Краткая информация о проекте

Наименование	BR24993001 «Создание большой языковой модели (LLM) для поддержки казахского языка и технологического прогресса»
Актуальность	Реализация программы будет способствовать укреплению статуса казахского языка как основополагающего элемента государственной идентичности, культурного наследия и межкультурного взаимодействия. Важным аспектом программы является создание первого в Казахстане высокопроизводительного дата-центра, предназначенного для решения задач искусственного интеллекта. Этот дата-центр будет функционировать как лаборатория коллективного пользования, предоставляя возможности для исследовательского сообщества страны.
Цель	Создание современной большой языковой модели (LLM), способствующей поддержке казахского языка, как государственного и языка межкультурного общения в РК, развитию технологий, обеспечению безопасности данных, развитию образования и совершенствованию научных исследований.
Задачи	1. Анализ существующих методов и технологий создания больших языковых моделей (Large language model). Решение задачи играет ключевую роль в уточнении и обновлении дальнейшей стратегии реализации Программы. Обновление стратегии возможно только, если будут открыты принципиально новые подходы к разработке LLM. 2. Сбор и анализ корпуса данных для обучения LLM. Решение задачи напрямую влияет на основной результат Программы, что включает качество, легитимность и этичность разрабатываемой LLM для казахского языка. 3. Разработка архитектуры и оптимизация производительности LLM модели. Решение задачи является ключевым в целях обеспечения производительности разрабатываемой LLM, а также в целях ее долгосрочного использования. 4. Развертывание LLM на кластере GPU. Решение задачи является ключевым в целях обеспечения производительности разрабатываемой LLM и в целях ее долгосрочного использования. 5. Интеграция разработанной языковой модели в различные системы и платформы. Решение задачи является критически важным для достижения цели Программы, обеспечивая широкое применение ИИ на

	казахском языке в различных сферах экономики и госуправления.
Ожидаемые и достигнутые результаты	1. Создание прототипа Retrieval-Augmented Generation (RAG) платформы и запуск пилотной версии модели-ассистента на архитектуре LLaMA (1,94 млрд параметров). 2. Разработка и развёртывание базовой инфраструктуры для взаимодействия компонентов LLM, retrieval-системы и хранилища данных. 3. Сбор и формирование корпуса казахоязычных текстов объёмом 400 миллионов слов, охватывающего широкий спектр жанров, стилей и тематики. 4. Предобработка текстовых данных: очистка, нормализация, аннотация и подготовка к дальнейшей лингвистической разметке. 5. Начало разработки архитектуры LLM-моделей с открытым исходным кодом (8 и 70 млрд параметров) с последующей оптимизацией производительности. 6. Исследование и реализация гибридного подхода к параллелизации обучения LLM (тензорная, модельная и конвейерная схемы). 7. Разработка решений по распределению вычислительных нагрузок и повышению эффективности использования GPU-кластера. 8. Подготовка научных публикаций и формирование методологической базы для дальнейшего масштабирования проекта.
Имена и фамилии членов исследовательской группы с их идентификаторами (Scopus Author ID, Researcher ID, ORCID, при наличии) и ссылками на соответствующие профили	Мансурова Мадина Есимхановна – научный руководитель проекта, к.ф.-м.н., доцент, заведующая кафедрой искусственного интеллекта и Big Data КазНУ им. аль-Фараби, ведущий научный сотрудник. Область научных интересов: высокопроизводительные вычисления, NLP, искусственный интеллект. Автор более 90 научных статей, 4 монографий, 2 учебников с грифом МОН РК. Scopus ID: 56617164900, Researcher ID: отсутствует, ORCID ID: 0000-0002-9680-2758. Индекс Хирша: Scopus – 6, Web of Science – 4. Тукеев Уалихан Абдикерович – исполнитель, доктор технических наук, профессор КазНУ им. аль-Фараби. Специалист по формальным грамматикам и машинному переводу, участник и руководитель множества грантовых и международных проектов, более 20 лет опыта. Scopus ID: 55701639900, Researcher ID (WoS): AAR-3668-2020, ORCID ID: 0000-0001-9878-981X. Индекс Хирша: 5.

	Рахимова Динара Рыскалиевна – исполнитель, PhD, старший преподаватель КазНУ. Участвует в разработке инновационных методов предобработки текстов, координатор магистерской программы по вычислительной лингвистике. Автор 27 публикаций, цитируемых в Scopus. Scopus ID: 57191430929, Researcher ID: P-4962-2017, ORCID ID: 0000-0003-0140-4181. Индекс Хирша: 4. Кәрібаева Аружан Сериковна – исполнитель, PhD, старший преподаватель КазНУ. Специализируется на морфологическом анализе, имеет опыт участия в международных проектах и стажировке в Эдинбурге. Scopus ID: 57196004542, ORCID ID: 0000-0002-2023-1573. Индекс Хирша: 3. Шормакова Айжан Нурлановна – исполнитель, магистр, старший преподаватель КазНУ, докторант PhD. Занимается LLM и NLP для казахского языка. Scopus ID: 55786942400, ORCID ID: 0000-0002-1637-4643. Индекс Хирша: 2. Карюкин Владислав – исполнитель, PhD, старший преподаватель кафедры информационных систем КазНУ им. аль-Фараби. Специализируется на предобработке текстов на казахском языке. Scopus ID: 57218952479, ORCID ID: 0000-0002-8768-0349. Индекс Хирша: 2. Жуманов Жандос Муратович – младший научный сотрудник, магистр, программист в ТОО "Alem Research". Специалист в области синтаксического анализа, автоматического распознавания и синтеза речи. Scopus ID: 55701348900, Researcher ID: B-2733-2015, ORCID ID: 0000-0002-0395-9730. Индекс Хирша: 4. Дүйсенбекқызы Жансая – младший научный сотрудник, магистр технических наук, докторант 1 курса КазНУ по специальности «Информационные системы». Занимается распознаванием казахской речи у детей. Researcher ID: JMQ-5368-2023, ORCID ID: 0000-0001-7743-0795. Нурмаганбет Даурен – исполнитель, техник, магистр 1 курса по направлению «Вычислительная лингвистика» КазНУ им. аль-

