әл-Фараби атындағы Қазақ ұлттық университеті

ӘОЖ 004.5

Қолжазба құқығында

САРСЕНБАЕВА АСИЯ КАЛИСЕИТОВНА

Қазақстан Республикасының заңнамасы саласындағы мемлекеттік тілге арналған семантикалық тәсілдер негізінде модельдерді зерттеу және интеллектуалды сұрақ-жауап жүйесін әзірлеу

8D06101-Ақпараттық жүйелер

Философия докторы (PhD) дәрежесін алу үшін дайындалған диссертация

Отандық ғылыми кеңесші:
Рахимова Диана Рамазановна,
PhD, қауымдастырылған профессор,
әл-Фараби атындағы Қазақ ұлттық университеті
Шетелдік ғылыми кеңесші:
Eśref Adali, профессор,
Стамбул Техникалық Университеті,
Түркия

Қазақстан Республикасы
Алматы, 2026

МАЗМҰНЫ

НОРМАТИВТІ СІЛТЕМЕЛЕР 4
БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР 5
КІРІСПЕ 6
1 ЗАҢНАМА САЛАСЫНДАҒЫ СҰРАҚ-ЖАУАП ЖҮЙЕЛЕРІН ЗЕРТТЕУДІ ТАЛДАУ 11
1.1. Сұрақ-жауап жүйелерінің даму эволюциясы және негізгі архитектуралары 11
1.2. Жасанды интеллект және LLM модельдерінің заңнама саласындағы қолданылуы 14
1.3. Заңнамалық мәтіндердің және қазақ тілінің лингвистикалық ерекшеліктері 18
1.4. Қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің қазіргі жағдайы 19
1.5. Отандық және шетелдік ғылыми зерттеулерді салыстырмалы талдау 22
1.6 1 бөлім бойынша қорытынды 26
2. ЗАҢНАМАЛЫҚ ДЕРЕКТЕР ҚОРЫН ҚАЛЫПТАСТЫРУ ЖӘНЕ LLM МОДЕЛЬДЕРІН ЗЕРТТЕУ 28
2.1 Қазақстан Республикасының құқықтық деректер кеңістігін талдау 28
2.1.1 Құқықтық ақпараттық ресурстар және цифрлық инфрақұрылым 28
2.1.2 Заңнамалық мәтіндер корпусын және сұрақ-жауап деректер жиынтығын қалыптастыру 31
2.1.3 Мәтіндерді жинау, алдын ала өңдеу және аннотациялау әдістері 32
2.2 Үлкен тілдік модельдерді зерттеу және бейімдеу әдістері 37
2.2.1 Базалық LLM модельдерінің салыстырмалы талдауы 37
2.2.2 Fine-tuning негізінде модельдерді домендік бейімдеу 41
2.3 RAG архитектурасын заңнамалық сұрақ-жауап жүйесіне интеграциялау 57
2.3.1 RAG жүйесіндегі іздеу модулінің құрылымы 58
2.3.2 Генеративті жауап қалыптастыру моделі 61
2.3.3 Интеллектуалды заңнамалық сұрақ-жауап жүйесінің RAG архитектурасы 63
3 ЗАҢНАМАЛЫҚ СҰРАҚ-ЖАУАП ЖҮЙЕСІ ҮШІН СЕМАНТИКАЛЫҚ ПРЕПРОЦЕССИНГ (SEMANTICPREP) ТӘСІЛІН ӘЗІРЛЕУ 66
3.1 SemanticPrep тәсілінің концептуалдық негіздері және архитектуралық моделі 67
3.1.1 Заңнамалық іздеудегі алшақтық мәселелері 67
3.1.2 SVD/LSA тәсілінің ғылыми негіздемесі 69
3.1.3 SemanticPrep тәсілінің архитектурасы 70
3.2 SemanticPrep тәсілінің бірінші кезеңі: сұрақтарды заңнамалық түрлендіру 72
3.2.1 mT5 моделін fine-tuning арқылы бейімдеу 72
3.2.2 Табиғи тілдегі сұрақтарды құқықтық құрылымға түрлендіру моделі 74
3.3 SemanticPrep тәсілінің екінші кезеңі: семантикалық кеңейту моделі 79
3.3.1 SVD/LSA негізіндегі семантикалық талдау алгоритмі 80
3.3.2 Семантикалық кеңейту әдісінің математикалық моделі 81
3.4 SemanticPrep тәсілін басқа LLM модельдермен салыстырмалы талдау 89

4 ҰСЫНЫЛҒАН ТӘСІЛДІ ЭКСПЕРИМЕНТТІК БАҒАЛАУ ЖӘНЕ ЗАҢНАМАЛЫҚ СҰРАҚ-ЖАУАП ЖҮЙЕСІНІҢ НӘТИЖЕЛЕРІН ТАЛДАУ 93
4.1 Эксперименттік деректер сипаттамасы және бағалау метрикалары 93
4.2 Базалық LLM модельдерінің эксперименттік салыстырмалы талдауы 101
4.2.1 Базалық LLM модельдерінің zero-shot режиміндегі бағалануы 101
4.2.2 LLM модельдерін fine-tuning арқылы бейімдеу нәтижелерін бағалау 103
4.3 Finetuning LLM және RAG интеграциясы негізіндегі жүйенің нәтижелерін бағалау және талдау 111
4.4 SemanticPrep негізіндегі ұсынылған тәсілдің тиімділігін бағалау 118
4.5 Әзірленген заңнамалық сұрақ-жауап жүйесінің архитектурасы 123
4.6 Әзірленген жүйенің практикалық апробациясы және қолданбалы нәтижелері 125
ҚОРЫТЫНДЫ 127
ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ 130
ҚОСЫМША А Енгізу актісі 138

НОРМАТИВТІ СІЛТЕМЕЛЕР

Берілген диссертацияда келесі стандарттарға сәйкес сілтемелер қолданылды: Қазақстан Республикасы МЖМБС 5.04.034-2011. «Қазақстан Республикасы

Мемлекеттік жалпыға білім беру стандарты. Жоғары оқу орнынан кейінгі білім. Докторантура». Негізгі ережелер ҚР БҒМ бекітілген. 17.06.2011 ж. №261. Астана, 2011 ж.

«Диссертацияны безендіру нұсқаулығы», Қазақстан Республикасы БҒМ ЖАК 28 қыркүйек, 2004 жыл. №377-3 ж.

ГОСТ 7.32-2001. Ғылыми зерттеу жұмысының есебі. Безендіру ережелері мен құрылымы.

ГОСТ 7.1-2003. Библиографиялық таспа. Библиографиялық сипаттау.

БЕЛГІЛЕУЛЕР МЕН ҚЫСҚАРТУЛАР

NLP – Natural Language Processing (Табиғи тілдерді өңдеу)

ML – Machine Learning (Машиналық оқыту)

NN – Neural Networks (Нейрондық желілер)

LLM – Large Language Model (Үлкен тілдік модель)

BLEU – Bilingual Evaluation Understudy (Мәтін апасын бағалау метрикасы)

AI- Artificial Intelligence (Жасанды интеллект)

QA - Question Answering (Сұрақ-жауап)

QAS - Question Answering System (Сұрақ-жауап жүйесі)

ROUGE - Recall-Oriented Understudy for Gisting Evaluation

METEOR - Metric for Evaluation of Translation with Explicit Ordering (Аударма сапасын анық реттілікпен бағалау метрикасы)

MRR - Mean Reciprocal Rank (Іздеу нәтижелерін бағалау метрикасы)

GPT - Generative Pre-Trained Transformer (Генеративті алдын ала оқытылған трансформер)

GPT-4o mini - Optimized version of GPT-4 for inference (Қорытынды шығару үшін оңтайландырылған GPT-4 нұсқасы)

Llama - Large Language Model Meta AI (Meta AI компаниясының үлкен тілдік үлгісі)

GEMINI - Google's Multimodal Large Language Model (Google компаниясының мультимодальды үлкен тілдік үлгісі)

KazLLM - Kazakh Large Language Model (Қазақ тіліндегі үлкен тілдік үлгі)

RAG - Retrieval-Augmented Generation (Іздеумен күшейтілген генерация)

IR - Information Retrieval (Ақпаратты іздеу)

mT5 - Multilingual Text-to-Text Transfer Transformer (Көптілді мәтіннен мәтінге трансформер)

SVD - Singular Value Decomposition (Сингулярлы мәнге жіктеу)

LSA - Latent Semantic Analysis (Латентті семантикалық талдау)

КІРІСПЕ

Зерттеу тақырыбының өзектілігі. Қазақстан Республикасының заңнамалық ақпараттарына мемлекеттік тілде қолжетімділікті қамтамасыз ету қажеттілігімен айқындалады. Қазіргі таңда азаматтардың заңнамалық ақпаратты алу процесі күрделі, көп уақытты талап етеді және көбінесе құқықтық мәтіндердің терминологиялық күрделілігіне байланысты түсінуді қиындатады. Осыған байланысты мағынасы бойынша талдап, нақты және дәл құқықтық жауаптарды автоматты түрде ұсынатын интеллектуалды сұрақ-жауап жүйелерін әзірлеу өзекті ғылыми мәселе болып табылады.

Жасанды интеллект технологияларының, әсіресе генеративті тілдік модельдердің (LLM) қарқынды дамуы заңнамалық мәтіндерді өңдеу мен талдаудың жаңа мүмкіндіктерін ашуда. GPT, LLaMA, Gemini сияқты модельдер мәтіндерді түсіну және генерациялау міндеттерінде жоғары нәтижелер көрсеткенімен, оларды қазақ тілі сияқты агглютинативті және ресурсы шектеулі тілдерге бейімдеу жеткілікті деңгейде зерттелмеген [1]. Бұл өз кезегінде морфологиялық күрделілік, сөз түрленуінің көптүрлілігі және заңнамалық терминологияның ерекшеліктерін ескере отырып, арнайы әдістерді әзірлеуді талап етеді [2].

Қазіргі қолданыстағы жүйелер (Adilet, eGov, LawVision) негізінен кілт сөздерге негізделген іздеу немесе ережелік тәсілдерді қолданады және пайдаланушы сұрағын семантикалық тұрғыдан толық түсінбейді [3]. Сонымен қатар, заманауи LLM модельдерінде нақты заңнамалық дереккөздермен байланыс болмауына байланысты hallucination мәселесі орын алады, бұл құқықтық жауаптардың дәлдігіне кері әсер етеді [4].

Осыған байланысты, қазақ тілінің морфологиялық ерекшеліктерін ескеретін, семантикалық талдау әдістерін, Retrieval-Augmented Generation (RAG) тәсілін және ірі тілдік модельдерді біріктіретін интеллектуалды сұрақ-жауап жүйесін әзірлеу қажеттілігі туындайды. Әсіресе табиғи тілдегі сұрақтарды заңнамалық терминологияға сәйкестендіретін семантикалық препроцессинг әдістерін енгізу ғылыми тұрғыдан өзекті болып табылады.

