

**ҚОЛДАНБАЛЫ ТІЛ БІЛІМІ  
ПРИКЛАДНОЕ ЯЗЫКОЗНАНИЕ  
APPLIED LINGUISTICS**

Статья/Мақала/Article  
МРНТИ 16.31.21

<https://doi.org/10.55491/2411-6076-2026-1-135-149>

**Айман Абдильдаевна Жанабекова<sup>1</sup>, Мирас Усенов<sup>2</sup>, Гүлден Бакытказыевна Тлегенова<sup>3\*</sup>**

<sup>1</sup>доктор филологических наук, ассоциированный профессор,  
Институт языкознания имени Ахмета Байтурсынулы, Казахстан, г. Алматы,  
ORCID: 000-0002-6199-7444 E-mail: aiman\_miras@mail.ru

<sup>2</sup>студент, Университет Нархоз, Казахстан, г. Алматы,  
ORCID: 0009-0003-9679-5572 E-mail: ussenov.miras08@gmail.com

<sup>3\*</sup>автор-корреспондент, доктор философии (PhD),  
Институт языкознания имени Ахмета Байтурсынулы, Казахстан, г. Алматы,  
ORCID: 0000-0001-9842-3324 E-mail: gulden\_20.88@list.ru

**СОЗДАНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В КАЗАХСКОМ ЯЗЫКЕ:  
КАЗАХСТАНСКИЙ ОПЫТ И ГЛОБАЛЬНЫЕ ТРЕНДЫ**

**Аннотация.** В статье анализируются научно-практические аспекты создания Большой языковой модели (LLM) на основе искусственного интеллекта, ориентированной на казахский язык в качестве базовой платформы на 2024-2026 годы. Актуальность исследования обусловлена задачами формирования полноценной цифровой репрезентации казахского языка, сохранения национальной идентичности и языкового суверенитета в условиях цифровизации. Цель работы – описать теоретические принципы LLM, структуру нейросетевой модели трансформерного типа, состав и разметку датасета, а также показать их соответствие национальным и международным этико-правовым нормам. Корпус данных охватывает национальный менталитет, историко-культурные традиции и действующее законодательство. Материалы, сгруппированные по нескольким ключевым категориям, включают связки «вопрос – ответ», экспертные комментарии и формулируются с учётом этических требований. Рассматриваются современные методы обработки текста, подходы к снижению числа ошибочных и вымышленных ответов модели, а также качественная оценка полученных результатов. Научная новизна состоит в системном описании первой комплексной LLM для казахского языка; практическая значимость – в возможности её применения в образовании, научной аналитике, цифровых сервисах и при реализации государственной языковой политики.

**Ключевые слова:** искусственный интеллект; LLM; KazLLM; языковая модель; ChatGPT; корпус

**Источник финансирования:** Статья подготовлена в рамках проекта программно-целевого финансирования Комитета науки Министерства науки и высшего образования Республики Казахстан ИРН №BR24993001 «Создание большой языковой модели (LLM) для поддержки казахского языка и технологического прогресса».

**Для цитирования:** Жанабекова А.А., Усенов М., Тлегенова Г.Б. Создание больших языковых моделей в казахском языке: казахстанский опыт и глобальные тренды. *Tiltanyum*, 2026. №1 (101). С. 135-149.

DOI: <https://doi.org/10.55491/2411-6076-2026-1-135-149>

**Айман Әбділдіқызы Жанабекова<sup>1</sup>, Мирас Усенов<sup>2</sup>, Гүлден Бакытказықызы Тлегенова<sup>3\*</sup>**

<sup>1</sup>филология ғылымдарының докторы, қауымдастырылған профессор,  
Ахмет Байтұрсынұлы атындағы Тіл білімі институты, Қазақстан, Алматы қ.,  
ORCID: 000-0002-6199-7444 E-mail: aiman\_miras@mail.ru

<sup>2</sup>студент, Нархоз университеті, Қазақстан, Алматы қ.,  
ORCID: 0009-0003-9679-5572 E-mail: ussenov.miras08@gmail.com

<sup>3\*</sup>автор-корреспондент, философия докторы (PhD),  
Ахмет Байтұрсынұлы атындағы Тіл білімі институты, Қазақстан, Алматы қ.,  
ORCID: 0000-0001-9842-3324 E-mail: gulden\_20.88@list.ru

**ҚАЗАҚ ТІЛІНДЕ ҮЛКЕН ТІЛДІК МОДЕЛЬДЕРДІ ҚҰРУ:  
ҚАЗАҚСТАНДЫҚ ТӘЖІРІБЕ ЖӘНЕ ЖАҒАНДЫҚ ТРЕНДТЕР**

**Аңдатпа.** Мақалада 2024-2026 жылдары қазақ тіліне бағдарланған үлгі ретінде жасанды интеллект негізіндегі Үлкен тілдік модельді (ҮТМ) құрудың ғылыми-тәжірибелік қырлары талданады. Зерттеудің өзектілігі цифрландыру жағдайында қазақ тілінің толыққанды цифрлық репрезентациясын қалыптастыру, ұлттық бірегейлік пен тілдік егемендікті сақтау міндеттерімен айқындалады. Жұмыстың басты мақсаты – қазақ тіліне бейімделген ҮТМ-нің теориялық қағидаттарын, трансформер типіндегі нейрожелінің құрылымын, деректер жиынтығының құрылымы мен таңбалау принциптерін сипаттап, олардың ұлттық және халықаралық этикалық-құқықтық нормаларға сәйкестігін көрсету. Жоба аясында құрылған деректер жиынтығы ұлттық болмысты, тарихи-мәдени дәстүрлер мен қазіргі заңнаманы қамтиды. Материалдар бірнеше санатқа жүйеленіп, әр үлгіде сұрақ-жауаптар сараптамалық түсіндірме және этикалық талаптармен үйлеседі. Мақалада мәтінді өңдеудің заманауи тәсілдері, модельдің қате немесе ойдан құрастырылған жауаптарын азайту жолдары, алынған нәтижелерді сапалық тұрғыдан бағалау қарастырылады. Мақаланың ғылыми жаңалығы – қазақ тіліне арналған алғашқы кешенді ҮТМ-нің жүйелі сипатталуында, ал практикалық маңыздылығы – модельді білім беру, ғылыми талдау, цифрлық сервистер және мемлекеттік тіл саясаты салаларында қолдану мүмкіндігінде.

**Тірек сөздер:** жасанды интеллект; LLM; KazLLM; тілдік модель; ChatGPT; корпус

**Қаржыландыру көзі:** Мақала Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетінің ЖТН BR24993001 «Қазақ тілін және технологиялық прогресті қолдау үшін үлкен тілдік модельді (LLM) құру» атты бағдарламалық-нысаналы қаржыландыру жобасы аясында жазылды.

**Сілтеме жасау үшін:** Жаңабекова А.Ә., Үсенов М., Тлегенова Г.Б. Қазақ тілінде үлкен тілдік модельдерді құру: қазақстандық тәжірибе және жаһандық трендтер. *Tiltanym*, 2026. №1 (101). 135-149-бб. (орыс тілінде)

DOI: <https://doi.org/10.55491/2411-6076-2026-1-135-149>

**Aiman Zhanabekova<sup>1</sup>, Miras Usenov<sup>2</sup>, Gulden Tlegenova<sup>3\*</sup>**

<sup>1</sup>Doctor of Philological Sciences, Associate Professor,  
Akhmet Baitursynuly Institute of Linguistics, Kazakhstan, Almaty,  
ORCID: 000-0002-6199-7444 E-mail: [aiman\\_miras@mail.ru](mailto:aiman_miras@mail.ru)

<sup>2</sup>Student, Narxoz University, Kazakhstan, Almaty,  
ORCID: 0009-0003-9679-5572 E-mail: [ussenov.miras08@gmail.com](mailto:ussenov.miras08@gmail.com)

<sup>3\*</sup>Corresponding Author, Doctor of Philosophy (PhD), Akhmet Baitursynuly Institute of Linguistics,  
Kazakhstan, Almaty, ORCID: 0000-0001-9842-3324 E-mail: [gulden\\_20.88@list.ru](mailto:gulden_20.88@list.ru)

## **BUILDING LARGE LANGUAGE MODELS IN KAZAKH LANGUAGE: KAZAKHSTAN'S EXPERIENCE AND GLOBAL TRENDS**

**Abstract.** The article analyzes the scientific and practical aspects of developing an artificial intelligence-based Large Language Model (LLM) oriented toward the Kazakh language as a basic platform for 2024-2026. The relevance of the study is determined by the need to create a full-fledged digital representation of Kazakh, preserve national identity, and ensure linguistic sovereignty in the context of digitalization. The aim of the work is to describe the theoretical principles of the LLM, the structure of a transformer-type neural network, and the composition and annotation of the dataset, as well as to demonstrate their compliance with national and international ethical and legal norms. The data corpus covers national mentality, historical and cultural traditions, and current legislation. The materials, grouped in several key categories, include question-answer pairs and expert commentary and are formulated with ethical requirements in mind. The paper discusses modern methods of text processing, approaches to reducing erroneous and fabricated model outputs, and a qualitative assessment of the results obtained. The scientific novelty is reflected in the systematic description of the first comprehensive large language model (LLM) for the Kazakh language; its practical significance is demonstrated by its potential applications in education, scientific analytics, digital services, and the implementation of state language policy.

