МРНТИ 20.53.19

10.51889/2959-5894.2026.93.1.022

Н.А. Тойганбаева[1*], Ғ. Әбдіманап[2], А. Муса[1], Н. Абдурахмонова[3]

[1]Әл-Фараби атындағы Қазақ Ұлттық университеті, Алматы қ., Қазақстан
[2]«ҚазМұнайГаз Инжиниринг» ЖШС, Астана қ., Қазақстан
[3]Мирзо Ұлықбек атындағы Өзбекстан Ұлттық Университеті, Ташкент қ., Өзбекстан
[*]*e-mail: nazkon@gmail.com*

ҚАЗАҚ ТІЛІНЕ АРНАЛҒАН LLM ҮШІН ДЕРЕКТЕРДІ OCR АРҚЫЛЫ ДАЙЫНДАУ

Аңдатпа

Соңғы жылдары жасанды интеллект және үлкен тілдік модельдер (LLM) қарқынды дамуда. Бұл модельдердің тиімділігі оларды үйретуге пайдаланылған деректер сапасына тәуелді. Қазақ тіліне арналған құрылымдалған мәтіндік ресурстардың тапшылығы LLM дамытуда қиындық тудырады. Мақалада қазақ тіліндегі мәтіндерді OCR технологиясы арқылы цифрландыру және олардан JSON форматында сапалы датасет жасау қарастырылады. Жұмыстың мақсаты – қазақ мәтіндерін автоматты өңдеп, LLM оқытуға жарамды құрылымдалған деректер дайындау. Бұл үшін сканерленген құжаттар жиналып, Tesseract OCR арқылы танылып, JSON құрылымына келтірілді. Нәтижесінде 37 062 құжат өңделіп, LLaMA3.2 3B моделін қазақ тілінде оқытуға қолданылды. Модель ұлттық стильді меңгеріп, поэтикалық мәтіндер құра алды. Train/loss графигі оқыту тұрақтылығын көрсетті.

Түйін сөздер: OCR, JSON, датасет, қазақ тілі, LLM, тану, өңдеу, модель.

Н.А. Тойганбаева[1], Г. Абдиманап[2], А. Муса[1], Н. Абдурахмонова[3]
[1]Казахский национальный университет имени аль-Фараби, г. Алматы, Казахстан
[2]ТОО «КазМунайГаз Инжиниринг», г. Астана, Казахстан
[3]Национальный университет Узбекистана имени Мирзо Улугбека, г. Ташкент, Узбекистан

ПОДГОТОВКА ДАННЫХ С ПОМОЩЬЮ OCR ДЛЯ LLM НА КАЗАХСКОМ ЯЗЫКЕ

Аннотация

В последние годы наблюдается стремительное развитие искусственного интеллекта и крупных языковых моделей (LLM). Эффективность таких моделей во многом зависит от качества данных, использованных для их обучения. Недостаток структурированных текстовых ресурсов на казахском языке представляет собой серьёзную проблему для развития LLM. В данной статье рассматривается процесс оцифровки казахскоязычных текстов с помощью технологии OCR и создание на их основе качественного датасета в формате JSON. Цель работы – автоматическая обработка казахских текстов и подготовка структурированных данных, пригодных для обучения LLM. Для этого были собраны отсканированные документы, распознаны с помощью Tesseract OCR и преобразованы в структуру JSON. В результате было обработано 37 062 документа, которые использовались для обучения модели LLaMA 3.2 3B на казахском языке. Модель успешно освоила особенности национального стиля и смогла генерировать поэтические тексты. График train/loss продемонстрировал стабильность обучения.

Ключевые слова*:* OCR, JSON, датасет, казахский язык, LLM, распознавание, обработка, модель.

N.A. Toiganbayeva[1], G. Abdimanap[2], A. Musa[1], N. Abdurakhmonova[3]
[1]Al-Farabi Kazakh National University, Almaty, Kazakhstan
[2]KazMunayGas Engineering LLP, Astana, Kazakhstan
[3]National University of Uzbekistan named after Mirzo Ulugbek, Toshkent, Uzbekistan

PREPARATION OF DATA WITH THE HELP OF OCR FOR LLM IN KAZAKH LANGUAGE

Abstract

In recent years, artificial intelligence and large language models (LLMs) have undergone rapid development. The effectiveness of these models largely depends on the quality of the training data. However, the scarcity of structured text resources in the Kazakh language poses a significant challenge for LLM

development. This paper explores the digitization of Kazakh-language texts using OCR technology and the creation of a high-quality dataset in JSON format. The main objective of the study is to automatically process Kazakh texts and prepare structured data suitable for training LLMs. For this purpose, scanned documents were collected, processed using Tesseract OCR, and converted into a structured JSON format. As a result, 37,062 documents were processed and used to train the LLaMA 3.2 3B model in the Kazakh language. The model demonstrated an understanding of national linguistic style and was capable of generating poetic texts. The train/loss graph indicated stable training performance.

Keywords: OCR, JSON, dataset, Kazakh language, LLM, recognition, processing, model.

