DOI: 10.37943/HRBD3124

Manas Yergesh
PhD Student, Department of Artificial Intelligence Technologies
Shymkent90@mail.ru, orcid.org/0000-0002-2180-293X
L.N. Gumilyov Eurasian National University, Kazakhstan
Kundyz Maxutova
PhD, Department of Artificial Intelligence Technologies
qunkabai@gmail.com, orcid.org/0000-0002-3216-0397
L.N. Gumilyov Eurasian National University, Astana, Kazakhstan.
Alibek Barlybayev
PhD, Director of Artificial Intelligence RSI
frank-ab@mail.ru, orcid.org/0000-0002-0188-5336
L.N. Gumilyov Eurasian National University, Kazakhstan
Arailym Tleubayeva
PhD Doctoral Student, School of Artificial Intelligence and Big Data
a.tleubayeva@astanait.edu.kz, orcid.org/0000-0001-9560-9756
Astana IT University, Kazakhstan
Ainur Orazayeva
PhD, Department of Information Technology
oar_is@mail.com, orcid.org/0000-0002-2899-9886
K. Kulazhanov Kazakh University of Technology and Business, Kazakhstan
Aigul Abildina
Master's Degree, Senior Lecturer, Department of Information Technology
abildina12@mail.ru, orcid.org/0009-0000-6597-7910
K. Kulazhanov Kazakh University of Technology and Business, Kazakhstan

A REPRODUCIBLE EVALUATION OF ONTOLOGY-GUIDED HYBRID EXTRACTIVE QUESTION ANSWERING FOR KAZAKH HISTORY

Abstract: Reproducibility and comparability remain important methodological concerns in domain-specific question answering, particularly in low-resource educational settings. This study examines the impact of ontology-guided context filtering on extractive question answering in the Kazakh language under controlled and reproducible conditions. A benchmark corpus on the History of Kazakhstan was constructed from 11 school textbooks for grades 5–10 and canonically indexed using deterministic section identifiers. Fixed train–test splits, standardized chunking, and unified preprocessing and evaluation scripts ensured consistency across experiments.

Three question answering systems were evaluated under identical constraints: a neural extractive baseline operating on unfiltered context, an ontology-only system relying on structured knowledge representations, and a hybrid system applying ontology-guided context filtering prior to neural answer extraction. Filtering was conducted under an inference setting that excludes access to gold document or section identifiers, preventing the use of privileged metadata.

The hybrid system demonstrates modest improvements over the neural-only baseline. Exact Match accuracy increases from 68.87% to 70.45%, while the token-level F1 score improves from 80.02% to 81.68%. Ranking-based retrieval metrics also show small gains, with Mean Reciprocal Rank increasing from 0.870 to 0.892 and Recall@10 from 95.1% to 96.9%. Inference latency is reduced from 20.01 ms to 6.83 ms per question. Paired statistical testing confirms that the improvement in F1 is statistically significant, whereas improvements in Exact Match are less conclusive.

Ablation experiments suggest that these differences are not attributable to context length reduction alone, but are associated with the interaction between ontology-guided context selection

Received: 28.05.2026 Accepted: 17.07.2026 Published: 30.06.2026

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

and neural reader training. Overall, ontology-guided filtering can provide incremental benefits for extractive question answering in a domain-specific, low-resource educational setting under transparent and reproducible experimental conditions.

Keywords: domain-specific question answering; extractive question answering; ontology-guided filtering; hybrid neuro-symbolic systems; low-resource languages; Kazakh language; reproducible benchmarking; educational question answering.

Introduction

In low-resource and educational settings, additional care is required to ensure that experimental results are reliable and comparable. Small differences in data preprocessing, train–test splits, or evaluation procedures may substantially affect reported performance, especially when datasets are limited in size and scope [1-2]. In addition, distributional shifts, such as differences in question wording, topic coverage, or corpus editions, can noticeably affect model behavior, reinforcing the need for fixed splits, documented indexing, and consistent evaluation scripts [3].

One important aspect concerns the construction of the input context provided to the neural reader. In practice, this context is often produced by auxiliary procedures such as filtering or retrieval rather than being directly given [4]. If these procedures are not consistently applied across systems, it becomes difficult to determine whether observed performance differences are caused by the QA model itself or by differences in the input context [5]. This issue is particularly relevant for hybrid approaches that combine neural models with structured knowledge, where access to document-level metadata may unintentionally simplify the task at inference time [6].

To reduce such confounding factors, this study adopts a unified and transparent experimental setup. The corpus is indexed in a canonical manner, data splits are fixed and deterministic, and all systems are evaluated using the same metrics and normalization rules, following established best practices in QA benchmarking [7]. Ontology-guided filtering is applied under a NO-GOLD setting, meaning that information about the correct book or section is not available during inference. This constraint is intended to ensure that context selection relies only on the content of the question and the available knowledge resources [8].

Within this framework, three types of systems are examined. The NN-only model represents a standard neural extractive QA baseline operating on unfiltered context, following the dominant reader-based paradigm [9]. The Ontology-only approach provides an interpretable reference that does not involve neural training and reflects classical knowledge-based QA methods [10]. The Hybrid-B system combines these two components by using ontology-based signals to guide context selection before neural answer extraction, consistent with recent neuro-symbolic approaches [11]. The neural reader architecture and training procedure are kept the same across neural and hybrid systems so that observed differences can be attributed primarily to the way context is constructed.

The aim of this work is to examine whether ontology-guided context filtering can provide measurable benefits in terms of answer quality, ranking stability, and inference efficiency under controlled conditions. Rather than proposing a new model architecture, the focus is on a careful comparison of existing approaches using a reproducible evaluation protocol in a specific educational domain [12].