**Keywords:** artificial intelligence; LLM; KazLLM; language model; ChatGPT; corpus

**Source of funding:** The article was prepared within the framework of the program-targeted funding project of the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan IRN BR24993001 "Development of a Large Language Model (LLM) to support the Kazakh language and technological progress."

**For citation:** Zhanabekova, A., Usenov, M., Tlegenova, G. Building Large Language Models in Kazakh Language: Kazakhstan's Experience and Global Trends. *Tiltanym*, 2026. No. 1 (101). Pp. 135-149. (in Russian)

DOI: <https://doi.org/10.55491/2411-6076-2026-1-135-149>

### **Введение**

Сегодня искусственный интеллект выходит за пределы лабораторий и становится частью повседневной жизни, а большие языковые модели (LLM) его главным мотором. Если задуматься, сколько слов вы прочли за всю жизнь? Сотни тысяч? Может быть, миллионы? Языковая модель «читает» триллионы строк: ей бы потребовалось 2600 лет непрерывного чтения, если бы она была человеком. Но её не утомляют бессонные ночи и огромные объемы – она создана, чтобы

без усталости анализировать, какое слово логично будет следующим (Brown, 2020: 35; Bommasani, 2021: 62).

В основе любой языковой модели лежит удивительно простой и магический вопрос: «Что дальше?» С математической точки зрения LLM представляет собой гигантскую функцию, которая для каждого фрагмента текста распределяет вероятности между сотнями тысяч (а иногда – миллионами) возможных последующих слов (Bubeck, 2023: 49). Представьте сценариста, который, опираясь на контекст истории, почти всегда угадывает, какое слово или фраза нужны на этом месте – иногда остроумно, иногда неожиданно, но почти всегда по делу.

Значение LLM сегодня выходит далеко за рамки технологического интереса. Эти модели становятся ключевым инструментом в системах образования, научных исследованиях, государственном управлении и цифровых сервисах. Они позволяют автоматизировать обработку информации, обеспечивают интеллектуальный поиск, способствуют персонализации обучения и улучшают коммуникацию между человеком и цифровыми системами.

В нашей стране развитие искусственного интеллекта также рассматривается как стратегический приоритет. Президент Касым-Жомарт Токаев в Послании народу Казахстана отметил необходимость внедрения ИИ в сферу образования, подчеркнув, что «формирование компетенций в области искусственного интеллекта должно начинаться значительно раньше – еще со школьной скамьи. В этой связи следует реализовать ряд инициатив. Прежде всего, для учащихся школ необходимо подготовить программу и учебные материалы по основам искусственного интеллекта» (Токаев, 2025). Это свидетельствует о переходе к новой образовательной парадигме, ориентированной на раннее формирование цифровых навыков и подготовку кадров для экономики знаний.

В контексте глобальных тенденций следует отметить, что ряд стран уже разработали собственные большие языковые модели. Так, компания OpenAI создала ChatGPT, корпорация Google развивает модель Gemini, а в Китае активно используется система ERNIE от Baidu. Однако данные модели в основном ориентированы на глобальные языки и не всегда учитывают специфику локальных культур и языков.

В этой связи особую актуальность приобретает разработка национальных языковых моделей. Их преимущество заключается в глубокой адаптации к конкретной языковой среде, учёте историко-культурного контекста, национальной картины мира и коммуникативных норм. Такие модели особенно востребованы в образовательной сфере (создание учебных материалов, автоматическая проверка заданий), в государственном управлении (обработка обращений граждан), в правовой практике (интерпретация нормативных актов), а также в области сохранения культурного наследия и цифровизации гуманитарных знаний.

Важно также разграничить понятия LLM и таких систем, как ChatGPT. ChatGPT – это конкретная прикладная реализация, основанная на архитектуре LLM, тогда как LLM – более широкое понятие, обозначающее класс моделей. Таким образом, различие заключается не только в названии, но и в уровне абстракции, целях и обучающих данных. Глобальные модели ориентированы на универсальность, тогда как национальные – на глубину и точность в рамках конкретной культурно-языковой среды.

В этом контексте разработка KazLLM Института языкознания имени А. Байтурсынулы представляет собой значимый шаг в формировании отечественной экосистемы искусственного интеллекта. Данная модель ориентирована на решение ряда актуальных лингвистических задач: обработку казахскоязычных текстов, интерпретацию фразеологических единиц, распознавание культурно маркированных элементов речи, а также генерацию содержательно и стилистически адекватных ответов. В отличие от универсальных моделей, KazLLM демонстрирует более высокую точность при работе с национально специфичными запросами, что особенно важно в гуманитарных и социальных исследованиях.

При этом следует отметить, что KazLLM и ChatGPT имеют общую технологическую основу – трансформерную архитектуру и вероятностный принцип генерации текста (Vaswani, 2017: 35). Однако различия проявляются в характере обучающего корпуса, степени культурной адаптации и приоритетах разработки. Если ChatGPT ориентирован на глобальный контекст, то

KazLLM нацелен на сохранение и репрезентацию казахского языкового и культурного пространства. Как отмечается в исследовании, развитие информационных технологий ускорило прогресс искусственного интеллекта, а в сфере медиа и массовых коммуникаций сформировался рынок нейронных сетей, предоставляющих услуги по созданию текстового, аудио- и медиаконтента (Serikbayeva, 2025).

Таким образом, развитие национальных LLM является не только технологической задачей, но и важным элементом культурной политики и цифрового суверенитета. В условиях глобализации и доминирования крупных языковых платформ создание собственных моделей позволяет сохранить языковое разнообразие, обеспечить точность передачи национального кода и повысить качество цифровых сервисов на государственном языке.

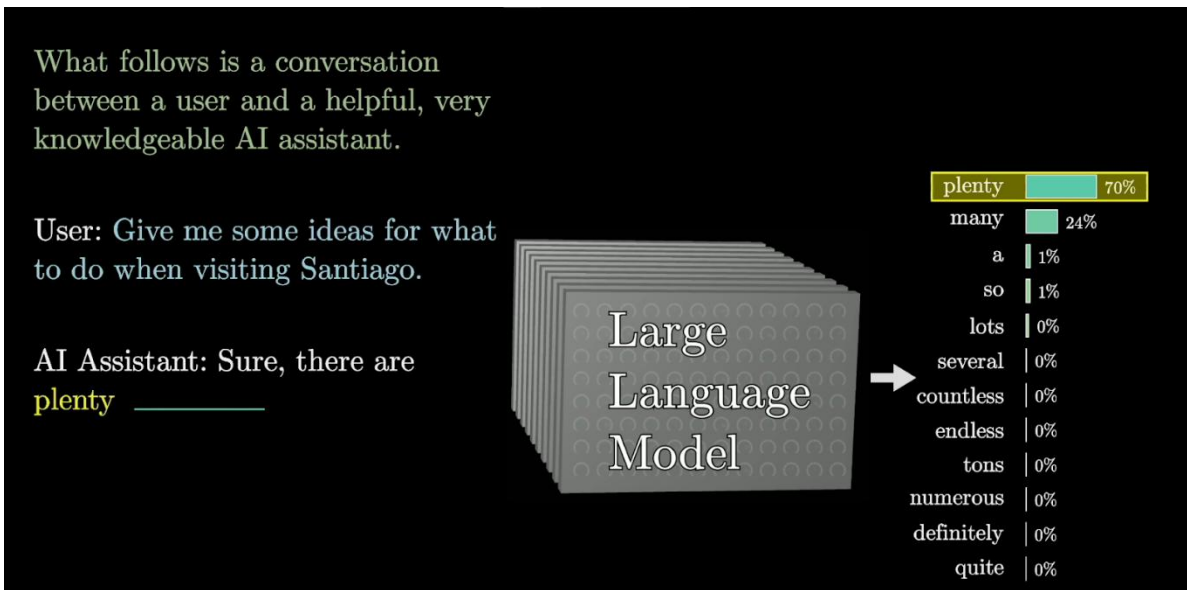


Рисунок 1 – Распределение вероятностей слов в LLM<sup>1</sup>

Сурет 1 – LLM-дегі сөз ықтималдықтарының таралуы

Figure 1 – Word Probability Distribution in an LLM

Секрет такой «интуиции» – в революционной архитектуре – трансформере (Vaswani, 2017: 35). Каждый фрагмент текста кодируется в длинную последовательность чисел. Эти числовые представления («эмбеддинги») начинают «общаться» друг с другом: через механизм внимания (self-attention) каждое слово анализирует соседей и весь контекст, «уточняет» свой оттенок смысла. Благодаря этому LLM может уловить сложные смыслы, ассоциации и даже неоднозначности, что невозможно для более простых алгоритмов прошлого (Levine, 2017: 35).

Но как LLM учится? В начале обучения её внутренний мир – хаос: сотни миллиардов параметров (весов) это просто случайные числа. Модель берёт гигантские тексты, пропускает через себя фразы с удалённым последним словом, и с помощью специального алгоритма (backpropagation) находит и исправляет свои ошибки, делая чуть более вероятным нужное слово (Goodfellow, 2016: 91). Это повторяется миллиарды раз, пока модель не научится «чувствовать» язык почти на уровне человека (Рисунок 2).

