Жоба туралы қысқаша ақпарат

Жоба аты	Қазақ тілін және технологиялық үрдісті қолдау үшін үлкен тілдік модель (LLM) әзірлеу
Жоба өзектілігі	Жасанды интеллект, әсіресе үлкен тілдік модельдердің (LLM) қарқынды дамуы білім экономикасына өту мен мемлекеттердің цифрлық трансформациясының негізгі факторы болып отыр. Әлемде GPT-4 және Gemini сияқты жүйелер белсенді дамып келеді, алайда олар негізінен ресурсы мол тілдерге бағытталған, бұл қазақ тілі үшін олардың тиімділігін шектейді. Мұндай жағдай технологиялық алшақтықты туындатып, Қазақстанның ұлттық контекстінде заманауи ИИ-шешімдерге қолжетімділікті төмендетеді. Осы жағдайларда қазақ тіліне арналған ұлттық LLM-ді әзірлеу стратегиялық маңызға ие болады. Жоба деректер корпусы құруға, трансформер архитектураларын бейімдеуге, арнайы токенизаторларды әзірлеуге және Retrieval-Augmented Generation (RAG) әдістерін енгізуге бағытталған, бұл тілдің морфологиялық және мәдени ерекшеліктерін ескеруге мүмкіндік береді. Жобаның аясында әдеби қорының, соның ішінде кітапхана ресурстарының цифрландырылуы ерекше мәнге ие, бұл мәдени мұраны сақтау мен цифрлық дамуын қамтамасыз етеді. Жобаның өзектілігі сонымен қатар жасанды интеллектіні білім беру, денсаулық сақтау, мемлекеттік басқару және медиа сияқты түйінді салаларға енгізу қажеттілігімен де байланысты. Қазақ тілді LLM (үлкен тілдік модель) әзірлеу аналитиканың, автоматтандырудың және пайдаланушы сервистерінің сапасын арттыруды қамтамасыз етеді. Сондай-ақ, бұл еліміздің технологиялық егемендігін нығайтуға және цифрлық ортада мемлекеттік тілдің қолданыс аясын кеңейтуге мүмкіндік береді.
Жоба мақсаты	Қазақстан Республикасында мемлекеттік тіл және мәдениетаралық қарым-қатынас тілі ретінде қазақ тілін қолдауға ықпал ететін, технологияларды дамытуға, деректер қауіпсіздігін қамтамасыз етуге, білім беруді жетілдіруге және ғылыми зерттеулерді дамытуға бағытталған заманауи үлкен тілдік модельді (LLM) құру
Жоба міндеттері	1. Үлкен тілдік модельдерді (Large Language Model, LLM) құрудың қолданыстағы әдістері мен технологияларын талдау; 2. LLM-ді оқытуға арналған деректер корпусын жинақтау және талдау; 3. LLM моделінің архитектурасын әзірлеу және оның өнімділігін оңтайландыру; 4. LLM-ді GPU кластерінде орналастыру; 5. Әзірленген тілдік модельді әртүрлі жүйелер мен платформаларға интеграциялау.
