

Article

Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation

Diana Rakhimova ^{1,2}, Assem Turarbek ^{1,2,*}, Vladislav Karyukin ¹ , Assiya Sarsenbayeva ¹ and Rashid Alieyev ¹ 

¹ Department of Information Systems, Al-Farabi Kazakh National University, Almaty 050040, Kazakhstan; diana.rakhimova@kaznu.edu.kz (D.R.); vladislav.karyukin@gmail.com (V.K.)

as.sarsenbayeva@gmail.com (A.S.); alieyev.rashid.sh@gmail.com (R.A.)

² Institute of Information and Computational Technologies, Almaty 050010, Kazakhstan

* Correspondence: turarbekasem1@gmail.com

Abstract

The research focuses on the development and evaluation of a legal question–answer system for the Kazakh language, a low-resource and morphologically complex language. Four datasets were compiled from open legal sources—Adilet, Zqai, Gov, and a manually created synthetic set—containing question–answer pairs extracted from official legislative documents and government portals. Seven large language models (GPT-4o mini, GEMMA, KazLLM, LLaMA, Phi, Qwen, and Mistral) were fine-tuned using structured prompt templates, quantization methods, and domain-specific training to enhance contextual understanding and efficiency. The evaluation employed both automatic metrics (ROUGE and METEOR) and expert-based manual assessment. GPT-4o mini achieved the highest overall performance, with ROUGE-1: 0.309, ROUGE-2: 0.175, ROUGE-L: 0.263, and METEOR: 0.320, and received an expert score of 3.96, indicating strong legal reasoning capabilities and adaptability to Kazakh legal contexts. The results highlight GPT-4o mini’s superiority over other tested models in both quantitative and qualitative evaluations. This work demonstrates the feasibility and importance of developing localized legal AI solutions for low-resource languages, contributing to improved legal accessibility, transparency, and digital governance in Kazakhstan.



Academic Editor: Paolo Bellavista

Received: 7 July 2025

Revised: 13 August 2025

Accepted: 19 August 2025

Published: 27 August 2025

Citation: Rakhimova, D.; Turarbek, A.; Karyukin, V.; Sarsenbayeva, A.; Alieyev, R. Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation. *Computers* **2025**, *14*, 354. <https://doi.org/10.3390/computers14090354>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The modern development of artificial intelligence and natural language processing technologies has led to the active implementation of various question–answer (QA) systems in numerous fields of activity, including jurisprudence. In the context of the growing volume of regulatory legal acts, jurisdictional practice, and the legal literature, automated systems are becoming an essential tool for lawyers, government agencies, businesses, and citizens. QA systems in the legal field enable you to quickly locate relevant legal provisions, analyze precedents, generate legal opinions, and provide users with expert advice on legal issues. They are based on machine learning algorithms, semantic search, and natural language understanding technologies, making them effective tools for legal analysis [1].

In Kazakhstan, QA technologies have been intensively developed over the last decade. Within the framework of the “Digital Kazakhstan” program [2], the digitalization of legal processes is growing, including the creation of chatbots and automated contract analysis systems. For example, the adilet.kz platform provides access to legislative acts, and

qamqor.gov.kz gives access to the legal protection of entrepreneurs [3]. The development of QA systems in Kazakhstan faces several linguistic and technological barriers related to the peculiarities of the Kazakh language. The primary issues are morphological complexity, syntactic variability, limited language corpora, the quality of machine translation, and the absence of localized models.

The Kazakh language is an agglutinative language, meaning it employs many affixes that alter the meaning of a word. It complicates tokenization and morphological analysis, as traditional natural language processing (NLP) algorithms are primarily focused on analytical languages, such as the English language [4], which highlights the morphological complexity. Another notable feature of the Kazakh language is its syntactic variability, which includes free word order and necessitates complex parsing models. Unlike the fixed order in English, which is typically expressed as Subject–Verb–Object (SVO), inversions are possible in Kazakh, affecting the meaning of a sentence [5]. QA in the Kazakh language faces limitations in language corpora [6], the poor quality of machine translation [7], and a lack of localized models [8].

Improving legal workflows is a particularly significant issue in Kazakhstan, as the number of lawsuits increases yearly, and the number of qualified lawyers does not always match demand. According to existing data, the legal market in Kazakhstan faces a shortage of specialists, which increases the importance of automation and the implementation of technologies that can speed up and improve processes. One of the most promising solutions to this problem is advances in deep learning, especially in the use of large language models (LLMs), which can significantly improve data processing and analysis in legal applications.

LLMs have significantly influenced the NLP field by enhancing their ability to analyze and generate text that resembles human language. Modern models such as OpenAI's GPT, Google's PaLM, and Meta's LLAMA are based on the transformer architecture, which makes them highly effective in understanding and generating text [9–12]. Recent research has increasingly focused on applying LLMs in specialized fields, such as law, medicine, and education [13].

Generally, the development of jurisdictional QA systems in Kazakhstan will allow for significant improvement in the correctness of legal answers, taking into account linguistic challenges specific to the Kazakh language, such as agglutination, syntactic variability, and a lack of language resources. Therefore, the existing QA models developed for the English language are insufficiently effective for Kazakh due to its agglutinative nature and syntactic variability. The lack of high-quality language corpora and localized language models limits the performance of legal QA systems in Kazakh. The integration of LLMs into legal QA systems also enhances the system's capacity for semantic understanding, legal text generation, and expert-level reasoning. The successful implementation of the systems in Kazakhstan aligns with national digitalization goals and addresses the shortage of qualified lawyers by automating routine legal tasks. These automated systems can reduce the workload of legal professionals and partially compensate for the shortage of qualified legal personnel in Kazakhstan. They will provide scalable and legal advice in routine inquiries. The integration of the QA systems into national platforms will also increase public access to legal information and improve citizen engagement with legal services.

This study aims to empirically evaluate the ability of modern multilingual LLMs to adapt to the legal domain in the Kazakh language through domain-specific fine-tuning. Despite the limited availability of linguistic resources for Kazakh, such fine-tuning can significantly improve the legal accuracy and contextual relevance of the model outputs. This paper also examines the fundamental principles of QA systems in jurisprudence, their

advantages and limitations, and their development prospects in the context of the digital transformation of the legal system.

2. Materials and Methods

The problem of building QA systems has many variations, ranging from factual answers to subjective opinions, from simple queries to multi-step reasoning, from single answers to lists of entities, from brief answers to detailed explanations, and from one-off questions to interactive dialogs [14,15]. Over the past decade, this field has advanced significantly due to the development of advanced technologies. The introduction of deep learning in retriever-reader architectures [16–18] has improved the accuracy of answers, and the development of LLMs [19,20] has given an even greater impetus to the development. With the advent of LLMs, chatbots have gained the ability to generate dialogs and perform translations [21] automatically.

However, since their work is based on pre-trained data, it may not reflect up-to-date information and may exhibit a low understanding of emerging topics and domains [22]. While 2023 marked the release of foundational LLMs, such as ChatGPT-4 and LLaMA-2, experts predict that 2024 will be an active year for the development of retrieval-augmented generation (RAG) models and AI agents [23].

The integration of artificial intelligence (AI) [24,25] into the legal field has evolved from the first rule-based expert systems to modern deep learning models. Early LegalAI projects [26] such as TAXMAN and HYPO used rule-based logic to mimic human legal reasoning [27]. While these systems demonstrated the capabilities of AI in legal applications, they were limited by a fixed set of knowledge and could not generalize information beyond pre-programmed scenarios.

Implementing deep learning to predict the outcome of legal cases has been an essential step in the development of LegalAI. Researchers [28–30] have pioneered the use of neural networks to analyze legal documents and predict outcomes. Research in LegalAI focuses on incorporating legal knowledge into AI models. For example, ref. [31] demonstrated how attention between facts and statutory clauses can improve charge predictions, and [32] utilized topological graphs to account for the relationships between different Legal Judgment Prediction (LJP) tasks, highlighting the importance of structured legal knowledge in enhancing the performance of models.

In addition to LJP, advances in legal entity recognition and classification, as achieved in research [33], have improved document analysis methods. Additionally, developments in court document creation and legal summarization, proposed by researchers [34], open new horizons for automating legal processes. Research [35,36] has significantly contributed to integrating legal knowledge with AI, providing valuable tools for both legal professionals and the general public.

Efforts to improve the performance of LJP have led to the use of more sophisticated neural network architectures. For example, researchers [37] implemented control mechanisms to improve penalty predictions, and [38] proposed multi-scale attention models to handle more complex cases with multiple defendants. The shift from traditional AI to deep learning in the legal field has been a substantial advance, significantly improving the efficiency, accuracy, and accessibility of legal advice and making legal expertise more accessible [39]. However, such models can inherit biases from the data they are trained on, raising ethical concerns about fairness and impartiality. Most legal QA systems are classified as closed-domain systems because they work with a limited and well-defined corpus of legal texts. In this paper, we focus on closed-domain conditions, specifically the legal system of Kazakhstan. This distinction is critical because it affects model training, search strategies, and evaluation criteria. While open-domain QA systems must handle diverse

and unpredictable topics, closed-domain systems can leverage the context and structure of the domain to improve accuracy and interpretability.

QA systems based on LLaMA, GPT, and DeepSeek models have shown significant advances in NLP. An overview of research studies with real-world results in this area is shown in Table 1.

Table 1. Analysis of scientific research.

Authors	Language	Model	Domain Type	Accuracy	Evaluation Methods	Limitations
Emrah Budur, Rıza Özcelik, Dilara Soylu, Omar Khattab, Tunga Güngör, Christopher Potts [40]	Turkish	ColBERT-QA model, trained on SQuAD-TR	Closed	Improvement in Exact Match (EM) accuracy by 24–32% compared to baseline models (BM25 and DPR). Improvement in F1-score by 22–29%.	Exact Match, F1	Focused only on Turkish; requires SQuAD-style annotation
Tikhomirov M. & Chernyshov D. [41]	Russian	LLaMA Tokenization Adaptation	General/Closed	Improved performance on Russian SuperGLUE, with fine-tuning speed increased by 35% and inference speed up to 60%.	Russian SuperGLUE Benchmarks	No legal-domain focus; technical improvements only
Hong Q. et al. [42]	English	GPT-4, Qwen-Max, Gemini, LLaMA-2	Open	GPT-4: 70.5%; Qwen-Max: 81.2%; Gemini: 77.2%; LLaMA-2: lower.	Accuracy (Open LegalBench)	Limited Kazakh evaluation; mostly English benchmarks
Tao, M.; Zhao, D.; Feng, Y. [43]	English, Chinese	Several open-source LLMs, including ChatGPT, are available	Open	In the GPT-4-based evaluation, the framework improves by 0.86% for N-Acc and 0.53% for O-Acc, respectively.	N-Acc and O-Acc metrics with GPT-4	Slight gain; lacks legal specialization
Togmanov et al. [44]	Kazakh	LLaMA-3.1, Qwen-2.5, GPT-4, DeepSeek	Closed	~40–60% across topics.	Topic-wise evaluation	Lack of fine-grained legal reasoning metrics

The analysis of scientific studies in Table 1 reveals that LLMs continue to demonstrate high potential in QA systems, particularly in complex tasks that necessitate evidential reasoning and operate with unstructured data. While Haystack and LlamaIndex demonstrate varying degrees of accuracy in Russian, GPT-4 and TAT-LLMs excel in English-language benchmarks. The Chain-of-Discussion framework yields improvements in results due to the use of multiple models.

A recent benchmark, KazMMLU [44], evaluates Kazakh, Russian, and English language models in multiple domains, including law, using a multiple-choice format. In the legal category, models like GPT-4 and DeepSeek achieved ~63% and ~58% accuracy, while the KazLLM reached ~45%. Our study differs by focusing on open-ended, generative legal QA over Kazakh legislative texts. Rather than selecting predefined options, models generate full responses, evaluated via ROUGE, METEOR, and expert review. GPT-4o mini led in factuality and legal accuracy. While KazMMLU tests recognition and domain recall, our approach assesses real-world legal reasoning, citation accuracy, and linguistic nuance. Thus, our study complements KazMMLU by extending evaluation from classification to generation. Both works highlight the KazLLM's limitations and the value of domain-specific adaptation.

In this study, we focus on a closed-domain QA system where all queries and responses are restricted to the legal framework of Kazakhstan. Unlike open-domain QA systems that work with general knowledge (e.g., Wikipedia), our system is designed to search and analyze structured legal sources, such as the Adilet database, government regulations, and frequently asked questions on legislation. This restriction provides a more accurate assessment but also creates specific problems related to legal semantics, terminology, and context dependence.

A QA system is a modern and specialized method for accessing information in various fields of activity. It overcomes the gap between information and relevant knowledge. QA

is an integral assistant to lawyers, as it enhances work efficiency and increases the availability of legal services. The development of QA systems in Kazakhstan can contribute to the modernization of the legal system and strengthen the citizens' trust in the judicial system. Continuing research and development in this area is crucial, taking into account both international experiences and local characteristics.

3. Results

Training LLMs for QA systems involves adapting models to perform effectively on domain-specific QA datasets. This process typically involves supervised training on high-quality question–answer pairs, enabling the models to learn context-aware answer extraction and generation. Training a language model for the legislative field consists of customizing it to comprehend legal terminology, interpret, and provide accurate responses to law-related queries. Training QA models involves several steps, as shown in Figure 1.

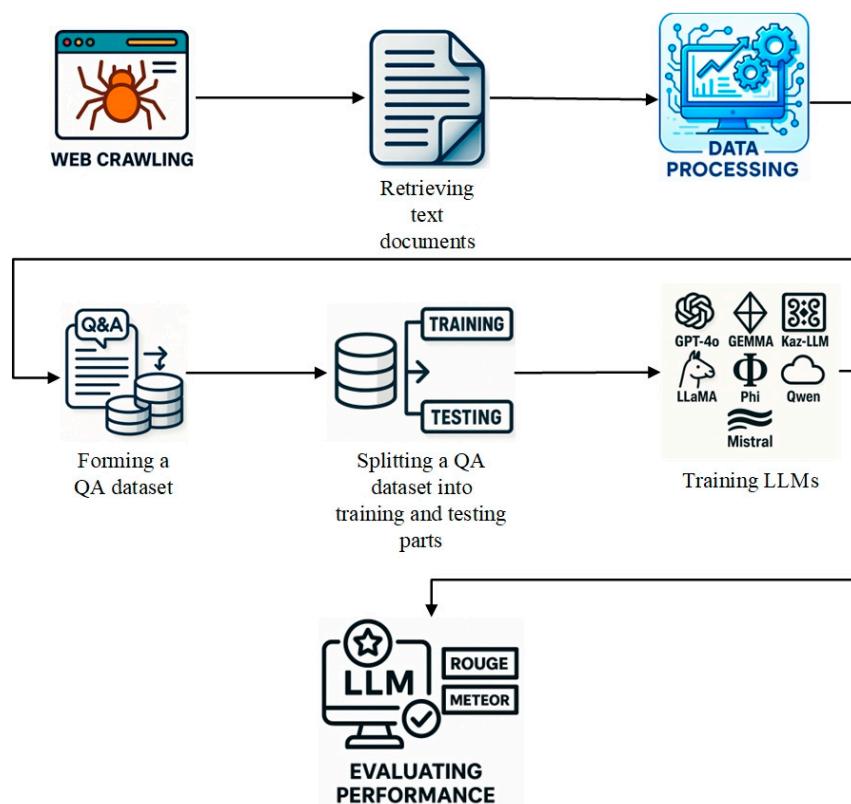


Figure 1. Training the QA LLMs.

The whole process of building QA systems with LLMs starts with configuring web crawlers for the required sources on which LLMs are built. Data collection with the use of web scraping tools generally implies deep analytics of electronic portals where the required legislative data is located. It means that it is essential to understand the structure of web pages, including blocks, styles, and scripts. If the markup and scripts of the web pages are complex, it is necessary to create dynamic processing programs. In the conducted research, the dataset was taken from "Adilet", "Zqai", and "Gov" sources. Experts synthetically created the other part of the full QA dataset. The scraped texts undergo a preprocessing phase, where they are cleaned and applied with various regular expressions and patterns to make them structured and prepared for subsequent steps. The preprocessed texts of all datasets are formed and combined for training with LLMs. They are split into training and testing parts to enable both learning and evaluation. The training part is used for fine-tuning GPT, GEMMA, Kaz-LLM, LLaMA, Phi, Qwen, and Mistral, tailoring them specifically for

the QA task. Finally, the performance of these trained models is evaluated using standard metrics like ROUGE and METEOR, which measure the quality and accuracy of the generated answers in comparison to reference responses. These steps of LLMs' training ensure a systematic development of effective QA systems. The detailed description and presentation of all shown steps are given in the subsequent subsections.

3.1. Data Collection

In Kazakhstan, all documents adopted by the Ministry of Justice of the Republic of Kazakhstan in electronic format are compiled in the Information and Legal System of Regulatory Legal Acts of the Republic of Kazakhstan, known as "Adilet." It contains a large number of legal documents collected between 1947 and 2025. These documents are partially translated into three languages. The largest number of these legal documents, 15,359 in total, was uploaded in 2023. Considering the division of documents into normative and non-normative [45], we can see some parts of all the documents. Unfortunately, the legislation of the Republic of Kazakhstan covers a wide range of areas of legal relations, and the number of legislative documents is growing every day; an example is provided in Appendix A, Table A1.

For a more detailed understanding of the distribution of legislative acts of the Republic of Kazakhstan by type, approving authority, legal domain, and year of adoption, structured Table A2 is presented in Appendix A. This information reflects the complexity and thematic diversity of the national legislation, which serves as an important foundation for building QA systems.

To obtain practical results from the research on QA systems in the field of legislation of the Republic of Kazakhstan, datasets in the state language were prepared. The construction of a QA system involves several key stages: data preparation, model selection, additional training, and quality evaluation. This process enables you to tailor the model to meet specific needs, thereby enhancing the accuracy and relevance of the answers. The data were gathered from open legislative sources: "Adilet" (<https://adilet.zan.kz/kaz/>), "Zqai" (<https://www.zqai.kz/ru/questions?>, accessed on 25 June 2025), and "Gov" (<https://www.gov.kz/>, accessed on 25 June 2025). These electronic portals have different structures but represent a competent, legally formalized knowledge base. The "Adilet" portal contains the legislative documents in the form of consecutive text. Therefore, the questions and answers were constructed manually from them.

On the "Zqai" and "Gov" portals, the data were already structured in the form of question-answer blocks, which simplified the formation of documents for training the system. Additionally, a synthetic dataset was manually constructed, comprising various questions and answers on different topics within the legislative field. The specification of this synthetic dataset is that it includes relatively short questions and answers, which demands a larger number of samples.

The Adilet and Zqai systems provide legal regulations but do not publish a categorical analysis of user queries. However, based on the data on the volume of documents and the popularity of topics, it is possible to compare and check to what extent our synthetic corpus (on 14 topics of the legislation of the Republic of Kazakhstan) corresponds to the topic and is relevant in national legal information systems. In the process of forming a synthetic corpus of legal questions and answers on the legislation of the Republic of Kazakhstan, an attempt was made to ensure its thematic completeness based on the analysis of key areas of law. A full comparative analysis is presented in Table 2.

Table 2. Comparison of the correspondence of the topics of the synthetic corpus with legislative portals.

