ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№3
2025

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

3 (355)

JULY – SEPTEMBER 2025

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

CHIEF EDITOR:

MUTANOV Galimkair Mutanovich, doctor of technical sciences, professor, academician of NAS RK, acting General Director of the Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/1423665

EDITORIAL BOARD:

KALIMOLDAYEV Maksat Nuradilovich, (Deputy Editor-in-Chief), Doctor of Physical and Mathematical Sciences, Professor, Academician of NAS RK, Advisor to the General Director of the Institute of Information and Computing Technologies of the CS MES RK, Head of the Laboratory (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=56153126500, https://www.webofscience.com/wos/author/record/2428551

Mamyrbaev Orken Zhumazhanovich, (Academic Secretary), PhD in Information Systems, Deputy Director for Science of the Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=55967630400, https://www.webofscience.com/wos/author/record/1774027

BAIGUNCHEKOV Zhumadil Zhanabaevich, Doctor of Technical Sciences, Professor, Academician of NAS RK, Institute of Cybernetics and Information Technologies, Department of Applied Mechanics and Engineering Graphics, Satbayev University (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=6506823633, https://www.webofscience.com/wos/author/record/1923423

WOICIK Waldemar, Doctor of Technical Sciences (Phys.-Math.), Professor of the Lublin University of Technology (Lublin, Poland), https://www.scopus.com/authid/detail.uri?authorId=7005121594, https://www.webofscience.com/wos/author/record/678586

SMOLARJ Andrej, Associate Professor Faculty of Electronics, Lublin polytechnic university (Lublin, Poland), https://www.scopus.com/authid/detail.uri?authorId=56249263000, https://www.webofscience.com/wos/author/record/1268523

KEILAN Alimkhan, Doctor of Technical Sciences, Professor (Doctor of science (Japan)), chief researcher of Institute of Information and Computational Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=8701101900, https://www.webofscience.com/wos/author/record/1436451

KHAIROVA Nina, Doctor of Technical Sciences, Professor, Chief Researcher of the Institute of Information and Computational Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=37461441200, https://www.webofscience.com/wos/author/record/1768515

OTMAN Mohamed, PhD, Professor of Computer Science Department of Communication Technology and Networks, Putra University Malaysia (Selangor, Malaysia), https://www.scopus.com/authid/detail.uri?authorId=56036884700, https://www.webofscience.com/wos/author/record/747649

NYSANBAYEVA Saule Yerkebulanovna, Doctor of Technical Sciences, Associate Professor, Senior Researcher of the Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=55453992600, https://www.webofscience.com/wos/author/record/3802041

BIYASHEV Rustam Gakashevich, doctor of technical sciences, professor, Deputy Director of the Institute for Informatics and Management Problems, Head of the Information Security Laboratory (Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=6603642864, https://www.webofscience.com/wos/author/record/3802016

KAPALOVA Nursulu Aldazharovna, Candidate of Technical Sciences, Head of the Laboratory cybersecurity, Institute of Information and Computing Technologies CS MES RK (Almaty, Kazakhstan), https://www.scopus.com/authid/detail.uri?authorId=57191242124,

KOVALYOV Alexander Mikhailovich, Doctor of Physical and Mathematical Sciences, Academician of the National Academy of Sciences of Ukraine, Institute of Applied Mathematics and Mechanics (Donetsk, Ukraine), https://www.scopus.com/authid/detail.uri?authorId=7202799321, https://www.webofscience.com/wos/author/record/38481396

MIKHALEVICH Alexander Alexandrovich, Doctor of Technical Sciences, Professor, Academician of the National Academy of Sciences of Belarus (Minsk, Belarus), https://www.scopus.com/authid/detail.uri?authorId=7004159952, https://www.webofscience.com/wos/author/record/46249977

TIGHINEANU Ion Mihailovich, Doctor of Physical and Mathematical Sciences, Academician, President of the Academy of Sciences of Moldova, Technical University of Moldova (Chisinau, Moldova), https://www.scopus.com/authid/detail.uri?authorId=7006315935, https://www.webofscience.com/wos/author/record/524462

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

© «Central Asian Academic Research Center» LLP, 2025

БАС РЕДАКТОР:

МҰТАНОВ Ғалымқайыр Мұтанұлы, техника ғылымдарының докторы, профессор, ҚР ҰҒА академигі, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты» бас директорының м.а. (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/1423665

РЕДАКЦИЯ АЛҚАСЫ:

ҚАЛИМОЛДАЕВ Мақсат Нұрәділұлы, (бас редактордың орынбасары), физика-математика ғылымдарының докторы, профессор, ҚР ҰҒА академигі, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты» бас директорының кеңесшісі, зертхана меңгерушісі (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=56153126500, https://www.webofscience.com/wos/author/record/2428551

МАМЫРБАЕВ Өркен Жұмажанұлы (ғалым хатшы), Ақпараттық жүйелер саласындағы техника ғылымдарының (PhD) докторы, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты» директорының ғылым жөніндегі орынбасары (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=55967630400, https://www.webofscience.com/wos/author/record/1774027

БАЙГУНЧЕКОВ Жұмаділ Жаңабайұлы, техника ғылымдарының докторы, профессор, ҚР ҰҒА академигі, Кибернетика және ақпараттық технологиялар институты, Қолданбалы механика және инженерлік графика кафедрасы, Сәтбаев университеті (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=6506823633, https://www.webofscience.com/wos/author/record/1923423

ВОЙЧИК Вальдемар, техника ғылымдарының докторы (физ-мат), Люблин технологиялық университетінің профессоры (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=7005121594, https://www.webofscience.com/wos/author/record/678586

СМОЛАРЖ Анджей, Люблин политехникалық университетінің электроника факультетінің доценті (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=56249263000, https://www.webofscience.com/wos/author/record/1268523

КЕЙЛАН Әлімхан, техника ғылымдарының докторы, профессор (ғылым докторы (Жапония)), ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институтының» бас ғылыми қызметкері (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=8701101900, https://www.webofscience.com/wos/author/record/1436451

ХАЙРОВА Нина, техника ғылымдарының докторы, профессор, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институтының» бас ғылыми қызметкері (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=37461441200, https://www.webofscience.com/wos/author/record/1768515

ОТМАН Мохаммед, PhD, Информатика, Коммуникациялық технологиялар және желілер кафедрасының профессоры, Путра университеті Малайзия (Селангор, Малайзия), https://www.scopus.com/authid/detail.uri?authorId=56036884700, https://www.webofscience.com/wos/author/record/747649

НЫСАНБАЕВА Сауле Еркебұланқызы, техника ғылымдарының докторы, доцент, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институтының» аға ғылыми қызметкері (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=55453992600, https://www.webofscience.com/wos/author/record/3802041

БИЯШЕВ Рустам Гакашевич, техника ғылымдарының докторы, профессор, Информатика және басқару мәселелері институты директорының орынбасары, Ақпараттық қауіпсіздік зертханасының меңгерушісі (Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=6603642864, https://www.webofscience.com/wos/author/record/3802016

КАПАЛОВА Нұрсұлу Алдажарқызы, техника ғылымдарының кандидаты, ҚР ҒЖБМ ҒК «Ақпараттық және есептеу технологиялары институты», Киберқауіпсіздік зертханасының меңгерушісі (Алматы, Қазақстан), https://www.scopus.com/authid/detail.uri?authorId=57191242124,