Күтілетін және қол жеткізілген нәтижелер	**1. ҚН:** Үлкен тілдік модельдерді (LLM) құрудың қолданыстағы әдістері мен технологиялары бойынша аналитикалық және техникалық есептер дайындалады. **ҚЖН:** Үлкен тілдік модельдерді (LLM) құрудың заманауи әдістері мен технологияларына кешенді талдау жүргізілді. GPT, BERT, LLaMA, T5 сияқты модельдердің архитектуралары мен экожүйелері зерттеліп, оларды түрлері, міндеттері және оқыту әдістері бойынша таксономиясы қалыптастырылды. Retrieval-Augmented Generation (RAG) тәсіліне ерекше назар

	аударылып, оның құрамдас бөліктері, шектеулері және сыртқы білім көздерінің рөлі қарастырылды. Деректерді дайындау үдерістері (тазалау, дедупликация, сүзгілеу) заманауи әдістерді (MinHash, fastText, RoBERTa) қолдану арқылы зерттелді. Қазақ тіліндегі мәтіндер корпусы әзірленіп, таңбаланды, сондай-ақ тілдің морфологиясын ескеретін арнайы стеммер жасалды. Жіктеу модельдері (LogReg, SVM, XGBoost) апробациядан өткізіліп, 98%-ға дейінгі дәлдік көрсеткішіне қол жеткізілді. LLM, іздеу жүйесі (retrieval) және деректер қоймасынан тұратын базалық инфрақұрылымы бар RAG-платформасының прототипі әзірленді. Сонымен қатар ашық LLM модельдері (OLMo, LLaMA 3/3.1) және модельдерді ауқымды түрде оқытуға мүмкіндік беретін жоғары өнімді инфрақұрылым (NVIDIA DGX H200) зерттелді. Этикалық және технологиялық аспектілерді қамтитын аналитикалық есеп дайындалды. **2. КН:** LLM-ді оқытуға арналған деректерді жинақтау және талдау жұмыстары басталады. **ҚЖН:** Көркем, ғылыми, публицистикалық және арнайы мәтіндерді қамтитын көлемі 400 млн сөзден тұратын ауқымды қазақ тіліндегі корпус қалыптастырылды. Деректерді алдын ала өңдеу әдістері әзірленді, оның ішінде нормализация, токенизация, тазалау, стемминг және қазақ тілінің агглютинативті құрылымын ескеретін толық жалғаулар жүйесі (CSE) негізіндегі морфологиялық сегментация бар. Бұл тәсіл өңдеу сапасын арттыруға және бірегей токендер санын қысқартуға мүмкіндік беріп, токенизаторлардың тиімділігін қысу коэффициенті бойынша 41–49%-ға дейін жақсартты. 139 мың сұрақ-жауап жұбын қамтитын стандартталған датасет, сондай-ақ RAG, мәтінді қысқаша мазмұндау (summarization), көпқадамды пайымдау (multi-step reasoning), қауіпсіздік, ЖИ этикасы және Қазақстанның мәдени контексі бағыттары бойынша мамандандырылған деректер жиынтықтары әзірленді. Лингвистикалық және мета-таңбалаудың автоматты және жартылай автоматты құралдары жасалып, KazBERT негізіндегі NER-таңбалаудың сапасы F1 = 0,9726 деңгейіне жетті. Kubernetes, PyTorch, Hugging Face, MinIO, MongoDB, Prometheus және Grafana технологияларын қолдана отырып, деректерді дайындау мен сақтауға арналған модульдік инфрақұрылым жобаланып, сынақтан өткізілді. Мәтіндерді автоматтандырылған түрде жинау, өңдеу және логтауға арналған скрипттер жүзеге асырылды. Алынған нәтижелер қазақ тіліндегі LLM және RAG жүйелерін одан әрі оқытуға, бейімдеуге және енгізуге арналған технологиялық әрі корпустық негіз қалыптастырды. **3. КН:** LLM моделінің архитектурасын әзірлеу және өнімділігін оңтайландыру бойынша жұмыстар басталады. **ҚЖН:** Үлкен және шағын тілдік модельдерді қазақ тіліне бейімдеу жұмыстары басталып, жүйелі түрде кеңейтілді. ISSAI платформасында ашық модельдер орнатылып, тестіленді, оның ішінде LLaMA 3 70B моделі бар, сондай-ақ сапа эталонын қалыптастыру мақсатында GPT-4, Claude және Gemini сияқты жабық жүйелерге салыстырмалы талдау жүргізілді. Қазақ тіліндегі корпустар негізінде көлемі 1 млрд токенге дейін алдын ала оқыту жүзеге асырылып, бұл тілдік бейімделу жылдамдығы мен оқытудың тұрақтылығын бағалауға мүмкіндік берді.