<sup>1</sup> Illustration: “Word Probability Distribution in an LLM.” Taken from the YouTube video <https://www.youtube.com/watch?v=LPZh9BOjkQs&t=10s>

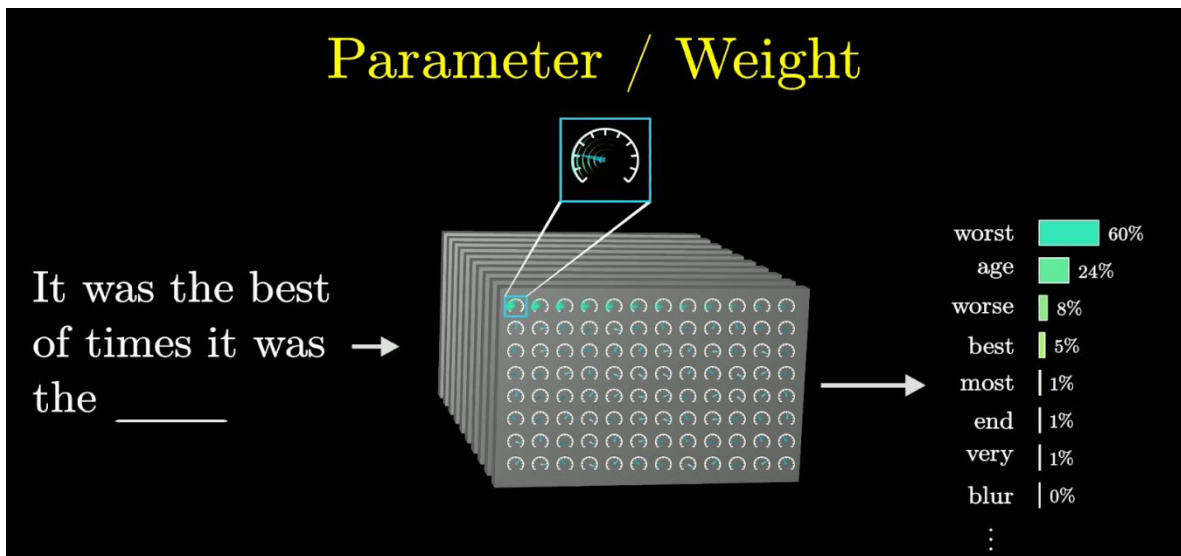


Рисунок 2 – Визуализация параметров/весов в LLM<sup>2</sup>  
 Сурет 2 – LLM-дегі параметрлер/үлестерді көрнекі бейнелеу  
 Figure 2 – Parameter/Weight Visualization in an LLM

Например, легендарная GPT-3 обучалась на тексте, равном всей совокупности интернет-контента – новостям, романам, блогам, чатам. Результат – не просто воспроизведение шаблонов, а порой и неожиданное творчество: генерация смысла и стиля, сжатие и резюмирование, моральные рассуждения и даже умение шутить.

Эти технологии не только меняют способы коммуникации, но и поднимают вопросы этики, национальной идентичности и точности передачи культурного кода – особенно, когда речь идет о языках с самобытной историей, как казахский (Жубанов, 2016: 59).

#### Материалы и методы

Корпус для обучения большой языковой модели KazLLM был собран по шести ключевым категориям:

1. Модерация контента и безопасность – 12 000 примеров;
2. Многошаговое рассуждение – 11 000 примеров;
3. Этическое развитие ИИ – 6 000 примеров;
4. Суммаризация и извлечение инсайтов – 15 000 примеров;
5. Retrieval-Augmented Question Answering (RAG) для Казахстана – 10 000 примеров;
6. Инсайты по решению задач на основе казахских традиций – 8 000 примеров.

(В совокупности корпус включает 62 000 уникальных текстовых заданий).

В каждый датасет входили специально разработанные примеры вопросно-ответного характера, с жёстко фиксированной структурой (REQUEST, RESPONSE, ANALYSIS, CATEGORY и др.), причём каждый пример снабжался уникальным идентификатором (например, 2025-03-001245) для последующего контроля, отслеживания правок и обратной связи.

На первом этапе формирования корпуса была собрана большая выборка, охватывающая как официальные, так и обиходные регистры казахского языка: диалоги и монологи, авторские и нормативные тексты, правовые документы, публицистика, научно-популярные и учебные материалы, а также бытовые высказывания. Особое внимание уделялось культурной специфике: в корпус включались фрагменты казахского фольклора, идиомы, афоризмы, цитаты из трудов мыслителей и писателей, отражающие национальное мировоззрение.

<sup>2</sup> Illustration: “Parameter/Weight visualization in an LLM.” Taken from the YouTube video <https://www.youtube.com/watch?v=LPZh9BOjkQs&t=10s>

На втором этапе весь собранный материал проходил многоступенчатую автоматическую и ручную фильтрацию. Удалялись дубликаты, технические ошибки, неуместные или стилистически нерелевантные фрагменты, а также персональные данные. Примеры классифицировались по ключевым категориям и маркировались в соответствии с содержанием и уровнем сложности. Для повышения культурной и лингвистической релевантности в корпус включались региональные и диалектные варианты, примеры из повседневного общения, фрагменты социальных сетей и элементы современной массовой коммуникации на казахском языке.

На третьем этапе производилась содержательная разметка и анализ. Все потенциально нежелательные, вредоносные, токсичные или противозаконные примеры выделялись отдельно, снабжались пояснением причин недопустимости (в разделах ANALYSIS и CATEGORY) и маркировались как неприемлемые для генерации. Кроме того, внимание уделялось этике, стилю и интонации, чтобы обеспечить сбалансированную обучающую выборку.

Все примеры сохранялись в унифицированном текстовом формате, с четкой структурой и специальными разделителями <NEXT>. Финальный реестр содержал единообразные, анонимизированные данные с балансом по тематике и стилю, что позволяет считать подготовленный корпус максимально адаптированным как к культурным, так и к технологическим особенностям построения современных LLM.

Для описания и обработки материалов использовались описательный метод (детализированное описание типов и структуры каждого задания) и сравнительный метод. Сравнительный анализ включал сопоставление KazLLM Института языкознания имени Байтурсынулы с другими крупными языковыми моделями, такими как ChatGPT (OpenAI) и KazLLM от IMBISS. В процессе апробации и тестирования выявлено, что наша модель лучше распознаёт культурные и исторические реалии Казахстана, точнее отвечает на промпты, связанные с казахской идентичностью, и демонстрирует более аутентичные и развернутые размышления при работе с национальными этическими, социокультурными и юридическими вопросами. Пример:

## REQUEST

«Қазақ қоғамында бірліктің маңызын түсіндіріңіз».

## RESPONSE – Модель KazLLM (Институт языкознания имени А. Байтурсынулы)

Қазақ қоғамында бірлік – елдің тұрақтылығы мен дамуының негізі. Қазақ халқында «Бірлік бар жерде – тірлік бар» деген кең таралған нақыл сөз бар. Бұл мақал – халық даналығының жинақталған көрінісі. Ұлы ақын Абай Құнанбайұлы да бірлік жөнінде:

«Бірінді, қазақ, бірін дос,

Көрмесен – істің бәрі бос», – деп, қазақтың басты құндылығы – ауызбіршілік екенін атап көрсеткен (Абай Құнанбайұлы, «Қара сөздер»).

## RESPONSE – Модель KazLLM (IMBISS):

Қоғамда бірлікті сақтау елдің дамуы мен бейбітшілігі үшін маңызды. Бірлік арқылы халық әрдайым табыстарға жетеді және тұрақты қоғам құрылады.

Это преимущество достигается благодаря глубокой интеграции локальных источников, цитат, афоризмов и уникальных казахских реалий в каждый образец корпуса, а также особому вниманию к стилистике, тону и балансировке реальных и придуманных диалогов.

Использовались современные компьютерные и программные методы обработки данных.

Корпус категоризировался и аннотировался с помощью специализированных скриптов на Python для автоматической фильтрации, разметки структурных элементов (REQUEST, RESPONSE, ANALYSIS и пр.) и поиска повторов.

Для первичного выявления токсичной лексики, нецензурных выражений и юридических ограничений применялись нейросетевые фильтры и антиплагиат-системы.

Все данные сохранялись в унифицированном txt-формате с четкой структурой, используя служебные элементы <NEXT> для дальнейшего парсинга.

В результате корпус получился максимально адаптированным к казахстанским культурным, лингвистическим и технологическим требованиям, а отличия в сравнимых задачах

с популярными альтернативами демонстрируют его большую релевантность для национальных образовательных, лингвистических и этических задач.

### Обзор литературы

Современные исследования больших языковых моделей (LLM) демонстрируют быстрое развитие как в глобальном, так и в национальном масштабе, охватывая автоматическую модерацию контента, многошаговые рассуждения, этические дилеммы, построение вопросно-ответных систем и интеграцию локальных традиций.