	Фараби. Участвует в разработке ПО для синтаксического анализа. Алиев Р. – исполнитель, техник, студент КазНУ по специальности «Информационные системы». Участвует в разработке ПО для морфологического анализа. Scopus ID: 00000000000, Researcher ID: JPA-3963-2023, ORCID ID: 0009-0003-5432-0480. Индекс Хирша: 0. Абдиахметова Зухра Муратовна – исполнитель, PhD in Computer Sciences, магистр, и.о. доцента НАО «КазНУ им. аль-Фараби». Специалист в области обработки данных, машинного обучения и анализа больших данных. Активно участвует в исследованиях и разработке решений в сфере предобработки текстов и создания Цифрового двойника. Scopus ID: 57193647310, ORCID ID: 0000-0001-7158-0995. Индекс Хирша: 4. Жайсанова Динара Сайлауовна – исполнитель, PhD, старший преподаватель кафедры «Искусственный интеллект и Big Data» НАО «КазНУ им. аль-Фараби». Участвует в экспериментах по оптимизации архитектуры LLM для казахского языка, анализирует и документирует результаты. Scopus ID: 57204395807, ORCID ID: 0000-0002-8116-6111, Researcher ID: CJV-4009-2022. Индекс Хирша: 2. Исабаева Даража Нагашыбаевна – исполнитель, кандидат педагогических наук, ассоциированный профессор, и.о. профессора кафедры «Искусственный интеллект и Big Data» КазНУ. Специалист в области применения машинного обучения в образовании, оптимизации преподавания. Scopus ID: 56366568600, ORCID ID: 0000-0002-9979-3121, Researcher ID: EZV-5601-2022. Индекс Хирша: 3. Қадырбек Нұрғали Қазбекұлы – исполнитель, магистр, старший научный сотрудник КазНУ им. аль-Фараби. Участвует в улучшении архитектуры LLM, тестировании и оптимизации производительности модели. Scopus ID: 57220749433, ORCID ID: 0000-0002-5461-8899, Researcher ID: AEK-7760-2022. Индекс Хирша: 3. Қайратұлы Бауыржан – исполнитель, магистр, докторант 1 курса КазНУ по специальности «Наука о данных». Занимается созданием и

	разметкой корпусных данных для обучения и тестирования LLM. Scopus ID: 00000000000, ORCID ID: 0000-0002-0341-8899. Индекс Хирша: 0. Қырғызбаева Маржан Ескендерқызы – исполнитель, магистр, докторант 3 курса по специальности «8D06103 – Компьютерные науки» КазНУ им. аль-Фараби. Участвует в разработке LLM, выборе и интеграции архитектур, обучении модели и оценке её эффективности. Scopus ID: 57220744243, ORCID ID: 0009-0004-5333-3252. Индекс Хирша: 2. Оспан Әсел Ғалымжанқызы – исполнитель, магистр, старший преподаватель НАО «КазНУ им. аль-Фараби». Оптимизирует производительность LLM, устраняет узкие места, внедряет улучшения и проводит тестирование. Scopus ID: 57238489800, ORCID ID: 0000-0002-1860-6997. Индекс Хирша: 2. Сарсембаева Талшын Сағдатбекқызы – исполнитель, докторант НАО «КазНУ им. аль-Фараби». Разрабатывает систему оценки качества LLM, метрики, процедуры тестирования и анализирует результаты. Scopus ID: 57224454827, ORCID ID: 0000-0001-7668-2640. Индекс Хирша: 3. Шоманов Адай Сакенович – исполнитель, PhD, инструктор Назарбаев Университета. Исследует методы безопасности LLM, разрабатывает защитные механизмы, проводит аудиты и тестирование на уязвимости. Scopus ID: 57195543732, ORCID ID: 0000-0001-8253-7474, Researcher ID: ABG-7891-2021. Индекс Хирша: 5. Жиенбаев Мейран Муратулы – исполнитель, докторант КазНУ им. аль-Фараби. Работает над дообучением LLM для специализированных чат-ботов, сбором и подготовкой данных, оценкой и оптимизацией моделей. ORCID ID: 0009-0005-0497-0965. Индекс Хирша: 0. Байбурин Ержан Мухаметкалиевич – исполнитель, научный сотрудник лабораторий цифровых технологий и моделирования. Разрабатывает инновационные методы

	взаимодействия с LLM, исследует интерфейсы и оценивает их эффективность. Scopus ID: 56111999400, ORCID ID: 0000-0002-1583-9912, Researcher ID: GRT-5436-2022. Индекс Хирша: 4. Абешев Куаныш Шурабатырович – исполнитель, PhD in Mathematics, ассоциированный профессор, декан Школы цифровых технологий, Алматы Менеджмент Университет. Эксперт в Data Science, Machine Learning, Computer Vision. Преподаёт курсы Machine Learning и Computer Vision, член Association for Symbolic Logic (ASL). Scopus ID: 56124436100, ORCID ID: 0000-0003-1140-7431, Researcher ID: N-9739-2014. Индекс Хирша: 5. Тұрар Олжас Нұрқонысұлы – исполнитель, декан факультета информационных технологий НАО «КазНУ им. аль-Фараби». Специалист по математическому моделированию и машинному обучению. Занимается архитектурой LLM на GPU-кластере. Scopus ID: 56523413700, ORCID ID: 0000-0002-6720-0045. Индекс Хирша: 4. Дарибаев Беимбет Серикович – исполнитель, заведующий кафедрой НАО «КазНУ им. аль-Фараби». Специалист по параллельным алгоритмам, более 10 лет опыта в НИР. Отвечает за выбор и настройку GPU-оборудования. Scopus ID: 57191892577, ORCID ID: 0000-0003-1313-9004. Индекс Хирша: 4. Лебедев Данил Владимирович – исполнитель, ТОО «Astana IT University». Специалист по распределённым и параллельным вычислениям, более 20 лет опыта. Исследует методы сжатия моделей (прунинг, квантование) и разрабатывает стратегию для LLM. Scopus ID: 56522634900, ORCID ID: 0000-0002-5186-6483. Индекс Хирша: 4. Ермек Алимжанов – исполнитель, КазНУ им. аль-Фараби. Специалист по автоматизации и распределённым вычислениям, более 10 лет опыта. Отвечает за разработку и отладку скриптов развертывания ПО на GPU-кластере. Scopus ID: 57191433356, ORCID ID: 0000-0002-8758-2220. Индекс Хирша: 2.