КОВАЛЕВ Александр Михайлович, физика-математика ғылымдарының докторы, Украина Ұлттық Ғылым академиясының академигі, Қолданбалы математика және механика институты (Донецк, Украина), https://www.scopus.com/authid/detail.uri?authorId=7202799321, https://www.webofscience.com/wos/author/record/38481396

МИХАЛЕВИЧ Александр Александрович, техника ғылымдарының докторы, профессор, Беларусь Ұлттық Ғылым академиясының академигі (Минск, Беларусь), https://www.scopus.com/authid/detail.uri?authorId=7004159952, https://www.webofscience.com/wos/author/record/46249977

ТИГИНЯНУ Ион Михайлович, физика-математика ғылымдарының докторы, академик, Молдова Ғылым Академиясының президенті, Молдова техникалық университеті (Кишинев, Молдова), https://www.scopus.com/authid/detail.uri?authorId=7006315935, https://www.webofscience.com/wos/author/record/524462

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).

Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.

Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*

Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*

Мерзімділігі: *жылына 4 рет.*

http://www.physico-mathematical.kz/index.php/en/

© «Орталық Азия академиялық ғылыми орталығы» ЖШС, 2025

ГЛАВНЫЙ РЕДАКТОР:

МУТАНОВ Галимкаир Мутанович, доктор технических наук, профессор, академик НАН РК, и.о. генерального директора «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6506682964, https://www.webofscience.com/wos/author/record/142 3665

Редакционная коллегия:

КАЛИМОЛДАЕВ Максат Нурадилович, (заместитель главного редактора), доктор физико-математических наук, профессор, академик НАН РК, советник генерального директора «Института информационных и вычислительных технологий» КН МНВО РК, заведующий лабораторией (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=56153126500, https://www.webofscience.com/wos/author/record/2428551

МАМЫРБАЕВ Оркен Жумажанович, (ученый секретарь), доктор философии (PhD) по специальности «Информационные системы», заместитель директора по науке РГП «Институт информационных и вычислительных технологий» Комитета науки МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=55967630400, https://www.webofscience.com/wos/author/record/1774027

БАЙГУНЧЕКОВ Жумадил Жанабаевич, доктор технических наук, профессор, академик НАН РК, Институт кибернетики и информационных технологий, кафедра прикладной механики и инженерной графики, Университет Сатпаева (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6506823633, https://www.webofscience.com/wos/author/record/1923423

ВОЙЧИК Вальдемар, доктор технических наук (физ.-мат.), профессор Люблинского технологического университета (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=7005121594, https://www.webofscience.com/wos/author/record/678586

СМОЛАРЖ Анджей, доцент факультета электроники Люблинского политехнического университета (Люблин, Польша), https://www.scopus.com/authid/detail.uri?authorId=56249263000, https://www.webofscience.com/wos/author/record/1268523

КЕЙЛАН Алимхан, доктор технических наук, профессор (Doctor of science (Japan)), главный научный сотрудник РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=8701101900, https://www.webofscience.com/wos/author/record/1436451

ХАЙРОВА Нина, доктор технических наук, профессор, главный научный сотрудник РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=37461441200, https://www.webofscience.com/wos/author/record/1768515

ОТМАН Мохамед, доктор философии, профессор компьютерных наук, Департамент коммуникационных технологий и сетей, Университет Путра Малайзия (Селангор, Малайзия), https://www.scopus.com/authid/detail.uri?authorId=56036884700, https://www.webofscience.com/wos/author/record/747649

НЫСАНБАЕВА Сауле Еркебулановна, доктор технических наук, доцент, старший научный сотрудник РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=55453992600, https://www.webofscience.com/wos/author/record/3802041

БИЯШЕВ Рустам Гакашевич, доктор технических наук, профессор, заместитель директора Института проблем информатики и управления, заведующий лабораторией информационной безопасности (Казахстан), https://www.scopus.com/authid/detail.uri?authorId=6603642864,https://www.webofscience.com/wos/author/record/3802016

КАПАЛОВА Нурсулу Алдажаровна, кандидат технических наук, заведующий лабораторией кибербезопасности РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), https://www.scopus.com/authid/detail.uri?authorId=57191242124,

КОВАЛЕВ Александр Михайлович, доктор физико-математических наук, академик НАН Украины, Институт прикладной математики и механики (Донецк, Украина), https://www.scopus.com/authid/detail.uri?authorId=7202799321, https://www.webofscience.com/wos/author/record/38481396

МИХАЛЕВИЧ Александр Александрович, доктор технических наук, профессор, академик НАН Беларуси (Минск, Беларусь), https://www.scopus.com/authid/detail.uri?authorId=7004159952, https://www.webofscience.com/wos/author/record/46249977

ТИГИНЯНУ Ион Михайлович, доктор физико-математических наук, академик, президент Академии наук Молдовы, Технический университет Молдовы (Кишинев, Молдова), https://www.scopus.com/authid/detail.uri?authorId=7006315935, https://www.webofscience.com/wos/author/record/524462

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*
Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**
Тематическая направленность: *информационно-коммуникационные технологии.*
В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*
Периодичность: *4 раза в год.*
http://www.physico-mathematical.kz/index.php/en/

© ТОО «Центрально-азиатский академический научный центр», 2025

CONTENTS

МАЗМҰНЫ

СОДЕРЖАНИЕ

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 3. Namber 355 (2025). 78–92

https://doi.org/10.32014/2025.2518-1726.365

ҒТМР 20.19.27:
ОӘЖ 004.8

Zh. Bazarbek, N. Toyganbaeva*, M. Mansurova, T. Sarsembayeva, M. Sakypbekova, 2025.
Al-Farabi Kazakh National University, Almaty, Kazakhstan.
E-mail: bodinaz@mail.ru

DEVELOPING A DATASET FOR CREATING A LARGE LANGUAGE MODEL (LLM) FOR THE KAZAKH LANGUAGE

Madina Mansurova — Candidate of Physico-mathematical Sciences, Professor, Head of the Department of Artificial Intelligence and Big Data at Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Madina.Mansurova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-9680-2758;
Toyganbaeva Nazgul — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: bodinaz@mail.ru, ORCID ID: https://orcid.org/0000-0003-2661-8661;
Bazarbek Zhaniya — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Zhaniya.Bazarbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-7838-2104;
Sarsembayeva Talshyn — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: talshyn.sagdatbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0001-7668-2640;
Sakypbekova Meruyert — senior lecturer at the Department of Artificial Intelligence and Big Data of Al-Farabi Kazakh National University, Almaty, Kazakhstan,
E-mail: Meruert.Sakypbekova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-6652-1357.

