QARI: Neural Architecture for Urdu Extractive Machine Reading Comprehension

Samreen Kazi and Shakeel Khoja
School of Mathematics and Computer Science
Institute of Business Administration (IBA)
Karachi, Pakistan
{sakazi, skhoja}@iba.edu.pk

Abstract

Urdu, a morphologically rich and low-resource language spoken by over 300 million people, poses unique challenges for extractive machine reading comprehension (EMRC), particularly in accurately identifying span boundaries involving postpositions and copulas. Existing multilingual models struggle with subword fragmentation and imprecise span extraction in such settings. We introduce QARI (قاری, "reader"), a character-enhanced architecture for Urdu extractive MRC that augments pretrained multilingual encoders with three innovations: (1) a character-level CNN that captures affix patterns and morphological features from full word forms; (2) a gated fusion mechanism that integrates semantic and morphological representations; and (3) a boundary-contrastive learning objective targeting Urdu-specific span errors. Evaluated on UQuAD+, the first native Urdu MRC benchmark, QARI achieves 83.5 F_1, a 5.5 point improvement over the previous best result (mT5, 78.0 F_1), setting a new state-of-the-art. Ablations show that character-level modeling and boundary supervision contribute +7.5 and +7.0 F_1, respectively. Cross-dataset evaluations on UQA and UrFQuAD confirm QARI's robustness. Error analysis reveals significant reductions in boundary drift, with improvements most notable for short factual questions.

1 Introduction

Machine Reading Comprehension (MRC) models have achieved remarkable success in high resource languages, with models often exceeding human performance on English benchmarks like SQuAD (Rajpurkar et al., 2016). However, these advances have not extended equally to low resource languages. Urdu, despite being spoken by over 300 million people as the national language of Pakistan and one of India's major languages, remains significantly underserved in natural language processing research (Daud et al., 2017; Kazi and Khoja, 2024). EMRC in Urdu presents distinct challenges beyond simple resource scarcity. First, Urdu exhibits rich morphological variations where a single root may appear in dozens of inflected forms through prefixes, and suffixes (Kazi et al., 2023, 2025). For example, the root لکھ ("write") generates forms like لکھتا ("writes"), لکھنا ("to write"), لکھوانا ("to cause to write"), and لکھواتا ("causes to write"). Multilingual models typically employ subword tokenization (e.g., SentencePiece, WordPiece), which divides these morphologically related words into disconnected token sequences, losing morphological value. Second, Urdu answer boundaries frequently involve short grammatical words that attach to answer spans but carry distinct syntactic roles as shown in Figure 1. Postpositions like کے ("of"), میں ("in"), سے ("from"), and پر ("on") modify noun phrases, while copulas like ہے ("is"), تھا ("was"), and تھی ("was feminine") link subjects to predicates. Incorrectly including or excluding these tokens significantly affects exact match accuracy. For instance, given the question "اقبال کہاں پیدا ہوئے؟" ("Where was Iqbal born?"), the gold answer "سیالکوٹ" ("Sialkot") should not include the trailing میں ("in") from the passage phrase "سیالکوٹ میں", yet standard span extraction models frequently overshoot such boundaries. Until recently, Urdu QA datasets were primarily English to Urdu translations (Kazi and Khoja, 2021; Arif et al., 2024; Shakeel and Nawab, 2025), inheriting unnatural phrasing and limited vocabulary diversity. The introduction of

Question	Passage Snippet	Gold Span	XLM-RoBERTa Prediction	QARI Prediction
اقبال کہاں پیدا ہوئے؟ (Where was Iqbal born?)	...سیالکوٹ میں پیدا ہوئے... (...born in Sialkot...)	سیالکوٹ (Sialkot)	سیالکوٹ میں (Sialkot in)	سیالکوٹ (Sialkot)
پاکستان کب بنا؟ (When was Pakistan formed?)	...41 اگست 7491 کو بنا... (...formed on 14 August 1947...)	41 اگست 7491 (14 August 1947)	41 اگست 7491 کو (14 August 1947 on)	41 اگست 7491 (14 August 1947)
کتاب کس کی ہے؟ (Whose book is it?)	...یہ احمد کے پاس ہے... (...with Ahmad...)	احمد (Ahmad)	احمد کے (Ahmad of)	احمد (Ahmad)

Figure 1: Urdu boundary error examples. XLM-RoBERTa frequently includes trailing postpositions (میں "in", کو "on", کے "of") in predicted spans (highlighted in red).