	Мәткерім Базаргүл – исполнитель, старший преподаватель НАО «КазНУ им. аль-Фараби». Специалист по распределённым системам, разрабатывает модель планирования ресурсов на кластерах GPU. Scopus ID: 55898795100, ORCID ID: 0000-0002-5336-687X. Индекс Хирша: 3. Аубакиров Санжар – исполнитель, Казахский национальный университет имени аль-Фараби. Специалист по разработке Web/мобильных приложений и архитектуре информационных систем. Будет разрабатывать алгоритмы управления очередями задач на кластере GPU и интегрировать их с существующей инфраструктурой. Scopus ID: 57190230321, ORCID ID: 0000-0002-8416-527X. Индекс Хирша: 3. Более 15 лет опыта в научных проектах. Нурахов Едил Сергазиевич – исполнитель, директор ДИТ, ТОО «Astana IT University». Будет подготавливать инфраструктуру для загрузки и хранения дообученной LLM и настраивать её параметры для развертывания на кластере GPU. Scopus ID: 57204717628. Индекс Хирша: 3. Более 5 лет опыта в ИТ-проектах и разработке информационных систем. Кенжебек Ержан Ғалымжанұлы – исполнитель, старший преподаватель КазНУ им. аль-Фараби. Будет устанавливать инструменты мониторинга производительности GPU и разрабатывать метрики и систему алертов. Scopus ID: 57221598108, ORCID ID: 0000-0002-6492-8292. Индекс Хирша: 2. 6+ лет опыта, 15 публикаций, из них 3 в Scopus. Опыт в нейросетях и задачах нефтеотдачи. Қасымбек Нұрислам Мұратбекұлы – исполнитель, старший преподаватель КазНУ им. аль-Фараби. Будет исследовать методы адаптивного обучения и реализовывать прототип адаптивного обучения LLM на GPU-кластере. Scopus ID: 57217825275, ORCID ID: 0000-0001-5663-2267. Индекс Хирша: 1. 7+ лет опыта, 13 публикаций, 4 в Scopus. Специалист в параллельных вычислениях (MPI, OpenMP).

	Мустафин Мақсат Бейбитович – исполнитель, старший преподаватель КазНУ им. аль-Фараби. Будет анализировать архитектуру LLM и разрабатывать методы её оптимизации под ограниченные ресурсы GPU. Scopus ID: 58693756000, ORCID ID: 0000-0002-3655-0771. Индекс Хирша: 1. 6+ лет опыта, 8 публикаций (2 в Scopus). Специалист по CUDA, визуализации, OpenGL/Vulkan. Муханбет Ақсултан Айтуарұлы – исполнитель, докторант КазНУ им. аль-Фараби. Будет разрабатывать систему управления ресурсами с учетом специфики задач LLM. Scopus ID: 57214259277, ORCID ID: 0000-0003-4699-0436. Индекс Хирша: 2. 5+ лет опыта, занимается квантовыми алгоритмами и ML. Махмут Ерлан – исполнитель, преподаватель КазНУ им. аль-Фараби. Будет исследовать методы распределенного обучения и настраивать соответствующую инфраструктуру для LLM. Scopus ID: 58042934300, ORCID ID: 0009-0002-3451-415X. Индекс Хирша: 0. 5+ лет опыта. Специалист по CUDA и визуализации данных. Аетова Багдат Алтынбековна – бухгалтер проекта, КазНУ им. аль-Фараби. Обеспечивает ведение финансовой отчетности. Научных идентификаторов нет. Adali Eşref – профессор, PhD, Стамбульский технический университет (Турция). Участвует в разработке методов синтаксического и морфологического анализа текстов. Один из основателей POS-маркировки и стемминга для агглютинативных языков. Scopus ID: 14834024800, ORCID ID: 0000-0002-1561-8255. Индекс Хирша: 9. 18 лет опыта, 49 публикаций (WoS, Scopus). Участвовал в проектах по машинному переводу (турецкий–английский, туркменский–турецкий). Барахнин Владимир Борисович – доктор техн. наук, доцент, ФИЦ ИВТ (Россия, Новосибирск). Участвует в формировании корпуса, настройке LLM, разработке тестов. Scopus ID: 6508258628, ORCID ID: 0000-0003-3299-0507, WoS ID: A-5856-2014. Индекс Хирша: Scopus – 7, WoS – 4. 200+ публикаций, автор монографий и учебников.

