

Sarsenbayeva Asiya, dissertation titled "Research of models and development of an intellectual question-and-answer system based on semantic approaches for the state language in the field of legislation of the Republic of Kazakhstan", submitted for the degree of Doctor of Philosophy (PhD) in the educational program 8D06101 - Information Systems

ABSTRACT

Relevance of the research topic.

The need to ensure access to legislative information in the official language of the Republic of Kazakhstan is driven by demand itself. Currently, obtaining legislative information for citizens is complex, time-consuming, and often difficult to understand due to the terminology of legal texts. Therefore, a pressing research challenge is the development of intelligent question-and-answer systems that analyze meaning and automatically provide accurate and understandable legal answers.

The rapid development of artificial intelligence technologies, particularly generative language models (LLMs), is opening up new possibilities for processing and analyzing legislative texts. Although models such as GPT, LLaMA, and Gemini have demonstrated high performance in text understanding and generation tasks, their adaptation to agglutinative and resource-poor languages such as Kazakh has been understudied. This, in turn, requires the development of specialized methods that take into account the morphological complexity, diversity of word formation, and specific features of legislative terminology.

Current systems (Adilet, eGov, LawVision) primarily use keyword searches or rule-based approaches and do not fully understand the semantics of the user's question. Furthermore, current LLM models suffer from the "hallucination" problem due to the lack of connection to actual legal sources, which negatively impacts the accuracy of legal responses.

In this regard, there is a need to develop an intelligent question and answer system that takes into account the morphological features of the Kazakh language, combining methods of semantic analysis, the method of generating with advanced search (RAG) and fine - tuning. LLM models. In particular, the implementation of semantic preprocessing methods that match natural language questions with legislative terminology is scientifically sound.

This research aims to improve the accuracy of question and answer systems on legislation in the Kazakh language, expand the availability of legal information, and develop intelligent systems at the national level.

Aim of the dissertation

Research and development of a legislative system of questions and answers that takes into account the morphological features of the Kazakh language and combines semantic analysis and a generative model.

Research objectives:

- To analyze the current state of question and answer systems on legislation in the Kazakh language and identify research problems.
- Collect documents from official portals (adilet.kz, zan.gov.kz) and create a specialized database of legislation.
- To propose a two-stage semantic pre-processing method that determines the semantic relationship of a question and transforms it into legislative form.
- To improve the level of understanding of legal terminology in the Kazakh language by fine - tuning large language models (LLM) in the legislative corpus.
- Develop a pipeline for an intelligent system of questions and answers on legislation.
- Compare the proposed system with basic models and experimentally prove its effectiveness.

Research methods

Natural language processing (NLP), large language models (LLM) based on transformer architecture, RAG technology, SVD/LSA semantic analysis, mT5 transformer, evaluation metrics and expert evaluation methods.

Object of research

Intelligent question-and-answer systems to provide access to the legislative framework of the Republic of Kazakhstan in the state language, as well as the search and processing of legal information based on semantic analysis.

Subject of research

Methods of semantic analysis that take into account the linguistic features and morphological structure of the Kazakh language, methods of formalizing legislative terminology and methods of fine-tuning large language models (LLM) based on a corpus of data in the field of legislation.

Scientific novelty of the obtained results:

1. Given the high complexity and terminological specificity of legal texts, a two-stage semantic preprocessing approach was developed to match natural language questions with legislative terminology.
2. Based on the developed methods, an intelligent question-and-answer system was created, combining semantic preprocessing, adaptive LLM, and RAG technologies. Its effectiveness was confirmed using experimental data and qualitative and quantitative metrics.

Statements submitted for defence:

1. A specialized dataset was compiled, consisting of a digital corpus of regulatory legal acts of the Republic of Kazakhstan with a volume of 5.3 GB and 8,000 question-answer pairs.
2. As a result of fine-tuning large language models based on the collected legislative dataset, the textual accuracy of responses increased by 1.5-2 times compared to the base ROUGE models.
3. A two-stage approach to semantic preprocessing was proposed that matches plain-language queries with legal terminology. This method improved information retrieval quality by 11-15% in terms of Recall@5 and by 16-18% in terms of MRR.
4. A system was developed that combines semantic preprocessing, LLM fine-tuning, and RAG technologies. This allowed for an increase in the semantic accuracy of BERTScore responses by 3-5% and Expert Score by 25-30%.

The volume and structure of the dissertation

The dissertation consists of an introduction, four main chapters, a conclusion, a list of references, and appendices. The total length of the work is 141 typewritten pages, including 40 figures, 41 tables, 108 references, and 1 appendice.

The first chapter presents an analytical review of intelligent question and answer systems in the field of legislation: the evolution of development and modern architectures of question and answer systems, the use of large language models (LLM) in legal information systems, linguistic and semantic features of processing legislative texts, the current state of development of legal question and answer systems in the Kazakh language, and also systematically considers a comparative analysis of domestic and foreign studies.