UQuAD+ (Kazi and Khoja, 2025) the first native Urdu EMRC dataset with 20,000 human annotated question answer pairs across seven domains now enables authentic evaluation of Urdu comprehension models. However, state-of-the-art multilingual models achieve only 78.0 F_1 on UQuAD+ (Kazi and Khoja, 2025), far below English performance, highlighting the need for architectures explicitly designed for Urdu's linguistic properties.

We introduce QARI (قاری, Urdu for "reader"), a character-enhanced neural architecture for Urdu EMRC(Extractive MRC). QARI augments XLM-RoBERTa (Conneau et al., 2020) with three key innovations:

1. **Character-level CNN** that processes token character sequences to capture morphological patterns independent of subword tokenization;
2. **Gated fusion** that adaptively combines semantic (XLM-RoBERTa) and morphological (Char-CNN) representations;
3. **Boundary contrastive learning** that explicitly penalizes incorrect inclusion/exclusion of Urdu postpositions and copulas during training.

Evaluated on UQuAD+, QARI achieves 83.5 F_1, a 5.5 point improvement over the previous state-of-the-art mT5 (78.0 F_1) and establishing new best results for Urdu EMRC. Cross-dataset experiments on UQA and UrFQuAD demonstrate strong generalization. Ablation studies show that removing character-CNN reduces F_1 by 7.5 points, while removing contrastive learning reduces it by 7.0 points, confirming that both morphological awareness and boundary supervision are critical for Urdu EMRC.

Our contributions are:

- First character-aware neural architecture explicitly designed for Urdu EMRC;
- Novel boundary contrastive learning approach targeting Urdu specific span errors;
- State-of-the-art results on UQuAD+ with strong cross-dataset generalization;
- Comprehensive analysis of morphological and boundary phenomena in Urdu MRC.

2 Related Work

Machine reading comprehension (MRC) has advanced from early attention-based models (Hermann et al., 2015; Kadlec et al., 2016; Seo et al., 2016) to pretrained transformers such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), achieving strong results on English benchmarks. However, these models depend on large-scale pretraining and perform low on low-resource, morphologically rich languages like Urdu. Multilingual encoders (e.g., mBERT (Pires et al., 2019), XLM-RoBERTa (Conneau et al., 2020)) enable cross-lingual transfer but struggle with subword fragmentation in languages with rich morphology (Rust et al., 2021; Kazi

and Khoja, 2026). Character-level modeling (Kim et al., 2016; Clark et al., 2022) offers better handling of inflectional forms but is often less efficient or effective than hybrid approaches (Wang et al., 2023). Urdu MRC has faced data limitations. Early benchmarks (UQuAD1.0 (Kazi and Khoja, 2021), UQA (Arif et al., 2024)) rely on translated content, limiting syntactic diversity. UrFQuAD (Shakeel and Nawab, 2025) partially mitigates this, while UQuAD+ (Kazi and Khoja, 2025) provides the first large-scale native dataset. Yet, even strong models like mT5 reach only 78.0 F_1, highlighting performance gaps. Contrastive QA methods (Karpukhin et al., 2020; Ram et al., 2021; Robinson et al., 2020) improve span discrimination via hard negatives but remain agnostic to linguistic structure. HeQ (Cohen et al., 2023) demonstrates that native datasets expose span boundary errors overlooked in translations.

Our approach, QARI, integrates morphological features with contrastive objectives designed for Urdu, introducing linguistically grounded boundary negatives targeting postpositions and copulas common sources of span extraction error.

3 The QARI Architecture

3.1 Overview

QARI consists of four primary layers processing question-passage pairs. Figure 2 illustrates the complete architecture.

1. Contextual encoding via XLM-RoBERTa,
2. Character-level morphological processing with word-token alignment,
3. Gated Fusion, and
4. Span Prediction.