	Руководил проектами РНФ, РФФИ. Эксперт в NLP и математическом моделировании. Амиргалиев Едилхан Несипханович – исполнитель, академик Национальной инженерной академии РК, член-корреспондент НАН РК, доктор технических наук, профессор, заведующий лабораторией искусственного интеллекта и робототехники ИИВТ КН МНВО РК. Выполняет сравнительный анализ современных LLM (GPT-3, BERT, LaMDA, BLOOM и др.) и определяет их преимущества и недостатки для обучения на казахском языке. Scopus ID: 56167524400, ORCID: 0000-0002-6528-0619, Researcher ID: C-6963-2015. Индекс Хирша: 15. Автор 8 монографий и более 350 научных публикаций, в том числе 95+ в журналах Scopus. Руководил проектами ПЦФ и ГФ МОН РК. Мамырбаев Оркен Жумажанович – исполнитель, PhD, профессор, заместитель генерального директора по науке ИИВТ КН МНВО РК. Определяет параметры обучения LLM (размер корпуса, количество токенов, GPU и др.), а также области её применения: образование, госуслуги, законотворчество, поддержка казахского языка. Scopus ID: 55967630400, ORCID: 0000-0001-8318-3794, Researcher ID: 0-1265-2017. Индекс Хирша: 11. Эксперт в распознавании речи и образов. Калижанова Алия Уалиевна – исполнитель, кандидат физ.-мат. наук, ассоциированный профессор, главный научный сотрудник ИИВТ КН МОН РК. Разрабатывает правила по фильтрации вредоносного контента в обучающих данных, отвечает за этические аспекты создания LLM. Scopus ID: 57034267000, ORCID: 0000-0002-5979-9756, Researcher ID: AAR-7411-2020. Индекс Хирша: 7. Руководитель проектов AP09259547, AP05132778 и др. Нарынов Сергазы Сакенович – исполнитель, кандидат технических наук, ведущий научный сотрудник ИИВТ КН МНВО РК. Отвечает за сбор данных, формирование корпуса на казахском языке и генерацию синтетических текстов с помощью GAN. Scopus ID: 56582431600, ORCID: 0000-0002-9611-9772.

	Индекс Хирша: 5. Руководитель проектов ГФ AP09259140, AP08855520, участник ПЦФ BR05236699-ОТ-19. Козбакова Айнур Холдасовна – исполнитель, PhD, ассоциированный профессор, ведущий научный сотрудник ИИВТ КН МОН РК. Формирует корпус данных на русском языке и генерирует синтетические тексты с помощью GAN. Scopus ID: 57195683902, ORCID: 0000-0002-5213-4882, Researcher ID: K-5077-2018. Индекс Хирша: 6. Участие в проектах ПЦФ BR05236693, ГФ AP05132648, AP08855903 и др. Оралбекова Дина Орымбаевна – исполнитель, PhD, старший научный сотрудник ИИВТ КН МНВО РК. Формирует корпус данных на английском языке и расширяет его с помощью GAN. Scopus ID: 57226648854, Web of Science: профиль. Индекс Хирша: Scopus – 5, Web of Science – 2. Эксперт в области распознавания образов и речи. Мерембаев Тимур Жумаканович – исполнитель, магистр математики, Research Assistant, АО "Международный университет информационных технологий". Разрабатывает и обучает модели на основе трансформеров для казахского языка, учитывая лингвистические и культурные особенности. Scopus ID: 57207766879, ORCID: 0000-0001-8185-235X. Индекс Хирша: 4. Опыт более 10 лет в дистанционном зондировании, ГИС и БД. Черикбаева Ляйля Шариповна – исполнитель, PhD, магистр информатики, и.о. доцента РГП на ПХВ "КазНУ им. аль-Фараби". Фиксирует основные риски генерации вредоносного контента LLM. Отвечает за контроль качества данных, используемых при обучении модели, для снижения рисков. Scopus ID: 57200073690. Индекс Хирша: 1. Опыт работы в задачах распознавания образов и компьютерного зрения. Участник проекта ПЦФ BR05236693. Қуанышбай Дархан Нургазыұлы – исполнитель, представитель Университета Сулеймана Демиреля. Осуществляет защиту персональных данных и конфиденциальности согласно требованиям ИТ-безопасности РК. Разрабатывает методы

	адаптации LLM к культурным и языковым особенностям казахского языка. Scopus ID: 57215602501. Индекс Хирша: 1. Опыт в робототехнике, компьютерном зрении и машинном обучении. Ыбытаева Галия Сейткалиевна – исполнитель, младший научный сотрудник РГП "Институт информационных и вычислительных технологий". Проводит лингвистический анализ казахского языка (морфология, синтаксис, семантика), формирует базу локальных выражений и идиом для обучения модели. Scopus ID: 57221333455. Индекс Хирша: 2. Опыт в робототехнике, компьютерном зрении и машинном обучении. Утегенова Анар Урантаевна – исполнитель, PhD, ведущий научный сотрудник РГП "Институт информационных и вычислительных технологий". Определяет направления интеграции LLM в образование, госуслуги, законодательство и др. Разрабатывает техническую спецификацию API. Scopus ID: 56826094200. Индекс Хирша: 5. Научные интересы: информационные и космические технологии, ГИС, телекоммуникации, машинное обучение. Автор 50+ публикаций, стаж более 20 лет. Сундетов Талғат Рысбекұлы – исполнитель, PhD-докторант, младший научный сотрудник РГП "Институт информационных и вычислительных технологий". Разрабатывает API-сервисы, документацию, и осуществляет адаптацию API под выбранные платформы. Scopus ID: 57211753744, ORCID: 0000-0003-1193-5517. Индекс Хирша: 1. Участник проекта ГФ №AP05132648. Опыт в робототехнике и компьютерном зрении. Атаниязова Айсулыу Саламатовна – исполнитель, PhD-докторант РГП "Институт информационных и вычислительных технологий". Участвует в разработке и адаптации API-сервисов и соответствующей документации. Scopus ID: 57298357700. Индекс Хирша: 0. Опыт в робототехнике, компьютерном зрении и машинном обучении. Хусейн Атакан Варол – исполнитель, генеральный директор ISSAI, АОО «Назарбаев