Кіріспе

Үлкен тілдік модельдер (LLM, Large Language Models) қазіргі уақытта жасанды интеллект жүйелерінің өзегі саналады. Үлкен тілдік модельдер мәтінді түсіну, генерациялау, аудару, сұрақтарға жауап беру, диалог жүргізу және күрделі лингвистикалық тапсырмаларды орындау қабілеттеріне ие. Бірақ аталған модельдердің тиімділігі оларды қандай деректермен және қандай тілде үйретілгеніне тікелей байланысты. Қазіргі таңда әлемдік LLM-дардың басым бөлігі ағылшын тіліне негізделген. Қазақ тіліне бейімделген LLM құру үшін алдымен үлкен, әртүрлі және сапалы мәтіндік деректер жинау керек. Цифрлық дәуірде жасанды интеллект, үлкен тілдік модельдер және жалпы тілдік технологиялар үшін сандық мәтіндердің мол болуы – негізгі талаптардың бірі. Алайда қазақ тілі үшін бұл бағытта елеулі қиындықтар бар. Қазақ тіліндегі цифрланған, құрылымданған және өңдеуге дайын мәтіндер көлемі басқа ірі тілдермен салыстырғанда өте аз. Кез келген жасанды интеллект жүйесі үйретілген деректердің сапасынан аспайтыны белгілі. Сондықтан деректердің мазмұны, тазалығы, алуан түрлілігі мен тілдік дұрыстығы – LLM өнімділігін айқындайтын негізгі факторлар. Қазақ тіліндегі дереккөздерден алынған мәліметтер әр түрлі форматта болады. Қазақ тілінде LLM жасау үшін датасет жинау – ең маңызды және жауапты қадам. Қазақ тілінде цифрланған мәтіндердің тапшылығын ескерсек, OCR (OCR, Optical Character Recognition) технологиясының маңызы өте жоғары. OCR – бұл мәтін бейнесін (мысалы, сканерленген құжаттар, фотосуреттер, скриншоттар) машиналық түрде оқуға болатын мәтінге айналдыратын технология. OCR мәтіндік мазмұнды цифрландырудың, жазбаша ақпаратты машинада оқылатын пішімдерге түрлендірудің және кең қолжетімділікті қамтамасыз етудің таптырмас технологиясына айналды. OCR технологиясы – қазақ тіліндегі мәтіндік деректерді жедел әрі тиімді цифрлау құралы. Оптикалық таңбаларды тану LLM моделін үйретуде корпусты сапалы кеңейтіп, қазақ тілінің цифрлық өрісін кеңейтуге үлкен үлес қосады. Осыдан зерттеу тақырыбының өзектілігі – қазақ тілін цифрлық кеңістікке бейімдеу, сапалы мәтіндік деректерді қолжетімді ету және тілдік модельдердің даму мүмкіндіктерін кеңейту қажеттілігімен анықталады. Бұл бағыттағы ғылыми жұмыстар тек техникалық емес, сонымен қатар мәдени, білім беру және ұлттық тіл саясаты тұрғысынан да ерекше мәнге ие.

İ. Çetıntas зерттеу жұмысында тарихи түрік құжаттарын OCR арқылы өңдеуде кездесетін қиындықтарды шешу туралы қарастырылған. Эксперимент 1930–1970 жылдар аралығындағы 30 түрлі газет бетіне жүргізілген. Алынған сөздер NLP құралдары арқылы тексеріліп, ұсынылған әдіс OCR дәлдігін орта есеппен 18%-ға арттыратыны анықталған. Бұл алдын ала бейнеөңдеудің тарихи мәтіндерді танудағы маңызын дәлелдейді [1].

Y. Chen қолмен жазылған мәтінді тану сапасын арттыру және доменге бейімделу мүмкіндігін кеңейту мақсатында OCR моделін тілдік модельмен біріктіруді ұсынған. TrOCR және CharBERT модельдерінің үйлесімі визуалды және лингвистикалық ақпаратты тиімді біріктіре отырып, мәтіннің түпнұсқалығын сақтайтынын көрсеткен [2].

T. Nguyen мақаласында тарихи құжаттардағы OCR нәтижелерінің сапасын арттырудың маңыздылығы және оның ақпаратты іздеу мен табиғи тілді өңдеу жүйелеріне әсері туралы қарастырылған. Пост-OCR өңдеудің мәселелері сипатталып, оның типтік пайплайны мен заманауи әдістеріне шолу жасалған. Сонымен қатар, бағалау метрикалары, қолжетімді

деректер жиынтықтары және қолдануға ыңғайлы құралдар ұсынылып, болашақ зерттеу бағыттары айқындалған [3].

Соңғы бес жылда қазақ тіліне арналған деректер жиынтықтарын (датасеттерді) құру бағытында бірқатар маңызды бастамалар жүзеге асырыла бастады. Жинақталып жатқан бұл деректер қорлары қазақ тілінің технологиялық дамуына серпін беріп, оны цифрлық экожүйеге толыққанды енгізудің алғышартына айналуда. 1-кестеде қазақ тіліндегі датасеттерге шолу жасалған.