Abstract. This study focuses on the development of a Large Language Model (LLM) for the Kazakh language and provides a comprehensive overview of its theoretical foundations, methodological aspects, and practical applications. A model designed with consideration of the morphological, syntactic, dialectal, and orthographic features of the Kazakh language represents a significant step in advancing artificial intelligence. The relevance of this research lies in the need to create intelligent systems capable of professionally processing Kazakh texts and providing users with accurate and well-grounded responses. Within the framework of the project, a specialized dataset was collected and structured. The purpose of this article is to create a premium dataset suitable for LLM training by collecting data in the Kazakh language and presenting it in a high-quality and accessible form. The dataset development addressed key challenges such as filling gaps in linguistic materials, covering dialectal diversity, incorporating orthographic variations, and

including diverse usage scenarios. The application of OCR technology enabled the digitization of materials from multiple sources and their transformation into formats convenient for processing. Furthermore, methods for annotation, structuring, and systematization of data were proposed, contributing to improved model reliability and accuracy. The study also analyzes advanced methodologies such as MBERT and GPT, emphasizing their limitations in processing the Kazakh language. The importance of building unique datasets for low-resource languages is particularly highlighted. The findings demonstrate the potential for applying AI in government, education, healthcare, business, and digital services. Thus, this work contributes to reducing informational inequality, integrating the Kazakh language into the global AI ecosystem, and fostering Kazakhstan's technological progress.

Keywords: Kazakh language, large language model (LLM), mBERT, dataset, natural language processing (NLP)

Acknowledgment: This research was funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001-OT-24).

Ж.П. Базарбек, Н.А. Тойганбаева*, М.Е. Мансурова,
Т.С. Сарсембаева, М.Ж. Сакыпбекова, 2025.
Әл-Фараби атындағы Қазақ ұлттық университеті.
E-mail: bodinaz@mail.ru

ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН ҮЛКЕН ТІЛ МОДЕЛІН (LLM) ЖАСАУ ҮШІН DATASET ӘЗІРЛЕУ

Мансурова Мадина Есимхановна — ф.-м. ғ.к., профессор, әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының меңгерушісі, Алматы, Қазақстан,
E-mail: Madina.Mansurova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-9680-2758;
Тойганбаева Назгуль Абеновна — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: bodinaz@mail.ru ORCID ID: https://orcid.org/0000-0003-2661-8661;
Базарбек Жания Пархатқызы — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: Zhaniya.Bazarbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-7838-2104;
Сарсембаева Талшын Сағдатбекқызы — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: talshyn.sagdatbek@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0001-7668-2640;
Сакыпбекова Меруерт Жумабековна — Әл-Фараби атындағы ҚазҰУ, Жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Қазақстан,
E-mail: Meruert.Sakypbekova@kaznu.edu.kz, ORCID ID: https://orcid.org/0000-0002-6652-*1357,*

Аннотация. Бұл зерттеу қазақ тілі үшін үлкен тіл моделін (LLM) құруға арналған және оның ғылыми-теориялық негіздерін, әдістемелік қырларын, сондай-ақ практикалық қолдану мүмкіндіктерін кешенді қарастырады. Қазақ тілінің морфологиялық, синтаксистік, диалектілік және

әртүрлі графикалық жүйелердегі ерекшеліктерін ескеріп жасалған модель жасанды интеллектті дамытудағы маңызды қадам болып саналады. Мұндай бағыттың маңыздылығы қазақ тілінде кәсіби деңгейде мәтіндермен өзара әрекеттесетін, пайдаланушы сұрақтарына толық әрі орынды жауап беретін интеллектуалды жүйелерді қалыптастыру қажеттілігімен байланысты. Жоба аясында арнайы dataset жиналып, құрылымдалды. Осы мақаланың мақсаты – қазақ тіліндегі деректерді жинақтап, сапалы әрі қолжетімді формада ұсыну арқылы LLM оқытуға жарамды премиум-жиынтық құру. Dataset әзірлеу барысында жетіспейтін мәтіндік ресурстарды толықтыру, диалектілерді қамту, түрлі орфографиялық нұсқаларды енгізу және әртүрлі сценарийлерді ескеру мәселелері шешілді. OCR технологиясын қолдану деректерді әртүрлі көздерден цифрландыруға және өңдеуге ыңғайлы пішімдерге түрлендіруге мүмкіндік берді. Сонымен бірге деректерді аннотациялау, құрылымдау және жүйелеу әдістері ұсынылып, олардың модельдің сенімділігі мен дәлдігін арттыруға ықпалы айқындалды. Зерттеу MBERT және GPT сияқты алдыңғы қатарлы әдіснамаларды талдауды қамтиды. Бұл модельдердің қазақ тілімен жұмыс істеудегі шектеулері көрсетіліп, ресурсы шектеулі тілдер үшін арнайы dataset қалыптастырудың маңыздылығы ерекше атап өтілді. Жобаның нәтижелері мемлекеттік басқару, білім беру, медицина, бизнес және цифрлық қызмет көрсету салаларында жасанды интеллектті кеңінен қолдануға мүмкіндік береді. Осылайша, атқарылған жұмыс қазақ тілінің ақпараттық теңсіздігін азайтуға, оны жаһандық жасанды интеллект экожүйесіне енгізуге және Қазақстанның технологиялық дамуына жаңа серпін беруге бағытталған.

Түйін сөздер: жасанды интеллект, қазақ тілі, табиғи тілді өңдеу, машиналық оқыту, үлкен тіл моделі (LLM), mBERT, GPT синтетикалық деректер

Алғыс: Бұл зерттеу Қазақстан Республикасының Ғылым және жоғары білім министрлігінің Ғылым комитеті тарапынан қаржыландырылды (Грант № BR24993001-ОТ-24).

Ж.П. Базарбек, Н.А. Тойганбаева*, М.Е. Мансурова, Т.С. Сарсембаева, М.Ж. Сакыпбекова, 2025.
Казахский Национальный университет им. аль-Фараби, Алматы, Казахстан.
E-mail: bodinaz@mail.ru

РАЗРАБОТКА ДАТАСЕТА ДЛЯ СОЗДАНИЯ БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ (LLM) ДЛЯ КАЗАХСКОГО ЯЗЫКА

Мансурова Мадина Есимхановна — к.ф.-м.-н., профессор, Заведующая кафедрой Искусственного интеллекта и Big Data КазНУ имени Аль-Фараби, Алматы, Казахстан,
E-mail: Madina.Mansurova@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0002-9680-2758;
Тойганбаева Назгуль Абеновна — старший преподаватель кафедры Искусственного интеллекта и Big Data КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: bodinaz@mail.ru, ORCID ID: https://orcid.org/0000-0003-2661-8661;

Базарбек Жания Пархатқызы — старший преподаватель кафедры Искусственного интеллекта и Big Data КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: Zhaniya.Bazarbek@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0002-7838-2104;
Сарсембаева Талшын Сағдатбекқызы — старший преподаватель кафедры Искусственного интеллекта и Big Data КазНУ имени аль-Фараби, Алматы, Казахстан,
E-mail: talshyn.sagdatbek@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0001-7668-2640;
Сакыпбекова Меруерт Жумабековна — Әл-Фараби атындағы ҚазҰУ жасанды интеллект және Big Data кафедрасының аға оқытушысы, Алматы, Казахстан,
E-mail: Meruert.Sakypbekova@kaznu.edu.kz , ORCID ID: https://orcid.org/0000-0002-6652-1357.