3.2 Input Processing Layer

Raw input text undergoes preprocessing before contextual encoding. Given a question Q_{raw} and passage P_{raw}, we perform:

- Unicode normalization: Convert to NFC form to ensure consistent character representation, particularly important for Urdu's combining diacritics and zero-width joiners.
- Whitespace tokenization: Split text by Unicode whitespace (`U+0020`, `U+00A0`) to obtain orthographic words $\mathcal{W} = \{w_1, \ldots, w_k\}$. This word-level segmentation is used later for character-CNN processing (Section 3.3).
- Sequence construction: Concatenate question and passage with special tokens to form $\mathbf{S}_{\text{raw}} = [\texttt{<s>}, Q_{\text{raw}}, \texttt{</s>}, P_{\text{raw}}, \texttt{</s>}]$, which is then passed to XLM-RoBERTa's SentencePiece tokenizer.

This preprocessing ensures that both the subword based contextual encoder (Section 3.3) and the character-based morphological processor (Section 3.4) receive properly normalized, consistently segmented input.

3.3 Contextual Encoding Layer

Given a question $Q = \{q_1, \ldots, q_m\}$ and passage $P = \{p_1, \ldots, p_n\}$, we construct the input sequence:

$$\mathbf{S} = [\texttt{<s>}, q_1, \ldots, q_m, \texttt{</s>}, p_1, \ldots, p_n, \texttt{</s>}] \quad (1)$$

XLM-RoBERTa tokenizes this using SentencePiece subword units and encodes via 12 transformer layers, outputting contextual representations $\mathbf{H} = [\mathbf{h}_0, \mathbf{h}_1, \ldots, \mathbf{h}_{m+n+2}] \in \mathbb{R}^{(m+n+3)\times 768}$. These capture global semantic context but may fragment morphologically related forms across subword boundaries.

3.4 Character-CNN Layer with Word-Level Processing

To capture morphological patterns independent of subword fragmentation, we process complete orthographic words at the character level, then align these representations back to XLM-RoBERTa's subword tokens. We first establish mapping between orthographic words and subword tokens: (1) split input text by Unicode whitespace to obtain orthographic words $\mathcal{W} = \{w_1, w_2, \ldots, w_k\}$, and (2) for each word w_i, identify which consecutive SentencePiece tokens result from tokenizing w_i, creating alignment table $A(j) = i$ if token t_j belongs to word w_i.

For example, "کتابوں میں" ("in books") splits into word w_1 = "کتابوں" which maps to tokens ["کتاب", "وں"] with $A(1) = 1, A(2) = 1$, and