	Университет». Высококвалифицированный специалист в области робототехники, машинного обучения, умных систем и ИИ. Автор более 100 научных публикаций, обладатель международных патентов. Руководитель завершённых проектов Министерства образования и науки РК и Назарбаев Университета. Scopus ID: 24484134500, ORCID ID: 0000-0002-4042-425X. Индекс Хирша: 28. Аскат Куздеуов – исполнитель, старший аналитик данных ISSAI, АОО «Назарбаев Университет». Специалист в области компьютерного зрения, распознавания речи, NLP и робототехники. Автор 17 научных статей, участник более 30 проектов. Scopus ID: 57208734478. Индекс Хирша: 9. Абдрахманова Мадина – исполнитель, аналитик данных ISSAI, АОО «Назарбаев Университет». Образование: Университет Иллинойса в Урбана-Шампейн (бакалавр), Калифорнийский университет, Ирвайн (магистр). Опыт в мультимодальном обучении, аугментации данных и управлении кластерными системами. Scopus ID: 57191414950. Индекс Хирша: 3. Полонская Алина – исполнитель, аналитик данных ISSAI, АОО «Назарбаев Университет». Выпускница магистратуры СПбГУ. Научные интересы: NLP, генерация текста, распознавание речи. Работает над многоязычными моделями ИИ. Scopus ID: 000000000, ORCID ID: 0000-0000-0000-0000. Индекс Хирша: 0. Мусаходжаева Саида – исполнитель, научный сотрудник ISSAI, АОО «Назарбаев Университет». Выпускница KAIST (Южная Корея), бакалавриат – Назарбаев Университет. Автор 12 научных статей, участвует в проектах по ИИ для казахского языка. Scopus ID: 57189046541. Индекс Хирша: 6. Ешпанов Рустем – исполнитель, аналитик данных ISSAI, АОО «Назарбаев Университет». Магистр Оксфордского университета по лингвистике. Специалист по NLP и ИИ, автор научных публикаций, участвовал в проектах по казахскому языку.

	Scopus ID: 57364291200. Индекс Хирша: 2. Аспандияр Нуриманов – исполнитель, аналитик данных ISSAI, АОО «Назарбаев Университет». Магистр Data Science, Frankfurt School of Finance & Management. Ранее работал в стартапах и консалтинге, исследует LLM и казахский язык. Scopus ID: 000000000, ORCID ID: 0000-0000-0000-0000. Индекс Хирша: 0. Серікбай Бауыржан – AI-специалист, магистр, научный сотрудник НАО «Национальный научно-практический центр «Тіл – Қазына» имени Шайсултана Шаяхметова». Опыт работы в направлении проекта – более 4 лет. Участвовал в разработке пяти подкорпусов Национального корпуса казахского языка (2021–2024 гг.), проектах TTS и STT для казахского языка, портала для детей «Бала тілі», а также в создании проекта «Экранный диктор» для незрячих. Специализируется на парсинге данных, автоматизации процессов обработки с использованием Python, разработке скриптов, нормализации и форматировании данных. Scopus ID: отсутствует, ORCID ID: 0000-0000-0000-0000, индекс Хирша – 0. Алшынбекова Мадина Серикболсыновна – исполнитель проекта, магистр гуманитарных наук Сианьского университета иностранных языков (КНР), докторант ЕНУ. Научный сотрудник отдела прикладной лингвистики НАО «Национальный научно-практический центр «Тіл – Қазына». Область научных интересов: сопоставительная фразеология. Опыт работы – более 10 лет. Автор более 10 научных публикаций, в том числе индексируемых в Scopus. Выполняет координацию работ по календарному плану, проверку и анализ контента, контроль подготовки статей к публикации. Scopus ID: отсутствует, ORCID ID: 0009-0002-9739-1775, индекс Хирша – 0. Серікбай Бауыржан – магистр, AI-специалист, научный сотрудник НАО «Национальный научно-практический центр "Тіл-Қазына" имени Шайсултана Шаяхметова». Опыт работы в проекте – более 4 лет. Участник разработки пяти подкорпусов Национального корпуса казахского языка (2021–2024 гг.), портала для детей “Бала тілі”, проекта TTS казахского языка, чат-ботов по синтезу речи, проекта “Экранный диктор” для незрячих. Основные

	задачи: распределение задач, парсинг данных, разработка скриптов, автоматизация сбора и обработки данных, нормализация. Scopus ID: — ORCID ID: 0000-0000-0000-0000 Индекс Хирша – 0. Алшынбекова Мадина Серикболсыновна – магистр, докторант ЕНУ, научный сотрудник отдела прикладной лингвистики НАО «Тіл-Қазына». Основной исполнитель проекта, занимается координацией, редактированием, анализом контента, контролем подготовки публикаций. Магистр гуманитарных наук Сианьского университета иностранных языков (КНР). Научные интересы – сопоставительная фразеология. Автор более 10 научных публикаций, в том числе в базе Scopus. Scopus ID: — ORCID ID: 0009-0002-9739-1775 Индекс Хирша – 0. Демесинова Ляйля Муратовна – PhD, научный сотрудник. Опыт преподавания – более 17 лет. Основные задачи в проекте: координация, редактирование, грамматическая и структурная проверка, метаразметка, анализ и подготовка научных статей. Автор более 40 научных публикаций, 2 монографий и 2 учебников, часть которых входит в базу Scopus и КОКСОН. Scopus ID: 58242361100 ORCID ID: 0000-0002-2414-7360 Индекс Хирша – 0. Участник проектов AP08855981 (2020–2022), AP19677903 (2023–2026). Исполнитель ряда проектов Центра «Тіл-Қазына», включая мобильное приложение «Тіл-Құрал», синтез устной речи, лексико-грамматический минимум. Гүлназ Төкенқызы – кандидат филологических наук, доцент, ведущий научный сотрудник отдела прикладной лингвистики НАО «Тіл-Қазына». Ассоциированный профессор, научный стаж – 30 лет. Автор более 100 научных публикаций, включая более 30 в КОКСОН и множество в Scopus. Автор и соавтор 17 учебников и электронных учебных пособий. ORCID ID: 0000-0003-4993-9927 Индекс Хирша – 3. Лауреат госпремии «Лучший преподаватель вуза» (2011), эксперт KazTest (2011–2023), член диссовета ЕНУ. Руководитель и исполнитель ряда научных проектов, в том числе по эпическому наследию, молодежной поэзии,