Literature Review

Reproducibility has become a central concern in machine learning research, especially for tasks where small implementation choices can change conclusions. Prior work emphasizes that “replicability” and “reproducibility” are not interchangeable, and that reported gains are difficult to trust without transparent experimental protocols and controlled conditions [1-2]. This problem is amplified in low-resource settings, where datasets are smaller and variance across splits and preprocessing pipelines is higher. In addition, distributional shifts (e.g., differences in question wording, topic coverage, or corpus editions) can noticeably affect model behavior, reinforcing the need for fixed splits, documented indexing, and consistent evaluation scripts [3].

DOI: 10.37943/HRBD3124

For question answering systems, reproducibility is not only about training determinism; it also concerns the construction of inference-time inputs (retrieval, filtering, truncation rules) and the evaluation pipeline (normalization, exact match definition, handling of unanswerable items). Benchmarks such as SQuAD 2.0 explicitly highlight the importance of unanswerable questions and calibrated confidence, motivating stricter evaluation assumptions when moving from open-domain to constrained, domain-specific settings [7].

Modern neural extractive QA is typically formulated as predicting an answer span within a provided context. Early "machine reading at scale" pipelines established the now-standard separation between (i) retrieving candidate passages and (ii) applying a reader to extract an answer span [4]. However, reader quality alone does not guarantee robust performance: when context is long, noisy, or only weakly related, span extractors may latch onto superficial lexical matches and produce unstable outputs. Multi-paragraph reading work showed that naïvely applying paragraph-level models across many candidate paragraphs is vulnerable to distraction by irrelevant text, motivating approaches that explicitly model paragraph selection and aggregation [5].

Transformer-based pretraining further improved extractive QA, with BERT-style encoders becoming dominant due to strong contextual representations [9]. At the same time, surveys of BERT behavior stress that performance gains can hide brittle heuristics and dataset artifacts, reinforcing the importance of controlled evaluation and ablations when reporting improvements [8]. This is particularly relevant in educational QA, where users expect consistent answers and where errors may be pedagogically harmful even if aggregate metrics look acceptable.

Because extractive QA depends heavily on the supplied context, retrieval and ranking have become core components. Dense Passage Retrieval (DPR) demonstrated that learned dense retrievers can outperform sparse baselines in open-domain settings by improving the recall of semantically relevant passages [13]. Retrieval-augmented generation (RAG) extended this idea by combining a retriever with a sequence-to-sequence generator, showing that explicit non-parametric memory can improve factual coverage in knowledge-intensive tasks [14]. Related generative QA models such as BART provide the sequence-to-sequence backbone for many retrieval-augmented pipelines and enable flexible answer generation when strict span extraction is insufficient [15].

A key observation from this literature is that "better QA" is often "better context": performance gains can originate from improved candidate selection rather than improved reasoning. Therefore, comparisons between systems must ensure identical constraints on retrieval depth, truncation, and access to metadata; otherwise, reported gains may reflect hidden leakage or an easier inference setting rather than a genuinely stronger QA method.

Ontology- and KG-based QA has long been studied as an alternative that prioritizes interpretability and structured reasoning. Ontology-driven QA systems have shown that mapping natural language questions onto ontology-aware representations can support explainable answers and structured query execution [10]. More broadly, knowledge graphs provide a schema-aware representation that can support entity-centric retrieval, constraint handling, and provenance tracking—properties that are especially valuable in educational domains [16].

Recent surveys of KGQA document strong progress but also emphasize persistent challenges in entity linking, multi-hop reasoning, constraint satisfaction, and dataset bias, particularly for complex questions [17]. These limitations are often more pronounced in low-resource languages, where high-quality entity dictionaries, relation inventories, and annotated KGQA datasets are scarce. As a result, ontology-only systems can be transparent but may struggle to reproduce exact textual answers or handle paraphrasing and incomplete coverage.

A growing body of work argues that integrating symbolic structure with neural representation learning can combine the strengths of both paradigms: the robustness of neural models and the controllability of symbolic reasoning [11, 18, 19]. Surveys of neural-symbolic learning highlight common integration patterns, including symbolic constraints during learning, symbolic retrieval followed by neural reading, and joint reasoning architectures. They also identify

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

open problems such as scalability, knowledge incompleteness, and evaluation methodology [18-19].

In QA specifically, a practical and widely used hybridization strategy is symbolic guidance for context construction followed by neural answer extraction. This design does not require inventing a new reader architecture; instead, it targets a frequent failure mode of neural QA—poor context quality—by reducing noise and improving the relevance of the input supplied to the reader. Importantly, for such hybrid systems, the fairness of evaluation depends on preventing inference-time access to gold document identifiers or section labels; otherwise, symbolic components may unintentionally bypass the true difficulty of the task.

For low-resource languages, multilingual transfer is a standard strategy. Work on multilingual BERT indicates that shared multilingual representations enable cross-lingual transfer even across scripts, but also reveals systematic weaknesses and uneven transfer quality depending on language pairs and typological similarity [20]. This motivates careful benchmarking in Kazakh: improvements should be reported under fixed splits and consistent preprocessing, and claims should be validated with sanity checks to separate genuine gains from artifacts of context selection or evaluation variance.

The aim and objectives of the study

The aim of this study is to evaluate whether ontology-guided context filtering can improve extractive question answering in a low-resource educational domain under controlled and reproducible conditions, without relying on gold document metadata at inference time (NO-GOLD setting).

To achieve this aim, the study addresses the following objectives:

– construct a canonically indexed Kazakh history corpus and a consistent extractive QA benchmark with fixed train–test splits;

– implement and evaluate three system classes (NN-only, Ontology-only, and Hybrid-B) under identical preprocessing and evaluation protocols;

– quantify answer quality, retrieval stability, and inference latency using standard metrics (EM, F1, MRR, Recall@K);

– validate interpretability and robustness through sanity checks, ablation experiments, and paired statistical testing.

