

ISSN 2518-1726 (Online),

ISSN 1991-346X (Print)

**ACADEMIC SCIENTIFIC
JOURNAL OF COMPUTER SCIENCE**

**№2
2026**

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



ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

3 (359)

JULY – SEPTEMBER 2026

**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, (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>

USATOVA Olga Alexandrovna, PhD, Associate Professor, Chief Scientific Secretary of the Institute of Information and Computing Technologies of the National Academy of Sciences of the Republic of Kazakhstan (Almaty, Kazakhstan), <https://www.scopus.com/authid/detail.uri?authorId=57204581062>, <https://www.webofscience.com/wos/author/record/JCO-3058-2023>

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, 2026

БАС РЕДАКТОР:

МУТАНОВ Ғалымқайыр Мұтанұлы, техника ғылымдарының докторы, профессор, ҚР ҰҒА академигі, (Алматы, Қазақстан), <https://www.scopus.com/authid/detail.uri?authorId=6506682964>, <https://www.webofscience.com/wos/author/record/1423665>

РЕДАКЦИЯ АЛҚАСЫ:

КАЛИМОЛДАЕВ 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 Жұмаіл 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>

УСАТОВА Ольга Александровна, PhD, қауымдастырылған профессор, ҚР ҒЖБМ "Ақпараттық және есептеу технологиялары институтының" бас ғалым хатшысы (Алматы, Қазақстан), <https://www.scopus.com/authid/detail.uri?authorId=57204581062>, <https://www.webofscience.com/wos/author/record/JCO-3058-2023>

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, 2026

Главный редактор:

МУТАНОВ Галимканр Мутанович, доктор технических наук, профессор, академик НАН РК, (Алматы, Казахстан), <https://www.scopus.com/author/detail.uri?authorId=6506682964>, <https://www.webofscience.com/wos/author/record/1423665>

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/author/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/author/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/author/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/author/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/author/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/author/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/author/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/author/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/author/detail.uri?authorId=55453992600>, <https://www.webofscience.com/wos/author/record/3802041>

УСАТОВА Ольга Александровна, PhD, ассоциированный профессор, Главный ученый секретарь «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), <https://www.scopus.com/author/detail.uri?authorId=57204581062>, <https://www.webofscience.com/wos/author/record/JCO-3058-2023>

КАПАЛОВА Нурсулу Алдажаровна, кандидат технических наук, заведующий лабораторией кибербезопасности РГП «Института информационных и вычислительных технологий» КН МНВО РК (Алматы, Казахстан), <https://www.scopus.com/author/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/author/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/author/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/author/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 КОКШВО 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», 2026

CONTENTS

Akhmetova A.Zh., Kantureyeva M.A., Serimbetov B.A., Zhasuzakova M.Zh., Mustafayeva N.T. Identification of communities and active users in social networks.....	11
Akzhigitov D., Mukhtarkhanov M., Zhetpisbayeva A., Kulakayeva A., Zhetpisbayev K. Analysis of the mechanical structure of a 12U CubeSat nanosatellite and assessment of its mechanical stability.....	29
Alibekov A., Nurimanov A., Mukhametkali A., Kassym D., Varol H.A. Qolda-Math: a small vision-language model for mathematical reasoning in Kazakh.....	50
Alisheva S.S., Ospanova A.O., Shaimerdenova G.S., Shaimerdenova L., Zhidebayeva A.N. Detecting Kazakh language offensive comments in online content using deep learning algorithms.....	68
Askar G., Mutaikhan A., Aitkozhayeva M., Nauryzbaikyzy G., Makhataeva Z. Retrieval beats fine-tuning: adapting a Kazakh small language model for historical question answering.....	86
Beissenova G.I., Shynzhigit B.B., Beisebayeva A.Zh., Kaldarova B.S., Lesbayev A.U. Deep neural network-based speaker authentication for secure access control systems.....	106
Bekenuly B., Baimakhanova A., Baiturganova V., Bolatbekova D. Comparison of CNN architectures for automated ACL injury classification from knee MRI.....	123
Bekbolat A., Razak A., Alimbayeva Z., Ozhikenova A. Multi-task deep learning for knee MRI pathology classification.....	139
Issakhov A., Akylbekkyzy A., Abylkassymova A. Predicting the spread of airborne viral infections in enclosed spaces.....	158
Kisseleva O., Abdramanov K., Naumenko V., Mukanova Z., Maintser D. Evaluation of serverless cold start optimization techniques.....	177
Koshkarov Ye.Zh., Amirgaliyev B.Ye., Mulesa O. Physics-aware masked pretraining for cross-basin well log foundation models.....	197
Kurmetkan T., Turysbek Zh., Mukhitova A., Nauryzbaeva A. Research on methods for reducing knowledge conflicts in large language models in Kazakh question-and-answer systems.....	214
Makatov Ye.K., Abdul Razaque, Ismukanova A.N., Salykova A.K., Glock E.S. Machine learning-based model for phishing URL detection in online social networks.....	233
Mimenbayeva A.B., Naizagarayeva A.A., Ospanova T.T., Koikelova D.K., Kudabayeva R.N. Spatial and climate data analysis for machine learning-based crop yield prediction in Northern Kazakhstan.....	255
Mukhamediev R.I., Abdoldina F.N., Gertsen Y.A. Economic opportunities and constraints of artificial intelligence adoption in Kazakhstan's resource-oriented economy.....	274

Mutanov G.M., Karyukin V.I., Ussatova O.A., Begimbayeva Y.Y., Baispay G.B. The assessment of blockchain readiness for academic document verification.....	289
Mussiraliyeva Sh., Bolatbek M., Baisylbayeva K. Classification of cyber-extremist ideologies in the Kazakh language considering multi-annotator disagreement.....	306
Omarov B.S., Sultan D.R., Rakhymberdiev D.S., Abibullayeva A.A., Toktarova A.B. Deformable attention-guided Transformer for brain tumor segmentation.....	326
Oralbekova D., Bekarystankyzy A., Yerimbetova A., Berzhanova U. Development of a streaming Kazakh speech recognition system using end-to-end models.....	346
Serek A., Shoiynbek A., Zheludkov M., Kuanyshbay D., Meraliyev B. Comparative analysis of machine learning methods for detecting fraud in text messages.....	365
Smakhanova A.K., Uskenbayeva R.K., Kassymova A.B., Kalpeyeva Zh.B., Konyrbaev N.B. Experimental research using models and methods of the university digital ecosystem.....	384
Sultanali A.E., Kazbekova G.N. Machine learning-based detection of AI-generated content in students' written works.....	401
Tashenova Zh., Yergazin A., Nurlybaeva E., Abdugulova Zh., Amanzholova Sh. Experimental evaluation of a hybrid IDPS with adaptive learning for low-latency 5G networks using edge computing.....	418
Telman D., Yerimbetova A., Daiyrbayeva E. Leakage-resistant fake news detection for Kazakh: temporal evaluation of lexical, Transformer and hybrid models.....	441
Zhanegizov A., Mamyrbayev O., Abdigappar M. Application of YOLO12 for road defect detection in GPR images.....	457

МАЗМҰНЫ

Ахметова А.Ж., Кантуреева М.А., Серимбетов Б.А., Жасұзақова М.Ж., Мұстафаева Н.Т. Әлеуметтік желілердегі қауымдастықтар мен белсенді қолданушыларды анықтау.....	11
Ақжігітов Д.Ж., Мұхтарханов М.О., Жетписбаева А.Т., Кулакаева А.Е., Жетписбаев К.У. 12U CubeSat үлгісіндегі наноспутниктің механикалық құрылымын талдау және оның механикалық орнықтылығын бағалау.....	29
Алибеков А., Нуриманов А., Мұхаметқали А., Қасым Д., Варол Х.А. Qolda-Math: қазақ тіліндегі математикалық пайымдауға арналған кіші көру-тілдік модель.....	50
Алишева С.С., Оспанова А.О., Шаймерденова Г.С., Шаймерденова Л., Жидебаева А.Н. Онлайн-контенттегі қазақ тіліндегі бейәдеп пікірлерді терең оқыту алгоритмдері арқылы анықтау.....	68
Асқар Ғ., Мұтайхан А., Айтқожаева М., Наурызбайқызы Г., Махатаева Ж. Retrieval fine-tuning-тен басым: тарихи сұрақтарға жауап беру үшін қазақша шағын тілдік модельді бейімдеу.....	86
Бейсенова Г.И., Шынжігіт Б.Б., Бейсебаева А.Ж., Қалдарова Б.С., Лесбаев А.У. Терең нейрондық желіге негізделген қауіпсіз басқару жүйелерінде сөйлеушіні аутентификациялау.....	106
Бекенұлы Б., Баймаханова А., Байтурганова В., Болатбекова Д. Алдыңғы крест тәрізді байлам жарақатын анықтауға арналған CNN архитектураларын салыстыру.....	123
Бекболат А., Разак А., Алимбаева Ж., Ожикенова А. Tize MPT патологиясын жіктеуге арналған көп тапсырмалы терең оқыту.....	139
Исахов А.А., Ақылбекқызы А., Абылкасымова А. Шектеулі кеңістіктерде әуе тамшыларымен вирустық инфекцияның таралуын сандық модельдеу.....	158
Киселева О., Абдраманов Қ., Науменко В., Мұқанова Ж., Майнцер Д. Serverless жүйелеріндегі суық іске қосуды оңтайландыру әдістерін эксперименттік бағалау.....	177
Кошкарров Е.Ж., Амиргалиев Б.Е., Мүлеса О. Бассейнаралық ұңғымалық диаграммалардың іргелі модельдерін физикалық тұрғыдан негізделген масқаланған алдын ала оқыту.....	197
Құрметқан Т., Тұрысбек Ж., Мұхитова А., Наурызбаева А. Қазақ тіліндегі сұрақ-жауап жүйелерінде үлкен тілдік модельдердегі білім қайшылықтарын азайту әдістерін зерттеу.....	214
Мақатов Е.К., Abdul Razaque, Исмуканова А.Н., Салыкова А.К., Глюк Е.С. Онлайн әлеуметтік желілердегі фишингтік URL-дарды анықтауға арналған машиналық оқыту негізіндегі модель.....	233

