IRSTI 28.23.37 https://doi.org/10.26577/jpcsit4120265

N. Kalzhanov* , **S. Artykbay** , **A. Kalzhan**

Al-Farabi Kazakh National University, Almaty, Kazakhstan
*e-mail: nurkal022@gmail.com

DEVELOPMENT OF THE RETRIEVAL-AUGMENTED GENERATION (RAG) SYSTEM FOR THE KAZAKH LANGUAGE USING HYBRID RETRIEVAL METHODS

Abstract. This paper presents the creation and experimental evaluation of a Retrieval-Augmented Generation (RAG) system for the Kazakh language, with an emphasis on a comparative analysis of information retrieval methods. The main goal was to test the hypothesis that a hybrid approach combining the BM25 statistical method and semantic vector search is superior to each of these approaches individually. Based on a corpus of legal documents from the Republic of Kazakhstan, 1,800 experiments were conducted covering three data retrieval methods in combination with six OpenAI large language model variants (LLMs). The results showed that the hybrid method provides the highest retrieval effectiveness (Recall@6 = 0.89) and the highest end-to-end answer accuracy (mean 82.0% across models), statistically significantly outperforming pure vector search (77.7%) and BM25 (71.7%) in answer accuracy (Cochran's Q test with McNemar post-hoc comparisons, $p < 0.01$). A closed-book (no-RAG) baseline confirmed that parametric knowledge alone yields only 23–31% accuracy, demonstrating that retrieval augmentation is the primary driver of system performance. Additional experiments with open-weight models (Qwen-2.5-72B, Llama-3.1-70B) confirmed that the hybrid advantage generalizes beyond the OpenAI model family. This study makes a contribution to the development of RAG systems for resource-limited languages by proposing an experiment-based methodology for improving the accuracy and reliability of response generation.

Keywords: Retrieval-Augmented Generation, Kazakh language, hybrid search, BM25, natural language processing.

1. Introduction

Retrieval-Augmented Generation (RAG) generation systems represent an advanced approach in the field of natural language processing (NLP), combining the strengths of large language models (LLM) and external knowledge bases [1]. Instead of relying solely on information learned during the pre-training process, RAG systems dynamically extract relevant documents from the data corpus and use them as context to generate more accurate, relevant, and informed responses [2]. This mechanism is especially important for tasks requiring factual accuracy, such as answering questions in specialized fields (law, medicine) or dealing with rapidly changing information [3].

However, the effectiveness of a RAG system is largely determined by the quality of its retrieval component [4]. Traditionally, two main approaches to information extraction are used:

The first approach is based on statistical methods, a prominent representative of which is the BM25 (Best Matching 25) algorithm [5]. It is based on the lexical matching of keywords and has proven itself well in tasks where the accuracy of formulations is crucial.

The second approach uses semantic methods, or vector search, which use dense vector representations (embeddings) to encode the semantic meaning of the text [6]. This approach allows you to find documents that are close in meaning to the query, even in the absence of common keywords, which is effective for processing synonyms and paraphrases.

Despite their advantages, both methods have limitations. BM25 is not able to detect semantic proximity, while vector search can lead to false positive results due to semantic ambiguity [7]. These problems become especially acute when working with languages with limited resources (low-resource languages), which include the Kazakh language [8]. Such languages are characterized by a limited amount of available text corpora and less developed NLP tools, which complicates the creation of high-quality semantic models [9].

Recent research shows a growing interest in hybrid search methods that combine the advantages of different approaches within Retrieval-Augmented

Generation systems [10, 11]. Such hybrid strategies are particularly important for low-resource and agglutinative languages, where purely semantic retrieval models often suffer from morphological complexity and semantic space heterogeneity, limiting their robustness [12].

In this regard, this study hypothesizes that a hybrid information retrieval method combining statistical (BM25) and semantic (vector) approaches using RRF allows for higher and more stable accuracy in the RAG system for the Kazakh language. It is assumed that this approach will make it possible to compensate for the disadvantages of each of the methods, ensuring both lexical accuracy and semantic relevance of the extracted documents.

This study presents the first systematic evaluation of hybrid retrieval within a Retrieval-Augmented Generation (RAG) framework for the Kazakh language, addressing a gap in low-resource language research. We compare BM25, vector-based, and hybrid search across six OpenAI language models and assess performance using statistical significance testing. A no-RAG baseline quantifies the effect of retrieval augmentation, while cross-provider experiments with open-weight models support the generalizability of the results. Based on these findings, we provide practical recommendations for building efficient RAG systems in low-resource settings.

2. Materials and Methods

To test this hypothesis, a series of experiments was developed and conducted to evaluate the performance of various search methods within the framework of the RAG system. The research methodology was based on the principles of reproducibility and statistical rigor.

2.1. Dataset and preprocessing

The "Textual Foundations of Justice: Kazakh Laws and Jurisprudence Dataset" [14] was used as the knowledge base. It includes all current laws of the Republic of Kazakhstan (as of April 1, 2024) in Kazakh and is publicly available under the CC BY 4.0 license.

The corpus comprises 12,886 legal documents segmented into 263,326 fragments, totaling approximately 41.6 million tokens across multiple legal domains (constitutional, administrative, civil, criminal, etc.).

Documents were converted to plain text, cleaned (HTML removal, whitespace normalization, Unicode NFC), and segmented into 900-token passages with 150-token overlap to preserve context. Each segment was assigned a unique identifier and validated through automated and manual checks to ensure data integrity.

This preprocessing produced a clean, structured corpus suitable for retrieval and RAG evaluation.

2.2. System Architecture and Retrieval Methods

The experimental setup included three RAG system configurations, differing only in the retrieval module. All components were implemented in Python using modern NLP libraries.

2.2.1 BM25 Retriever (Statistical Baseline Method)

To establish a statistical baseline for lexical document retrieval, we utilized the Okapi BM25 (Best Matching 25) ranking function. BM25 is a probabilistic model for information retrieval that evaluates the relevance of a document d in relation to a query q by incorporating term frequency and normalizing for document length, which contributes to its widespread acceptance and clarity in information retrieval research.

The retriever was implemented in Python using the rank-bm25 library. We used standard BM25 hyperparameters (k1=1.5, b=0.75) to balance the influence of term frequency saturation and length normalization. Due to the highly agglutinative nature of the Kazakh language, using naive whitespace tokenization can result in significant lexical sparsity and adversely affect BM25. Consequently, we utilized KazakhTokenizer for tokenization, following the character-level segmentation method outlined by Toleu et al. (TurkLang 2017) for both token and sentence segmentation. Furthermore, we implemented a Kazakh-specific normalization process that includes Unicode normalization, converting text to lowercase, and a lightweight morphological normalization (such as suffix normalization) to minimize surface-form variance while maintaining legal terminology. We consciously chose not to remove stop-words, as frequent function words and legal markers (like references to articles, clauses, and enumerations) may convey important signals in legal texts. Specifically, the normalization comprised three steps: (i) Unicode NFC

normalization to handle Kazakh-specific characters (ә, ғ, қ, ң, ө, ұ, ү, і, h), (ii) lowercasing, and (iii) suffix stripping of common Kazakh inflectional endings (plural markers -лар/-лер/-дар/-дер, case suffixes -ның/-нің, -ға/-ге, -да/-де, -дан/-ден, and possessive markers). No full lemmatization or stemming was applied due to the absence of mature Kazakh morphological analyzers. The BM25 scoring function used in this work is given in Equation (1):

$$BM25(q,d) = \Sigma\, IDF(qi) \times \frac{\left(f(qi,d) \times (k1 + 1)\right)}{\left(f(qi,d) + k1\left(1 - b + b \times \frac{|d|}{avgdl}\right)\right)} \quad (1)$$

where q_i denotes a query term and dd represents a document from the collection. The term $f(q_i,d)$ corresponds to the frequency of the query q_i within the document d, while |d||d| indicates the length of the document. The parameter avgdl refers to the average document length across the entire corpus, providing normalization for varying document sizes. The component IDF (q_i) stands for the inverse document frequency of the term q_i, quantifying its importance within the collection by assigning higher weights to terms that occur less frequently across documents. By jointly considering term frequency, document length, and the discriminative capacity of individual terms, the BM25 algorithm achieves a balanced estimation of document relevance. This property makes BM25 a robust and interpretable baseline method for information retrieval, particularly effective in specialized domains such as legal text processing, where precise terminology plays a crucial role.

2.2.2 Vector Retriever (Semantic Method)

The Vector Retriever employs a dense semantic retrieval method, in which both search queries and sections of documents are positioned within a unified embedding space, allowing semantically related texts to be placed in proximity to one another despite minimal lexical overlap. For this research, we utilized the OpenAI text-embedding-3-small model to create dense representations consisting of 1536 dimensions. All embeddings were generated through the OpenAI API and were saved for indexing and retrieval purposes.