Кесте 1. Қазақ тіліндегі датасеттерге шолу

№	Датасет атауы	Авторлар	Сипаттамасы	Сілтеме
1	KazNERD [4]	Yeshpanov Y., Khassanov H., Atakan V.	Қорытынды деректер жиынтығы 112 702 сөйлем мен 25 атаулы мән санатына жататын 136 333 аннотациядан тұрады.	https://issai.nu.edu.kz/ru/kaznerd-rus/ (11.07.2025)
2	Kazakh Speech Corpus (KSC2) [5]	Mussakhojayeva S., Khassanov Y., Huseyin A.V.	Жалпы алғанда, KSC2-де 600 мыңнан астам мәлімдемені қамтитын шамамен 1,2 мың сағаттық жоғары сапалы транскрипцияланған деректер бар.	https://issai.nu.edu.kz/kz-speech-corpus/ (11.07.2025)
3	KOHTD [6]	Toiganbayeva N, Kasem M., Abdimanap G., Bostanbekov K., Abdallah A, Alimova A., Nurseitov D.	Жинақ 3000 қолмен жазылған емтихан парағын, 140 335-тен астам кескінделген жолақтарды қамтиды және шамамен 922 010 таңбадан тұрады.	https://github.com/abdoelsayed2016/KOHTD (11.07.2025)
4	KazSAnDRA [7]	Yeshpanov R., Varol H. A.	Қазақ тіліндегі пікірталдау (sentiment analysis) үшін арнайы әзірленген және өз саласында алғаш рет ашық түрде жарияланған ең ірі деректер жиынтығы. Бұл жиынтық әртүрлі дереккөздерден алынған 180 064 пікірді қамтиды және әр пікірге 1-ден 5-ке дейінгі сандық баға берілген, бұл тұтынушылардың көзқарасын сандық түрде бейнелеуге мүмкіндік береді.	https://issai.nu.edu.kz/2024/07/01/kazsandra-kazakh-sentiment-analysis-dataset-of-reviews-and-attitudes/ (11.07.2025)
5	QThink-Task Dataset [8]	Kadyrbek N., Tuimebayev Z., Mansurova M., Viegas V.	Бұл LLM модельдеріне арналған, түрлі пайымдау мен тілдік түсіну қабілеттерін бағалайтын, қазақстандық контекске негізделген көптапсырмалы деректер жиынтығы.	https://huggingface.co/datasets/nur-dev/QThink-Task (11.07.2025)

Датасет – үлкен тілдік модельдердің құрылымындағы негізгі компоненттердің бірі және модельдің меңгеретін біліміне, тілдік құрылымына, логикалық пайымына және мазмұн сапасына тікелей әсер етеді. Сондақтан деректерді мұқият іріктеу мен сапалы өңдеу – LLM әзірлеу үдерісіндегі шешуші қадамдардың бірі болып табылады.

Зерттеу әдіснамасы

OCR (Optical Character Recognition, оптикалық таңбаларды тану) – бұл мәтін бар кескіндерді машинамен оқуға болатын мәтіндік деректерге түрлендіру технологиясы. Бұл технология құжаттарды цифрландыруда, қолжазбаларды тануда және баспа материалдарын өңдеуде кеңінен қолданылады. OCR жұмысының негізгі принципі – кескінді талдау, мәтіндік аймақтарды бөліп алу және оларды мәтіндік таңбаларға айналдыру. Бұл компьютерлік көру алгоритмдері, атап айтқанда мәтінді сегменттеу, таңба белгілерін шығару және оларды жіктеу арқылы жүзеге асырылады.

Осы зерттеу жұмысында OCR жүзеге асыру үшін Tesseract OCR құралы таңдалды. Tesseract-ті таңдаудың себептері:

- Қазақ тілін қолдау: Tesseract – қазақ тіліндегі мәтінді тану үшін тілдік пакеттерді ұсынатын санаулы ашық кітапханалардың бірі, бұл жоба үшін басты талап болып табылады.
- Ашық бастапқы код: ашық бастапқы кодқа ие құрал ретінде Tesseract лицензиялық шектеулерден бос, бұл оны жобаға еш қосымша шығынсыз біріктіруге мүмкіндік береді.
- Жоғары өнімділік: Tesseract алдын ала кескіндерді өңдеу арқылы шу мен бұрмаланулардың әсерін азайтып, мәтінді жоғары дәлдікпен тануды қамтамасыз етеді. Tesseract қазақ тілінен басқа да көптілдермен жұмыс жасауға мүмкіндігі бар, бұл көптілді құжаттармен жұмыс істегенде маңызды.
- Икемділік пен баптау мүмкіндігі: Tesseract тілдік модельдер мен OCR параметрлерін баптауға мүмкіндік береді, бұл ең жақсы нәтижелерге қол жеткізуге көмектеседі.
- Белсенді әзірлеушілер қауымдастығы: құралды қолданушылардың кең қауымдастығы мен үнемі жаңартылып отыруы сенімді және ыңғайлы шешімге қабылдауға үлес қосады.

Мәтінді тану дәлдігін арттыру үшін кескіндерді алдын ала өңдеу үдерісі қолданылды. Бұл үдерісте суреттерді сұр түске түрлендіру, морфологиялық өңдеу арқылы көлеңкелерді жою және жарықтылықты қалыпқа келтіру қадамдары қамтылды. Бұл кезеңдер артефактілерді жоюға және мәтіндік аймақтардың контрастылығын арттыруға мүмкіндік беріп, OCR сапасын едәуір жақсартады. Құжаттардың көптілді сипатын ескере отырып, Tesseract OCR баптауларында kaz+rus+eng тілдік пакеті қолданылды. OCR нәтижелері артық таңбалардан тазартылып, құрылымды мәтін алу үшін пішімделді (2-кесте).