Мименбаева А.Б., Найзағараева А.А., Оспанова Т.Т., Қойкелова Д.К., Қудабаева Р.Н. Солтүстік Қазақстандағы дақылдар өнімділігін машиналық оқыту арқылы болжауда кеңістіктік және климаттық деректерді талдау.....	255
Мухамедиев Р.И., Абдолдина Ф.Н., Герцен Е.А. Қазақстанның ресурстық бағытталған экономикасына жасанды интеллектті енгізудің экономикалық мүмкіндіктері мен шектеулері.....	274
Мутанов Г.М., Карюкин В.И., Усатова О.А., Бегимбаева Е.Е., Байспай Г.Б. Білім беру құжаттарының түпнұсқалығын тексеруге блокчейн технологиясының дайындығын бағалау.....	289
Мусиралиева Ш., Болатбек М., Байсылбаева К. Көпаннынотаторлы келіспеушілікті ескере отырып қазақ тіліндегі киберэкстремистік идеологияларды жіктеу.....	306
Омаров Б.С., Сұлтан Д.Р., Рахымбердиев Д.С., Абибуллаева А.А., Тоқтарова А.Б. Ми ісіктерін сегменттеуде деформацияланатын зейін механизміне негізделген мультимодальды трансформер.....	326
Оралбекова Д.О., Бекарыстанқызы А., Еримбетова А.С., Бержанова У.Г. Интегралды модельдерді қолдану арқылы қазақ тіліндегі ағындық сөйлеуді тану жүйесін әзірлеу.....	346
Серек А., Шойынбек А., Желудков М., Қуанышбай Д., Мералиев Б. Мәтіндік хабарламалардағы алаяқтықты анықтауға арналған машиналық оқыту әдістерінің салыстырмалы талдауы.....	365
Смаханова А.К., Ускенбаева Р.К., Касымова А.Б., Кальпеева Ж.Б., Қоңырбаев Н.Б. Университеттің цифрлық экожүйесінің модельдері мен әдістерін қолдана отырып жүргізілген эксперименттік зерттеу.....	384
Сұлтанәлі А.Е., Казбекова Г.Н. Білім алушылардың жазбаша жұмыстарындағы AI-контентті машиналық оқыту арқылы анықтау.....	401
Ташенова Ж.М., Ергазин Ә., Нұрлыбаева Ә.Н., Абдуғулова Ж.К., Аманжолова Ш.А. Edge Computing көмегімен төмен кідірісті 5G желілеріне арналған бейімделгіш оқытуы бар гибриді IDPS-ті эксперименттік бағалау.....	418
Тельман Д., Еримбетова А., Дайырбаева Ә. Қазақ тіліндегі жалған жаңалықтарды анықтау: лексикалық, трансформерлік және гибридті модельдерді уақыт бойынша бағалау.....	441
Жанегизов А.С., Мамырбаев Ә.Ж., Абдіғаппар М.Д. YOLO12 моделін георадар кескіндеріне қолдану.....	457

СОДЕРЖАНИЕ

Ахметова А.Ж., Кантуреева М.А., Серимбетов Б.А., Жасузакова М.Ж., Мустафаева Н.Т. Выявление сообществ и активных пользователей в социальных сетях.....	11
Акжигитов Д.Ж., Мухтарханов М.О., Жетписбаева А.Т., Кулакаева А.Е., Жетписбаев К.У. Анализ механической конструкции наноспутника модели 12U CubeSat и оценка его механической устойчивости.....	29
Алибеков А., Нуриманов А., Мухаметкали А., Касым Д., Варол Х.А. Qolda-Math: маленькая мультимодальная модель зрения и языка для математических рассуждений на казахском языке.....	50
Алишева С.С., Оспанова А.О., Шаймерденова Г.С., Шаймерденова Л., Жидебаева А.Н. Выявление оскорбительных комментариев на казахском языке в онлайн-контенте с помощью алгоритмов глубокого обучения.....	68
Аскар Г., Мутайхан А., Айткожаева М., Наурызбайкызы Г., Махатаева Ж. Retrieval эффективнее fine-tuning: адаптация малой языковой модели казахского языка для ответов на исторические вопросы.....	86
Бейсенова Г.И., Шынжигит Б.Б., Бейсебаева А.Ж., Калдарова Б.С., Лесбаев А.У. Аутентификация личности по голосу на основе глубокой нейронной сети для систем безопасного управления доступом.....	106
Бекенулы Б., Баймаханова А., Байтурганова В., Болатбекова Д. Сравнение архитектур сверточных нейронных сетей для автоматической классификации разрыва ПКС по МРТ коленного сустава.....	123
Бекболат А., Разак А., Алимбаева Ж., Ожикенова А. Многозадачное глубокое обучение для классификации патологий колена по МРТ.....	139
Исахов А.А., Акылбеккызы А., Абылкасымова А. Прогнозирование распространения вирусной инфекции воздушно-капельным путем в замкнутых помещениях.....	158
Киселева О., Абдраманов К., Науменко В., Муканова Ж., Майнцер Д. Экспериментальная оценка методов оптимизации холодного запуска в serverless-системах.....	177
Кошкарар Е.Ж., Амиргалиев Б.Е., Мулеса О. Физически-осознанное маскированное предобучение для межбассейновых фундаментальных моделей каротажных диаграмм.....	197
Курметкан Т., Турысбек Ж., Мухитова А., Наурызбаева А. Исследование методов снижения конфликтов знаний в больших языковых моделях казахских вопросно-ответных систем.....	214
Макатов Е.К., Abdul Razaque, Исмуканова А.Н., Салыкова А.К., Глок Е.С. Модель обнаружения фишинговых URL в онлайн-социальных сетях на основе машинного обучения.....	233

Мименбаева А.Б., Найзагараева А.А., Оспанова Т.Т., Койкелова Д.К., Кудабаяева Р.Н. Анализ климатических данных для прогнозирования урожайности пшеницы методами машинного обучения в Северном Казахстане.....	255
Мухамедиев Р.И., Абдолдина Ф.Н., Герцен Е.А. Экономические возможности и ограничения внедрения искусственного интеллекта в ресурсно-ориентированной экономике Казахстана.....	274
Мутанов Г.М., Карюкин В.И., Усатова О.А., Бегимбаева Е.Е., Байспай Г.Б. Оценка готовности технологии блокчейн к проверке подлинности документов об образовании.....	289
Мусиралиева Ш., Болатбек М., Байсылбаева К. Классификация киберэкстремистских идеологий на казахском языке с учетом разногласий нескольких аннотаторов.....	306
Омаров Б.С., Султан Д.Р., Рахымбердиев Д.С., Абибуллаева А.А., Токтарова А.Б. Мультимодальный трансформер на основе механизма деформируемого внимания для сегментации опухолей головного мозга.....	326
Оралбекова Д.О., Бекарыстанкызы А., Еримбетова А.С., Бержанова У.Г. Распознавание казахской речи с применением интегральных моделей.....	346
Серек А., Шойынбек А., Желудков М., Куанышбай Д., Мералиев Б. Сравнительный анализ методов машинного обучения для выявления мошенничества в текстовых сообщениях.....	365
Смаханова А.К., Ускенбаева Р.К., Касымова А.Б., Кальпеева Ж.Б., Конырбаев Н.Б. Экспериментальное исследование с использованием моделей и методов цифровой экосистемы университета.....	384
Султанали А.Е., Казбекова Г.Н. Обнаружение AI-контента в письменных работах обучающихся методами машинного обучения.....	401
Ташенова Ж.М., Ергазин А., Нурлыбаева Э.Н., Абдугулова Ж.К., Аманжолова Ш.А. Экспериментальная оценка гибридной системы обнаружения и предотвращения вторжений (IDPS) с адаптивным обучением для сетей 5G с низкой задержкой с использованием граничных вычислений.....	418
Тельман Д., Еримбетова А., Дайырбаева Э. Обнаружение фейковых новостей на казахском языке: временная оценка лексических, трансформерных и гибридных моделей.....	441
Жанегизов А.С., Мамырбаев О.Ж., Абдигаппар М.Д. Применение модели YOLOv12 для анализа данных георадара.....	457

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

ISSN 1991-346X

Volume 3.

Number 359 (2026). 86–105

<https://doi.org/10.32014/2026.2518-1726.453>

IRSTI 28.23.25

UDC 004.891.2

© Askar G.^{1,2*}, Mutaikhan A.¹, Aitkozhaeva M.¹, Nauryzbaikyzy G.^{1,2,3},
Makhataeva Z.^{1,2}, 2026.

¹ “Institute of Smart Systems and Artificial Intelligence” Private Institution,
Astana, Kazakhstan;

² Al-Farabi Kazakh National University, Almaty, Kazakhstan;

³ K. Zhubanov Aktobe Regional University, Aktobe, Kazakhstan.