Thematic Direction of the Synthetic Corpus	Representation in Adilet	Representation in Zqai
Constitutional law	Yes. Legislative changes are presented.	Yes (constitutional law and political analysis).
Civil law	Yes. It is in great demand among users.	Yes (corporate, inheritance, procedure).
Administrative law	Yes. Code of Administrative Offenses.	Yes (administrative and legal sciences).
Criminal law	Yes. The Criminal and Criminal Procedure Codes are actively presented.	Yes.
Labor law	Yes. The Labor Code is in great demand among users.	Yes.
Medical and pharmaceutical law	Yes. Patient rights issues, licensing, bioethics.	Yes.
Education	Yes. Includes regulation of UNT, universities, and professional standards.	Yes.
Family law	There is no explicit code, but there are regulatory documents in the legislation.	Yes. Reflected in Zqai research.
Environmental law	Yes. NPA and regulations.	Not fully. There are Zqai thematic articles.
Financial and tax law	Yes. Tax Code and other documents.	Yes.
Land and agrarian law	Yes. Land Code in the list.	Yes.
Customs and business law	Yes. Entrepreneurship Code.	Partially presented through corporate law.
Information and digital law	Yes. Strengthened with the introduction of digital services.	Yes. Zqai publishes topics: e-services and cyberlaw.
International law and foreign relations	A few documents directly.	Yes. Zqai research and international topics.

A comparative analysis with the portals adilet.zan.kz and zqai.kz showed a high degree of coverage of classical branches of law: constitutional, civil, administrative, criminal, labor, family, environmental, financial, and agrarian. At the same time, some inconsistencies and gaps were identified: topics such as information and digital law, as well as international law, are presented in the above-mentioned state and commercial sources partially or fragmentarily.

In adilet.zan.kz, the greatest interest of portal visitors is related to civil, labor, criminal, administrative, financial, and land law, which is confirmed by the popularity of the relevant codes. Our synthetic set fully covers these areas, including subcategories (inheritance, corporate law, contracts, liability). And in the Zqai scientific materials, the emphasis is on constitutional law, international law, information/digital law, and environmental law. Our set also includes these topics, expanding the educational and analytical value.

The comparative analysis showed that the current version of the synthetic corpus covers the main areas of Kazakhstani legislation. However, a number of significant legal topics remained outside the scope of consideration. Among them are migration legislation, pension and social security, intellectual property, anti-corruption policy, civil service, transport and aviation regulation, as well as religious and budgetary legal aspects.

In this regard, in the future, it is planned to expand the corpus by including missing topics. This will ensure a wider coverage of legal areas, increase the relevance of the

synthetic material, and make it more suitable for training intelligent systems aimed at automated legal assistance for the population.

After collecting the question–answer data, they were validated by removing duplicate, irrelevant, or poorly formulated data. The statistics on the number of QA pairs, sentences, words, and datasets are presented in Table 3.

Table 3. The statistics of the datasets.

Dataset	Number of QA Pairs	Sentences	Words	Size
Adilet	1012	7448	59,402	912 Kb
Zqai	602	15,591	219,125	756 Kb
Gov	738	9421	127,565	1.89 Mb
Synthetic	3764	22,829	67,926	328 Kb

The structures of the QA datasets are shown in Figure 2 and also in Figures A1–A3 in Appendix A.

id		Questions	Answers
0	0	Халық билікті қалай жүзеге асырады?	тікелей республикалық референдум және еркін са...
1	1	Қандай мемлекеттік органдарда партия үйымдарын...	мемлекеттік органдарда партия үйымдарын құруға...
2	2	ҚР Азаматтығынан айыруға жол беріледі мә?	террористік қылмыстар жасағаны, сондай-ақ Қаза...
3	3	Кімдерге саяси партияларда, көсіптік одақтарда...	Республика Конституциялық Сотының, Жоғары Сот...
4	4	Еріксіз енбекке жол беріледі мә?	қылмыстық немесе әкімшілік құқық бұзушылық жас...
...
1007	1007	Біріккен Үлттар Үйымының Бас хатшысы комитет м...	1. Бастапқы сайлау, осы Пактің күшіне енгіннен ...
1008	1008	Біріккен Үлттар Үйымының Бас хатшысы комитет м...	1. Комитет мүшелері төртжылдық мерзімге сай...
1009	1009	Адам құқыбы жеңіндегі комитеттің отырысы қайда...	37-бап!н 1. Комитеттің бірінші отырысы Бі...
1010	1010	Комитет езінің лауазымды адамдарын қандай мер...	39-бап!н 1. Комитет езінің лауазымды адам...
1011	1011	Пактіге қатысушы мемлекет Пактінің міндеттерін...	1. Осы бапқа сәйкес осы Пактіге қатысушы мемле...

1012 rows × 3 columns

Figure 2. The Adilet dataset.

All these datasets were thoroughly processed and prepared for subsequent processing by the different QA systems. First, the analysis of various QA architectures was conducted, and then the utilized LLMs were described.

3.2. Question–Answer Systems Classification

The architecture of QA systems can be classified by various criteria, including the source of data (structured or unstructured), the type of model used (rule-based, statistical, or neural), and the processing stages. The main types of QA architectures include the following:

- Rule-based QA systems;
- Information retrieval (IR)-based QA;
- Machine Reading Comprehension (MRC) QA;
- End-to-End Neural QA;
- Knowledge-based QA;
- Multi-hop QA.

Taking a look at each architecture separately, the earliest rule-based QA systems work based on predefined rules, patterns, and regular expressions. An example is the ELIZA System (1966), simulating a psychotherapist. It substituted phrases based on keywords without analyzing the meaning [46]. Information retrieval (IR)-based QA systems employ in-

formation retrieval approaches similar to those used by search engines. Unlike rule-based, they use relevance metrics to select text containing the answer. For example, IBM Watson used IR + NLP for the Jeopardy game [47].

Machine Reading Comprehension (MRC) QA systems read the text context and search it for an accurate answer to a question. A notable example is BiDAF (Bidirectional Attention Flow), which utilizes a combination of character-level, word-level, and context-level embeddings, along with a bidirectional attention mechanism, to align question and passage representations [48]. More recent MRC models, such as BERT, RoBERTa, and ELECTRA, further improve performance by pre-training on large corpora with masked language modeling and then fine-tuning on QA datasets like SQuAD. These models enable deep contextual understanding by considering the entire sentence and question simultaneously, making them highly effective for extractive QA tasks where the answer is a text span within the passage.

End-to-End Neural QA architectures generate answers from scratch using language models that operate without specific context, using knowledge encoded in the model parameters [49,50]. Knowledge-based QA systems use formal knowledge bases (Wikidata, DBpedia, Freebase, etc.). The query is transformed into a logical form (SPARQL and others) and executed over the knowledge graph structure [51]. These models, typically based on transformer architectures such as GPT or T5, are trained on massive text corpora, allowing them to internalize factual and contextual knowledge within their parameters. When a question is input, the model processes it through self-attention mechanisms and outputs a response in natural language, leveraging its embedded knowledge rather than querying structured data sources.

Multi-hop question-answering (QA) systems are designed to answer complex queries that require reasoning over multiple pieces of evidence, often scattered across different parts of a document or multiple documents. Unlike single-hop QA, which finds answers within a single sentence or paragraph, multi-hop QA involves a chain of reasoning steps where intermediate information must be retrieved and connected to arrive at the final answer [52]. A complete comparison of the architectures described above is shown in Table 4.

Each QA system architecture uses its own approach to query processing. Key parameters, including accuracy, implementation complexity, and computational costs, are considered and presented in Table 5 to evaluate the effectiveness of different QA systems.

Table 4. The comparison of QA structures.

Type	Used Data	Model	Example
Rule-based	Structured	Template	ELIZA
IR-based	Unstructured	Searcher + NLP	IBM Watson
MRC	Text and context	BERT, BiDAF	SQuAD
End-to-End [53]	Text	GPT, T5	ChatGPT
Knowledge-based	Structured graphs	SPARQL + NLG	Wolfram Alpha
Multi-hop	Text and graphs	GNN, Memory Networks	HotpotQA

Table 5. The comparison of QA effectiveness [54].

Parameter	Factoidal	Deep Analytical	Chatbots	IR-Based	LLM-Based
Accuracy	High	Very high	Medium	High	Model-dependent
Context	Low	High	Very high	Medium	High
Computational complexity	Low	High	Medium	Medium	Very high
NLP use	Limited	Full	Full	Part	Full

Each architecture has tradeoffs across accuracy, context dependence, computational complexity, and NLP capabilities, making them suitable for different application domains—from chatbots and factoid systems to deep analytical engines. All six major types of QA—rule-based, information retrieval (IR)-based, Machine Reading Comprehension (MRC), End-to-End Neural, knowledge-based, and multi-hop—have their own specifications. Rule-based systems rely on hand-crafted patterns and templates (e.g., ELIZA), offering limited language understanding [55]. IR-based QA systems retrieve relevant documents using search algorithms and extract answers using NLP techniques, as seen in IBM Watson. Comparative tables in the document highlight differences in data use, model complexity, example systems, accuracy, computational cost, and NLP capabilities across these architectures.

A comparison of approaches for QA systems in the Kazakh language. Rule-based /IR-based systems, such as the ElasticSearch implementation, work on the keyword principle and BM25 ranking. It is suitable for simple factual queries, where answers can be found by keywords in a document or regulatory text database. This approach requires minimal computations and is easy to deploy and interpret. It also supports local rules. However, it does not understand the query's meaning, is sensitive to synonyms and grammar, and has weak processing of agglutinative forms of Kazakh. Dense Passage Retrieval [56] is a classical IR approach with dual-encoder retrieval, which shows an accuracy gain of 9–19% over BM25.

Compact BERT-like models are suitable for more complex questions requiring an understanding of context or agglutinative forms and extractive QA tasks (extracting an answer from a text). BERT is used in combination with expert classification and morphological rules for a regular question-answering system for short questions in the Kazakh language [57].

The advantages of BERT include deep contextual understanding, resistance to word form changes, and the ability to work with a small corpus after fine-tuning [58]. But they also require GPUs and more data and are sometimes less interpretable than IR. For simple queries like “What is the VAT rate?”, an IR solution (ElasticSearch + keys) will be enough. For queries answering a question within a long context (for example, “What is civil liability?” within an article), BERT models are better suited. Despite the effectiveness of the BERT model in information retrieval and text understanding tasks, its application in narrowly specialized areas such as law faces several limitations. First, BERT is trained on a general language corpus and, as a rule, does not capture all the nuances of legal terminology and logical relationships in complex cases. Second, the model is poorly adapted to understanding various Kazakh dialects and colloquial forms of the language, which is critical for analyzing legal issues in a regional context. In addition, BERT models do not scale well to analyze long documents, since they are limited in the number of input tokens (usually 512), which makes them less suitable for parsing complex legal texts and contextual dependencies in them.

In this regard, this study decided to use more powerful LLMs, which are capable of processing long input data and demonstrate better results in tasks requiring deep semantic interpretation, logical inference, and contextual response generation. With their help, it is possible to cope with both official legal discourse and simplified queries in dialectal forms of the Kazakh language, providing more accurate and interpretable answers.

3.3. The Utilized Large Language Models

In this work, seven LLMs were deployed to build the QA systems. These LLMs are GPT, GEMMA, Kaz-LLM, LlaMA, Phi, Qwen, and Mistral. Each of them has its configuration specifications and training mechanisms.

Using pre-trained GPT models to build an improved question-answering system, an approach was developed. The architecture of the GPT model is based on the transformer decoder structure and consists of a stack of identical layers with a masked multi-head self-attention mechanism, a feed-forward neural network (FFN), layer normalization, and residual connections. The masked self-attention ensures that the model predicts each token based only on preceding tokens, enabling autoregressive generation. Each attention block is followed by an FFN that expands and contracts the dimensionality of the input, using ReLU or GELU activation functions. To preserve the sequential nature of language, GPT adds positional embeddings to the input token embeddings, allowing the model to encode word order. The final output layer projects the hidden states to a vocabulary-sized vector, from which the next token is predicted using a softmax function [59].

Using pre-trained GPT models to build an improved question-answering system, an approach was developed. Among the GPT models, the most advanced are GPT-3.5-turbo, GPT-4, and GPT-4o mini, which offer complex data analysis capabilities. GPT-3.5-Turbo supports long contextual queries, which allows it to process extended user dialogs and document data. With the development and update of versions, the GPT-3.5-Turbo model was replaced by o4-mini since this model is cheaper, more functional, and faster. It accepts text and image data, making it ideal for fine-tuning small datasets. GPT-4o is a universal flagship model that can solve a wide range of problems. This model allows you to achieve excellent results, but it is more expensive. During the retraining process of the question-answering system in the legal field, the GPT-4o mini model was selected because it is fast, highly functional, and more cost-effective compared to GPT-4. The further training process involved the following steps: preliminary data preparation, uploading the data to the OpenAI service, and initiating the training process.

GEMMA is an open language model developed by Google DeepMind based on the Gemini architecture [60]. It is based on a transformer architecture similar to LLaMA and PaLM, specifically optimized for dialog tasks that involve the generation and processing of natural languages. This model has a configuration of 2, 7, and 9 billion parameters. The architecture of the GEMMA model includes Multi-Query Attention, revealing respective attention variants, Rotary Positional Embedding (RoPE) for reducing a model's size, GeGLU Activations that replace standard ReLU non-linearity, and RMSNorm which normalizes transformer sublayers. In building a legal question–answer system, a model with 9 billion parameters was used, which is capable of accurately perceiving and interpreting user queries to provide logical, informative, and relevant answers. GEMMA is supported in the Hugging Face library and is well-suited for Supervised Fine-Tuning (SFT), Low-Rank Adaptation (LoRA), and Prompt Tuning methods. Another advantage of this model is its efficient deployment on local and server machines.

The KazLLM is a specialized language model focused on the Kazakh language. It is designed for text understanding and task generation. The KazLLM is based on the architectures of BERT, RoBERTa, and the LLaMA family, as well as other transformers. It employs advanced architectural features such as RoPE, RMSNorm, and SwiGLU activations [61]. The tokenizer uses a vocabulary of 128,000. The KazLLM was trained on a large-scale multilingual corpus of 150 billion tokens, which includes filtered web data, Wikipedia, and diverse Kazakh-specific sources such as news and educational content. Following a pre-training phase focused on next-token prediction, the model underwent Supervised Fine-Tuning and Direct Preference Optimization to improve alignment and instruction-following capabilities. The architecture supports a context length of 8192 tokens and integrates performance optimizations like Grouped-Query Attention and FlashAttention v2 for improved efficiency during inference. The KazLLM considers various features of the language's morphology, syntax, and semantics, enabling it to outperform multilingual

analogs in tasks that require in-depth knowledge of the Kazakh language, and is retrained on Kazakh texts from various sources. Due to language adaptation, the KazLLM copes well with processing complex grammatical structures, which is especially important in building the legal sphere in Kazakhstan. The model with 8 billion parameters was chosen for training to build the QA system.

LLaMA is a series of language models developed by Meta and designed for a wide range of natural language processing tasks, including text generation, summarization, and question-answering systems [44]. The LLaMA architecture belongs to the class of decoder-only transformer models similar to GPT, which makes it effective in generative tasks. In this architecture, the input of each transformer sublayer is normalized with the RMSNorm function. The SwiGLU activation function replaces the ReLU non-linearity. The absolute positional embeddings are removed in favor of RoPE embeddings. With an open architecture and high compatibility with the Hugging Face Transformers and DeepSpeed libraries, LLaMA integrates well into modern ML pipelines and supports retraining using LoRA, QLoRA, and other effective techniques. In training on legal data, the LLaMA 3 model was utilized, which is distinguished by its higher accuracy, improved integration, scalability, and the ability to be retrained. In the experiments, the LLaMA of 3 billion parameters was utilized for training the QA system.

Phi is a language model developed by Microsoft Research to efficiently solve problems in the field of natural language processing [62]. The Phi model employs full attention across the context window. It is trained with a curriculum emphasizing high-quality synthetic data, including techniques such as multi-agent prompting, instruction reversal, and self-revision. Post-training involves Supervised Fine-Tuning (SFT) and two rounds of Direct Preference Optimization (DPO), with one stage targeting pivotal tokens. Despite the simplicity of the core transformer design, Phi achieves high performance across QA reasoning through innovations in training data and alignment. Phi demonstrates high efficiency in text analysis, question-answering systems, and mathematical reasoning. Phi-1, Phi-1.5, and Phi-2 focused on code generation, improving logical thinking, and enhancing language understanding. Phi-3 models were trained on 3.3 to 4.8 trillion tokens, including a wide range of synthetic and web data, ensuring their high knowledge density. A model with 14 billion parameters was used to develop a question-answering system in the legal field of the Republic of Kazakhstan.

Qwen is a language model developed by Alibaba Cloud in 2023. Since then, several generations of models have been released, Qwen 2, Qwen 2.5, and Qwen 3, each improving in performance, multi-language support, and reasoning ability [63]. The architecture of Qwen is based on the LLaMA model, which is widely used as a top open-source LLM. The modifications of the architecture include RoPE embeddings that use FP32 precision for achieving higher accuracy, pre-normalization (RMSNorm) that replaced the traditional layer normalization technique, and the SwiGLU activation function based on Swish and Gated Linear Unit (GLU). Qwen 2 showed excellent results on 15 benchmarks, including language comprehension, text generation, question-answering systems, programming, and logical reasoning. Introduced in January 2025, Qwen 2.5 was trained on 18 trillion tokens, significantly improving its knowledge of programming and mathematics. Qwen 3, released in April 2025, is the family's latest model and includes dense and sparse architectures. It was trained on 36 trillion tokens. The Qwen 2.5 model with 7 billion parameters was used to develop the legal question-answering system.

Mistral is another family of LLMs developed by Mistral AI in 2023. The primary goal of developing these models is to create highly efficient and accessible open-source models that can match or surpass OpenAI's performance. The models are based on a transformer architecture that utilizes Grouped-Query Attention (GQA) and Sliding Window Attention

(SWA) mechanisms, offering high performance and efficient text processing. GQA reduces the number of key-value heads (8) while maintaining the number of query heads (32). This design cuts memory usage during decoding and speeds up inference, especially useful in large-scale deployments. Mistral adopts SWA, where each token can only attend to a fixed-size window ($W = 4096$) of previous tokens. It enables scalable attention while allowing deeper layers to propagate information beyond the window (up to $k \times W$ tokens for k layers), maintaining long-range context efficiently. Another feature of the Mistral model is the Rolling Buffer Cache, where the oldest entries are overwritten in a circular fashion. This dramatically reduces memory requirements without sacrificing quality. In the pre-fill and chunking, Mistral chunks the input and fills the attention cache in parts, improving memory handling during generation. Attention is computed over both the current chunk and the sliding cache window. As this study employs the Mistral model, which has 7 billion parameters comparable in performance to LLaMA, the proposed enhancements allow Mistral 7B to outperform even larger models like LLaMA 2 13B and LLaMA 1 34B on many benchmarks while maintaining smaller size and faster throughput, making it an efficient yet powerful alternative for real-world language model deployment.

The agglutinative nature of Kazakh, characterized by polysyllabic word forms with numerous suffixes, makes the tokenization task more challenging compared to Indo-European languages. Such morphologically rich structures lead to a high variability of word forms, which poses a challenge for both frequency models and transformers with fixed vocabularies. SentencePiece, which is used in GEMMA, KazLLM, LLaMA, Phi, and others, is based on statistical byte-pair encoding (BPE) or unigram model training, where no pre-splitting of text into words is required. It automatically identifies frequency substrings, regardless of morphology. However, in agglutinative languages, this leads to a fragmentation of word forms into uninterpretable subparts. This can lead to excessive chunking, which prevents models from generalizing effectively.

The GPT-3.5, GPT-4, and GPT-4o mini tokenizers use byte-level BPE, where all input data is encoded in bytes. This makes them language-neutral, but they do not take into account morphological boundaries. Therefore, in the Kazakh language, they often divide words into artificial fragments that do not reflect the structure of the root and affixes. The Qwen and Phi models in the basic version are not adapted for agglutinative languages but can be re-tokenized through a customized SentencePiece or used dictionaries. They exhibit low efficiency in complex forms of agglutination, particularly without the localization of tokenizers.