Кесте 2. OCR үрдісіне дейінгі және кейінгі файлдар

OCR -ге дейінгі файлдар	*OCR -ден кейінгі файлдар*
Бұл файлдар PDF немесе сурет (JPEG, PNG, TIFF) форматында болады.	*Құжаттарды сандық архивке енгізу жеңілдейді*
Файлдарда мәтінді таңдап көшіру, өңдеу, іздеу мүмкін болмайды.	*Құжаттың ішіндегі мәтінді көшіріп, өзгертіп, іздеуге болады.*
Файлда суреттер болғандықтан, файл көлемі көбіне үлкен болады.	*Файлдың көлемі кішірейіп, құрылымы дұрысталады, тек қажет ақпарат сақталып, артық визуалды элементтер жойылады.*
Көзі нашар көретін адамдарға немесе автоматтандырылған жүйелерге бұл файлдармен жұмыс істеу қиын.	*Мәтіндік файлдар экран оқитын бағдарламалармен және басқа ассистивті технологиялармен жақсы үйлеседі.*

OCR (оптикалық тану) технологиясы – қағаз негізіндегі құжаттар мен сканерленген бейнелерді құрылымдалған және іздеуге жарамды сандық деректерге түрлендірудің тиімді тәсілі болып табылады. Бұл технологияның қолданылуы құжаттармен жұмыс істеу үдерісін айтарлықтай оңтайландырып, ақпараттық ресурстардың қолжетімділігін арттырады. OCR технологиясын енгізу мәліметтерді сақтау, өңдеу және іздеу мүмкіндіктерін кеңейтіп, сандық архивтер мен ақпараттық жүйелердің функционалдық әлеуетін арттырады. Осыған байланысты ұйымдар мен жеке пайдаланушылар үшін деректерді тиімді басқару мақсатында

OCR технологиясын қолдану заманауи ақпараттық инфрақұрылымның маңызды құрамдас бөлігі болып табылады.

OCR дереккөздері:

- әл-Фараби атындағы ҚазҰУ кітапханасынан алынған оқулықтар, оқу құралдары монографиялар, диссертациялар;
- газет-журналдар.

Жалпы көлемі 198 ГБ болатын мәтіндік деректер OCR-сервисі арқылы өңделіп, құрылымдалған JSON форматына түрлендірілді.

OCR процесі аяқталғаннан кейін алынған мәтіндік файлдар құрылымдалған JSON форматындағы деректерге түрлендірілді. Әрбір JSON файлы келесі негізгі өрістерден тұрды (Сурет 1):

1. author – құжат авторы;
2. name – құжаттың атауы;
3. annotation – аннотация (қысқаша сипаттама);
4. text – құжаттың толық мәтіні.

1-суретте JSON (JavaScript Object Notation) форматында жазылған құрылымды мәтін берілген. Бұл құрылым белгілі бір кітап туралы мәліметті сипаттайды және төрт негізгі кілттен (параметрден) тұрады:


```
{
    "autor": "Досжан Д.",
    "name": "Жазмыштың формуласы",
    "annotation": "Кейіпкер бойындағы алып- жұлып бара жатқан құштарлық, ашыну
мен жеріну, ұрынып жүріп ырзық- несібені теріп жеу – осы мінездердің қазақ
кейіпкерінің бойындағы – сегіз қыры, бір сыры деп білгейсіз",
    "text": "Дүкенбай\nДОСЖАН\nКЕР ннен нні\nЖАЗМЫПЛТЫР\nФОРМУЛАС\nЖАЗМЫЙЛЫН
\nФОРМУЛАСЫ\nХикаяттар\nБАСПАСЫ\n64 Досжан Дүкенбай\nЖазмыштың формуласы.
Хикаяттар. Астана:  Фолиант ,\n536 бет.\nӨмір тауға шыққандай хал. Биік басына
көтерілген кезде:  Жеттім бе, жетпедім бе?  деп соңына қарасаң жалаңаяқ шалғынды
кешіп жүгірген балалық шақ көз алдыңа мөлдіреп келе қалады әлі алынбаған асулар
бас айналдырып, көңіл өрекпітеді, өліп-өшіп тырмысып биіктей бересің. Сіз бен
бізге ешқашан жан шақырып аялдау жазбаған, жазмыш формуласы солай.\nӘр кітаптың
мінезі болады. Сонғы жылдары жазған жеті хикаяттан құралған қолыңыздағы кітаптың
мінезі кемелдену мен ілгерілеу, кешегіні бүгін міске тұтпау, көбіне адалдықтан
опық жеу. Кейіпкер бойындағы алыпжұлып бара жатқан құштарлық, ашыну мен жеріну,
ұрынып жүріп ырзықнесібені теріп жеу осы мінездердің қазақ кейіпкерінің
бойындағы сегіз қыры, бір сыры деп білгейсіз.\nДосжан Дукенбай, 2008\n6
```


Сурет 1. JSON файлы

Аннотацияны бөліп алу жұмыстың маңызды кезеңдерінің бірі болды, себебі аннотация мәтіннің мазмұнын қысқаша сипаттап, оның тақырыптық бағытын көрсетеді. Аннотацияны анықтау үшін мәтіндегі арнайы үлгілерге негізделген эвристикалық ережелер қолданылды: мысалы, «Аннотация», «Қысқаша сипаттама», «Түйін» сияқты кілтсөздер мен олардың мәтіндегі орналасу реті ескерілді. Егер аннотация нақты берілмесе немесе ашық көрсетілмесе, мәтіннің алғашқы абзацтарына негізделген табиғи тілді өңдеу әдістері арқылы автоматты түрде аннотация жасалды.

Барлық элементтерді бөліп алғаннан кейін, деректер стандарттау кезеңінен өтті: артық таңбалар мен бос орындар жойылды, OCR қателері түзетілді, мәтін бірыңғай UTF-8 кодтау форматына келтірілді. Нәтижесінде алынған JSON файлдары құрылымдалған, машинамен оқуға жарамды дереккөз ретінде сақталып, әрі қарайғы талдау мен ақпараттық жүйелерге енгізуге дайын болды.