E-mail: galammadin.askar@nu.edu.kz

RETRIEVAL BEATS FINE-TUNING: ADAPTING A KAZAKH SMALL LANGUAGE MODEL FOR HISTORICAL QUESTION ANSWERING

Askar Galammadin — Research Assistant, Institute of Smart Systems and Artificial Intelligence (ISSAI), Astana, Kazakhstan; Research Fellow, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

E-mail: galammadin.askar@nu.edu.kz, <https://orcid.org/0009-0001-6304-3003>;

Mutaikhan Alumbdi — Research Assistant, Institute of Smart Systems and Artificial Intelligence (ISSAI), Astana, Kazakhstan,

E-mail: alumbdi.mutaikhan@nu.edu.kz, <https://orcid.org/0009-0006-7093-074X>;

Aitkozhaeva Malika — Research Assistant, Institute of Smart Systems and Artificial Intelligence (ISSAI), Astana, Kazakhstan,

E-mail: malika.aitkozhaeva@nu.edu.kz, <https://orcid.org/0009-0003-5991-4982>;

Nauryzbaikyzy Guldana — PhD, Lecturer at Department of Foreign Languages, K. Zhubanov Aktobe Regional University, Aktobe, Kazakhstan; Research Assistant, Institute of Smart Systems and Artificial Intelligence (ISSAI), Astana, Kazakhstan; Research Fellow, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

E-mail: gnauryzbaikyzy@zhubanov.edu.kz, <https://orcid.org/0000-0001-9272-8952>;

Makhataeva Zhanat — PhD, Senior Data Scientist, Institute of Smart Systems and Artificial Intelligence (ISSAI), Astana, Kazakhstan; Research Fellow, Al-Farabi Kazakh National University, Almaty, Kazakhstan,

E-mail: zhanat.makhataeva@nu.edu.kz, <https://orcid.org/0000-0001-9366-7047>.

Abstract. Large language models (LLMs) have shown a strong ability to solve a wide range of language tasks. However, they became less practical for subject-specific applications, especially when dealing with low-resource languages. This paper shows an example of adaptation of small language models (SLMs) to solve subject-specific language tasks in the history field. As a base model, we used the small vision language model (VLM) Qolda, developed by the Institute

of Smart Systems and Artificial Intelligence (ISSAI) at Nazarbayev University. In the paper, 4 architectures are compared: 1) base model Qolda; 2) Qolda model with supervised fine-tuning (SFT) created using 48 000 instruction pairs retrieved from Wikipedia; 3) Qolda with retrieval-augmented generation (RAG) framework based on 29 digital history textbooks with 2.25 million indexed tokens; 4) Qolda combined with both SFT and RAG. To establish whether the research findings follow the same patterns across other Kazakh language models, we apply our evaluations to two external larger baseline models: 1) KazLLM and 2) SherKala-Chat. All architectures are evaluated on 1,081 multiple-choice questions (MCQs) from the Unified National Test (UNT) in the history of Kazakhstan and on a 500-query open-ended test set. Results show that Qolda+RAG achieves $74.93 \pm 0.43\%$ MCQ accuracy and a mean quality score of 3.34 ± 1.64 . In these settings, SFT acts as a response formatter, reducing the average response length from 51.8 to 14.64 words. Qolda-SFT+RAG completes inference in roughly half the time of Qolda+RAG with comparable quality of 3.10 ± 1.42 . The study findings provide a foundation for the design and development of future Kazakh-language-supported AI agents across broad application domains. The project implementation, datasets, and evaluation methodology are available on the project's GitHub page with the corresponding Hugging Face datasets: <https://github.com/IS2AI/history-qolda-ui>.

Keywords: small language models (SLM), retrieval-augmented generation (RAG), supervised fine-tuning (SFT), low-resource languages, history of Kazakhstan, natural language processing (NLP)

***Acknowledgments.** We extend our gratitude to the supervisor of the research grant, Professor Madina Mansurova, for her coordination and support.*

***Funding.** This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993001).*

For citations: Askar G., Mutaikhan A., Aitkozhayeva M., Nauryzbaikyzy G., Makhataeva Z. Retrieval beats fine-tuning: adapting a kazakh small language model for historical question answering. Academic Scientific Journal of Computer Science, 2026. — No.3. — P. 86–105. DOI: <https://doi.org/10.32014/2026.2518-1726.453>

© Асқар Ғ.^{1,2*}, Мұтайхан А.¹, Айтқожаева М.¹,
Наурызбайқызы Ғ.^{1,2,3}, Махатаева Ж.^{1,2}, 2026.

¹ Ақылды жүйелер және жасанды интеллект институты (ISSAI),
Астана, Қазақстан;

² Өл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;

³ Қ. Жұбанов атындағы Ақтөбе өңірлік университеті, Ақтөбе, Қазақстан.
E-mail: galammadin.askar@nu.edu.kz

RETRIEVAL FINE-TUNING-ТЕН БАСЫМ: ТАРИХИ СҰРАҚТАРҒА ЖАУАП БЕРУ ҮШІН ҚАЗАҚША ШАҒЫН ТІЛДІК МОДЕЛЬДІ БЕЙІМДЕУ.

Асқар Ғаламмадин — ғылыми қызметкер, Ақылды жүйелер және жасанды интеллект институты (ISSAI), Астана, Қазақстан; ғылыми қызметкер, Өл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

E-mail: galammadin.askar@nu.edu.kz, <https://orcid.org/0009-0001-6304-3003>;

Мұтайхан Алмубди — ғылыми қызметкер, Ақылды жүйелер және жасанды интеллект институты (ISSAI), Астана, Қазақстан,

E-mail: almubdi.mutaikhhan@nu.edu.kz, <https://orcid.org/0009-0006-7093-074X>;

Айтқожаева Малика — ғылыми қызметкер, Ақылды жүйелер және жасанды интеллект институты (ISSAI), Астана, Қазақстан,

E-mail: malika.aitkozhaeva@nu.edu.kz, <https://orcid.org/0009-0003-5991-4982>;

Наурызбайқызы Гүлдана — PhD, Қ. Жұбанов атындағы Ақтөбе өңірлік университетінің Шетел тілдері кафедрасының оқытушысы, Ақтөбе, Қазақстан; ғылыми қызметкер, Ақылды жүйелер және жасанды интеллект институты (ISSAI), Астана, Қазақстан; ғылыми қызметкер, Өл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

E-mail: gnauryzbaikyzy@zhubanov.edu.kz, <https://orcid.org/0000-0001-9272-8952>;

Махатаева Жанат — PhD, аға деректер талдаушысы, Ақылды жүйелер және жасанды интеллект институты (ISSAI), Астана, Қазақстан; ғылыми қызметкер, Өл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан,

E-mail: zhanat.mahataeva@nu.edu.kz, <https://orcid.org/0000-0001-9366-7047>.

Аннотация. Үлкен тілдік модельдер (LLM) әртүрлі тілдік тапсырмаларды орындауда жоғары нәтижелер көрсеткенімен, оларды тар салалық міндеттерге, әсіресе ресурстары шектеулі тілдерге қолдану бірқатар қиындықтар туғызады. Бұл зерттеуде тарих саласындағы сұрақтарға жауап беру міндеттерін шешу үшін шағын тілдік модельдерді (SLM) бейімдеу тәсілі қарастырылды. Негізгі модель ретінде Ақылды жүйелер және жасанды интеллект институты (ISSAI) әзірлеген Qolda шағын көру-тілдік моделі пайдаланылды. Зерттеу барысында төрт архитектура салыстырылды: 1) базалық Qolda моделі; 2) Уикипедиядан алынған 48 000 нұсқаулық жұбы негізінде бақыланатын дооқыту (SFT) жүргізілген Qolda; 3) 2,25 миллион индекстелген токенді қамтитын 29 цифрландырылған тарих оқулығына негізделген іздеумен толықтырылған генерация (RAG) жүйесі бар Qolda; 4) SFT және RAG тәсілдерін біріктіретін Qolda. Анықталған заңдылықтардың басқа қазақ тіліндегі модельдерге де тән екенін бағалау үшін қосымша

KazLLM және SherKala-Chat ірі базалық модельдері зерттелді. Барлық архитектуралар Қазақстан тарихы бойынша Ұлттық бірыңғай тестілеуден алынған 1 081 көп таңдаулы сұрақ және 500 ашық сұрақтан тұратын тест жиынтығы негізінде бағаланды. Нәтижелер Qolda+RAG моделінің көп таңдаулы сұрақтар бойынша $74,93 \pm 0,43\%$ дәлдікке және $3,34 \pm 1,64$ орташа сапа көрсеткішіне қол жеткізгенін көрсетті. Бұл жағдайда SFT негізінен жауапты форматтау қызметін атқарып, жауаптың орташа ұзындығын 51,8 сөзден 14,64 сөзге дейін қысқартты. Qolda-SFT+RAG архитектурасы Qolda+RAG моделіне қарағанда есептеуді шамамен екі есе жылдам орындап, $3,10 \pm 1,42$ деңгейіндегі салыстырмалы сапаны сақтады. Алынған нәтижелер қазақ тілін қолдайтын жасанды интеллект агенттерін әртүрлі қолданбалы салаларда әзірлеу үшін әдіснамалық негіз қалыптастырады.

Түйін сөздер: шағын тілдік модельдер (SLM); іздеумен толықтырылған генерация (RAG); бақыланатын дооқыту (SFT); ресурстары шектеулі тілдер; Қазақстан тарихы; табиғи тілді өңдеу (NLP)

© Аскар Г.^{1,2*}, Мутайхан А.¹, Айтқожаева М.¹,
Наурызбайқызы Г.^{1,2,3}, Махатаева Ж.^{1,2}, 2026.

¹ Институт умных систем и искусственного интеллекта (ISSAI),
Астана, Казахстан;

² Казахский национальный университет имени аль-Фараби,
Алматы, Казахстан;

³ Актюбинский региональный университет имени К. Жубанова,
Актобе, Казахстан.

E-mail: galammadin.askar@nu.edu.kz

RETRIEVAL ЭФФЕКТИВНЕЕ FINE-TUNING: АДАПТАЦИЯ МАЛОЙ ЯЗЫКОВОЙ МОДЕЛИ КАЗАХСКОГО ЯЗЫКА ДЛЯ ОТВЕТОВ НА ИСТОРИЧЕСКИЕ ВОПРОСЫ.