Methods and Materials

A. Corpus Composition and Canonical Indexing of Sources

The experimental corpus covers the History of Kazakhstan in the Kazakh language and aggregates materials from 11 educational textbooks corresponding to grades 5–10. After loading and integrity verification of all sources, a canonical indexing of sections was performed. A unified registry of sections was constructed (sections_loaded = 488) with deterministic identifiers and associated metadata (book_id, section_id, stable_id). For each section, both the original text (text) and its normalized representation (text_n) were stored, together with length statistics measured in characters and tokens.

This canonicalization step is essential for strict reproducibility. All downstream objects—including chunks, queries, relevance judgments (qrels), and training splits—are uniquely anchored to stable section keys, ensuring that experimental artifacts can be deterministically reconstructed across runs.

Description and motivation. Figure 1 illustrates the distribution of corpus volume across sources and highlights an inherent imbalance in coverage. The largest number of sections originates from Kazakh_History_6 (54 sections), followed by Kazakh_History_5 (46), then textbooks for grades 7–9 (28–29 each), and Kazakh_History_10 (22). This diagnostic is methodologically important, as such imbalance may influence entity frequency, topic distribution, and question difficulty. Consequently, it must be considered during split construction and error analysis in subsequent experiments.

DOI: 10.37943/HRBD3124

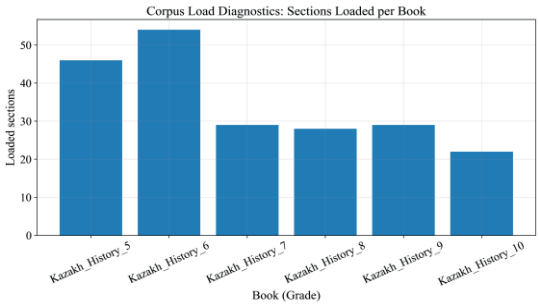

Figure 1. Corpus loading diagnostics: number of sections per textbook (grade level)

B. Section Length Statistics and Motivation for Chunking

The corpus exhibits substantial variability in section length across grade levels, which is critical for transformer-based models with limited input context. The average section length (in tokens) is as follows: grade 5 — 516, grade 6 — 853, grade 7 — 679, grade 8 — 1645, grade 9 — 508, and grade 10 — 1013. The longest sections, predominantly from grade 8, pose a risk of exceeding the model's context window and lead to an increased number of sliding windows during inference. This, in turn, raises computational cost and complicates consistent alignment between questions, contexts, and answers.

Fig. 2 quantitatively confirms the heterogeneity of section lengths across sources and motivates the use of chunking as a unifying transformation. Segmenting long sections into fixed-size overlapping fragments reduces the risk of information truncation, stabilizes input sizes for both retrieval and reader components, and enables the construction of consistent relevance tables (qrels) at the fragment level.

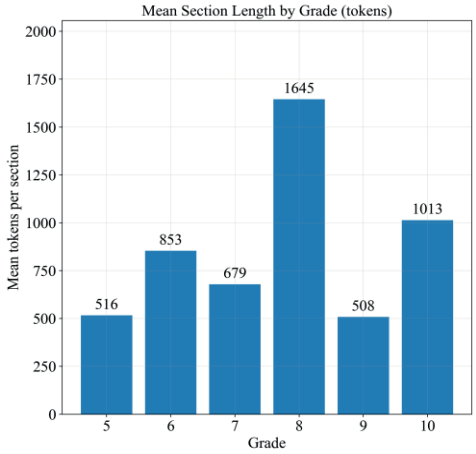

Figure 2. Average section length by grade level (in tokens)

The pronounced peak in section length for grade 8 (1645 tokens) indicates that, without chunking, a non-negligible portion of context would systematically fall outside the model's input window. Such behavior would be methodologically unacceptable for reproducible system comparison.

C. Ontological Artifacts and Graph-Based Knowledge Diagnostics

To support the interpretable knowledge layer and ontology-guided context filtering, an ontological index was constructed from extracted triplets using a confidence threshold of conf_min = 0.40, resulting in 8396 assertions (defs). These assertions form a compact representation of domain facts and relations.

A knowledge graph built from the triplets contains 11 164 nodes and 8396 edges (undirected projection: 11 161 nodes, 7904 edges). The graph exhibits a sparse connectivity pattern typical for educational historical domains, with an average degree of 1.42 (median: 1, 99th

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

percentile: 7, maximum: 54). Despite overall sparsity, the graph includes a large connected component of 1757 nodes alongside many smaller components (n_components = 3408), reflecting a dense thematic core surrounded by a broad periphery of weakly connected entities.

For methodological inspection, two complementary graph views are considered. The degree-core view isolates the highly connected core (2233 nodes, 1295 edges; Fig. 3), demonstrating that the ontology contains connected chains among central historical concepts rather than isolated facts. The community (dense-neighborhood) view highlights locally dense thematic clusters (450 nodes, 556 edges; Fig. 4), emphasizing groups of entities related to specific periods, states, or personalities.