	Негізгі нәтижелердің бірі – қазақ тіліне бейімделген токенизаторды әзірлеу болды. Ол сөздік құрамды ішінара алмастыру және кеңейту арқылы орташа бір сөзге шаққандағы токен санын шамамен 7-ден 2-ге дейін қысқартуға мүмкіндік берді, нәтижесінде оқыту мен инференс тиімділігі айтарлықтай артты. Есептеулерді жеделдету мақсатында packing техникасы енгізіліп, өткізу қабілеті ұлғайып, есептеу шығындары азайтылды. 1B, 8B және 70B көлеміндегі модельдер оңтайландырылды, оның ішінде PTQ/QAT, GPTQ, SmoothQuant, DoReFa, Brevitas сияқты кванттау әдістері қолданылды. Бұл жадты тұтынуды 45–70%-ға қысқартуға және инференс жылдамдығын 1,5–2,2 есеге арттыруға мүмкіндік берді. Әзірленген конфигурациялар жергілікті және мобильді ортада орналастыруға бейімделген, галлюцинация деңгейін төмендетуге және дәлдікті арттыруға ықпал етеді, сондай-ақ қазақ тіліндегі LLM, SLM және мультимодальды жүйелерді одан әрі дамытуға негіз қалайды. **4. ҚН:** LLM-ді GPU кластерінде орналастыру бойынша жұмыстар басталады. **ҚЖН:** Жоғары өнімді және ауқымды түрде кеңейтілетін есептеу инфрақұрылымын құру бойынша кешенді жұмыстар орындалды. NVIDIA DGX H200 (8×H100, NVLink 4.0) базасындағы жүйе, сондай-ақ NVIDIA L40 GPU-ымен жабдықталған HYPERPC FORCE сервері орналастырылып, тестіленді. Бұл жоғары есептеу тығыздығын және жүктемелерге төзімділікті қамтамасыз етті. Бағдарламалық орта (Ubuntu Server, CUDA, NVIDIA драйверлері), деректерді сақтау жүйелері (RAID, MinIO) және VPN арқылы қорғалған қолжетімділік бапталды. Параллелдеу әдістеріне (тензорлық, модельдік, конвейерлік) зерттеу жүргізіліп, гибридті тәсілдің ең тиімді екені анықталды. DeepSpeed, Megatron-LM және Colossal-AI құралдарын қолдану арқылы тензорлық шардтау, 2D/3D-параллелизм және гибридті оптимизаторлар іске асырылды. Бұл қосымша шығындарды азайтып, GPU жүктемесін арттыруға мүмкіндік берді (оңтайлы микробатч b=4). INT8/INT4, GPTQ, SmoothQuant және QAT сияқты кванттау әдістері әзірленіп, енгізілді, нәтижесінде жад тұтынуы 45–70%-ға қысқарып, инференс жылдамдығы 2,2 есеге дейін артты. Kubernetes, Volcano және GPU Operator негізінде тапсырмаларды жоспарлау жүйесі іске асырылып, жүктемені теңгеру және көппайдаланушылық режим қамтамасыз етілді. Қосымша түрде мониторинг жүйелері (Prometheus, Grafana), JupyterHub ортасы және GPU температураларын бақылау тетіктері енгізілді. Алынған нәтижелер инфрақұрылымның тұрақтылығын, ауқымдалуын және LLM-дерді оқыту мен орналастыруға толық дайын екенін дәлелдейді. **КН:** Әзірленген тілдік модельді әртүрлі жүйелер мен платформаларға интеграциялау жүзеге асырылады. **ҚЖН:** Retrieval-Augmented Generation (RAG) архитектурасына негізделген қолданбалы жүйелерге тілдік модель орналастырылып, енгізілді. Толық функционалды RAG-платформа іске асырылды, оған фронтенд (React) және бекенд (FastAPI) кіреді, бұл деректерді жүктеу, семантикалық іздеу және диалогтық өзара әрекеттесуді қамтамасыз етеді. Пайдаланушы интерфейсі Upload, Search және

	Chat сценарийлерін қолдап, жүйенің функционалына ыңғайлы қолжетімділікті қамтамасыз етеді. Бекенд оқиғаларға негізделген архитектурада Celery және Redis құралдарын пайдалана отырып құрылған, бұл сұрауларды асинхронды өңдеуді және жоғары ауқымдалуды қамтамасыз етеді. /upload, /search және /chat негізгі эндпоинттерін қамтитын REST API жүзеге асырылып, құжаттарды өңдеу, эмбеддингтерді генерациялау және интеллектуалды іздеуді қолдайды. Векторлық ұсынуларды сақтау үшін әртүрлі embedding модельдерін (OpenAI, KazRoberta) қолдайтын Milvus базасы интеграцияланды. KazQAD-Retrieval (~6000 сұрақ) датасетін пайдалана отырып жүйеге кешенді тестілеу жүргізілді, нәтижесінде іздеу уақыты ~50 мс-тен ~20 мс-ге дейін қысқарып, нәтижелердің өзектілігі айтарлықтай артты. Reranking (LaBSE, MiniLM), кэштеу (80%-ға дейін hit rate) және логтау механизмдері енгізілді. Жүйе тұрақтылық, ауқымдылық танытып, QA-жүйелер, аналитикалық платформалар және қазақ тіліндегі AI-ассистенттер сияқты интеллектуалды сервистерге интеграциялауға дайын екенін көрсетті.