In terms of preprocessing, the following steps were taken:

- Cleaning the text from noise and incorrect artifacts;
- Bringing texts to a uniform spelling (including normalization of dialect forms when possible);
- Minimal lemmatization when training separate versions of models to compare the impact of morphemic regularity.

Such measures allow adapting universal models (e.g., GEMMA, LLaMA, and Qwen) to the peculiarities of Kazakh grammar and using them in real applications.

4. Experiments

The collected Adilet, Zqai, Gov, and synthetic datasets were all split into 95% training and 5% testing parts. The quantity of every dataset is shown in Table 6. During the training stage, the Zqai and Gov datasets were combined, as their structure is generally similar, allowing for an increase in size.

Table 6. The statistics of the training and testing parts of the datasets.

Dataset	Train	Test	Total (Files)
Adilet	961	51	1012
Zqai and Gov	1273	67	1340
Synthetic	3576	188	3764

4.1. Model Training

The following models were trained on the described prepared datasets: GPT, GEMMA, Kaz-LLM, LlaMA, Phi, Qwen, and Mistral. Each of them has its configuration parameters.

In training a GPT model, the GPT-4o mini model was configured using the official OpenAI API and a message format corresponding to the chat scheme between the user and the assistant. Each QA training pair was formed as a message structure containing two consecutive elements: a user request with the “user” role and a model answer with the “assistant” role. The recording format looked like this:

```
QA.append({
  "messages": [
    {"role": "user", "content": question},
    {"role": "assistant", "content": answer}
  ]
})
```

The resulting QA list was converted into a JSON file that complied with the OpenAI API requirements for fine-tuning. Each line of the file represented one full-length training session of the model.

```
with open("QA_finetune_adilet.jsonl", "rb") as f:
    response = client.files.create(
        file=f,
        purpose="fine-tune"
    )
    fine_tune_response = client.fine_tuning.jobs.create(
        training_file=file_id,
        model="gpt-4o-mini-2024-07-18"
    )
```

During the fine-tuning of the GPT-4o mini model, the configuration followed OpenAI’s official API format, using a structured chat-based messaging scheme where each QA pair consisted of a “user” prompt followed by an “assistant” response. While this format ensures compatibility with the API and simulates real-world dialog, the effectiveness of the fine-tuning heavily depends on the prompt engineering strategy—particularly how user inputs are phrased. The default approach employed a zero-shot prompting style, where prompts were direct and concise questions, such as “Қазақстанда апелляция беру тәртібі қандай?” (“What is the appeal procedure in Kazakhstan?”) or “Сот шешіміне қалай шағымдануға болады?” (“How can I appeal a court decision?”). These forms assume that the model can infer the correct structure and tone without any contextual guidance. To improve generalization and output quality, alternative prompt strategies should be considered. One such approach is few-shot prompting, where the prompt includes an example before the target question. For example: “Азаматтық талап қалай беріледі? Талапкер тиісті сотқа өтініш беруі тиіс... Енді: Апелляциялық шағым қалай рәсімделеді?” (“Example: How is a civil claim filed? The plaintiff must file an application with the appropriate court... Now: How is an appeal filed?”). This structure primes the model with an expected answer format. Another effective method is chain-of-thought prompting, where the question explicitly encourages a step-by-step explana-

tion, such as “Апелляция беру үшін қандай қадамдарды орындау қажет? Әр қадамды түсіндіріп беріңіз” (“What are the steps to follow to file an appeal? Please explain each step”). This guides the model to reason logically through the process. Additionally, instructional prompts can be used to shape tone and audience awareness: “Апелляциялық процесті бірінші рет естіп тұрған адамға түсіндіретін зангер ретінде жауап беріңіз” (“Respond like a lawyer explaining the appeals process to someone who is hearing it for the first time”). In future fine-tuning iterations, it is advisable to include a mix of prompt styles to better prepare the model for diverse user inputs. This includes varying the tone (e.g., formal, casual, urgent), audience assumptions (e.g., student, citizen, lawyer), and clarity levels (e.g., vague or compound queries). Such diversity in prompt engineering will not only improve robustness but also enhance the model’s usability in real-world legal QA applications.

The GEMMA model was fine-tuned with 9 billion parameters. The training dataset was composed of structured prompt-response pairs. Each interaction was wrapped in dialog-style tags (<start_of_turn>user, <start_of_turn>model) to match the conversational nature of the model’s architecture. The model was trained using causal language modeling (CLM) objectives on a number of examples, with an EOS token marking the end of each interaction. The tokenizer used was SentencePiece, which was consistent with GEMMA’s original pre-training. The fine-tuning process focused on improving factual consistency, domain grounding, and fluency in the Kazakh language. This model demonstrates strong capabilities in providing precise legal references, procedural clarifications, and domain-specific responses within a conversational context. It is deployed via a local server/API and optimized for GPU inference with quantized weights for reduced latency.

The configuration template is shown as follows:

```
"""<start_of_turn>user
{}\\n{}<end_of_turn>
<start_of_turn>model
{}<end_of_turn>"""+EOS_TOKEN
```

Despite the structural consistency, prompt wording and engineering have a significant role in ensuring the model responds effectively across diverse query types.

For example, a base prompt could be framed as follows:

```
<start_of_turn>user
Күкүқтық көмек көрсетеу үшін қайда жүгінуге болады? (Where can I turn for legal assistance?) <end_of_turn>
<start_of_turn>model
Сіз заң көмегін алу үшін тұргылышты жеріндеңігі адвокаттар алқасына немесе мемлекеттік құқықтық ақпарат порталдарына жүгіне аласыз (You can contact your local bar association or state legal information portals for legal assistance) <end_of_turn>
```

This reflects a zero-shot strategy, where the model infers its behavior from the conversational pattern without additional guiding examples. However, alternative prompt strategies can significantly enhance performance in complex or ambiguous cases. For example, chain-of-thought prompting could help the model reason through steps logically:

```
<start_of_turn>user
Сол шешіміне шағым беру процесін кезең-кезеңімен түсіндіріп беріңіз (Explain the process of appealing a court decision step by step) <end_of_turn>
```

The fine-tuning of the KazLLM and Qwen model was implemented using the supervised instruction-tuning approach with legal and administrative QA data in Kazakh. The dataset was formatted using a structured prompt template that separates the instruction, input, and expected response, enabling models to generalize across a wide range of task

types and input formulations. The format of the instructions enabled better task generalization, response grounding, and improved interpretability for domain experts. The models were trained using a CLM loss with EOS token boundaries, and the inference was structured to support both single-turn and multi-turn completion.

The configuration template is shown as follows:

```
### Instruction:  
{}  
### Input:  
{}  
### Response:  
{}"""+EOS_TOKEN
```

The structured format allowed the use of various prompt engineering strategies to optimize performance for diverse user intents. A simple zero-shot instruction is like

```
### Instruction:  
Сұраққа нақты, қысқаша жауап беріңіз (Answer the question clearly and concisely)  
### Input:  
Сом шешіміне апелляцияны қай мерзімде беру керек? (When should an appeal against a court decision be filed?)  
### Response:  
Апелляция сом шешімі шыққан күннен бастап 15 күн ішінде берілуді түсік (The appeal must be filed within 15 days from the date of the court decision)
```

To support more complex reasoning, chain prompting can be applied by explicitly requesting step-by-step explanations:

```
### Instruction:  
Сом ісін қайта қаралу процесін кезең-кезеңімен түсіндіріңіз (Explain the judicial review process step by step)  
### Input:  
Қайта қаралу бастамасын қалаї көтеруге болады? (How to raise the initiative for reconsideration?)  
### Response:  
1. Ең алдымен, шешімнің күшіне енгеніне қараластан, жаңа мән-жайлар анықталуы керек (First of all, new circumstances must be identified, regardless of whether the decision has entered into force)  
2. Содан кейін өтініш сомқа жазбаша түрде беріледі (The application is then submitted to the court in writing)
```

This structured setup enhances the model to find linguistic variation from task semantics, supporting prompt diversity and enabling more robust generalization across legal tasks in the Kazakh language.

The LLaMA-based QA model was built using a 3-billion-parameter model aligned for chat-based instruction following. Fine-tuning was performed using Meta's ChatML prompt structure with special tokens like <|begin_of_text|>, <|start_header_id|>, and <|eot_id|>. This format segmented roles explicitly into system, user, and assistant, which helped the model maintain coherence across multiple conversational turns. The dataset was composed of Kazakh legal QA entries, where context was delivered through linked legal documents, and responses followed formal stylistic norms. This configuration offered compactness, low resource demand, and responsiveness suitable for GPU-light deployments.

The LLaMA model's configuration is presented as follows:

```
"""<|begin_of_text|><|start_header_id|>system
```

```
{|<|end_header_id|>{|<|eot_id|>
<|start_header_id|>user
{|<|end_header_id|>{|<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
{|<|eot_id|>"""+EOS_TOKEN
```

The prompt structure was crucial for the effectiveness of the model, heavily dependent on how instructions and user queries were phrased. For instance, a basic zero-shot prompt with system role clarification might appear as follows:

```
<|begin_of_text|><|start_header_id|>system
Сіз Қазақстан заңнамасы бойынша сұрақтарға ресми және нақты жауап беретін көмекші боласыз (You will be an assistant who will provide official and accurate answers to questions on Kazakhstani legislation)<|end_header_id|><|eot_id|>
<|start_header_id|>user
Жер теліміне менисік құқығын қалай алуға болады? (How to obtain ownership of a land plot?)<|end_header_id|><|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
Жеке менисікке жер телімін алу үшін сіз жергілікті атқаруышы органға өтініш беруіңіз керек. Сонымен қатар, жер кадастрынан қажетті құжаттарды ရәсімдеу қажет. (To obtain a land plot for private ownership, you must apply to the local executive body. In addition, you must obtain the necessary documents from the land cadastre)<|eot_id|>
```

The Phi-based QA model was configured using Microsoft's ChatML-lite format, utilizing tags such as `<|im_start|>`, `<|im_sep|>`, and `<|im_end|>` to mark role boundaries. This model, likely based on Phi-2 with around 1.3 billion parameters, was instruction-tuned for Kazakh legal queries. The system prompt defined the assistant's behavior, while user questions and assistant answers were clearly demarcated using lightweight separators. Fine-tuning was performed on legal corpora with EOS-based sequence termination. Due to its compact size and efficient tokenization, the Phi model proved ideal for edge-device deployment and lightweight chatbot integrations while still retaining high accuracy for short, factual responses in Kazakh.

The Phi model's configuration is presented as follows:

```
"""<|im_start|>system<|im_sep|>
{|<|im_end|>
<|im_start|>user<|im_sep|>
{|<|im_end|>
<|im_start|>assistant<|im_sep|>
{|<|im_end|>"""+EOS_TOKEN
```

The prompt design was central to ensuring that the model maintained accurate and formal responses. A standard prompt is structured as follows:

```
<|im_start|>system<|im_sep|>
Сіз Қазақстан заңнамасы бойынша сұрақтарға қысқа әрі нақты жауап беретін заң кеңесшіcіz (You are a legal advisor who provides concise and clear answers to questions on Kazakhstani legislation)<|im_end|>
<|im_start|>user<|im_sep|>
Неке қио үшін қандай құжаттар қажет? (What documents are needed to get married?)<|im_end|>
<|im_start|>assistant<|im_sep|>
Неке қио үшін жеке күәлік, неке қио туралы өтініш және мемлекеттік баж салығын төлеңені туралы түбіртек қажет (To get married, you need an identity card, a marriage application, and a receipt for paying the state fee)<|im_end|>
```

In chain responses, the user prompt can explicitly request structured answers:

<|im_start|>user<|im_sep|>

Апелляциялық шағым беру қадамдарын ретімен сипаттаңыз (Describe the steps in order of filing an appeal)<|im_end|>

<|im_start|>assistant<|im_sep|>

1. Сом шешімімен келіспеген жағдайда апелляция жазылады (If you disagree with the court's decision, you can file an appeal)

2. Шағым шешім шыққан күннен бастап 15 күн ішінде берілуі керек (The appeal must be filed within 15 days from the date of the decision)<|im_end|>

By incorporating varied prompts, this configuration allows the Phi model to perform reliably in constrained setups without compromising legal accuracy or user clarity.

The 7 billion Mistral model was fine-tuned using a ChatML-inspired prompt format, structured with the special tokens <|im_start|>, <|im_end|>, and an end-of-sequence EOS_TOKEN. Each training instance consisted of a clearly defined system role prompt, a user query, and an assistant response. This formatting enabled the model to maintain a proper conversational structure and ensured clean separation between roles, enhancing alignment to task instructions. The training objective was CLM, and the fine-tuning approach preserved the original Mistral tokenizer and positional encoding. This configuration supported the development of a fast, compact, and accurate assistant capable of delivering fluent and legally valid responses in formal Kazakh.

The Mistral model's configuration is presented as follows:

```
<|im_start|>system
{}<|im_end|>
<|im_start|>user
{}<|im_end|>
<|im_start|>assistant
{}<|im_end|> + EOS_TOKEN
```

Prompt design for the Mistral model had a critical role in aligning the model's behavior with domain-specific expectations. A standard prompt is structured as follows:

<|im_start|>system

Сіз Казақстан заңдарына сүйеніп, заңды сұрақтарға нақты және ресми түрде жауап беретін кеңесші боласыз (You will be a consultant who will provide clear and official answers to legal questions based on the laws of Kazakhstan)<|im_end|>

<|im_start|>user

Жеке қасінкерді тіркелу үшін не ісмей керек? (What do I need to do to register as a sole proprietor?)<|im_end|>

<|im_start|>assistant

Жеке қасінкер ретінде тіркелу үшін eGov порталында өтініш толтырып, ӘЦҚ арқылы қол қою қажет. Сонымен қатар, салық органдарына тіркелу жүргізіледі (To register as an individual entrepreneur, you need to fill out an application on the eGov portal and sign it using an electronic signature. In addition, registration with the tax authorities is carried out)<|im_end|>

A chain response is like this:

<|im_start|>user

Неке бұзы процесін кезең-кезеңімен түсіндіріп беріңіз (Explain the divorce process step by step)<|im_end|>

<|im_start|>assistant

1. Алдымен некені бұзы туралы өтініши АХАТ бөліміне немесе сотқа беріледі (First, an application for divorce is filed with the Civil Registry Office or the court)

2. Егер екі тарапта келіссе, процесс жөнілдемілген түрде өтеді (If both parties agree, the process will be simplified) <|im_end|>

Through this structure, the Mistral model achieved a high degree of controllability and response clarity, essential for reliable deployment in formal or advisory settings.

The models were fine-tuned using the Quantized Low-Rank Adaptation (QLoRA) approach in 4-bit quantization (quantization type: nf4, with double quantization and float16 computations) for efficient training on limited resources. The model architecture was supplemented with adaptive LoRA layers, with the rank parameter r set to 16, the lora_alpha coefficient set to 16, and zero dropout, focusing on the projections q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, and down_proj. Data preparation included tokenization and formatting using EOS tokens in the style of Alpaca prompts.

4.2. Results and Score Evaluation

The training was performed over 60 steps using a batch size of 2 and gradient accumulation over four iterations, with an initial learning rate of 2×10^{-4} and the AdamW (8-bit) optimizer. The control metrics (loss, perplexity, and learning rate) were also logged at each step and saved in a .csv file for subsequent analysis.

The quality of the trained models was assessed using various metrics, such as ROUGE and METEOR, which measure similarity to reference answers. They enable us to measure the similarity between the model's answers and reference answers, which helps us objectively assess the accuracy and completeness of the generation.

The ROUGE metric evaluates how well the generated answers match the reference one by the number of matching words or phrases. Several variants of this metric, such as ROUGE-1, ROUGE-2, ROUGE-N, and ROUGE-L, are used as the unigram matches of the generated and reference texts and are measured by (1)

$$\text{ROUGE}-1 = \frac{\sum_{w \in \text{Reference}} \text{Count}_{\text{match}}(w)}{\sum_{w \in \text{Reference}} \text{Count}(w)} \quad (1)$$

where w is a unigram, and $\text{Count}_{\text{match}}(w)$ is the number of unigrams of the generated text that appeared in the referenced text.

For ROUGE-2, we compute the overlap of two-word sequences (bigrams) as (2)

$$\text{ROUGE}-2 = \frac{\sum_{bg \in \text{Reference}} \text{Count}_{\text{match}}(bg)}{\sum_{bg \in \text{Reference}} \text{Count}(bg)} \quad (2)$$

where bg is a two-word sequence (bigram), $\text{Count}_{\text{match}}(bg)$ is the number of bigrams of the generated text that appear in the reference text, and $\text{Count}(bg)$ is the total number of bigrams in the generated text.

ROUGE-N measures the exact N-gram match as (3)

$$\text{ROUGE}-N = \frac{\sum_{n-\text{gram} \in \text{Reference}} \text{Count}_{\text{match}}(n-\text{gram})}{\sum_{n-\text{gram} \in \text{Reference}} \text{Count}(n-\text{gram})} \quad (3)$$

where $\text{Count}_{\text{match}}(n-\text{gram})$ is the number of $n-\text{gram}$ in the candidate that also appear in the reference, $\text{Count}(n-\text{gram})$ is the total number of $n-\text{gram}$ in the reference, and N is the $n-\text{gram}$ size.

ROUGE-L takes into account the Longest Common Subsequence (LCS), making it useful for natural language evaluation. ROUGE-L precision, recall, and F1-score are computed as (4)–(6)

$$\text{ROUGE}-L = \frac{\text{LCS}(X; Y)}{\text{length}(X)} \quad (4)$$

$$ROUGE - L = \frac{LCS(X; Y)}{\text{length}(Y)} \quad (5)$$

$$ROUGE - L_{F1} = \frac{(1 + \beta)^2 \times \text{Precision} \times \text{Recall}}{\text{Recall} + \beta^2 \times \text{Precision}} \quad (6)$$

where X is the candidate sentence, Y is the reference sentence, $\text{length}(X)$ is the number of tokens in the candidate sentence, $\text{length}(Y)$ is the number of tokens in the reference sentence, $LCS(X; Y)$ is the length of the Longest Common Subsequence between X and Y , and β is a balance coefficient.

The METEOR metric was developed for evaluating machine translation but is also effectively used in question-answering systems. Unlike ROUGE, it takes into account not only exact word matches but also their synonymous forms and morphological variations (for example, different forms of the same verb). METEOR features lexical flexibility (accounting for synonymy and morphological variation), word order (penalizing word rearrangements if they change the meaning of the sentence), and semantic matching (incorporating dictionaries and thesauri for better matching).

The metric is calculated by (7)

$$\text{METEOR} = F_{\text{mean}}(1 - \text{Penalty}) \quad (7)$$

where F_{mean} is a harmonic mean value between precision and recall,

$\text{Penalty} = \gamma \times \left(\frac{ch}{m}\right)^\theta$ is a penalty score, ch is the number of chunks, m is the number of matched unigrams, and γ and θ are empirically defined coefficients.

The ROUGE and METEOR scores for all models are shown in Table 7.

Table 7. ROUGE and METEOR scores for models.