Осы зерттеу жұмысы қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің дәлдігін арттыруға, құқықтық ақпараттың қолжетімділігін кеңейтуге және ұлттық деңгейдегі интеллектуалды жүйелерді дамытуға бағытталған.

Диссертациялық жұмыстың мақсаты – қазақ тілінің морфологиялық ерекшеліктерін ескеретін семантикалық талдау әдістері негізінде Қазақстан Республикасының заңнамасы саласына арналған интеллектуалды сұрақ-жауап жүйесін зерттеу және әзірлеу.

Зерттеудің міндеттері. Қойылған мақсатқа қол жеткізу үшін келесі міндеттерді орындау қарастырылады:

1) Қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің қазіргі жағдайын талдап, ғылыми зерттеу мәселелерін анықтау.

2) Ресми порталдардан (adilet.kz, zan.gov.kz) құжаттарды жинап, мамандандырылған заңнамалық деректер қорын құрастыру.

3) Сұрақтың мағыналық байланысын анықтап, оны заңнамалық формаға келтіретін екі кезеңді семантикалық препроцессинг әдісін ұсыну.

4) Ірі тілдік модельдерді (LLM) заңнамалық корпуста алдын ала оқыту (fine-tuning) арқылы олардың қазақ тіліндегі құқықтық терминологияны түсіну деңгейін арттыру

5) Интеллектуалды заңнамалық сұрақ-жауап жүйесінің пайплайнын әзірлеу

6) Ұсынылған жүйені baseline модельдермен салыстырып, эксперименттік түрде тиімділігін дәлелдеу

Зерттеу нысаны. Қазақстан Республикасының заңнамалық базасына мемлекеттік тілде қолжетімділікті қамтамасыз ету және құқықтық ақпаратты семантикалық талдау негізінде іздеу мен өңдеуге арналған интеллектуалды сұрақ-жауап жүйелері..

Зерттеу пәні. Қазақ тілінің агглютинативті құрылымын ескеретін семантикалық талдау әдістері, құқықтық терминологияны түрлендіру тәсілдері, заңнамалық мәтіндерді интеллектуалды өңдеу модельдері және табиғи тілдегі сұраныстарды құқықтық контексте интерпретациялау әдістері.

Зерттеу әдісі. Табиғи тілдерді өңдеу (NLP), Transformer архитектурасына негізделген ірі тілдік модельдер (LLM), Retrieval-Augmented Generation (RAG) технологиясы, бағалау метрикалары (BERTScore, ROUGE, METEOR, MRR, Recall@5) және эксперттік бағалау әдістемелері.

Алынған нәтижелердің ғылыми жаңалығы:

1) Құқықтық мәтіндердің жоғары күрделілігі мен терминологиялық нақтылығын ескере отырып, табиғи тілдегі сұрақтарды заңнамалық терминологиямен сәйкестендіруге арналған екі кезеңді семантикалық препроцессинг тәсілі әзірленді.

2) Әзірленген тәсілдер негізінде семантикалық препроцессинг, бейімделген LLM және RAG технологияларын біріктіретін интеллектуалды сұрақ-жауап жүйесінің пайплайны құрылды. Оның тиімділігі эксперименттік деректерде сапалық және сандық метрикалар арқылы расталды.

Жұмыстың теориялық маңыздылығы. Жұмыстың теориялық маңыздылығы қазақ тілінің агглютинативті табиғатын ескеретін семантикалық талдау тәсілдерін құқықтық доменде қолданудың теориялық негіздерін кеңейтумен сипатталады. Диссертациялық зерттеу нәтижелері заңнамалық мәтіндерді интеллектуалды өңдеу, табиғи тілдегі сұраныстарды құқықтық контексте интерпретациялау және low-resource (табиғи тілдік ресурстары жеткіліксіз) тілдер үшін интеллектуалды сұрақ-жауап жүйелерін құру салаларындағы ғылыми-әдіснамалық негіздерді толықтырады.

Жұмыстың практикалық маңыздылығы. Жұмыстың практикалық маңыздылығы зерттеу нәтижелерін Қазақстан Республикасының заңнамасы бойынша мемлекеттік тілдегі интеллектуалды құқықтық сервистерді, электрондық үкімет жүйелерін және құқықтық ақпараттық платформаларды әзірлеуде қолдану мүмкіндігімен сипатталады. Ұсынылған тәсілдер құқықтық ақпаратқа қолжетімділікті арттыруға, пайдаланушы сұраныстарын

семантикалық өңдеу сапасын жақсартуға және заңнамалық сұрақтарға автоматтандырылған жауап қалыптастыру тиімділігін жоғарылатуға мүмкіндік береді.

Қорғауға шығарылатын негізгі тұжырым. қазақ тілінің агглютинативті және морфологиялық ерекшеліктерін ескеретін семантикалық тәсілдер негізінде Қазақстан Республикасының заңнамасы саласына арналған интеллектуалды сұрақ-жауап жүйесінің моделі әзірленді. Зерттеу барысында құқықтық доменге арналған 8 000-нан астам сұрақ-жауап жұбы мен 5.3 ГБ көлеміндегі заңнамалық құжаттардан тұратын мамандандырылған деректер жиыны қалыптастырылды. Табиғи тілдегі пайдаланушы сұраныстарын құқықтық терминологиямен сәйкестендіруге, заңнамалық контексті интерпретациялауға және ақпараттық іздеу сапасын арттыруға бағытталған екі кезеңді семантикалық препроцессинг әдісі ұсынылды. Сонымен қатар семантикалық талдау, ақпараттық іздеу және генеративті модельдерді біріктіретін интеллектуалды заңнамалық сұрақ-жауап жүйесінің архитектурасы әзірленді. Жүргізілген эксперименттік зерттеулер нәтижесінде ұсынылған тәсілдің baseline модельдермен салыстырғанда ROUGE-L көрсеткішін 15–20%-ға, BERTScore мәнін 3–5%-ға, Recall@5 көрсеткішін 11–15%-ға, MRR метрикасын 16–18%-ға және сараптамалық бағалау нәтижелерін 30–40%-ға дейін жақсартуға мүмкіндік беретіні анықталды.

Сенімділік дәрежесі мен апробациялау нәтижелері. Диссертациялық жұмыстың ғылыми нәтижелері теориялық талдау, эксперименттік зерттеулер және салыстырмалы бағалау арқылы негізделді. Ұсынылған әдістер мен модельдердің тиімділігі халықаралық деңгейде кең қолданылатын ROUGE-L, BERTScore, Recall@5, MRR және сараптамалық бағалау метрикалары негізінде тексерілді. Эксперименттік нәтижелер ұсынылған тәсілдің baseline модельдермен салыстырғанда заңнамалық сұрақ-жауап міндеттерін орындауда жоғары тиімділік көрсететінін дәлелдеді.

Диссертациялық зерттеудің негізгі нәтижелері халықаралық және республикалық ғылыми конференцияларда, семинарларда және ғылыми басылымдарда баяндалып, ғылыми қауымдастық тарапынан талқыланып апробациядан өтті. Зерттеу нәтижелері бойынша Web of Science және Scopus дерекқорларында индекстелетін халықаралық басылымдарда, сондай-ақ Қазақстан Республикасы Ғылым және жоғары білім министрлігінің ұсынған ғылыми журналдарында мақалалар жарияланды.

Зерттеушінің жеке үлесі. Докторант диссертациялық зерттеудің барлық кезеңдерін өз бетінше орындады және қойылған ғылыми міндеттердің шешімін жүзеге асырды. Зерттеу барысында қазақ тіліндегі заңнамалық сұрақ-жауап міндеттеріне арналған 8 000-нан астам сұрақ-жауап жұбы мен 5.3 ГБ көлеміндегі заңнамалық құжаттардан тұратын мамандандырылған деректер жиынын жинақтау, алдын ала өңдеу және құрылымдау жұмыстары орындалды. Заңнамалық доменге бейімделген үлкен тілдік модельдерді (LLM) fine-tuning негізінде оқыту және салыстырмалы эксперименттік зерттеу жүргізілді. Табиғи тілдегі пайдаланушы сұраныстарын құқықтық терминологиямен

сәйкестендіруге арналған екі кезеңді семантикалық препроцессинг әдісі әзірленді. Сонымен қатар семантикалық талдау, ақпараттық іздеу және генеративті модельдерді біріктіретін интеллектуалды заңнамалық сұрақ-жауап жүйесінің бағдарламалық архитектурасы мен функционалдық пайплайны құрастырылды. Ұсынылған модельдер мен әдістердің тиімділігі baseline тәсілдермен салыстырмалы түрде эксперименттік бағалау арқылы дәлелденді.

Диссертация тақырыбының ғылыми-зерттеу жұмыстарының жоспарларымен байланысы. Диссертация Қазақстан Республикасы жоғары білім және ғылым министрлігінің (қаржыландыру көзі: Қазақстан Республикасы Білім және ғылым министрлігінің Ғылым комитеті) ғылыми-зерттеу гранттық қаржыландыру AP 19677835 және BR24993001 жобасы аясында орындалды.

Басым бағыт: Ақпараттық технологиялар және жасанды интеллект.

Мамандандырылған бағыт: Табиғи тілді өңдеу және интеллектуалды сұрақ-жауап жүйелері.