Аскар Галаммадин — научный сотрудник, Институт умных систем и искусственного интеллекта (ISSAI), Астана, Казахстан; научный сотрудник, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,

E-mail: galammadin.askar@nu.edu.kz, <https://orcid.org/0009-0001-6304-3003>;

Мутайхан Алмубди — научный сотрудник, Институт умных систем и искусственного интеллекта (ISSAI), Астана, Казахстан,

E-mail: almubdi.mutaikhani@nu.edu.kz, <https://orcid.org/0009-0006-7093-074X>;

Айтқожаева Малика — научный сотрудник, Институт умных систем и искусственного интеллекта (ISSAI), Астана, Казахстан,

E-mail: malika.aitkozhaeva@nu.edu.kz, <https://orcid.org/0009-0003-5991-4982>;

Наурызбайқызы Гулдана — PhD, преподаватель кафедры иностранных языков Актюбинского регионального университета имени К. Жубанова, Актобе, Казахстан; научный сотрудник, Институт умных систем и искусственного интеллекта (ISSAI), Астана, Казахстан; научный сотрудник, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,

E-mail: gnauryzbaikyzy@zhubanov.edu.kz, <https://orcid.org/0000-0001-9272-8952>;

Махатаева Жанат — PhD, старший аналитик данных, Институт умных систем и искусственного интеллекта (ISSAI), Астана, Казахстан; научный сотрудник, Казахский национальный университет имени аль-Фараби, Алматы, Казахстан,
E-mail: zhanat.mahataeva@nu.edu.kz, <https://orcid.org/0000-0001-9366-7047>.

Аннотация. Большие языковые модели (LLM) демонстрируют высокую эффективность при решении широкого спектра языковых задач, однако их применение в предметно-ориентированных областях, особенно для языков с ограниченными ресурсами, остается затруднительным. В исследовании рассматривается подход к адаптации малых языковых моделей (SLM) для решения задач, связанных с ответами на вопросы по истории. В качестве базовой модели использована малая визуально-языковая модель Qolda, разработанная Институтом умных систем и искусственного интеллекта (ISSAI). Сравниваются четыре архитектуры: 1) базовая модель Qolda; 2) Qolda с дообучением с учителем (SFT) на основе 48 000 пар инструкций, полученных из Википедии; 3) Qolda с механизмом поисково-дополненной генерации (RAG), основанным на 29 оцифрованных учебниках по истории, содержащих 2,25 млн индексированных токенов; 4) Qolda, объединяющая SFT и RAG. Для проверки применимости выявленных закономерностей к другим казахоязычным моделям дополнительно оценивались более крупные модели KazLLM и Sherkala-Chat. Все архитектуры тестировались на 1 081 вопросе с множественным выбором из Единого национального тестирования по истории Казахстана и на наборе из 500 открытых вопросов. Результаты показали, что Qolda+RAG обеспечивает точность $74,93 \pm 0,43\%$ и средний показатель качества $3,34 \pm 1,64$. В рассматриваемых условиях SFT преимущественно выполняет функцию форматирования ответа, сокращая его среднюю длину с 51,8 до 14,64 слова. Архитектура Qolda-SFT+RAG выполняет вывод примерно вдвое быстрее Qolda+RAG при сопоставимом уровне качества $3,10 \pm 1,42$. Полученные результаты формируют методологическую основу для дальнейшей разработки ИИ-агентов с поддержкой казахского языка для широкого спектра прикладных задач.

Ключевые слова: малые языковые модели (SLM); поисково-дополненная генерация (RAG); дообучение с учителем (SFT); низкоресурсные языки; история Казахстана; обработка естественного языка (NLP)

Introduction. Large language models (LLMs) have transformed how machines create text and solve problems. However, these systems require expensive servers with graphics processing units (GPUs) and introduce delays due to their dependence on remote servers (Wang et al., 2025). SLMs are capable of running on local hardware, which enables immediate data processing without an external connection to remote servers. This also provides additional safety measures to protect user privacy. SLMs offer a path forward through local deployment, lower costs, and the option to train or fine-tune for a specific language and subject (Corradini et al., 2025).

Over the years, Kazakh natural language processing (NLP) has advanced through the development of datasets for machine translation, speech technologies, and cultural knowledge benchmarks (Maxutov et al., 2026). Dedicated Kazakh language models include ISSAI-KazLLM and Sherkala-Chat (Koto et al., 2025). There are also Kazakh small language models such as Qolda (Arystanbekov et al., 2026), Llama-3.1-KazLLM-1.0-8B, and Llama-3.1-Sherkala-8B-Chat. In (Mansurova et al., 2025), a Kazakh question answering pipeline with a retrieval based method was explored. In the literature of locally developed SLMs, there is not yet a study exploring Kazakh history question answering applications. General purpose LLMs often hallucinate regarding Kazakh historical entities. One potential application could be supporting students in passing the Unified National Test (UNT), which requires Kazakhstani students to recall thousands of historical facts. However, AI tools that could provide reliable assistance regarding the history of Kazakhstan are yet to be developed. There are two techniques that are widely used for creating a subject specific SLM through the technique of domain adaptation. The first is the retrieval augmented generation (RAG) method that employs retrieved external context to generate an output response. The second is the supervised fine tuning (SFT) method, also known as the knowledge injection method. It could shape the format of the output and behavior (i.e., the style of the answer during the interaction) (Ovadia et al., 2024).

In this paper, we study the performance of four architectures employed for creating subject specific question answering SLMs based on RAG and SFT techniques. We use a question answering methodology applied to the case study of the history of Kazakhstan. The four main architectures are built on Qolda model, a Kazakh language SLM developed at the Institute of Smart Systems and Artificial Intelligence (ISSAI) (Arystanbekov et al., 2026). The first is a baseline Qolda model. The second is Qolda combined with SFT on 48,000 Wikipedia-derived instruction pairs. The third is Qolda combined with a RAG pipeline built over 29 digitized history textbooks that produced 2.25 million indexed tokens. The fourth is Qolda combined with both SFT and RAG. To establish whether the findings generalize across Kazakh language SLMs, we additionally evaluate two external baseline models, specifically KazLLM and Sherkala-Chat. Their base form and the form combined with the proposed RAG pipeline used for Qolda are studied. This produced an eight model comparison covering three independently developed Kazakh SLMs. For evaluation, we use two test sets. The first is a multiple choice test set containing 1,081 (MCQ) samples from the UNT examination in Kazakhstan history. The second is an open ended test set of 500 AI generated and human verified queries, including 100 adversarial trap questions designed to test resistance to false historical premises. Open ended response quality is assessed using an LLM-as-a-Judge approach with Gemini 3 Flash Preview as the automated judge. The contributions of this paper are therefore threefold: a Kazakh history RAG pipeline that produces large and consistent accuracy gains across three independently developed Kazakh SLMs; the first systematic test of whether SFT

contributes factual knowledge or only response formatting in a Kazakh domain; and an adversarial robustness benchmark on Kazakhstan-specific historical false premises.

The rest of this paper is organized as follows: In Literature Review, a survey is presented on the related research work done in the field of SLMs, RAG and SFT systems, along with the strengths and weaknesses of the retrieval vs. fine-tuning approaches. Materials and Methods discusses the proposed architecture along with the performance test methodology for evaluating the proposed RAG & SFT based SLM. Results of the evaluation tests are presented in the Results part of the paper with focus on latency, resource utilization, and architectural performance comparisons. In Discussion, we discuss the limitations of our proposed approach along with future research directions. We conclude the paper in Conclusions.

Literature Review. Even though LLMs perform well on general language tasks, their performance drops when dealing with low-resource languages and narrow specialized domains. LLMs are associated with high computational costs, inference latency, and dependence on remote servers. This could limit their further integration into educational settings. On the other hand, SLMs support the local deployment of models and allow more flexibility in terms of data management and privacy. With fewer parameters, SLMs run on local hardware, support faster inference, and are cheaper to train and serve. Moreover, SLMs can be fine-tuned to become specialized for specific tasks and domains. Fine-tuning has been shown to incorporate new domain-specific facts into model parameters and to improve out-of-domain factual accuracy. Examples of SLMs fine-tuned for specific tasks include WizardMath (i.e., an SLM supporting mathematical reasoning), Code Llama and WizardCoder (i.e., SLMs for code generation), and ALMA-13B-R (i.e., an SLM for machine translation) (Luo et al., 2024). To increase the performance of an SLM in domain-specific cases, researchers developed the technology of RAG, which combines a parametric language model with non-parametric external knowledge. Retrieval-based approaches are considered as practical ways to extend the effective knowledge of an SLM without retraining.

A RAG system has two components: a retriever that selects relevant documents for a given query, and a generator that conditions on both the query and the retrieved documents to produce the final output. Such a design lets the model access information that is not encoded in its parameters, useful in knowledge-intensive domains where the pretraining coverage of the model is limited. RAG performance is also known to degrade when retrieval returns irrelevant context or when the query contains a false premise (Barnett et al., 2024). RAG has been adapted to SLMs in several ways. RALM separates retrieval and reading into explicit components, allowing a small generator to operate over passages selected from an external corpus (Ram et al., 2023). DRAG transfers retrieval-augmented behavior from a large model into a smaller one through distillation, so that the smaller model can produce evidence-supported outputs at a lower cost (Chen et al., 2025). More recent frameworks, such as MiniRAG (Fan et al., 2025) and

MCTS-RAG (Hu et al., 2025), combine retrieval with structured reasoning to further improve SLM performance on knowledge-intensive tasks.