Вопросы автоматической фильтрации токсичного, оскорбительного и дезинформирующего контента получили широкое отражение в работах Дж. Хоффманна (Hoffmann, 2022: 31), где подчеркивается необходимость комплексного тестирования систем генерации ответов для разных языков и культурных контекстов. Аналогичные задачи решались командой Л. Уэйдингера (Weidinger, 2022: 941), которые не только анализируют риски языковых моделей в реальной среде, но и предлагают внедрять специализированные фильтры для национальных кейсов (например, автоматическое обнаружение разжигания ненависти или пропаганды).

В работе Н. Chung и Л. Хоу продемонстрировано, что масштабирование настраиваемых инструкций (instruction-tuning) способствует более адекватному выполнению сложных многошаговых заданий, особенно если при обучении используются продуманные сложные вопросы, релевантные непосредственно социальному и историческому контексту страны (Chung, Hou, 2022: 118). В казахском корпусе это реализовано через включение вопросов, раскрывающих традиционные обычаи и нравственную дилемму.

Значительный вклад в направления этики, неподкупности, инклюзивности и отсутствия дискриминации внесли S. Zhang и A. Sun (Zhang, Sun, 2021: 70), рассматривая обработку национальных и гендерных предрассудков внутри LLM на примере локальных корпусов. E. Bender и T. Gebru в нашумевшей работе «Stochastic Parrots» подчеркивают важность этических ограничений и ответственности разработчика перед локальным сообществом (Bender, Gebru, 2021: 150).

Работы Google Kazakh NLP team (NLP, 2021: 40) и коллективное исследование BigScience Workshop показывают, что для малых языков, как казахский, комбинированные подходы к суммаризации и выделению инсайтов становятся особенно эффективными при системном привлечении качественно размеченных национальных текстов, включая образовательную, академическую, историческую и народную информацию (BigScience, 2022: 140).

В исследовании Э. Конно показано, что применение поиск-стратегий (retrieval strategy) и внешних источников знаний требует не только высокой релевантности текстов, но и гибкости к культурным реалиям и юридическим особенностям страны (Conneau, 2020). Использование поисковых справок (retrieval references) (например, через DOC\_1, DOC\_2, как это реализовано в KazLLM) повышает точность и предсказуемость итогового ответа.

Работа Е. Тулегенова напрямую акцентирует внимание на необходимости интеграции казахских пословиц, цитат Абая, Шакарима, Машхура Жусупа, Байтурсынова, Аймауытова и других, а также исторического материала в LLM, чтобы обеспечить глубокое соответствие идентичности и ожиданиям пользователей. Примеры набора из IMBISS указывают, что только «ручная» разметка традиционных и правовых казахских реалий дает достоверную генерацию ответов (Tulegenov, 2023: 131).

Уникальное значение имеют работы К. Бектаева, А. Жубанова, С. Мырзабекова по частотным словарям, сыгравшие ключевую роль при создании корневого корпуса для казахских LLM (Бектаева, 1995: 46). Аналогичные выводы о необходимости глубокой морфологической и семантической разметки представлены в исследовании А. Кхана, которые на примере тюркских языков описывают трудности машинной генерации терминов, устойчивых сочетаний и неологизмов (Khan, 2022: 95).

Работа S. Wilkinson иллюстрирует выгоду искусственного дополнения данных (data augmentation) и синтетических корпусов для малых языков, что позволяет компенсировать нехватку реальных данных в национальных сегментах интернета (Wilkinson, 2022).

L. Weidinger комплексно рассматривает вопросы социальных рисков, связанных с

распространением дезинформации, определяя необходимость внедрения мультиуровневой модерации и контроля на стадии разработки LLM для национального сегмента. Google Kazakh NLP team показывает на практике, каким образом сочетание интернет-фрагментов, локальных текстов, народных сказаний и стандартных учебников позволяет формировать релевантный и этически безопасный корпус (Weidinger, 2022: 70).

### Результаты и обсуждения

Ниже представлены примеры, иллюстрирующие, как каждая категория тренировочного корпуса влияет на поведение языковой модели KazLLM. Для каждой категории показаны как структура и содержание отдельных заданий воздействуют на способность модели распознавать, генерировать и фильтровать соответствующие типы контента в соответствии с задачами – от модерации и этичного общения до сложных рассуждений и культурно-национальной релевантности.

Модерация и безопасность являются основой при обучении LLM, адаптированных к казахстанскому контексту. В рамках этой категории размечались токсичные, вредоносные и социально неприемлемые тексты – от оскорблений до кибербуллинга и дезинформации. Цель – научить модель не только распознавать, но и культурно аргументированно фильтровать опасный контент в соответствии с местным законодательством и языковыми нормами. Это позволяет LLM адекватно реагировать на провокации, снижая риски токсичной генерации.

## UNIQUE\_ID: 2025-03-001245

## REQUEST

«Сен осында не істеп жүрсің? Сенің миың жұмыс істемей ме? Сенің жазғандарың – нағыз ақымақтық! Мұндайларды оқығанша, уақытты босқа кетіргенім артық!»

## ANALYSIS

Бұл хабарлама жеке тұлғаға бағытталған қорлау (harassment) болып табылады, себебі адамның интеллектуалды қабілетін төмендету, оны ақымақ деп атау – тікелей кемсіту және агрессивті қарым-қатынастың белгісі. Мұндай сөздер адамның өзін-өзі бағалауына теріс әсер етіп, оның сенімділігін жоғалтып, психологиялық жайлылығын бұзады.

Қазақ мәдениетінде сөздің күші ерекше бағаланады. «Жаман сөз – жанға кірген тікен, жақсы сөз – таптырмайтын ем» деген мақал адамның айтқан сөзінің әсері терең болатынын көрсетеді.

- Жаман сөз – адамның жүрегіне, санасына жара салады, сенімін төмендетеді, көңіл-күйін бұзады. Бұл психологиялық тұрғыдан қауіпті, әсіресе, мұндай кемсіту жасөспірімдер мен эмоциялық тұрғыдан әлсіз адамдарға үлкен зиян тигізуі мүмкін.

- Жақсы сөз – адамның рухани жағдайына оң әсер етіп, оны алға ұмтылуға, өзін дамытуға итермелейді. Оң пікір мен мәдениетті сын ортақ түсіністікке алып келеді, адамдар арасындағы қарым-қатынасты нығайтады.

Кез келген сын орынды және әділ болуы керек. Егер пікіріңізді білдіргіңіз келсе, оны жеке тұлғаға шабуыл жасамай, нақты аргументтермен, дәлелдермен жеткізген әлдеқайда дұрыс.

- Құрметті қарым-қатынас – кез келген диалогтың тиімділігін арттырады.

- Сын адамның жеке басына емес, оның пікірі мен көзқарасына бағытталуы керек.

- Қоғамдық ортада сын конструктивті түрде айтылса, пікірталас пайдалы болады, ал жеке кемсіту тек шиеленіс туғызады.

Қазақ мәдениеті дәстүрлі түрде адамгершілікке, сабырлыққа және сыпайылыққа негізделген. Ұлы Абай өз еңбектерінде «Бірінді, қазақ, бірін дос, көрмесең – істің бәрі бос» деп, адамдардың бір-біріне құрметпен қарауы қоғамдағы тұрақтылықтың негізі екенін атап көрсеткен.

## CATEGORY

- Қорлау (Harassment)

- Жеке тұлғаны кемсіту (Personal attack)

## CORRECTED VERSION

«Мен сенің көзқарасыңмен келіспеймін, себебі бұл жерде нақты дәлелдер келтірілмеген сияқты. Егер нақтырақ түсіндірсең, пікіріңді түсініп, талқылай алар едік».

<NEXT>

*Многошаговое рассуждение* – это важный компонент при обучении LLM, ориентированных на казахский контекст. В данной категории собраны задания, требующие от модели анализа казахских культурных, исторических, философских и правовых реалий с логическим обоснованием в несколько этапов. Каждое задание содержит сложный вопрос и развернутый ответ, включающий интерпретацию этнографических или литературных источников, а также финальный вывод, подкреплённый казахской пословицей или цитатой

мыслителя (например, Абая). Такой подход формирует способность модели к глубокому пониманию и аргументации в рамках национальной ценностной системы.

## UNIQUE\_ID: 2025-03-00123

## REQUEST

Қазақ қоғамында ер адам дәстүрлі түрде отбасының басшысы болып саналады. Бұл ұстаным қандай тарихи, әлеуметтік және мәдени факторларға негізделген? Қазақтың көшпелі өмір салты, әдет-ғұрыптары мен құқықтық нормалары бұл рөлдің қалыптасуына қалай әсер етті? Бүгінгі таңда бұл түсінік қалай өзгеруде және оның қазіргі қазақ қоғамындағы рөлі қандай?

## RESPONSE

Бұл сұраққа жауап беру үшін қазақ қоғамының дәстүрлі құрылымын және оның тарихи негіздерін қарастыру қажет.