Model	Adilet	Zqai + Gov	Synthetic	Average Scores
GPT-4o mini	ROUGE-1: 0.365	ROUGE-1: 0.209	ROUGE-1: 0.3536	ROUGE-1: 0.3092
	ROUGE-2: 0.251	ROUGE-2: 0.0783	ROUGE-2: 0.1968	ROUGE-2: 0.1754
	ROUGE-L: 0.329	ROUGE-L: 0.1277	ROUGE-L: 0.3324	ROUGE-L: 0.2630
	METEOR: 0.370	METEOR: 0.190	METEOR: 0.400	METEOR: 0.3200
GEMMA	ROUGE-1: 0.081	ROUGE-1: 0.105	ROUGE-1: 0.208	ROUGE-1: 0.131
	ROUGE-2: 0.019	ROUGE-2: 0.027	ROUGE-2: 0.062	ROUGE-2: 0.036
	ROUGE-L: 0.073	ROUGE-L: 0.101	ROUGE-L: 0.197	ROUGE-L: 0.124
	METEOR: 0.071	METEOR: 0.068	METEOR: 0.186	METEOR: 0.108
KazLLM	ROUGE-1: 0.094	ROUGE-1: 0.106	ROUGE-1: 0.157	ROUGE-1: 0.119
	ROUGE-2: 0.03	ROUGE-2: 0.026	ROUGE-2: 0.075	ROUGE-2: 0.044
	ROUGE-L: 0.084	ROUGE-L: 0.099	ROUGE-L: 0.157	ROUGE-L: 0.113
	METEOR: 0.066	METEOR: 0.066	METEOR: 0.093	METEOR: 0.075
LLaMA	ROUGE-1: 0.069	ROUGE-1: 0.075	ROUGE-1: 0.15	ROUGE-1: 0.098
	ROUGE-2: 0.038	ROUGE-2: 0.01	ROUGE-2: 0.075	ROUGE-2: 0.041
	ROUGE-L: 0.069	ROUGE-L: 0.066	ROUGE-L: 0.15	ROUGE-L: 0.095
	METEOR: 0.119	METEOR: 0.047	METEOR: 0.165	METEOR: 0.110
Phi	ROUGE-1: 0.081	ROUGE-1: 0.079	ROUGE-1: 0.146	ROUGE-1: 0.102
	ROUGE-2: 0.019	ROUGE-2: 0.012	ROUGE-2: 0.051	ROUGE-2: 0.027
	ROUGE-L: 0.077	ROUGE-L: 0.072	ROUGE-L: 0.138	ROUGE-L: 0.096
	METEOR: 0.083	METEOR: 0.043	METEOR: 0.152	METEOR: 0.093
Qwen	ROUGE-1: 0.065	ROUGE-1: 0.091	ROUGE-1: 0.183	ROUGE-1: 0.113
	ROUGE-2: 0.017	ROUGE-2: 0.017	ROUGE-2: 0.07	ROUGE-2: 0.035
	ROUGE-L: 0.063	ROUGE-L: 0.087	ROUGE-L: 0.177	ROUGE-L: 0.109
	METEOR: 0.055	METEOR: 0.05	METEOR: 0.172	METEOR: 0.092
Mistral	ROUGE-1: 0.067	ROUGE-1: 0.114	ROUGE-1: 0.176	ROUGE-1: 0.119
	ROUGE-2: 0.018	ROUGE-2: 0.029	ROUGE-2: 0.083	ROUGE-2: 0.043
	ROUGE-L: 0.065	ROUGE-L: 0.103	ROUGE-L: 0.176	ROUGE-L: 0.115
	METEOR: 0.048	METEOR: 0.078	METEOR: 0.168	METEOR: 0.098

In the experimental results, it is notable that GPT-4o mini outperformed other models in all three datasets, Adilet, Zqai + Gov, and synthetic, as in the ROUGE and METEOR metrics, with the highest average scores of ROUGE-1: 0.3092, ROUGE-2: 0.1754, ROUGE-L: 0.2630, and METEOR: 0.3200 exceeding the scores of models such as GEMMA, KazLLM, and LLaMA, and Phi GPT-4o mini also showed a superior performance in capturing both lexical overlap and semantic alignment, particularly excelling on the synthetic dataset, where it achieved a METEOR score of 0.400. The GEMMA achieved 0.131 in ROUGE-1, 0.036 in ROUGE-2, 0.124 in ROUGE-L, and 0.108 in METEOR, indicating that GPT-4o mini significantly surpassed this model. The KazLLM, a localized model of the Kazakh language, demonstrated moderate performance, scoring 0.119 in ROUGE-1, 0.044 in ROUGE-2, 0.113 in ROUGE-L, and 0.075 in METEOR. This model outperformed Phi, Qwen, and Mistral but fell short of the top-tier systems. The Mistral, Qwen, and KazLLMs exhibited comparable mid-range results, with ROUGE-1 ranging between 0.110 and 0.120 and METEOR between 0.075 and 0.098, reflecting moderate alignment with the ground truth. Although LLaMA achieved the lowest score of 0.098 on ROUGE-1, it achieved the highest METEOR score of 0.110 among non-GPT models, suggesting a stronger focus on semantic rather than lexical accuracy. Phi consistently showed the weakest performance, particularly on ROUGE-2 and ROUGE-L, indicating difficulty in generating coherent multi-word expressions and structured sequences. These baseline models consistently yield lower scores across all metrics, offering challenges in processing legal texts in a morphologically rich and syntactically variable language like Kazakh. Overall, the findings highlight the effectiveness of large general-purpose models like GPT-4o mini in legal language tasks, while also emphasizing the need for further development and fine-tuning of localized models such as the KazLLM to bridge the performance gap.

In addition to the counted ROUGE and METEOR scores, the generated output answers were also observed for finding syntactic, semantic, and morphological errors with various hallucination cases. The GEMMA and Qwen models have limited semantic accuracy without additional training on legal texts. Thus, GEMMA outputs morphological errors and hallucinations, since it uses SentencePiece, which is not tailored for agglutinative structures. In this model, the answers contained repeated words and phrases. Some answers were also too long or too short. They could be either extended or made concise, retrieving the most important aspects of the questions. The Qwen model is not sufficiently adapted to the Kazakh legal context. The predictions of the Qwen model were mostly without serious hallucinations, only giving some long answers. Among the errors, some answers do not correspond to the topic at all, in the form of hallucinations. The answers of the LLaMA model were characterized by the presence of technical elements such as <|end_of_text|>, which might be easily eliminated by a regular expression. Some answers were also quite long, making them suitable for shortening. The Mistral model kept the instructions markup as <|im_start|>system\nCategory: Тікелей сұрақтар\n<|im_start|>user\n in its response. Some answers also included word and phrase repetitions that made the answers unclear. The Phi model's answers did not have many hallucinations. Its responses were clear, but some of them were longer or shorter than the average length of other responses. The KazLLM was especially significant in forming precise and grammatically correct answers. Nevertheless, the fine-tuning stage of the KazLLM played a significant role, as the basic model was unable to respond without errors. The most common mistake was duplicating the question instead of providing a real, valuable response. In some answers, there were also many repeated words, or they were generally very brief. The answers of the most advanced presented models, GPT-4o mini, demonstrated the absence of any serious mistakes. Moreover, the model aimed to pro-

vide the most comprehensive response possible, which enhanced its efficiency compared to other models.

The reason for conducting experiments on different models, rather than just the KazLLM, is related to the following cases. Many general multilingual LLMs, such as GEMMA, Qwen, Mistral, and Phi, were pre-trained on massive multilingual corpora and are widely available with open weights and efficient inference capabilities. This made them an efficient tool for prototyping and benchmarking in Kazakh, without the need for a large-scale training infrastructure from scratch. At the time of development, the availability of high-quality LLM systems trained on various corpora of Kazakh texts was limited and not widely available. In contrast, general multilingual models had already seen multilingual data, including Kazakh, during pre-training, and they were able to cover a wide range of different topics. In addition, general models served as a foundation to validate feasibility, evaluate baseline performance, and identify Kazakh-specific linguistic challenges. This helped inform later efforts to create and refine dedicated Kazakh-optimized models like the KazLLM. Therefore, other LLMs offered a wide range of output data, compared to the KazLLM, initially trained on a restricted number of topics. It is worth noting that examples of errors for the models and datasets used are presented in Table A3 and Figures A4–A6 in Appendix A.

The plots of the loss function for every model are also shown in Figures 3–8.

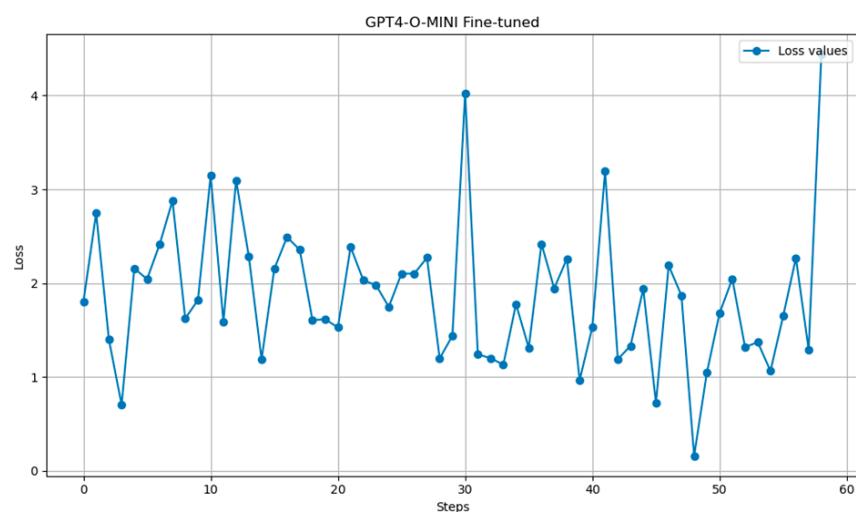


Figure 3. The loss plot of the GPT-4o mini model.

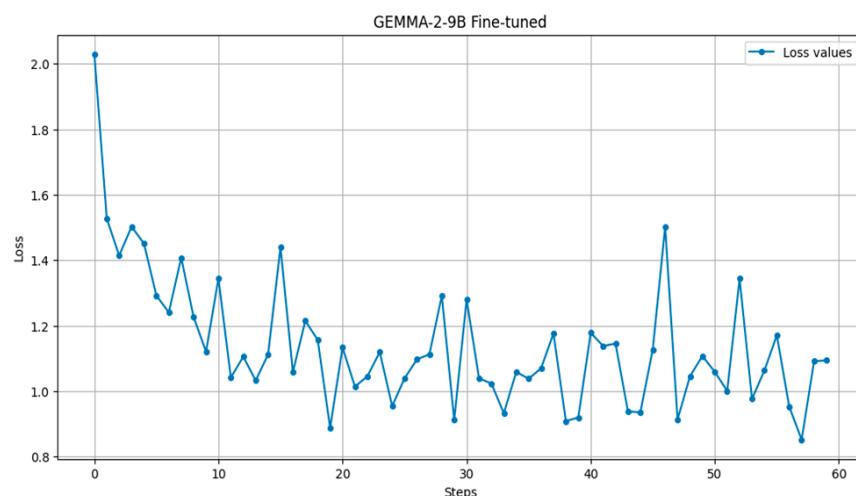


Figure 4. The loss plot of the GEMMA model.

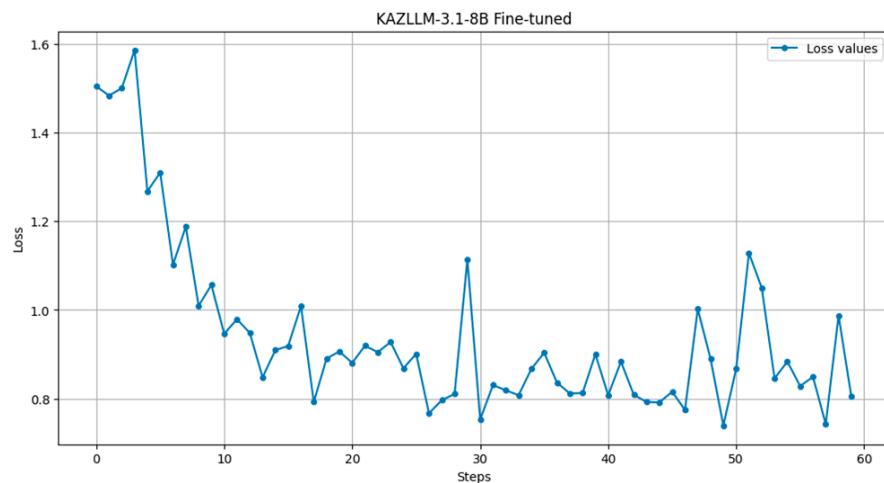


Figure 5. The loss plot of the KazLLM.

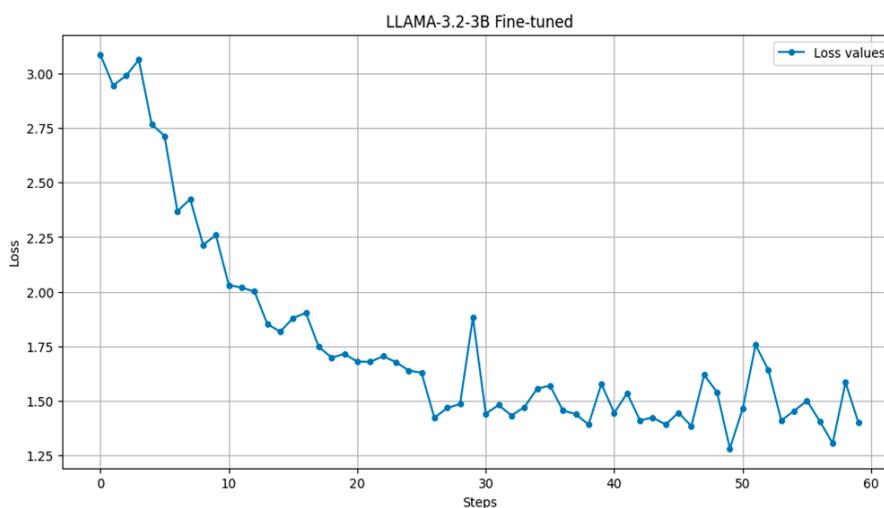


Figure 6. The loss plot of the LLaMA model.

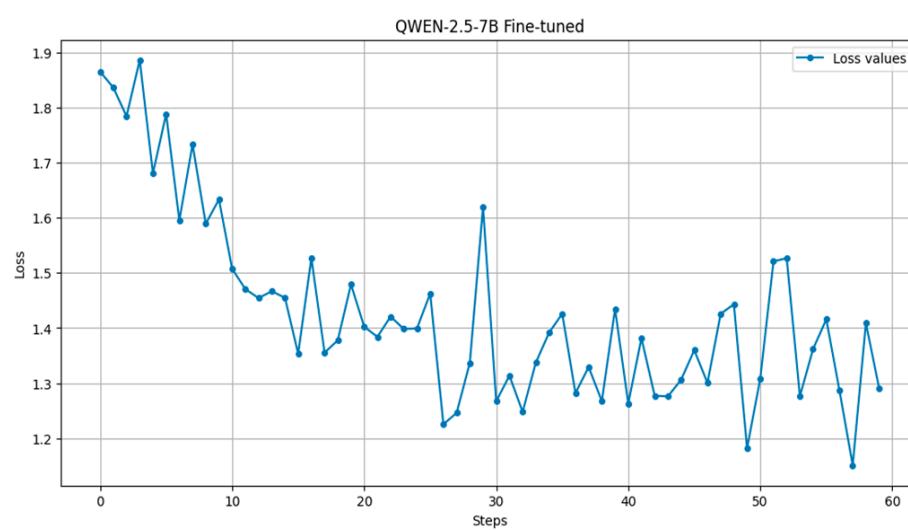


Figure 7. The loss plot of the Qwen model.

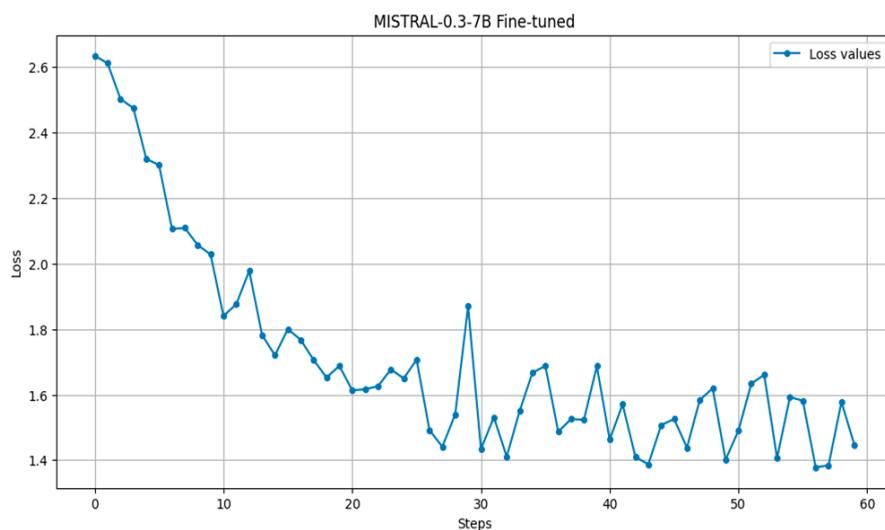


Figure 8. The loss plot of the Mistral model.

These metric trends are further supported by the loss plots (Figures 3–8), which illustrate the convergence behavior of the models. GPT-4o mini exhibits a smooth, steady decline with low-amplitude oscillations and no prolonged plateaus, signaling stable gradients, continued learning late in training, and strong generalization. In contrast, GEMMA and KazLLM exhibited more unstable convergence patterns, characterized by plateaus and fluctuations, which may indicate overfitting or sensitivity to learning rate schedules. GEMMA falls steeply within the first dozen steps and then hovers around 1.0–1.2 values with occasional spikes, while the KazLLM declines rapidly from 1.5 to 0.8 in the first 10 steps and maintains the lowest, steadiest curve thereafter with only brief blips. LLaMA exhibits a large initial descent from 3.1 to 1.5 and then a mildly noisy plateau around 1.4–1.6. Models such as Mistral and Qwen demonstrated moderate stability, but their curves hinted at early saturation and limited potential for further optimization. Qwen drops gradually from 1.9 to 1.3 by step 20 and remains relatively stable with a few mid-training spikes. In addition, Mistral quickly falls from 2.6 to 1.6 within 15 steps before settling into a stable 1.45–1.6 band with intermittent outliers. All models were trained using the CLM loss function based on cross-entropy, combined with QLoRA-based optimization and gradient accumulation. While these techniques improved training stability and efficiency under limited resources, the results suggest that loss minimization alone is insufficient for evaluating domain-specific generation quality in the legal field.

All models employed the standard CLM loss function, based on cross-entropy, which measures the negative log-likelihood of the predicted tokens against the actual next tokens in the sequence. Despite its simplicity, CLM remains effective for training autoregressive language models, particularly in instruction-tuned settings. The training process further benefited from gradient accumulation and QLoRA-based optimization with 4-bit quantization, which enabled the efficient use of limited computational resources while maintaining model quality.

It is important to emphasize that a lower training loss does not always guarantee higher output quality. For example, the LLaMA model displayed only moderate loss reduction but performed competitively in semantic evaluations, achieving the highest METEOR score among lightweight models. This highlights the necessity of interpreting loss curves in conjunction with downstream evaluation metrics, particularly for domain-specific tasks such as legal QA, where linguistic precision and contextual correctness are crucial.

Future improvements may include experimenting with alternative loss functions, such as reinforcement learning from human feedback (RLHF) or contrastive loss functions, which can better capture long-range dependencies, semantic coherence, and legal interpretability beyond token-level accuracy.

In conclusion, the GPT-4o mini model has demonstrated clear superiority in terms of both training efficiency and output quality. Alongside GPT-4o mini, the LLaMA and GEMMA models were selected for additional expert evaluation based on their strong performance across key evaluation metrics. Although the GPT-4o mini proved itself to be the strongest model, its wide productive implementation is related to some aspects that must be paid attention to. One of the concerns is the stability of the Internet connection, which is a critical part of this model's deployment stage. A stable and high-speed connectivity for cloud-based API interactions is a mandatory part of working with this model. Unfortunately, some regions of Kazakhstan cannot boast of a powerful network infrastructure, which could limit seamless integration or widespread adoption. Another critical factor is the financial aspect associated with commercial API access, which generates ongoing costs of its usage. The budget limitations of the public sector may become unsustainable for many companies and organizations, leaving the priority of its use to the large state-owned enterprises. Nevertheless, this research primarily focuses on comparing the models' capabilities in building QA systems, and the provided possibilities of the GPT-4o mini's use are comparatively inexpensive for dozens and hundreds of requests, even for users and organizations with a low budget. So, until the implementation of the full commercial deployment of a QA system, the use of any of the presented models is mostly suitable for everybody.