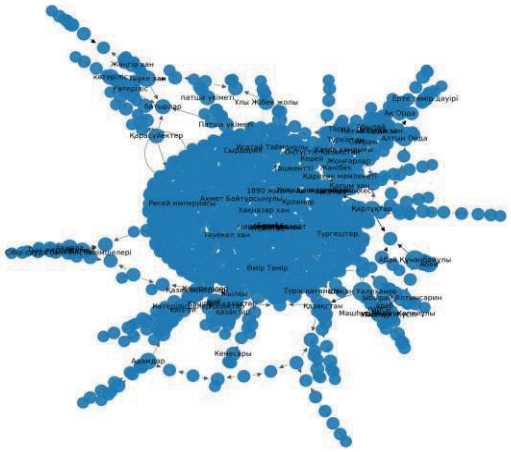

Figure 3. Ontology graph: degree-core (high-connectivity) subgraph

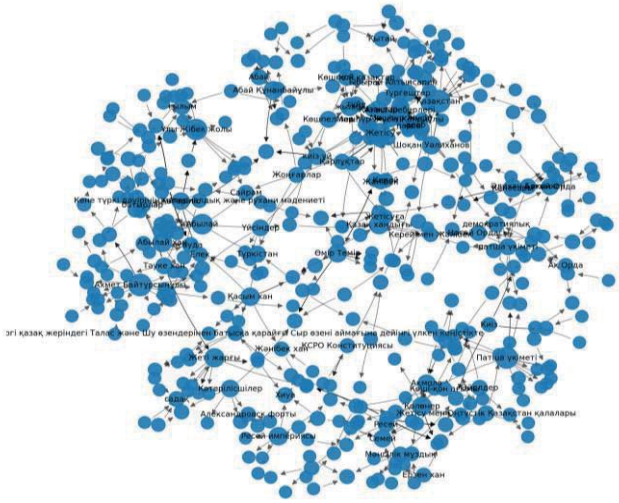

Figure 4. Ontology graph: community-aware (dense-neighborhood) subgraph

Together, these diagnostics indicate that the ontology supports both global connectivity and local density, providing a suitable structural basis for interpretable hybrid QA and for ontology-guided context filtering in a school-level historical corpus.

D. QA Dataset Construction and Consistency Control

The initial pool of question–answer pairs contains 5916 entries. After filtering based on completeness criteria and valid linkage to source sections, 5686 examples were retained. For each question, a strict context mapping was performed using the canonical section key; all mappings are exact (exact_key), ensuring deterministic recovery of contexts from the corpus and eliminating ambiguity in the "question → section" association.

DOI: 10.37943/HRBD3124

To control the quality of answer span extraction with respect to the evidential fragment and/or the section text, several diagnostic categories were evaluated, including cases with available evidence but no exact span match, answers appearing adjacent to the evidence, or the absence of a confident match. These checks serve as an internal validation mechanism and ensure that the resulting datasets are suitable for strict extractive QA settings and fair comparisons.

A key outcome of this consistency control is the absence of span mismatches in the final datasets (Span mismatches = 0) for both section-level and chunk-level representations. This result confirms the internal consistency of the "context–answer" linkage in all datasets used in the experiments.

To support retrieval-style evaluation and to standardize the candidate space across systems, the section texts are transformed into a chunk corpus. Chunks inherit stable identifiers from their parent sections, enabling consistent aggregation while keeping the candidate set fixed across retrieval baselines, neural readers, and hybrid pipelines.

E. QA Systems

This study evaluates three classes of question answering systems designed to operate on the same canonically indexed corpus and under identical experimental constraints. The systems differ only in how the input context is constructed and processed, allowing their behavior to be compared in a controlled manner.

Let a question be denoted as q, and let the set of candidate context chunks be

$$\mathcal{C} = \{c_1, c_2, \dots, c_N\} \tag{1}$$

Each system defines its own effective context set $C' \subseteq C$ and answer selection mechanism.

The NN-only system is a standard neural extractive QA model based on a multilingual transformer encoder. Given a question q and an unfiltered set of context chunks C, the model predicts an answer span by estimating the start and end token positions within each context chunk.

For a given chunk $c \in C$, the neural reader computes:

$$p_{\text{start}}(i \mid q, c), \quad p_{\text{end}}(j \mid q, c) \tag{2}$$

where i and j denote token indices. The predicted answer span is obtained as:

$$(\hat{\imath}, \hat{\jmath}) = \arg\max_{i \le j} p_{\text{start}}(i \mid q, c)\; p_{\text{end}}(j \mid q, c) \tag{3}$$

Long contexts are handled using a sliding-window strategy with fixed window length and stride. The neural reader processes each window independently, and the final answer is selected as the span with the highest confidence score across all windows and all chunks in C.

Training and inference are performed using fixed hyperparameters and identical optimization settings across all neural systems considered in this study. As a result, the NN-only baseline provides a clean reference point for assessing the effect of context construction strategies without interference from architectural or training-related differences.

The Ontology-only system answers questions by relying exclusively on structured knowledge represented as a set of ontological assertions:

$$\mathcal{O} = \{(e_h, r, e_t)\} \tag{4}$$

where e_h and e_t are entities and r denotes a relation.

Given a question q, candidate answers are generated by matching question terms to entities and relations in O. The answer selection process can be expressed as:

$$\hat{a} = \arg\max_{a \in O} \text{sim}(q, a) \tag{5}$$

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

where $sim(\cdot)$ denotes a symbolic or lexical compatibility function.

No neural reader is employed in this setting, and no parameter training is required. This baseline serves as an interpretable reference, isolating the contribution of structured knowledge and enabling a direct comparison with neural and hybrid approaches.

The Hybrid-B system integrates ontology-guided context filtering with neural answer extraction. For a given question q, ontology-based signals — such as matched entities and relations—are used to identify a reduced set of context chunks:

$$\mathcal{C}' = \{c \in \mathcal{C} \mid s_{\text{onto}}(q, c) \geq \tau\} \tag{6}$$

where $s_{\text{onto}}(q, c)$ measures the relevance of a chunk to the question based on ontological evidence, and τ is a fixed threshold.

The neural reader then operates on the filtered context $\mathcal{C}'$ using the same span prediction mechanism as in the NN-only system:

$$(\hat{\imath}, \hat{\jmath}) = \arg\max_{c \in \mathcal{C}', i \leq j} p_{\text{start}}(i \mid q, c)\, p_{\text{end}}(j \mid q, c) \tag{7}$$

Crucially, ontology-guided filtering is applied under a strict NO-GOLD constraint: information about the correct book or section is not available at inference time. If filtering fails to produce a valid context, the system falls back to the unfiltered set $\mathcal{C}$ to ensure coverage.