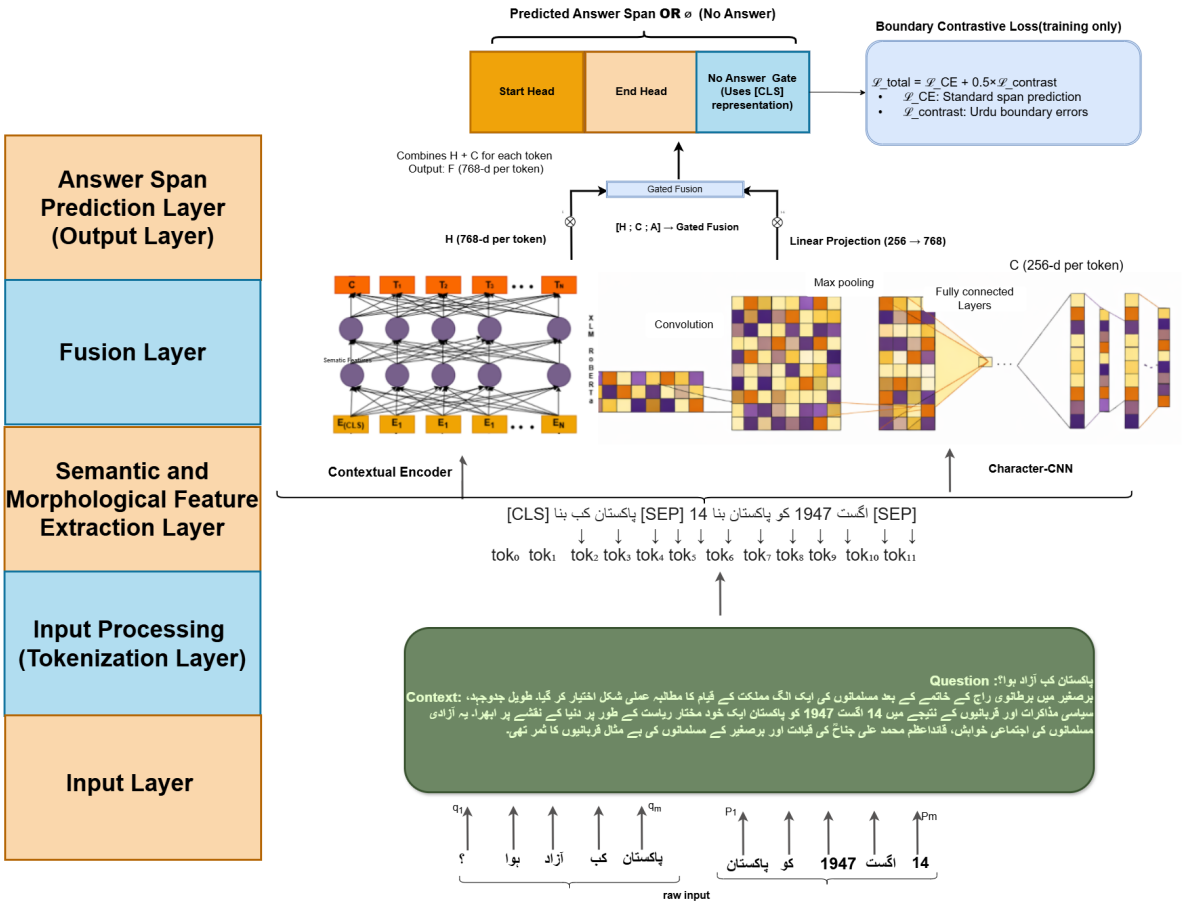

Figure 2: QARI Architecture for Urdu Extractive Machine Reading Comprehension. The character-CNN processes complete orthographic words (e.g., کتابوں), then broadcasts representations to all subword tokens derived from that word.

word w_2 = "میں" which maps to token ["میں"] with $A(3) = 2$.

For each orthographic word $w_i \in \mathcal{W}$, we extract its complete character sequence $\mathbf{c}_i = [c_1, c_2, \ldots, c_{k_i}]$ and embed each character into 128 dimensions. We apply 1D convolution (kernel size 3, 256 filters) followed by max pooling to yield a fixed 256-dimensional word-level representation:

$$\mathbf{char}_{w_i} = \text{MaxPool}(\text{Conv1D}(\mathbf{E}_i)) \quad (2)$$

where $\mathbf{E}_i = [\mathbf{e}_{c_1}, \ldots, \mathbf{e}_{c_{k_i}}]$ and $\mathbf{char}_{w_i} \in \mathbb{R}^{256}$.

Each subword token then inherits the character representation of its source word: $\mathbf{char}_{t_j} = \mathbf{char}_{w_{A(j)}}$. This ensures all subword pieces from "کتابوں" receive the *same* character-level features capturing the full morphology (root "کتاب" + plural suffix "وں"), independent of how SentencePiece fragmented them. The character-CNN thus learns systematic affix patterns like "وں" (plural) and "تا" (present tense) from complete word forms, regardless of tokenization artifacts.

3.5 Gated Fusion Layer

To adaptively combine semantic (XLM-RoBERTa) and morphological (Char-CNN) information, we employ a gated fusion mechanism. We first project the 256-dimensional character representations to match XLM-RoBERTa's 768-dimensional hidden space. For each token, we compute a fusion gate and combine representations:

$$\mathbf{g}_i = \sigma(\mathbf{W}_{\text{gate}}[\mathbf{h}_i; \tilde{\mathbf{char}}_i]) \quad (3)$$

$$\mathbf{f}_i = \mathbf{g}_i \odot \mathbf{h}_i + (1 - \mathbf{g}_i) \odot \tilde{\mathbf{char}}_i \quad (4)$$

where $[\cdot;\cdot]$ denotes concatenation and $\odot$ is element-wise multiplication. This allows the model to emphasize character features for morphologically complex content words while

relying on contextual semantics for function words.