Нәтиженің жарияланымдары. Диссертация тақырыбы бойынша 9 ғылыми жұмыс жарияланды, оның ішінде Scopus және Web of Science базаларына индексделетін импакт-факторы бар басылымдарда 4 мақала жарияланды, Қазақстан Республикасы Білім және ғылым министрлігі Білім және ғылым саласындағы бақылау комитеті ұсынатын журналдарда 3 мақала жарияланды:

1) Rakhimova D., Karyukin V., Turarbek A., **Sarsenbayeva A.K**., Alieyev R. (2025). Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation. Computers (MDPI). 14(9), 354. https://doi.org/10.3390/computers14090354 WoS (SCIE), Q1, 2025; 5-Year IF: 2.7; SJR: 0.492; Percentile: 84 (Computer Science, Applications)
2) D.Rakhimova, **A.Sarsenbayeva**, E. Adali, A. Kulzhanova, Classification of questions in the intellectual question-answer system for the state language, Вестник НИА РК №4 Алматы, 2024, – С. 166-179 (КОКСНВО МНВО РК) https://doi.org/10.47533/2024.1606-146X.014
3) Рахимова Д.Р., **Сарсенбаева А.К.,** Турарбек А., Ауезова А.С. Применение глубокого обучения для повышения точности ответов в мультиязычных вопросно-ответных системах. Вестник НАН РК, физико-математическая серия, № 2, Алматы, 2025, – С. 196-211 (КОКСНВО МНВО РК) https://doi.org/10.32014/2025.2518-1726.353
4) **Сарсенбаева А.К.,** Рахимова Д.Р., Шормакова А.Н., Мансурова М. Агглютинативті тілдердің заңды мәтіндеріндегі мағыналық ұқсастықты анықтаудың контекстке негізделген тәсілі: қазақ тіліндегі жағдайды зерттеу. Қазақстан Республикасы Ұлттық ғылым академиясының Хабаршысы, - физика математикалық сериясы, № 1, Алматы 2026 – С. 354-372 https://doi.org/10.32014/2026.2518-1726.417
5) Rakhimova D., Karyukin V., Karibayeva A., **Sarsenbayeva A.K**. Developing advanced question-answering models for Legal Kazakh texts: A comparative study of modern approaches. ACIIDS 2025: 17th Asian Conference on

Intelligent Information and Database Systems. DOI: 10.1007/978-981-96-5881-7_19 (Scopus, процентиль 39)

6) Rakhimova, D., Karyukin, V., Amirova, D., **Sarsenbayeva**, A. Collection and Preprocessing of Data for LLM in the Kazakh Language in the Field of Legislation. Communications in Computer and Information Science, 2024, 2211 CCIS, P.129–139. Springer. DOI:10.1007/978-3-031-72260-8_11 (Scopus, процентиль 39)
7) **Sarsenbayeva A.K**., Rakhimova D.R., Turarbek A. Multilingual Large Language Models in the Legal Domain: An Effectiveness Analysis in Kazakh, Turkish, and English. 2025 10th International Conference on Computer Science and Engineering (UBMK), Istanbul, Turkiye, 2025, pp. 410-415, doi: 10.1109/UBMK67458.2025.11207002. (Scopus, процентиль 39)
8) **Сарсенбаева А.К.,** Рахимова Д.Р. Қазақ, түрік және ағылшын тілдеріндегі заңнама сұрақтарына жауап беру сапасын бағалау: көптілді LLM үлгілерінің тәжірибелік зерттеуі. X Международная научно-практическая конференция «Информатика и прикладная математика», Алматы, 2025. – С. 293-302
9) Рахимова Д., **Сарсенбаева А**. Заңнама саласына байланысты сұрақ-жауап жүйелеріндегі заманауи жағдайды зерттеу Информатика и прикладная математика: Мат. IX Межд. науч.-практ. конф. (31.10 - 01.11.2024 г.). – Алматы, 2024, – С. 76-83.

Диссертация құрылымы және көлемі. Диссертациялық жұмыс кіріспеден, 4 бөлімнен, қорытынды, әдебиеттер тізімі және қосымшадан тұрады. Диссертацияның толық көлемі: 141 бет машиналық мәтіні, оның ішіне 40 сурет, 41 кесте, 108 пайдаланылған әдебиеттер тізімі және қосымша кіреді.

1 ЗАҢНАМА САЛАСЫНДАҒЫ СҰРАҚ-ЖАУАП ЖҮЙЕЛЕРІН ЗЕРТТЕУДІ ТАЛДАУ

Заңнамалық сұрақ-жауап жүйелерін зерттеудің өзектілігі қазіргі цифрлық мемлекеттік қызмет, құқықтық хабардарлық және мемлекеттік тілді дамыту бағыттарымен тығыз байланысты. Бұл салада негізгі қиындықтар: 1) қарапайым тілдегі азаматтық сұрақтар мен заңнамалық мәтіндер арасындағы терминологиялық және семантикалық алшақтық; 2) қазақ тілінің агглютинативті морфологиясынан туындайтын өңдеу күрделілігі; 3) қазақ тіліндегі заңнамалық NLP үшін мамандандырылған деректер жиынының жетіспеушілігі.

Бұл тарауда заңнамалық сұрақ-жауап жүйелерінің даму тарихы мен негізгі архитектуралары, заманауи LLM модельдерінің ерекшеліктері, қазақ тілінің лингвистикалық ерекшеліктері, заңнамалық мәтіндерді өңдеудің мәселелері, сондай-ақ отандық және шетелдік зерттеулер жүйелі түрде талданады.

1.1. Сұрақ-жауап жүйелерінің даму эволюциясы және негізгі архитектуралары

Сұрақ-жауап жүйелері - пайдаланушының табиғи тілде жазған сұрағына автоматты түрде дәл жауап беруге арналған интеллектуалды жүйелер. Олардың даму тарихы алты онжылдықты қамтиды: қарапайым ережелерге негізделген шешімдерден бастап, бүгінгі заманауи нейрожелілік архитектураларға дейін жетті. QAS дамуын бес негізгі кезеңге бөлуге болады. Әр кезең өзінің технологиялық серпілісімен және шектеулерімен ерекшеленеді.

Бірінші кезең - ережелерге негізделген жүйелер (1960-1980-жылдар). Алғашқы QAS қарапайым шаблондар мен ережелер жиынтығына сүйенді. 1961 жылы BASEBALL жүйесі бейсбол статистикасы туралы тар шеңберлі сұрақтарға жауап бере алды [5]. 1966 жылы Джозеф Вейценбаумның ELIZA бағдарламасы адаммен диалогты еліктеу қабілетімен таң қалдырды, алайда нақты мағынаны түсінбейтін [6]. 1971 жылы LUNAR жүйесі Аполлон миссиясының геологиялық деректері бойынша сұрақтарға жауап беріп, логикалық-семантикалық тәсілдің алғашқы үлгісін көрсетті. Бұл кезеңнің басты шектеуі - тар домен: жүйелер тек алдын ала анықталған тақырыптармен жұмыс істей алды.

Екінші кезең - статистикалық тәсілдер (1990-жылдар). 1990-шы жылдары TREC QA Track байқауы QAS зерттеулеріне серпін берді. Статистикалық модельдер, ақпаратты іздеу (Information Retrieval) алгоритмдері және TF-IDF салмақтау тәсілдері кеңінен қолданыла бастады [7]. Бұл кезеңде жүйелер кеңірек тақырыптармен жұмыс істей алатын болды, бірақ семантикалық байланыстарды түсіне алмады.

Үшінші кезең - машиналық оқыту (2000–2015-жылдар). 2011 жылы IBM Watson жүйесі Jeopardy! телешоуында адам чемпиондарын жеңіп, QAS тарихында бетбұрыс сәт болды. Watson миллиондаған құжатты өңдеп, мультидеңгейлі талдау жасай алды. SQuAD (Stanford Question Answering Dataset, 2016) бенчмаркі зерттеушілерге стандартты бағалау платформасын берді [8].

Осы кезеңде Support Vector Machines (SVM), Random Forest және рекуррентті нейрондық желілер (RNN) белсенді қолданылды.

Төртінші кезең - Transformer архитектурасы (2017–2020-жылдар). 2017 жылы Google-дің «Attention Is All You Need» мақаласы Transformer архитектурасын ұсынды - бұл NLP тарихындағы революциялық жаңалық болды [9]. BERT (Google, 2018) екі жақты кодтауды қолданып, контекстті терең түсіне алды [10]. GPT сериясы (OpenAI, 2018–2020) генеративті мүмкіндіктерді ашты [11]. T5 (Google, 2020) барлық NLP тапсырмаларын мәтін-мәтін форматына келтірді. XLNet, RoBERTa, ALBERT модельдері BERT-ті жетілдіріп, бенчмарктерде жаңа рекордтар орнатты [12-15].

Бесінші кезең - ірі тілдік модельдер дәуірі (2020-жылдан бүгінге дейін). GPT-3 (2020, 175 млрд параметр), GPT-4 (2023), Gemini (Google, 2023), Claude (Anthropic, 2023) модельдері нөлдік (zero-shot) және аз нысанды (few-shot) оқытуда таңғажайып нәтиже берді. RAG архитектурасы іздеу мен генерацияны біріктіріп, галлюцинация мәселесін азайтты [16]. Қазіргі таңда LLM және RAG интеграциясы QAS-нің де-факто стандартына айналды [17]. Кесте 1.1-де сұрақ-жауап жүйелерінің тарихи даму кезеңдері, қолданылған негізгі технологиялар, өкілдік модельдер мен олардың шектеулері көрсетілген. Кестеде ережелік жүйелерден бастап қазіргі LLM және RAG архитектураларына дейінгі эволюция сипатталады.

Кесте 1.1 - Сұрақ-жауап жүйелерінің даму кезеңдері

Кезең	**Жылдар**	**Негізгі технология**	**Өкілді жүйелер**	**Шектеулер**
I	1960–1980	Ережелер, шаблондар	BASEBALL, ELIZA, LUNAR	Тар домен, морфология жоқ
II	1990–2000	TF-IDF, статистика	TREC QA системалары	Семантика түсінілмейді
III	2000–2017	ML, RNN, SVM	IBM Watson, SQuAD	Контекст шектеулі
IV	2017–2020	Transformer, BERT	BERT, T5, GPT-2	Домендік баптау қажет
V	2020–қазір	LLM, RAG	GPT-4, Gemini, Claude	Галлюцинация мәселесі

Қазіргі QAS архитектурасы жүйенің барлық компоненттері - пайдаланушы интерфейсі, тілдік өңдеу модульдері, іздеу механизмі және білім базасы - арасындағы өзара байланысты анықтайды. Тарихи даму барысында алты негізгі архитектура қалыптасты.

Монолитті (ережелерге негізделген) архитектура - бірінші буын жүйелердің негізі. Пайдаланушы сұрағы алдын ала жазылған шаблондармен салыстырылып, сәйкес жауап қайтарылады. Артықшылығы: жылдам, болжамды. Кемшілігі:

икемсіз, жаңа сұрақтарға бейімделе алмайды. Қазіргі таңда тек өте шектеулі сценарийлерде (FAQ боттары) қолданылады.

Классикалық екі деңгейлі архитектура «пайдаланушы - білім базасы» моделіне негізделген. Сұрақ өңдеу модулі мәтінді структурланған форматқа түрлендіреді де, деректер базасынан сәйкес жауапты іздейді. Логикалық-семантикалық іздеуге мүмкіндік берді, бірақ контекстуалды түсінік шектеулі болып қалды.

Үш деңгейлі архитектурада пайдаланушы мен білім базасы арасына NLP модулі бар қосымша қабат (application server) енгізілді. Бұл қабат синтаксистік талдау, семантикалық интерпретация және ниетті анықтауды (intent detection) жүзеге асырады. Заманауи диалогтік жүйелердің (Siri, Google Assistant алғашқы нұсқалары) негізі болды.

Таратылған және гибридті архитектуралар бірнеше деректер көзімен - ішкі базалар, веб-ресурстар, онтологиялар, API - бір мезгілде жұмыс істейді. Интеграция модулі нәтижелерді біріктіріп, ең дәл жауапты қалыптастырады. Бұл тәсіл ауқымды корпоративтік жүйелерде кеңінен қолданылады.