The second chapter describes the methods for forming the legislative database of the Republic of Kazakhstan and adapting LLM models. It analyzes legal information resources and digital infrastructure; proposes methods for forming a corpus of legislative texts and a set of questions and answers, collecting texts, pre-processing, and annotation; considers a comparative analysis of basic LLM models, subject-specific adaptation of models based on fine-tuning, embedding, and semantic representation methods; and describes the RAG architecture search module, the generative answer generation model, and the overall architectural diagram of the system.

Chapter three describes the conceptual and architectural model of the proposed two-stage semantic preprocessing method, SemanticPrep, the formalization of question and answer computation, the adaptation of the mT5 model through fine-tuning, and the model for converting natural language questions into legal structures; the mathematical model of the SVD/LSA-based semantic expansion method and

semantic similarity assessment methods, as well as the overall algorithm and computational complexity of the proposed method.

Chapter four presents the experimental evaluation of the proposed SemanticPrep method: a comparative analysis of LLM models, the results of implementing the RAG system, the impact of semantic preprocessing on the response quality, and the results of testing the proposed system in a practical environment.

Approbation of the work

The research results were discussed at scientific seminars of Al-Farabi Kazakh National University, presented at international conferences, and underwent practical approbation at LLP "Almaly Corporation".

Publications.

On the topic of the dissertation, 9 scientific papers have been published, including 4 articles in journals indexed in the Scopus and Web of Science databases, 2 papers in the proceedings of international conferences, and 3 papers in journals recommended by the Ministry of Science and Higher Education of the Republic of Kazakhstan.

1. Rakhimova D., Karyukin V., Turarbek A., Sarsenbayeva AK, Aliyev R. (2025). Legal AI in resource-constrained languages: Creation and evaluation of question-answering systems for Kazakhstani legislation. Computers (MDPI). 14(9), 354. <https://doi.org/10.3390/computers14090354> WoS (SCIE), Q1, 2025; 5-year impact factor: 2.7; SJR: 0.492; Percentile: 84 (Computer Science, Applied Sciences)
2. D. Rakhimova, A. Sarsenbaeva, E. Adali, A. Kulzhanova, Classification of questions in the intelligent question and answer system for the state language, Bulletin of the NIS RK No. 4 Almaty, 2024, - P. 166-179 <https://doi.org/10.47533/2024.1606-146X.014>
3. Rakhimova D.R., Sarsenbaeva A.K., Turarbek A., Auezova A.S. Application of deep learning to improve the accuracy of answers in multilingual question-and-answer systems. Bulletin of the NAS RK, Physics and Mathematics Series, No. 2, Almaty, 2025, - P. 196-211 <https://doi.org/10.32014/2025.2518-1726.353>
4. Sarsenbaeva A.K., Rakhimova D.R., Shormakova A.N., Mansurova M. Context-oriented approach to identifying semantic similarity in legal texts of agglutinative languages: the example of the Kazakh language. Bulletin of the National Academy of Sciences of the Republic of Kazakhstan, series "Physics and Mathematics", No. 1, Almaty 2026 – P. 354-372 <https://doi.org/10.32014/2026.2518-1726.417>

5. Rakhimova D., Karyukin V., Karibayeva A., Sarsenbayeva A.K. Developing advanced question answering models for Kazakh legal texts: A comparative study of modern approaches. ACIIDS 2025: 17th Asian Conference on Intelligent Information and Database Systems. DOI: 10.1007/978-981-96-5881-7_19 (Scopus, percentile 39)
6. Rakhimova, D. , Karyukin, V. , Amirova, D. , Sarsenbaeva, A. Data collection and preprocessing for the Kazakh language master's program in law. Communications in Computer and Information Science, 2024, 2211 CCIS, pp . 129–139. Springer. DOI: 10.1007/978-3-031-72260-8_11 (Scopus, percentile)
7. Sarsenbaeva A.K., Rakhimova D.R., Turarbek A. Multilingual Large Language Models in the Legal Field: Performance Analysis in Kazakh, Turkish, and English. 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkey, 2025, pp. 410–415, doi: 10.1109/UBMK67458.2025.11207002. (Scopus, 39th percentile)
8. Sarsenbaeva A.K., Rakhimova D.R. Assessing the quality of answers to questions about legislation in Kazakh, Turkish, and English: An experimental study of multilingual LLM models. X International Scientific and Practical Conference "Informatics and Applied Mathematics", Almaty, 2025. – P. 293-302
9. Rakhimova D., Sarsenbaeva A. Study of the current situation in question and answer systems related to the field of legislation, computer science and applied mathematics: Proc. IX Interscientific and Practical Conference (10/31 - 11/01/2024). - Almaty, 2024, - P. 76-83.