Аннотация. Данное исследование посвящено созданию большой языковой модели (LLM) для казахского языка и комплексно рассматривает её научно-теоретические основы, методологические аспекты, а также практические возможности применения. Модель, разработанная с учётом морфологических, синтаксических, диалектологических и графических особенностей казахского языка, является важным шагом в развитии искусственного интеллекта. Актуальность данного направления обусловлена необходимостью формирования интеллектуальных систем, способных профессионально работать с текстами на казахском языке и давать пользователям полные и обоснованные ответы. В рамках проекта был собран и структурирован специализированный dataset. Цель данной статьи – создание премиального набора данных, пригодного для обучения LLM, за счёт систематизации и представления казахскоязычных текстов в удобной форме. В процессе разработки были решены задачи восполнения недостатка языковых материалов, охвата диалектов, включения различных орфографических вариантов и сценариев использования. Применение технологии OCR позволило оцифровать материалы из различных источников и преобразовать их в удобные для обработки форматы. Кроме того, были предложены методы аннотирования, структурирования и систематизации данных, способствующие повышению точности и надёжности моделей. Исследование включает анализ передовых методологий, таких как MBERT и GPT, с указанием их ограничений при работе с казахским языком. Особо подчёркнута значимость формирования уникальных наборов данных для языков с ограниченными ресурсами. Полученные результаты открывают возможности широкого применения ИИ в сфере государственного управления, образования, медицины, бизнеса и цифровых услуг. Таким образом, проделанная работа направлена на сокращение информационного неравенства, интеграцию казахского языка в глобальную экосистему искусственного интеллекта и придание нового импульса технологическому развитию Казахстана.

Ключевые слова: искусственный интеллект, казахский язык, обработка естественного языка, машинное обучение, большая языковая модель (LLM), mBERT, GPT синтетические данные

Благодарность: Данное исследование было профинансировано Комитетом науки Министерства науки и высшего образования Республики Казахстан (Грант № BR24993001-ОТ-24).

Кіріспе. Үлкен Тілдік Модельдер (LLM) — бұл табиғи тілді адамға жақын деңгейде құруға және өңдеуге қабілетті жасанды интеллект. Олар автоматты түрде аудармада, мәтінді талдауда, дауыстық көмекшілерде және басқа салаларда қолданылады. Қазақ тілінің бірегей морфологиялық және синтаксистік ерекшеліктері бар. Осыған сүйене отырып, ағылшын және басқа да әлемдік тілдерге бағытталған қолданыстағы LLM-ді өңдеу қиын. Мамандандырылған модельді құру қазақ тіліндегі мәтіндерді неғұрлым нақты және табиғи түсінуді және қалыптастыруды қамтамасыз етеді. Тілдік модельдің сапасы оқыту деректерінің көлемі мен оның түрлілігіне тікелей байланысты. Деректер қоры неғұрлым бай және таза болса, модельдің нәтижелері соғұрлым жақсы болады. Дереккөздердің сан түрлілігін ескеру және әртүрлі стильдер мен диалектілер арасындағы тепе-теңдікті сақтау маңызды.

Үлкен тілдік модельдер саласындағы зерттеулер табиғи тілдерді өңдеу технологияларының қарқынды дамуын көрсетеді. mBERT сияқты трансформаторлық модельдердің пайда болуы маңызды кезеңдердің бірі болды (Devlin et al., 2018). GPT-3 және GPT-4 модельдері жоғары сапалы мәтіндерді жасауды, автоматты аударманы, чатбот жасауды және басқа да көптеген тапсырмаларды орындай алды (OpenAI 2023). Тиісті зерттеу бағыттарының ішінде мыналар бар: Дәл баптау: мамандандырылған деректер корпусында үлкен үлгілерді оқыту медицина, құқық және техникалық аударма сияқты жоғары мамандандырылған салаларда дәлдікті арттыруға мүмкіндік береді (Howard & Ruder, 2018). Көптілді модельдер үшін зерттеулерде көрсеткендей, mBERT және BLOOM сияқты модельдер мәтіндерді бірнеше тілде өңдей алады, бірақ ресурстары аз тілдерді өңдеу сапасы әлі де үлкен мәселе болып қала береді. LLM-нің біржақтылығы мен этикасында ғалымдар деректердің модельдік бейімділікке әсерін және деректерді сүзу және теңгерімдеу арқылы оны азайту жолдарын зерттейді (Bender et al., 2021). Никбахт Р. Және басқа ғалымдардың ұсынған ғылыми жұмысында 3gpp (3-Generation Partnership Project) техникалық сипаттамаларын түсіну үшін үлкен тілдік үлгілерді (LLM) оқытуға арналған ашық деректер жинағын ұсынады. Зерттеушілер деректер жиынтығының сапасын бағалау үшін сұрақтарға жауап бермес бұрын тиісті ақпаратты алу үшін RAG жүйесін қолданған (Nikbakht et al. 2024). Есептеу шығындарын оңтайландыру және азайтуда жаңа зерттеулер модельдердің көлемін азайтуға және олардың сапасын айтарлықтай жоғалтпай тездетуге бағытталған. Зерттеудің бұл бағыттары қазақ тіліне арналған мамандандырылған модельдерді қоса алғанда, LLM-нің дамуына негіз болады.

Материалдар мен әдістер.

Қазақ тілінің морфологиялық және синтаксистік күрделілігі. Қазақ тілі агглютинативті болып табылады. Яғни, бұл сөздердің мағынасын өзгертетін көптеген аффикстердің болуын білдіреді. Бұл мәтінді лемматизациялау мен токенизациялауды қиындатады. Латын, кириллица, қысқартылған жазу мәселелері бойынша қазақ тілі бірнеше алфавиттерді қолданады: тарихи контексте кириллица, латын және араб жазуы. Осылайша аралас пайдалану мәтіндерді өңдеуде қиындықтар туғызады. Диалектілер және стилистикалық ерекшеліктер, диалектілердегі айырмашылықтар мен сөйлеу мәнері (ресми, ауызекі, жастар жаргоны) әмбебап модель құру үшін деректерді жинауды және қалыпқа келтіруді талап етеді. Бұл морфологиялық күрделілік LLM әзірлеуде ерекше назар аударуды талап етеді:

- Көптік аффикстер жүйесі: кітап → кітаптар → кітаптарым
- Көптеген септік формалары: кітап → кітапқа, кітаптан, кітаппен
- Диалектілік және стилистикалық айырмашылықтар: ресми (Үкімет шешім қабылдады), ауызекі (Бүгін кешке кездесеміз бе?), жастар жаргоны (Түсінбедім, қазір тексеремін)
- Әліпбилердің әртүрлілігі: кириллица, латын графикасы, араб жазуы

Бұл ерекшеліктерді ескермей құрылған модель қазақ тіліндегі мәліметтерді дәл өңдей алмайды. №1 кестеде LLM мен басқа да машиналық оқыту жобалары үшін қажетті дерекқорларды құруда қолданылатын негізгі дерек көздері әртүрлі болады.