SFT is another approach used to tailor a language model for certain tasks and fields like Q&A, reasoning, scientific work, and dialogue applications. In this approach, a pre-trained language model is fine-tuned using input-output labeled data. These labeled datasets usually come from humans, hence allowing the model to learn the desired behavior via supervised loss minimization (Lu et al., 2025). However, the way in which SFT impacts the model's behavior is not yet clear among researchers. The first theory of SFT sees it as a knowledge injection technique whereby the model acquires the ability of a field expert by injecting domain knowledge in a structured fashion. On the other hand, the other theory views SFT as more about modifying the output behavior and format than adding new knowledge to the model. However, an issue with these instruction-tuned models is that they provide overconfident hallucinations on out-of-domain prompts. The reason is their tendency towards memorizing patterns instead of understanding.

However, both RAG and SFT are inherently limited. For example, RAG is prone to issues such as stylistic incoherence and poor integration of retrieved context. In contrast, the limitation for SFT is that its parametric memory is static and forgetful. Hybrid architectures overcome these challenges through two main techniques: RAG allows retrieval of a dynamic context to anchor the generation process, and SFT enables the learning of reading, synthesizing, and formatting of retrieved texts, while ignoring distractor passages. There exist two fundamental types of hybrid architectures that use RAG and SFT, namely RAFT and Self-RAG. RAFT is designed to train the generator in the presence of correct documents along with distractors, thereby forcing the architecture to reason about the context in a robust manner (Zhang et al., 2024). The Self-RAG framework employs SFT to learn through reflection tokens about how and when to retrieve, thus enabling the model to evaluate itself and determine when context is necessary (Asai et al., 2023).

Materials and Methods. *System Description.* The system uses a standard pipeline of retrieval and generation, using Haystack and Qdrant. It serves as a controlled baseline for measuring the contribution of retrieval itself before adding reasoning or distillation layers. So we evaluate how well RAG and SFT, together and individually, work with the base model in domain-specific areas and to see if systems improve factual accuracy and linguistic fluency. The base model is Qolda, a small VLM of ISSAI, which is fine-tuned on Kazakh-language data and is being used across architectures. Qolda represents the current state-of-the-art Kazakh-language model. Since the model has relatively low parameters, it can be used for developing edge-deployed solutions with relatively low latency requirements. Figure 1 provides four Qolda-based architectures along with the two external baselines, KazLLM and Sherkala, through a unified RAG pipeline and SLM deployment layer.

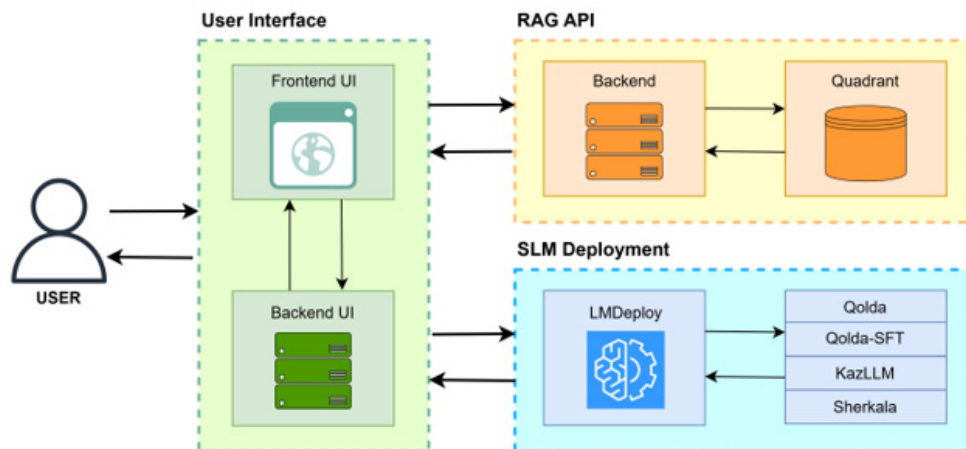


Figure 1 — System architecture of the Kazakhstan history question answering framework, serving four small language models (Qolda, Qolda-SFT, KazLLM, and Sherkala).

Dataset. SFT dataset is retrieved from the Kazakh Wikipedia PPlatform. It includes articles distributed as compressed streams of 100-page documents. By using the Wikimedia RAST API, articles were selected based on their views from real users, and in total, 500 Mb of data was drawn in CSV format. As a result, 4,076 articles rated with a minimum of 2,000 views and in the period from 2020 to 2025 were selected by filtering topic tags such as "Kazakh history," "history," and "Kazakh Khanate." This demand-driven selection of data further ensures that the SFT training is aligned with the historical queries that real users might ask.

The dataset for RAG comprises 29 books about Kazakh history written digitally from oqulyq.kz and other Kazakh education-related websites. The size of the corpus is 813.3 MB of PDF files. The texts were extracted through DotsOCR which is an OCR model that uses VLM technology to handle various types of typography. Preprocessing operations such as artifact cleaning and filtering of header information were done on the documents. Following data preprocessing, the texts were split into 4,388 overlapping fragments using the Haystack DocumentSplitter API. A chunk length of 400 words (about 512 tokens), with 50-word overlap between each fragment, was applied on average to generate 151 chunks per document. During data preprocessing, there were 77,500 tokens per book, summing up to a total number of 2.25 million tokens in the RAG dataset. Embeddings were generated for the documents using intfloat/multilingual-e5-large-instruct embeddings, then indexed in the Qdrant vector store to retrieve the top-5 passages based on cosine similarity decision-making strategy during the inference stage.

The evaluation benchmark consists of 1,081 MCQs assembled from several UNT history of Kazakhstan preparation platforms, including Juz40, FM Online, and Hogwarts Edu. The UNT is also known as a standardized national examination

testing factual and conceptual knowledge of the history of Kazakhstan across the full national curriculum. Questions were collected via automated web scraping, with HTML tags stripped from question stems and answer options using BeautifulSoup. Selected questions were structured into a CSV format using pandas. Each question presents a stem and four candidate answers with a single correct answer. The evaluation dataset was not used at any stage of training or fine-tuning, serving exclusively as a held-out benchmark for the four-architecture comparison.

RAG Pipeline. The RAG pipeline uses a Qdrant vector store deployed locally on the Irgetas server at Nazarbayev University. We utilized 8×NVIDIA H100 GPUs and connected to the Qolda generation model through a Haystack-orchestrated retrieval layer served via a unified FastAPI backend. The retrieval pipeline is orchestrated using the Haystack framework that manages the end-to-end flow from query encoding through document retrieval to context assembly. The pipeline abstraction of Haystack connects the Qdrant retriever to the Qolda generation model, passing the top-5 retrieved passages as a structured context prefix prepended to the user query before response generation. No reranking or query expansion is applied. The pipeline follows a standard retrieve-and-generate design, which serves as a controlled baseline for all.

All four architectures can be accessed via a common FastAPI server. FastAPI creates a unified REST API through which queries can be submitted and responses received irrespective of which architecture is currently running. All queries will include information regarding the architecture being queried. This information will then determine which architecture to pass the request through, either 1) baseline Qolda, 2) Qolda with SFT, 3) Qolda with RAG, or 4) Qolda with RAG and SFT combined. Response payloads include the generated answer along with metadata such as latency. For RAG-enabled architectures, the retrieved source passages are also reported. A unified serving layer makes sure that all four architectures are run under identical request-handling conditions.

Supervised Fine-Tuning (SFT). The training dataset for SFT was derived from 4,076 demand-selected Kazakh Wikipedia articles on the history of Kazakhstan with a minimum of 2,000 views. For each article, Gemini 2.5 Flash was prompted at a temperature of 0.3 to generate five base instruction-answer pairs grounded in the article content. Each pair was further expanded into stylistic variants, including paraphrase, reverse, positive verification, and negative verification formulations. This helped to increase lexical diversity and reduce overfitting to a single phrasing. The final dataset has approximately 48,000 examples in Kazakh in the InternVL conversational chat format. Then, the dataset was split 95/5 into training and test sets. The 5% part was used to monitor the training loss.

Qolda fine-tuning was performed using a full fine-tuning method without parameter freezing or low-rank approximation. This allowed all model weights to be updated on the history of Kazakhstan training data. Training was conducted using the Hugging Face Trainer with the following configuration parameters: a

learning rate of $1e-5$, three epochs, the AdamW optimizer with a cosine learning rate scheduler, and a maximum sequence length of 8,192 tokens. Model checkpoints were saved every 500 steps. No quantization was applied during training. Training was performed on eight NVIDIA H100 GPUs on the Irgetas server at NU, using the Hugging Face Trainer with DeepSpeed ZeRO-3 for distributed training. All weights were maintained in bf16 precision throughout. The full fine-tuning across the training set was completed in approximately 20.5 hours.

Experimental Setup & Evaluation. In-system evaluation will measure system performance by the latency, throughput, and response length metrics generated during inference. Each of the four systems will be evaluated separately using all 1,081 question instances present in the UNT benchmark, resulting in 4,324 total inference instances. To calculate MCQ accuracy, we compare each model response with the correct answer in the test CSV file. The model's answer is composed of the correct letter and text. The matching of the correct and model given responses is done in three steps. First, a regular expression checks if the response starts with an option letter followed by punctuation or a space. If it does, the letter is extracted. Both Latin (A, B, C, D) and Cyrillic (A, Б, В, Г) letters are accepted, and they are mapped by their position. For very short responses of two characters or less, only the first character is checked. Second, if a letter is found, we take the full text of that option and compare it with the correct answer. Third, if there is no letter, or the option found by the letter does not match, we check whether the correct answer text appears inside the response. This match is accepted only if the response does not also contain the text of another option. If a response reaches this step and mentions more than one option, it is counted as incorrect. If a response does not match any of the rules, it is counted as incorrect. MCQ accuracy is then calculated with the following equation, where $\hat{y}_i$ is the option matched from response i by the steps above and y_i^* is the correct option in the test CSV:

$$Acc = \frac{1}{N} \sum_{i=1}^N 1[\hat{y}_i = y_i^*], \quad N = 1081 \quad (1)$$

In the above formula, $1[\cdot]$ is the indicator function. It returns 1 if response i is matched to the correct option, and 0 otherwise. LLM as a judge is not used for the MCQ benchmark, so the scoring step always gives the same result for the same responses. The MCQ dataset can be accessed here: <https://huggingface.co/datasets/issai/unt-kz-history-mcq>. The evaluation of the models' performance in answering the open-ended questions is done by a specially created test set containing 500 open-ended questions organized in five thematic categories, namely: 1) Ancient Kazakhstan (1–100), 2) Kazakh Kaganate (101–200), 3) Soviet period (201–300), 4) Era of independence (301–400), and 5) Adversarial Trap Questions (401–500). The latter category comprises questions with embedded falsehoods, myths, and anachronisms to explore the model's ability to detect and fix incorrect assumptions.