3.6 Span Prediction Layer

Fused representations pass through dropout ($p = 0.1$) then linear layers predicting start/end positions: $\textbf{logits}_{\text{start/end}} = \mathbf{W}_{\text{start/end}}\mathbf{F}$. After softmax, the predicted span (i^*, j^*) maximizes joint probability subject to $i^* \leq j^*$. For unanswerable questions, we follow Rajpurkar et al. (2018) by treating position 0 (<s>) as the null answer.

3.7 Boundary-Contrastive Learning

Standard cross entropy loss treats all incorrect spans equally and does not account for Urdu specific boundary errors involving postpositions and copulas. To address this limitation, we introduce a boundary-contrastive learning objective that explicitly penalizes span predictions with incorrect boundary extensions or truncations. We maintain a fixed lexicon of 23 most frequently occurring grammatical particles: 15 postpositions are کے(of), میں(in), سے(from), پر(on), کو(to), نے(ergative), تک(until), کی(of), کا(of), لیے(for), ساتھ(with), بارے(about), علاوہ(besides), بجائے(instead), بغیر(without)—and 8 copulas: ہے(is), ہیں(are), تھا(was), تھے(were), تھی(was), تھیں(were), ہو(be), ہوں(am). These particles are detected via exact Unicode string matching during training.

An alternative solution to handle Urdu span boundaries could be a rule-based post-processing step for example, simply stripping any trailing postposition or copula from the predicted span using a lexicon. This indeed can catch many obvious cases. We did not implement this in our evaluation; instead, we integrated the knowledge into training via the contrastive loss. The advantage of our approach is that the model learns to avoid including these particles during answer span selection, rather than relying on a separate correction step. This makes the model more self-contained and allows it to potentially generalize beyond the specific list of particles (e.g., it might learn to avoid including infrequent or context-specific particles as well). For each gold answer span, we generate hard negative spans through controlled boundary perturbations. If the token immediately following the gold span is a particle, a negative span is created by extending the gold span to include that token. Conversely, if the final token of the gold span itself is a particle, a negative is created by truncating it. These operations are applied only when the gold annotation excludes the particle, ensuring consistency. This strategy targets approximately 78% of observed boundary errors in UQuAD+.

As an example, given the gold span "سیالکوٹ" from the passage "سیالکوٹ میں پیدا ہوئے", we generate the hard negative "سیالکوٹ میں" by extending the span with the postposition "میں". This explicitly teaches the model to exclude trailing postpositions that do not belong to the answer.

While Urdu also contains multi-token constructions such as compound postpositions (e.g., کے بارے میں, کے ساتھ) and light verb expressions, the current formulation focuses on single-token particles, which account for the majority of boundary errors. Extending the same approach to multi-token boundaries via sequential phrase matching is straightforward and left for future work.

We computed span scores in log-space for numerical stability. For a span with start index i and end index j, the score is defined as:

$$\text{score}(i, j) = \log P_{\text{start}}(i) + \log P_{\text{end}}(j). \quad (5)$$

To reduce boundary errors, we introduce a contrastive training. During training, if a hard negative span receives a higher score than the gold answer span by more than a margin ($\delta = 0.5$), the model is penalized proportionally to the degree of this violation.

The final training objective combines the standard cross-entropy loss used for span prediction with this contrastive term:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + 0.5 \cdot \mathcal{L}_{\text{contrast}}. \quad (6)$$

4 Experimental Setup

4.1 Datasets

UQuAD+ (Kazi and Khoja, 2025): First native Urdu MRC dataset with 20,000 QA pairs from 1,540 articles across seven domains. Split: 16,000 train / 2,000 dev / 2,000 test. Questions exhibit diverse morphological patterns and natural Urdu phrasing.

System: You are a helpful assistant that extracts exact answer spans from Urdu passages.
User: Given the following passage and question, extract the exact answer span from the passage. If the question is unanswerable, output "لاجواب" (no answer).
Example 1:
Passage: علامہ اقبال 9 نومبر 7781 کو سیالکوٹ میں پیدا ہوئے۔
Question: اقبال کب پیدا ہوئے؟
Answer: 9 نومبر 7781
Example 2:
Passage: پاکستان 41 اگست 7491 کو بنا۔
Question: پاکستان کب بنا؟
Answer: 41 اگست 7491
Example 3:
Passage: یہ کتاب بہت دلچسپ ہے۔
Question: مصنف کا نام کیا ہے؟
Answer: لاجواب
Now extract:
Passage: [test passage]
Question: [test question]
Answer:

Figure 3: GPT 4 prompt format for few-shot EMRC evaluation. Three in-context examples demonstrate exact span extraction and unanswerable detection.