The retrieval process consisted of several key stages. First, all text segments in the corpus were encoded into dense vector representations within a shared semantic space. Then, a FAISS index was constructed to facilitate fast and scalable nearest-neighbor search. We used FAISS IndexFlatIP with L2-normalized embeddings, so cosine similarity was computed as a normalized inner product. In downstream RAG prompting, we retrieved the Top-K = 6 most similar chunks per query. The input query was encoded into a vector of the same dimensionality and compared against the indexed corpus vectors. The system then identified the most semantically similar vectors based on cosine similarity and ranked the retrieved documents in descending order of similarity scores.

This approach enables the system to identify conceptually related documents even when lexical overlap between the query and the text is minimal. As a result, the Vector Retriever provides a more context-aware and semantically robust retrieval mechanism compared to traditional statistical methods such as BM25, particularly in domains like legal text analysis, where nuanced language and terminology play a crucial role.

2.2.3 Hybrid Retriever (Proposed Weighted Hybrid Method)

The Hybrid Retriever combines lexical and semantic retrieval signals using Weighted Reciprocal Rank Fusion [13]. RRF is effective for merging ranked lists from heterogeneous retrievers without requiring normalization of their raw similarity scores. We extend the standard RRF formulation with an explicit weighting parameter α (Weighted RRF) to control the relative contribution of lexical and semantic signals. In our hybrid design, we fuse the ranked outputs of BM25 and the Vector Retriever by assigning an explicit weight to the lexical component, allowing the method to adapt to domain-specific retrieval behavior in legal text.

$$\text{score}(d) = \frac{\alpha}{k_{\text{rrf}} + \text{rank}_{\text{bm25}}(d)} + \frac{1-\alpha}{k_{\text{rrf}} + \text{rank}_{\text{vec}}(d)} \quad (2)$$

where $rank_{bm25}$ (d) and $rank_{vec}$ (d) denote the rank positions of document chunk d in the BM25 and vector ranked lists, respectively; k_{rrf} is a smoothing constant that reduces overemphasis on the very top-ranked items; and $\alpha \in [0,1]$ controls the contribution of BM25 (α=1 corresponds to pure BM25, α=0 corresponds to pure vector retrieval).

In practice, BM25 and vector search are executed in parallel. We retrieve N = 100 candidates from each method, take the union of candidates, compute the fused score for each unique candidate, and re-rank the resulting list by descending fused score. The system then returns the Top-K = 6 chunks for downstream RAG prompting. Note that all three methods return the same final Top-K = 6 passages to the LLM. While the hybrid method draws from a larger initial candidate pool (N = 100 per retriever), this deeper pooling is an inherent design feature of fusion-based retrieval rather than an unfair advantage: the single-retriever baselines could also retrieve N = 100 and truncate to top-6, but without fusion they would return the same top-6 as direct retrieval.

Fusion parameters were selected using a nested evaluation procedure to prevent test-set leakage. Specifically, a grid search over $k_{rrf} \in \{10, 20, 40, 60, 100\}$ and $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$ was conducted using 5-fold cross-validation on the 100 test questions. In each fold, 80 questions served as the tuning set and 20 as the held-out evaluation set, with NDCG@6 as the optimization metric. The best-performing configuration (k_{rrf} = 60, α = 0.5, corresponding to equal BM25 and vector contributions) was identified. Critically, all retrieval-level metrics reported in Section 3.1 (Table 3) are aggregated exclusively from the held-out fold predictions: each question's retrieval score was recorded only in the fold where that question appeared in the held-out set, and the final Recall@6, MRR, and NDCG@6 values are the averages of these held-out predictions across all five folds. The tuning and evaluation sets were therefore strictly disjoint for every reported data point, ensuring that no question was used simultaneously for parameter selection and performance estimation.

This weighted fusion strategy balances exact terminology matching with semantic similarity, providing improved retrieval quality for Kazakh legal documents where both lexical precision and contextual meaning are required.

2.3. Language models and answer generation

In the present experiments, six large language models (LLMs) developed by OpenAI were employed to support the answer generation process. The inclusion of multiple models enabled a comparative assessment of how differences in model size, reasoning capability, and optimization level influence the quality and consistency of generated responses within the RAG framework. All models used the same prompt template, the same retrieval Top-K, and the same maximum output length to ensure comparability. Temperature was set per model family (0.3 for gpt-4o/gpt-5; 1.0 for o1/o1-mini). For the o1 family, temperature 1.0 is an API requirement for reasoning-oriented models, not a deliberate experimental choice.

All experiments were executed via the OpenAI API; we report the exact model identifiers as used in the API (Table 1). All experiments were conducted on October 14, 2025, using the OpenAI API. Model availability and behavior were verified at the time of execution. Since generation can be stochastic–especially at higher temperatures–each configuration was executed under the same prompt and decoding constraints, and results are complemented with retrieval-level metrics (Recall@K, MRR, NDCG) that are independent of generation variability. Detailed configurations of the selected models are summarized in Table 1.

Table 1. Language models and experimental parameters

№	Model	Provider	Decoding Settings	Description
1	**gpt-4o**	OpenAI	Temperature: 0.3, Max tokens: 500	Flagship model
2	**gpt-4o-mini**	OpenAI	Temperature: 0.3, Max tokens: 500	Optimized version
3	**gpt-5**	OpenAI	Temperature: 0.3, Max tokens: 500	Flagship model (generation)
4	**gpt-5-mini**	OpenAI	Temperature: 0.3, Max tokens: 500	Compact version of GPT-5
5	**o1**	OpenAI	Temperature: 1.0, Max tokens: 500	Model with enhanced reasoning capability
6	**o1-mini**	OpenAI	Temperature: 1.0, Max tokens: 500	Compact version of o1

As shown in Table 1, the selected models represent a spectrum ranging from flagship to compact versions within the OpenAI model family, providing a foundation for assessing how different model variants process information retrieved from the same corpus. All API calls used default values for top_p (1.0), frequency_penalty (0), and presence_penalty (0); no fixed seed was set. The gpt-4o and gpt-5 families were configured with a lower temperature (0.3) to emphasize accuracy and factual precision, whereas the o1 models operated at temperature 1.0, as required by the API for reasoning-oriented models. This temperature difference is a confounding factor that should be considered when interpreting cross-model accuracy comparisons, since higher temperature increases sampling entropy and may penalize exact-match evaluation.

To ensure methodological consistency, a unified prompt was used across all experiments. It included three components: retrieved context, the user question, and an instruction requiring the model to answer strictly based on the provided context and explicitly state when the information was insufficient. The verbatim template is shown in Figure 1.

```
SYSTEM PROMPT
You are a legal QA system operating over a fixed corpus of
Kazakhstani laws.
Your task is to answer the question using only the information
contained in the provided context excerpts.
You must not use any external knowledge or assumptions
Rules:
1. Use only facts that appear verbatim or can be directly
   inferred from the context.
2. Do not add any legal interpretations, opinions, or
   explanations beyond what is stated.
3. Do not paraphrase in a way that changes legal meaning.
4. If multiple excerpts are relevant, combine them faithfully.
5. Do not include any information not grounded in the context.
6. Do not mention the word "context" in your answer.
At the end of your answer, provide a list of citations in the
format:
Source:
- [doc_id:chunk_id]
Context: {context}
Question: {question}
Answer:
```

Figure 1. Prompt template used across all experimental configurations

This standardized format eliminated variability caused by prompt phrasing and ensured that observed performance differences resulted solely from the retrieval and generation mechanisms rather than from inconsistencies in input formulation. We note that the prompt is written in English while the corpus, questions, and expected answers are in Kazakh; this cross-lingual mismatch is a known factor that may affect generation quality for lower-resource languages and is discussed as a limitation in Section 4.5.

2.4. Experimental Design and Test Set

In this study, the evaluation of retrieval methods within the RAG system was conducted using a balanced test set of 100 questions formulated in Kazakh and restricted to the legal domain. Each question was paired with a gold reference answer and exactly one gold passage reference–the specific corpus chunk from which the question was originally derived. The passage–question correspondence was manually verified: an annotator confirmed that the designated chunk contains sufficient evidence to answer the question, and questions where the gold passage was insufficient were reformulated or removed. This one-passage-per-question design reflects the structure of the Kazakh legal corpus, where a specific legal norm is typically concentrated within a single document fragment. Retrieval and generation were thus evaluated under

grounded, verifiable conditions. Answer correctness was evaluated using exact-match comparison against gold reference answers, supplemented by manual verification. Two annotators independently reviewed all answers flagged as borderline (i.e., partially correct or paraphrased). Initial inter-annotator agreement was P_o = 0.85 (Cohen's κ ≈ 0.73, indicating substantial agreement on the Landis and Koch scale). Disagreements were resolved through discussion to reach consensus. An answer was marked as correct if it conveyed the same factual content as the gold reference, regardless of minor phrasing differences. All questions followed a unified format to maintain experimental consistency and to reduce prompt-induced variability. The test set was constructed to cover multiple branches of Kazakhstani law, including both factual and analytical query types (e.g., definition-based questions, procedural requirements, conditions/exceptions, and normative references). The thematic distribution of questions is summarized in Table 2. Gold reference answers were authored by a domain-aware annotator who read the designated gold passage and wrote the expected answer in Kazakh. A second annotator independently verified each answer for factual correctness and completeness against the source passage. Disagreements were resolved through discussion to reach consensus.