OCR-сервисін оңтайландыру шеңберінде екі түрлі өңдеу механизмі енгізілді:

1. Локалды режимде – ProcessPoolExecutor көмегімен көппроцессорлық өңдеу
2. Серверлік режимде – Celery және RabbitMQ негізінде асинхронды өңдеу.

Жүргізілген жүктемелік тестілеу нәтижелері өңдеу уақытының едәуір қысқарғанын көрсетті: локалды режимде 4-8 процессорлық ағын қолданылғанда 100 файлды өңдеу уақыты

шамамен 180 секундтан 40 секундқа дейін қысқарды. Ал серверлік режимде бұл көрсеткіш бір файл үшін орта есеппен 1,3 секундқа дейін азайды. Өңдеу тиімділігіне әсер ететін негізгі шектеуші факторлар ретінде дискілік енгізу/шығару (I/O), үлкен бейнелерді өңдеудегі CPU-ге түсетін жүктеме, және хабарламалар кезегіне (message queue) түсетін жүктеме анықталды. Аталған шектеулерді азайту мақсатында жүйеге буферлеу механизмі, параллель өңдеуді шектеу, воркерлерді автоматты масштабтау, сондай-ақ метрикаларды бақылау және мониторинг жүргізу құралдары енгізілді.

Зерттеу нәтижелері

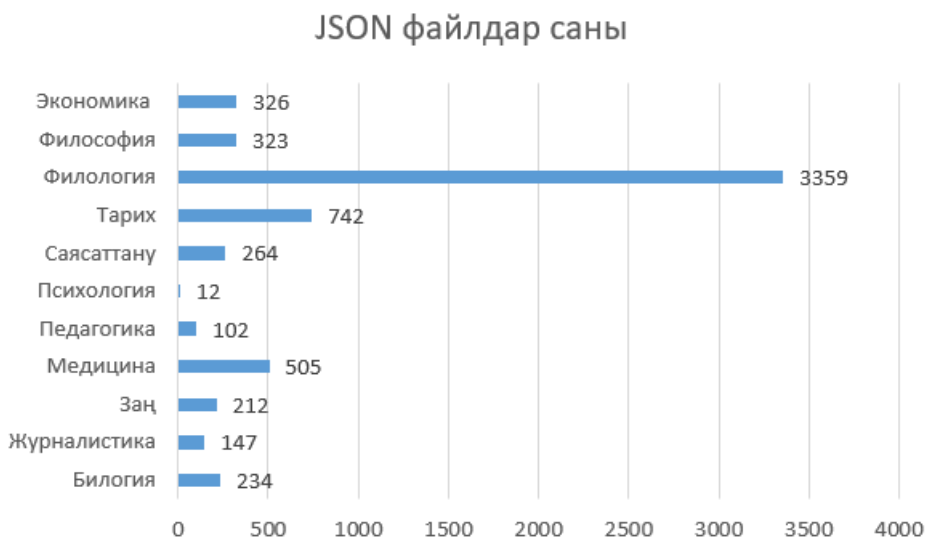
Қазақ тіліндегі үлкен көлемді мәтіндік деректерді жинау және өңдеу мақсатында арнайы OCR-сервис әзірленіп, FastAPI платформасында жүзеге асырылды. Аталған сервис PDF-файлдар мен сурет форматындағы деректерді қабылдап, олардың ішіндегі мәтінді Tesseract OCR кітапханасы арқылы таниды. Жоғары өнімділікке қол жеткізу үшін көппроцессорлық өңдеу әдісі қолданылды. OCR нәтижелері құрылымдалған JSON форматында сақталды. 3-кестеде көрсетілген нәтижесінде, жалпы саны 37 062 файл OCR-сервисі арқылы өңделіп, JSON форматына автоматты түрде түрлендірілді.

Кесте 3. Әртүрлі сала бойынша JSON файлдар саны

Сала атауы	*JSON файлдар саны*
Биология	*234*
Журналистика	*147*
Заң	*212*
Медицина	*505*
Педагогика	*102*
Психология	*12*
Саясаттану	*264*
Тарих	*742*
Филология	*3359*
Философия	*323*
Экономика	*326*
Егемен Қазақстан газеті	*29492*
Басқа сала	*1344*
Барлығы	*37062*

Егемен Қазақстан газетіндегі әрбір мақала JSON форматында жеке файл ретінде бөлінді. 2-суреттегі гистограммада сүйенсек, филология саласындағы JSON файлдар саны – 3359, бұл басқа барлық салалардан айтарлықтай жоғары. Бұл көрсеткіш филология саласында, әсіресе қазақ тілінде жазылған материалдардың (кітаптар, мақалалар, оқулықтар және т.б.) көптігін білдіреді.

Қазіргі ақпараттық технологиялар дәуірінде ұлттық мәдени және тілдік мұраны сақтау – қоғамның рухани қауіпсіздігін қамтамасыз етудің маңызды бағыттарының бірі болып табылады. Қазақ тіліндегі әдеби шығармалар, фольклорлық мұралар, тарихи құжаттар және тілтанымдық материалдар – ұлттың болмысын айқындайтын баға жетпес құндылықтар. Бұл мұраларды болашақ ұрпаққа жеткізудің ең тиімді жолы – құнды ақпараттарды сандық форматқа көшіру. Атап айтқанда, материалдарды JSON (JavaScript Object Notation) секілді құрылымдалған, өңдеуге және сақтауға ыңғайлы форматтарда цифрландыру қазіргі таңда кең таралып келеді. Мұндай тәсіл мұрағаттық құжаттарды, әдеби мәтіндерді және тіл ресурстарын оңай іздеуге, қайта өңдеуге, тіпті жасанды интеллект жүйелерінде қолдануға мүмкіндік береді.