The neural reader architecture, tokenization scheme, and training protocol are kept identical to those of the NN-only baseline. Consequently, observed differences between NN-only and Hybrid-B can be attributed primarily to context selection rather than to architectural or optimization changes.

F. Experimental Protocol and Evaluation

All experiments are conducted using a unified and reproducible protocol. Preprocessing steps, tokenization parameters, chunking strategy, and evaluation scripts are shared across all systems. Evaluation is performed on a fixed test set using Exact Match (EM) and token-level F1 as primary metrics, complemented by ranking-based measures such as Mean Reciprocal Rank (MRR) and Recall@K.

Let aa denote the ground-truth answer and â the predicted answer.

Exact Match is defined as:

$$EM = \begin{cases} 1, & if\ \hat{a} = a; \\ 0, & otherwise. \end{cases} \tag{8}$$

Token-level precision and recall are computed as:

$$\text{Precision} = \frac{|\text{Tokens}(\hat{a}) \cap \text{Tokens}(a)|}{|\text{Tokens}(\hat{a})|}, \text{Recall} = \frac{|\text{Tokens}(\hat{a}) \cap \text{Tokens}(a)|}{|\text{Tokens}(a)|} \tag{9}$$

The F1 score is the harmonic mean of precision and recall:

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{10}$$

Let $Rq = (d1, d2, \ldots)$ denote the ranked list of candidate chunks for question q, and let rank_q be the position of the first relevant chunk.

Mean Reciprocal Rank is defined as:

DOI: 10.37943/HRBD3124

$$\text{MRR} = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{\text{rank}_q} \tag{11}$$

Recall@K is computed as:

$$\text{RecallK} = \frac{1}{|Q|}\sum_{q\in Q} oI\backslash big\Big(\exists\, d \in \mathcal{R}_q^{(\mathcal{K})}\colon d \text{ is relevant}\backslash big\Big) \tag{12}$$

where $\mathcal{R}_q^{(\mathcal{K})}$ denotes the $top - K$ ranked candidates and $I(\cdot)$ is the indicator function.

Inference latency is measured as the end-to-end processing time per question, including tokenization, model inference, and answer aggregation. It is expressed as:

$$T_{\text{infer}} = T_{\text{token}} + T_{\text{model}} + T_{\text{agg}} \tag{13}$$

Reported latency values correspond to the average over the full test set:

$$\overline{T}_{\text{infer}} = \frac{1}{|Q|}\sum_{q\in Q} T_{\text{infer}}(q) \tag{14}$$

computed after a warm-up phase to ensure stable timing.

To support interpretation, additional sanity checks and ablation experiments are conducted, including random and lexical overlap baselines, as well as length-matched context truncation. These controls help verify that observed improvements are not artifacts of context length or candidate pool size.

Statistical significance of observed differences is assessed using paired testing on per-question scores. Given paired measurements x_q and y_q for two systems, the null hypothesis of no difference is tested on the distribution of differences:

$$\Delta_q = x_q - y_q \tag{15}$$

using an appropriate paired test depending on score distribution assumptions.

Results

This section presents the experimental results obtained under the unified evaluation protocol described in the Methods section. All systems are evaluated on the same fixed test split and using identical metrics, ensuring direct and fair comparison.

A. Overall QA Performance and Retrieval Stability

Table 1 summarizes the main metrics of extractive QA quality, retrieval stability, and inference efficiency.

Table 1. QA quality, retrieval stability, and inference latency

System	EM (%)	F1 (%)	EM@K (%)	F1@K (%)	MRR	Recall@5	Recall@10	Latency (ms)
NN-only (plain context)	68.87	80.02	80.83	92.38	0.870	0.923	0.951	20.01
Hybrid-B (filtered context)	70.45	81.68	82.23	93.92	0.892	0.941	0.969	6.83

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Hybrid-B improves answer quality relative to the NN-only baseline, with +1.58 percentage points in EM and +1.66 percentage points in F1. Retrieval stability also improves, as indicated by higher MRR and Recall@10, meaning that relevant context is more frequently ranked among top candidates. Inference latency is reduced from 20.01 ms to 6.83 ms, corresponding to an approximately 2.9× speedup, which is practically important for interactive educational use.

B. Statistical Significance of Improvements

To assess whether the observed improvements are robust, paired statistical tests were applied on a per-question basis. A paired bootstrap test provides a 95% confidence interval (CI) for the mean difference, and a paired randomization test yields a two-sided p-value. The results of these tests are presented in Table 2.

Table 2. Statistical significance (Hybrid-B − NN-only)

Metric	**Δ mean**	**95% CI**	**p (two-sided)**
EM	+0.0158	[+0.0009, +0.0308]	0.055
F1	+0.0166	[+0.0026, +0.0309]	0.025

The improvement in F1 is statistically significant: the confidence interval is entirely positive and the p-value is below 0.05. For EM, the mean improvement is positive and the confidence interval does not cross zero, but the p-value lies slightly above the conventional threshold. This is consistent with the stricter nature of EM compared to F1. Accordingly, Hybrid-B shows a statistically confirmed gain in F1 and a positive, though less confidently significant, trend in EM.

C. NO-GOLD Inference and Fairness of Filtering

A critical concern for hybrid systems is potential leakage of gold metadata during inference. In this study, ontology-guided filtering is applied under a strict NO-GOLD regime: both book_id and section_id are explicitly set to Noneduring filtering and inference. As a result, the system cannot restrict candidates to the correct source section.