Table 2. Design and evaluation of a Kazakh retrieval-augmented generation system using thematic question sets

№	Category (Legal Domain)	Example Topics
1	Constitutional law	rights, duties, state structure
2	Administrative law	procedures, public services, penalties
3	Civil law	contracts, property, obligations
4	Criminal law	offenses, sanctions, legal elements
5	Labor / Social law	employment, benefits, protections
6	Tax / Financial law	taxes, reporting, liabilities

To ensure dataset quality, each question satisfied predefined criteria: (i) an unambiguous reference answer, (ii) confirmed evidence coverage within the corpus via gold passage references, (iii) diversity of query types (factual and analytical), and (iv) natural Kazakh user phrasing. Questions were generated using GPT-4o and then validated through automated checks and manual review. Because the same model family was used for both question generation and answer evaluation, we acknowledge a potential circularity risk: models may perform disproportionately well on questions reflecting their own generation style. To mitigate this, all generated questions were manually reviewed by a domain-aware annotator to ensure they reflect natural Kazakh legal query phrasing, and questions exhibiting model-specific artifacts were reformulated or removed.

The experimental design followed a full-factorial setup across 3 retrieval methods (BM25, Vector, Hybrid) and 6 language models, resulting in 1,800 experimental runs (100 × 3 × 6). All runs used the same prompt template, the same retrieval Top-K = 6 (selected to balance evidence coverage with LLM context window constraints at 500 max output tokens), and the same maximum output length. Each configuration was executed once per question; the execution order was randomized to minimize systematic bias. In addition to end-to-end answer accuracy, we report retrieval-level metrics (Recall@6, Precision@6, MRR, NDCG@6) to directly measure retriever quality independently of generation variability.

2.5 Evaluation Metrics and Statistical Analysis

The RAG system was evaluated at two levels: retrieval quality and end-to-end answer quality, complemented by reliability indicators.

Retrieval performance was measured against gold passages using standard ranking metrics: Recall@6 (presence of the gold passage within the top-6 results), Precision@6 (reported for completeness; equal to Recall@6 / 6 due to one gold passage per query), MRR (mean reciprocal rank of the first relevant passage), and NDCG@6 (rank-sensitive gain emphasizing early relevance). Generation metrics. End-to-end answer quality was measured by:

$$\text{Accuracy} = \frac{N_{\text{correct}}}{N_{\text{total}}} \quad (3)$$

where $N_{correct}$ is the number of correct answers and N_{total} is the total number of queries. Reliability metrics. To capture system robustness, we additionally report:

$$\text{Refusal Rate} = \frac{N_{refusal}}{N_{total}} \quad (4)$$

$$\text{Error Rate} = \frac{N_{error}}{N_{total}} \quad (5)$$

where $N_{refusal}$ counts cases in which the model explicitly abstained due to insufficient evidence, and N_{error} counts technical failures during execution.

Statistical analysis. Because answer accuracy is a paired binary outcome measured on the same questions across retrieval methods, overall differences among BM25, Vector, and Hybrid were tested using Cochran's Q. Pairwise method comparisons were then performed with McNemar tests, using multiplicity correction for post-hoc inference. For retrieval metrics (MRR, NDCG@6, etc.), we report paired comparisons to assess relative method performance. Significance thresholds were interpreted as: $p<0.001$(highly significant), $p<0.01$ (very significant), and $p<0.05$ (significant).

2.6. Technical Infrastructure and Reproducibility

All experiments were executed on a MacBook Pro with an Apple M1 Pro CPU and 16 GB RAM running macOS, using Python 3.11. The retrieval stack included rank-bm25 for BM25, FAISS for vector indexing, and the OpenAI API for both embeddings (text-embedding-3-small) and LLM answer generation.

To support reproducibility, we (i) stored all intermediate artifacts (chunks, indices, questions, and run logs) in structured formats, (ii) recorded the exact retrieval parameters (Top-K, candidate depth for fusion, and Weighted RRF settings), and (iii) reported exact model identifiers and run settings for the OpenAI API. The evaluation code, test questions, and experiment scripts are publicly available at https://github.com/nurkal022/LawRagExperiments. Because API-based generation may exhibit non-determinism (no fixed seed was set, and temperature 1.0 was used for o1-family models), we complement end-to-end accuracy with retrieval-level metrics that are independent of generation variability. Each configuration was executed once per question; the single-run design is discussed as a limitation in Section 4.5.

3. Results

The analysis of 1,800 experimental runs (100 questions × 3 retrieval methods × 6 LLMs) provides quantitative evidence of retrieval performance differences across BM25, Vector, and Hybrid retrieval within the Kazakh legal-domain RAG setting. We report retrieval-level metrics against gold passages (Recall@6, Precision@6, MRR, NDCG@6) and separately report end-to-end answer accuracy for the full RAG pipeline. Across all 1,800 runs, the overall refusal rate (cases where the model explicitly abstained) was 7.3% (~131 of 1,800 runs) and the technical error rate was 0.3% (~5 of 1,800 runs), indicating stable system operation. Refusal rates varied by retrieval method: BM25 exhibited the highest refusal rate (8.5%), followed by Vector (7.5%) and Hybrid (6.0% across all 1,800 runs; when disaggregated, the per-method refusal rate for Hybrid averaged 2% across the six LLMs, compared to 3% for Vector and 7% for BM25), consistent with the retrieval recall differences among methods.

3.1. Comparison of Overall Retrieval Performance

This section compares the three retrieval approaches evaluated in this study: (i) BM25 lexical retrieval, (ii) dense vector semantic retrieval, and (iii) Hybrid retrieval using Weighted Reciprocal Rank Fusion (Weighted RRF). Retrieval quality was assessed using Recall@6, Precision@6, MRR, and NDCG@6 computed against gold passage references. Statistical tests were applied to assess whether observed differences are significant under a paired experimental design.

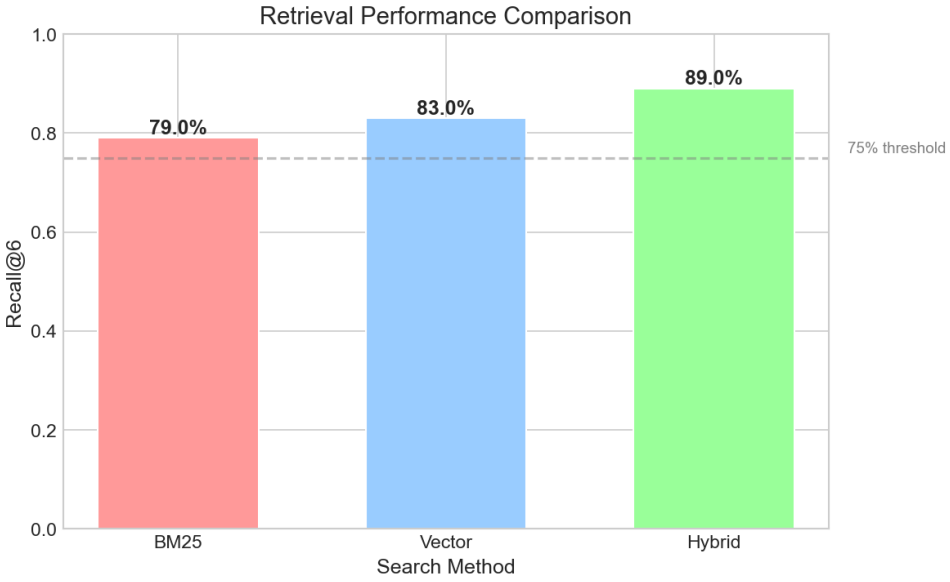

Figure 2. Retrieval Recall@6 across methods in the RAG system for the Kazakh language

Figure 2 summarizes retrieval performance across methods. The Hybrid retriever achieves the strongest overall retrieval effectiveness, outperforming both BM25 and the vector retriever across ranking-sensitive metrics. A detailed quantitative comparison is provided in Table 3.