Зерттеу тобы мүшелерінің аты-жөні, идентификаторлары (Scopus Author ID, Researcher ID, ORCID, бар болса) және сәйкес профильдерге сілтемелер	1. Мансурова Мадина Есимхановна к.ф.-м.н.,профессор h-index Scopus: 8, WOS Researcher ID: PBN-6811-2025, ORCID: 0000-0002-9680-2758, Scopus Author ID: 56617164900 2. Түймебаев Жансейіт Қансейітұлы, д.фил.н., профессор h-index Scopus: 3, WOS Researcher ID: MTU-2794-2025, ORCID: 0000-0001-5495-1686, Scopus Author ID: 58785893400 3. Оспан Әсел Ғалымжанқызы магистр, старший преподаватель кафедры искусственного интеллекта и Big Data, h index 2, WOS Researcher ID: KLD-3000-2024, ORCID: 0000-0002-1860-6997, Scopus Author ID: 57238489800 4. Тұрар Олжас Нұрқонысұлы h-index Scopus: 4, WOS Researcher ID: OS-6918-2024, ORCID: 0000-0002-6720-0045, Scopus Author ID: 56523413700 5. Сарсембаева Талшын Сағдатбекқызы старший преподаваетль кафедры искусственного интеллекта и Big Data, h index 5, WOS Researcher ID: 2 MGB-4560-2025, ORCID: 0000-0001-7668-2640, Scopus Author ID: 57224454827 6. Бопеев Тимур Маратович, магистр, преподаватель кафедры искусственного интеллекта и Big Data, h-index: 2, ResearcherID: EOS-2566-2022, ORCID: 0009-0009-9813-6953, Scopus Author ID: 55960187200 7. Елубаев Сулеймен Актлеуович, магистр, заведующий лабораторией «Интернет вещей» кафедры искусственного интеллекта и Big Data, h-index: 2, ResearcherID: EIR-4189-2022, ORCID: 0009-0005-0758-9406, Scopus Author ID: 6504255748 8. Байпакбаева Салтанат Туркестанкызы, магистр, h-index: 5, ResearcherID: CEK-0233-2022, ORCID: 0000-0001-7251-0585, Scopus Author ID: 57208569081

	9. Тукеев Уалшер Ануарбекович - д.т.н., профессор, h-index Scopus: 7, WOS Researcher ID:AAR-3668-2020, ORCID: 0000-0001-9878-981X, Scopus Author ID: 55701639900 10. Тойганбаева Назгуль Абеновна, PhD, преподаватель, кафедры искусственного интеллекта и Big Data, h-index Scopus: 1, WOS Researcher ID O-2399-2014, https://orcid.org/0000-0003-2661-8661, Scopus ID 56069463400 11. Аман Мұса, магистр, преподаватель, кафедры искусственного интеллекта и Big Data, h-index Scopus: 1, WOS Researcher ID: OLP-9019-2025, ORCID: 0009-0001-9972-7677, Scopus Author ID: 60085728100 12. Базарбек Жания Пархатқызы, магистр, старший преподаватель кафедры искусственного интеллекта и Big Data 13. Рахимова Диана Рамазановна - PhD, доцент исследователь, h-index Scopus: 7, WOS Researcher ID: D-8421-2012, ORCID: 0000-0003-1427-198X, Scopus Author ID: 55682794500 14. Шормакова Асем Ноябревна - PhD, и.о.доцента, h-index Scopus: 3, WOS Researcher ID: LNF-7076-2024, ORCID: 0000-0002-1637-4643, Scopus Author ID: 55786942400 15. Байдулла Аяулым Мейірханқызы магистр технических наук, h-index Scopus: 0, WOS Researcher ID:PJB-9416-2026, ORCID:0009-0002-3083-9715, Scopus Author ID: - . 16. Бажибаева Айнур Бахытбековна магистр, h-index Scopus: 0, WOS Researcher ID:OTH-2331-2025,ORCID: 0009-0009-5432-5795, Scopus Author ID: - . 