	личности Ахмета Байтурсынова. Исполнитель проектов по созданию Национального корпуса казахского языка и контент-редактор проекта «Qazgramma». Дана Оспанова Жаңабекқызы – магистр («Қазақ тілі мен әдебиеті», 2017), докторант ЕНУ. Научный сотрудник отдела прикладной лингвистики НАО «Национальный научно-практический центр "Тіл – Қазына" имени Шайсултана Шаяхметова», руководитель проекта. Обладает более чем 8-летним опытом научной работы. Руководитель проекта пяти подкорпусов Национального корпуса казахского языка (2021–2024 гг.). Выполняет задачи по координации проектных работ, редактированию, проверке и анализу контента, контролю подготовки научных публикаций. Автор более 15 научных публикаций, включая статьи в базе данных Scopus. Участвовала как контент-корректор в проекте «Qazgramma», а также как исполнитель по сбору и подготовке материалов для журнала «Тіл және қоғам» и портала qazcorpora.kz. Scopus ID: — ORCID ID: 0000-0001-8901-8060 Индекс Хирша – 0. Омарова София Кожиакпаровна – кандидат филологических наук, заведующий отдела прикладной лингвистики НАО «Национальный научно-практический центр "Тіл – Қазына" имени Шайсултана Шаяхметова». Автор более 45 научных публикаций, 2 монографий и 20 патентов в сфере методики преподавания казахского языка. Обладает богатым опытом участия в научных проектах, включая разработку латинографического алфавита, переводческую деятельность (в том числе перевод «Тюркского альманаха» и книги «Тюркское воспитание» в 2013 г.), методику преподавания казахского языка как родственного, создание обучающих словарей и видеоуроков. Участник и руководитель ряда крупных проектов: Проект Министерства образования РК по созданию лексико-грамматического минимума для 1–11 классов (2017, 2023 гг.); Руководитель проекта «Qazgramma.kz» (лингвистическая часть, 2023 г.); Метаразметчик подкорпуса Национального корпуса казахского языка «Qazcorpora.kz» (2022 г.); Советник проекта по универсальному лексико-грамматическому минимуму (2024 г.). Scopus ID: —

	ORCID ID: 0009-0009-0819-0543 Индекс Хирша – 0. Имеет государственные награды и почётные знаки. Меруерт Ермахан – магистр, научный сотрудник НАО «Национальный научно-практический центр "Тіл – Қазына" имени Шайсултана Шаяхметова». Образование: магистр в области международного и европейского права, Римский университет Ла Сапиенца, Италия (стипендиат программы LazioDisco). Обладает компетенциями в сборе, анализе, преобразовании данных и переводе с английского на казахский язык. Участвовала в проектах Центра «Тіл-Қазына». Специализируется на редактировании, проверке и форматировании контента. Scopus ID: — ORCID ID: 0000-0000-0000-0000 Индекс Хирша – 0. Бакиткызы Молдир - руководитель отдела Til-Hub, НАО «ННПЦ "Тіл – Қазына" им. Ш. Шаяхметова», магистр. Роль: Исполнитель. Функции: координация проектных работ согласно календарному плану, обработка контента, орфографическая и грамматическая корректура, структурный анализ текста, метаразметка и подготовка статей для публикации. Лиза Есбосын Қожаққызы - должность: Исполнительный директор, НАО «ННПЦ "Тіл – Қазына"», магистр, исполнитель. Функции: координация проектной работы, проверка и редактирование научного контента, контроль публикационной деятельности Нурсултан Еламанов Алибекович, магистр IT-специалист «Национальный научно-практический центр «Тіл –Қазына» имени Шайсултана Шаяхметова». Разработка и поддержка базы данных для хранения и управления большими объемами данных, Настройка резервного копирования и восстановления данных. Имеет 7 лет стажа в разработке веб-порталов. Сандугаш Кенжебаева Баймаханқызы, магистр. Руководитель центра управления проектами НАО «Национальный научно-практический центр «Тіл – Қазына» имени Шайсултана Шаяхметова». Координация работ в соответствии с заданиями и календарного плана Проверка, редактирование, анализ собранного контента. Эксперт по медиа, а также

	сертифицированный медиа-тренер. Есть большой опыт работы в разного рода международных проектах. Обладатель гранта Digital Communication Network в США, организованном Государственным Департаментом США. Имеет знания и опыт работы с различными источниками данных (аудио, видео, текстовые), конвертацией, метаразметкой данных. Scopus ID: отсутствует, ORCID ID: 0000-0000-0000-0000, индекс Хирша – 0. Scopus ID: отсутствует, ORCID ID: 0000-0000-0000-0000, индекс Хирша – 0. Олжас Ибраев Пазылбекович – магистр, советник генерального директора НАО «ННПЦ «Тіл – Қазына». Исполнитель проекта. Выполняет координацию, проверку и редактирование контента. Имеет управленческий и предпринимательский опыт: директор компании DRIFT (по настоящее время), генеральный директор Green Innovation (2018–2024), исполнительный директор Market Connect (2017–2018), главный специалист в ННТЭО (2014–2016). Асель Исмайлова Амангелдыевна – магистр экономических наук, специалист IT-отдела НАО «ННПЦ «Тіл – Қазына». Исполнитель проекта. Участвует в координации и редактировании контента, в том числе в рамках проекта по синтезу казахской устной речи. Координатор проекта по разработке Национальной модели LLM. Организует конкурсы, семинары и конференции, ведёт взаимодействие с госорганами по научно-методическим и организационным вопросам. Была корректором-редактором сборника конференции по латинской графике. Scopus ID: отсутствует, ORCID ID: 0000-0000-0000-0000, индекс Хирша – 0. Жазира Искакова Максутовна – кандидат филологических наук, заведующая методическим отделом НАО «Национальный научно-практический центр «Тіл – Қазына» имени Шайсултана Шаяхметова». Исполнитель проекта. В рамках проекта осуществляет координацию работы по заданиям и календарному плану, проверку и редактирование собранного контента, контроль подготовки научных статей к публикации. Имеет более 20 лет опыта научно-методической деятельности. Автор свыше 60 научных статей и 5 учебных пособий по обучению казахскому языку как второму и казахской литературе. ORCID ID: 0000-0002-0940-615X