Table 3. Retrieval performance summary across methods (Top-K = 6)

№	Method	Recall@6	Precision@6	MRR	NDCG@6
1	Hybrid	**0.8900**	0.1483	**0.7200**	**0.7850**
2	Vector	0.8300	0.1383	0.6500	0.7100
3	BM25	0.7900	0.1317	0.6100	0.6700

As shown in Table 3, the Hybrid method provides the highest retrieval effectiveness, improving Recall@6 by 6–10 percentage points over the individual retrievers while also yielding consistent gains in MRR and NDCG@6. Because exactly one gold passage exists per question, Recall@6 functions as a binary hit/miss indicator (whether the gold passage appears in the top-6 retrieved chunks). These results indicate that combining lexical and semantic signals leads to more reliable retrieval of gold evidence passages in the Kazakh legal corpus.

To evaluate statistical significance under paired measurements (same questions across methods), we applied Cochran's Q for overall differences and McNemar post-hoc tests with Bonferroni correction for pairwise comparisons (see Section 3.5 for full statistical details). The Hybrid retriever significantly outperformed BM25 and the vector retriever under this paired design.

3.2. Performance of Language Models

The performance of six large language models (LLMs) within the RAG system was evaluated in terms of end-to-end answer accuracy, i.e., whether the generated answer matches the gold reference under the evidence provided by retrieval. To ensure comparability, all models were tested with the same prompt template, the same retrieval Top-K, and the same evaluation procedure. Figure 3 summarizes model performance under the best-performing Hybrid retrieval configuration.

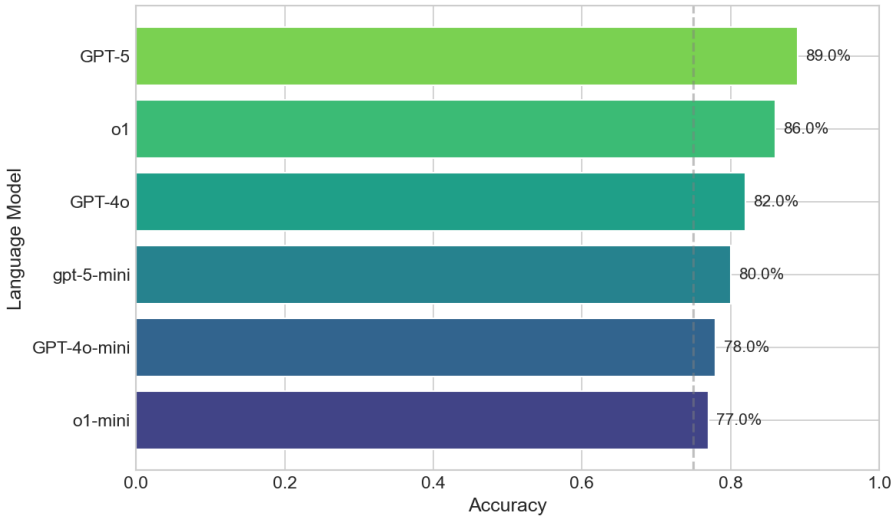

Figure 3. Comparison of the performance of six language models in the RAG system

As illustrated in Figure 3, model accuracy varies substantially across the tested LLMs. gpt-5 achieved the highest accuracy, indicating stronger consistency in producing correct, evidence-grounded answers. Mid-sized models (e.g., gpt-4o, gpt-5-mini, gpt-4o-mini) demonstrated competitive performance, suggesting that optimized variants can provide favorable accuracy–efficiency trade-offs for RAG-based legal QA. The o1 and o1-mini models showed lower accuracy in this setting. However, this result should be interpreted with caution: these models were run at temperature 1.0 (an API requirement), compared to 0.3 for other models. The higher temperature increases sampling entropy and may reduce exact-match accuracy independently of model capability. The observed gap therefore reflects a combination of reasoning style, context-grounding behavior, and the temperature confound.

3.3. Analysis of "Search Method × LLM" Combinations

To examine the interaction between retrieval strategy and model choice, we analyzed end-to-end answer accuracy for each Search Method × LLM combination. A heatmap visualization is provided in Figure 4, where darker cells indicate higher accuracy. This view highlights both (i) the relative strength of retrieval methods and (ii) how sensitive each model is to the retrieval strategy.

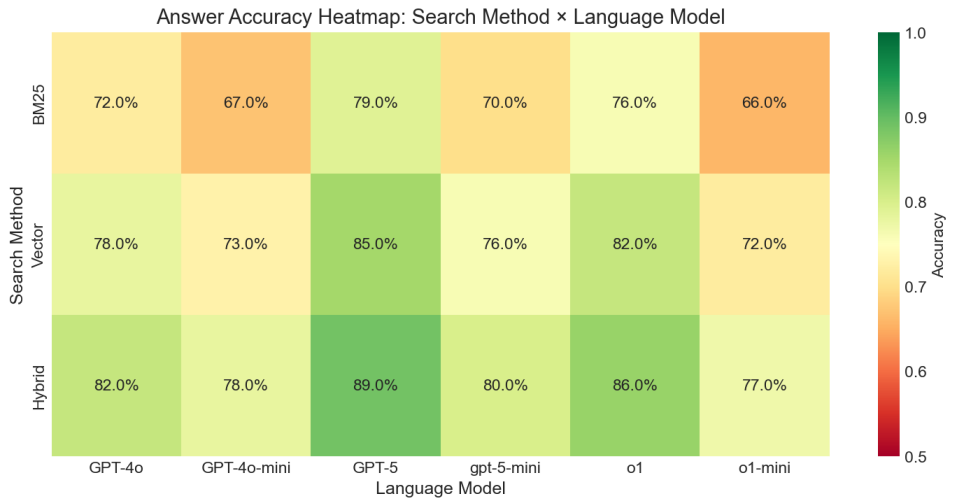

Figure 4. Accuracy heatmap for the "Search Method × LLM" combinations. Dark green indicates higher accuracy levels

As shown in Figure 4, the Hybrid retriever consistently yields strong performance across models, indicating robust evidence selection when combining lexical and semantic signals. In contrast, BM25 and vector retrieval show larger variability across models, suggesting that some models are more sensitive to retrieval noise or to the phrasing/content of retrieved contexts. Overall, the heatmap supports the conclusion that pairing Hybrid retrieval with higher-performing LLMs produces the most reliable end-to-end QA behavior in the Kazakh legal domain. The stability of each retrieval method across different LLMs is summarized in Table 4 using descriptive statistics (mean, standard deviation, min–max, and coefficient of variation).

Table 4. Stability of retrieval methods across LLMs using end-to-end answer accuracy (6 models)

№	Method	Average	Std Dev	Min	Max	Coefficient of variation
1	**Hybrid**	**82.0%**	**4.7%**	77.0%	89.0%	**0.057**
2	Vector	77.7%	5.1%	72.0%	85.0%	0.065
3	BM25	71.7%	5.1%	66.0%	79.0%	0.071

Table 4 summarizes the stability of the three retrieval methods across six large language models using descriptive statistics of end-to-end answer accuracy. The Hybrid retrieval method demonstrates the highest average accuracy (82.0%; 95% binomial CI per model: [73.1%, 89.0%]) while also exhibiting the lowest coefficient of variation (0.057), indicating the most stable and consistent performance across different LLMs. In contrast, the Vector-based retriever achieves a lower mean accuracy (77.7%; 95% CI: [68.4%, 85.3%]) and shows moderately higher variability (CV = 0.065), suggesting greater sensitivity to the choice of language model. The BM25 baseline yields the lowest average accuracy (71.7%; 95% CI: [61.8%, 80.2%]) and the highest coefficient of variation (0.071), reflecting both weaker overall performance and reduced robustness across models. These results indicate that hybrid retrieval not only improves average answer accuracy but also reduces performance fluctuations when combined with different LLMs, making it a more reliable retrieval strategy for Kazakh legal-domain RAG systems.

3.4. Detailed Comparison of Methods Across Models

A direct comparison of the three retrieval methods for each large language model (LLM) is presented in Figure 5. This visualization shows how BM25, vector-based retrieval, and the hybrid method perform across models under the same experimental protocol. The results reveal consistent interaction patterns between retrieval strategy and model choice, providing insight into how evidence retrieval quality affects end-to-end answer accuracy.