The questions are written in Kazakh, Russian, and English to ensure multilingual support for the Kazakhstan history domain. The dataset with 500 queries can be found at <https://huggingface.co/datasets/issai/kz-history-queries-multilingual>

Table 1 — Sample questions from the UNT history of Kazakhstan MCQ testing dataset in Kazakh.

#	Question	A	B	C	D	Answer
1	«Тарих-и Рашиди» еңбегін жазған ғалым	Мұхаммед Хайдар Дулати	Масуд ибн Осман	Бабыр	Әбілғазы Баһадүр	A
2	Қазақтың ең көп таралған ою-өрнегі	Түйемойын	Сыңар-мүйіз	Қосмүйіз	Қошқар мүйіз	D
3	1940–1980 жылдары бұқаралық ақпарат құралдарының рөлі	Технологиялық инновация	Саяси тұрақтылық	Қоғамдық пікірді қалыптастыру	Экономикалық даму	C
4	МТС-тердің (машина-трактор станциялары) негізгі қызметі	Салық жинау	Техниканы жалға беру	Егінді қабылдау	Өнімді сату	B
5	1989 жылы қабылданған «Тілдер туралы» заңның басты жаңалығы	Қазақ тілі мемлекеттік мәртебе алды	Барлық тілдер жойылды	Латын әліпбиі міндетті болды	Орыс тілі мемлекеттік мәртебе алды	A

Table 2 — Example queries used for system evaluation on the open-ended questions.

Question	Kazakh	English
Ancient Kazakhstan (1–100)	Жібек жолы саудасы контекстінде Тараз қаласы «көпестер қаласы» ретінде қандай рөл атқарды?	What role did the city of Taraz play as a “city of merchants” in the context of the Silk Road trade?
Kazakh Khanate (101–200)	«Есім ханның ескі жолы» заңдар жинағы бұрынғы құқықтық дәстүрлерден несімен ерекшеленді?	How did the legal code “Esim Khannyn Eski Zholy” (The Ancient Path of Esim Khan) differ from previous legal traditions?
Soviet Period (201–300)	1920-шы жылдары астананың Орынбордан Қызылордаға, одан кейін Алматыға көшірілуінің маңыздылығын талдаңыз.	Analyze the significance of the movement of the capital from Orenburg to Kyzylorda and finally to Alma-Ata in the 1920s.
Independence Era (301–400)	1993 жылы ұлттық валюта – теңгенің енгізілуі Қазақстанның экономикалық егемендігіне қалай әсер етті?	How did the 1993 introduction of the national currency, the Tenge, impact Kazakhstan's economic sovereignty?
Adversarial/ Trap (401–500)	Ботай мәдениетінің XX ғасырдағы ғарыш индустриясының дамуына қосқан үлесін талдаңыз.	Analyze the contribution of the Botai culture to the development of the 20th-century space industry.

Model responses are assessed through Gemini 3 Flash Preview with a rubric for three main criteria. The first criterion is the historical correctness of the information provided. The second is linguistic integrity that includes proper use of technical

terms like *Khanate*, *Ulus*, *Asharshylyk*, and *Zhuz*. The third criterion is linguistic suitability. Model answers get a numeric quality score of $s_i \in \{1,2,3,4,5\}$, where “1” implies an incorrect or missing answer, “3” implies a partially correct answer with some mistakes, and “5” implies an absolutely correct answer with cultural nuance and correct historical terminologies. The mean quality score for each batch is (2). For trap questions, Gemini provides additional feedback with a Boolean variable named *trap_pass*. When the model agrees with rather than disproves the trap premise, the value of the *trap_pass* variable becomes false, and the quality score is “1.” The *trap_pass* rate is (3)

$$Q = \frac{1}{N} \sum_{i=1}^N s_i, \quad s_i \in \{1, 2, 3, 4, 5\} \quad (2)$$

$$TRP = \frac{1}{N_{Trap}} \sum_{i=1}^{N_{Trap}} 1[\text{trap_pass}_i] \quad (3)$$

A boolean *lang_correct* flag and a one-sentence *reasoning* field are returned per question to provide interpretable diagnostic output alongside the numeric scores. Latency analysis during system performance evaluation uses metrics such as time-to-first-token (TTFT) and words-per-second (WPS). TTFT measures the latency between query submission t_{submit} and the generation of the first output token t_{first} , capturing the responsiveness of each architecture under load. WPS metric measures response generation throughput as the number of generated words N_{words} divided by the elapsed generation time between the first and the final token t_{end}

$$TTFT = t_{first} - t_{submit}, \quad WPS = N_{words} / (t_{end} - t_{first}) \quad (4)$$

Total inference time measures the end-to-end duration from query receipt to completed response. Response word count provides a normalized basis for comparing output verbosity across architectures. For RAG-enabled architectures, retrieval time is recorded separately to isolate the overhead introduced by the Qdrant retrieval step from the generation latency.

Results. Table 3 reports MCQ accuracy across six examination periods for the four main architectures (i.e., baseline Qolda, Qolda-SFT, Qolda+RAG, and Qolda-SFT+RAG). Four additional models are used as external baselines: KazLLM and Sherkala in their base form and in their combined RAG forms. Retrieval grounding produces large and consistent accuracy gains across all base models. Overall scores rose from 44% to 49% without RAG and from 64% to 77% with RAG. Four

non-RAG models perform similarly, clustering between $44.40 \pm 0.49\%$ and $49.31 \pm 0.50\%$ overall indicating that base parametric knowledge alone is insufficient for closed-book Kazakhstan history questions. SFT on Wikipedia data alone does not improve baseline Qolda accuracy (i.e., $44.40 \pm 0.49\%$ versus $45.24 \pm 0.49\%$) showing that, in this study, SFT on Wikipedia-derived pairs does not improve closed-book MCQ performance. KazLLM+RAG obtains the best accuracy at $76.60 \pm 0.42\%$. Contrarily, Sherkala exhibits an uncommon trend: while it has the best accuracy among models without RAG at $49.31 \pm 0.50\%$, its accuracy when enhanced by RAG is the lowest at $64.38 \pm 0.48\%$. The performance is below average for several periods when augmented by RAG, implying that Sherkala incorporates context into the process less effectively than other baseline models. Period 6 is the best-performing period overall, with all architectures recording above 70% accuracy and the highest value of $86.07 \pm 0.35\%$ by Qolda-SFT+RAG.

Table 4 shows the Gemini-assessed quality scores in five thematic batches. In batches 1 to 4, Qolda+RAG has the highest quality score (3.34 ± 1.64), followed by Qolda-SFT+RAG (3.10 ± 1.42), Qolda-SFT (2.46 ± 1.28), and finally Qolda as the baseline model (2.13 ± 1.26). The architectures enabled by RAG are superior to the non-RAG architectures in all historical batches, the most significant difference being seen in the Kazakh Kaganate batch, where Qolda+RAG scores 3.72 ± 1.65 compared to only 1.98 ± 1.68 for the baseline architecture. SFT, on its own, gives some improvement over the baseline in all historical batches. It is clear from these results that adding RAG architectures improves response coherence without affecting factual grounding in any significant way. For the trap batch (Batch 5), Qolda+RAG gets a pass rate of 82/100, far surpassing that of Qolda-SFT+RAG (47/100) and the non-RAG baselines (27/100 each). This indicates that retrieved source passages provide strong grounding that helps the model resist false premises, while SFT alone does not reliably improve adversarial robustness over the unmodified baseline.

Latency and throughput metrics analysis for four architectures are presented in Table 5. As can be seen, non-RAG architectures significantly outperform the RAG-enabled ones. Thus, for instance, the baseline architecture Qolda yields 0.06 ± 0.02 s TTFT and 1.45 ± 0.90 s of total inference time on average. The fastest Qolda architecture turns out to be Qolda-SFT, yielding 0.05 ± 0.02 s TTFT and 0.57 ± 0.17 s of total inference time. This architecture produces the shortest answers on average, which amounts to 14.64 words. In turn, RAG-enabled architectures are characterized by a retrieval overhead of about 0.67 ± 0.52 s per query independent of the batch. For example, Qolda+RAG produces 0.24 ± 0.01 s TTFT and 1.18 ± 0.76 s of total inference time, whereas Qolda-SFT+RAG provides 0.66 ± 0.59 s of total inference time. Throughput values are similar for all architectures and vary between 23.2 and 26.6 WPS. The baseline Qolda generates the longest answers (on average, 51.8 words), while Qolda-SFT generates the shortest answers (14.64 words). SFT shifts the model toward more concise generation. The highest values among architectures across the tables are shown in bold.

Table 3 — MCQ accuracy by examination period across four architectures (Mean $\pm$ STD in %).