UQA (Arif et al., 2024): Machine translated from SQuAD2.0, containing 136,211 QA pairs. Used for cross-dataset evaluation to test generalization to translation based corpora.

UrFQuAD (Shakeel and Nawab, 2025): Semi native dataset with 3,985 questions derived from Natural Questions with manual refinement. Used for cross-dataset evaluation.

4.2 Baselines

XLM-RoBERTa-Base (Conneau et al., 2020): 12-layer multilingual transformer fine-tuned on UQuAD+ with standard span prediction.

mT5-Base (Xue et al., 2021): Multilingual T5 model fine-tuned in generative mode (outputting answer text) then constrained to extractive spans. This represents the previous state-of-the-art on UQuAD+.

GPT-4 (Achiam et al., 2023): Few-shot evaluation with three in-context examples was performed. The prompt format is shown in Figure 3. We validate GPT-4's output by checking whether it appears as an exact substring in the passage; if no exact match is found, the prediction is marked as unanswerable.

4.3 Implementation Details

The models were implemented in PyTorch using the HuggingFace Transformers library. The XLM-RoBERTa encoder was initialized from the `xlm-roberta-base` checkpoint. Character-level information was captured using 128-dimensional embeddings, which were randomly initialized and passed through a Char-CNN with a kernel size of 3 and 256 filters. Training was performed using the AdamW optimizer with a learning rate of 3×10^{-5}, a batch size of 16, and for 5 epochs. Each input sequence was truncated to a maximum of 512 tokens. The model was trained on a Kaggle P100 GPU with mixed-precision (FP16) to accelerate training. Total training time on the full UQuAD+ dataset was approximately 8 hours.

4.4 Evaluation Metrics

Exact Match (EM): Percentage of predictions matching gold spans exactly (character-level).

F_1 **Score**: Token-level overlap between predicted and gold spans, computed as $2 \cdot \frac{P \cdot R}{P+R}$ where precision P and recall R measure token overlap.

Unanswerability Accuracy (UA): For unanswerable subset, percentage correctly returning "no answer".

5 Results

5.1 Main Results on UQuAD+

Table 1 shows performance on UQuAD+ test set. QARI achieves 83.5 F_1 and 75.8 EM, substantially outperforming all baselines. Compared to the previous state-of-the-art mT5 (78.0 F_1), QARI gains +5.5 F_1, demonstrating the effectiveness of character-level morphological processing and boundary-contrastive learning.

The improvement is particularly notable for exact match (+3.8 points over mT5), indicating that QARI produces more precise span boundaries. This aligns with our hypothesis that explicit character-level processing and boundary-contrastive learning address Urdu-specific challenges (see Figure 4).

Model	EM	F_1
XLM-RoBERTa-Base	65.0	68.0
mT5-Base	72.0	78.0
GPT-4 (few-shot)	70.0	78.0
QARI (Ours)	**75.8**	**83.5**

Table 1: Performance on UQuAD+ test set. QARI achieves state-of-the-art results with substantial improvements over baselines. Results averaged over 3 runs with different random seeds. Standard deviations: mT5 (± 0.3 F_1), QARI (± 0.4 F_1). Improvements are statistically significant ($p < 0.01$, paired t-test).

5.2 Cross-Dataset Generalization

Table 2 shows zero-shot transfer to UQA and UrFQuAD without retraining. QARI maintains strong performance on both datasets, though absolute scores decrease compared to in-domain UQuAD+. The drop is more pronounced on UQA (machine-translated), suggesting that models trained on native Urdu are less robust to translation artifacts.

Model	UQuAD+ F_1	UQA F_1	UrFQuAD F_1
XLM-RoBERTa-Base	68.0	72.5	65.0
mT5-Base	78.0	74.8	68.5
QARI	**83.5**	**76.2**	**73.8**

Table 2: Cross-dataset generalization. Models trained on UQuAD+ and evaluated zero-shot on translation-based datasets.