Нейрожелілік (LLM негізіндегі) архитектура - қазіргі QAS-тің негізгі бағыты. Transformer архитектурасына сүйенген ірі тілдік модельдер контекстті, мағынаны және тілдік нюанстарды терең түсінеді. Жауапты тек іздеп қана қоймай, өздігінен генерациялай алады. Алайда галлюцинация, жоқ ақпаратты ойлап шығару, бұл архитектураның басты кемшілігі.

RAG архитектурасы нейрожелілік генерацияның икемділігі мен нақты дереккөзден іздеудің сенімділігін біріктіреді. Сұрақ алдымен векторлық іздеу арқылы релевантты үзінділерді табады, содан кейін LLM осы үзінділер негізінде жауап генерациялайды. Бұл тәсіл галлюцинацияны айтарлықтай азайтады және жауаптың дереккөзін көрсетуге мүмкіндік береді. Заңнамалық QAS үшін RAG архитектурасы ең тиімді шешім болып табылады.

Көпагентті архитектурада бірнеше мамандандырылған агент өзара әрекеттесіп, күрделі сұрақтарды кезең-кезеңімен шешеді: бір агент сұрақты талдаса, екіншісі іздеу жасайды, үшіншісі жауап генерациялайды. Бұл тәсіл зерттеу сатысында, алайда болашағы зор.

Кесте 1.2 - QAS архитектураларының салыстырмалы сипаттамасы

Архитектура	**Негізгі тәсіл**	**Артықшылығы**	**Кемшілігі**	**Галлюцинация**	**Заңнамаға жарамдылығы**
Монолитті	Ережелер	Жылдам, болжамды	Икемсіз	Жоқ	Өте төмен
Екі деңгейлі	ДБ іздеу	Нақты	Контекст шектеулі	Жоқ	Төмен
Үш деңгейлі	NLP + ДБ	Икемді	Күрделі баптау	Минимал ды	Орташа

Кесте 1.2 жалғасы

Гибридті	Көп дереккөз	Толық қамту	Баяу	Минималды	Орташа
LLM негізді	Генерация	Икемді, жақсы	Галлюцинация	Жоғары	Орташа
RAG	Іздеу + генерация	Дәл + икемді	Индексация	Төмен	Жоғары
Көпагенттi	Агенттер желісі	Күрделі сұрақтар	Зерттеу сатысы	Орташа	Перспективалы

Кесте 1.2-де QAS негізгі архитектуралары салыстырмалы түрде көрсетілген. Әр архитектураның жұмыс істеу тәсілі, артықшылықтары мен шектеулері, сондай-ақ галлюцинацияға бейімділігі және заңнамалық жүйелерде қолдануға жарамдылығы талданған.

1.2. Жасанды интеллект және LLM модельдерінің заңнама саласындағы қолданылуы

Соңғы жылдары жасанды интеллект технологиялары, әсіресе үлкен тілдік модельдер (LLM), құқықтық ақпаратты өңдеу, заңнамалық құжаттарды талдау және сұрақ-жауап жүйелерін құру саласында кеңінен қолданыла бастады. LLM модельдерінің дамуы Transformer архитектурасымен тікелей байланысты. Transformer моделі рекуррентті және сверткалық желілерге сүйенбей, self-attention механизмін қолдану арқылы мәтіндегі алыс контекстік байланыстарды тиімді өңдеуге мүмкіндік берді. Бұл архитектура қазіргі GPT, LLaMA, Gemini, Claude, mT5 сияқты ірі тілдік модельдердің негізін құрайды.

Құқық саласында LLM модельдері заңнамалық мәтіндерді түсіндіру, құқықтық құжаттарды іздеу, келісімшарттарды талдау, сот шешімдерін болжау, құқықтық қорытынды жасау және пайдаланушы сұрақтарына жауап беру сияқты міндеттерде қолданылады [18]. 2025 жылғы шолу жұмыстары LLM модельдерінің құқықтық құжаттарды дайындау, құқықтық зерттеу, compliance monitoring, case analysis және құқықтық білім беру бағыттарында белсенді зерттеліп жатқанын көрсетеді. Сонымен қатар, бұл зерттеулер құқық саласында LLM қолданудың негізгі тәуекелдері ретінде дәлдік, интерпретациялану, этикалық жауапкершілік және сенімділік мәселелерін атап өтеді [19].

Заңнамалық сұрақ-жауап жүйелері үшін LLM модельдерінің басты артықшылығы - табиғи тілде қойылған сұрақты түсініп, пайдаланушыға түсінікті форматта жауап бере алуы. Дегенмен құқық саласында жауап тек тілдік тұрғыдан дұрыс болуы жеткіліксіз. Жауап нақты заң нормаларына, баптарға,

тармақтарға және қолданыстағы құқықтық дереккөздерге сүйенуі тиіс. Сондықтан құқықтық LLM жүйелері жай генеративті модель ретінде емес, заңнамалық деректер қорымен байланысқан, дәлелді және тексерілетін жауап беретін интеллектуалды жүйе ретінде қарастырылады.

Қазіргі зерттеулер құқықтық LLM модельдерін бірнеше бағытта қарастырады: жалпы мақсаттағы LLM модельдерін құқықтық тапсырмаларда қолдану; құқықтық мәтіндермен қосымша оқытылған domain-specific модельдер; retrieval-based немесе RAG архитектурасымен толықтырылған генеративті жүйелер; құқықтық reasoning және benchmark-based бағалау жүйелері. Мысалы, LegalBench benchmark жұмысы құқықтық reasoning қабілеттерін бағалау үшін 162 тапсырмадан тұратын кешенді benchmark ұсынды және 20 open-source және commercial LLM моделін салыстырды. Бұл зерттеу LLM модельдерінің құқықтық reasoning қабілетін жалпы NLP метрикаларымен ғана емес, құқықтық ойлау түрлері бойынша бөлек бағалау қажеттігін көрсетті [20].

Қытай құқықтық домені үшін ұсынылған LawBench benchmark жұмысы LLM модельдерінің құқықтық білімін үш деңгейде бағалайды: құқықтық білімді есте сақтау, құқықтық мәтінді түсіну және құқықтық білімді нақты тапсырмаларда қолдану. Бұл зерттеуде 51 LLM моделі бағаланып, GPT-4 басқа модельдерден жоғары нәтиже көрсеткенімен, құқықтық тапсырмаларда толық сенімді және қолдануға дайын LLM модельдеріне әлі қол жеткізілмегені атап өтіледі [21].

Құқықтық сұрақ-жауап жүйелерінде LLM модельдерінің сенімділігін арттыру үшін Retrieval-Augmented Generation (RAG) архитектурасы кеңінен қолданылады. RAG тәсілі генеративті модельдің параметрлік білімін сыртқы дереккордан алынған нақты құжаттармен толықтырады. Lewis et al. ұсынған RAG моделі генерация процесін pre-trained seq2seq model және dense vector index арқылы алынған сыртқы біліммен біріктіреді. Бұл тәсіл әсіресе knowledge-intensive NLP tasks үшін тиімді болып саналады [16].

Құқықтық салада RAG архитектурасының маңызы ерекше. Себебі заңнамалық ақпарат жиі жаңарып отырады, ал LLM модельдерінің ішкі білімі белгілі бір уақытқа дейінгі деректермен шектеледі. Сонымен қатар, құқықтық жауап дәл бапқа, тармаққа немесе нормативтік актіге негізделуі керек. Сондықтан RAG құқықтық сұрақ-жауап жүйелерінде үш негізгі мәселені шешуге мүмкіндік береді: біріншіден, жауапты заңнамалық дереккөздермен негіздеу; екіншіден, модель галлюцинациясын азайту; үшіншіден, пайдаланушыға тексерілетін құқықтық дәлел ұсыну.

Бұл бағытта Antoine Louis және әріптестері ұсынған long-form legal question answering жұмысы маңызды орын алады. Авторлар retrieve-then-read pipeline негізінде ұзын форматтағы құқықтық жауаптар генерациялау әдісін ұсынды және 1868 сараптамалық аннотацияланған француз тіліндегі құқықтық сұрақтан тұратын LLeQA dataset әзірледі. Бұл жұмыс құқықтық сұрақ-жауап жүйесінде retrieval кезеңі мен генерация кезеңін бөлек бағалау қажет екенін көрсетеді [22].

LLM модельдерінің құқық саласындағы ең үлкен шектеулерінің бірі - hallucination, яғни модельдің нақты құқықтық дереккөздерге сәйкес келмейтін, бірақ сенімді көрінетін қате жауап генерациялауы. Stanford зерттеушілері жүргізген Large Legal Fictions зерттеуінде жалпы мақсаттағы LLM модельдері құқықтық сұрақтарға жауап беру кезінде елеулі деңгейде hallucination жіберетіні көрсетілді. Зерттеу ChatGPT-4, GPT-3.5 және Llama 2 модельдерінің нақты тексерілетін құқықтық сұрақтарда қате құқықтық фактілерді генерациялау қаупі бар екенін дәлелдейді [23].

2024 жылғы Stanford HAI талдауында құқықтық AI модельдері кейбір сұрақтарда кемінде әрбір алтыншы жағдайда hallucination жіберетіні көрсетілді. Бұл құқықтық салада LLM модельдерін адам бақылауынсыз қолдану қауіпті екенін дәлелдейді. Құқықтық жауаптың қате болуы тек ақпараттық қателік емес, нақты азаматтық, әкімшілік немесе құқықтық салдарға әкелуі мүмкін [24].

Осы себепті қазіргі құқықтық LLM зерттеулерінде жауаптың groundedness, faithfulness және citation accuracy көрсеткіштері маңызды орын алады. Яғни жүйе тек дұрыс жауап беріп қана қоймай, сол жауаптың қай құқықтық дереккөзге сүйенгенін де көрсетуі тиіс. 2026 жылы ұсынылған Legal RAG Bench жұмысы құқықтық RAG жүйелерін end-to-end бағалау үшін retrieval failure, reasoning error және hallucination сияқты қателерді бөлек талдауды ұсынады [25]. Бұл зерттеу құқықтық RAG жүйесінде retrieval сапасы жалпы жауап сапасының негізгі шектеуші факторы екенін көрсетеді.

LLM модельдерін құқықтық доменге бейімдеудің тағы бір бағыты - құқықтық корпус негізінде domain-specific pretraining немесе continued pretraining жүргізу. MultiLegalPile жұмысы 24 тіл мен 17 юрисдикцияны қамтитын 689GB көлеміндегі көптілді құқықтық корпус ұсынды [26]. Авторлар құқықтық доменге арналған жоғары сапалы деректердің LLM және legal NLP модельдерін оқыту үшін маңызды екенін көрсетеді. Бұл жұмыс құқықтық мәтіндерге арналған көптілді деректердің көп жағдайда ағылшын тілімен шектелетінін және low-resource тілдер үшін арнайы корпус құру қажеттігін дәлелдейді.