17. Тешебаев Нұрдаулет Мұратбекұлы магистр, h-index Scopus: 0, WOS Researcher ID: GYA-0067-2022, ORCID:0000-0002-6560-6578, Scopus Author ID: 60405426200 18. Абдиахметова Зухра Муратовна, h-index Scopus: 6, WOS Researcher ID: MEO-9276-2025, ORCID:0000-0001-8702-511X, Scopus Author ID: 57193647310 19. Толқын Өрден, магистр,преподаватель кафедры искусственного интеллекта и Big Data,h-index 0, ORCID: 0009-0007-6501-4613, scopes ID: 59563392300 20. Жайсанова Динара Сайлауовна, PhD, и. о доцента кафедры искусственного интеллекта и Big Data, h index 2, WOS Researcher ID: R-4415-2017, ORCID: 0000-0002-8116-6111, Scopus Author ID: 57204395807 21. Солтангельдинова Мадина Кызырханкызы магистр технических наук, h-index Scopus 0, Web of Science ResearcherIDPSM-3693-2026, ORCID: 0009-0008-7263-2337, Scopus Author ID: - 22. Әбдіманап Ғалымжан Сейтахметғалиұлы магистр, программист, h-index Scopus: 4 , WOS Researcher ID: AFH-9860-2022, ORCID: 0000-0003-1676-4075, Scopus Author ID: 57219870474

	23. Amirhosein Mosavi h-index Scopus: 83 , WOS Researcher ID :I-7440-2018, ORCID:0000-0003-4842-0613, Scopus Author ID: 57191408081 24. Ян Ченхан, Кандидат технических наук, аспирант кафедры искусственного интеллекта и больших данных, h-index Scopus: 3 , WOS Researcher ID:PHF-3684-2026, ORCID: 0000-0003-0997-6280, Scopus Author ID: 59377116300 25. Чжао Байдун, Кандидат технических наук, аспирант кафедры искусственного интеллекта и больших данных, h-index Scopus: 3 , WOS Researcher ID: OMK-5686-2025, ORCID: 0009-0002-5572-7306, Scopus Author ID: 58924501700 26. Қадырбек Нұрғали h index 5, WOS Researcher ID: AEK-7760-2022, ORCID: 0000-0002-5461-8899, Scopus Author ID: 57220749433 27. Дарибаев Беимбет Серикович, h index 6, WOS Researcher ID: AAS-2995-2020, ORCID: 0000-0003-1313-9004, Scopus Author ID: 57191892577 28. Перепёлкин Владислав Александрович6 h-index Scopus: 6 , Scopus Author ID: 36662921000 29. Тұрсынбек Бақытжан 30. Peng Wang, h index 57, WOS Researcher ID: S-5603-2019, Scopus Author ID: 57309244300 31. Фу Ян Вэй6 h index 52, WOS Researcher ID: KHA-1853-2024, ORCID: 0000-0002-6595-6893, Scopus Author ID: 57213791184 32. Varol, Huseyin Atakan, Scopus Author ID: 24484134500, https://orcid.org/0000-0002-4042-425X
Қолдану	Әзірленген LLM және RAG-платформа қазақ тіліндегі деректерді талдау және өңдеуге арналған интеллектуалды жүйелерді құру үшін қолданылады. Бұл шешім білім беру саласында (курс ұсыныстары жүйесі, MOOC-платформалар, мектеп пен жоғары оқу орындарын қолдау жүйелері), журналистикада (AI-QazMedia, жаңалықтарды талдау, мазмұнды қысқаша жинақтау және кестелерді генерациялау), мемлекеттік секторда (нормативтік ақпаратты іздеу, деректер аналитикасы), сондай-ақ ғылыми зерттеулерде (корпустарды өңдеу, білімдер мен онтологияларды құру) қолданылуы мүмкін. Жобаның аясында Қазақ ұлттық университетінің кітапханасындағы қазақ тіліндегі әдеби қорының цифрландыру жұмыстары басталып, жалғасуда. Бұл деректер корпусын айтарлықтай кеңейтіп, мәдени мұраны цифрлық форматта сақтауға мүмкіндік береді. Құрылған ИИ-агенттер кең қолдануға бағытталған: денсаулық сақтау саласында (дәрігерлер мен пациенттерді қолдау үшін интеллектуалды ассистенттер), абитуриенттер үшін (білім беру бағдарламалары бойынша кеңес беру), журналистер үшін (аналитика және контент генерациясы), сондай-ақ білім беру жүйесінде. Платформа веб-қосымшалар мен корпоративтік жүйелерге API арқылы интеграциялауды қолдайды, семантикалық іздеуді, диалогтық интерфейстерді және аналитикалық тапсырмаларды автоматтандыруды қамтамасыз етеді. Ерекше мәні — қазақ тілі мен

	мәдени контекстке бейімделуі, бұл заманауи ИИ-технологияларға қолжетімділікті кеңейтеді.