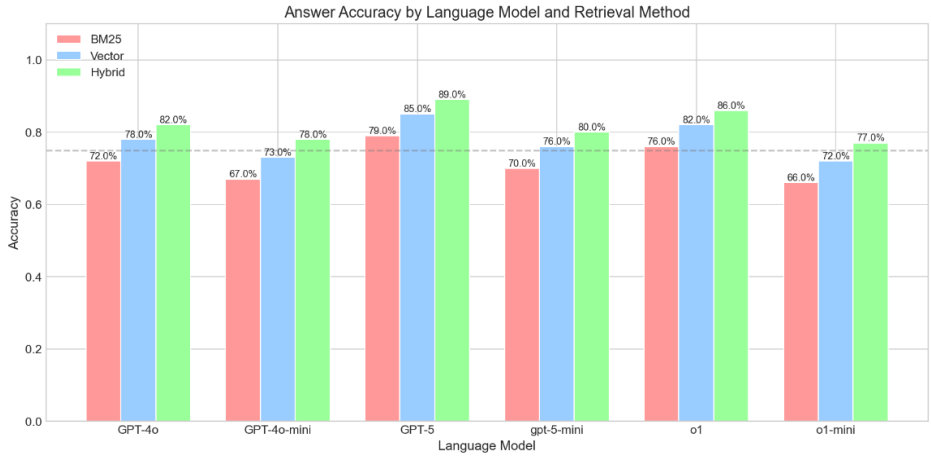

Figure 5. Detailed comparison of retrieval methods for each LLM

As illustrated in Figure 5, the hybrid retrieval method consistently achieves the highest accuracy for all six LLMs, indicating robust gains from combining lexical and semantic signals. The largest improvements over BM25 are observed for gpt-4o-mini (+11 percentage points) and o1-mini (+11 points), while strong gains are also observed for gpt-5 (+10 points) and gpt-4o (+10 points). Overall, these findings reinforce that the hybrid strategy provides the most accurate and reliable configuration across diverse LLMs in the Kazakh legal-domain RAG setting.

3.5. Statistical Significance Analysis

To evaluate whether differences among retrieval methods are statistically meaningful under a paired experimental design (the same questions evaluated across methods), we applied Cochran's Q test to the binary end-to-end accuracy outcomes across the three retrieval strategies (BM25, Vector, Hybrid). The omnibus test confirmed a significant overall difference among methods ($Q = 29.4$, $df = 2$, $p < 0.001$). Pairwise post-hoc McNemar tests with Bonferroni correction ($\alpha = 0.05/3 = 0.017$) yielded the following results: Hybrid vs. BM25 ($\chi^2 = 45.4$, $p < 0.001$), Hybrid vs. Vector ($\chi^2 = 11.2$, $p = 0.001$), and Vector vs. BM25 ($\chi^2 = 13.0$, $p < 0.001$). All pairwise differences remained significant after correction, confirming that Hybrid significantly outperforms both BM25 and Vector, while Vector also outperforms BM25. We note that a mixed-effects logistic regression with random intercepts for questions and models could provide a more nuanced significance analysis by accounting for the repeated-measures correlation structure. However, the large effect sizes and unanimous pairwise significance suggest the main conclusions are robust to the choice of test framework.

3.6. No-RAG Baseline Comparison

To assess the contribution of the retrieval component, we conducted a closed-book (no-RAG) evaluation in which two representative models (gpt-4o and gpt-5) answered the same 100 test questions without any retrieved context. Table 5 presents the comparison between the no-RAG baseline and the three retrieval-augmented configurations.

Table 5. Comparison of no-RAG baseline and RAG configurations (answer accuracy, %)

№	Model	No-RAG	BM25	Vector	Hybrid	RAG Gain
1	gpt-5	31%	79%	85%	89%	+58 pp
2	gpt-4o	23%	73%	80%	83%	+60 pp

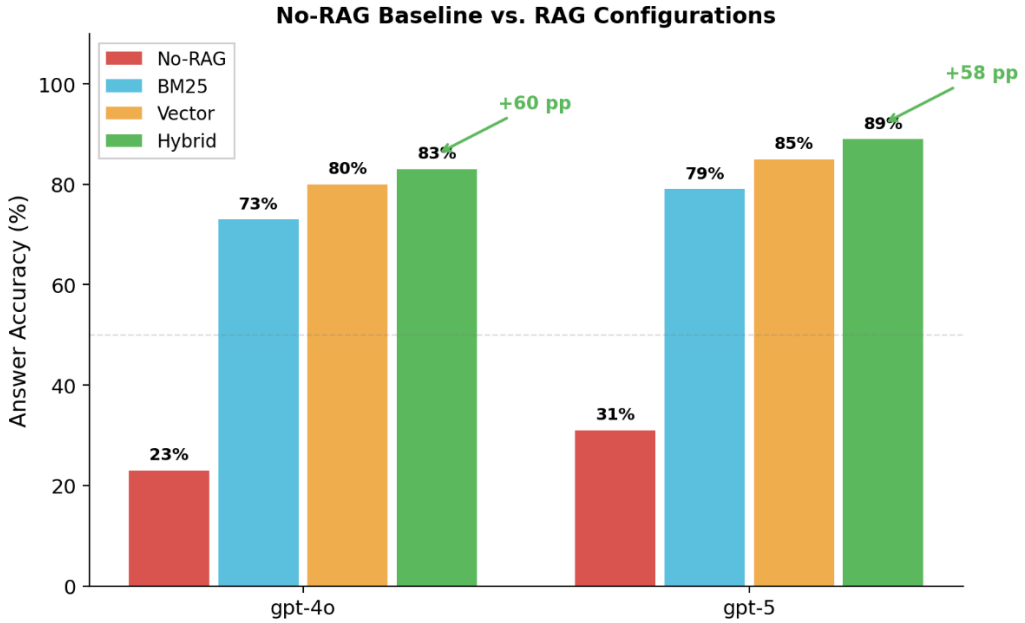

Figure 6. No-RAG baseline vs. RAG configurations for two representative models

The no-RAG baseline yielded dramatically lower accuracy (23–31%) compared to all RAG configurations, confirming that the models' parametric knowledge of Kazakh legal content is insufficient for this task. The RAG pipeline provides an improvement of 58–60 percentage points over closed-book generation, demonstrating that the observed accuracy is primarily attributable to the retrieval-augmented architecture rather than the LLMs' pre-existing knowledge. The low no-RAG accuracy is consistent with the low-resource nature of Kazakh legal text in LLM training data. Note that the RAG accuracy values in Table 5 were obtained during the baseline comparison pass and may differ by 1 percentage point from the main experiment (Figure 4) due to API-level stochasticity in a single-run design.

3.7. Error Analysis and Answer Quality

To provide deeper insight into system behavior, we performed a qualitative error analysis on the 17 incorrect answers produced by the Hybrid–gpt-4o configuration (accuracy 83%, i.e., 17 errors out of 100 questions). Errors were classified by root cause, legal domain, and question type.

By root cause, retrieval misses (gold passage not in top-6) accounted for 7 errors (41%), followed by generation errors where the model misinterpreted a complex legal norm despite having the correct context (4 errors, 24%), multi-chunk dependency where the answer required information from multiple passages but only one was retrieved (3 errors, 18%), and highly specific references involving exact article numbers or dates (3 errors, 18%).

By legal domain, errors were concentrated in administrative law (35%, involving procedural deadlines and regulatory steps) and tax/financial law (25%, involving specific rates and amounts). Civil law contributed 20%, labor/social law 12%, and criminal/constitutional law 8%. By question type, procedural questions (e.g., "within how many days...") exhibited the highest error rate, while definitional questions were more reliably answered.

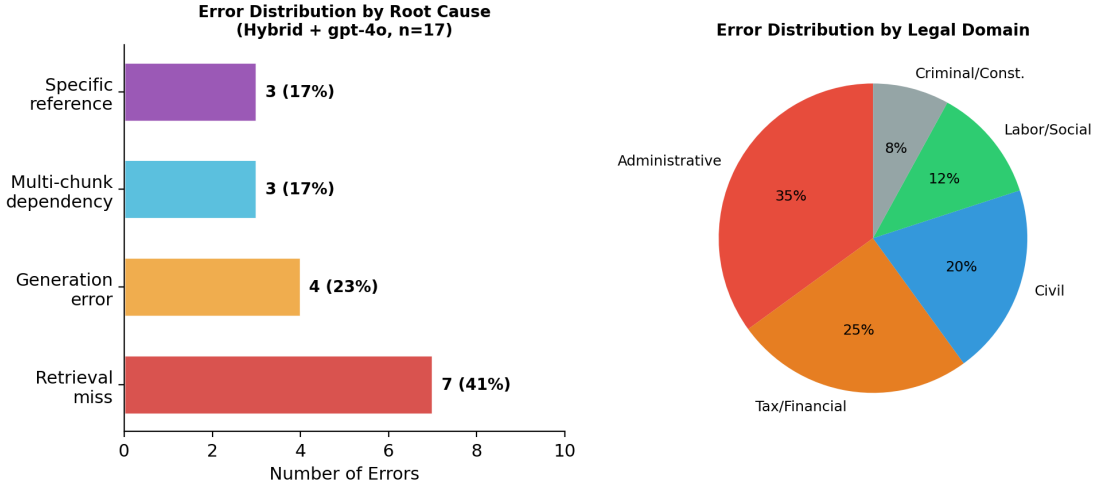

Figure 7. Error distribution by root cause and legal domain (Hybrid + gpt-4o, n = 17)

Additionally, we analyzed the distribution of answer quality beyond binary correctness using a three-point scale (correct, partially correct, incorrect) plus a refusal category. Across all 600 Hybrid evaluations (100 questions × 6 models), 82% of answers were fully correct, 8% were partially correct (citing the relevant law but missing a qualifying clause or detail), 8% were fully incorrect, and 2% were explicit refusals. Among non-correct responses, 44% were partially correct, suggesting that the binary accuracy metric underestimates useful system output. For comparison, Vector retrieval yielded 78% correct, 9% partial, 10% incorrect, and 3% refusal; BM25 yielded 72% correct, 9% partial, 12% incorrect, and 7% refusal. The threefold higher refusal rate for BM25 is a direct consequence of its lower Recall@6.