Кесте 1.3 - Құқықтық LLM жүйелерінің негізгі бағыттары

Бағыт	**Негізгі мақсаты**	**Артықшылығы**	**Шектеуі**
Жалпы LLM	Табиғи тілдегі сұрақтарға жауап беру	Тілдік генерациясы күшті	Заңнамалық дәлдікке кепілдік бермейді
Legal LLM	Құқықтық мәтіндерге бейімделген модель	Терминологияны жақсы түсінеді	Көп тілдерде дайын модель жоқ
LLM және RAG	Жауапты нақты дереккөзбен негіздеу	Галлюцинацияны азайтады	Retrieval сапасына тәуелді
Legal benchmark	Модельдің құқықтық қабілетін бағалау	Нақты task-based бағалау береді	Қазақ тілі үшін benchmark жоқ

Қазақ тілі үшін бұл мәселе одан да өзекті. Қазақ тілі агглютинативті

құрылымға ие, құқықтық мәтіндері ресми стильде жазылады, ал пайдаланушы сұрақтары көбіне бейресми, қысқа немесе тұрмыстық тілде беріледі. Сондықтан жалпы мақсаттағы LLM немесе қарапайым RAG архитектурасы қазақ тіліндегі заңнамалық сұрақтарға тұрақты және дәл жауап беруге әрдайым жеткіліксіз болуы мүмкін. Мұндай жағдайда сұрақты заңи стильге түрлендіру, құқықтық терминдермен кеңейту және retrieval кезеңіне дейін семантикалық алдын ала өңдеу қажет. Кесте 1.3-те құқықтық сұрақ-жауап жүйелерін әзірлеуде қолданылатын негізгі бағыттар салыстырылып көрсетілген.

Жалпы LLM модельдері табиғи тілдегі генерация мүмкіндіктерімен ерекшеленгенімен, құқықтық дәлдікке толық кепілдік бере алмайды. Сондықтан соңғы зерттеулерде Legal LLM және LLM мен RAG бағыттарына басымдық берілуде. Legal LLM модельдері құқықтық терминология мен заңнамалық құрылымды жақсы түсінеді, алайда қазақ тілі үшін арнайы бейімделген модельдердің болмауы маңызды мәселе болып табылады. Ал RAG архитектурасы генерацияны нақты құқықтық дереккөздермен байланыстыру арқылы галлюцинацияны азайтады және заңнамалық жауаптардың сенімділігін арттырады. Сонымен қатар халықаралық legal benchmark жүйелері модельдердің құқықтық reasoning қабілетін бағалауға мүмкіндік бергенімен, қазақ тілінде мұндай benchmark-тардың жоқтығы отандық QA dataset пен арнайы бағалау метрикаларын әзірлеу қажеттігін көрсетеді.

Кесте 1.4 - Құқықтық LLM зерттеулері

Зерттеу	**Жылы**	**Негізгі үлесі**
Vaswani et al. – Transformer [9]	2017	Self-attention негізіндегі Transformer архитектурасы
Lewis et al. -RAG [16]	2020	Генерацияны сыртқы біліммен біріктіру
LegalBench [20]	2023	162 құқықтық reasoning task
LawBench [21]	2024	Құқықтық білімді есте сақтау, түсіну, қолдану деңгейлері
Louis et al. -LleQA [22]	2024	Long-form Legal QA және retrieve-then-read pipeline
Dahl et al. - Legal hallucinations [23]	2024	LLM галлюцинациясын құқықтық доменде жүйелі талдау
MultiLegalPile [26]	2024	689GB көптілді құқықтық корпус
Legal RAG Bench [25]	2026	RAG жүйесінде retrieval, reasoning және hallucination қателерін бөлек бағалау

Кесте 1.4-те құқықтық жасанды интеллект және LLM негізіндегі сұрақ-жауап жүйелері саласындағы негізгі ғылыми зерттеулер көрсетілген. Transformer архитектурасынан бастап қазіргі Legal RAG және Legal Benchmark бағыттарына дейінгі зерттеулер құқықтық доменде генеративті модельдерді қолданудың теориялық және практикалық негізін қалыптастырды. Аталған жұмыстар заңнамалық QA жүйелерінде retrieval сапасы, reasoning қабілеті және галлюцинация мәселелері негізгі шектеуші факторлар екенін көрсетеді. Сонымен қатар көптілді құқықтық корпустар мен арнайы benchmark жүйелерінің дамуы low-resource тілдер үшін, соның ішінде қазақ тілі үшін, құқықтық деректер қорын қалыптастыру және RAG архитектурасын семантикалық препроцессингпен толықтыру қажеттігін дәлелдейді.

Жүргізілген талдау LLM модельдерінің құқықтық сұрақ-жауап жүйелерінде үлкен мүмкіндікке ие екенін көрсетеді. Алайда құқық саласында LLM-ді тікелей қолдану жеткіліксіз, себебі құқықтық жауап дәл, тексерілетін және нақты нормативтік дереккөздерге негізделуі тиіс. Қазіргі зерттеулер RAG архитектурасының құқықтық доменде галлюцинацияны азайту және жауапты дереккөзбен негіздеу үшін маңызды екенін дәлелдейді. Сонымен бірге low-resource және агглютинативті тілдерде, соның ішінде қазақ тілінде, retrieval сапасы тек RAG арқылы емес, сұрақты алдын ала семантикалық өңдеу арқылы да жақсартылуы тиіс. Сондықтан осы диссертацияда ұсынылатын SemanticPrep тәсілі қазақ тіліндегі заңнамалық сұрақ-жауап жүйелері үшін ғылыми және практикалық тұрғыдан негізделген шешім болып табылады.

1.3. Заңнамалық мәтіндердің және қазақ тілінің лингвистикалық ерекшеліктері

Қазақ тілі түркі тілдер тобына жатады және агглютинативті морфологиялық құрылыммен сипатталады. Агглютинативті тілдерде грамматикалық мағыналар сөз түбіріне аффикстерді дәйекті жалғау арқылы беріледі, нәтижесінде көптеген сөз тұлғалары мен жоғары морфологиялық вариативтілік пайда болады. Табиғи тілді өңдеу жүйелері үшін бұл ерекшелік токенизация, лемматизация, морфологиялық талдау және ақпараттық іздеу кезінде елеулі қиындықтар туғызады [4].

Негізгі мәселелердің бірі - мүмкін сөз тұлғаларының көптігінен туындайтын лексикалық сирек кездесушілік (lexical sparsity). Құқықтық мәтіндерде бұл мәселе күрделі заңдық терминдердің, нормативтік актілердің ресми атауларының және ұзақ синтаксистік конструкциялардың болуынан бұрыннан да күрделене түседі. Тіпті шамалы морфологиялық өзгерістер мәтіннің заңдық мағынасына айтарлықтай әсер етуі мүмкін. Мысалы, «заңды» және «заңсыз» сөздері ортақ түбірге ие бола отырып, қарама-қарсы құқықтық мағына білдіреді.

Қазақстандық зерттеулері дәстүрлі статистикалық әдістердің қазақ тілі үшін сөздердің күрделі морфотактикалық құрылымын жеткілікті дәрежеде ескермейтінін, ал тілдің агглютинативті табиғаты мәтінді өңдеуге морфологияны ескеретін (morphology-aware) тәсілдерді қолдануды талап ететінін көрсетті [27]. Авторлар морфема тізбектері мен грамматикалық санаттар арасындағы

ауысуларды ескеретін деректерге негізделген (data-driven) морфологиялық талдау үлгісін ұсынды.

Синтаксистік тұрғыдан қазақ тілі SOV негізгі құрылымында (бастауыш–толықтауыш–баяндауыш) салыстырмалы түрде еркін сөз тәртібімен сипатталады. Септік жүйесінің арқасында сөйлем мүшелерінің семантикалық рөлдері сыңарлардың орын тәртібі өзгерген жағдайда да сақталады. Алайда ақпаратты автоматты түрде алу және құқықтық нормаларды талдау міндеттері үшін мұндай вариативтілік «субъект–іс-қимыл–объект» қатынастарын анықтауды айтарлықтай қиындатады [28].

Қосымша күрделілік мыналарды қамтитын заңдық мәтіндерден туындайды:

– деонтикалық конструкциялар («міндетті», «құқылы», «тыйым салынады»);

– баптар мен заңдар арасындағы сілтемелер;

– нормалардың қолданылу мерзіміне қатысты уақыттық шектеулер;

– күрделі атаулы нысандар (бап нөмірі, күні, билік органы, нормативтік акт).

Қарапайым мәтіндерден айырмашылығы, құқықтық дискурс тек беткейлік лексикалық талдауды ғана емес, заңнама нормалары арасындағы логикалық-семантикалық қатынастарды да түсінуді талап етеді.

Қазақ тілін морфосинтаксистік талдау үшін ақырлы а́втоматтарға (Finite-State Transducer, FST) негізделген ережелер жиынтығы бойынша (rule-based) әдістер де, заманауи нейрондық желілік тәсілдер де қолданылады [29]. Universal Dependencies Kazakh Treebank (UD_KTB) ресурсы ерекше рөл атқарады: ол қазақ мәтіндерінің бірыңғай синтаксистік аннотациясын қамтамасыз етеді және синтаксистік талдау үлгілерін үйретуде пайдаланылады [30].

Қазіргі зерттеулер агглютинативті тілдер үшін ең тиімді тәсілдердің морфологиялық сезімтал мәтін өңдеуі мен контекстік нейрондық желілік репрезентацияларды біріктіретін гибридті әдістер екенін көрсетеді [31]. Заңдық Question Answering міндеттерінде бұл аса маңызды, өйткені морфологиялық талдаудың қателіктері retrieval-механизмдерінің сапасын төмендетуге және заңдық жағынан сенімсіз жауаптар генерациялауға әкелуі мүмкін.

Заң саласындағы заманауи үлкен тілдік үлгілердің негізгі мәселелерінің бірі - галлюцинациялар, яғни іс жүзінде немесе заңдық тұрғыдан дұрыс емес ақпаратты генерациялау. Қазақ тілі үшін бұл мәселе мамандандырылған корпустардың шектеулілігі мен жоғары морфологиялық вариативтіліктен бұрыннан да күрделене түседі. Осыған байланысты заманауи заңдық сұрақ-жауап жүйелері барған сайын RAG технологиясын қолдануда, мұнда генеративті үлгі тиісті нормативтік құжаттарды іздеу механизмдерімен толықтырылады.

1.4. Қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің қазіргі жағдайы

Қазақстанда соңғы жылдары мемлекеттік қызметтерді цифрландыру, электрондық үкімет жүйелерін дамыту және құқықтық ақпаратқа қолжетімділікті арттыру бағытында бірқатар жобалар жүзеге асырылуда. «Цифрлық Қазақстан» бағдарламасы аясында құқықтық ақпараттық жүйелерді автоматтандыру, заңнамалық құжаттарды цифрландыру және азаматтарға

құқықтық кеңес беру қызметтерін электрондық форматқа көшіру жұмыстары жүргізілуде.