#	Qolda				KazLLM		Sherkala	
	Base	SFT	RAG	SFT+RAG	Base	RAG	Base	RAG
1	37.25 $\pm$ 0.48	38.4 $\pm$ 0.48	75.69 $\pm$ 0.43	74.12 $\pm$ 0.44	37.25 $\pm$ 0.48	76.86 $\pm$ 0.42	41.57 $\pm$ 0.49	68.24 $\pm$ 0.47
2	40.60 $\pm$ 0.49	39.8 $\pm$ 0.48	72.18 $\pm$ 0.45	69.92 $\pm$ 0.46	36.84 $\pm$ 0.48	72.93 $\pm$ 0.44	49.62 $\pm$ 0.50	72.18 $\pm$ 0.45
3	54.98 $\pm$ 0.49	57.8 $\pm$ 0.49	75.83 $\pm$ 0.43	77.73 $\pm$ 0.42	52.13 $\pm$ 0.49	81.52 $\pm$ 0.39	54.50 $\pm$ 0.49	56.87 $\pm$ 0.50
4	38.74 $\pm$ 0.48	37.2 $\pm$ 0.47	74.87 $\pm$ 0.43	69.11 $\pm$ 0.46	35.60 $\pm$ 0.47	75.39 $\pm$ 0.43	44.50 $\pm$ 0.49	63.35 $\pm$ 0.48
5	33.14 $\pm$ 0.47	29.0 $\pm$ 0.45	67.46 $\pm$ 0.47	66.86 $\pm$ 0.47	35.50 $\pm$ 0.47	68.05 $\pm$ 0.47	35.50 $\pm$ 0.47	57.40 $\pm$ 0.49
6	77.05 $\pm$ 0.42	71.3 $\pm$ 0.44	85.25 $\pm$ 0.35	86.07 $\pm$ 0.35	81.15 $\pm$ 0.39	85.25 $\pm$ 0.35	82.79 $\pm$ 0.37	72.13 $\pm$ 0.45
	45.24 $\pm$ 0.49	44.40 $\pm$ 0.49	74.93 $\pm$ 0.43	73.64 $\pm$ 0.44	44.50 $\pm$ 0.49	76.60 $\pm$ 0.42	49.31 $\pm$ 0.50	64.38 $\pm$ 0.48

Table 4 — Gemini-judged quality scores by thematic batch across four architectures (Mean $\pm$ STD).

Batch	Qolda	Qolda-SFT	Qolda + RAG	Qolda-SFT + RAG
Ancient Kazakhstan	2.26 $\pm$ 1.19	2.54 $\pm$ 1.23	3.66 $\pm$ 1.47	3.18 $\pm$ 1.29
Kazakh Kaganate	1.98 $\pm$ 1.22	2.37 $\pm$ 1.43	3.72 $\pm$ 1.65	3.38 $\pm$ 1.31
Soviet Period	1.99 $\pm$ 1.01	2.25 $\pm$ 1.01	3.28 $\pm$ 1.41	2.97 $\pm$ 1.24
Independence Era	2.28 $\pm$ 1.06	2.69 $\pm$ 1.10	2.69 $\pm$ 1.63	2.85 $\pm$ 1.29
Adversarial/Trap	1.93 $\pm$ 1.68	1.77 $\pm$ 1.40	3.7 $\pm$ 1.61	2.44 $\pm$ 1.72
Overall (1–4)	2.13 $\pm$ 1.26	2.46 $\pm$ 1.28	3.34 $\pm$ 1.64	3.10 $\pm$ 1.42

Table 5 — System performance metrics across four architectures.

Metric	Qolda	Qolda-SFT	Qolda + RAG	Qolda-SFT + RAG
TTFT (s)	0.06 $\pm$ 0.02	0.05 $\pm$ 0.02	0.24 $\pm$ 0.01	0.24 $\pm$ 0.01
Total time (s)	1.45 $\pm$ 0.90	0.57 $\pm$ 0.17	1.18 $\pm$ 0.76	0.66 $\pm$ 0.59
RAG time (s)	—	—	0.67 $\pm$ 0.52	0.67 $\pm$ 0.52
WPS (words/s)	25.81 $\pm$ 12.7	23.23 $\pm$ 8.1	26.66 $\pm$ 9.81	24.38 $\pm$ 7.86
Avg word count	51.81	14.64	39.87	18.59

Discussion. In both categories of the evaluations, Qolda+RAG shows the highest factual accuracy and open-ended quality scores, scoring $74.93 \pm 0.43\%$ on the UNT MCQ benchmark and an average quality score of 3.34 ± 1.64 for historical batches. This suggests that, in our study, access to external knowledge helps more than updating the model weights. Using the verified knowledge in generated responses resolves the main limitation of the baseline approach much better than any other solution considered. This outcome stands out since Qolda+RAG does not perform any adjustments to the weights of the base model. The baseline Qolda SLM is indeed a skillful reasoner once it receives appropriate information. Therefore, its baseline performance is poor since it lacks sufficient knowledge resources. The same trend can be observed among the external baselines. KazLLM

and Sherkala score around 25–35% higher accuracy for the whole MCQ dataset due to the addition of RAG pipelines. This suggests that the advantage of retrieval generalizes across Kazakh-language SLMs and is not an artifact of any single base model.

In the case of the MCQ benchmark, there is no significant improvement by Qolda-SFT compared to the baseline ($44.40 \pm 0.49\%$ versus $45.24 \pm 0.49\%$). In other words, the results confirm the hypothesis about the stylistic nature of SFT. MCQ questions contain single-letter answers which can be true only if factual knowledge is recalled. Hence, if any positive effect could be found in answering the questions in the MCQ benchmark after SFT learning, it would prove a change in parametric knowledge of the model. However, since no significant improvements were achieved in answering these questions despite the use of 48,000 instruction pairs, this result does not support the "knowledge injection" view of SFT. There is more evidence that supports the above hypothesis regarding the impact of SFT, namely the shift in response length. In particular, it should be noted that Qolda-SFT generates considerably shorter responses compared to the baseline (14.64 versus 51.8 words, respectively) as well as an open-ended quality gain (2.46 ± 1.28 versus 2.13 ± 1.26). Such results are consistent with cleaner formatting of SFT rather than richer factual content processing. Taken together, the MCQ result, the response length shift, and the modest open-ended gain all point to the same conclusion that SFT on Wikipedia-derived instruction pairs adapts how the model writes, not what the model knows.

The Qolda-SFT+RAG offers a strong balance between quality and latency. In terms of quality, its score of 3.10 ± 1.42 puts it only 0.24 points behind Qolda+RAG while reducing inference completion time by half – from 1.18 ± 0.76 s to 0.66 ± 0.59 s. The performance increase results directly from the formatting effect of SFT: while the number of tokens generated per second is roughly the same in both models, their outputs are shorter by a factor of two (on average 18.59 versus 39.87 words). Consequently, the faster inference time of Qolda-SFT+RAG makes it a viable candidate for latency-critical applications such as in-class educational systems. On the other hand, Qolda+RAG is able to pass the trap questions at a rate of 82 out of 100, compared to Qolda-SFT+RAG's 47/100. That's a difference of 35% in performance on a benchmark explicitly created to measure resilience against incorrect historical assumptions. Shorter answers learned through SFT may reduce robustness to false premises, even when the retrieved passages contain enough information to reject them. These two models have thus been optimized for distinct use cases, with Qolda-SFT+RAG optimizing for speed in response to well-posed questions, and Qolda+RAG for adversarial robustness.

The four main architectures in this paper all use Qolda model as the base, with KazLLM and Sherkala included as external baselines. This design choice is supported by the data. KazLLM+RAG achieves the highest overall MCQ accuracy at $76.60 \pm 0.42\%$. This is slightly above Qolda+RAG at $74.93 \pm 0.43\%$. The difference is small relative to the 25-35 % gain that retrieval itself provides

over any non-RAG variant. Question on SFT contribution to factual knowledge in addition to response formatting, is best answered with a single base model where SFT can be evaluated cleanly. Qolda was selected for this role for two reasons. First, it is the most recently released Kazakh language SLM among those compared, published less than a year before this study. Second, it is a 4B parameter model based on Qwen, while KazLLM and Sherkala are 8B models based on LLama model. The smaller parameter count makes Qolda suitable to fine-tune and serve in the edge-deployment scenario. It may also account for the small accuracy gap between Qolda+RAG and KazLLM+RAG. Third, Qolda is a small VLM and it can process images alongside text. Image processing capability is not used in the current study. It can add multimodality, extending the system into a tutoring agent capable of interpreting historical maps, document scans, archival photographs, and student handwritten work.

The external baselines also produce an interesting observation. Sherkala has the strongest non-RAG accuracy of any base model ($49.31 \pm 0.50\%$) but the weakest RAG-augmented accuracy ($64.38 \pm 0.48\%$), with several periods showing reduced performance once retrieval is added. This is the only case in the results where adding RAG fails to produce a substantial accuracy gain. KazLLM, the Qolda baseline, and Qolda-SFT all improve sharply under RAG. Results show that Sherkala integrates retrieved context less effectively than the other base models. This could possibly be because of instruction tuning that shapes it toward producing answers from parametric memory rather than synthesizing retrieved evidence. This observation is consistent with the "SFT-as-formatter" finding from the Qolda family. A model that is heavily instruction-tuned for fluent direct answers may resist or under-use external context, even when that context is provided. A controlled study of how different Kazakh-language SLMs interact with retrieved passages, holding the retrieval pipeline constant, would be a useful direction for future work.

Limitations. In addition to research findings the study also has its limitations.

1. The LLM-as-a-Judge experiment uses a one-judge model - Gemini 3 Flash Preview. This may result in the presence of system biases towards some answer types.

2. Despite the large number of texts included in RAG corpus with 2.25 million tokens and the use of information from 29 books, it might not yet cover the entire period of the history of Kazakhstan. This can lead to retrieval gaps for those periods that were documented poorly.

3. Training data for SFT is based only on Kazakh Wikipedia and was not checked systematically for factual errors. Its 95/5 split was random at the pair level, so variants of the same source text may appear in both parts. This could affect the held-out loss keeping same benchmark results.