№1 кесте - Деректерді жинау және өңдеу

Ашық дереккөздер	
Дереккөз түрі	Мысалдар
Кітаптар, Мақалалар, энциклопедиялар	Әдеби шығармалар, ғылыми басылымдар, газеттер
Жаңалықтар сайттары, блогтар, форумдар	Онлайн басылымдар, талқылау платформалары
Әлеуметтік желі	Платформаларда пайдаланушы жасаған мазмұнды мәтіндер Twitter, Facebook
Жабық дереккөздер	
Диссертациялық жұмыстар	Философия докторы (PhD) дәрежесін алу үшін дайындалған диссертациялар
Кітаптар, энциклопедиялар	Әл-Фараби атындағы ҚазҰУ кітапханасынан әдеби шығармалар, ғылыми басылымдар, кітаптар

Мәтіндерді алдын ала өңдеу қадамдары. Бірінші, дубликаттарды жою – бірдей мәтіндердің бірнеше рет қолданылуын болдырмау. Екінші, орфографиялық және грамматикалық қателерді түзету. Дұрыс белгіленген деректер сапаны арттырады. Үшінші, қалыпқа келтіру (Normalization) – жазу стилін біріздендіру. Төртінші форматтау және құрылымдау – JSON, XML немесе CSV форматтарында сақтау.

Датасетті "docx" форматында құжаттарды автоматты түрде өңдеуге арналған бағдарлама әзірленді және python ортасында өңделді. Негізгі мақсат – мәтінді шығару, оны жеке сөйлемдерге бөлу және одан әрі талдау немесе өңдеу үшін `txt` форматында сақтау. Және ол сол құжаттағы басқа тілдерді, формула, суреттерді жойып, жаңа форматқа сақтайды.

Негізгі жұмысы келесідей:

1. Бағдарлама 'docx' файлын ашады және барлық мәтінді шығарады.

2. Содан кейін мәтін артық бос орындардан, жолдарды тасымалдаудан және басқа қажетсіз таңбалардан тазартылады.

3. Алынған мәтін тыныс белгілері ережелері арқылы сөйлемдерге бөлінеді.

4. Барлық сөйлемдер ' txt ' файлында сақталады, мұнда әр сөйлем бөлек жолда орналасады.

Бұл тәсіл мәтіндік ақпаратты құрылымдалған түрде ыңғайлы сақтауға, сондай-ақ оны әрі қарай талдау, табиғи тілді өңдеу және басқа да тапсырмалар үшін пайдалануға мүмкіндік береді.

Қазақ тіліне арналған LLM құрудағы негізгі қиындықтардың бірі — деректер көлемінің шектеулі болуы. Осы мәселені шешу үшін синтетикалық деректерді генерациялау әдістері қолданылады, оның ішінде мәтіндерді парафразалау, синонимдерді қолдану, статистикалық және нейрондық генерациялау тәсілдері қолданылады. Сондай-ақ, деректерді қолмен аннотациялау және белгілеу процесі маңызды рөл атқарады, бұл деректердің сапасын арттырып, модельдің түсіну қабілетін жақсартады. Сонымен қатар, интернеттен және түрлі сандық архивтерден автоматты түрде деректер жинау әдістері де қарастырылады.

Dataset сапасын бағалау үшін толықтық, әртүрлілік, синтаксистік дұрыстық және белгілеу стандарттарына сәйкестік көрсеткіштері қолданылады. Деректердің теңгерімділігі, стильдік вариативтілігі және тақырыпты қамту деңгейі де маңызды. Сонымен қатар, корпус сапасын жақсарту үшін автоматтандырылған валидация және адамдық тексеру әдістері қатар қолданылады. Машиналық оқыту жүйелері үшін деректерді бөлудің (train/test/validation split) дұрыстығын сақтау модельдің генерализациялау қабілетін арттырады.

LLM үшін dataset таңдау және дайындау – оның өнімділігі мен тиімділігіне тікелей әсер ететін маңызды қадам. №2 кестеде көптеген зерттеу мен өндірістік жобаларда қолданылатын дайын Dataset-тер.

№2 - Датасеттерге шолу

Датасеттер	**Сипаттамасы**	**Жылы**	**Қолжазба тілі**
Common Crawl (Common Crawl, 2025)	шамамен 1 триллион сөздер бар.	2007	ағылшын
The Pile (EleutherAI, EleutherAI. The Pile: GitHub repository, 2025)	800 ГБ дереккөзі, (кітаптар, код, ғылыми мақалалар).	2020	ағылшын

PubMed (PubMed, 2025)	биомедициналық әдебиеттеріне, жаратылыстану журналдарына және онлайн кітаптарға 37 миллионнан астам сілтемелерді қамтиды.	1997	ағылшын
Open Corpora (OpenCorpora, 2025)	Орыс тіліндегі мәтіндердің үлкен коллекциясы. Әдеби шығармалар, ресми құжаттар, энциклопедиялық мәтіндер.	2009	орыс
Yandex Toloka Datasets (Toloka.ai, 2025)	Toloka платформасы арқылы жиналған және өңделген орыс тіліндегі деректер. Қамтитын мазмұн: Жаңалықтар, техникалық мәтіндер, пікірлер.	2014	орыс
Қазақ тілінің ұлттық корпусы (Qazcorpus, 2025)	қазақ тілінің лексика-грамматикалық жүйесін толық қамтыған (терең аннотацияланған) миллиондаған сөзқолданыстан тұратын электронды пішіндегі көлемді мәтіндер жинағы, Жалпы сөзқолданыс саны – 65 000 000. 16 ішкорпустан тұрады.	2012	қазақ
KOHTD	3000 қолжазба емтихан жұмысы мен 140335-тен астам сегменттелген суреттері бар және шамамен 922010 таңбадан тұратын қазақ тіліндегі офлайн қолжазба мәтіндік деректер жинағы (Kazakh offline Handwritten Text dataset - KOHTD) (Toiganbayeva et al, 2022)	2022	қазақ
Kaz_txt_Dataset	Әр түрлі сала бойынша қазақ тіліндегі кең датасет	2024	қазақ

Қазақ тілі үшін сапалы дереккөздердің жетіспеуі LLM өнімділігін шектейді. Қазақ тілі үшін сапалы dataset құру – үлкен тілдік модельдерді (LLM) оқыту мен дамытудың маңызды аспектілерінің бірі. Бұл өзектілікті бірнеше негізгі факторлармен түсіндіруге болады:

- Тілдік ресурстар тапшылығы – қазақ тілінде үлкен әрі әртүрлі деректер жиынтығы жоқтың қасы.

- LLM қазақ тілінде нашар жұмыс істейді – Қолданыстағы GPT, LLaMA, Mistral модельдері қазақ тіліндегі мәтіндерді өңдеуде жиі қателеседі.

- Ақпараттық теңсіздік – Қазақ тілді қолданушылар үшін жасанды интеллект мүмкіндіктері шектеулі болып отыр.

- Сапалы dataset қазақ тілінде дамыған GPT сияқты модельдерді жасауға мүмкіндік береді.

- Мемлекеттік қызмет, білім беру, бизнес, медицина және тағы басқа салаларда жасанды интеллект құралдарын тиімді қолдануға жол ашады.

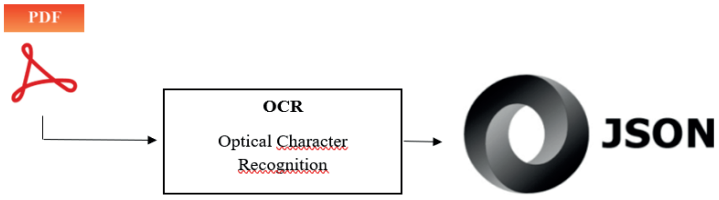
- Егер қазақ тілі үшін сапалы деректер жинақталмаса, ол үлкен тілдік модельдердің көлеңкесінде қалып қоюы мүмкін.