Қазіргі уақытта Қазақстанда құқықтық ақпараттың негізгі цифрлық көздері ретінде:

- «Әділет» нормативтік-құқықтық актілердің ақпараттық жүйесі adilet.zan.kz [32];
- eGov мемлекеттік қызметтер порталы https://egov.kz [33];
- Құқықтық ақпарат бірыңғай жүйесі https://zan.gov.kz [34];
- Paragraph ақпараттық жүйесі https://prg.kz/ [35];
- Zqai.kz платформасы https://zqai.kz/ [36];
- Akorda.kz сияқты сервистер https://www.akorda.kz/ru [37] қолданылады.

Алайда бұл жүйелердің басым бөлігі классикалық ақпараттық іздеу қағидаларына негізделген және толыққанды интеллектуалды Question Answering (QA) жүйелері болып табылмайды. Көптеген платформалар пайдаланушыға тек нормативтік құжаттар тізімін немесе алдын ала дайындалған FAQ үлгілерін ұсынады. Табиғи тілде қойылған күрделі сұрақтарды түсіну, құқықтық контекст ескеру және түсіндірмелі жауап генерациялау мүмкіндіктері шектеулі.

Қазақстандағы құқықтық цифрлық инфрақұрылымның ең маңызды элементтерінің бірі - «Әділет» ақпараттық-құқықтық жүйесі. Бұл жүйе Қазақстан Республикасының нормативтік-құқықтық актілеріне қолжетімділік береді және мемлекеттік құқықтық ақпараттың негізгі ресми көзі болып табылады. Дегенмен жүйе негізінен keyword-based search қағидасына сүйенеді және семантикалық сұраныстарды өңдеуде белгілі бір шектеулерге ие. Мысалы, пайдаланушының бейресми немесе тұрмыстық тілде қойылған сұрағы заң мәтініндегі ресми құқықтық терминологиямен сәйкес келмеуі мүмкін.

Қазақ тіліндегі заңнамалық сұрақ-жауап жүйелерінің дамуын тежейтін негізгі факторлардың бірі - low-resource language мәселесі. Қазақ тілінде:

- сапалы құқықтық QA dataset-тердің аздығы;
- annotated legal corpus көлемінің шектеулілігі;
- domain-specific language model-дердің болмауы;
- құқықтық мәтіндерге арналған benchmark жүйелерінің жетіспеуі байқалады.

Сонымен қатар қазақ тілінің агглютинативті табиғаты да құқықтық QA жүйелерін құруды қиындатады. Бір түбірге көптеген аффикстердің жалғануы сөз формаларының күрт көбеюіне алып келеді және lexical sparsity мәселесін туындатады. Мысалы, пайдаланушы «жұмыстан шығарып жіберді» деп сұраса, заң мәтінінде дәл осы форма емес, «еңбек шартын бұзу», «қызметкерді жұмыстан босату» сияқты ресми құқықтық терминдер қолданылуы мүмкін. Мұндай жағдайда классикалық keyword matching әдістері қажетті құқықтық норманы таба алмауы ықтимал.

Қазақстанда қазақ тіліндегі құқықтық QA жүйелеріне қатысты алғашқы ғылыми зерттеулер негізінен BERT, multilingual transformer және rule-based NLP тәсілдеріне сүйенді. Мысалы, 2021 жылы ұсынылған BERT негізіндегі қазақ тіліндегі Question Answering жүйесінде құқықтық сұрақтарды өңдеу үшін

multilingual BERT архитектурасы пайдаланылды [38]. Алайда бұл зерттеулер көбіне шағын деректер жиынтығына негізделіп, толыққанды retrieval-based legal QA архитектурасын қамтымады.

Соңғы жылдары үлкен тілдік модельдер мен RAG архитектурасына негізделген құқықтық жүйелерге қызығушылық арта бастады. 2025 жылы жарияланған «Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation» зерттеуінде қазақ тіліндегі құқықтық сұрақ-жауап жүйесін әзірлеу үшін Adilet, Gov, Zqai және synthetic dataset негізінде QA корпус құрастырылды. Жұмыста GPT-4o mini, Gemini, KazLLM, LLaMA, Phi, Qwen және Mistral модельдері fine-tuning арқылы бейімделіп, ROUGE және METEOR метрикалары бойынша бағаланды. GPT-4o mini моделі ең жоғары нәтижені көрсетті [4].

Бұл зерттеу қазақ тіліндегі құқықтық QA жүйелерін құрудың техникалық мүмкіндігін көрсеткенімен, бірнеше маңызды мәселелер әлі толық шешілмеген:

– retrieval сапасының тұрақсыздығы;
– бейресми сұрақтарды құқықтық стильге түрлендірудің болмауы;
– құқықтық терминдерді семантикалық кеңейтудің жеткіліксіздігі;
– hallucination мәселесі;
– evidence-based verification механизмдерінің әлсіздігі;
– morphology-aware preprocessing кезеңінің болмауы.

Қазақстанда құқықтық AI бағытында мемлекеттік деңгейде де жобалар іске қосылуда. 2025 жылы e-ZanKomegi платформасына жасанды интеллект негізіндегі құқықтық көмек көрсету жүйесін енгізу басталды. Жоба азаматтардың құқықтық кеңеске тәулік бойы қол жеткізуін қамтамасыз етуге бағытталған. Сонымен қатар chatbot және AI assistant элементтері тестіленуде. Соған қарамастан қазіргі жүйелердің басым бөлігі:

– retrieval-based reasoning қолданбайды;
– құқықтық дәлелдерді генерацияланған жауаппен байланыстыра алмайды;
– пайдаланушы сұрағын құқықтық терминология тұрғысынан қайта түрлендірілмейді;
– қазақ тілінің морфологиялық ерекшеліктерін толық ескермейді.

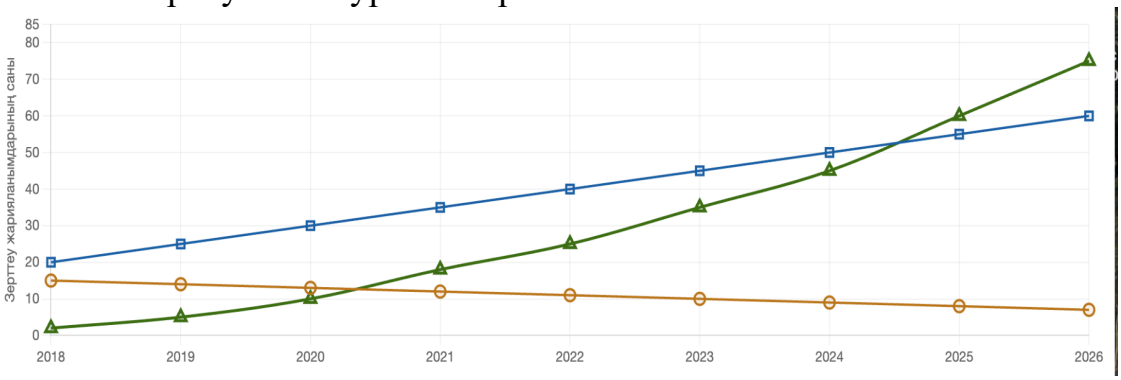
Халықаралық тәжірибемен салыстырғанда Қазақстандағы құқықтық QA зерттеулері әлі бастапқы даму кезеңінде екенін байқауға болады. Мысалы, қытай тілінде Lawformer сияқты құқықтық ұзын мәтіндерге бейімделген арнайы transformer модельдері ұсынылса, француз тілінде LLeQA және statutory retrieval dataset-тері жасалған, ал Urdu-English құқықтық QA үшін арнайы LEGAL-UQA dataset әзірленген [39, 40].

Қазақ тілінде де соңғы жылдары KazQAD dataset-і ұсынылды. Бұл ашық домендегі қазақ тіліндегі Question Answering dataset-і retrieval және ODQA тапсырмалары үшін маңызды ресурс болып табылады [41]. Алайда ол жалпы энциклопедиялық және open-domain QA бағытына арналған және құқықтық доменге арнайы бейімделмеген.

Сондықтан қазіргі таңда қазақ тіліндегі заңнамалық сұрақ-жауап жүйелері үшін келесі ғылыми мәселелер өзекті болып отыр:

– құқықтық domain-specific dataset қалыптастыру;
– morphology-aware retrieval жүйелерін әзірлеу;
– бейресми сұрақтарды құқықтық стильге семантикалық түрлендіру;
– құқықтық терминдерді автоматты кеңейту;
– RAG архитектурасын low-resource language жағдайына бейімдеу;
– evidence-based legal answer generation тәсілдерін дамыту.

Соңғы жылдары құқықтық AI және retrieval-based question answering жүйелеріне қызығушылықтың артуы байқалады. Бұл үрдіс әсіресе LLM және RAG архитектураларының пайда болуынан кейін күшейді. «Scopus, Web of Science және Google Scholar дерекқорларынан жиналған деректер негізінде авторлық талдау жасалды. «Kazakh NLP», «legal question answering» және «retrieval-augmented generation» кілт сөздері бойынша анықталған 85 ғылыми жазбаның жыл бойынша таралуы 1.1-суретте берілген.



Сурет 1.1 Ғылыми жарияланымдардың динамикасы көрсетілген.

Көк сызық жалпы Kazakh NLP зерттеулерінің тұрақты өсуін көрсетсе, жасыл сызық заңнамалық QA және LLM-based Legal AI бағытындағы жұмыстар санының соңғы жылдары күрт артқанын көрсетеді. Әсіресе 2023 жылдан кейін RAG, legal reasoning және LLM fine-tuning бағытындағы зерттеулердің белсенді дамуы байқалады.

Сонымен қатар дәстүрлі rule-based және keyword-based жүйелерге арналған зерттеулер саны біртіндеп азаюда. Бұл қазіргі құқықтық AI жүйелерінің semantic retrieval және large language models негізіндегі архитектураларға көшу үрдісін дәлелдейді.

Жүргізілген салыстырмалы талдау қазіргі уақытта қазақ тіліндегі құқықтық сұрақ-жауап жүйелері негізінен baseline LLM және keyword-based retrieval деңгейінде екенін көрсетті. Сонымен қатар morphology-aware semantic preprocessing және evidence-based retrieval мәселелері жеткілікті зерттелмеген.