Сурет 2. Әртүрлі сала бойынша JSON файлдар саны

Сонымен қатар, цифрланған материалдар отандық және халықаралық ғылыми айналымға енуге жол ашады, әрі заманауи білім беру мен зерттеу үдерістеріне тиімді ықпал етеді.

Соңғы жылдары жасанды интеллект саласында үлкен тілдік модельдер қарқынды дамып келеді. Бұл модельдер – мәтінді түсіну, генерациялау және талдау сияқты күрделі тілдік міндеттерді орындай алатын қуатты құралдар. Алайда мұндай модельдердің тиімділігі мен сапасы оларды оқыту үшін қолданылатын датасеттерге тікелей тәуелді. Бұл орайда мемлекеттік қазақ тіліндегі тілдік ресурстардың жетіспеушілігі айтарлықтай мәселе туғызады. OCR технологиясының көмегімен бұрын тек қағаз күйінде сақталған әдеби және ғылыми еңбектер, тарихи мұрағаттар мен оқулықтар цифрлы форматқа көшіріліп, олардан автоматты түрде JSON құрылымындағы мәтіндік датасеттер жасақталуда. Бұл қазақ тілінің табиғи мәтіндердегі қолданыс ерекшеліктерін ескере отырып, LLM модельдерін ана тілімізде оқытуға және дамытуға зор мүмкіндік береді.

Brown T. және басқалар ұсынған мақалада GPT-3 моделінің үлкен тілдік модельдер көмегімен машинаны табиғи тілде оқытуға болатынын дәлелдейді [9]. Бұл жұмыстың нәтижелері LLM саласында тұтас жаңа парадигманың бастамасы болды.

GPT-4 – OpenAI әзірлеген модель – көпмодальды, көптілді және кең контекстік мүмкіндіктерге ие заманауи үлкен тілдік модель. Ол мәтінмен қатар визуалды ақпаратты да өңдеуге қабілетті, логикалық пайымдау мен тапсырмаларды орындауда жоғары нәтижелер көрсетеді. GPT-4 моделі табиғи тіл өңдеу саласындағы күрделі мәселелерді шешуге арналған тиімді құрал болып табылады [10].

PaLM моделі 6144 TPU v4 чипін пайдалана отырып Pathways жүйесі негізінде үйретілді және көлемі аса үлкен, әртүрлі көздерден жинақталған көптілді деректер жиынтығында жаттықтырылды. Ауқымды және әртараптандырылған датасет үлгінің күрделі тілдік тапсырмаларда, соның ішінде көптілді пайымдау мен код генерациясында жоғары нәтижелер көрсетуіне ықпал етті [11].

Xue және басқалар ұсынған модель mT5 — T5 моделінің көптілді нұсқасы болып табылады және ол Common Crawl негізінде жасалған 101 тілді қамтитын деректер жиынтығында алдын ала үйретілген. Модель көптілді бенчмарктерде жоғары нәтижелер көрсетіп, нөлдік-shot жағдайында кездейсоқ аударманы болдырмау үшін арнайы әдіс ұсынылды. Зерттеуде қолданылған код пен модель салмақтары ашық түрде қолжетімді [12].

Үлкен тілдік модельдердің тиімділігі мен генеративті қабілеттері тікелей түрде оларды үйретуде қолданылатын деректер жиынтығының көлемі, сапасы және алуан түрлілігіне тәуелді. Ауқымды және әртүрлі тілдік корпус модельдің лексикалық қамтылуын арттырып

қана қоймай, семантикалық түсіну, көптілділік және контексті дұрыс болжау сияқты жоғары деңгейлі қабілеттерін дамытады. Сондықтан аз ресурсты тілдер үшін сапалы әрі теңгерімді деректер жинау – үлкен тілдік модельдердің әділ, бейтарап және көптілді қолдану мүмкіндігін қамтамасыз етудің негізгі алғышарты болып саналады.

OCR технологиясы қолданылып, қазақ тіліндегі кітаптар мен құжаттар сандық форматқа көшірілді. Бұл мәтіндер негізінде құрылымдалған датасеттер жасалып, олар LLM (мысалы, LLaMA 3.2 3B) моделін оқытуда пайдаланылды. Нәтижесінде, модель қазақ тіліндегі мәтіндерді жақсы түсініп, өндей алатын деңгейге жетті. 3-суретте LlaMA моделі қазақ халқының дүниетанымдық ерекшеліктерін, поэтикалық құрылымын және күйші ретінде жазылған тілдік стильдерді тануға мүмкіндік берді.