- Өзіміздің dataset-іміз болмаса, AI қазақ тілін қате немесе бұрмаланған түрде түсіндіруі мүмкін.

Міне, осы өзектілі мәселелер «Қазақ тілі мен технологиялық прогресті қолдау үшін үлкен тіл моделін (LLM) құру» жобасы аясында шешімін тауып,

Kaz_txt_Dataset құрылып жатыр. Kaz_txt_Dataset құру үшін pdf форматтағы материалдар OCR технологиясына негізделіп, txt, JSON форматтарына аударылып (1-сурет.), жинақталған мәтіндер әр сала бойынша іріктелген.

1-сурет. pdf форматтағы материалдар OCR технологиясына негізделіп JSON форматтарына аудару



OCR технологиялары dataset-ті жаңа деректер көздерімен толықтыру арқылы баспа және қолжазба мәтіндерін цифрландыруға көмектеседі. OCR жүйелері латын, кириллица және араб әліпбиіндегі мәтіндерді тану және конвертациялау мүмкіндігін ескеруі керек. Бұған қоса, тарихи және сирек кездесетін құжаттардан алынған мәтіндерді тану, түзету және валидациялау әдістерін жетілдіру маңызды. OCR-дің қазақ тіліне бейімделуі үшін арнайы нейрондық желілерді оқыту қажет, бұл кириллица мен латын графикасын қатар тануға мүмкіндік береді.

Dataset болашақта өзін-өзі оқытатын модельдерге арналған нейрондық желілермен интеграциялануы мүмкін, бұл деректердің сапасын автоматты түрде жақсартуға мүмкіндік береді. Сонымен қатар, әртүрлі салаларға арналған мамандандырылған модельдерді дамыту (құқықтану, медицина, білім беру) маңызды бағыттардың бірі болып табылады. Бұдан бөлек, мультимодальды модельдер құру — яғни мәтін, аудио және бейнені біріктіру арқылы кеңейтілген LLM жасау да өзекті. Мысалы, сөйлеу тілін тану және генерациялау мүмкіндіктерін жетілдіру қазақ тіліндегі қолданушыларға ыңғайлырақ жүйелер жасауға мүмкіндік береді. Қазақ тілінде LLM дамытудағы маңызды қадамдардың бірі — диалектілік ерекшеліктерді есепке алатын және морфологиялық құрылымға сезімтал модельдер жасау, бұл өз кезегінде қазақ тілінің цифрлық трансформациясына айтарлықтай үлес қосады.

Қазақ тілінде LLM құруға байланысты бірнеше зерттеу жұмыстары бар. MBERT және GPT-4 сияқты көп тілді модельдер қазақ тілімен жұмыс істеу қабілетін көрсетсе де, олар негізінен ағылшын және басқа да жоғары ресурстық тілдерге бейімделген. Қазақ тілінің лингвистикалық ерекшеліктерін неғұрлым терең және дәл өңдеу үшін мынадай тәсілдер әзірленді:

- KazNERD (Yeshpanov et al., 2022) - атаулар, ұйымдар және географиялық атаулар сияқты субъектілердің қолмен аннотациясын қамтитын қазақ тілінде аталған субъектілерді тану міндеті үшін құрылған датасет.

- KazGPT-жергілікті бастамалар шеңберінде әзірленген қазақ мәтінін генерациялауға бағытталған прототиптік тілдік модель.

- Қолмен белгіленген корпустарға негізделген модельдер — морфологиялық және синтаксистік талдау үшін, соның ішінде semeval және Universal Dependencies халықаралық жобалары аясында қолданылды.

Біз ұсынатын әдіс мамандандырылған қазақ тілді деректер жиынтығын және бейімделген трансформаторлық архитектураны қолдану есебінен мәтінді генерациялау және түсіну сапасын арттыру мақсатында аталған модельдердің артықшылықтарын біріктіруді көздейді. Қазақ тіліне арналған LLM әзірлеу үшін бірнеше эксперимент жүргізілді.

Эксперимент барысында әртүрлі дереккөздерден алынған мәтіндер өңделді. Негізгі қадамдар:

- Мәтіндерді лемматизациялау және токенизациялау – қазақ тілінің морфологиялық ерекшеліктеріне сәйкес арнайы өңдеу әдістері қолданылды.

- Диалектілерді сәйкестендіру – әртүрлі стильдер мен аймақтық ерекшеліктерді ескеру үшін корпусқа әртүрлі көздерден алынған мәтіндер қосылды.

- Деректерді аннотациялау – POS-тегтеу, синтаксистік құрылымдарды белгілеу және семантикалық аннотациялар енгізілді.

Модельді баптау (Fine-tuning)

Зерттеуде mBERT және GPT-3 модельдері қазақ тіліндегі деректер жиынтығында нақтылап бапталды. Баптау кезеңдері:

1. Алдын ала оқытылған модельді пайдалану – mBERT және BLOOM көптілді модельдері қазақ тіліндегі мәтіндерге бейімделді.

2. Қазақ тіліндегі корпус негізінде қайта оқыту – арнайы жинақталған dataset қолданылды.

3. Тілдік ерекшеліктерге бейімдеу – аффиксация, септеу және морфологиялық өзгешеліктер ескерілді.

Зерттеу нәтижелері мен талқылау.

Модельдердің тиімділігі бірнеше метрика бойынша бағаланды:

- Perplexity (PPL) – тілдік модельдің қаншалықты сәйкес келетінін өлшеу.
- BLEU score – аударма және генерацияланған мәтін сапасын бағалау.
- F1-score – атау есімдерді тану (NER) және морфологиялық талдау дәлдігін өлшеу.

Зерттеу барысында қазақ тіліне арналған LLM әртүрлі үлгілері талданды. Пысықтау нәтижелері төмендегі №3 кестеде келтірілген.

№3 кесте - Нәтижелер

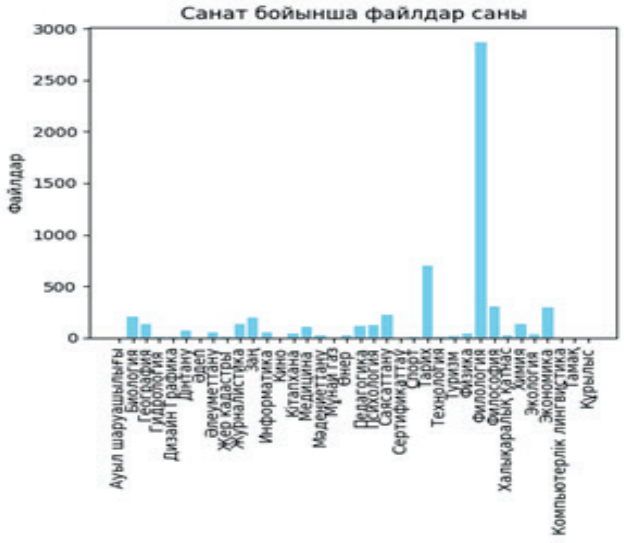
Модель	Perplexity ↓	BLEU ↑	F1-score ↑
mBERT	25.3	32.7	78.4
GPT-3	19.8	45.2	85.1
BLOOM	22.1	38.5	81.3
LLaMA	18.5	47.1	86.4
Mistral	17.3	50.5	88.2

Қазақ тіліне арналған GPT-3 және LLaMA модельдерін дәл баптау процесін қолдана отырып жетілдіру кезінде бірнеше тест мәтіндері қолданылды. Мысалы, bert моделі бастапқыда "Бүгінгі ауа-райы қандай? деген сұраққа "Мен бүгін ауа райы деректерін көрсете алмаймын" деп жауап берді. Ал, Mistral моделі "Бүгін Алматыда күн ашық, ауа температурасы +10°c" дәлірек жауап берді. Бұл қазақ тіліндегі мәтіндердің мәнмәтінін түсінудегі соңғы үлгілердің артықшылықтарын көрсетеді.