	Жамал Айткалиевна Манкеева – главный научный сотрудник Института языкознания имени А. Байтурсынова, доктор филологических наук, один из ведущих специалистов в области лексикологии, лексико-семантики, корпусной лингвистики. Автор более 200 научных публикаций, включая монографии и статьи в международных изданиях. В проекте занимается подбором и обоснованием источников данных из широкого спектра языковых структур. Scopus ID: 57204363069 ORCID: 0000-0002-7390-4778 h-индекс: 2 Рустембек Нусхабекович Шойбеков – главный научный сотрудник Института языкознания имени А. Байтурсынова, ведущий специалист в области лексикологии, этнолингвистики и лингвокультурологии. Автор фундаментальных трудов энциклопедического характера. В проекте осуществляет отбор и обоснование источников данных из языковых структур различных стилей. Scopus ID: отсутствует ORCID: 0000-0002-8979-3566 h-индекс: 0 Мырзаберген Малбакович Малбаков – главный научный сотрудник Института языкознания имени А. Байтурсынова, специалист в области тюркологии, истории языка и лексикографии. Один из ведущих отечественных исследователей в данной области. В проекте отвечает за отбор и обоснование источников из лексикографических ресурсов. Scopus ID: 56195379100 ORCID: 0000-0001-8485-0544 h-индекс: 0 Бағдан Қатайқызы Мамынова – главный научный сотрудник Института языкознания имени А. Байтурсынова. Известный ученый в области грамматики, текстоведения, морфологии и прагмалингвистики. Автор более 250 научных публикаций, включая 15 монографий и 3 фундаментальные статьи, индексируемые в Scopus. В проекте занимается обоснованием языковых источников в различных исторических и социальных контекстах. Scopus ID: 56176636300 ORCID: отсутствует h-индекс: 1

	Сарсенбай Куантаевич Кулманов – ведущий научный сотрудник Института языкознания имени А. Байтурсынова, руководитель Центра терминологии. Специалист в области грамматики, терминоведения и переводоведения. Автор более 100 научных работ. В проекте участвует в обосновании и систематизации терминологических данных для обеспечения разнообразия в обучающих моделях. Scopus ID: отсутствует ORCID: 0000-0001-8289-9299 h-индекс: 0 Альфия Абдыкеновна Солтанбекова – ведущий научный сотрудник Института языкознания имени А. Байтурсынова, специалист в области морфологии, синтаксиса, функциональной грамматики и лексикографии. Автор около 100 научных работ. В рамках проекта занимается подбором и обоснованием источников данных из языковых структур, представленных в различных типах и контекстах. Scopus ID: отсутствует ORCID: 0000-0002-3513-4012 h-индекс: 0 Айнур Аташбековна Сейтбекова – ведущий научный сотрудник Института языкознания имени А. Байтурсынова, специалист по истории языка, тюркологии, арабскому и персидскому языкам. Участвует в создании исторического подкорпуса Национального корпуса казахского языка. В проекте занимается отбором языковых данных из исторических источников. Scopus ID: 572055459500 ORCID: 0000-0002-3513-4012 h-индекс: 0 Эльмира Абдугалиевна Утебаева – ведущий научный сотрудник Института языкознания имени А. Байтурсынова. Специалист по этнолингвистике, концептуологии, лексикографии и неографии. Автор более 60 научных статей. Один из разработчиков культурно-репрезентативного внутрикорпуса казахского языка. В проекте осуществляет анализ языковых данных с этнокультурным контекстом. Scopus ID: 57200244748 ORCID: отсутствует h-индекс: 1 Нұрсауле Мақсұтқызы Рсалиева – ведущий научный сотрудник Института языкознания имени А. Байтурсынова, специалист в области терминологии, переводоведения, ономастики и

	лексикографии. Главный редактор и составитель казахско-английского и англо-казахского словаря Oxford. Один из разработчиков параллельного и писательского подкорпусов НККЯ. В проекте отвечает за проверку и анализ источников данных в разных стилях, жанрах и типах. Scopus ID: 410650707 ORCID: 0000-0003-2065-0505 h-индекс: 0 Гулжихан Байказиевна Кубденова – ведущий научный сотрудник Института языкознания имени А. Байтурсынова, специалист по сравнительному языкознанию тюркских языков. Автор около 50 научных статей. Один из разработчиков исторического подкорпуса Национального корпуса казахского языка. В проекте осуществляет отбор данных из различных исторических текстов. Scopus ID: отсутствует ORCID: 0000-0001-7514-8710 h-индекс: 0 Арайлым Ботановна Шормакова – старший научный сотрудник Института языкознания имени А. Байтурсынова, специалист по этнолингвистике, лингвокультурологии, лексикологии и лексикографии. Автор более 30 научных публикаций. В рамках проекта занимается группировкой отобранных языковых данных. Scopus ID: 57211905258 ORCID: 0000-0001-9551-3818 h-индекс: 1 Ақмарал Зинол-Ғабденқызы Бисенғали – старший научный сотрудник Института языкознания имени А. Байтурсынова. Специалист в области тюркологии, терминологии, лексикологии и этнолингвистики. Автор около 40 научных статей. В проекте осуществляет выравнивание и сопоставление данных из терминологических источников, отражающих реальные языковые проявления в общественной жизни. Scopus ID: 57200075552 ORCID: 0000-0002-1964-9948 h-индекс: 1 Нурлан Кунпияевич Шуленбаев – старший научный сотрудник Института языкознания имени А. Байтурсынова, специалист в области переводоведения и терминологии. Автор более 40 научных статей. В проекте занимается подбором источников данных из электронных ресурсов.