Нақты аналогтармен салыстырғанда басымдылық	Әзірленген жүйе GPT-4, Claude және Gemini сияқты жетекші коммерциялық шешімдерге қарағанда айтарлықтай артықшылықтарға ие. Бұл модельдер негізінен жаһандық тілдер мен әмбебап сценарийлерге бағытталған болса, ұсынылған LLM арнайы қазақ тілін, оның агглютинативтік құрылымын және мәдени контекстін ескере отырып әзірленген, бұл локалдық тапсырмаларда жауаптардың дәлдігі мен өзектілігін айтарлықтай арттырады. Негізгі артықшылықтардың бірі – бейімделген токенизатор мен морфологиялық сегменттеу әдістерін қолдану. Бұл токендер санын айтарлықтай азайтып, оқыту мен инференс тиімділігін арттырады. Бұл әсіресе ресурсы шектеулі тілдер үшін маңызды, себебі стандартты токенизаторлар тиімділігі төмен бола береді. Жабық API-шешімдерден айырмашылығы, әзірленген платформа ашық модельдерге негізделген және жергілікті орналастыруды қолдайды, бұл деректерді бақылауға, қауіпсіздік талаптарына сәйкестікті қамтамасыз етуге және сыртқы провайдерлерге тәуелсіз болуға мүмкіндік береді. Retrieval-Augmented Generation (RAG) архитектурасын интеграциялау өз білім базаларын, соның ішінде цифрланған ресурстарды пайдалануға мүмкіндік береді, бұл нәтижелердің сенімділігін және түсініктілігін арттырады. Қосымша басымдық – шектеулі есептеу ресурстарына бейімделу: кванттау, гибридтік параллелдеу және edge-құрылғыларды қолдау әдістері енгізілген, бұл жүйені кең ауқымды ұйымдар үшін қолжетімді етеді. Сондай-ақ, ерекше қасиеті — ұлттық корпустар мен цифрландыру жобаларымен (мысалы, ҚазҰУ қорлары) интеграция, бұл қазақ тіліндегі ИИ-сервистердің тұрақты экожүйесін қалыптастырып, технологиялық егеменділікті нығайтады.