Корпус деректерін визуалды талдау. Қазақ тіліне арналған үлкен тілдік корпусты әзірлеу кезінде деректердің жалпы көлемі бойынша және оларды тақырыптық санаттар бойынша бөлу қажет. Дереккөздердің әртүрлілігі және тақырыптар бойынша деректердің тепе-теңдігі қорытынды тіл моделінің сапасына тікелей әсер етеді. Бұл жұмыста жиналған корпусқа талдау жасалды және оны алынған мәліметтер негізінде датасеттің құрылымын бейнелейтін үш диаграмма арқылы көрсетеміз.

2- суреттегі гистограммада әртүрлі тақырыптық санаттар бойынша файлдар санын көрсетміз. Бірқатар санаттарда мәтіндердің едәуір саны бар екендігі байқалады, ал басқаларында деректер айтарлықтай аз. Мысалы, филология, тарих, экономика санаттары ең жоғары көрсеткіштерге ие, бұл осы салалардағы мәтіндердің жоғары қолжетімділігін көрсетеді. Сонымен қатар, компьютерлік лингвистика, құрылыс, кино санаттарында әлдеқайда аз деректер бар, бұл осы салалардағы мәтіндерді өңдеу кезінде модель сапасының жеткіліксіздігіне әкелуі мүмкін.

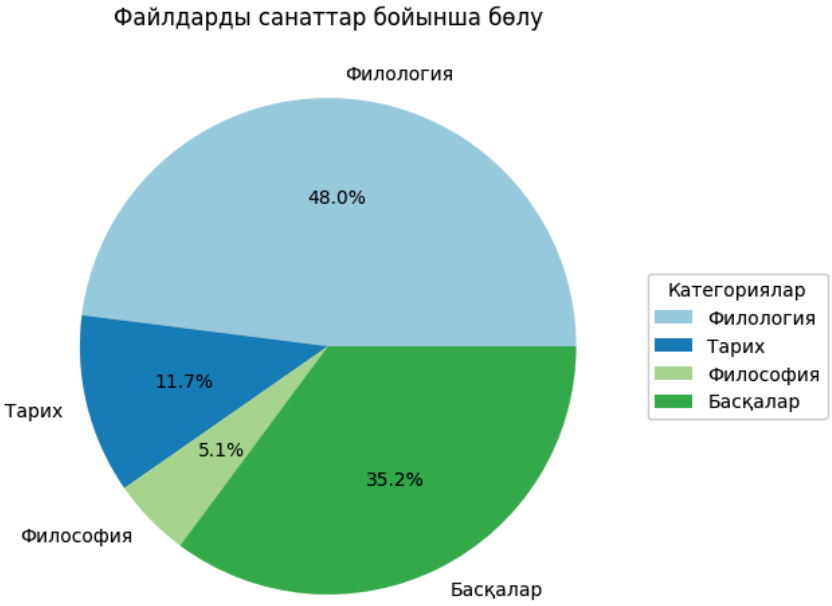
2-сурет. Деректерді тарату гистограммасы



Бұл талдау корпусты теңдестіру үшін бірқатар бағыттар бойынша қосымша мәліметтер жинау қажеттілігі туралы қорытынды жасауға мүмкіндік береді.

3-суреттегі дөңгелек диаграмма тақырыптық санаттар бойынша файлдардың пайыздық таралуын көрсетеді. Визуализацияның бұл түрі корпуста қандай тақырыптар басым екенін және деректердің қаншалықты біркелкі бөлінгенін бағалауға мүмкіндік береді.

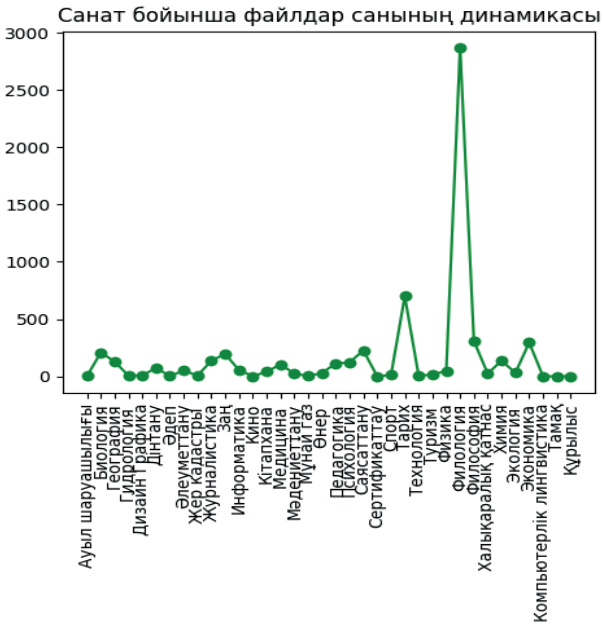
3-сурет. Корпус құрылымының дөңгелек диаграммасы



Диаграммада көрсетілгендей, филология, тарих, экономика корпустың көп бөлігін алады, ал қалған санаттар аз ғана пайызды құрайды. Мұндай теңгерімсіздік тілдік модельдің сапасына әсер етуі мүмкін, өйткені тек тақырыптар басым болатын корпуста оқытылған Машиналық оқыту алгоритмдері аз ұсынылған салалардағы мәтіндерді өңдеу кезінде нашар өнімділікті көрсетуі мүмкін. Оңтайлы шешім қосымша деректерді мақсатты түрде жинау, сондай-ақ санаттар бойынша файлдар санын қалыпқа келтіру арқылы корпусты теңестіру болуы мүмкін.

4-суретте сызықтық график санаттар бойынша файлдар санының өзгеруін көрсетеді. Бұл корпус құрылымындағы трендтерді анықтауға және әртүрлі бағыттар бойынша деректерді жинау қарқынын нақты салыстыруға мүмкіндік береді. Деректер көлемі біркелкі бөлінбегенін көруге болады, бұл корпусты теңестіру бойынша қосымша жұмыс істеу қажеттілігін тағы бір рет растайды.