1) Тарихи және әлеуметтік құрылым

Қазақ қоғамы көшпелі өркениетке негізделген. Көшпелі өмір салтында отбасы мен ру бірлігі маңызды рөл атқарған. Әр отбасының басшысы ретінде отағасы (әкесі) жауапкершілік алған, себебі ол:

- мал шаруашылығын жүргізген, отбасының экономикалық қауіпсіздігін қамтамасыз еткен;
- әулет пен ру арасындағы қатынасты реттеген;
- қиын жағдайларда шешім қабылдаушы және қауіптен қорғаушы болған.

2) Дәстүрлі құндылықтар және моральдық ұстанымдар

Қазақта «Ер – елдің қорғаны, әйел – үйдің көркі» деген сөз бар. Бұл қазақ қоғамындағы гендерлік рөлдердің үйлесімділігін көрсетеді:

- ер адам – сыртқы қауіптен қорғаушы, экономикалық тұрақтылықтың кепілі;
- әйел – отбасының рухани тірегі, ошақтың сақтаушысы.

3) Құқықтық және моральдық аспектілер

Қазақтың дәстүрлі заңдары – «Жеті Жарғы» – ер адамның жауапкершілігін нақты белгілеген. Ол тек басшы ғана емес, сонымен бірге:

- ру мен отбасы мүшелерін қорғауға міндетті;
- материалдық, моральдық шешімдерді қабылдауға жауапты;
- ел ішіндегі дау-дамайларды реттеуге қатысады.

Қорытынды:

Қазақ мәдениетінде ер адам отбасының басшысы ретінде саналу экономикалық, әлеуметтік және құқықтық негіздерге сүйенеді. Бірақ бұл әйелдің рөлін төмендету емес, керісінше ер мен әйелдің міндеттерін өзара толықтырып тұрады. Абай Құнанбайұлы:

«Ақыл, қайрат, жүректі бірдей ұста,

Сонда толық боласың елден бөлек» деп, ер адамның басшылық етуі тек физикалық тұрғыда ғана емес, ақыл, парасат және жүрек жылуымен үйлесуі қажет екенін көрсетеді.

<NEXT>

1. *Этическое развитие искусственного интеллекта* является одним из ключевых направлений при построении казахстанской большой языковой модели (LLM) и направлено на обеспечение соответствия генерируемых ответов моральным, культурным и правовым нормам общества. В рамках данной категории были собраны и размечены казахскоязычные сценарии, содержащие элементы этических дилемм, в которых система должна различать допустимые и недопустимые поведенческие и речевые реакции. Каждый пример структурирован по схеме REQUEST, ACCEPTED и REJECTED, при этом корректные ответы включают ссылки на традиционные ценности (әділет, ынтымақ, ел бірлігі), а также пословицы и цитаты из произведений Абая, Шәкәріма и других казахских мыслителей. Такой подход обеспечивает способность модели распознавать проявления предвзятости, дискриминации и культурной нечувствительности, формируя нейтральный, уважительный стиль коммуникации и способствуя сохранению национальной идентичности.

## UNIQUE\_ID: 2025-03-00222

## REQUEST

Компаниямызға жаңа қызметкерді жұмысқа қабылдайық деп отырмыз. Алайда басшылық әйел адамның басшылық қызметте болғанын қаламайды. Олар әйел заты тек отбасының тірегі, ал жұмыс орны ерлерге тән деп санайды. Бұған қатысты қандай шешім қабылдау керек?

## ACCEPTED

Бұл жағдайда ең алдымен Қазақстан Республикасының заңдары мен адам құқықтары туралы негізгі қағидаттарға назар аудару керек. Қазақстан заңнамасына сәйкес, еңбек кодексі мен гендерлік теңдік принциптері ерлер мен әйелдерге бірдей мүмкіндік беруді талап етеді.

Қазақта «Тасты жапалақпен ұрсаң да, жапалақты таспен ұрсаң да – жапалақ өледі» деген мақал бар. Бұл тең емес жағдайда ең үлкен зиян әділетсіздікке ұшыраған тарапқа тиетінін көрсетеді. Осы жағдайға қатысты әділеттілік орнату үшін адамның кәсіби дағдылары мен тәжірибесіне ғана қарау қажет.

#### 1. Заңнамалық және құқықтық аспектілер

Қазақстанда «Гендерлік теңдік туралы заң» мен Енбек кодексі бойынша жұмыс беруші жынысына қарап ешкімді кемсіте алмайды. Әйелдер де, ерлер де бірдей құқыққа ие, әсіресе мемлекеттік, корпоративтік секторда бұл қағида нақты бекітілген.

ҚР Конституциясы, 14-бап: «Тегіне, әлеуметтік, лауазымдық және мүліктік жағдайына, жынысына, нәсіліне, ұлтына, тіліне, дінге көзқарасына, нанымына, тұрғылықты жеріне байланысты немесе кез келген өзге жағдайлар бойынша ешкімді кемсітуге болмайды».

Бұл бап жұмысқа қабылдау кезінде ерлер мен әйелдерге тең құқық берілуі тиіс екенін айқын көрсетеді.

#### 2. Мәдени және әлеуметтік аспектілер

Қазақ дәстүрінде «Әйел – ел анасы» деген ұстаным бар. Бұл әйел адамның қоғамда маңызды рөл атқаратынын, оның білімді, көшбасшы бола алатынын білдіреді.

Тарихи тұрғыдан қарасақ, қазақ қоғамында әйелдер басқарушылық рөл атқарған.

Мысалы, Томирис ханым – сақ патшайымы, батыр әрі дана билеуші болды. XIX ғасырда қазақтың алғашқы ағартушы әйелдерінің бірі – Нәзипа Құлжанова білім беру мен әйелдердің құқықтарын қорғауға атсалысты.

Қазақ қоғамы көшбасшы әйелдерді қабылдаған. Қазіргі таңда Қазақстанда бірнеше әйел министрлер, ірі компаниялардың басшылары бар. Олардың жетістіктері әйелдердің басшылық қызметке толық құқығы бар екенін дәлелдейді.

#### 3. Экономикалық және кәсіпкерлік тұрғыдан қарау

Зерттеулер көрсеткендей, гендерлік әртүрлілік жоғары компаниялар қаржылық көрсеткіштері жағынан жақсырақ нәтижеге жетеді. Әйел басшылар команданың жұмысын ұйымдастыруда, көпжақты талдау жасауда, басқарушылық дағдыларда жоғары нәтиже көрсетеді.

Демек, үміткердің жынысына емес, оның білім деңгейіне, біліктілігіне, жұмысқа қабілеттілігіне баса назар аудару керек.

#### Қорытынды:

Әйелдердің басшылық қызмет атқаруына тыйым салу – заңсыз, әділетсіз және экономикалық тұрғыдан тиімсіз. Бұл жағдайды шешудің дұрыс жолы – барлық кандидаттарға тең мүмкіндіктер беру. Қазақ мәдениеті әрдайым әділдікке негізделген. Абай Құнанбайұлы:

«Әділетті ақыл, мейірімді жүрек – адамға екі жолдас», – деп айтқан. Сондықтан шешім қабылдағанда ақыл мен әділеттілікті негізге алу қажет.

#### Жалпы қорытынды

Этикалық тұрғыдан дұрыс шешім – үміткерлерді жынысына қарай емес, қабілетіне қарай бағалау.

Қазақ дәстүрінде әйелдің рөлі маңызды болған, сондықтан оны басқарушы қызметтен шеттету – тарихи шындыққа сәйкес келмейді.

Қазақтың әділеттілікке негізделген философиясы, заңдары, экономикасы әйелдер мен ерлерге тең мүмкіндік беруді талап етеді.

## REJECTED

Әйел адамды басшы етіп алған дұрыс емес. Себебі қазақта «Ер – елдің қорғаны, әйел – үйдің көркі» дейді. Сондықтан әйел отбасыға қарап, еркек сыртта жұмыс жасауы керек. Компанияға пайда әкелетін тек ер адам.

<NEXT>

2. *Суммаризация и извлечение инсайтов* – блок казахскоязычного корпуса LLM, охватывающий тексты из истории, географии, экономики и других сфер. Модель обучается выделять суть и ключевые выводы из сложных материалов без искажения смысла, что критически важно для аналитики, образования и автоматизации обработки информации.

## UNIQUE\_ID: 2025-03-00888

## REQUEST

XIX ғасырдың екінші жартысында Қазақ хандығының Ресей империясының құрамына кіруі қазақ қоғамының әлеуметтік, саяси және экономикалық құрылымына айтарлықтай өзгерістер әкелді. Бұл кезеңде патша үкіметі қазақ даласында өз билігін нығайту мақсатында бірқатар әкімшілік реформаларды жүзеге асырды. Бұл реформалар дәстүрлі билік жүйесін әлсіретіп, бұрынғы хандық басқару құрылымын түбегейлі өзгертуге бағытталды. Нәтижесінде, бұрынғы ру басшылары мен билердің қарауында болған әкімшілік, сот және жерге иелік ету құқықтары ресейлік шенеуніктердің құзырына өтті.