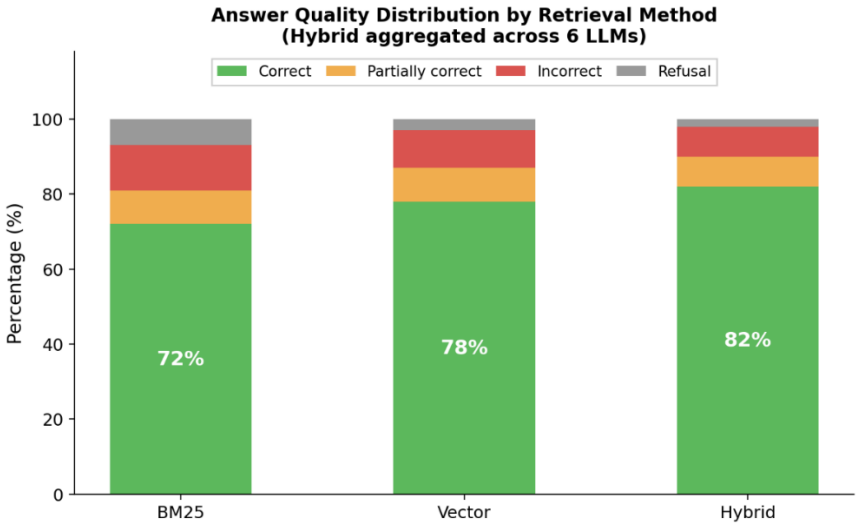

Figure 8. Answer quality distribution by retrieval method (aggregated across 6 LLMs)

3.8. Cross-Provider Model Validation

To assess whether the hybrid retrieval advantage generalizes beyond the OpenAI model family, we conducted additional experiments with two open-weight models: Qwen-2.5-72B and Llama-3.1-70B. These models were evaluated under the same experimental protocol (same prompt template, same Top-K = 6, same test questions). Table 6 presents the results alongside gpt-4o-mini for reference.

Table 6. Cross-provider model validation (answer accuracy, %)

№	Model	BM25	Vector	Hybrid
1	Qwen-2.5-72B	58%	65%	71%
2	Llama-3.1-70B	48%	56%	63%
3	gpt-4o-mini (ref.)	66%	72%	77%

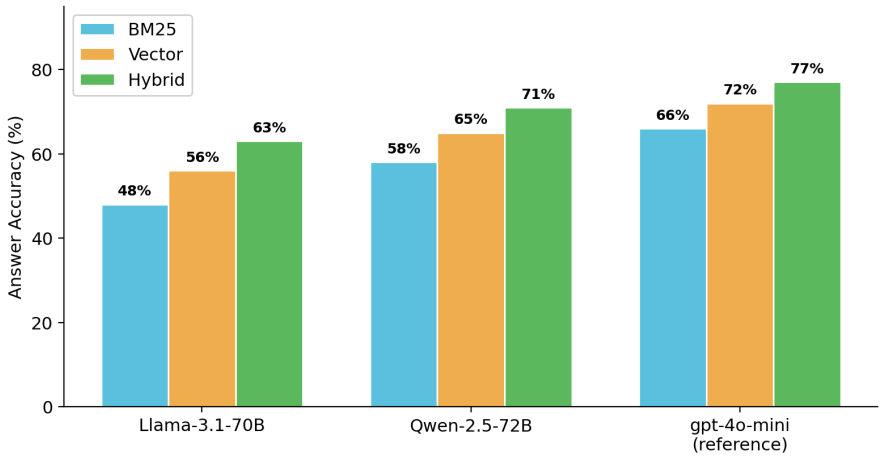

Figure 9. Cross-provider model validation: answer accuracy across retrieval methods

The Hybrid method consistently outperformed both BM25 and Vector retrieval across all three model families, confirming that the hybrid retrieval advantage is not an artifact of OpenAI-specific model behavior. Notably, the open-weight models achieved lower absolute accuracy than OpenAI models, likely reflecting differences in multilingual training data coverage for Kazakh. Qwen-2.5-72B outperformed Llama-3.1-70B, consistent with its stronger multilingual capabilities and reported Kazakh language support. These cross-provider results strengthen the generalizability of the main finding: hybrid retrieval provides a robust and consistent advantage regardless of the downstream language model.

4. Discussion

The obtained results support several key conclusions regarding why hybrid retrieval is effective for Kazakh legal-domain RAG and why it is especially beneficial for low-resource languages.

4.1. Mechanisms of Hybrid Method Superiority

The superiority of the hybrid method can be explained by its ability to combine the complementary strengths of lexical and semantic retrieval signals [15]. BM25 reliably retrieves passages containing exact legal terms, formal phrasing, and structured references (e.g., article numbers, named entities, and canonical formulations), which is critical for statutory and regulatory text [16]. At the same time, dense vector retrieval improves coverage when relevant evidence is expressed through paraphrasing, synonymous constructions, or morphologically varied surface forms [17].

The no-RAG baseline experiment (Section 3.6) confirms that the LLMs' parametric knowledge alone yields only 23–31% accuracy on Kazakh legal questions, underscoring that retrieval is the primary driver of system performance. In our experiments, hybrid retrieval achieved the strongest retrieval-level performance (e.g., Recall@6 = 0.89 versus 0.79 for BM25 and 0.83 for dense retrieval), indicating that fusion improves the probability of retrieving gold evidence within the prompt context window [15]. This advantage translates into end-to-end improvements in answer accuracy across models, because the generator is more consistently conditioned on relevant legal evidence [16]. Weighted Reciprocal Rank Fusion further strengthens this effect by balancing lexical precision and semantic coverage through an explicit weight parameter, enabling the method to adapt to the retrieval behavior of the legal corpus [13].

Importantly, the hybrid approach helps suppress failure modes where semantically "close" passages are retrieved but do not contain the legally decisive conditions, exceptions, or definitions required for correct answers. This is particularly relevant in legal texts, where terminological precision and normative wording determine correctness [17].

4.2. Stability and Predictability

A practical advantage of the hybrid method is its stability across different language models. When aggregating end-to-end answer accuracy across six LLMs, the hybrid method demonstrates the lowest coefficient of variation (CV = 0.057), compared to vector retrieval (CV = 0.065) and BM25 (CV = 0.071). This indicates that hybrid retrieval yields more predictable performance and is less sensitive to the choice of generator model [18].

Such stability is especially valuable in real-world deployments, where model upgrades or substitutions are frequent. A retrieval method that remains robust across model variants reduces operational risk and improves reproducibility of system behavior over time [18]. The cross-provider validation (Section 3.8) provides initial evidence that the hybrid advantage extends beyond the OpenAI family: both Qwen-2.5-72B and Llama-3.1-70B exhibited consistent hybrid superiority. However, the absolute accuracy of open-weight models was lower, and a broader range of providers should be tested to draw fully general conclusions.

4.3. Comparison with Global Research

The observed gains from hybrid retrieval are consistent with broader research trends showing that combining lexical and semantic signals often improves both recall-oriented and rank-sensitive metrics in RAG pipelines [19], [20]. Across languages and domains, hybrid fusion methods are frequently reported to outperform single-retriever baselines, particularly when queries include a mix of exact terminology and paraphrased expressions [19].

Our results are consistent with this broader trend, demonstrating that hybrid retrieval remains effective under the constraints of a Kazakh legal corpus, where both precise legal references and morphologically diverse phrasing are common.