This guarantees that Hybrid-B does not gain an artificial advantage from prior knowledge of the answer location. Its improvements arise from selecting more relevant context based solely on the question text and the ontological index.

D. Ablation Study: Filtering vs. Context Length Reduction

To disentangle the effect of meaningful context selection from simple context shortening, several ablation baselines were evaluated. All comparisons are performed on the same test set. The results of these ablation experiments are summarized in Table 3.

Table 3. Context filtering ablations (n = 1137)

Model	**EM (%)**	**F1 (%)**	**MRR**	**Recall@10**	**Latency (ms)**
NN-only (plain context)	68.87	80.02	0.870	0.951	20.01
NN-only + Ontology-filter (NO-GOLD)	65.61	76.50	0.827	0.893	3.70
NN-only + Lexical-only filter	63.85	74.68	0.808	0.875	5.07
NN-only + Length-matched random	33.25	42.09	0.422	0.480	3.55
Hybrid-B (filtered context, trained)	70.45	81.68	0.892	0.969	6.83

Figure 5 illustrates that Hybrid-B achieves the highest EM and F1 scores, while random length-matched truncation leads to a severe degradation in performance, confirming that context relevance rather than length alone is critical.

DOI: 10.37943/HRBD3124

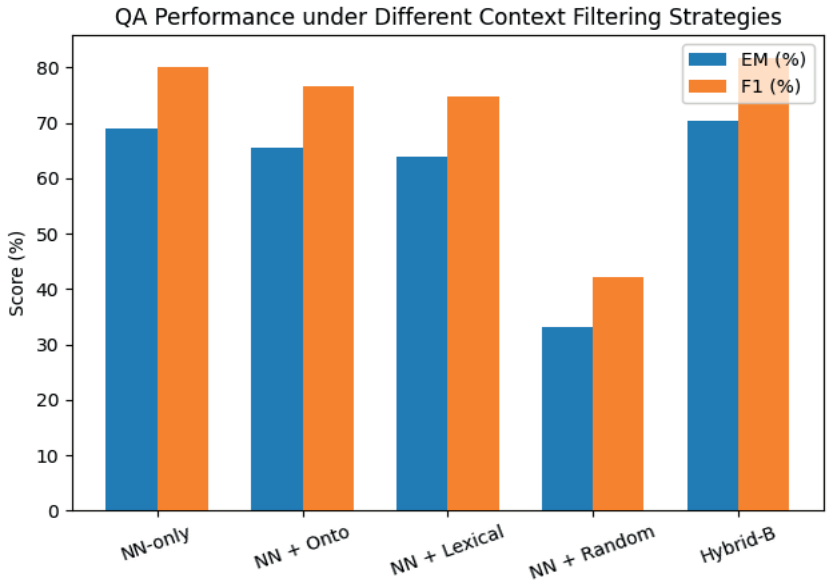

Figure 5. QA Performance under Different Context Filtering Strategies

Random length-matched context causes a severe drop in performance, demonstrating that reducing context length alone is insufficient. Lexical filtering improves speed but substantially degrades accuracy, indicating that surface-level token overlap does not preserve sufficient semantic evidence. Ontology-based filtering without joint training also reduces accuracy, showing that filtering is a strong intervention that can remove useful information. Hybrid-B outperforms all ablations because the neural reader is trained specifically on filtered context, allowing it to adapt to the altered input distribution.

E. Stability with Respect to Random Seeds

To verify robustness, ontology-guided filtering was evaluated under multiple random seeds. The results are reported in Table 4.

Table 4. Stability across filter seeds (n = 1137)

Seed	Onto_F1	Onto_MRR	LenMatch_F1	LenMatch_MRR	Onto_used_rate (%)
42	76.50	0.8267	42.09	0.4219	99.38
7	76.50	0.8267	42.00	0.4329	99.38
123	76.50	0.8267	41.83	0.4277	99.38

Ontology-based filtering exhibits highly stable behavior across seeds, while the length-matched random baseline remains consistently poor. This indicates that the observed effects are not artifacts of favorable random initialization.

F. Performance by Question Type

To better understand where filtering helps or hurts, performance differences were analyzed by question type. The results of this analysis are presented in Table 5.

Table 5. Mean ΔF1 (percentage points) relative to plain baseline

Question type	Count	ΔF1 Onto-filter	ΔF1 Lexical-only	ΔF1 LenMatch random
who	91	+0.56	−5.53	−43.33
where	51	−2.01	−0.58	−37.59
when	221	−3.35	−5.71	−42.06
other	774	−4.16	−5.54	−36.15

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

Ontology-based filtering performs best on “who” questions, consistent with the ontology’s strength in representing entities and person-centric relations. For “where” and “when” questions, filtering alone may remove precise contextual phrasing needed for exact answers. This further supports the central conclusion that filtering alone is insufficient, and that the gains of Hybrid-B arise from the joint design of ontology-guided selection and reader training.

G. Qualitative Analysis of System Behavior

G.1. Dialogue-style Examples of Successful and Failed Predictions

To make the qualitative analysis more interpretable, several representative question-answering cases were examined in a dialogue-style format. Each example compares the expected answer, the NN-only prediction, and the Hybrid-B prediction, followed by a short error analysis. These examples illustrate both the strengths and limitations of ontology-guided filtering.

Example 1. Successful prediction for a “who” question

Question: Қазақ хандығын кімдер құрды?

Expected answer: Керей мен Жәнібек хандар.

NN-only prediction: Қазақ хандығы XV ғасырда құрылды.

Hybrid-B prediction: Қазақ хандығын Керей мен Жәнібек хандар құрды.