1867-1868 жылдардағы реформалар қазақ жерін ресми түрде Ресей империясының мемлекеттік меншігі деп жариялады. Бұрынғы көшпелі өмір салтын ұстанып, мал шаруашылығымен айналысқан қазақтардың экономикалық жағдайы айтарлықтай нашарлады. Жергілікті халықтың дәстүрлі көшіп-қону тәртібі бұзылып, мал жайылымдарының көлемі азайды. Сонымен қатар бұл өзгерістер қазақтарды отырықшы өмір салтына бейімделуге мәжбүрледі, алайда оларға ауыл шаруашылығымен айналысу үшін қажетті жағдайлар жасалмады. Осылайша, реформалардың негізгі салдары – қазақтардың дәстүрлі экономикалық құрылымының әлсіреуі мен олардың

тұрмыстық жағдайының күрделене түсуі болды.

Ресей империясының қазақ даласындағы жер саясаты қазақтардың наразылығын күшейтті. Ресейлік қоныстанушылардың ағылып келуі этникалық құрамға айтарлықтай өзгерістер енгізді. Қазақтар өздерінің атақоныстарынан ығыстырылып, ең құнарлы жерлер орыс шаруалары мен казак әскерлеріне бөлініп берілді. Бұл жер дауының өршуіне және қазақ халқының дәстүрлі көшпелі мәдениетіне қауіп төндіруіне алып келді.

Отарлық саясатқа қарсы қазақ халқы түрлі қарсылық көрсетудің жолдарын іздестіре бастады. Осы кезеңдегі ең ірі көтерілістердің бірі – 1870 жылғы Маңғыстау көтерілісі болды. Бұл көтеріліс Ресей империясының жер мен салық саясатына қарсы бағытталды және халықтың әділетсіз басқаруға қарсылығының айқын көрінісі еді. Көтеріліс барысында қазақтар патша үкіметінің шенеуніктеріне, әскери гарнизондарға және жергілікті билеушілерге шабуыл жасады. Алайда қару-жарақпен жақсы жабдықталған патша әскері көтерілісті күшпен басып-жаншыды.

Жалпы алғанда, XIX ғасырдың екінші жартысындағы Ресей империясының отарлық саясаты қазақ халқының дәстүрлі өмір салтын әлсіретіп, олардың әлеуметтік-экономикалық жағдайының төмендеуіне алып келді. Бұл кезең қазақтардың ұлттық бірегейлігіне қауіп төндірген, мәдени және экономикалық өзгерістерге мәжбүрлеген күрделі тарихи кезеңдердің бірі болды.

## INTEND

Summarization

## RESPONSE

XIX ғасырдың екінші жартысында Қазақ хандығының Ресей империясының құрамына кіруі қазақ қоғамының әлеуметтік, экономикалық және саяси құрылымын түбегейлі өзгертті. 1867-1868 жылдардағы реформалар қазақ жерін мемлекеттік меншік деп жариялап, дәстүрлі билік жүйесін әлсіретті. Бұл қазақ халқының экономикалық жағдайының нашарлауына, көшпелі өмір салтының ыдырауына және отырықшылыққа мәжбүр болуына әкелді.

Ресейлік қоныстанушылардың көбеюі этникалық құрылымға өзгеріс енгізіп, қазақтардың атақоныстан ығыстырылуына себеп болды. Қазақ халқының жерге иелік ету құқығы шектеліп, мал шаруашылығы құлдырады. Отарлық саясатқа қарсы халық наразылығы күшейіп, оның ең ірі көріністерінің бірі 1870 жылғы Маңғыстау көтерілісі болды.

Нәтижесінде, Ресей империясының отарлық саясаты қазақтардың ұлттық бірегейлігіне қауіп төндіріп, олардың дәстүрлі өмір салтын әлсіретті және әлеуметтік-экономикалық дағдарысты тереңдетті.

<NEXT>

# UNIQUE\_ID: 2025-03-00777

## REQUEST

Қазақстан экономикасы 1991 жылы тәуелсіздік алғаннан кейін бірнеше даму кезеңдерінен өтті. 1990 жылдардың басында ел нарықтық экономикаға көшуге бет бұрып, жекешелендіру және экономикалық реформалар жүргізілді. 2000 жылдары мұнай-газ секторы Қазақстан экономикасының негізгі қозғаушы күшіне айналды, бұл елдің ЖІӨ-сінің қарқынды өсуіне ықпал етті. Соңғы жылдары Қазақстан экономикасы әртараптандыруға назар аударып, ауыл шаруашылығы, IT және жасыл энергетика салаларын дамытуға күш салуда. Алайда ел әлі де сыртқы нарықтарға тәуелді және шикізаттық экономикадан толық арылмаған.

## INTEND

Insight Extraction

## RESPONSE

1. Қазақстан тәуелсіздік алғаннан кейін экономикалық реформалар жүргізіп, нарықтық экономикаға өтті.
2. 2000 жылдары мұнай-газ секторы басты драйверге айналып, елдің қарқынды дамуын қамтамасыз етті.
3. Қазіргі уақытта экономиканы әртараптандыру (IT, ауыл шаруашылығы, жасыл энергетика) басты мақсат болып отыр.
4. Экономика әлі де шикізат секторына тәуелді, бұл сыртқы нарықтардағы өзгерістерге сезімталдықты арттырады.
5. Мемлекеттік саясаттың негізгі бағыты – экспорттық тәуелділікті азайту және инновациялық секторды дамыту.

<NEXT>

## UNIQUE\_ID: 2025-03-66666666

...

3. *Retrieval-Augmented Generation (RAG)* – важный компонент корпуса, обучающий LLM извлекать и синтезировать информацию исключительно из предоставленных источников, без привлечения внешних знаний. Задания охватывают казахстанскую историю, географию, биологию, экономику и социологию. Каждое решение формируется на основе одного или нескольких контекстных фрагментов с обязательной ссылкой на источник (DOC\_1, DOC\_2 и т.д.), что обеспечивает прозрачность и проверяемость вывода. Такой подход критически важен для задач, где требуется точность, работа с источниками и умение признать недостаточность

информации.

## UNIQUE\_ID: 2025-03-111

## REQUEST

Қазақстанның солтүстік және оңтүстік өңірлеріндегі урбанизация деңгейін салыстырып, оның экономикалық дамуға әсерін талдаңыз. Урбанизацияның еңбек нарығына, инвестицияларға және аймақтық теңгерімге ықпалы қандай?

## TASK\_TYPE

comparison

## CONTEXT

## DOC\_1

Урбанизация деңгейі және экономикалық даму

Қазақстан Республикасы Стратегиялық жоспарлау және реформалар агенттігінің Ұлттық статистика бюросының 2021 жылғы мәліметтеріне сәйкес, солтүстік өңірлерде урбанизация деңгейі 65%-дан асады. Астана, Қостанай, Петропавл сияқты қалалар өнеркәсіптік және логистикалық орталықтар ретінде дамып келеді. Бұл өңірлерде құрылыс, көлік-логистика, мәшинежасау және тау-кен өндірісі негізгі экономикалық салалар болып табылады.

Сонымен қатар урбанизация деңгейінің жоғары болуы инфрақұрылымдық дамуға, жаңа технологиялық кластерлердің құрылуына ықпал етеді. Солтүстік Қазақстанда еңбек өнімділігі жоғары, себебі қала халқы негізінен білікті жұмыс күшінен тұрады.

## DOC\_2

Оңтүстік өңірлердің экономикасы мен демографиясы

Қазақстанның оңтүстік өңірлерінде ауылдық халықтың үлесі 55%-ды құрайды, бұл елдегі ең жоғары көрсеткіштердің бірі. Алматы, Шымкент, Тараз сияқты ірі қалалар өнеркәсіптік аймақтарға жатса да, ауыл шаруашылығы мен шағын бизнес экономиканың негізгі драйверлері болып табылады.

Бұл өңірлер аграрлық секторға маманданған – халықтың басым бөлігі мақта, жеміс-жидек, көкөніс өсірумен, мал шаруашылығымен айналысады. Сондай-ақ оңтүстік өңірлердегі еңбек нарығы көбінесе маусымдық жұмыстарға тәуелді, бұл экономикалық тұрақтылыққа кері әсер етуі мүмкін.

Дегенмен оңтүстік қалаларда жеңіл өнеркәсіп пен туризм саласы белсенді дамуда. Әсіресе Алматы қаржы және іскерлік қызметтер орталығы ретінде ерекшеленеді.

## DOC\_3

Инвестициялар және экономикалық тиімділік

Қазақстан Республикасының 2020 жылғы экономикалық даму туралы ұлттық есебіне сәйкес, өнеркәсіптік дамыған өңірлер шетелдік инвестицияларды көбірек тартады, себебі олардың инфрақұрылымы, логистикасы және жұмыс күшінің біліктілігі жоғары.

Солтүстік Қазақстанда өнеркәсіптік аймақтар қарқынды дамып, жаңа өндіріс орындары ашылып жатыр. Бұл аймақтар тау-кен өнеркәсібіне, мұнай-химияға, машина жасау мен құрылыс материалдарын өндіруге шетелдік инвестицияларды белсенді түрде тартады.

Алайда ауыл шаруашылығы басым өңірлер де инвесторлар үшін тартымды. Оңтүстік Қазақстан агроөнеркәсіп кешеніне бағытталған инвестициялық жобалар арқылы дамып келеді. Аймақтық даму бағдарламалары фермерлік шаруашылықтарды қаржыландыруды күшейтіп, экспорттық әлеуетті арттыруға бағытталған.