4.3. Manual Evaluation

Based on the results, an additional manual evaluation of the models was carried out. In order to do the manual evaluation, 100 questions and answers shown on the Adilet website were taken. Automatic evaluation of such systems is traditionally based on the F1 metric, but these indicators do not always reflect all aspects of the quality of the generated answers. In this regard, several studies introduce expert evaluation, which allows for considering the semantic accuracy, completeness, clarity of presentation, and relevance of the answer. Expert evaluation is typically carried out manually, meaning that a group of specialists with the necessary knowledge in the field assesses the system's answers against the specified criteria. This approach provides an opportunity to identify nuances that automatic metrics may overlook [64]. Here are the aspects of evaluation:

1. Qualitative evaluation by multi-aspect criteria. The literature widely discusses the use of multi-aspect expert evaluation, where the following criteria assess answers:
 - Accuracy and completeness: The correspondence of the answer to the facts and all the key aspects of the question.
 - Clarity and logic: Evaluation of the answer's structure, consistency, and comprehensibility.
 - Relevance: The degree to which the answer corresponds to the question asked.
 - Grammatical and stylistic correctness: Compliance with regulatory language requirements and the absence of spelling errors.

Such criteria enable experts to assess whether the standard answer aligns with the quality of the presentation, which is crucial for systems operating in the legal field.

2. Paired comparison method: In this method, experts compare the answers received from different models in pairs, choosing the higher-quality option.

3. Ranking method: Experts are asked to rank all the answers on a quality scale (for example, from 0 to 5). This method enables you to obtain generalized evaluations, which can then be aggregated to determine the average performance of the model.
4. Multiple comparison method: When using this approach, experts evaluate groups of answers, which reduces the load and increases the accuracy of the evaluation, especially when there are many options. This method combines the advantages of paired comparisons and ranking.
5. Delphi method: The Delphi method is an iterative anonymous survey among experts, followed by discussing the results and reaching a consensus. This method is widely used in studies related to forecasting and assessing complex systems, but it requires significant time and organizational costs.

For an expert evaluation of the question–answer system in the Kazakh language, the method of qualitative assessment based on multi-aspect criteria is suitable for the following reasons:

1. Comprehensive quality reflection: This method enables you to evaluate not only the accuracy of matching with standard answers but also additional aspects important to the user's perception of the answer, such as clarity of presentation, logical argumentation, completeness, and relevance of the answer.
2. Taking into account the specifics of the Kazakh language and legal context, when working with the Kazakh language, especially in the legal sphere, it is essential to evaluate how accurately and clearly legal information is conveyed.
2. Flexibility and adaptability: A multi-aspect evaluation allows you to tailor the criteria to the system's specific requirements. This is essential for questions related to complex legal norms, where not only is factual accuracy important but also the structure and completeness of the answer.
3. Increasing the reliability of evaluation: Using multiple criteria reduces the influence of subjectivity in individual evaluations. Aggregating evaluations by various parameters enables a more objective and comprehensive understanding of the system's performance quality.

Thus, qualitative evaluation by multi-aspect criteria is the most suitable method for expert evaluation of the QA system in the Kazakh language since it provides a comprehensive analysis and takes into account all important aspects of the quality of answers, which cannot be achieved using only automatic metrics.

Table 8 shows the criteria for expert evaluation of the legal QA system in the Kazakh language.

Table 8. The criteria for QA system evaluation.

Criterion	1 (Bad)	2 (Below Average)	3 (Average)	4 (Good)	5 (Excellent)
Legal accuracy	Contains serious legal errors and misinterpretations.	Partially correct but includes major inaccuracies.	Generally accurate with some unclear legal aspects.	Accurate with only minor discrepancies.	Fully accurate and reflects Kazakhstani legislation.
Legal completeness	Misses key aspects required to understand the issue.	Covers some aspects but omits critical legal details.	Covers main legal points, lacks details.	Covers nearly all key legal aspects in detail.	Exhaustive and fully discloses legal aspects.
Clarity and logical structure	Unclear and poorly structured.	Inconsistent structure and unclear logic.	Mostly clear but has minor logic issues.	Well-structured and mostly clear.	Impeccably structured and easy to understand.
Relevance and contextuality	Irrelevant and disconnected from the Kazakh legal context.	Partially relevant with weak contextualization.	Mostly relevant, some deviations from context.	Relevant, with minor contextual issues.	Fully relevant, adapted to the Kazakh context.
Legal correctness and professional style	Unprofessional style with numerous terminology and grammar errors.	Poor style with noticeable errors.	Acceptable style with minor terminology issues.	Professional with correct terminology and few errors.	Impeccable legal style and terminology.

The manual evaluation was conducted by a panel of three legal experts, each with at least 7–10 years of professional experience in the Kazakh legal practice. The panel consisted of the following: one legal academic affiliated with a national university law faculty, one government legal advisor involved in citizen legal services, and one head of the state and legal department of the Zhetysu region administration. All experts are fluent in the Kazakh language, including its professional legal terminology. Each expert was given identical model responses to 100 legal questions. The expert evaluation was conducted on a five-point scale based on the following criteria: legal accuracy, legal completeness, clarity and logical structure, relevance, and compliance. For each model, an average score was calculated for each criterion and an overall average score, allowing for the identification of the strengths and weaknesses of each model within the framework of a given task.

Methodology for calculating the average score:

1. Each expert gave scores on a five-point scale for each criterion and for each model.
2. The average score for each criterion was calculated as the arithmetic mean of all scores given by the experts.
3. The overall score for the model was calculated as the arithmetic mean of all scores for all criteria.

The average expert score is calculated as (8)

$$C_i = \frac{\sum_{j=1}^n O_{ij}}{n} \quad (8)$$

where O_{ij} is a grade put by j expert on i criteria, and n is the number of experts.

The full grade of the model is calculated as (9)

$$Total_{score} = \frac{\sum_{k=1}^m C_k}{M} \quad (9)$$

where C_k is an average score on each criterion of m , and m is the number of criteria.

To quantify the agreement between the three raters who manually checked the quality of the generated responses, the Fleiss Kappa coefficient, a statistic that measures the degree of agreement between more than two annotators when using a discrete scale, was used.

The steps for calculating the Fleiss coefficient are the following:

1. Forming the observation matrix: For each criterion in each question, the assessments of three experts were recorded. For example, if, according to the criterion “legal accuracy” for the first question, the scores were 4, 4 and 5, then the corresponding observation row in the table had the following form: [0, 0, 0, 2, 1], where the values represent the number of votes for assessments from 1 to 5.
2. Calculating the agreement for each observation: For each row, the agreement level was calculated using Formula (10):

$$P_i = \frac{1}{n(n-1)} \times \sum n_{ik}(n_{ik} - 1) \quad (10)$$

where $n = 3$ is the number of experts, and n_{ik} is the number of experts who chose category k . This allowed us to determine how often the experts agreed in their judgments on a specific criterion.

3. Forming an average observed agreement (11):

$$\bar{P} = \frac{1}{N} \times \sum P_i, \quad (11)$$

where N is the total number of criteria (questions \times 5).

4. Calculating the expected agreement P_e based on the shares of votes for all categories (12):

$$P_e = \sum (p_k^2) \quad (12)$$

where p_k is the proportion of all ratings given to category k (from 1 to 5).

5. Forming the final Formula (13):

$$k = \frac{\bar{P} - P_e}{1 - P_e} \quad (13)$$

In contrast to the classical scale proposed by Landis and Koch, which is often criticized for its overly optimistic interpretation of low Kappa values, the present study uses a more rigorous approach to assessing inter-rater agreement. This is due to the fact that κ values in the range of 0.2–0.4, although formally designated as “satisfactory agreement,” may in practice reflect significant differences between annotators [65].

Modern research also emphasizes that the interpretation of the Kappa coefficient must take into account the context of the task and sensitivity to data imbalance, especially in the case of complex categorical tasks such as scoring responses in legal question-answering systems [66,67]. In this regard, in the present study, the threshold of “substantial agreement” is set at $\kappa > 0.60$, as shown in Table 9.

Table 9. Interpretation of Fleiss' Kappa values in the present study.

Values	Agreement Interpretation
$\kappa \leq 0.00$	No agreement (below chance level)
0.01–0.20	Weak agreement
0.21–0.40	Restricted or unstable agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement (threshold in this study)
0.81–1.00	Almost complete or complete agreement

The Fleiss coefficient

Fleiss' Kappa calculations are presented in Table 10. The best inter-rater agreement was recorded for the GPT mini model on the Adilet dataset ($\kappa = 0.75$), which indicates high predictability and structural stability of the generated answers. The LLaMA and Gemma models on other datasets showed significant agreement ($\kappa = 0.61$ –0.69), which is also considered acceptable for expert assessment tasks.

Table 10. Result of inter-expert assessment by Fleiss' Kappa coefficient.

Model Name	Adilet	Zqai and Gov	Synthetic
GPT-4o mini	0.75	0.62	0.65
LLaMA	0.61	0.63	0.64
GEMMA	0.62	0.67	0.69

To illustrate the achieved level of inter-annotator agreement, Table A5 in Appendix B presents an example of the highest degree of consensus among experts. This instance was recorded during the evaluation of a response generated by the GPT mini model trained on the Adilet dataset. The ratings provided by all three experts across the five evaluation criteria—legal accuracy, completeness, clarity, relevance, and professional style—varied by no more than one point. This led to a high Fleiss' Kappa value ($\kappa \approx 0.75$), which corresponds to the level of substantial agreement according to contemporary interpretation standards.

In contrast, Table A6 demonstrates an example of the lowest agreement between annotators. It presents one question and a model-generated answer produced by Gemma, trained on the synthetic dataset. Expert evaluations varied considerably across several criteria, particularly in terms of completeness and legal accuracy. These divergences may reflect ambiguity or insufficient specificity in the response. As a result, the Fleiss' Kappa score was lower ($\kappa \approx 0.45$), corresponding to a moderate or inconsistent level of agreement and highlighting the interpretive variability inherent to human judgment in legal question-answering tasks. Moreover, the analysis of disagreement cases highlights the importance of linguistic precision and domain-specific grounding in ensuring reliable expert validation. These findings underscore the critical role of strong inter-rater agreement in evaluating legal QA systems—especially in low-resource language settings, where linguistic nuances can amplify interpretive variation.

Achieving a high level of inter-rater agreement across different datasets confirms the stability of the evaluation criteria and indicates that the model outputs were sufficiently interpretable and structured to support consistent judgments by independent experts. This provides a solid foundation for analyzing the overall expert scores, which highlight the strengths and limitations of each model in greater detail.

The models' answers were also evaluated using an F1-score, which presents a balance between precision and recall. The F1-score metric is used to assess the degree of match between the extracted text and the reference answer. It measures lexical and partial semantic similarity, calculated as a harmonic mean between precision, i.e., the proportion of relevant tokens among all extracted ones, and recall, reflecting the coverage of reference tokens in the extracted text. The F1-score value allows for partially correct matches, in which the found fragment contains deviations from the reference or covers it only partially, thereby providing a more flexible and informative evaluation of the extraction quality compared to strict match metrics. It is calculated as (14)

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (14)$$

Thus, using the proposed expert evaluation method, it is possible to conduct a comparative analysis of the GPT, LLaMA, and GEMMA models and determine the most effective model for a legal QA system in Kazakh. The comparison of the models' F1-scores and expert scores is presented in Table 11, and examples of evaluated questions and answers in the Kazakh language, along with their English translations, are provided in Table A4.

Table 11. The expert evaluation of models.

Model Name	F1-Score Synthetic	F1-Score Adilet	F1-Score Zqai+ Gov	F1-Score Total	Expert Score Synthetic	Expert Score Adilet	Expert Score Zqai + Gov	Expert Score Total
GPT-4o mini	0.171	0.114	0.096	0.127	2.8	3.697	2.893	3.13
LLaMA	0.077	0.082	0.066	0.075	1.06	1.33	1.11	1.17
GEMMA	0.118	0.086	0.049	0.084	2.07	3.16	2.67	2.63

The presented F1 and expert scores were interpreted in the following way. F1-scores of GPT-4o mini significantly outperformed both LLaMA and GEMMA, with the highest total score of 0.127 and the best performance of 0.171 on the synthetic dataset. GEMMA and LLaMA showed noticeably lower F1-scores of 0.084 and 0.075, respectively. These results indicate GPT-4o mini's superior capability in text classification or information retrieval tasks, especially on synthetic datasets. However, it is important to note that all models demonstrated decreased performance on the Zqai + Gov dataset. The expert scores further reinforce the dominance of GPT-4o mini, which earned the highest overall expert rating of

3.13, clearly surpassing scores of GEMMA and LLaMA models. GPT-4o mini was consistently rated higher across all datasets, particularly on the Adilet dataset, where it reached a score of 3.697. GEMMA was second-best in this evaluation, maintaining relatively strong scores across each domain, while LLaMA was consistently lagging behind, especially on the synthetic and Zqai + Gov datasets. Overall, GPT-4o mini exhibited higher performance metrics, proving its effectiveness in handling diverse datasets.

The expert score highlights that automatic metrics are insufficient for evaluating legal QA systems: expert verification reveals aspects that are not reflected in numerical indicators (accuracy of terminology, logical coherence, correctness of legal references). GPT-4o mini demonstrates confident leadership in both automatic metrics and expert assessment, especially in tasks that require working with clear legal wording. The disadvantage in rare cases is excessive paraphrasing, which can reduce the accuracy of citing the law, especially evident on the synthetic dataset. The average expert score is 3.13, which confirms the model's ability to adapt to the legal context, providing high-quality answers both in content and style. LLaMA showed weak results in expert assessment. The model has difficulty processing complex legal syntax and produces less informative answers. For example, in the examples of the Zqai+Gov dataset, it does not maintain the logical structure of the answer, and there are stylistic errors. The lowest score for synthetic = 1.06 due to the high frequency of incomplete or general answers and loss of key details. On the positive side, the model sometimes successfully copes with simple, short questions. GEMMA demonstrates more stable results than LLaMA but is inferior to GPT-4o mini in terms of quality and quantitative metrics. The gap is especially noticeable when working with corpora containing freer syntax and complex formulations (Zqai + Gov). On the corpus, Adilet correctly cites regulations and maintains the official style. And on the synthetic dataset, it showed a decrease in accuracy in rare or non-standard questions, with a simplified presentation of the answer without sufficient detail, which led to a low expert score of 2.07.

In summary, the consolidated analysis confirms that GPT-4o mini currently provides the most accurate, interpretable, and legally valid responses for Kazakh-language legal QA tasks. The combination of automatic and expert human evaluation yielded robust results and underscored the importance of expert judgment in validating domain-specific AI applications in high-stakes fields such as law.

5. Conclusions and Future Work

Today, quality control systems are becoming an integral part of lawyers' work processes, increasing the efficiency of work and the availability of legal services. In Kazakhstan, their development can contribute to the modernization of the legal system and the strengthening of citizens' trust in the judicial system. This study was devoted to the construction of a question-and-answer system for ensuring primary consultation and quality in the legislative field of the Republic of Kazakhstan, taking into account both linguistic features and technological limitations.

Seven LLMs, GPT-4o mini, GEMMA, KazLLM, LLaMA, Phi, Qwen, and Mistral, were trained on separate datasets (Adilet, Zqai, Gov, and synthetic) collected from legislative portals. Each dataset was used in isolation for training and testing, which allowed us to identify the strengths and weaknesses of the models depending on the format and complexity of the source. This made it possible to determine how well each model copes with the following:

- formalized and clearly structured formulations (Adilet);
- complex logical constructions and looser syntax (Zqai);
- citizen-oriented information materials (Gov);
- diverse and non-standard formulations (synthetic).

Quantitative evaluation using ROUGE and METEOR metrics showed that GPT-4o mini confidently outperformed the other models on all datasets, achieving the highest values of ROUGE-1—0.309, ROUGE-2—0.175, ROUGE-L—0.263, and METEOR—0.320. This indicates a high degree of lexical–semantic correspondence with the reference answers. The GEMMA and LLaMA models showed moderate results, while the KazLLM, Mistral, and Phi faced serious difficulties in processing complex legal syntax and semantics.

In addition to the automatic metrics, manual expert evaluation was carried out on five criteria: legal accuracy, completeness, clarity, relevance, and professional style. The results confirmed the quantitative findings: GPT-4o mini received the highest average score of 3.13 out of 5, reflecting its adaptability to legal reasoning and contextual language. GEMMA scored in the middle at 2.63, and LLaMA scored low at 1.17, highlighting the need to complement automated metrics with expert analysis.

According to the F1 metric, GPT-4o mini 0.127 also remained the leader, while GEMMA and LLaMA showed lower values, 0.084 and 0.075, respectively.

The shortcomings identified during the experiments included the following:

- The tendency of individual models (KazLLM and LLaMA) to hallucinations, i.e., generating unreliable facts;
- Repetition of words and phrases in long responses (especially in basic versions without fine-tuning);
- Decreased accuracy when processing legal texts with a complex syntactic structure;
- Low resistance to non-standard or incomplete query formulations.

The aggregated results demonstrate the feasibility of developing high-performance, language-adapted quality control systems for Kazakh-language legal applications. GPT-4o mini is currently the most suitable candidate for implementation in legal support platforms due to its combination of high accuracy, robustness to different text types, and good context adaptability.

The future work will focus on the following:

1. Expanding the training corpora, including collecting real legal queries from citizens and representatives of legal organizations.
2. Improving multilingual adaptation, with priority for the Kazakh language.
3. Integrating methods for generating augmented search information to improve the factual validity and explainability of results.
4. Reducing the number of hallucinations and eliminating repetitions in long answers through additional training and optimization of the architecture.

As part of the development of our legal Q&A system, we plan to create a separate online platform accessible to a wide range of users: legal consultants, civil servants, and citizens. One of the priorities will be integration with national legal portals (for example, Adilet.kz), which will expand the coverage and increase the practical value of the system. At the prototyping stage, load and latency tests have already been conducted, showing an acceptable level of delays and throughput when using optimized LLM architectures and caching of frequently asked queries. In subsequent iterations, a full-fledged API for integration with external services will be implemented, ensuring seamless data exchange between sources of legal information and the intelligent query processing system. Thus, integration with existing legal platforms is a logical and important step, and the future online platform will be the basis for the scalable implementation of intelligent legal advice and support in Kazakhstan.

Author Contributions: Conceptualization, D.R.; Data Curation, D.R., V.K., A.T., and A.S.; Formal Analysis, A.T. and R.A.; Investigation, D.R. and R.A.; Methodology, D.R.; Resources, V.K., A.T., and A.S.; Software, V.K. and R.A.; Validation, A.T.; Visualization, A.S.; Writing—Original Draft, D.R. and A.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the grant project IRN AP 19677835 “Research of models and development of an intelligent question-answer system based on semantic approaches for the state language in the field of legislation of the Republic of Kazakhstan” of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

Data Availability Statement: This study analyzed publicly available datasets. The results obtained and datasets can be found here: <https://github.com/AssiyaSarsenbayeva/QA-system>, accessed on 2 August 2025.

Acknowledgments: We sincerely thank the experts for their professional evaluation and valuable recommendations, which have contributed to improving the quality of the experiment and the reliability of its results.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
NLP	Natural Language Processing
QA	Question Answering
LLM	Large Language Model
ASR	Automatic Speech Recognition
ML	Machine Learning
DL	Deep Learning
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit Ordering
GPT	Generative Pre-Trained Transformer
GPT-4o mini	Optimized version of GPT-4 for inference
LLaMA	Large Language Model Meta AI
GEMMA	Google Efficient Multilingual Model Architecture
Qwen	Alibaba’s Chinese-Centric LLM
Phi	Lightweight language model by Microsoft
KazLLM	Kazakh Large Language Model (locally trained)
RAG	Retrieval-Augmented Generation
IR	Information Retrieval

Appendix A

Table A1. The collection of documents.

