCNLP 2021* (Vol. 1, pp. 2416–2429). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.188

Yusufu, A., Liu, J., Aziz, K., Ainiwaer, A., Li, B., Li, F., et al. (2026). LASQ: A low-resource aspect-based sentiment quadruple extraction dataset. *Journal of King Saud University Computer and Information Sciences*, *38*, Article 337. https://doi.org/10.1007/s44443-026-00701-x

Zhang, W., Deng, Y., Li, X., Yuan, Y., Bing, L., & Lam, W. (2021). Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of EMNLP 2021* (pp. 9209–9219). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.726

Zhang, Y., Yang, Y., Li, Y., Liang, B., Chen, S., Dang, Y., et al. (2022). Boundary-driven table-filling for aspect sentiment triplet extraction. In *Proceedings of EMNLP 2022* (pp. 6485–6498). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.435

Appendix A Category Inventory

Table 11 preserves the saved category vocabulary and row denominators. All thirty categories occur in every partition. Counts describe the analyzed release and are not population prevalence estimates.

Table 11 Category vocabulary and record counts

Category	Train	Dev	Test	Total
TEACHER COMPETENCE	203	26	25	254
TEACHING CLARITY	203	25	25	253
TEACHING METHOD	204	25	26	255
TEACHER ATTITUDE	197	25	24	246
COURSE CONTENT	204	25	26	255
CURRICULUM	204	25	25	254
PRACTICAL TRAINING	203	25	25	253
ASSESSMENT FAIRNESS	203	25	25	253
EXAM PROCESS	203	25	26	254
FEEDBACK	202	25	25	252
ASSIGNMENT WORKLOAD	202	26	25	253
SCHEDULE	202	25	26	253
ATTENDANCE DISCIPLINE	204	25	25	254
LMS PLATFORM	204	25	25	254
TECHNICAL SUPPORT	197	24	25	246
CLASSROOM INFRASTRUCTURE	203	25	25	253
LEARNING RESOURCES	203	25	26	254
ADMIN SERVICE	202	26	25	253
PAYMENT CONTRACT	203	25	25	253
SCHOLARSHIP SUPPORT	203	26	26	255
STUDENT SUPPORT	191	24	24	239
INTERNSHIP CAREER	191	24	24	239
LEARNING OUTCOME	191	24	24	239
ACADEMIC ENVIRONMENT	191	24	24	239
INCLUSION EQUALITY	191	24	24	239
SAFETY SECURITY	192	24	24	240
DORMITORY CAMPUS	189	24	24	237
LANGUAGE COMMUNICATION	190	24	24	238
DIGITAL CONTENT	188	24	23	235
OVERALL SATISFACTION	190	24	24	238

The saved ontology uses uppercase category names with underscores. Spaces are used in this display for readability; the machine-readable labels retain their original spelling.

Appendix B Individual Seed Results

The tables below report every final test seed. Values are F1 percentages, not confidence limits. Candidate scores are available only for the discriminative systems; an unavailable metric is not assigned zero. Tables 12 and 13 give strict tuple and candidate aspect scores, respectively.

Table 12 Strict tuple F1 for every final seed

System	13	29	42	71	101
MorphoSpan-Uz	19.45	33.09	30.48	32.37	15.37
XLM-R span	10.70	26.04	24.82	22.45	11.44
Without bytes	10.98	24.73	30.18	23.73	7.46
Without suffix features	17.22	26.69	30.48	31.21	8.10
Without boundary objective	11.20	25.26	17.46	21.90	7.68
Target-only LoRA	0.00	0.00	0.00	0.00	0.00
mT5 constrained	1.51	1.10	1.23	1.08	0.00
ByT5 constrained	0.00	1.09	3.78	0.00	5.94

Table 13 Candidate aspect exact F1 for every final seed

System	13	29	42	71	101
MorphoSpan-Uz	59.39	84.87	83.33	81.88	56.08
XLM-R span	58.43	82.69	77.93	72.81	36.19
Without bytes	50.92	76.04	77.58	77.13	29.69
Without suffix features	53.38	78.51	77.44	83.16	37.87
Without boundary objective	52.98	73.24	61.93	76.89	43.62
Target-only LoRA	0.00	0.00	0.00	0.00	0.00

Appendix C Reproducibility Record

The partition files contain 5,953 training, 743 development, and 744 test rows. Model inputs are raw texts grouped into complete source examples. Predictions contain source IDs, allocation-group IDs, reference tuples, final tuples, and, for discriminative systems, retained candidate lists. The scorer deduplicates sets within a source and aggregates exact counts before calculating micro-F1. Saved prediction identity was matched independently against the test data, rather than inferred solely from legacy configuration fields. Table 14 records the data fingerprints and model revisions.

Table 14 Frozen data fingerprints and model revisions

Object	SHA-256 or model commit
Original v4 train.jsonl	ba035dfed5294d493261db8321ec28d4db6b7535d34e48909fcfa1dd9ad9f868
Original v4 validation.jsonl	7b3d72feccef5f024209062d7d0618265e3425cfd0712f60868a276bc57b338b
Original v4 test.jsonl	ca462be0a8e27bde0fb0f7be6cdf8b24fbf7060a4a4c7bb80b31889266633ed4
Original v4 split_manifest.jsonl	b9bc1ddd48fd6735d149ac34c643a0e51a11fb44072dafe0f6e6ce394d89b84f
Compact test.jsonl	88743eb6ce80966f2c5c332bd35a77481b41e374bc224c326c25f127e6f798b8
Compact split manifest	1da1b57f0187ecbbc951d6fb8ac1c8899402546ab1438063688441b2e4f0ae4f
FacebookAI/xlm-roberta-base	e73636d4f797dec63c3081bb6ed5c7b0bb3f2089
google/mt5-small	73fb5dbe4756edadc8fbe8c769b0a109493acf7a
google/byt5-small	68377bdc18a2ffec8a0533fef03b1c513a4dd49d

Original v4 data and compact evaluation data have different byte fingerprints after packaging; their model inputs, targets, and groups match. Data values are SHA-256 fingerprints and model values are repository commit revisions.

The executable components are models.py, span_training.py, generation.py, and metrics.py under src/morphospan. The final comparison uses configs/experiments_span_v4.json, yakuniy_test/TEST_PLAN.json, per-run original_training_config.json, compatibility_audit.json, and predictions.jsonl. The v4 dataset metadata records mixed span provenance; the compact public packaging updates readiness metadata but does not turn those spans into a uniformly human-reviewed set. A readiness flag is therefore not a substitute for the explicit review audit.

The retrospective bootstrap uses paired seed resampling and complete allocation-group resampling, 10,000 replicates, random seed 1729, and percentile intervals of the mean-seed F1 difference. Checkpoint regeneration additionally requires the omitted weights or new training. Any new retraining, transfer study, threshold change, or enlarged annotation set must receive a new configuration and evaluation record.