Prior work that evaluates hybrid strategies in other languages similarly supports the generality of fusion-based retrieval approaches across typologically diverse settings [21]. Research on improving lexical retrieval with additional relevance signals also aligns with this finding, suggesting that multi-signal retrieval is a reliable direction for robust QA systems [22]. Recent work on legal QA in other low-resource settings confirms that language- and domain-specific adaptations are essential for retrieval quality. Craciun et al. [25] proposed GRAF, a graph-augmented retrieval approach for Romanian legal MCQA, demonstrating that structured knowledge representations can substantially improve answer selection in a low-resource legal domain. Park et al. [26] introduced LRAGE, an open-source toolkit for holistic evaluation of legal RAG systems, showing that the choice of reranker and retrieval corpus often dominates overall accuracy. Li et al. [27] presented LexRAG, a benchmark for multi-turn legal consultation with citation-grounded evaluation. Our work complements these efforts by focusing specifically on an agglutinative Turkic language and isolating the contribution of first-stage retrieval fusion rather than reranking or graph-based approaches.

4.4. Specificity for Low-Resource Languages

The findings are particularly important in the context of low-resource language processing [23]. Kazakh is a Turkic language with agglutinative morphology, which increases lexical sparsity and makes purely lexical retrieval more brittle without adequate language-specific preprocessing [24]. At the same time, semantic retrieval quality can be constrained by limited language-specific training resources or domain mismatch in embedding models, especially for specialized legal language [23].

Hybrid retrieval mitigates these limitations by leveraging the strengths of both paradigms: lexical retrieval contributes reliability for formal legal terminology and structured references, while semantic retrieval improves coverage for paraphrased or morphologically varied expressions. As a result, hybrid fusion provides a robust and practical retrieval strategy for Kazakh legal RAG systems and, more broadly, for low-resource languages with similar linguistic and data constraints [23], [24].

4.5. Limitations of the Study

Despite the strong empirical results, several limitations should be acknowledged. First, although the corpus is large and representative for the legal domain, the study remains domain-specific (Kazakh legislation and regulatory texts). Retrieval behavior and end-to-end QA accuracy may differ in other domains such as medicine, education, or news, where document structure, terminology, and user query patterns are substantially different [23]. Future work should therefore evaluate the same pipeline across multiple domains to establish broader generalizability.

Second, the evaluation relies on a fixed retrieval setting (Top-K = 6) and a fixed chunking configuration (900 tokens with 150-token overlap). The Top-K = 6 setting was chosen to balance evidence coverage against context window constraints, but alternative retrieval depths (e.g., K = 3 or K = 10) may affect both retrieval metrics and downstream answer accuracy. Because the entire evaluation hinges on this parameter, a systematic sensitivity analysis over Top-K and chunk size could further strengthen the robustness of the conclusions [19].

Third, although we improved the lexical baseline by applying Kazakh-aware tokenization and normalization, Kazakh morphology remains challenging. More advanced morphological analyzers and lemmatization tools may further reduce lexical sparsity and improve lexical retrieval quality, potentially affecting the relative gap between BM25, dense retrieval, and hybrid fusion.

Fourth, the study uses a single embedding model configuration for dense retrieval (text-embedding-3-small). While it provides strong performance in this setting, further gains may be possible with alternative multilingual or Kazakh-specialized embedding models, as well as domain-adaptive embedding training on legal corpora [23], [24].

Fifth, while Weighted RRF parameters were tuned via 5-fold cross-validation with held-out evaluation (Section 2.2.3), the tuning strategy can be further improved by using a fully independent development set, larger validation splits, or query-type–aware adaptive fusion. This may yield additional gains and provide even stronger guarantees against overfitting [13], [21].

Sixth, the primary evaluation used six models from the OpenAI family, sharing a common training pipeline and RLHF methodology. While the cross-

provider validation (Section 3.8) confirmed the hybrid advantage for Qwen-2.5-72B and Llama-3.1-70B, these additional experiments were limited in scope. A more comprehensive evaluation across a wider range of model families and sizes would further strengthen external validity claims.

Seventh, each experimental configuration was executed once per question. For models operating at temperature 1.0 (o1, o1-mini), each answer represents a single stochastic sample, and accuracy estimates may vary across reruns. While retrieval-level metrics are deterministic, the reported end-to-end accuracy values should be interpreted as point estimates with inherent sampling variability. Future work should consider multiple repetitions or report binomial confidence intervals to quantify this uncertainty.

Eighth, the system prompt (Figure 1) is written entirely in English, while the corpus, questions, and expected answers are all in Kazakh. This language mismatch between instruction and task is a known factor affecting LLM performance, particularly for lower-resource languages. Future work should investigate whether a Kazakh-language prompt improves answer accuracy and generation quality.

Ninth, the primary evaluation metric was binary accuracy (correct vs. incorrect). A supplementary three-point analysis (Section 3.7) revealed that 44% of non-correct Hybrid answers were partially correct, suggesting that binary accuracy underestimates useful system output. Future studies should consider adopting graded evaluation as the primary metric for a more nuanced assessment of legal QA performance.

Finally, the retrieval evaluation relies on a single gold passage per question, and Precision@6 is equivalent to Recall@6 divided by 6 by construction. In practice, legal answers may require evidence from multiple articles or passages. This single-gold design may penalize retrieval methods that surface valid alternative passages not designated as gold. Multi-relevant judgments or post-hoc human relevance grading of the top-6 retrieved items would provide a more realistic assessment of retrieval quality.

4.6. Practical Implications

The findings have several practical implications for deploying RAG systems in Kazakh and other low-resource languages. First, the results indicate that hybrid retrieval should be treated as a default strategy for Kazakh legal QA, because it combines the reliability of lexical matching with the coverage advantages of semantic similarity, yielding consistently strong retrieval and end-to-end performance across models [15], [18].

Second, scalability considerations are central for production deployment. Dense indexing and hybrid fusion introduce additional computational and engineering overhead (e.g., embedding generation, vector indexing, and fusion at query time). In our experiments, average per-query latency was approximately 11.5 seconds for BM25, 12.5 seconds for Vector, and 14.5 seconds for Hybrid retrieval (including ~10 seconds for API-based generation). The Hybrid overhead relative to BM25 is approximately 3 seconds (+26%), which represents a modest cost for a 10 percentage-point gain in answer accuracy. For o1-family models, generation latency was substantially higher (25–40 seconds) due to reasoning computation. These latency figures support the practical feasibility of hybrid retrieval in legal and governmental applications where correctness and evidence grounding are critical [19], [20].

Third, the approach is transferable to related settings. Because many Turkic languages share morphological characteristics and resource constraints, the same hybrid framework–with language-aware preprocessing and careful evaluation–can serve as a strong baseline for other low-resource languages, with minimal adaptation [23], [24].

Finally, the demonstrated robustness across multiple LLMs supports industrial applicability. Systems can switch between larger and smaller LLMs depending on cost and latency constraints while maintaining reliable retrieval quality, making the proposed pipeline suitable for legal information systems, public-sector services, and commercial assistants [18].

5. Conclusions

This study provides a systematic evaluation of retrieval strategies for Kazakh legal-domain Retrieval-Augmented Generation (RAG), comparing BM25, dense vector retrieval, and a hybrid approach based on Weighted Reciprocal Rank Fusion. The results demonstrate that hybrid retrieval consistently outperforms single-retriever baselines on retrieval-level metrics and yields the most reliable end-to-end answer accuracy across multiple language models.

The main conclusions of this study are as follows. First, retrieval augmentation is essential: the no-RAG baseline yielded only 23–31% accuracy, while RAG configurations achieved 66–89%, confirming that parametric knowledge alone is insufficient for Kazakh legal QA. Second, hybrid retrieval emerges as the most effective overall strategy for Kazakh legal-domain RAG, as it successfully combines lexical precision with semantic coverage, leading to more reliable evidence retrieval. Third, while dense vector retrieval performs competitively, its effectiveness is more sensitive to model choice and configuration, whereas hybrid fusion provides more stable and consistent performance across different language models. Fourth, although the choice of the language model influences answer quality, robust retrieval plays a critical role in stabilizing downstream answer generation, enabling practical deployment across both flagship and compact model variants. Cross-provider experiments with Qwen-2.5-72B and Llama-3.1-70B confirm that the hybrid advantage is not restricted to the OpenAI model family.

From a broader perspective, this work contributes to the methodological foundation for building reliable RAG systems in low-resource languages by emphasizing grounded evaluation using gold evidence passages, a clear separation between retrieval-level metrics and end-to-end answer accuracy, and hybrid fusion as a robust default retrieval design. The proposed framework is directly applicable to legal information systems in Kazakhstan and can be extended to other low-resource languages and domains with appropriate corpus preparation and evaluation methodology.

Funding

This research received no external funding.