Жарияланымдар тізімі (URL, DOI көрсетілген)	1. Omarova S., Ospanova D., Aitova N., Tokenkyzy G., Ormanova A., Alshynbekova M. A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh // Forum for Linguistic Studies. – 2025. – Vol. 7, No. 2. – P. 869–881. – DOI: https://doi.org/10.30564/fls.v7i2.8317. (Q2, процентиль ≈ 55%, SJR 2024: 0.234) 2. Kadyrbek N., Tuimebayev Z., Mansurova M., Viegas V. The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh and Direct Preference Optimization // Big Data and Cognitive Computing. – 2025. – Vol. 9, No. 5. – Art. 137. – DOI: https://doi.org/10.3390/bdcc9050137. (Q1, процентиль ≈ 78%, SJR 2024: 0.820) 3. Искакова Ж., Демессинова Л., Зинедина А., Есбосын Л., Оспанова Д., Жұмабай М., Ербатырова Ж. The

	Significance of the Lexical-Grammatical Minimum in Teaching Kazakh at the A2 Level // Theory and Practice in Language Studies. 2025. Vol. 15, No. 10. P. 3477–3488. DOI: 10.17507/tpls.1510.36. URL: https://tpls.academypublication.com/index.php/tpls/article/view/11035/9185 (Q1) 4. Күмісбек М.Н., Шормакова А.Н. Зиянды спам-хабарламаларды тиімді анықтауға арналған машиналық оқыту алгоритмдерін зерттеу және салыстырмалы талдау // Вестник Торайгыров университета. Серия: физика, математика и компьютерные науки. – Выпуск №2, 2025. – DOI: https://doi.org/10.48081/PJFQ3449. (ККСОН) 5. Рахимова Д., Жігер А., Малых В., Карюкин В., Бекарыстанқызы А. Ағылшын-қазақ тілі жұбы үшін нейрондық машиналық аударма // Вестник КБТУ. – Выпуск №2(73), 2025. – DOI: https://doi.org/10.55452/1998-6688-2025-22-2-54-66. (ККСОН) 6. Рахимова Д., Жігер А., Малых В., Карюкин В., Бекарыстанқызы А. Тілдер арасындағы аударма сапасын арттыру: ағылшын-қазақ тілдеріндегі жетістіктер мен мүмкіндіктер // Вестник Академии гражданской авиации. Серия: Компьютерные науки, приборостроение и автоматизация. – Выпуск №2 (37), 2025. – DOI: https://doi.org/10.53364/24138614_2025_37_2_12. (ККСОН) 7. Zhadil Y., Sarsembayeva T., Mansurova M., Duyssekeeva B. An AI-Driven Smart Digital Assistant Bot for Improving Student Support Services // 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST). – May 14–16, 2025. Astana, Kazakhstan. – DOI: 10.1109/SIST61657.2025.11139316. (Scopus Proceedings) 8. Assel Ospan, Aman Mussa, Madina Mansurova, Talshyn Sarsembayeva. LLM Agents for Enhanced Tabular Data Interpretation: A Perspective // 2025 IEEE 5th International Conference on Smart Information Systems and Technologies (SIST). – May 14–16, 2025. Astana, Kazakhstan. – DOI: 10.1109/SIST61657.2025.11139242. (Scopus Proceedings) 9. Mussa A., Mansurova M. Parameter-Efficient Fine-Tuning of LLaMA 2 for the Kazakh Language: Advancing Low-Resource Language Models // 10th International Congress on Information and Communication Technology. – February 18–21, 2025. –

	DOI: https://doi.org/10.1007/978-981-96-6935-6_42. 10. Mussina A., Aubakirov S., Trigo P., Mansurova M. From What-If Scenarios to Event Associations: A Novel Approach to Social Media Event Analysis // 14th International Conference on Data Science, Technology and Applications (DATA 2025). – June 10–12, 2025. Bilbao, Spain. – DOI: 10.5220/0013645900003967. (Scopus Proceedings) 11. Мансурова М., Леонидас С., Шормакова А., Кумисбек М. Enhancing Spam Detection with Large Language Models and Hybrid Learning Approaches // 34th European Conference on Operational Research (EURO 2025). – Accepted. – Session: MC-43: Digital Philosophy and OR. 12. Aman Mussa, Zhanseit Tuimebayev, Madina Mansurova. Make Large Language Models Efficient: A Review // IEEE Access. – 2025. – Vol. 13. – DOI: 10.1109/ACCESS.2025.3605110. (Q1, процентиль = 86%, SJR 2024 = 0.849) 13. Ж.П. Базарбек, Н.А. Тойганбаева, М.Е. Мансурова, Т.С. Сарсембаева, М.Ж. Сакыпбекова. Қазақ тіліне арналған үлкен тіл моделін (LLM) жасау үшін dataset әзірлеу // Academic Scientific Journal of Computer Science. – №3, 2025. – DOI: 10.32014/2025.2518-1726.365. (ККСОН) 14. Assel Ospan, Kanat Auyesbay, Talshyn Sarsembayeva, Aman Mussa. Application of Rule-Based Method for Automatic Extraction of Tags from Column-Style PDF-Documents // Journal of Problems in Computer Science and Information Technologies. – №3(3), 2025. – DOI: https://doi.org/10.26577/jpcsit2025336. (ККСОН) 15. Diana Rakhimova, Assem Shornakova, Madina Mansurova. Data Collection and Processing for Large Language Models in Low-Resource Languages: A Case Study on Kazakh // Международный форум «Русистика и тюркология: литературные традиции, культурный трансфер и перевод», 2025. 16. Diana Rakhimova, Aliya Zhiger, Madina Mansurova, Valentin Malykh, Magzhan Kairanbay. Improving Post-Editing of Kazakh Translations with Fine-Tuned Large Language Models: Dataset and Evaluation // International Journal of Innovative Research and Scientific Studies. – 2025. – Vol. 8, No. 8. – P. 220–233. – DOI: 10.53894/ijirss.v8i8.10583. (Q2) 17. Құрметқан Т., Мамырбаев Ө.Ж. Қазақ сөйлеулерін тануда transformer моделінің жетілдірген түрлерін қолдану ерекшеліктері // ШҚТУ хабаршысы. – №2, 2025. – P. 163–171. – DOI: 10.51885/1561-4212_2025_2_163.