In this case, the NN-only model retrieves a thematically relevant context about the establishment of the Kazakh Khanate, but extracts temporal information instead of the required personal names. Hybrid-B performs better because ontology-guided filtering preserves the relation between the Kazakh Khanate and the historical figures associated with its foundation. As a result, the neural reader receives a more focused context and extracts the correct person-centered answer.

Example 2. Successful prediction for a role-identification question

Question: Абылай хан кім болды?

Expected answer: Абылай хан — XVIII ғасырдағы Қазақ хандығының белгілі ханы.

NN-only prediction: Абылай хан жоңғарларға қарсы күресті.

Hybrid-B prediction: Абылай хан — XVIII ғасырдағы Қазақ хандығының ханы.

Analysis. The NN-only model selects a context related to Ablai Khan’s military activity, which is topically relevant but does not directly answer the question “кім болды?”. Hybrid-B selects a more appropriate context because the ontology captures the relation between the entity “Абылай хан” and the role “хан”. This example shows that ontology-guided filtering is particularly useful when the question requires identifying a historical person’s role or status.

Example 3. Successful prediction for a named historical concept

Question: Қасым хан тұсындағы заңдар жинағы қалай аталды?

Expected answer: Қасым ханның қасқа жолы.

NN-only prediction: Қасым хан тұсында Қазақ хандығы күшейді.

Hybrid-B prediction: Қасым хан тұсындағы заңдар жинағы «Қасым ханның қасқа жолы» деп аталды.

The NN-only model retrieves a general passage about Kasym Khan but does not identify the specific legal concept. Hybrid-B narrows the context by using ontology-based relations between Kasym Khan and the legal code. This reduces irrelevant context and allows the neural reader to extract the correct named historical term.

Example 4. Failed prediction for a “when” question

Question: Қазақстан тәуелсіздігін қашан жариялады?

Expected answer: 1991 жылғы 16 желтоқсанда.

NN-only prediction: 1991 жылғы 16 желтоқсанда.

Hybrid-B prediction: 1991 жылы.

In this example, the NN-only model extracts the complete date from the broader context. Hybrid-B produces a partially correct but less precise answer. This suggests that ontology-guided filtering preserved the general context about independence but may have removed or down-ranked the fragment containing the exact day and month. This type of error is especially important for “when” questions because Exact Match requires the full temporal expression.

DOI: 10.37943/HRBD3124

Example 5. Failed prediction for a “where” question

Question: Орбұлақ шайқасы қай жерде болды?

Expected answer: Орбұлақ шатқалында.

NN-only prediction: Орбұлақ шатқалында.

Hybrid-B prediction: Орбұлақ шайқасында.

Hybrid-B correctly identifies the relevant event but fails to extract the precise location. The filtered context remains topically related to the Battle of Orbulak, but it does not preserve the exact locational expression required by the question. This illustrates a limitation of ontology-guided filtering for “where” questions, especially in Kazakh, where place names may appear in different locative or morphological forms.

Example 6. Failed prediction for a broader explanatory question

Question: Жоңғар шапқыншылығы қазақ халқына қалай әсер етті?

Expected answer: Жоңғар шапқыншылығы қазақ халқына ауыр зардап әкелді, халық босқыншылыққа ұшырады, шаруашылық әлсіреді және елдің бірлігіне қауіп төнді.

NN-only prediction: Жоңғар шапқыншылығы қазақ халқына ауыр зардап әкелді.

Hybrid-B prediction: Жоңғарлар қазақ жеріне шабуыл жасады.

Analysis. This example shows a limitation for broader explanatory questions. The expected answer requires several causal and consequence-related elements, such as social disruption, migration, economic weakening, and political instability. Hybrid-B focuses mainly on the event itself and loses part of the broader explanatory context. Therefore, the answer remains thematically relevant but incomplete.

These dialogue-style examples show that ontology-guided filtering improves answer extraction when the required evidence is clearly represented through entities and relations in the ontology. This is especially visible in “who” questions and role-identification questions. However, filtering may reduce answer quality when the answer depends on exact dates, locations, morphological forms, or broader explanatory context. Therefore, the error analysis supports the need for adaptive question-type-aware filtering, where entity-based filtering is used for “who” questions, date-preserving filtering for “when” questions, and location-preserving filtering for “where” questions.

Discussion

The results show that ontology-guided context filtering can improve domain-specific extractive question answering when it is combined with neural reader training. Hybrid-B outperforms the NN-only baseline in EM, F1, MRR, and Recall@10, while also reducing inference latency. Since the neural architecture and training configuration remain unchanged, these improvements can be attributed mainly to more effective context construction.

The strict NO-GOLD setting is an important part of the evaluation design. Since gold book and section identifiers are not used during inference, the improvements cannot be explained by metadata leakage. Instead, Hybrid-B selects relevant context based on the question and the ontological representation.

The ablation results show that the gains are not caused by shorter context alone. Random length-matched truncation and lexical-only filtering reduce performance, while ontology-based filtering without joint reader training also lowers accuracy. This confirms that the best results are achieved when structured context selection and neural answer extraction are combined under the same filtered-context setting.

The qualitative examples further show that ontology-guided filtering is most useful when the required evidence is clearly represented through entities and relations in the ontology. However, filtering may reduce answer completeness when the answer depends on exact dates, locations, morphological forms, or broader explanatory context. Thus, the effectiveness of filtering depends on the alignment between the question type, the ontology, and the evidence needed for answer extraction.

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

From a practical perspective, reducing latency from 20.01 ms to 6.83 ms per question is important for educational QA systems, where fast and stable responses are required. The study also demonstrates the value of reproducible evaluation for Kazakh-language NLP through canonical indexing, fixed data splits, standardized preprocessing, and unified evaluation scripts.

The main limitations are the focus on a single educational domain, the use of one multilingual transformer-based reader, and the construction of the ontology from the same corpus. Future studies should test the approach on other domains, external knowledge sources, and larger QA models.