	Scopus ID: отсутствует ORCID: 0000-0003-2203-9822 h-индекс: 0 Зубайра Боранбековна Садырбаева – научный сотрудник Института языкознания имени А. Байтурсынова, специалист в области ономастики и этнолингвистики. В проекте занимается разработкой лингвистической разметки ономастических имен. Scopus ID: отсутствует ORCID: 0000-0002-2729-4198 h-индекс: 0 Аида Серикхановна Барменкулова – научный сотрудник Института языкознания имени А. Байтурсынова, специалист по стилистическому анализу поэтического текста. В рамках проекта занимается разработкой лингворазметки данных, предназначенных для обучения больших языковых моделей. Scopus ID: отсутствует ORCID: 0000-0002-1471-5846 h-индекс: 0 Талгат Бекбулатович Рамазанов – научный сотрудник Института языкознания имени А. Байтурсынова, специалист по грамматике, истории языка и сравнительному языкознанию. В проекте занимается разработкой мета- и лингворазметки данных, предназначенных для обучения больших языковых моделей. Scopus ID: отсутствует ORCID: 0000-0002-4405-5914 h-индекс: 0 Гульден Бакытказыевна Тлегенова – научный сотрудник Института языкознания имени А. Байтурсынова, молодой лингвист, специализирующийся в области прикладной лингвистики. В проекте занимается созданием метаразметки языковых данных для обучения больших языковых моделей. Scopus ID: отсутствует ORCID: 0000-0001-9842-3324 h-индекс: 0 Әсел Қазбекқызы Сейдамат – младший научный сотрудник Института языкознания имени А. Байтурсынова, PhD-докторант, один из ключевых специалистов по разработке исторического подкорпуса Национального корпуса казахского языка. В проекте занимается лингворазметкой текстов различных стилей. Scopus ID: отсутствует ORCID: 0000-0003-1893-4920

	h-индекс: 0 Меруерт Ануаровна Имангазина – младший научный сотрудник Института языкознания имени А. Байтурсынова, специалист по грамматике и сравнительному языкознанию. Свободно владеет английским языком (IELTS – 7.0, уровень C1), автор около 10 научных публикаций. В проекте занимается лингворазметкой данных для обучения языковых моделей. Scopus ID: отсутствует ORCID: 0000-0001-8473-2812 h-индекс: 0 Айкерим Мурсал – младший научный сотрудник Института языкознания имени А. Байтурсынова, PhD-докторант, один из квалифицированных специалистов по разработке исторического подкорпуса Национального корпуса казахского языка. В проекте отвечает за лингворазметку данных исторических текстов. Scopus ID: отсутствует ORCID: 0000-0001-8035-4236 h-индекс: 0 Қымбат Берікұлы Слямбеков – младший научный сотрудник Института языкознания имени А. Байтурсынова, специалист по грамматике и стилистике. Автор около 10 научных статей. В проекте занимается разработкой мета- и лингворазметки данных, предназначенных для обучения больших языковых моделей. Scopus ID: отсутствует ORCID: 0000-0002-9731-0448 h-индекс: 0 Маржан Серікқызы – младший научный сотрудник Института языкознания имени А. Байтурсынова, PhD-докторант, специалист в области этнолингвистики. Автор около 10 научных публикаций. В проекте занимается мета- и лингворазметкой этнолингвистических данных. Scopus ID: отсутствует ORCID: 0000-0002-4547-1928 h-индекс: 0 Айдана Мухымбайқызы Махамбетова – младший научный сотрудник Института языкознания имени А. Байтурсынова, специалист в области социолингвистики. В проекте занимается сбором и систематизацией социолингвистических материалов. Scopus ID: отсутствует

	ORCID: 0009-0005-5846-7433 h-индекс: 0 Аягуль Данияровна Омарова – младший научный сотрудник Института языкознания имени А. Байтурсынова, PhD-докторант, специалист в области терминологии и терминографии. Автор около 10 научных публикаций. В проекте работает над мета- и лингворазметкой терминологических данных. Scopus ID: отсутствует ORCID: 0000-0002-4948-6764 h-индекс: 0
Список публикаций со ссылками на них	1. Sofya Omarova. A Corpus Approach in Language Discovery: A Word Frequency Analysis Based on the Corpus Outcomes in Kazakh. Forum for Linguistic Studies (FLS). View Vol. 7 , Iss. 4 (April 2025): In Progress ISSN: 2705-0602 (Online) 2. Akbayan Bekarystankyzy, Abdul Razaque, Orken Mamyrbayev Integrated end-to-end multilingual method for low-resource agglutinative languages using Cyrillic scripts // Journal of Industrial Information Integration, Volume 43, January 2025, 100750, https://doi.org/10.1016/j.jii.2024.100750 3. Shaikhiyeva, Z., Mansurova, M., & Amirkhanova, G. (2024). Text to SQL Transformation Using LLM: A Comparative Research of T5, Seq2Seq, and SQLNet Models. Proceedings of the IX International Conference on Computer Science and Engineering (UBMK-2024) (pp. 967–972). IEEE. ISBN: 979-8-3503-6587-0, p.967-972. 4. Kairatuly, B., & Mansurova, M. (2024). Named Entity Recognition from Kazakh Speech. Proceedings of the IX International Conference on Computer Science and Engineering (UBMK-2024). IEEE. ISBN: 979-8-3503-6587-0, p.153-156. 5. Шайхиева, Ж., Мансурова, М., Амирханова, Г., & Оспан, А. (2024). Кластеризация комментариев с использованием большой языковой модели в финансовых технологиях. Информатика и прикладная математика: Материалы IX Международной научно-практической конференции (31 октября – 1 ноября 2024 г.) (сс. 444–451). Алматы. 6. Mansurova, M. E., Rakhimova, D. R., & Duisenbekkyzy, Zh. (2024). Morphological Parsing of Kazakh Texts with Deep Learning Approaches. Journal of Mathematics, Mechanics and Computer Science (JMMCS). ISSN: 1563–0277, eISSN: 2617-4871. Retrieved from https://bm.kaznu.kz. DOI: https://doi.org/10.26577/JMMCS.2022.v113.i1.13.

	7. Rakhimova, D., Mansurova, M., Matanov, N., & Yerik, A. (2024). Machine Learning-Based Synonymous Word Detection in Kazakh. Proceedings of the IX International Conference on Computer Science and Engineering (UBMK-2024). IEEE. ISBN: 979-8-3503-6587-0.