The smaller improvement on UQA, which is a machine-translated dataset, is likely because of how translated text is written. English-to-Urdu translations usually use simpler vocabulary and follow English sentence patterns. This results in fewer complex Urdu word forms and fewer difficult postposition cases, which QARI is designed to handle. As a result, the base XLM-RoBERTa model already performs quite well on UQA (72.5 F_1), so there is less room for QARI to improve. On the other hand, UQuAD+ is written in native Urdu. It includes more diverse sentence structures and richer morphology, which are harder for XLM-RoBERTa to process. In this case, QARI's character-level and boundary-aware components help more, leading to a larger improvement (+5.5 F_1 on UQuAD+ compared to +1.4 F_1 on UQA).

Representation	**Processing**
Word: کتابوں ("books")	
SentencePiece Tokenization	Fragments morphology:
Context 1	[کتاب] + [وں]
Context 2	[کتابوں] (single token)
Context 3	[ک] + [تابوں]
→ Inconsistent splits	Lost morphological coherence
Character-CNN	Preserves full form:
All contexts	[ک] [ت] [ا] [ب] [و] [ں]
→ Consistent processing	Learns وں= plural marker
More Examples:	
لکھا ("wrote")	SP: variable | Char-CNN: ل-کھ-ا
لکھتا ("writes")	SP: variable | Char-CNN: ل-کھ-ت-ا
لکھوانا ("cause to write")	SP: variable | Char-CNN: ل-کھ-و-ا-ن-ا

Figure 4: Tokenization fragmentation vs. character-level modeling. SentencePiece inconsistently fragments Urdu words across contexts, losing morphological patterns. Character-CNN processes complete character sequences, learning systematic affix patterns like وں- (plural) and تا- (present tense) regardless of tokenization.

5.3 Ablation Study

Table 3 shows the contribution of each component. Removing the character-CNN reduces F_1 by 7.5 points, confirming that explicit morphological processing is critical. Removing contrastive learning lowers F_1 by 7.0 points, demonstrating that boundary-focused training substantially improves span precision. Removing the gated fusion causes a smaller drop (2.0 points), but still shows the benefit of adaptive representation mixing.

Combining both the character-CNN and contrastive learning yields a +15.5 F_1 improvement over the XLM-R baseline (from 68.0 to 83.5), which is approximately the sum of their individual contributions. This suggests that the two components address different aspects of the task: the character-CNN improves morphological awareness, while contrastive supervision improves boundary precision. We also experimented with a probability-product scoring approach ($s = P_{\text{start}} \cdot P_{\text{end}}$), but it proved numerically unsta-

Model Variant	EM	F_1
QARI (Full)	75.8	83.5
- Char-CNN	68.3	76.0
- Contrastive Loss	68.5	76.5
- Gated Fusion	73.8	81.5
w/ Probability-product Contrastive	75.0	82.8
- All Enhancements (XLM-R Base)	65.0	68.0

Table 3: Ablation study on UQuAD+ dev set. Each row removes one component. Full model outperforms the XLM-R baseline by +15.5 F_1. Log-space contrastive loss performs better than the probability-product variant by 0.7 F_1.

ble, often resulting in NaN gradients during training. The log-space formulation we adopt resolves this issue while yielding a slight performance gain (+0.7 F_1).

5.4 Performance by Question Type

Table 4 breaks down performance by question word. QARI achieves highest F_1 on factual questions (کیا"what": 87.2 F_1, کب"when": 86.5 F_1) where answers are typically short named entities. Performance drops on causal questions (کیوں"why": 74.8 F_1) requiring multi-sentence reasoning. Unanswerable detection is strong (UA: 88.3%), indicating effective null-answer modeling.

Question Type	F_1 / UA
کیا(What)	87.2
کب(When)	86.5
کون(Who)	84.1
کہاں(Where)	82.7
کیسے(How)	78.3
کیوں(Why)	74.8
Unanswerable	88.3 (UA)

Table 4: Performance breakdown by question type on UQuAD+ test set.