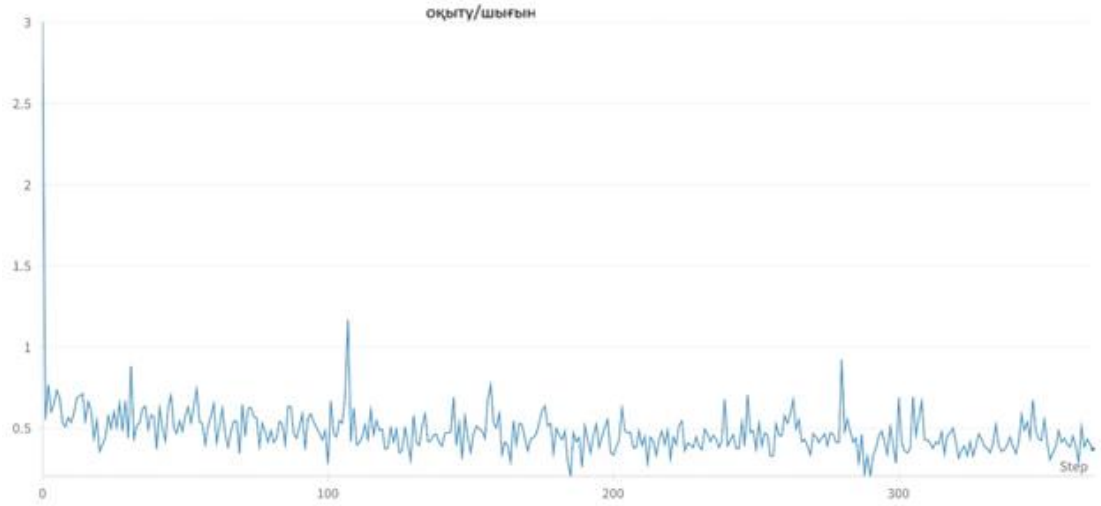
```
model_input = tokenizer(eval_prompt, return_tensors="pt").to("cuda")

model.eval()
with torch.inference_mode():
    print(tokenizer.decode(model.generate(**model_input, max_new_tokens=100)[0], skip_special_tokens=True))
```

```
Setting `pad_token_id` to `eos_token_id`:128001 for open-end generation.


күйші стил ретінде өмір туралы жазып беріңіз

---
Қазір де, ақырда да
Біздің мұның арасында
Ешқайда да жоқ емес
Өмірдің бұл мүшелінің
Біздің бұл мүшелінің
Өмір туралы жазып беріңіз
θ
```

Сурет 3. Модельлің берген жауабы

Бұл тәжірибе LLama 3.2 3B моделін пайдалану арқылы орындалған. Модель tokenizer арқылы мәтіндік сұранысты (prompt) өңдеп, generate әдісімен жаңа мәтін құраған. LLaMA 3.2 3B (Large Language Model Meta AI) – Meta компаниясы әзірлеген, 3.2 миллиард параметрден тұратын, қуатты және жеңілдетілген үлкен тілдік модель. Бұл модель LLaMA 3 сериясының шағын, бірақ тиімді нұсқасы болып саналады және төмен ресурсты құрылғыларда (мысалы, бір A10 немесе T4 GPU-да) қолдануға бейімделген [13].

4-суретте LLM моделін оқыту барысындағы **train/loss** (оқыту шығыны) метрикасының графигі көрсетілген. X осінде – оқыту қадамдары (step), Y осінде – шығын мәні (loss) бейнеленген.



Сурет 4. LLM моделін оқыту барысындағы train/loss (оқыту/шығын) метрикасының графигі

Алғашқы оқыту қадамдарында шығын мәні (loss) жоғары болғанымен (~2.8), бірнеше он қадам ішінде күрт төмендеп, тұрақты мәндерге жақындағаны байқалады. Оқыту процесі барысында loss шамамен 0.4–0.6 аралығында тербеліп отырды, бұл модельдің деректерге бейімделу деңгейінің жеткілікті екенін көрсетеді.

Дискуссия

Сирек кездесетін биік шыңдар — жеке қадамдардағы қиын үлгілерге немесе шағын батч өлшеміне байланысты ауытқулар. Дегенмен, жалпы тренд бойынша модельдің тұрақты үйренгені және артық үйрену (overfitting) белгілерінің байқалмағаны көрініп тұр. Бұл нәтижелер модельдің қазақ тіліндегі OCR арқылы алынған мәтіндермен және ұлттық стильдегі (мысалы, күйші тілі) тілдік құрылымдармен сәтті оқытылғанын дәлелдейді. Модель генерация барысында бейнелі, поэтикалық мәтіндер құра алатын деңгейге жетті, бұл оның тілдің терең құрылымдарын меңгеру қабілетін көрсетеді.

Қорытынды

Құрастырылған датасет LLM модельдерін оқыту үшін жоғары құндылыққа ие. Біріншіден, мәтіндер тақырыптар мен жанрлар тұрғысынан кең ауқымды қамтиды, бұл модельдің әртүрлі деректерге бейімделіп үйренуіне мүмкіндік береді. Екіншіден, метаақпараттың енгізілуі модельге мәтіннің контексін, авторын және жанрлық ерекшеліктерін жақсырақ түсінуге көмектеседі. Бұл әсіресе мәтін генерациясы, машиналық аударма және тақырыптық құрылымды талдау сияқты міндеттер үшін аса маңызды. JSON форматы өңдеуге ыңғайлы әрі қазіргі NLP фреймворктарына үйлесімді болғандықтан таңдалды.

Атқарылған жұмыстың нәтижелері жасалған тәсілдің масштабталатын және әмбебап екенін көрсетеді. Өңделген деректер тек LLM модельдерін оқытуда ғана емес, сонымен қатар арнайы тілдік ресурстарды – тақырыптық корпустарды, жиілік сөздіктерді және тіл талдауға арналған дерекқорларды жасау үшін де қолдануға болады. Бұл қазақ тіліндегі NLP бағыттарын дамытудың жаңа мүмкіндіктерін ашады, соның ішінде машиналық аударманы жақсарту, мәтіндерді автоматты түрде рефераттау және интеллектуалды іздеуді дамыту.

Алғыс

Бұл зерттеу жұмысы Қазақстан Республикасы Ғылым және жоғары білім министрлігі Ғылым комитетінің BR24993001 бағдарламалық-нысаналы қаржыландыруы аясында «Қазақ тілін және технологиялық прогресті қолдау үшін үлкен тілдік модельді (LLM) құру» жобасы бойынша қаржылық қолдауымен орындалды.