1.5. Отандық және шетелдік ғылыми зерттеулерді салыстырмалы талдау

Құқықтық сұрақ-жауап жүйелерін зерттеу соңғы жылдары бірнеше бағытта дамып келеді: құқықтық ақпаратты іздеу, заң мәтіндерін семантикалық талдау, үлкен тілдік модельдерді құқық саласына бейімдеу, RAG архитектурасын қолдану және құқықтық жауаптардың дәлдігін сараптамалық бағалау. Шетелдік зерттеулерде бұл бағыттар салыстырмалы түрде жүйелі дамыған: арнайы

құқықтық датасеттер, benchmark жүйелері, retrieval-based модельдер және long-form legal QA жүйелері ұсынылған. Ал қазақ тіліндегі зерттеулер әлі бастапқы даму кезеңінде және негізінен деректер жинау, LLM модельдерін бастапқы бағалау және low-resource жағдайында QA жүйесін бейімдеу мәселелеріне бағытталған.

Отандық зерттеулердің негізгі ерекшелігі - қазақ тілінің ресурсы шектеулі және морфологиялық жағынан күрделі тіл екендігін ескеру қажеттілігі. Қазақ тіліндегі ашық доменді QA үшін KazQAD датасеті ұсынылған. Бұл ресурс шамамен 6000 бірегей сұрақты және 12000-ға жуық passage-level relevance judgement қамтиды, алайда ол құқықтық доменге емес, жалпы open-domain QA міндеттеріне арналған (ACL Anthology) [41]. Сондықтан заңнамалық сұрақ-жауап жүйесін құру үшін құқықтық мәтіндерге негізделген арнайы корпус пен domain-specific QA dataset қажет.

Қазақ тіліндегі құқықтық QA бағыты бойынша маңызды зерттеулердің бірі - Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation. Бұл жұмыста Adilet, Zqai, Gov және synthetic дереккөздері негізінде қазақ тіліндегі құқықтық сұрақ-жауап жүйесі үшін деректер жиынтығы құрастырылып, бірнеше LLM моделі бағаланған [4]. Зерттеу қазақ тіліндегі заңнамалық QA жүйесін құрудың мүмкіндігін көрсетеді, бірақ retrieval сапасы, сұрақты құқықтық стильге келтіру, галлюцинацияны азайту және дәлелге негізделген жауап беру мәселелері әлі толық шешілмеген.

Шетелдік зерттеулерде құқықтық QA жүйелері әлдеқайда тереңірек дамыған. Мысалы, Antoine Louis және әріптестері ұсынған LLeQA датасеті француз тіліндегі құқықтық сұрақ-жауап жүйелеріне арналған. Бұл датасет 1868 сараптамалық аннотацияланған құқықтық сұрақты қамтиды және retrieve-then-read pipeline негізінде ұзын форматтағы құқықтық жауаптарды генерациялауға бағытталған [22]. Бұл жұмыс құқықтық QA жүйесінде retrieval кезеңін, жауап генерациясын және жауаптың құқықтық дереккөздермен байланысын бөлек бағалау қажеттігін көрсетеді.

Құқықтық reasoning қабілетін бағалау үшін LegalBench benchmark ұсынылған. Ол 162 құқықтық тапсырманы қамтиды және 20 open-source және commercial LLM моделін салыстыруға мүмкіндік береді [20]. Бұл зерттеу құқық саласында LLM модельдерін тек жалпы NLP метрикаларымен емес, құқықтық reasoning түрлері бойынша бағалау қажет екенін дәлелдейді.

Кесте 1.5 - Отандық және шетелдік зерттеулерді салыстырмалы талдау

Бағыт	**Отандық зерттеулердегі жағдай**	**Шетелдік зерттеулердегі жағдай**
QA dataset	KazQAD бар, бірақ құқықтық доменге арналмаған	LLeQA, BSARD сияқты құқықтық датасеттер бар

Кесте 1.5 жалғасы

Legal QA	Алғашқы LLM-based зерттеулер бар	Retrieval + long-form QA жүйелері дамыған
Retrieval	Көбіне keyword немесе baseline RAG деңгейінде	Dense retrieval, hybrid retrieval, graph-based retrieval қолданылады
LLM бағалау	GPT, Gemini, LLaMA, KazLLM сияқты модельдер бастапқы бағаланған	LegalBench сияқты benchmark жүйелері бар
RAG	Қолданыла бастады, бірақ evidence verification әлсіз	Legal RAG, retrieve-then-read, grounded answer generation дамыған
Тілдік ерекшелік	Агглютинативтілік және low-resource мәселе бар	Көп зерттеу ағылшын, француз, қытай тілдеріне бағытталған

Кесте 1.5 көрсетілген салыстырмалы талдау нәтижесінде келесі ғылыми олқылықтар анықталды:

– Қазақ тіліндегі заңнамалық сұрақ-жауап жүйелері үшін арнайы, сараптамалық тексерілген QA dataset көлемі жеткіліксіз.

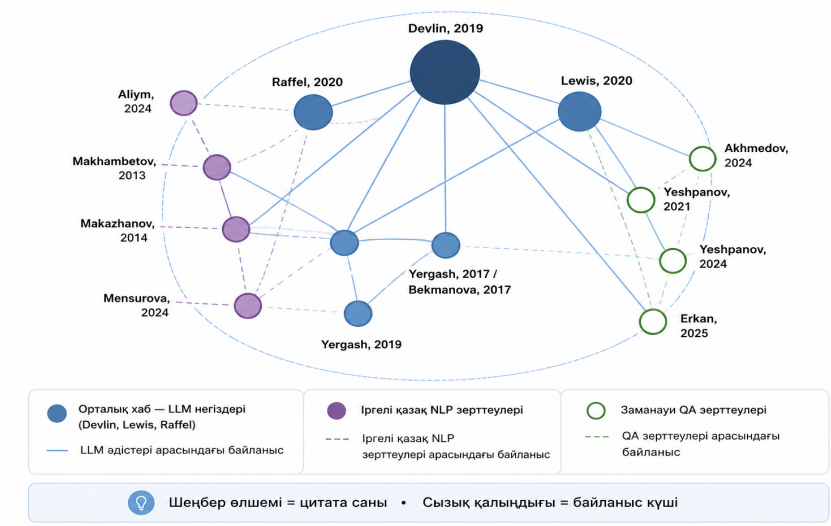
– Қолданыстағы жүйелер пайдаланушының бейресми сұрағын ресми құқықтық стильге автоматты түрде түрлендіруді толық шешпейді.

– Қазақ тілінің агглютинативті морфологиясы retrieval сапасына әсер етеді, бірақ morphology-aware semantic preprocessing жеткілікті зерттелмеген.

– RAG архитектурасы қолданылғанымен, retrieval сапасын жақсартатын алдын ала семантикалық өңдеу кезеңі әлсіз дамыған.

– Құқықтық жауаптардың дәлдігін бағалау үшін автоматты метрикалармен қатар сараптамалық бағалау және evidence-based verification қажет.

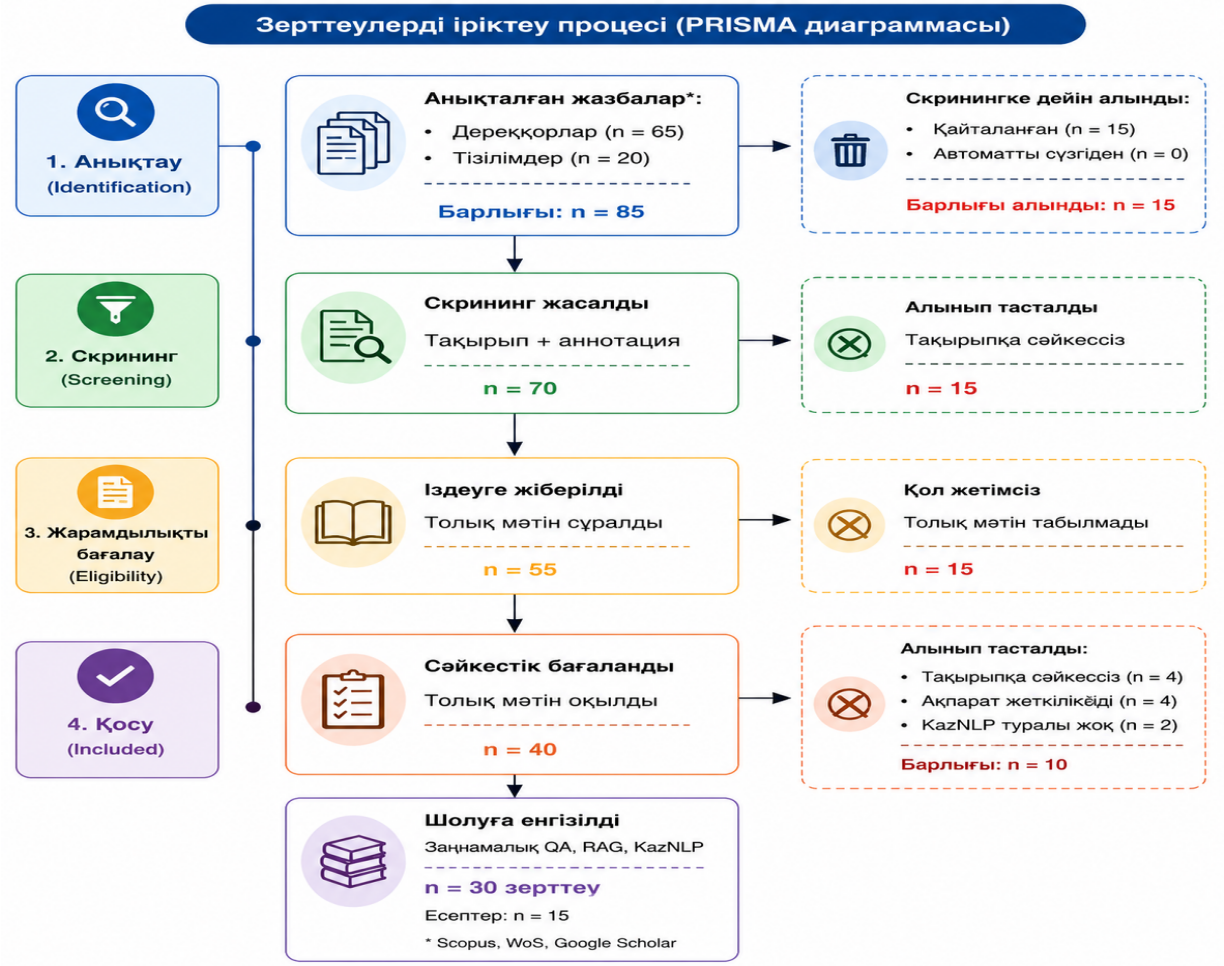
«Scopus» дерекқорынан алынған citation деректері негізінде авторлық визуализация жасалды, бұл 1.2-суретте көрсетілген. Желідегі әрбір түйін бір ғылыми жұмысты білдіреді; түйін көлемі сол жұмыстың цитатаалану санына пропорционал, ал сызықтар зерттеулер арасындағы тікелей сілтеме байланысын көрсетеді.