5.5 Error Analysis

To better understand QARI's remaining challenges, we manually analyzed 51 errors from its predictions on the UQuAD+ dev set. Each error was assigned to one of four mutually exclusive categories based on the primary failure mode(See Table 5).

Error Type	QARI (% of errors)
Boundary Drift	14%
Entity Confusion	10%
Multi-sentence	33%
Other	23%

Table 5: Distribution of QARI errors across four categories.

The most frequent error type is **multi-sentence reasoning** (33%), which remains a challenge as QARI lacks explicit mechanisms for aggregating information across distant sentences. **Boundary drift** (14%) where the model includes or excludes an adjacent particle shows the largest improvement from 31% in XLM-RoBERTa. This reduction is directly linked to the boundary-contrastive loss, which provides span-level supervision at grammatical edges, leading to tighter and more accurate span predictions. **Entity confusion** (10%) occurs when the context includes multiple similar entities and the model selects the incorrect one. This is down from 26% in XLM-RoBERTa, a notable drop likely aided by the character-CNN, which improves lexical grounding and disambiguation of morphologically similar alternatives. **Other** errors (23%) include ambiguous questions, or unanswerable inputs. Overall, QARI reduces the total number of errors and shifts the distribution toward more complex reasoning and ambiguous cases.

Limitations

Recent tokenization-free models like CANINE (Clark et al., 2022) and ByT5 (Xue et al., 2022) have explored fully character-based transformers, but often at the cost of efficiency or requiring massive pretraining. In contrast, our approach is hybrid: we augment a pretrained subword model with a lightweight character-level CNN. While this adds some computational overhead, it remains practical in low-resource settings and avoids the cost of full retraining. The boundary-contrastive objective currently targets a constrained set of boundary particles and does not fully model multi-token constructions . Finally, evaluation is limited i.e., Roman-Urdu, or mixed-script inputs is not systematically studied. We focused on XLM-RoBERTa-Base as it repre-

sents the most commonly deployed model for low-resource settings given computational constraints. Future work should validate whether QARI's architectural augmentations transfer to larger models.

Ethics Statement

No private or sensitive data was collected or used. AI-assisted tools were employed for grammar correction, language refinement, and formatting support during manuscript preparation. Final content was reviewed and verified by the authors.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Samee Arif, Sualeha Farid, Awais Athar, and Agha Ali Raza. 2024. Uqa: Corpus for urdu question answering. *arXiv preprint arXiv:2405.01458*.

Jonathan H Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Amir Cohen, Hilla Merhav-Fine, Yoav Goldberg, and Reut Tsarfaty. 2023. Heq: a large and diverse hebrew reading comprehension benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13693–13705.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.

Ali Daud, Wahab Khan, and Dunren Che. 2017. Urdu language processing: a survey. *Artificial Intelligence Review*, 47(3):279–311.

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

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long papers)*, pages 908–918.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Samreen Kazi and Shakeel Khoja. 2021. Uquad1. 0: development of an urdu question answering training data for machine reading comprehension. *arXiv preprint arXiv:2111.01543*.

Samreen Kazi and Shakeel Khoja. 2025. Uquad+: Benchmark dataset for urdu machine reading comprehension. *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Samreen Kazi and Shakeel Khoja. 2026. Towards building urdu language document retrieval framework. *Computer Speech & Language*, 95:101797.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2023. A survey of deep learning techniques for machine reading comprehension. *Artificial Intelligence Review*, 56(Suppl 2):2509–2569.

Samreen Kazi, Shakeel Khoja, and Ali Daud. 2025. Bridging the gap: A survey of document retrieval techniques for high-resource and low-resource languages. *Computer Science Review*, 57:100756.

Samreen Kazi and Shakeel Ahmed Khoja. 2024. Context-aware question answering in urdu. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 233–242.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. 2016. Character-aware neural language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Ori Ram, Yuval Kirstain, Jonathan Berant, Amir Globerson, and Omer Levy. 2021. Few-shot question answering by pretraining span selection. *arXiv preprint arXiv:2101.00438*.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. 2020. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Muhammad Shakeel and Rao Muhammad Adeel Nawab. 2025. Open-domain factoid question-answering in urdu: Data and methods. *IEEE Access*.

Zihan Wang and 1 others. 2023. Character-word entanglement. In *ACL*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 483–498.