Taken together, the findings suggest that ontology-guided filtering is an interpretable and efficient context construction strategy for low-resource educational QA. Its value lies in complementing neural models through structured evidence selection rather than replacing them.

Conclusion

This study presented a controlled and reproducible evaluation of ontology-guided hybrid extractive question answering for the Kazakh history domain. The main objective was to examine whether structured knowledge can improve context selection and answer extraction in a low-resource educational setting under a strict NO-GOLD inference scenario.

The results show that the Hybrid-B system achieves moderate but consistent improvements over the NN-only baseline in answer quality, retrieval stability, and inference efficiency. The improvement in F1, together with higher MRR and Recall@10, indicates that ontology-guided filtering helps select more relevant evidence for neural answer extraction. At the same time, the reduction in inference latency demonstrates the practical value of the approach for educational QA applications.

The ablation experiments confirm that these gains are not caused by context shortening alone. Random length-matched truncation, lexical-only filtering, and ontology-based filtering without joint reader training all reduce answer quality. This shows that the effectiveness of Hybrid-B arises from the combination of structured context selection and a neural reader trained under the same filtered-context conditions.

The qualitative analysis further demonstrates that ontology-guided filtering is most beneficial when the required evidence is clearly represented through entities and relations in the ontology. However, its effectiveness may decrease when the answer depends on exact dates, locations, morphological forms, or broader explanatory context. This finding suggests that future work should explore adaptive question-type-aware filtering strategies, including entity-based filtering for “who” questions, date-preserving filtering for “when” questions, and location-preserving filtering for “where” questions.

Taken together, this work contributes not a new neural architecture, but a reproducible methodological framework and empirical evidence on the role of structured knowledge in hybrid QA systems. The findings suggest that ontology-guided filtering can serve as an interpretable and efficient context construction strategy for improving extractive question answering in low-resource educational domains.

References

[1] Pineau, J., Vincent-Lamarre, P., Sinha, K., Larivière, V., Beygelzimer, A., d’Alché-Buc, F., Fox, E., & Larochelle, H. (2021). Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research*, *22*(164), 1–20. http://jmlr.org/papers/v22/20-303.html

[2] Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *379*(2197), 20200210. https://doi.org/10.1098/rsta.2020.0210

[3] Polo, F. M., Izbicki, R., Lacerda, E. G., Ibieta-Jimenez, J. P., & Vicente, R. (2023). A unified framework for dataset shift diagnostics. *Information Sciences, 649*, Article 119612. https://doi.org/10.1016/j.ins.2023.119612

[4] Chen, D., Fisch, A., Weston, J., & Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (pp. 1870–1879). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1171

[5] Clark, C., & Gardner, M. (2018). Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 845–855). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1078

[6] Lin, J., Nogueira, R., & Yates, A. (2021). Pretrained transformers for text ranking: BERT and beyond. *Synthesis Lectures on Human Language Technologies*, *14*(4), 1–325. Morgan & Claypool. https://doi.org/10.2200/S01123ED1V01Y202108HLT053

[7] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (pp. 784–789). Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2124

[8] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. https://doi.org/10.1162/tacl_a_00349

[9] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (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* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

[10] Diefenbach, D., Lopez, V., Singh, K., & Maret, P. (2018). Core techniques of question answering systems over knowledge bases: A survey. *Knowledge and Information Systems*, *55*(3), 529–569. https://doi.org/10.1007/s10115-017-1100-y

[11] Ali, M., Jabeen, H., Hoyt, C. T., & Lehmann, J. (2019). The KEEN Universe: An ecosystem for knowledge graph embeddings with a focus on reproducibility and transferability. In *The Semantic Web – ISWC 2019* (*Lecture Notes in Computer Science*, Vol. 11779, pp. 1–16). Springer. https://doi.org/10.1007/978-3-030-30796-7_1

[12] Tleubayeva, A., & Shomanov, A. (2024). Comparative analysis of multilingual QA models and their adaptation to the Kazakh language. *Scientific Journal of Astana IT University*, *19*, 89–97. https://doi.org/10.37943/19WHRK2878

[13] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W. (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 6769–6781). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.550

[14] Mansurova, A., Tleubayeva, A., Nugumanova, A., Shomanov, A., & Seker, S. E. (2025). A systematic evaluation of large language models and retrieval-augmented generation for the task of Kazakh question answering. *Information, 16*(11), Article 943. https://doi.org/10.3390/info16110943

[15] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7871–7880). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

[16] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutiérrez, C., Kirrane, S., Labra Gayo, J. E., Navigli, R., Neumaier, S., Ngonga Ngomo, A.-C., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., & Zimmermann, A. (2021). Knowledge graphs. *ACM Computing Surveys, 54*(4), Article 71. https://doi.org/10.1145/3447772

Scientific Journal of Astana IT University
ISSN (P): 2707-9031 ISSN (E): 2707-904X
VOLUME 26, JUNE 2026

[17] Song, Y., Zhang, H., Liu, Y., & Tang, J. (2023). Advancements in complex knowledge graph question answering: A survey. *Electronics, 12*(21), Article 4395. https://doi.org/10.3390/electronics12214395

[18] Yu, D., Jiang, Y., Wang, Z., & Zhang, J. (2023). A survey on neural-symbolic learning systems. *Neural Networks, 166*, 364–382. https://doi.org/10.1016/j.neunet.2023.06.028

[19] Bhuyan, B. P., Ramdane-Cherif, A., Tomar, R., & Singh, T. P. (2024). Neuro-symbolic artificial intelligence: A survey. *Neural Computing and Applications, 36*(21), 12809–12844. https://doi.org/10.1007/s00521-024-09960-z

[20] Pires, T., Schlinger, E., & Garrette, D. (2019). How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4996–5001). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1493