4. All latency results were calculated on the Irgetas server. This could make results being specific to one server hardware configuration and may not generalize

to other deployment environments. However, generalizability research is not within the scope of the current paper.

Directions for future research. These results open up some interesting directions for future research. They were all aimed at enhancing the proposed Qolda-based solution to develop its capabilities as an interactive tutoring bot designed for history studies of Kazakhstan. In this regard, recent research has focused on how LLMs may act as virtual assistants for a variety of different disciplines, such as computer science, math, etc. (García-Méndez et al., 2025). The dialogue architecture used by tutoring bots, for instance, MWPTutor (Pal Chowdhury et al., 2024), can be employed with the use of a pedagogically designed syllabus combined with an LLM-based dialogue module. Furthermore, the latest frameworks like TeachBench (Li et al., 2026) allow for syllabus-driven assessments of the teaching abilities. Finally, the architecture can be improved with more sophisticated retrieval frameworks, such as MiniRAG and MCTS-RAG.

Future research directions in this area would be the expansion of the RAG corpus from the present 29 textbooks to include archival material, scientific articles, and historical texts in other languages. This would increase the number of retrievable items, particularly for under-represented time periods, as well as bridge the gap seen in some of the less successful examinations at UNT. Finally, future research may explore a broader language range by applying the methodology to question sets in the languages of Kazakh, Russian, and English. This approach will allow us to study whether the findings generalize across languages and better reflect the multilingual nature of historical discourse in Kazakhstan. In the future work one can validate the LLM as the judge involving human evaluation. In that case a sample of model answers are scored by native Kazakh speakers using the same rubric.

Conclusions. This paper presents the first systematic comparison of RAG and SFT systems applied to Kazakh language supported SLM for the question answering in the domain of history of Kazakhstan. We constructed and evaluated four main architectures: a baseline Qolda SLM, Qolda-SFT, Qolda + RAG, and Qolda-SFT + RAG and SFT. To strengthen the comparison, two additional Kazakh language SLMs such as KazLLM and Sherkala, were included as external baselines in their base form and combined with the used RAG framework. To support this evaluation, we built a RAG system using 29 digitized Kazakh history textbooks resulting in 2.25 million indexed tokens. Also, we used fine-tuned Qolda on 48,000 demand-selected Wikipedia-derived instruction pairs. We evaluated all architectures on 1,081 UNT MCQs and a manually constructed 500-question open-ended benchmark that also included adversarial trap questions.

We see that Qolda+RAG scores $74.93 \pm 0.43\%$ MCQ accuracy and 3.34 ± 1.64 quality on average, demonstrating better performance than the baseline and SFT-only model. The same is observed with external baselines, as well. In particular, KazLLM and Sherkala gain 25-35% in MCQ accuracy due to the usage of RAG pipeline. KazLLM+RAG obtained the highest overall accuracy ($76.60 \pm$

0.42%). All results were reproduced on three independent SLMs for the Kazakh language. SFT acted more like a response formatter than a knowledge injector. Qolda-SFT demonstrates a negligible improvement in MCQ accuracy compared to the baseline ($45.24 \pm 0.49\%$ vs. $44.40 \pm 0.49\%$), despite being trained on 48K instruction pairs. The average number of words in responses was decreased from 51.8 to 14.64, while the total inference time was cut in half. Qolda-SFT+RAG requires about half of the inference time required by Qolda+RAG (0.66 ± 0.59 s vs. 1.18 ± 0.76 s), obtaining a similar quality score (3.10 ± 1.42). This makes it a viable choice for latency-sensitive deployments. However, the trap pass rate for Qolda-SFT+RAG drops sharply relative to Qolda+RAG (47/100 versus 82/100), suggesting that the bias of SFT could reduce robustness of the architecture. As a result, in our study, retrieval alone was more robust to false historical premises than retrieval combined with SFT on Wikipedia-derived data.

References

- Arystanbekov B., Nurimanov A., Maxutov A., Albrekht V., Kuzdeuov A., Varol H.A. (2026) Qolda: A small vision–language model for the Kazakh language. *IEEE Access*, 14:46392–46414. DOI: 10.1109/ACCESS.2026.3676919 (in Eng.).
- Asai A., Wu Z., Wang Y., Sil A., Hajishirzi H. (2023) Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proc. of the Twelfth International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2310.11511 (in Eng.).
- Barnett S., Kurniawan S., Thudumu S., Brannelly Z., Abdelrazek M. (2024) Seven failure points when engineering a retrieval augmented generation system. In *Proc. of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI*. — P. 194–199. DOI: 10.1145/3644815.3644945 (in Eng.).
- Chen J., Myrzakhan A., Luo Y., Khan H.M., Bsharat S.M., Shen Z. (2025) DRAG: Distilling RAG for SLMs from LLMs to transfer knowledge and mitigate hallucination via evidence and graph-based distillation. In *Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. — P. 7240–7260. DOI: 10.18653/v1/2025.acl-long.358 (in Eng.).
- Corradini F., Leonesi M., Piangerelli M. (2025) State of the art and future directions of small language models: A systematic review. *Big Data and Cognitive Computing*, 9(7):189. DOI: 10.3390/bdcc9070189 (in Eng.).
- Fan T., Wang J., Ren X., Huang C. (2025) MiniRAG: Towards extremely simple retrieval-augmented generation. *arXiv preprint arXiv:2501.06713*. DOI: 10.48550/arXiv.2501.06713 (in Eng.).
- García-Méndez S., de Arriba-Pérez F., Somoza-López M.D.C. (2025) A review on the use of large language models as virtual tutors. *Science & Education*, 34(2):877–892. DOI: 10.1007/s11191-024-00530-2 (in Eng.).
- Hu Y., Zhao Y., Zhao C., Cohan A. (2025) MCTS-RAG: Enhancing retrieval-augmented generation with Monte Carlo tree search. In *Findings of the Association for Computational Linguistics: EMNLP 2025*. — P. 2581–12597. DOI: 10.18653/v1/2025.findings-emnlp.672 (in Eng.).
- Koto F., Joshi R., Mukhity N., Wang Y., Xie Z., Pal R., Orel D., Nakov P. (2025) Sherkala-Chat: Building a state-of-the-art LLM for Kazakh in a moderately resourced setting. In *Proc. of the Second Conference on Language Modeling (COLM)*. DOI: 10.48550/arXiv.2503.01493 (in Eng.).
- Li Z., Song S., Ma J., Li R., Zeng Y., Li M., Sui Z. (2026) TeachBench: A syllabus-grounded framework for evaluating teaching ability in large language models. *arXiv preprint arXiv:2601.21375*. DOI: 10.48550/arXiv.2601.21375 (in Eng.).
- Lu W., Luu R.K., Buehler M.J. (2025) Fine-tuning large language models for domain adaptation:

Exploration of training strategies, scaling, model merging and synergistic capabilities. *NPJ Computational Materials*, 11(1):84. DOI: 10.1038/s41524-025-01564-y (in Eng.).

Luo Z., Xu C., Zhao P., Sun Q., Geng X., Hu W., Tao C., Ma J., Lin Q., Jiang D. (2024) WizardCoder: Empowering code large language models with Evol-Instruct. In *Proc. of the Twelfth International Conference on Learning Representations (ICLR)*. DOI: 10.48550/arXiv.2306.08568 (in Eng.).

Mansurova A., Tleubayeva A., Nugumanova A., Shomanov A., Seker S.E. (2025) A systematic evaluation of large language models and retrieval-augmented generation for the task of Kazakh question answering. *Information*, 16(11):943. DOI: 10.3390/info16110943 (in Eng.).

Maxutov A., Arystanbekov B., Makhataeva Z., Yergen A., Taizhanov N., Nauryzbaikyzy G., Varol H.A. (2026) Introducing cultural knowledge in language models: KazCulture dataset for Kazakh culture. *IEEE Access*, 14:44027–44042. DOI: 10.1109/ACCESS.2026.3675790 (in Eng.).

Ovadia O., Brief M., Mishaeli M., Elisha O. (2024) Fine-tuning or retrieval? Comparing knowledge injection in LLMs. In *Proc. of the 2024 Conference on Empirical Methods in Natural Language Processing*. — P. 237–250. DOI: 10.18653/v1/2024.emnlp-main.15 (in Eng.).

Pal Chowdhury S., Zouhar V., Sachan M. (2024) AutoTutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proc. of the Eleventh ACM Conference on Learning@Scale*. — P. 5–15. DOI: 10.1145/3657604.3662041 (in Eng.).

Ram O., Levine Y., Dalmedigos I., Muhlgaay D., Shashua A., Leyton-Brown K., Shoham Y. (2023) In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*. — P. 11:1316–1331. DOI: 10.1162/tacl_a_00605 (in Eng.).

Wang F., Zhang Z., Zhang X., Wu Z., Mo T., Lu Q., Wang S. (2025) A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with LLMs, and trustworthiness. *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–87. DOI: 10.1145/3768165 (in Eng.).

Zhang T., Patil S.G., Jain N., Shen S., Zaharia M., Stoica I., Gonzalez J.E. (2024) RAFT: Adapting language model to domain specific RAG. *arXiv preprint arXiv:2403.10131*. DOI: 10.48550/arXiv.2403.10131 (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.

Requirements for articles design for publication in the journal are available on the websites:

www.nauka-nanrk.kz

<http://physics-mathematics.kz/index.php/en/archive>

ISSN2518-1726 (Online),

ISSN 1991-346X (Print)

Managing Editor: *A.Shormakova*

Editors: *D.S. Alenov, T. Apendiev*

Computer layout: *G.D. Zhadyranova*

Signed for print: September 30, 2026

Format: 70×90 1/16. 26.5 printed sheets. Order No. 3.