Пайдаланылған дереккөздердің тізімі

[1] Çetintaş İ., Müngen A. A. Using Pre-Processing Methods to Improve OCR Performances of Digital Historical Documents // Proceedings of the 2021 Innovations in Intelligent Systems and Applications Conference (ASYU). – 2021. – P. 1–5.

[2] Chen Y., Ströbel P. B. TrOCR Meets Language Models: An End-to-End Post-correction Approach // Proceedings of the IEEE International Conference on Document Analysis and Recognition. – 2024.

[3] Nguyen T., Jatowt A., Coustaty M., Doucet A. Survey of Post-OCR Processing Approaches // ACM Computing Surveys. – 2021. – Vol. 54. – P. 1–37.

[4] Yeshpanov R., Khassanov Y., Varol H. A. KazNERD: Kazakh Named Entity Recognition Dataset // Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022). – Marseille, France: European Language Resources Association, 2022. – P. 417–426.

[5] Mussakhojayeva S., Khassanov Y., Varol H. A. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus // Proceedings of the 23rd INTERSPEECH Conference. – 2022. – P. 1367–1371.

[6] Toiganbayeva N. A., Kasem M., Abdimanap G., Bostanbekov K., Abdelrahman A., Alimova A., Nurseitov D. KOHTD: Kazakh Offline Handwritten Text Dataset // Signal Processing: Image Communication. – 2022. – Vol. 108. https://www.sciencedirect.com/science/article/pii/S0923596522001217

[7] Yeshpanov R., Varol H. A. KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – Torino, Italy: ELRA, ICCL, 2024. – P. 9657–9667.

[8] Kadyrbek N., Tuimebayev Z., Mansurova M., Viegas V. The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh and Direct Preference Optimization // Big Data and Cognitive Computing. – 2025. – Vol. 9, № 5. – Art. 137. – DOI: 10.3390/bdcc9050137.

[9] Brown T., Mann B., Ryder N., et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems (NeurIPS 2020). – 2020. – Vol. 33. – P. 1877–1901.

[10] Achiam O. J., Adler S., Agarwal S., et al. GPT-4 Technical Report. – 2023.

[11] Chowdhery A., Narang S., Devlin J., et al. PaLM: Scaling Language Modeling with Pathways // arXiv. – 2022. – arXiv:2204.02311.

[12] Xue L., Constant N., Roberts A., et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer // Proceedings of the North American Chapter of the Association for Computational Linguistics. – 2020.

[13] Meta AI. LLaMA: Open Foundation and Instruction Models [Электронный ресурс]. – 2024. – URL: https://ai.meta.com/llama/ (дата обращения: 17.12.2025).

References

[1] Çetintaş İ., Müngen A. A. Using Pre-Processing Methods to Improve OCR Performances of Digital Historical Documents // Proceedings of the 2021 Innovations in Intelligent Systems and Applications Conference (ASYU). – 2021. – P. 1–5.

[2] Chen Y., Ströbel P. B. TrOCR Meets Language Models: An End-to-End Post-correction Approach // Proceedings of the IEEE International Conference on Document Analysis and Recognition. – 2024.

[3] Nguyen T., Jatowt A., Coustaty M., Doucet A. Survey of Post-OCR Processing Approaches // ACM Computing Surveys. – 2021. – Vol. 54. – P. 1–37.

[4] Yeshpanov R., Khassanov Y., Varol H. A. KazNERD: Kazakh Named Entity Recognition Dataset // Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC 2022). – Marseille, France: European Language Resources Association, 2022. – P. 417–426.

[5] Mussakhojayeva S., Khassanov Y., Varol H. A. KSC2: An Industrial-Scale Open-Source Kazakh Speech Corpus // Proceedings of the 23rd INTERSPEECH Conference. – 2022. – P. 1367–1371.

[6] Toiganbayeva N. A., Kasem M., Abdimanap G., Bostanbekov K., Abdelrahman A., Alimova A., Nurseitov D. KOHTD: Kazakh Offline Handwritten Text Dataset // Signal Processing: Image Communication. – 2022. – Vol. 108.
https://www.sciencedirect.com/science/article/pii/S0923596522001217

[7] Yeshpanov R., Varol H. A. KazSAnDRA: Kazakh Sentiment Analysis Dataset of Reviews and Attitudes // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – Torino, Italy: ELRA, ICCL, 2024. – P. 9657–9667.

[8] Kadyrbek N., Tuimebayev Z., Mansurova M., Viegas V. The Development of Small-Scale Language Models for Low-Resource Languages, with a Focus on Kazakh and Direct Preference Optimization // Big Data and Cognitive Computing. – 2025. – Vol. 9, № 5. – Art. 137. – DOI: 10.3390/bdcc9050137.

[9] Brown T., Mann B., Ryder N., et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems (NeurIPS 2020). – 2020. – Vol. 33. – P. 1877–1901.

[10] Achiam O. J., Adler S., Agarwal S., et al. GPT-4 Technical Report. – 2023.

[11] Chowdhery A., Narang S., Devlin J., et al. PaLM: Scaling Language Modeling with Pathways // arXiv. – 2022. – arXiv:2204.02311.

[12] Xue L., Constant N., Roberts A., et al. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer // Proceedings of the North American Chapter of the Association for Computational Linguistics. – 2020.

[13] Meta AI. LLaMA: Open Foundation and Instruction Models [Electronic resource]. – 2024. – URL: https://ai.meta.com/llama/ (accessed December 17, 2025).