	18. Z. Makhataeva, H.A. Varol. Evaluation of Typing Speed, User Experience, and Cognitive Load Across Kazakh, Russian, and English Languages Among Kazakhstani Users // Казахский национальный университет имени Аль-Фараби. Вестник. Серия «Педагогические науки». – №2(83), 2025. – DOI: https://doi.org/10.26577/JES202583214. (ККСОН)
	19. Tókenqyzy G., Omarova S., Qadırkhan A. K. Қазіргі телебағдарламалар тілі (Қазақ тілі ұлттық корпусының кіші корпустары негізінде) // Вестник Евразийского гуманитарного института / Bulletin of the Eurasian Humanities Institute. 2025. № 2. С. 78–84. DOI: 10.55808/1999-4214.2025-2.09. URL: https://doi.org/10.55808/1999-4214.2025-2.09 (ККСОН)
	20. Omarova S. K., Iskakova Zh. M., Ormanova A. A., Abilmazhinova B. Zh., Yerbatyrova Zh. K. Қазақ тілін үйретудің корпустық әдісі (A corpus-based method for teaching the Kazakh language) // *Pedagogical University Bulletin*. 2025. № 3. С. 287–296. DOI: 10.48081/UKFQ5345. URL: https://doi.org/10.48081/UKFQ5345 (ККСОН)
	21. Makhataeva Z., Atymtay N., Meiramov R., Nauryzbaikyzy G., Sadirova K., Varol H. A. Multilingual Automatic Speech Recognition Interface for Typing: Usability Study and Performance Evaluation for Kazakh, Russian, and English // *Scientific Journal of Astana IT University*. 2025. № 24. С. XX–XX. ISSN 2707-9031, eISSN 2707-904X. DOI: 10.37943/24AHNP6638. (Принято к печати.) (ККСОН)
	Авторское свидетельство:
	1. Жәділ Е. Н., Сарсембаева Т. С. AIFarabi. – № 56242 от 1 апреля 2025 года.
	2. Әбдіманап Ғ. С., Тойганбаева Н. А. Программный комплекс для оптического распознавания текста (OCR) с поддержкой казахского языка. – № 61658 от 25 августа 2025 года.
	3. Рахимова Д. Р., Шормакова А. Н., Алиев Р. Ш. Автоматическая лингворазметка казахских текстов для обучения больших языковых моделей на основе NER. – № 62715 от 6 октября 2025 года.
	4. Кәрібаева А. С., Карюкин В. И. Автоматическая мета-разметка веб-данных для LLM казахского языка. – № 62786 от 8 октября 2025 года.
	5. Жанегизов А. С., Мамырбаев О. Ж., Нарынов С. С. RAG платформасының базасына векторларды сақтау қондырғысы мен API қабаттары. – № 60942 от 22 июня 2025 года.

	Жанегизов А. С., Мамырбаев О. Ж., Нарынов С. С. RAG платформасының базасына векторларды сақтау қондырғысы мен API қабаттары. – № 61054 от 28 июля 2025 года.
Патент туралы ақпарат	