4-сурет. Деректер динамикасының сызықтық графигі



Корпустың визуалды деректерін талдау бірнеше негізгі қорытынды жасауға мүмкіндік береді:

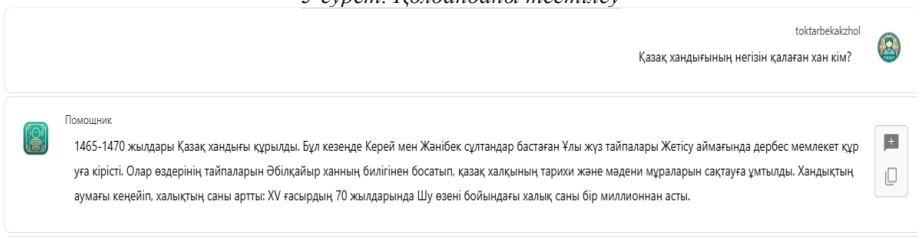
1. Деректер теңгерімсіздігі- белгілі бір санаттарда басқаларға қарағанда едәуір көп файлдар бар, бұл модельдің басым тақырыптарға ауысуына әкелуі мүмкін.

2. Корпусты кеңейту қажеттілігі — ең аз мәтіндері бар Санаттар датасеттің жалпы өкілдігін жақсарту үшін қосымша толтыруды қажет етеді.

3. Деректер құрылымын оңтайландыру- қайта ұсынылған санаттардағы қайталанатын деректерді сүзу немесе жетіспейтін тақырыптардағы деректерді жасанды түрде көбейту сияқты теңдестіру әдістері мүмкін.

5-суретте модель құрылған датасет негізінде сұраққа жауап берді:

5-сурет. Қолданбаны тестілеу



«Қазақ хандығының негізін қалаған хан кім?» сұрағына AI Assistant берген жауабы өте мазмұнды әрі құрылымды болды.

Бұл нәтижелер үлкен тілдік корпусты құруда маңызды рөл атқарады және оны одан әрі жақсартудың негізгі бағыттарын анықтауға көмектеседі. Санаттар арасында деректерді біркелкі бөлуге қол жеткізу әртүрлі тақырыптық салалардағы мәтіндерді тиімді өңдеуге қабілетті жоғары сапалы және әмбебап тілдік модельді қамтамасыз етеді.

Қорытынды. Қазақ тілі үшін үлкен тіл моделін (LLM) құру осы тілдегі мәтіндермен тиімді жұмыс істеуге қабілетті жасанды интеллектті дамытудағы маңызды қадам болып табылады. Қазақ тілінің морфологиялық, синтаксистік және диалектілік ерекшеліктерін зерделеу, сондай-ақ әртүрлі алфавиттерді есепке алу табысты модельді әзірлеу кезінде шешуші болып табылады. Бұл жолдағы басты проблемалардың бірі Қазақ тілі үшін тілдік деректердің шектелуі болып табылады, бұл LLM-ді оқытуды қиындатады және олардың өнімділігін төмендетеді.

Осы жоба аясында dataset жиналды, оның мақсаты — қазақ тілді модельді оқыту үшін деректерді жинау және құрылымдау. OCR сияқты заманауи технологияларды пайдалану әртүрлі көздерден материалдарды цифрландыруға және өңдеуге оңай пішімдерге түрлендіруге мүмкіндік береді, бұл қол жетімді деректер көлемін айтарлықтай кеңейтеді. AI Assistant сұрақтарға орынды және толық жауап беру үшін KazLLM-ге арналған датасетті әлі де толықтыру жұмыстарын жүргізу керек.

Сонымен қатар, қазақ тілі үшін сапалы және алуан түрлі dataset құру жасанды интеллектті дамыту, оны мемлекеттік басқару, білім беру, медицина және бизнес сияқты салаларда қолдану үшін мүмкіндіктер ашады. Толыққанды және аннотацияланған деректердің болуы қазақ тілімен адам түсінігіне жақын деңгейде жұмыс істей алатын дәл және тиімді модельдерді әзірлеуге мүмкіндік береді.

Осылайша, осы жоба шеңберінде атқарылған жұмыс ақпараттық теңсіздікті жоюға ықпал етеді және оның жасанды интеллекттің жаһандық әзірлемелеріне толыққанды қатысуын қамтамасыз ете отырып, қазақ тілін технологиялық прогресте қолдауды қамтамасыз етеді.

References

Bender E.M. et al. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. P. 610–623. Available at: https://dl.acm.org/doi/10.1145/3442188.3445922 (in Eng.)

Brown T. et al. Language Models are Few-Shot Learners [Electronic resource]. arXiv:2005.14165. Available at: https://arxiv.org/abs/2005.14165 (in Eng.)

Common Crawl [Electronic resource]. — Access mode: https://commoncrawl.org (date accessed: 08.04.2025) (in Eng.)

Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resource]. — arXiv:1810.04805. – Access mode: https://arxiv.org/abs/1810.04805 (in Eng.)

EleutherAI [Electronic resource]. — Access mode: https://www.eleuther.ai (date accessed: 08.04.2025) (in Eng.)

Howard J., Ruder S. Universal Language Model Fine-tuning for Text Classification. Proceedings

of the 56th Annual Meeting of the Association for Computational Linguistics, 2018. — P. 328–339. — Access mode: https://aclanthology.org/P18-1031 (in Eng.)

Nikbakht S.R., Benzaghta M., Geraci G. TSpec-LLM: An Open-source Dataset for LLM Understanding of 3GPP Specifications [Electronic resource]. – arXiv:2406.01768. — Access mode: https://arxiv.org/abs/2406.01768 (in Eng.)

OpenAI. GPT-4 Technical Report [Electronic resource]. – Access mode: https://openai.com/research/gpt-4 (in Eng.)

OpenCorpora [Electronic resource]. — Access mode: https://opencorpora.org (date of access: 08.04.2025) (in Eng.)

PubMed [Electronic resource]. – Access mode: https://pubmed.ncbi.nlm.nih.gov (date accessed: 08.04.2025) (in Eng.)

QazCorpus – Kazakh Language Corpus, version 2 [Electronic resource]. – Access mode: https://v2.qazcorpus.kz (date accessed: 08.04.2025) (in Eng.)

The Pile – GitHub repository [Electronic resource]. — Access mode: https://github.com/EleutherAI/the-pile (date accessed: 08.04.2025) (in Eng.)

Toloka.ai [Electronic resource]. — Access mode: https://toloka.ai (access date: 04/08/2025) (in Eng.)

Toiganbayeva N, Kasem M., Abdimanap G, Bostanbekov K., Abdelrahman A., Alimova A., Nurseitov D. Toiganbayeva, N. A. et al. (2022) "KOHTD: Kazakh Offline Handwritten Text Dataset.". Signal Processing: Image Communication. — Volume 108. https://www.sciencedirect.com/science/article/pii/S0923596522001217 (in Eng.)

Yeshpanov R., Khassanov Y., Varol H. A. KazNERD: Kazakh Named Entity Recognition Dataset. Proceedings of the Language Resources and Evaluation Conference (LREC 2022). — P. 406–414. — Access mode: https://aclanthology.org/2022.lrec-1.44/ (in Eng.)

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Ж.Ш.Әден*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 25.09.2025.
Формат 60x881/8. Бумага офсетная.
Печать –ризограф. 20,0 п.л. Заказ 3.

Национальная академия наук РК
050000, Алматы, ул. Кунаева 132