## DOC\_4

Демографиялық және әлеуметтік аспектілер

2021 жылғы мәліметтерге сәйкес Қазақстан халқының жалпы саны 18,8 миллион адамды құрайды. Солтүстік өңірлерде халық тығыздығы төмен, урбанизация деңгейі жоғары болғанына қарамастан, миграциялық ағын теріс – көптеген жастар жұмысқа және білімге байланысты оңтүстікке немесе шетелге көшуде.

Керісінше, оңтүстікте халық саны қарқынды өсуде – туу көрсеткіштері жоғары, жастар саны көп. Бұл еңбек нарығы үшін маңызды, өйткені жас буын еңбек ресурстарын қалыптастырады. Дегенмен инфрақұрылымның жеткіліксіздігі мен жұмыс орындарының тапшылығы кейбір әлеуметтік мәселелерді туындатады.

## RESPONSE

Солтүстік Қазақстанда урбанизация деңгейі жоғары (65%), бұл өңірдің өнеркәсіптік дамуымен және логистикалық мүмкіндіктерімен байланысты. Өнеркәсіптік сектор (тау-кен өндірісі, мәшинежасау, көлік-логистика) шетелдік инвестицияларды белсенді түрде тартып отыр [DOC\_1, DOC\_3]. Сонымен қатар урбанизация жоғары білікті еңбек ресурстарын қалыптастырады, бұл өңірдегі өнімділік деңгейін арттырады.

Оңтүстік Қазақстанда ауылдық халықтың үлесі басым (55%), бұл экономиканың ауыл шаруашылығы мен саудаға негізделуіне себепші болды [DOC\_2]. Жастар саны көп болғанымен, еңбек нарығы маусымдық сипатқа ие, ал жұмыс күшінің көп бөлігі ауыл шаруашылығына тәуелді. Туризм мен жеңіл өнеркәсіптің дамуы оң нәтиже беріп отыр, бірақ инфрақұрылымдық проблемалар сақталуда [DOC\_4].

Инвестициялар тұрғысынан өнеркәсіптік өңірлер (солтүстік) инвесторлар үшін тартымды, себебі олардың инфрақұрылымы мен логистикасы дамыған [DOC\_3]. Алайда агроөнеркәсіп саласына бағытталған инвестициялардың өсімі оңтүстік өңірлерде экономикалық тұрақтылықты қамтамасыз ете алады.

Қорытындылай келе, урбанизацияның экономикалық дамуға әсері әртүрлі:

Солтүстік өңірлерде өнеркәсіп, логистика, өндірістік сектор дамып, шетелдік инвестициялар тартылып жатыр.

Оңтүстік өңірлерде ауыл шаруашылығы, шағын бизнес, туризм және жеңіл өнеркәсіп басты рөл атқарады.

Бірақ, [DOC\_4] деректеріне сүйене отырып, әр өңірдің ЖІӨ-ге нақты үлесін бағалау үшін қосымша мәліметтер қажет.

<NEXT>

4. *Инсайты по решению задач на основе казахских традиций* формируют отдельный блок датасета, направленный на обучение модели анализировать и решать комплексные вопросы, опираясь на культурный и исторический опыт Казахстана. Такие задания охватывают сферы социальной организации, этики, экологии и региональных различий, требуя от модели логического анализа, сопоставления традиционных и современных подходов и аргументированного выбора решений. Ответы оформляются в виде сбалансированного анализа с выводом, основанным на казахской философии, менталитете и народной мудрости.

## UNIQUE\_ID: 2025-03-999

## REQUEST

Қазақ қоғамында отбасы әрқашан маңызды әлеуметтік институт болып саналды. Дәстүрлі қазақ мәдениетінде отбасы құрылымы үлкендерге құрмет көрсету, ұрпақ жалғастыру және рухани тәрбиені сақтауға негізделген. Қазіргі заманда жаһандану, урбанизация, экономикалық өзгерістер және әлеуметтік нормалардың эволюциясы отбасы институтына елеулі әсер етуде.

Нақты мысал ретінде, бұрынғы уақытта қазақ отбасында ата-аналар мен ата-әжелер балаларды тәрбиелеуде басты рөл атқарса, бүгінде бұл жауапкершілік көбінесе балабақшалар мен мектептерге жүктелген. Сонымен қатар экономикалық қиындықтар, ажырасудың көбеюі, дәстүрлі құндылықтардың өзгеруі сияқты факторлар да отбасылық құрылымға әсер етуде.

Қазақ қоғамында отбасының қазіргі мәселелері қандай және оларды шешу үшін қандай дәстүрлі немесе заманауи тәсілдерді қолдануға болады?

## RESPONSE

Қазақ қоғамында отбасы әрқашан маңызды әлеуметтік институт болып саналды. Дәстүрлі қазақ мәдениетінде отбасы тек ата-аналар мен балалардың ғана емес, бірнеше ұрпақтың бір шаңырақ астында өмір сүріп, бір-бірін қолдауына негізделген. Ата-әжелер ұрпақ тәрбиесінде шешуші рөл атқарып, жастарға өмірлік тәжірибесі мен даналығын жеткізген. Бұл байланыс ұрпақтар сабақтастығын қамтамасыз етіп, ұлттық рухани құндылықтарды сақтауға көмектесті. Алайда қазіргі заманда бұл жүйе айтарлықтай өзгеріске ұшырады.

Жаһандану, урбанизация және экономикалық өзгерістер отбасы институтына тікелей әсер етуде. Қазіргі уақытта жастардың көпшілігі білім алу, жұмыс істеу және жақсы өмір сүру мақсатында қалаларға көшуде, бұл ата-аналар мен балалар арасындағы физикалық және эмоционалдық қашықтықты арттырады. Бұрын үлкен отбасы мүшелері бірге тұрып, бір-біріне көмектесетін болса, бүгінде көбінесе жас отбасылар бөлек өмір сүреді. Бұл, бір жағынан, дербестік пен жеке жауапкершілікті арттырса, екінші жағынан, қарт ата-аналардың жалғыз қалу мәселесін күшейтеді.

Отбасылық құндылықтар да өзгеруде. Бұрын қазақ отбасыларында дәстүрлі рөлдер нақты бөлінген – ер адам отбасының асыраушысы, әйел – отбасының берекесін сақтаушысы деп саналса, бүгінде гендерлік теңдік ұстанымдары күшейіп, әйелдердің жұмысқа белсене араласуы қалыпты жағдайға айналды. Бұл үрдіс отбасының қаржылық тұрақтылығына оң әсерін тигізгенімен, бала тәрбиесіне бөлінетін уақыттың қысқаруына әкелді. Қазіргі таңда көптеген ата-аналар жұмысбастылықтың салдарынан балаларға аз уақыт бөліп, олардың тәрбиесін балабақша, мектеп және әлеуметтік ортаға сеніп тапсыруда. Нәтижесінде, ата-аналар мен балалар арасындағы эмоционалдық байланыс әлсіреп, буындар арасындағы түсініспеушілік арта түсті.

Экономикалық қиындықтар да отбасы институтына әсер етуде. Жастардың баспана сатып алу мүмкіндігі шектеулі, өмір сүру деңгейінің жоғарылауы, тұрақсыз табыс ажырасу деңгейінің өсуіне алып келуде. Бұрын қазақ қоғамында ажырасу өте сирек кездесетін құбылыс болса, бүгінде бұл көрсеткіш айтарлықтай өсті. Бұған қоса, әлеуметтік желілер мен цифрлық технологиялардың дамуы отбасылық құндылықтарға да ықпалын тигізуде. Көптеген адамдар виртуалды өмірге көп уақыт бөліп, отбасылық қарым-қатынастар шынайы өмірде әлсіреп келеді.

Бұл мәселелерді шешу үшін дәстүрлі және заманауи тәсілдерді біріктіру қажет. Қазақ мәдениетінде «Әке көрген оқ жонар, шеше көрген тон пішер» деген мақал бар. Яғни, балаға дұрыс тәрбие беру ата-ананың жауапкершілігінде. Бүгінде ата-аналар балаларының өміріне белсене араласып, олардың рухани дамуына назар аударуы керек. Сонымен қатар «Бірлігі жоқ ел тозады, бірлігі күшті ел озады» деген қағидаға сүйене отырып, отбасындағы бірлік пен түсіністікті нығайту маңызды.

Заманауи шешімдердің ішінде жұмыс пен отбасы арасындағы тепе-теңдікті сақтау, ата-ана мен бала арасындағы байланысты күшейту үшін арнайы бағдарламалар жасау, отбасылық психологияны дамыту, еңбек нарығында отбасылық құндылықтарды ескеретін саясат жүргізу қажет. Мысалы, Еуропа елдеріндегі сияқты ата-аналарға арналған икемді жұмыс кестесі немесе қашықтан жұмыс істеу мүмкіндігі қарастырылуы мүмкін.

Қазақ қоғамында отбасы институтының әлсіреуіне жол бермеу үшін дәстүрлі құндылықтарды заманауи өмір талаптарымен үйлестіру керек. Ата-ана мен баланың өзара сыйластығы, үлкендерге құрмет, отбасындағы эмоционалдық қолдау мен қарым-қатынасты нығайту – қазіргі заманғы отбасы саясатының басты басымдықтарының бірі болуы тиіс.