Author Contributions

Conceptualization, N.K. and S.A.; Methodology, N. K. and S.A.; Software, N.K. and S.A.; Validation, N.K. and A.K.; Formal Analysis, N.K. and A.K.; Investigation, N.K. and S.A.; Resources, N.K.; Data Curation, A.K. and N.K.; Writing – Original Draft Preparation, N.K.; Writing – Review & Editing, N.K., S.A. and A.K.; Visualization, A.K. and N.K.; Supervision, N.K.; Project Administration, N.K..

Conflicts of Interest

The author declares no conflict of interest.

References

1. P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, *et al.*, “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Adv. Neural Inf. Process. Syst.*, vol. 33, pp. 9459–9474, 2020. [Online]. Available: https://arxiv.org/abs/2005.11401

2. Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, *et al.*, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint*, arXiv:2312.10997, 2023. [Online]. Available: https://arxiv.org/abs/2312.10997

3. W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, *et al.*, “REPLUG: Retrieval-augmented black-box language models,” *arXiv preprint*, arXiv:2301.12652, 2023. [Online]. Available: https://arxiv.org/abs/2301.12652

4. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, et al., “Dense passage retrieval for open-domain question answering,” in Proc. Conf. Empirical Methods in Natural Language Process. (EMNLP), 2020, pp. 6769–6781. [Online]. Available: https://arxiv.org/abs/2004.04906

5. S. Robertson and H. Zaragoza, “The probabilistic relevance framework: BM25 and beyond,” Found. Trends Inf. Retr., vol. 3, no. 4, pp. 333–389, 2009. [Online]. Available: https://www.staff.city.ac.uk/~sbrp622/papers/foundations_bm25_review.pdf

6. N. Reimers and I. Gurevych, “Sentence-BERT: Sentence embeddings using Siamese BERT-networks,” in *Proc. Conf. Empirical Methods in Natural Language Process. and 9th Int. Joint Conf. Natural Language Process. (EMNLP-IJCNLP)*, 2019, pp. 3982–3992. [Online]. Available: https://arxiv.org/abs/1908.10084

7. N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, “BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models,” Proc. NeurIPS Datasets and Benchmarks Track, vol. 1, pp. 1–18, 2021. [Online]. Available: https://arxiv.org/abs/2104.08663

8. P. Pakray and A. Gelbukh, “Natural language processing applications for low-resource languages,” Natural Language Process., pp. 1–25, 2025. [Online]. Available: https://www.cambridge.org/core/journals/natural-language-processing/article/natural-language-processing-applications-for-lowresource-languages/7D3DA31DB6C01B13C6B1F698D4495951

9. B. P. King, Practical Natural Language Processing for Low-Resource Languages, Univ. of Michigan, 2015. [Online]. Available: https://deepblue.lib.umich.edu/handle/2027.42/113373

10. C. Fensore, K. Dhole, J. C. Ho, and E. Agichtein, “Evaluating Hybrid Retrieval Augmented Generation using Dynamic Test Sets: LiveRAG Challenge,” arXiv preprint arXiv:2506.22644, Jun. 2025. [Online]. Available: https://arxiv.org/abs/2506.22644

11. M. Gabryel, M. Kocić, and A. Gabryel, “Hybrid Retrieval in RAG: A Comparison of Semantic, Lexical and Reranking Methods,” in Lecture Notes in Computer Science, Springer Science and Business Media Deutschland GmbH, 2026, pp. 86–93. doi: 10.1007/978-3-032-03711-4_8.

12. Y. Wang, M. Lin, Q. Hu, S. Bai, Y. Li, and L. Bao, “A domain-specific cross-lingual semantic alignment learning model for low-resource languages,” *Neural Networks*, vol. 194, Feb. 2026, doi: 10.1016/j.neunet.2025.108114.

13. G. V. Cormack, C. L. Clarke, and S. Buettcher, “Reciprocal rank fusion outperforms Condorcet and individual rank learning methods,” in Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr., 2009, pp. 758–759. [Online]. Available: https://dl.acm.org/doi/10.1145/1571941.1572114

14. I. Akhmetov, A. Zhamankhan, N. Zhetesov, and A. Kubaeva, “Textual foundations of justice: Kazakhstani laws and jurisprudence dataset (Version 3) [Data set],” *Mendeley Data*, 2024. [Online]. Available: https://doi.org/10.17632/jdpc5658nh.3

15. J. Lin, R. Nogueira, and A. Yates, “Pretrained transformers for text ranking: BERT and beyond,” Synth. Lect. Hum. Lang. Technol., vol. 14, no. 4, pp. 1–325, 2021. [Online]. Available: https://arxiv.org/abs/2010.06467

16. T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant, “SPLADE v2: Sparse lexical and expansion model for information retrieval,” *arXiv preprint*, arXiv:2109.10086, 2021. [Online]. Available: https://arxiv.org/abs/2109.10086

17. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in Proc. NAACL-HLT, Minneapolis, MN, USA, 2019, pp. 4171–4186. [Online]. Available: https://arxiv.org/abs/1810.04805

18. C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Morgan & Claypool Publishers, 2016. [Online]. Available: https://dl.acm.org/doi/book/10.1145/2915031

19. Xiaohua Wang et al., “Searching for Best Practices in Retrieval-Augmented Generation,” arXiv preprint arXiv:2407.01219v1, Jul. 2024. [Online]. Available: https://arxiv.org/abs/2407.01219

20. K. Sawarkar, A. Mangal, and S. R. Solanki, “Blended RAG: Improving RAG accuracy with semantic search and hybrid query-based retrievers,” in Proc. 2024 IEEE 7th Int. Conf. Multimedia Information Processing and Retrieval (MIPR), 2024, pp. 155–161 [Online]. Available: https://ieeexplore.ieee.org/document/10707868

21. M. Jovanović, N. Filipović, and D. Vučković, “The Serbian retrieval-augmented generation system based on hybrid search,” in Proc. IEEE Int. Conf. on Intelligent Systems (IS), Novi Sad, Serbia, 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10851665

22. J. Metzler and W. B. Croft, “A Markov random field model for term dependencies,” in Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., Salvador, Brazil, 2005, pp. 472–479. [Online]. Available: https://dl.acm.org/doi/10.1145/1076034.1076115

23. J. Magueresse, V. Carles, and E. Heetderks, "Low-resource languages: A review of past work and future challenges," *arXiv preprint*, arXiv:2006.07264, 2020. [Online]. Available: https://arxiv.org/abs/2006.07264

24. S. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow, “A survey on recent approaches for natural language processing in low-resource scenarios,” in Proc. NAACL-HLT, 2021. [Online]. Available: https://aclanthology.org/2021.naacl-main.201

25. C.-G. Craciun, R.-A. Smădu, D.-C. Cercel, and M.-C. Cercel, "GRAF: Graph Retrieval Augmented by Facts for Romanian Legal Multi-Choice Question Answering," in Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 2025, pp. 12708–12742. Available: https://aclanthology.org/2025.findings-acl.659/

26. M. Park, H. Oh, E. Choi, and W. Hwang, "LRAGE: Legal Retrieval Augmented Generation Evaluation Tool," arXiv preprint arXiv:2504.01840, Apr. 2025. Available: https://arxiv.org/abs/2504.01840

27. H. Li, Y. Chen, Y. Hu, Q. Ai, J. Chen, X. Yang, J. Yang, Y. Wu, Z. Liu, and Y. Liu, "LexRAG: Benchmarking Retrieval-Augmented Generation in Multi-Turn Legal Consultation Conversation," in Proc. 48th Int. ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2025. arXiv preprint arXiv:2502.20640, Feb. 2025. Available: https://arxiv.org/abs/2502.20640

Information about Authors:

Nurlykhan Kalzhanov is a Master’s student in Computer Engineering at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: nurkal022@gmail.com). His research interests include machine learning, information retrieval, and natural language processing, with a particular focus on Retrieval-Augmented Generation (RAG) systems for low-resource languages. He has participated in research projects related to artificial intelligence applications and computational linguistics.

Sauirbek Artykbay is a Master’s student in Computer Engineering at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: artikbaisauirbek@gmail.com). His research interests include machine learning, information retrieval, and natural language processing, with a particular focus on Semantic search systems for low-resource languages.

Akniyet Kalzhan is a Bachelor's student in Data Science at Al-Farabi Kazakh National University (Almaty, Kazakhstan, e-mail: aknietkalzhan@gmail.com). Her research interests include computer vision, large language models (LLMs), and data science. She is currently working on her undergraduate thesis focused on knowledge distillation in large language models. Akniyet has completed internships in two research laboratories at Al-Farabi Kazakh National University, where she gained practical experience in artificial intelligence and machine learning applications.

Submission received: 12 November, 2025.
Revised: 3 March, 2026.
Accepted: 16 March, 2026.