Date	6 March 2025	10 March 2025	14 March 2025	20 March 2025	26 March 2025
Total documents:	418,131	418,385	418,648	419,149	419,599
In Kazakh:	208,260	208,392	208,535	208,804	209,017
In Russian:	207,716	207,838	207,958	208,190	208,427
In English:	2154	2154	2154	2154	2154

id	Question	Answer	Question (English translation)	Answer (English translation)
0 0	«Инватахси» қызметін кім пайдалана алады?	Бүгінде қалалар мен облыс орталықтарында инватахси...	Who can use the "Invataxi" service?	Today, invataxi operates in cities and regions...
1 1	Деструктивті діни бірлестіктердің / ағымдардың...	Егер сізді келесі адамдар талап ететін болса, ...	Signs of destructive religious groups/ movements	"If you are being pursued by the following peo...
2 2	Өтінімді қаруа ресімі	Егер адам езінің жеке басын күмелдірдін құж...	Application review procedure	If a person declares that he does not have any...
3 3	АЭС орнына көмір стансиялары мен жаңартылышты...	Көмір және газ стансияларының экология мен хал...	Wouldn't it be more efficient to build coal pl...	"Given the impact of coal and gas power plants...
4 4	Қандай жағдайларда алимент төлеуден босатылуы ...	Егер алимент төлеу туралы көлісім болмagan кез...	In what cases can alimony be exempted?	"If, in the absence of an agreement on the pay...
5 733	Одан арі жинақтау үшін тұрғын үй құрылышы жинақ...	Қазақстандықтардың зейнетеки жинақтары жеткіл...	Top up your housing construction savings accou...	"Pension savings of Kazakhstanis can be used f...
6 734	Непіктен бейтаңыс адамдарға телефоннан қоңырау...	Сіздің телефоныңыздың біраз уақыт алғаннан кей...	Why you can't let strangers call you on the phone	After taking your phone for some time, the att...
7 735	Форвардтық сатып алушу қызыгу үшін құжаттар т...	Форвардтық сатып алушу қызыгу үшін етінін бер...	List of documents for participation in forward...	"When submitting an application for participat...
8 736	Қазақстан Республикасында резидент емештер (ш...	«Жеке сәйкестендіру немірі» ЖСН – Қазақстанда...	Obtaining an IIN for non-residents (foreigners...	"Personal Identification Number" TIN is requir...
9 737	Әскери қызметке шақыру және ету	18 жастан 27 жасқа дейінгі азаматтар әскери қы...	Military service call-up and passage	"Citizens aged 18 to 27 are subject to militar...

Figure A1. The Zqai dataset.

id	Question	Answer	Question (English translation)	Answer (English translation)
0 0	2024ылғы мұғалімдерге берілін жатқан 50мың сы...	Құрметті Рыстың! Өтінішіздің қарал...	Will the 50,000 bonus given to teachers in 202...	Dear Rysty!_x000D_\nWe have reviewed ...
1 1	Қайырлы күн. Мениң ағам осыдан 4 жыл бұрын қай...	Құрметті Елена Владимировна! Өтінішіздің қар...	Good day. My brother died 4 years ago. He owed...	Dear Elena Vladimirovna! Having considered you...
2 2	Сәлеметсіз бе, сізге Надежда Сергеевна Винокоров...	Құрметті Винюкова Н.С.! Өтінішіздің қарал, б...	Hello, this is Nadezhda Sergeevna Vinyukova. I...	Dear Vinyukova N.S.! Having considered your ap...
3 3	Қоғамдық қор ашу туралы мәселе. 1.Тұстары қор...	Құрметті Алинур! Өтінішіздің қарал, бұл мәселе...	The question of opening a public foundation. 1...	Dear Alinur! Having considered your applicatio...
4 4	Қайырлы күн. Қайтыс болған адамның (әкемнің) а...	Құрметті Шагай Е.В.! Өтінішіздің қарал, бул ...	Good day. Please tell me how to get the medica...	Dear Shagay E.V.! Having considered your appli...
5 597	Қайырлы күн Сұрақ мынадай: 2005 жылдың біздің ...	Құрметті Аскар Зулхарнайұлы, Өтінішіздің 05.08.2019 жы...	Good day. The question is this: in 2005, our e...	Dear Askar Zulharnaiuly, Having considered yo...
6 598	Өнерге сәйкес. «Халықты жұмыспен қамтыйтурылым...	Құрметті Байнамаев Женис, Өтінішіздің 08.04.2019 жы...	According to Art. Article 1 of the Law of the ...	Dear Baynamayev Zhenis, Having reviewed your a...
7 599	Қайырлы күн 19.06.19 күнінде күйеүім екеуміз неке...	Құрметті Карина Талгатқызы, Өтінішіздің 2019 жыл...	Good day, on 19.06.19, my husband and I went t...	Dear Karina Talgatkyzy, Having considered your...
8 600	Бұрынғы күйеүі баланың Ресей Федерациясына түр...	Құрметті Татьяна Александрова, Өтінішіздің 2019 жы...	The ex-husband does not give his consent for t...	Dear Tatyana Alexandra, Having considered yo...
9 601	Сәлем. Жақында мен супермаркетке дүкен арапағ...	Жаян: Құрметті Пономарева Александра, Өтінішізді...	Hello. Recently I went shopping at a supermarket...	Answer: Dear Ponomareva Alexandra, Having revi...

Figure A2. The Gov dataset.

id	Questions	Answers	Questions (English translation)	Answers (English translation)
0 0	Поликлиникаға қалай тіркелуге болады?	Таңдалған медициналық мекемеге жеке күелікпен ...	How can I register at a polyclinic?	You need to visit the selected medical institu...
1 1	Медициналық көмек алу үшін қандай құрлаттар қажет?	Медициналық көмек алу үшін жеке күелік пен мін...	What documents are required to receive medical...	To receive medical care, you need an ID card ...
2 2	Тегін дәрі-дәрмекті қалай алуға болады?	Тегін дәрі-дәрмекті дәрігердің рецепті бойынша...	How can I get free medicines?	Free medicines can be obtained with a doctor's...
3 3	Кепілді тегін медициналық көмектің қандай қызмет...	Кепілді тегін медициналық көмекке алғашқы меби...	What services are included in the guaranteed f...	The guaranteed free medical care package inclu...
4 4	Дәрігерге онлайн қабылдауға қалай жазылуға бол...	Дәрігерге онлайн қабылдауға eGov.kz портала...	How can I make an online appointment with a do...	You can make an online appointment with a doc...
5 3759	Әскери есепке алу жүйесінде деректерди қорғау...	Мемлекеттік органдар мен аудит үйімдерде заңнам...	How is compliance with legislative regulations...	Government bodies and audit organizations cons...
6 3760	Қазақстан Республикасында әскери есепке алу жү...	Кесіби стандарттар, тренингтер және этикалық к...	How is adherence to ethical standards by emplo...	This is ensured through professional standards...
7 3761	Әскери есепке алу жүйесінде деректерди қорғау...	Аудит нәтижелері, қолданушылар пікірі мен тех...	How is the long-term sustainability of innovat...	Evaluation is based on audit results, user fee...
8 3762	Қазақстан Республикасында әскери есепке алу жү...	Халықаралық сертификаттар, аудит нәтижелері ж...	How are the results of international cooperati...	Results are measured through international cer...
9 3763	Әскери есепке алу жүйесінде деректерди қорғауд...	Стратегия инновациялық технологиялар, халықара...	What is the general strategy and future develo...	The strategy is aimed at sustainable developme...

Figure A3. The synthetic dataset.

Table A2. The legislative branches of the Republic of Kazakhstan.

Act Form	Agency of Act Approval	Legal Relations Area	Date of Approval
Constitution (1)	133,000,000,000 (10)	Agriculture (60)	2024 (16)
Constitutional law (48)	134,000,000,000 (1)	Civil right (78)	2023 (10)
Code (31)	135,000,000,000 (17)	Commonwealth of Independent States (4)	2022 (8)
Law (353)	144,000,000,000 (1)	Communication (23)	2021 (177)
Order (52)	International bodies and organizations (82)	Conservation and use of fauna (16)	2020 (330)
Decree (631)	The Accounts Committee for Control over Execution of the Republican Budget (5)	Conservation and use of forests (8)	2019 (290)
Order (963)	The Agency of the Republic of Kazakhstan on Government Service Affairs (1)	Conservation and use of lands (18)	2018 (275)
AGR (35)	The Central Election Commission of the Republic of Kazakhstan, the CEC of the RK	Conservation and use of waters (8)	2017 (128)
AGT (1)		Constitutional system and foundations of public administration (671)	2016 (203)
Agreement (17)		Construction and architecture (22)	2015 (94)
Amendments (13)		Cooperation (4)	2014 (49)

Table A3 shows real-world examples of generated model responses on three different datasets. In some cases, the answers contain factual distortions, logical inconsistencies, or excessive redundancy. For example, Gemma often gives too long and confusing answers with repetitions and inserts that are not relevant to the question ("To receive the service online..."). The LLaMA model is more often limited to fragmentary or incomplete explanations, which leads to a loss of legal accuracy ("Where are housing disputes heard in Kazakhstan?"). Qwen demonstrates a penchant for factual errors and inventing details that are not in the law ("President of the Republic of Kazakhstan... and Astana will be considered in the blessed month.)").

It is important to note that the table retains the actual formats and structure of responses, including technical failures and repetitions. This allows us to objectively show the nature of typical errors that occur when applying models to legal issues in the Kazakh language.

Table A3. Identifying errors in answering questions on different models and datasets.

Dataset	Question	Predictions_Gemma	Predictions_LLaMA	Predictions_Qwen
Adilet	Ұйым халықаралық ұйым ретінде толыққанды заңды тұлға мәртебесін пайдаланатын болады ма? (Will the organization enjoy the status of a full-fledged legal entity as an international organization?)	Жоқ, ұйым заңды тұлға мәртебесін алмауга шешім қабылдады. (No, the organization decided not to obtain the status of a legal entity.)	Қазақстан Республикасының заңдары мен халықаралық халықаралық ұйымдарға қатысуга күкіктік жағдайларды белгілейтін заңдарында көрсетілген күкіктік жағдайлармен айқындалады. < end_of_text > (The legal conditions for participation in international organizations are determined by the laws of the Republic of Kazakhstan and the laws of the Republic of Kazakhstan establishing the legal conditions for participation in international organizations. < end_of_text >).	Ұйым халықаралық ұйым ретінде толыққанды заңды тұлға мәртебесін пайдаланатын болады. (The organization will enjoy the status of a full legal entity as an international organization.)
Adilet	Қызметті онлайн режимінде қалай алуға болады? (How can I get the service online?)	Кызметті онлайн режимінде алу үшін ұсынылатын қызметтердің тізімінен белгілеген қызметті таңдаңыз, алдын ала келіп түсken ақпараттардың тізімін толтырыңыз, әкімшілік тіркеудін бастапқы күжаттарын таңдаңыз, қызметті қалай алуға арналған ақпаратты таңдаңыз. Егер қызметті өзге адамның атынан алуға болатын болса, қабылдаушының аты, әкесінің аты, тегі, азаматтығы, азаматтық хал, күйеуінің аты, әкесінің аты, тегі, азаматтығы, (To receive the service online, select the service you want from the list of services offered, fill in the list of previously received information, select the primary documents for administrative registration, select information on how to receive the service. If the service can be received on behalf of another person, enter the recipient's name, patronymic, surname, citizenship, civil status, husband's name, patronymic, surname, citizenship,)	Қызметті онлайн режимінде алу үшін «Қызметті онлайн режимінде алу» батырмасын таңдаңыз. Сіз онлайн режимінде алуға құқылы екенізді растаңыз.< end_of_text > (To receive the service online, select the “Receive the service online” button. Confirm that you are eligible to receive it online.< end_of_text >)	<p>Кызметті онлайн режимінде қалай алуға болады?</p> <ol style="list-style-type: none"> Басқарушы туралы қызмет алу үшін қызметтін алуға жүзеге асырылатын қызметтер туралы қызметтің отінішін күкірьшін таңдаңыз. Қызметтің отінішін күрү қызыметтің отінішін күрү (қызыметтің отінішін күрү) күкірьшін таңдаңыз. Басқарушы туралы қызмет алу үшін қызметтің отінішін күрүта қызыметтің отінішінде қызметтің отінішін күрү (қызыметтің отінішін күрү) күкірьшін таңдаңыз. (How to get the service online?) <ol style="list-style-type: none"> To get the service about the manager, select the service request right for the services to be carried out to receive the service. To create a service request, select the service request creation right for the participant (create a service request). To get the service about the manager, select the service request creation right for the participant (create a service request).)
Zqai and Gov	Қазақстан Республикасының Президенті Конституциялық Сот Төрағасының орынбасарын Конституциялық Сот Төрағасының ұсынуымен кімдердің арасынан тағайындауды? (From whom does the President of the Republic of Kazakhstan appoint the Deputy Chairman of the Constitutional Court upon the recommendation of the Chairman of the Constitutional Court?)	Қазақстан Республикасының Президенті Конституциялық Сот Төрағасының орынбасарын Конституциялық Сот Төрағасының ұсынуымен кімдердің арасынан тағайындауды? (From whom does the President of the Republic of Kazakhstan appoint the Deputy Chairman of the Constitutional Court upon the recommendation of the Chairman of the Constitutional Court?)	Қазақстан Республикасының Президенті(President of the Republic of Kazakhstan)	<p>Қазақстан Республикасының Президенті Конституциялық Сот Төрағасының орынбасарын Конституциялық Сот Төрағасының ұсынуымен Қазақстан Республикасының Президенті және Парламенттің халықаралық мүдделерінің орынбасарымен тағайындауды. (The President of the Republic of Kazakhstan appoints the Deputy Chairman of the Constitutional Court upon the recommendation of the Chairman of the Constitutional Court and the Deputy for International Interests of the President of the Republic of Kazakhstan and the Parliament.)</p>

Table A3. Cont.

Dataset	Question	Predictions_Gemma	Predictions_LLaMA	Predictions_Qwen
Zqai and Gov	XКБ жобасына қалай өтініш беруге болады? (How to apply for an ICB project?)	XКБ жобасын өзгерту және жаңа жобаларды дайындауды талап ететін келісімдердің ұсыныстарымен катар, өтініштердің қабылдауы оқімдік органдары мен дауазымда адамдарының өтініштерімен жүргізіледі. Казакстан Республикасының 2017 жылғы 30 шілдедегі № 336-V Қауіпсіздік және азаматтық корғаныс туралы Заңының 50-бабының 1-тармағына сойкес, қауіпсіздік пен азаматтық корғаныс органдары, жүргізушилер, есептілікті қамтамасызы өтетін тұлғалар мен жеке тұлғалар. өтініштің сапасын және болуы туралы жауапкершілігі бар және оның орналасуы туралы өтініштерді қабылдайды. (How to apply for a draft of the ISB? Along with proposals for agreements requiring changes to the ISB draft and the preparation of new drafts, applications are accepted by akimat bodies and officials. In accordance with paragraph 1 of Article 50 of the Law of the Republic of Kazakhstan No. 336-V dated 30 July 2017 "On Security and Civil Defense", security and civil defense bodies, drivers, persons providing reporting and individuals, are responsible for the quality and availability of the application and accept applications for its location.)	«Қазақстан Республикасының Құқықтық академиясы» РМК БАСҚА ҚР Үкіметінің 2020 жылғы 15 қаңтардағы № 44 қаулысының 1-бабына сойкес, Қазақстан Республикасының Әділет министрлігінің 2017 жылғы 5 қазандагы № 125 қаулысының 1-бабының 11) тармағына сойкес, тұрақты тұрғылықты орны Қазақстан Республикасының аумағында және Қазақстан Республикасының аумағында тұрақты тұрғылықты орны болмаса, Қазақстан Республикасының аумағында немесе шетелде тұрақты тұрғылықты орны болғанын үшін қызметті жүзеге асыратын жеке қасіпкерлеріне Қазақстан Республикасының заңнамасына сойкес келетін, Қазақстан Республикасының заңнамасында айқындалған және Қазақстан Республикасының заңнамасына сойкес қамтамасызы өтілтін заңнамалық талаптарға сойкес күрьылған қасіпкерлік субъектілеріне Қазақстан Республикасының заңнамасы (RSE "Legal Academy of the Republic of Kazakhstan" OTHER In accordance with Article 1 of the Resolution of the Government of the Republic of Kazakhstan No. 44 dated 15 January 2020, in accordance with paragraph 11) of Article 1 of the Resolution of the Ministry of Justice of the Republic of Kazakhstan No. 125 dated 5 October 2017, individual entrepreneurs carrying out activities in the territory of the Republic of Kazakhstan with a permanent place of residence and in the absence of a permanent place of residence in the territory of the Republic of Kazakhstan, with a permanent place of residence in the territory of the Republic of Kazakhstan or abroad, business entities established in accordance with the legislative requirements established by the legislation of the Republic of Kazakhstan and provided for in accordance with the legislation of the Republic of Kazakhstan)	XКБ-да білім беру, жұмысқа кімігін алу және қызметкерлік көрсету арқылы жұмысқа өтініш беруге болады. (You can apply for a job at the IKB by providing education, obtaining a work permit, and providing employment.)

Table A3. Cont.

Dataset	Question	Predictions_Gemma	Predictions_LLaMA	Predictions_Qwen
Synthetic	Зейнетакы жинактарын аудару туралы өтінішпен қайда жүгінуге болады(Where to apply for a pension savings transfer)	<p>Зейнетакы жинактарын аудару үшін Қазақстан Республикасының «Зейнетакы жүйесі туралы» Занының 8-бабына сәйкес жинактардың негізгі операторына (Жанактық және зейнетакы жинактары корпорациясына) етінішті беру қажет. 2020 жылдан 1 кантарынан бастап бұл етініш ҚР Заннама және құқықтық ақпарат институтының электрондық порталы арқылы (www.egov.kz, accessed on 25 June 2025) берілу мүмкін. Қазақстан Республикасының Заннама және құқықтық ақпарат институты электрондық порталында тіркелген ақпараттық жүйесінің қауіпсіздігін және ақпараттық құқық бузушылықтың алдын алу туралы қауіпсіздік көзделдердің көзделген жағдайларда, әдеттегі толемдер санынан кем төленеді. Бұл жағдайда жүйелік көзделдерді төлеу қажет емес. Жеке тұлғалар үшін жинактардың қалған бөлігінде Қазақстан Республикасының заннамасына сәйкес төленетін жеке тұлғаларға өндірістік көзделдердің 25%-ынан аз колемде төленеді (In the case of system income, after the individual is identified, 25% of income from savings for individuals is paid in accordance with the legislation of the Republic of Kazakhstan. 75% of system income is paid in cases provided for by law, less than the number of regular payments. In this case, system income is not required to be paid. The remaining part of the savings for individuals is paid in accordance with the legislation of the Republic of Kazakhstan in an amount less than 25% of production income for individuals)</p>	<p>Жүйелік кірістер кезінде жеке тұлға анықталғаннан кейін Қазақстан Республикасының заннамасына сәйкес жеке тұлғалар үшін жинактардан кірістердің 25%-ы төленеді. Жүйелік кірістердің 75%-ы заннамада көзделген жағдайларда, әдеттегі толемдер санынан кем төленеді. Бұл жағдайда жүйелік кірістерді төлеу қажет емес. Жеке тұлғалар үшін жинактардың қалған бөлігінде Қазақстан Республикасының заннамасына сәйкес төленетін жеке тұлғаларға өндірістік көзделердің 25%-ынан аз колемде төленеді (In the case of system income, after the individual is identified, 25% of income from savings for individuals is paid in accordance with the legislation of the Republic of Kazakhstan. 75% of system income is paid in cases provided for by law, less than the number of regular payments. In this case, system income is not required to be paid. The remaining part of the savings for individuals is paid in accordance with the legislation of the Republic of Kazakhstan in an amount less than 25% of production income for individuals)</p>	<p>Зейнетакы жинактарын аудару туралы өтінішпен жүгінуге болады даму саласы, қауіпсіздік көнеші туралы өтінішпен. (You can apply for a transfer of pension savings by applying to the development sector or the security council.)</p>
Synthetic	Қазақстанда тұрғын үйге қатысты сот дауларын қайда қарастырады? (Where are housing disputes heard in Kazakhstan?)	Қазақстанда тұрғын үйге қатысты сот дауларын жергілікті жергілікті соттар қарастырады. (In Kazakhstan, housing disputes are considered by local courts.)	Соттың төмөнгі және жоғары инстанциялары қарастырады. (Considered by lower and higher courts.)	Астана қайырлы айында қарастырылады. (Astana will be considered in the blessed month.)