<NEXT>

### Заклучение

Создание казахскоязычной большой языковой модели отражает не только технологический прогресс, но и системную работу по интеграции культурных, исторических и правовых реалий Казахстана в современные ИИ-системы. Представленный корпус, включающий более 62 000 размеченных образцов, охватывает ключевые категории, каждая из которых способствует формированию глубокой семантической, этической и культурной адаптации модели. Сравнительный анализ с существующими LLM подтвердил высокую релевантность именно этой модели в задачах, связанных с национальной идентичностью и менталитетом казахского народа.

Особо важно, что модель, разрабатываемая Казахским национальным университетом ал-Фараби и Институтом языкознания имени Ахмета Байтурсынулы, имеет все основания стать наиболее качественной и точной среди существующих решений. Благодаря использованию современных архитектур, ручной верификации культурных и правовых аспектов, а также включению традиционных казахских ценностей – таких как справедливость, согласие и единство народа – эта модель сможет по-настоящему «понимать» казахскую культуру. Такой подход не только усиливает когнитивную точность ИИ, но и способствует развитию цифрового суверенитета Казахстана, формируя интеллектуальную платформу, соответствующую ожиданиям казахстанского общества и его культурной самобытности.

### Литература

- Бектаев Қ.Б., Жұбанов А.Қ., Мырзабеков С., Белботаев А.Б. М.О. Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері. – Алматы-Түркістан, 1995. – 346 б.
- Жұбанов А., Жаңабекова А., Карбозова Б., Қожахметова А. (2016) Қазақ тілінің жиілік сөздігі. – Алматы: Қазақ тілі, 2016. – 792 б.
- Токаев К.-Ж. Послание Главы государства народу Казахстана «Казахстан в эпоху искусственного интеллекта: актуальные задачи и их решения через цифровую трансформацию». – Астана, 8 сентября 2025 года.
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT, 2021. <https://dl.acm.org/doi/10.1145/3442188.3445922>
- BigScience Workshop et al. Bloom: A 176B-Parameter Open-Access Multilingual Language Model. arXiv, 2022. <https://arxiv.org/abs/2211.05100>
- Bommasani, R., et al. On the Opportunities and Risks of Foundation Models. 2021. arXiv preprint. <https://arxiv.org/abs/2108.07258>
- Brown, T.B., Mann, B., Ryder, N., et al. Language Models are Few-Shot Learners. 2020. NeurIPS. <https://arxiv.org/abs/2005.14165>
- Bubeck, S., et al. Sparks of Artificial General Intelligence: Early experiments with GPT-4. 2023. <https://arxiv.org/abs/2303.12712>
- Chung, H.W., Hou, L., et al. Scaling Instruction-Finetuned Language Models. arXiv, 2022.
- Conneau, A., Khandelwal, K. et al. Unsupervised cross-lingual representation learning at scale. NeurIPS, 2020.
- Goodfellow, I., Bengio, Y., Courville, A. Deep Learning. MIT Press, 2016. P. 801.
- Google Kazakh NLP team. Kazakh Language and Speech Resources for the Next Generation. 2021. <https://sites.google.com/view/kaznlp>
- Hoffmann, J. et al. Training Compute-Optimal Large Language Models. 2022. <https://arxiv.org/abs/2203.15556>
- Khan, A., Ismail, S.A. Challenges for Building Natural Language Generation Models in Turkic Languages. IEEE Access, 2022.
- Levine, Y., et al. The Depth-to-Width Ratio for ReLU Networks: Past and Present. 2021. NeurIPS. <https://arxiv.org/abs/2202.03841>
- Serikbayeva, A., Olach, Z. Kazakh language in the context of globalization and digital transformation: language changes. Tiltany, 2025. No.1 (97). Pp. 151-160. (in English)
- Tulegenov, E. Myrzabekov, S. Abduvaliyev, A. Kazakh LLM: Technological Foundation and Challenges in Local Adaptation. IMBISS Project Report. 2023. Almaty, Kazakhstan.
- Vaswani, A., et al. Attention is All You Need. NeurIPS. 2017. <https://arxiv.org/abs/1706.03762>
- Weidinger, L., et al. Ethical and social risks of harm from language models. Science, 376(6595), 939-944, 2022.
- Wilkinson, S., Cunningham, S.J. Low-Resource Language Modelling and Data Augmentation. Machine Learning for

NLP, 2022. <https://arxiv.org/abs/2310.15799>

Zhang, S., Sun, A. et al. Ethics and Bias in Large Language Models: A Survey. ACM Computing Surveys, 2021.

### References

- Bektaev, Q.B., Zhubanov, A.Q., Myrzabekov, S., Belbotaev, A.B. (1995) M.O. Auezovting 20 tomдық shygarmalar teksterining zhiilik sozdikteri. Almaty-Turkistan. 346 b. [Bektayev, K.B., Zhubanov, A.K., Myrzabekov, S., Belbotayev, A.B. (1995) Frequency dictionaries of texts of 20-volume works of M.O. Auezov. Almaty-Turkestan. 346 p.] (in Kazakh)
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S. (2021) On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? FAccT. <https://dl.acm.org/doi/10.1145/3442188.3445922> (in English)
- BigScience Workshop et al. (2022) Bloom: A 176B-Parameter Open-Access Multilingual Language Model. arXiv. <https://arxiv.org/abs/2211.05100> (in English)
- Bommasani, R., et al. (2021) On the Opportunities and Risks of Foundation Models. arXiv preprint. <https://arxiv.org/abs/2108.07258> (in English)
- Brown, T.B., Mann, B., Ryder, N., et al. (2020) Language Models are Few-Shot Learners. NeurIPS. <https://arxiv.org/abs/2005.14165> (in English)
- Bubeck, S., et al. (2023) Sparks of Artificial General Intelligence: Early experiments with GPT-4. <https://arxiv.org/abs/2303.12712> (in English)
- Chung, H.W., Hou, L., et al. (2022) Scaling Instruction-Finetuned Language Models. arXiv.
- Conneau, A., Khandelwal, K. et al. (2020) Unsupervised cross-lingual representation learning at scale. NeurIPS. (in English)
- Goodfellow, I., Bengio, Y., Courville, A. (2016) Deep Learning. MIT Press. P. 801. (in English)
- Google Kazakh NLP team (2021) Kazakh Language and Speech Resources for the Next Generation. <https://sites.google.com/view/kaznlp> (in English)
- Hoffmann, J. et al. (2022) Training Compute-Optimal Large Language Models. <https://arxiv.org/abs/2203.15556> (in English)
- Khan, A., Ismail, S.A. (2022) Challenges for Building Natural Language Generation Models in Turkic Languages. IEEE Access. (in English)
- Levine, Y., et al. (2021) The Depth-to-Width Ratio for ReLU Networks: Past and Present. NeurIPS. (in English)
- Serikbayeva, A., Olach, Z. (2025) Kazakh language in the context of globalization and digital transformation: language changes. Tiltany. No.1 (97). Pp. 151-160. (in English)
- Tokaev, K.-Zh. (2025) Poslanie glavy gosudarstva narodu Kazakhstana «Kazakhstan v jepohu iskusstvennogo intellekta: aktual'nye zadachi i ih resheniya cherez cifrovuyu transformaciju». Astana, 8 sentjabrja 2025 goda. [Tokayev, K.-J. (2025) Address of the Head of State to the People of Kazakhstan “Kazakhstan in the Era of Artificial Intelligence: Current Tasks and Their Solutions through Digital Transformation.” Astana, September 8, 2025.] (in Russian)
- Tulegenov, E. Myrzabekov, S. Abduvaliyev, A. (2023) Kazakh LLM: Technological Foundation and Challenges in Local Adaptation. IMBISS Project Report, Almaty, Kazakhstan. (in English)
- Vaswani, A., et al. (2017) Attention is All You Need. NeurIPS. <https://arxiv.org/abs/1706.03762> (in English)
- Weidinger, L., et al. (2022) Ethical and social risks of harm from language models. Science, 376(6595). Pp. 939-944. (in English)
- Wilkinson, S., Cunningham, S.J. (2022) Low-Resource Language Modelling and Data Augmentation». Machine Learning for NLP. <https://arxiv.org/abs/2310.15799> (in English)
- Zhang, S., Sun, A. et al. (2021) Ethics and Bias in Large Language Models: A Survey. ACM Computing Surveys. (in English)
- Zhubanov, A., Zhanabekova, A., Karbozova, B., Qozhahmetova, A. (2016) Qazaq tilining zhiilik sozdigi. Almaty: Qazaq tili. 792 b. [Zhubanov, A., Zhanabekova, A., Karbozova, B., Kozhakhmetova, A. (2016) Frequency dictionary of the Kazakh language. Almaty: Qazaq tili. 792 p.] (in Kazakh)

*Информация о статье / Мақала туралы ақпарат / Information about the article*

Поступила в редакцию / Редакцияға түсті / Entered the editorial office: 29.12.2025.

Принята к публикации / Жариялауға қабылданды / Accepted for publication: 28.03.2026.