Table A4. Examples of evaluated questions and answers.

Question (Kaz)	Question (Eng)	Answer (Kaz)	Answer (Eng)
Климаттың өзгеруіне бейімделу процесі қандай сатылардан тұрады және оны жүзеге асыру үшін қай органдар жауапты?	What are the stages of the climate change adaptation process, and which bodies are responsible for its implementation?	<p>Климаттың өзгеруіне бейімделу процесі мына сатылардан тұрады:</p> <ol style="list-style-type: none"> Бағалау: Климаттың өзгеруінің қаупі мен асерін анықтау (мысалы, су тасқынның жиілігін артуы). Жоспарлау: Бейімделу шараларын жоспарлау және дамыту (мысалы, бөгөттер салу немесе егістік аумағын өзгерту). Іске асыру: Жоспарланған шараларды жүзеге асыру (мысалы, бөгет күрылымын бастау). Бақылау және бағалау: Бейімделу шараларының тиімділігін бақылау және оларды қажет болған жағдайда түзету (мысалы, бөгеттің суды үстаяу қабылетін тексеру). <p>Бұл процесті жүзеге асыру үшін жауапты органдар:</p> <p>Үкімет: Үлттық деңгейде саясатты өзірлеу және үйлестіру.</p> <p>Жергілікті билік: Аумактық деңгейде бейімделу жоспарларын іске асыру.</p> <p>Мемлекеттік органдар: (Мысалы, Су ресурстары министрлігі) — салалық бағдарламаларды жүзеге асыру.</p> <p>Ғылыми зерттеушілер: (Мысалы, өзгерістерді зерттеу және бейімделу шараларын өзірлеу).</p> <p>Қоғамдық өргенеуде жүзеге асыру үшін жауапты органдар:</p> <p>Құқықтық тұзғалық органдар: (Мысалы, қоғамдық өргенеуде жүзеге асыру үшін жауапты органдар).</p>	<p>The process of adapting to climate change consists of the following stages:</p> <ol style="list-style-type: none"> Evaluation: Identifying the risks and impacts of climate change (e.g., increasing the frequency of floods). Planning: Planning and developing adaptation measures (e.g., building dams or changing the area of crops). Implementation: Implementing the planned measures (e.g., starting dam construction). Monitoring and Evaluation: Monitoring the effectiveness of adaptation measures and adjusting them if necessary (e.g., checking the water retention capacity of a dam). The bodies responsible for implementing this process are: <ul style="list-style-type: none"> - Government: Developing and coordinating policies at the national level. - Local authorities: Implementing adaptation plans at the territorial level. - State bodies: (e.g., the Ministry of Water Resources) — implementing sectoral programs. - Scientific institutions: Researching climate change and developing adaptation measures. - Public organizations: Ensuring public participation and education.
Климаттың өзгеруіне бейімделу процесі қандай сатылардан тұрады және оны жүзеге асыру үшін қай органдар жауапты?	What are the stages of the climate change adaptation process, and which bodies are responsible for its implementation?	<p>Ст. 314 КР Экологиялық кодексі 2021 жылғы 2 қантардағы № 400-VI КРЗ</p> <ol style="list-style-type: none"> Климаттың өзгеруіне бейімделу процесі мынадай сатыларды қамтиды: <ol style="list-style-type: none"> (1) акпарат жинау және климаттың өзгеруіне осалықты бағалау; (2) климаттың өзгеруіне бейімделу жөніндегі шараларды өзірлеу; (3) климаттың өзгеруіне бейімделу жөніндегі шараларды жүзеге асыру; (4) климаттың өзгеруіне бейімделу жөніндегі шаралардың тиімділігін мониторингтеу және бағалау (5) климаттың өзгеруінің асер етуі және климаттың өзгеруіне бейімделу жөніндегі шаралардың тиімділігін туралы есептілік; (6) мониторингтеу мен бағалаумен нәтижелері негізінде климаттың өзгеруіне бейімделу жөніндегі шараларды түзет Климаттың өзгеруіне бейімделу процесін климаттың өзгеруіне бейімделу үшін басым ретінде айқындалған мемлекеттік басқару салалар бойынша үәкілетті орталық атқарушы органдар және облыстыардың, республикалық манызы бар қалалардың, астананың жергілікті атқарушы органдарды жүзеге асырады. Осы Кодекстің 313-бабының 2-тармағында көрсетілген, климаттың өзгеруіне бейімделу үшін басым мемлекеттік басқару салалары бойынша тиісті мемлекеттік бағдарламаларды өзірлеу және іске асыру шеберінде үәкілетті орталық атқарушы органдар және облыстыардың, республикалық манызы бар қалалардың, астананың жергілікті атқарушы органдарды осы баптың 1-тармағында көрсетілген климаттың өзгеруіне бейімделу процесінің сатыларын жүзеге асырады. Осы баптың 1-тармағында көрсетілген сатыларды жүзеге асыру жөніндегі талаптар коршаган органдар корғау саласындағы үәкілетті органдар көйткен климаттың өзгеруіне бейімделу процесін үйімдастыру және іске асыру қагидаларына сойкес іске асырылады. Қоршаган органдар корғау саласындағы үәкілетті органдар климаттың өзгеруін мәселелері жөніндегі халықаралық шарттарға тартағаттарға сойкес кің өзгеруіне бейімделу нәтижелері бойынша есептілкіт жүзеге ас Толығырақ оқыныш: https://adilet.zan.kz/kaz/docs/K2100000400, accessed on 25 June 2025. 	<p>Legal act: Article 314 Environmental Code of the Republic of Kazakhstan No. 400-VI, 2 January 2021</p> <ol style="list-style-type: none"> The process of adaptation to climate change includes the following stages: <ol style="list-style-type: none"> (1) information collection and evaluation of vulnerability to climate change; (2) planning for adaptation to climate change; (3) development of measures to adapt to climate change; (4) implementation of measures to adapt to climate change; (5) monitoring and evaluation of the effectiveness of measures to adapt to climate change; (6) reporting on the impact of climate change and the effectiveness of measures to adapt to climate change; (7) adjustment of measures to adapt to climate change based on the results of monitoring and evaluation. The process of adaptation to climate change is carried out by authorized central executive bodies in the areas of state administration identified as priorities for adaptation to climate change, as well as local executive bodies of regions, cities of republican significance, and the capital. Within the framework of the development and implementation of relevant state programs in priority areas of state management for adaptation to climate change, specified in paragraph 2 of Article 313 of this Code, authorized central executive bodies and local executive bodies of regions, cities of republican significance, the capital shall implement the stages of the climate change adaptation process specified in paragraph 1 of this Article. The requirements for the implementation of the stages specified in paragraph 1 of this Article shall be implemented in accordance with the rules for organizing and implementing the climate change adaptation process approved by the authorized body in the field of environmental protection. The authorized body in the field of environmental protection shall carry out reporting on the results of climate change adaptation in accordance with international treaties on climate change issues. <p>Read more: https://adilet.zan.kz/kaz/docs/K2100000400, accessed on 25 June 2025.</p>

Table A4. Cont.

Question (Kaz)	Question (Eng)	Answer (Kaz)	Answer (Eng)
Салық салу обьектісі ретінде тіркелген кәсіпкер патент негізінде қандай есептіліктер тапсыруы тиіс және бұл Салық кодексінің кай бабында көрсетілген?	What types of reporting must an entrepreneur registered as a tax entity submit under the patent-based tax regime, and which article of the Tax Code regulates this?	<p>Патент негізінде жұмыс істейтін дара кәсіпкер салықтық есептілікті женилдетілген тартиппен тапсырады. Бұл режим—арналық салық режимдерінің бірі болып табылады және негізінен шағын бизнес субъектілері үшін арналған.</p> <p>1. Есептіліктің негізі түрі:</p> <p>Салық есептілігі—Патент негізінде жұмыс істейтін кәсіпкер тек патент алуға отініш пен толемді растайтын күжаттығана тапсырады. Бұл салық режимінде жылдық немесе айлық табыс болжамы көрсетіледі және сонын негізінде біржолы толем жасалады.</p> <p>Қосымша декларациялар немесе тоқсан сайынғы/жылдық есептіліктер тапсырылмайды.</p> <p>2. Үакыт мерзімі мен тәртібі:</p> <ul style="list-style-type: none"> – Патент алуға отініш патенттің қолданылу мерзімі басталғанға дейін кемінде 5 жұмыс күні бүрін ұсынылуы керек. – Патент сомасы табыс салығы мен әлеуметтік толемдерді қамтиды және сол отінішпен бірге төленеді. – Егер кәсіпкер қызметтің тоқтататын болса немесе басқа салық режиміне отсе, бұл туралы салық органдына хабарлануы тиіс. <p>3. Құқықтық негіз:</p> <p>Бұл талаптар Қазақстан Республикасының Салық кодексінің 77-бабында нақты көрсетілген.</p> <p>Ол жерде “арналық салық режимін патент негізінде қолданатын дара кәсіпкер салық органына есептіліктің басқа түрлерін</p> <p>Патент негізінде салық толейтін кәсіпкер тек патент алуға отініш береді және тиісті толемді жүргізеді. Қосымша есептіліктерді тапсыру міндетті емес. Бұл КР Салық кодексінің 77-бабында нақты жазылған.</p> <p>Құқықтық дереккөз: https://adilet.zan.kz/kaz/docs/K1700000120, accessed on 25 June 2025.</p>	<p>An individual entrepreneur operating under the patent-based tax regime submits tax reports in a simplified manner. This regime is one of the special tax regimes and is primarily intended for small business entities.</p> <p>1. Main type of reporting: Under this regime, the entrepreneur submits only an application for obtaining a patent and a document confirming payment. The application includes an estimate of the annual or monthly income, and a one-time payment is made based on this estimate.</p> <p>No additional tax declarations or quarterly/annual reports are required.</p> <p>2. Deadlines and procedure: The application for the patent must be submitted at least 5 working days before the start of the patent's validity period.</p> <p>The patent amount includes individual income tax and social payments, and must be paid together with the application.</p> <p>If the entrepreneur ceases activities or switches to another tax regime, they must inform the tax authority accordingly.</p> <p>3. Legal basis: These requirements are clearly specified in Article 77 of the Tax Code of the Republic of Kazakhstan.</p> <p>It states that an individual entrepreneur using the patent-based special tax regime is not required to submit any other types of tax reports to the tax authority. In summary, a patent-based taxpayer only submits an application for the patent and makes the appropriate payment. Additional reporting is not required. This is directly stated in Article 77 of the Tax Code.</p> <p>Legal source: https://adilet.zan.kz/kaz/docs/K1700000120, accessed on 25 June 2025.</p>

df.iloc[[28, 37]]									
Nº	input	instruction	Predictions_Mstral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_KazLLM	Predictions_Phi	
28	Category: Тікелей сұраптар	Зейнетакы жинақтарын аудару туралы етінішпен к...	< im_start >system\nCategory: Тікелей сұраптар...	Зейнетакы жинақтарын аудару туралы етінішпен ж...	Жүйелік кірістер кезінде жеке тұлға анықталған...	Зейнетакы жинақтарын аудару үшін Қазақстан Рес...	Құрметті Мәденин, біз сізге Қазақстан Республи...	Зейнетакы жинақтарын аудару үшін Елік...	
37	Category: Тікелей сұраптар	ХКБ жобасына қалай етініш беруге болады?	< im_start >system\nCategory: Тікелей сұраптар...	ХКБ-да білім беру, жұмысқа кімігін алу және қы...	«Қазақстан Республикасының Құқықтық академиясы...	ХКБ жобасына қалай етініш беруге болады? ХКБ ж...	Қазақстан Республикасындағы Құқықтық акпаратты...	Әділет министрлігінің қаулысына сайкес, қалай ...	
df[df['instruction'] == 'Зейнетакы жинақтарын аудару туралы етінішпен қайда жүгінгүе болады']									
Nº	input	instruction	Predictions_Mstral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_KazLLM	Predictions_Phi	
28	Category: Тікелей сұраптар	Зейнетакы жинақтарын аудару туралы етінішпен к...	< im_start >system\nCategory: Тікелей сұраптар...	Зейнетакы жинақтарын аудару туралы етінішпен ж...	Жүйелік кірістер кезінде жеке тұлға анықталған...	Зейнетакы жинақтарын аудару үшін Қазақстан Рес...	Құрметті Мәденин, біз сізге Қазақстан Республи...	Зейнетакы жинақтарын аудару үшін Елік...	
df[df['instruction'] == 'ХКБ жобасына қалай етініш беруге болады?']									
Nº	input	instruction	Predictions_Mstral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_KazLLM	Predictions_Phi	
37	Category: Тікелей сұраптар	ХКБ жобасына қалай етініш беруге болады?	< im_start >system\nCategory: Тікелей сұраптар...	ХКБ-да білім беру, жұмысқа кімігін алу және қы...	«Қазақстан Республикасының Құқықтық академиясы...	ХКБ жобасына қалай етініш беруге болады? ХКБ ж...	Қазақстан Республикасындағы Құқықтық акпаратты...	Әділет министрлігінің қаулысына сайкес, қалай ...	

Figure A4. Example errors of the models using the Zqai and Gov datasets.

df.iloc[[1, 10]]									
Nº	input	instruction	Predictions_Mstral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_GPT	Predictions_KazLLM	Predictions_Phi
1	Category: Тікелей сұраптар	Қазақстанда тұрғын үйге катысты сот дауларын к...	< im_start >system\nCategory: Тікелей сұраптар...	Астана қайырлы айында қарастырылады.	Соттың төменгі және жоғары инстанциялары қарас...	Қазақстанда тұрғын үйге катысты сот дауларын ж...	Жергілікті соттарда	Қазақстанда тұрғын үйге катысты сот дауларын т...	Қазақстанда тұрғын үйге катысты сот дауларын т...
10	Category: Тікелей сұраптар	Тұрғын үйді жалға алу шартын бзу қандай салда...	< im_start >system\nCategory: Тікелей сұраптар...	Тұрғын үйді жалға алу шартын бзу қандай салда...	Жалға төлемдердің шұғын төлемінен қалдыру, шар...	Тұрғын үйді жалға алу шарттарының бұзылуы жалд...	Егер шарта көрсетілген талаптар бұзылса, айып...	Тұрғын үйді жалға алу шартын бзу көлісімшарт...	Тұрғын үйді жалға алу шартын бзу құқықтық сал...
df[df['instruction'] == 'Қазақстанда тұрғын үйге катысты сот дауларын қайда қарастырады?']									
Nº	input	instruction	Predictions_Mstral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_GPT	Predictions_KazLLM	Predictions_Phi
1	Category: Тікелей сұраптар	Қазақстанда тұрғын үйге катысты сот дауларын к...	< im_start >system\nCategory: Тікелей сұраптар...	Астана қайырлы айында қарастырылады.	Соттың төменгі және жоғары инстанциялары қарас...	Қазақстанда тұрғын үйге катысты сот дауларын ж...	Жергілікті соттарда	Қазақстанда тұрғын үйге катысты сот дауларын т...	Қазақстанда тұрғын үйге катысты сот дауларын т...
df[df['instruction'] == 'Тұрғын үйді жалға алу шартын бзу қандай салдарға әкеледі?']									
Nº	input	instruction	Predictions_Mstral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_GPT	Predictions_KazLLM	Predictions_Phi
10	Category: Тікелей сұраптар	Тұрғын үйді жалға алу шартын бзу қандай салда...	< im_start >system\nCategory: Тікелей сұраптар...	Тұрғын үйді жалға алу шартын бзу қандай салда...	Жалға төлемдердің шұғын төлемінен қалдыру, шар...	Тұрғын үйді жалға алу шарттарының бұзылуы жалд...	Егер шарта көрсетілген талаптар бұзылса, айып...	Тұрғын үйді жалға алу шартын бзу көлісімшарт...	Тұрғын үйді жалға алу шартын бзу құқықтық сал...

Figure A5. Example of errors of models using the synthetic dataset.

df.iloc[[4, 9]]									
Nº	input	instruction	Predictions_Gemma	Predictions_LLaMA	Predictions_Qwen	Predictions_Msitrail	Predictions_GPT	Predictions_KazLLM	Predictions_Phi
4	5	Category: Тікелей сұраптар Кызметті онлайн режимінде қалай алуға болады?	Кызметті онлайн режимінде қалай алу үшін үсынылатын ...	Кызметті онлайн режимінде алу үшін «Кызметті о...»	Кызметті онлайн режимінде қалай алуға болады?...	< im_start >system\nCategory:	Косымшада азаматтық алу туралы қызмет көрсетед...	Кызметті онлайн режимінде алу үшін сізге мына ...	Кызметті онлайн режимінде алу үшін, сіз осы ве...
9	10	Category: Тікелей сұраптар Казакстан Республикасының Президенті Конституц...	Казакстан Республикасының Президенті Конституц...	Казакстан Республикасының Президенті< end_of_t...	Казакстан Республикасының Президенті Конституц...	< im_start >system\nCategory:	Казакстан Республикасының Президенті Конституц...	Казакстан Республикасының Президенті Конституц...	Жарлық қабылдау белгімдегі арнайы қызметшілер...
df[df['instruction'] == 'Кызметті онлайн режимінде қалай алуға болады?']									
Nº	input	instruction	Predictions_Gemma	Predictions_LLaMA	Predictions_Qwen	Predictions_Msitrail	Predictions_GPT	Predictions_KazLLM	Predictions_Phi
4	5	Category: Тікелей сұраптар Кызметті онлайн режимінде қалай алу үшін үсынылатын ...	Кызметті онлайн режимінде алу үшін «Кызметті о...»	Кызметті онлайн режимінде қалай алуға болады?...	Кызметті онлайн режимінде қалай алуға болады?...	< im_start >system\nCategory:	Косымшада азаматтық алу туралы қызмет көрсетед...	Кызметті онлайн режимінде алу үшін сізге мына ...	Кызметті онлайн режимінде алу үшін, сіз осы ве...
"Казакстан Республикасының Президенті Конституциялық Сот Тәрағасының орынбасарын Конституциялық Сот Тәрағасының үсынуымен кімдердің арасынан тәғайындауда?"									
Nº	input	instruction	Predictions_Gemma	Predictions_LLaMA	Predictions_Qwen	Predictions_Msitrail	Predictions_GPT	Predictions_KazLLM	Predictions_Phi
9	10	Category: Тікелей сұраптар Казакстан Республикасының Президенті Конституц...	Казакстан Республикасының Президенті Конституц...	Казакстан Республикасының Президенті< end_of_t...	Казакстан Республикасының Президенті Конституц...	< im_start >system\nCategory:	Казакстан Республикасының Президенті Конституц...	Казакстан Республикасының Президенті Конституц...	Жарлық қабылдау белгімдегі арнайы қызметшілер...

Figure A6. Example errors of the models using the Adilet dataset.

Appendix B

Table A5. Example of the highest degree of consensus among experts.

Question	Answer of Model GPT Mini with Dataset Adilet	Criterion	Experts		
			№ 1	№ 2	№ 3
Рұқсат беру құжаттарын алу үшін лауазымды тұлға паралар беруді мензесе не істеге керек? (What should you do if an official suggests paying a bribe to obtain authorization documents?)	Егер рұқсат беру құжаттарын алу үшін лауазымды тұлға паралар беруді мензесе, сіз бұл әрекетті құқық корғау органдарына хабарлауыңыз керек. Соран байланысты сіздің азаматтық борыштыңыз—құқық корғау органдарының қызметкерлеріне лауазымды адамыңың паралар туралы сөздері мен іс-әрекеттері жөнінде дәлел мен айғартарынызды (ескертпелер мен аудио мен бейнетаста жақаларын коса алғанда) баяндай отырып, арыз жазу. Өйткени паралар бар—сүйбайлар жемқорлық қылмыстарының бірі. Ол үшін сізге құқық корғау органдары гимаратына жүргін немесе жағдай түрлерінде тұтқындалуын қамтыса) құру үшін жақын полиция бөлімшесіне хабарласу қажет. Инстанцияға жағаша етінші берген кезде хабарландырудың мисалдары, арызының тіркеу белгісін беру, арызының тіркеу белгісін бар арызындың көшірмесін алу (сүрай) үшін қабылдау болмасынан көзекін поштаниң лауазымды тұлғага арызындың үсынтынын немесе арыз берушінің өздері қабылдайтынын түсіндіру қажет. Арыз жаузадағы басты талаптардың бірі—оз деректерінізді, оқиғаның күнін, уақытын, орнын және паралар берумен байланысты болған лауазымды адамыңың әрекеттерін егжей-тегжейлі баяндау. Арыз беру кезінде паралар беру немесе паралар аудиен болған оқиғалардың барлығын егжей-тегжейлі сипаттаған жауз манызды. (If an official suggests giving a bribe to obtain permits, you must report this action to law enforcement agencies. In this regard, your civic duty is to file a complaint with law enforcement officers, providing them with evidence and testimony (including notes and audio and video recordings) about the words and actions of the official about taking a bribe. After all, bribery is one of the corruption crimes. To do this, you need to contact the building of a law enforcement agency or the police station nearest to your home to create an operational-investigative group (if this involves the arrest of an official at the time of illegal bribery, words about taking a bribe or other actions). When submitting a written application to an authority, examples of notification, to issue a registration mark to your application, to receive (request) a copy of your application with a registration mark to your application, it is necessary to explain to the duty officer in the reception room that the post office will submit your application to the official or will accept the applicant themselves. One of the main requirements for filing a complaint is to provide a detailed description of your personal details, the date, time, and location of the incident, and the actions of the official involved in the bribery.)	Accuracy	4	3	4
		Completeness	4	3	3
		Clarity	4	4	5
		Relevance	5	4	4
		Style	5	4	5

Table A6. Example of the lowest degree of consensus among experts.

Question	Answer of Model Llama 3.2 with Synthetic Dataset	Criterion	Experts		
			№ 1	№ 2	№ 3
Кедендік тексерудің түрлөрі, мерзімдері және тексеру нысандары қандай? (What are the types, terms, and forms of customs inspections?)	Кедендік тексерулер алдын ала, жоспарлы және кездейсок түрде жүзеге асады. Мерзімдері: жыл сайын, 3 жыл және 5 жыл. Мерзімдері және тексеру нысандары алеуметтік-экономикалық факторларға байланысты. (Customs inspections are carried out in advance, planned and random manner. Periods: annually, every 3 years and every 5 years. The periods and forms of inspections depend on socio-economic factors.)	Accuracy	3	1	2
		Completeness	2	1	2
		Clarity	4	2	2
		Relevance	3	2	3
		Style	3	1	2

References

1. Abdallah, A.; Piryani, B.; Jatowt, A. Exploring the State of the Art in Legal QA Systems. *J. Big Data* **2023**, *10*, 127. [CrossRef]
2. Ministry of Digital Development of Kazakhstan. *Strategic Plan for AI and Language Processing in Kazakhstan 2024–2030*; Ministry of Digital Development of Kazakhstan: Astana, Kazakhstan, 2024.
3. Issayev, B.; Makhambetov, K. Morphological Challenges in Kazakh NLP. *J. Comput. Linguist.* **2022**, *38*, 45–63.
4. Baimakhan, R.; Yessimbekova, A. Syntax Variability in Kazakh Language Processing. *Cent. Asian Linguist. Stud.* **2023**, *12*, 101–118.
5. Kazakh NLP Initiative. *Advances in Kazakh Language Processing*; Almaty Research Institute: Almaty, Kazakhstan, 2023.
6. Smagulova, D. Machine Translation Accuracy for Legal Kazakh Texts. *Kazakhstan J. AI Res.* **2023**, *9*, 77–95.
7. Tolegen, B.; Yermagambet, A.; Kassenov, M. Developing BERT-Based NLP Models for the Kazakh Language. *IEEE Trans. AI NLP* **2023**, *15*, 221–237.
8. Jang, D. Enhancing Search-Augmented Generation (RAG) Performance Using Korean Reranker. Available online: <https://aws.amazon.com/ko/blogs/tech/korean-reranker-rag/> (accessed on 1 July 2025).
9. Kim, J. A Study on Data Chunking Strategies to Enhance LLM Service Quality Using RAG Techniques. Master’s Thesis, Korea University, Seoul, Republic of Korea, 2024.
10. Angels, B.; Vinamra, B.; Renato, L.; Estevão, R.; Hendry, T.; Holstein, D.; Marsman, J.; Mecklenburg, N.; Malvar, S.; Nunes, L.O.; et al. RAG vs. Fine-Tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture. *arXiv* **2024**, arXiv:2401.08406. [CrossRef]
11. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Kiela, D. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
12. Ruder, S.; Sil, A. Multi-Domain Multilingual Question Answering. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts, Punta Cana, Dominican Republic & Online, 7–11 November 2021; Available online: <https://api.semanticscholar.org/CorpusID:245289877> (accessed on 1 July 2025).
13. Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y.J.; Madotto, A.; Fung, P. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* **2023**, *55*, 1–38. [CrossRef]
14. Mialon, G.; Dessì, R.; Lomeli, M.; Nalmpantis, C.; Pasunuru, R.; Raileanu, R.; Rozière, B.; Schick, T.; Dwivedi-Yu, J.; Celikyilmaz, A.; et al. Augmented Language Models: A Survey. *arXiv* **2023**, arXiv:2302.07842. [CrossRef]
15. Lazaridou, A.; Gribovskaya, E.; Stokowiec, W.; Grigorev, N. Internet-Augmented Language Models through Few-Shot Prompting for Open-Domain Question Answering. *arXiv* **2022**, arXiv:2203.05115.
16. Kamalloo, E.; Dziri, N.; Clarke, C.L.A.; Rafiei, D. Evaluating Open-Domain Question Answering in the Era of Large Language Models. *arXiv* **2023**, arXiv:2305.06984. [CrossRef]
17. Rogers, A.; Gardner, M.; Augenstein, I. QA Dataset Explosion: A Taxonomy of NLP Resources for Question Answering and Reading Comprehension. *ACM Comput. Surv.* **2023**, *55*, 1–45. [CrossRef]
18. Roy, R.S.; Anand, A. *Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases and Text Collections*; Synthesis Lectures on Information Concepts Retrieval and Services 2021; Morgan & Claypool Publishers: San Rafael, CA, USA, 2022.
19. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Vancouver, BC, Canada, 30 July–4 August 2017.
20. Izacard, G.; Grave, E. Leveraging Passage Retrieval with Generative Models for Open-Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2021), Kiev, Ukraine, 21–23 April 2021.
21. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.S.H.; Wu, L.; Edunov, S.; Chen, D.; Yih, W. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), Online, 16–20 November 2020.
22. Minaee, S.; Mikolov, T.; Nikzad, N.; Chenaghlu, M.; Socher, R.; Amatriain, X.; Gao, J. Large Language Models: A Survey. *arXiv* **2024**, arXiv:2402.06196v3.

23. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A Survey of Large Language Models. *arXiv* **2023**, arXiv:2303.18223v16.
24. Jeong, C.S. A Case Study in Applying Hyperautomation Platform for E2E Business Process Automation. *Inf. Syst. Rev.* **2023**, *25*, 31–56. [CrossRef]
25. Jeong, C.S. A Study on the Service Integration of Traditional Chatbot and ChatGPT. *J. Inf. Technol. Appl. Manag.* **2023**, *3*, 11–28. [CrossRef]
26. Skelter Labs. 2024 Year of the RAG: Reasons for RAG’s Attention and Future Trends. Available online: <https://www.skelterlabs.com/blog/2024-year-of-the-rag> (accessed on 1 July 2025).
27. Microsoft. Retrieval Augmented Generation Using Azure Machine Learning Prompt Flow. Available online: <https://learn.microsoft.com/en-us/azure/machine-learning/concept-retrieval-augmented-generation?view=azureml-api-2> (accessed on 1 July 2025).
28. Li, Z.; Zhang, N.; Yao, Y.; Wang, M.; Chen, X.; Chen, H. Unveiling the Pitfalls of Knowledge Editing for Large Language Models. *arXiv* **2023**, arXiv:2310.02129. [CrossRef]
29. Miladi, F.; Psyché, V.; Lemire, D. Evaluating Generative Pre-Trained Transformers in MOOC Assessments: A Comparative Study of GPT Models. In Proceedings of the International Conference on Artificial Intelligence in Education, Xiamen, China, 22–24 November 2024.
30. Trofimov, E. Application of Computer Techniques and Systems in the Study of Law, Intellectual Analysis and Modeling of Legal Activity: A Systematic Review. 2020. Available online: https://www.researchgate.net/publication/343631670_Application_of_Computer_Techniques_and_Systems_in_the_Study_of_Law_Intellectual_Analysis_and_Modeling_of_Legal_Activity_A_Systematic_Review (accessed on 1 July 2025).
31. Hu, Z.; Li, X.; Tu, C.; Liu, Z.; Sun, M. Few-Shot Charge Prediction with Discriminative Legal Attributes. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), Santa Fe, NM, USA, 20–26 August 2018; pp. 487–498.
32. Liu, Z.; Tu, C.; Liu, Z.; Sun, M. Legal Cause Prediction with Inner Descriptions and Outer Hierarchies. In *Chinese Computational Linguistics: Proceedings of the 18th China National Conference, CCL 2019, Kunming, China, 18–20 October 2019*; Springer: Kunming, China, 2019; pp. 573–586.
33. Wang, H.; He, T.; Zou, Z.; Shen, S.; Li, Y. Using Case Facts to Predict Accusation Based on Deep Learning. In Proceedings of the 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C), Sofia, Bulgaria, 22–26 July 2019; IEEE: New York, NY, USA, 2019; pp. 133–137.
34. Chen, H.; Cai, D.; Dai, W.; Dai, Z.; Ding, Y. Charge-Based Prison Term Prediction with Deep Gating Network. *arXiv* **2019**, arXiv:1908.11521. [CrossRef]
35. Pan, S.; Lu, T.; Gu, N.; Zhang, H.; Xu, C. Charge Prediction for Multi-Defendant Cases with Multi-Scale Attention. In *ChineseC-SCW 2019*; Springer: Kunming, China, 2019; pp. 766–777.
36. Angelidis, I.; Chalkidis, I.; Koubarakis, M. Named Entity Recognition, Linking and Generation for Greek Legislation. In Proceedings of the JURIX 2018: The Thirty-First Annual Conference, Groningen, The Netherlands, 12–14 December 2018; pp. 1–10.
37. Ye, H.; Jiang, X.; Luo, Z.; Chao, W. Interpretable Charge Predictions for Criminal Cases: Learning to Generate Court Views from Fact Descriptions. *arXiv* **2018**, arXiv:1802.08504. [CrossRef]
38. Cardellino, C.; Teruel, M.; Alemany, L.A.; Villata, S. Legal NERC with Ontologies, Wikipedia and Curriculum Learning. In Proceedings of the EACL, Valencia, Spain, 3–7 April 2017; pp. 254–259.
39. Hachey, B.; Grover, C. Extractive Summarisation of Legal Texts. *Artif. Intell. Law* **2006**, *14*, 305–345. [CrossRef]
40. Budur, E.; Özçelik, R.; Soylu, D.; Khattab, O.; Güngör, T.; Potts, C. Building Efficient and Effective OpenQA Systems for Low-Resource Languages. *arXiv* **2024**, arXiv:2401.03590. [CrossRef]
41. Tikhomirov, M.; Chernyshev, D. Impact of Tokenization on LLaMa Russian Adaptation. In Proceedings of the 2023 Ivannikov ISPRAS Open Conference (ISPRAS), Moscow, Russia, 4–5 December 2023; IEEE: New York, NY, USA, 2023; pp. 163–168.
42. Hong, Q.; Liu, S.; Wu, L.; Lu, Q.; Yang, P.; Chen, D.; Cheng, S. Evaluating the Performance of Large Language and Visual-Language Models in Cervical Cytology Screening. *NPJ Precis. Oncol.* **2025**, *9*, 1–10. [CrossRef]
43. Tao, M.; Zhao, D.; Feng, Y. Chain-of-Discussion: A Multi-Model Framework for Complex Evidence-Based Question Answering. *arXiv* **2024**, arXiv:2402.16313.
44. Togmanov, M.; Mukhitaly, N.; Turmakhhan, D.; Mansurov, J.; Goloburda, M.; Sakip, A.; Koto, F. KazMMLU: Evaluating Language Models on Kazakh, Russian, and Regional Knowledge of Kazakhstan. *arXiv* **2025**, arXiv:2502.12829.
45. Rakhimova, D.; Karyukin, V.; Amirova, D.; Sarsenbayeva, A. Collection and Preprocessing of Data for LLM in the Kazakh Language in the Field of Legislation. In *Modeling and Simulation of Social-Behavioral Phenomena in Creative Societies*, MSBC 2024; Springer: Cham, Switzerland, 2024; Volume 2211, pp. 152–166. [CrossRef]
46. Weizenbaum, J. ELIZA—A Computer Program for the Study of Natural Language Communication Between Man and Machine. *Commun. ACM* **1966**, *9*, 36–45. [CrossRef]

47. Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A.A.; Lally, A.; Murdock, J.W.; Nyberg, E.; Prager, J.; et al. Building Watson: An Overview of the DeepQA Project. *AI Mag.* **2010**, *31*, 59–79. [CrossRef]
48. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language Models Are Few-Shot Learners. *arXiv* **2020**, arXiv:2005.14165. [CrossRef]
49. Unger, C.; Bühlmann, L.; Lehmann, J.; Ngonga Ngomo, A.; Gerber, D.; Cimiano, P. Template-Based Question Answering over RDF Data. In Proceedings of the 21st International Conference on World Wide Web (WWW 2012), Lyon, France, 16–20 April 2012; ACM: New York, NY, USA, 2012; pp. 639–648. [CrossRef]
50. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering. In Proceedings of the EMNLP 2018, Brussels, Belgium, 31 October–4 November 2018; pp. 2369–2380.
51. Kim, M.-Y.; Goebel, R. Two-Step Cascaded Textual Entailment for Legal Bar Exam Question Answering. In Proceedings of the 16th International Conference on Artificial Intelligence and Law, London, UK, 12–16 June 2017; pp. 283–290.
52. Taniguchi, R.; Kano, Y. Legal Yes/No Question Answering System Using Case-Role Analysis. In *New Frontiers in Artificial Intelligence, JSAI-isAI 2016 Workshops*; Springer: Berlin/Heidelberg, Germany, 2017; pp. 284–298.
53. da Silva, J.W.F.; Venceslau, A.D.P.; Sales, J.E.; Maia, J.G.R.; Pinheiro, V.C.M.; Vidal, V.M.P. A Short Survey on End-to-End Simple Question Answering Systems. *Artif. Intell. Rev.* **2020**, *53*, 5429–5453. [CrossRef]
54. Wu, P.; Zhang, X.; Feng, Z. A Survey of Question Answering over Knowledge Base. In *Knowledge Graph and Semantic Computing: Knowledge Computing and Language Understanding, CCKS 2019*; Springer: Singapore, 2019; Volume 1134. [CrossRef]
55. Kwon, M.; Bang, J.; Hwang, S.; Jang, J.; Lee, W. A Dynamic-Selection-Based, Retrieval-Augmented Generation Framework: Enhancing Multi-Document Question-Answering for Commercial Applications. *Electronics* **2025**, *14*, 659. [CrossRef]
56. Zhao, W.X.; Liu, J.; Ren, R.; Wen, J.-R. Dense Text Retrieval Based on Pretrained Language Models: A Survey. *ACM Trans. Inf. Syst.* **2024**, *42*, 1–60. [CrossRef]
57. Rakhimova, D.; Kassymova, D.; Isabaeva, D. Research and Development of a Question Answer System Based on the BERT Model for the Kazakh Language. Bulletin of Abai Kazakh National Pedagogical University. Ser. Phys. Math. **2021**, *76*, 119–127. [CrossRef]
58. Yeshpanov, R.; Alimseitova, Z.; Mussakhojayaeva, A.; Serikbay, A.; Kassenov, O. KazQAD: Kazakh Open-Domain Question Answering Dataset. *arXiv* **2024**, arXiv:2401.12345. [CrossRef]
59. Meng, W.; Li, Y.; Chen, L.; Dong, Z. Using the Retrieval-Augmented Generation to Improve the Question-Answering System in Human Health Risk Assessment: The Development and Application. *Electronics* **2025**, *14*, 386. [CrossRef]
60. Gemma Team. Gemma: Open Models Based on Gemini Research and Technology. *arXiv* **2024**, arXiv:2403.08295. [CrossRef]
61. Goloburda, M.; Laiyk, N.; Turmakhambayev, D.; Wang, Y.; Togmanov, M.; Mansurov, J.; Sametov, A.; Mukhituly, N.; Wang, M.; Orel, D.; et al. Qorǵau: Evaluating LLM Safety in Kazakh-Russian Bilingual Contexts. *arXiv* **2025**, arXiv:2502.13640. [CrossRef]
62. Abdin, M.; Aneja, J.; Behl, H.; Bubeck, S.; Eldan, R.; Gunasekar, S.; Harrison, M.; Hewett, R.J.; Javaheripi, M.; Kauffmann, P.; et al. Phi-4 Technical Report. *arXiv* **2024**, arXiv:2412.08905. [CrossRef]
63. Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, Y.; Huang, F.; et al. Qwen Technical Report. *arXiv* **2023**, arXiv:2309.16609. [CrossRef]
64. Hu, R.; Cheng, Y.; Meng, L.; Xia, J.; Zong, Y.; Shi, X.; Lin, W. Training an LLM-as-a-Judge Model: Pipeline, Insights, and Practical Lessons. Companion. In Proceedings of the ACM Web Conference 2025 (WWW Companion ’25), Sydney, NSW, Australia, 28 April–2 May 2025; p. 18. [CrossRef]
65. McHugh, M.L. Interrater reliability: The kappa statistic. *Biochemia Medica* **2012**, *22*, 276–282. [CrossRef]
66. Rahman, A.; Ali, A.; Rahman, A. Comparison of Inter-Rater Agreement Coefficients for Ordinal Data in the Presence of Skewed Distributions. *Symmetry* **2022**, *14*, 262. [CrossRef]
67. Barrett, S.; Bisson, J.; Raghavan, D. Measuring Inter-Rater Agreement for Categorical and Ordinal Data: When and How to Use Kappa. *BMC Cancer* **2023**, *23*, 317. [CrossRef]

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