Сурет 1.2 – Ғылыми жұмыстар арасындағы байланыстар

Citation network орталығында Devlin et al. (BERT) [10], Lewis et al. (RAG) [42] және Raffel et al. (T5/mT5) [43] жұмыстары орналасқан. Бұл зерттеулер қазіргі Legal QA және LLM-based жүйелерінің негізгі архитектуралық негізін құрайды.

Сонымен қатар Makhambetov, Makazhanov, Yergesh және Bekmanova сияқты қазақстандық және түркі тілдері бойынша зерттеулер қазақ тілінің агглютинативті морфологиясын өңдеу мәселелерін қарастырады. Соңғы KazQAD және low-resource Legal AI бағытындағы жұмыстар retrieval-based және semantic preprocessing тәсілдерінің өзектілігін көрсетеді.



Сурет 1.3 – Ғылыми жұмыстар арасындағы байланыстар

Әдебиеттерді жүйелі талдау барысында қолданылған PRISMA әдістемесінің схемасы Сурет 1.3-те көрсетілген. Ғылыми жұмыстар Scopus, Web of Science және Google Scholar дерекқорларынан жиналып, тақырыптық сәйкестік, толық мәтіннің қолжетімділігі және Kazakh NLP мен заңнамалық QA бағыттарына қатыстылығы бойынша іріктелді.

Жалпы 85 ғылыми жазба анықталып, қайталанатын және тақырыпқа сәйкес келмейтін жұмыстар алынып тасталды. Соңғы шолуға 30 ғылыми зерттеу енгізілді. Бұл жұмыстар Kazakh NLP, retrieval-augmented generation, legal question answering және low-resource language бағыттарын қамтиды.

1.6 1 бөлім бойынша қорытынды

Бірінші бөлімде жүргізілген зерттеулерді талдау нәтижесінде заңнамалық сұрақ-жауап жүйелерінің даму тенденциялары, Қазақстандағы қолданыстағы жүйелердің жағдайы және олардың шектеулері жан-жақты талдау жасалды.

Қолданыстағы жүйелердің нақты сапасын бағалау мақсатында авторлармен 100 сұрақ-жауап жұбынан тұратын эксперименттік датасет құрастырылды. Бұл датасет заңнамалық тақырыптардың кең ауқымын қамтиды: еңбек құқығы, тұрғын үй заңнамасы, әкімшілік тәртіп, азаматтық-құқықтық қатынастар және т.б. Сұрақтар бейресми, күнделікті тілде тұжырымдалып, жауаптар Қазақстан Республикасының нормативтік-құқықтық актілерінің нақты баптарына негізделді. Аталған датасет негізінде Қазақстандағы қолданыстағы алты жүйеге эксперименттік бағалау жүргізілді. Нәтижелер 1.6-кестеде көрсетілген.

Кесте 1.6 – Қазақ тілін қамтитын заңнамалық сұрақ-жауап жүйелерін эксперименттік салыстыру

Жүйе / Модель атауы	**Техникалық шешімі**	**Метрикалар**	**Эксперименттік талдау бойынша жасалған тұжырым**
Әділет / Параграф https://prg.kz/ https://adilet.zan.kz	Elasticsearch / BM25 негізіндегі ақпараттық іздеу жүйесі; құжаттарды индекстеу және кілт сөздер арқылы іздеу	MRR: 0.32 Recall@5: 0.25	Семантикалық түсіну жоқ; сұрақ-жауап емес, тек құжаттар тізімін қайтарады; пайдаланушы өзі интерпретация жасауы керек
eGov / eGovKenes (1414) https://t.me/Kenes1414Bot GOV.KZ	Rule-based жүйе, Decision Tree және сценарийлік логика	-	Еркін тілдегі сұрақтарды түсінбейді; тек алдын ала анықталған сценарийлер; генеративті емес
LawVision https://lawvision.kz	-	BERTScore: 0.60 ROUGE-L: 0.35 METEOR: 0.30	Сұранысты дұрыс түсінбейді, 20% сұрақтарға жауаптың болмауы және заңнамалық дәлсіздіктер кездеседі. Жүйе артық ақпарат генерациялап, жауап құрылымын күрделендіреді және қайталанатын мазмұн береді.Бұл факторлар жауаптың релеванттылығын төмендетіп, жалпы жүйе сапасының төмен болуына әкеледі. Қазақ тілінде морфологиялық қателіктер кездеседі.
ChatGPT (GPT-4o)	Transformer (LLM), генеративті QA, zero-shot / few-shot reasoning	BERTScore: 0.67 ROUGE-L: 0.58 METEOR: 0.62	ҚР заңнамасына тікелей қосылмаған; hallucination болуы мүмкін; нақты құқықтық дәлдікке кепілдік жоқ

Кесте 1.6 жалғасы

Gemini (Google)	Multimodal Transformer, контекстке негізделген генерация	BERTScore: 0.72 ROUGE-L: 0.45 METEOR: 0.50	Қазақ тілін жақсы түсінеді, бірақ ҚР заң базасына тікелей қолжетімділік жоқ; құқықтық нақтылық тұрақсыз
Claude 3.5 (Anthropic)	Constitutional AI + Transformer, логикалық reasoning	BERTScore: 0.55 ROUGE-L: 0.40 METEOR: 0.35	Заңгерлік reasoning жақсы, бірақ нақты ҚР заңнамасы бойынша дереккөз жоқ; жаңартылған заңдарды білмеуі мүмкін

Эксперименттік бағалау нәтижелері бірқатар маңызды тұжырымдарға негіз болды. Іздеуге негізделген жүйелер (Әділет/Параграф) MRR: 0.32, Recall@5: 0.25 деңгейінде ғана нәтиже беріп, семантикалық түсіну мүмкіндіктерінің жоқтығын айқын көрсетті. Бұл жүйелер пайдаланушы сұрағын заңнамалық терминологиямен сәйкестендіре алмайды және тек кілт сөздер негізінде іздеу жүргізеді. Ережелік жүйелер (eGov/eGovKenes) тек алдын ала анықталған сценарийлерге ғана жауап бере алатындықтан, бейресми немесе күрделі сұрақтарды өңдеуде тиімсіз болып шықты. Жалпы мақсаттағы LLM модельдері (ChatGPT, Gemini, Claude) тілдік генерация сапасы бойынша салыстырмалы түрде жоғары нәтиже көрсеткенімен (BERTScore: 0.55–0.72), Қазақстан Республикасының нақты заңнамалық базасына тікелей қолжетімдігі жоқтықтан жауаптардың заңдылық дәлдігіне кепілдік берілмейді. LawVision жүйесі жергілікті заңнамамен жұмыс жасауға бағытталғанымен, BERTScore: 0.60, ROUGE-L: 0.35, METEOR: 0.30 деңгейіндегі нәтижелері жеткіліксіз деп бағаланды; жүйе қазақ тілінің морфологиялық ерекшеліктерін толық ескермейтіндіктен бейресми сұрақтарды дұрыс өңдей алмайды.

Осылайша, жүргізілген эксперименттік талдау қолданыстағы жүйелердің барлығында ортақ бір мәселенің шешілмеген күйінде қалып отырғанын дәлелдейді: пайдаланушының қарапайым тілдегі сұрағы мен заңнамалық мәтіндегі ресми терминология арасындағы семантикалық алшақтық. Бұл нәтижелер қазақ тіліндегі заңнамалық сұрақ-жауап жүйелері саласында ғылыми зерттеу мәселесінің өзектілігі мен шешілмеген сипатын растайды және осы диссертациялық жұмыстың бағытын ғылыми тұрғыдан негіздейді.

2. ЗАҢНАМАЛЫҚ ДЕРЕКТЕР ҚОРЫН ҚАЛЫПТАСТЫРУ ЖӘНЕ LLM МОДЕЛЬДЕРІН ЗЕРТТЕУ

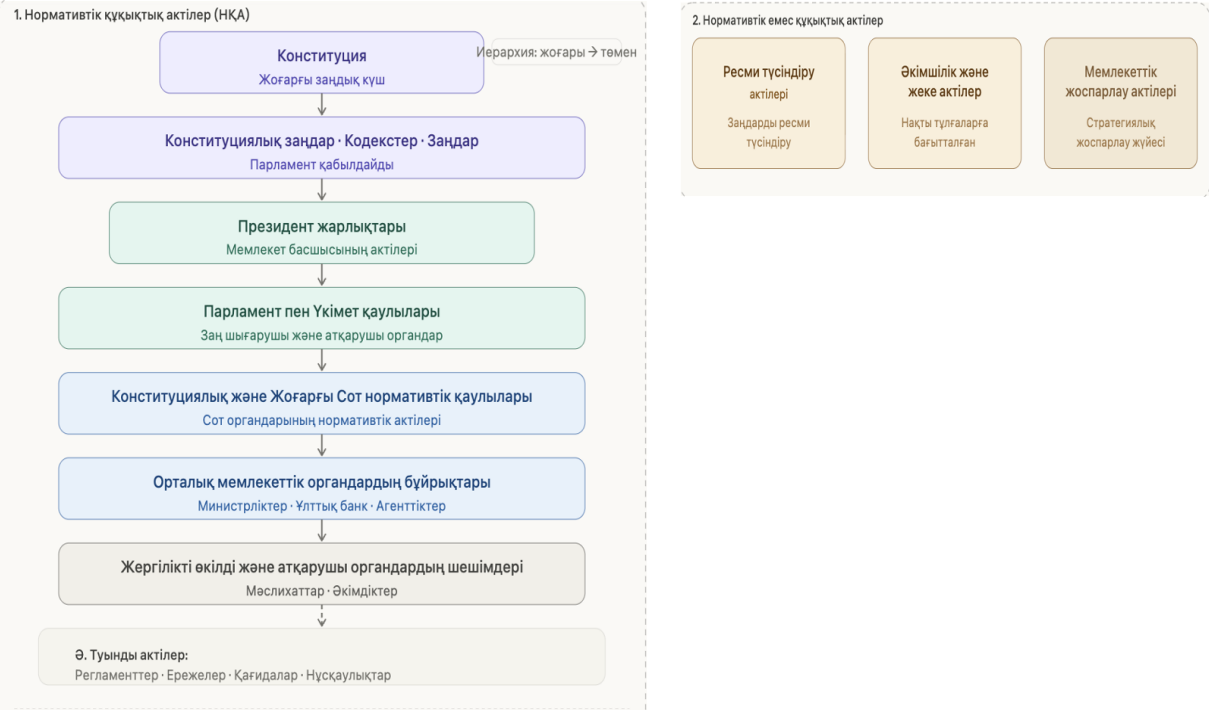
Бұл бөлімде Қазақстан Республикасының заңнамалық мәтіндері негізінде мамандандырылған деректер қорын қалыптастыру және үлкен тілдік модельдерді (LLM) құқықтық доменге бейімдеу әдістері қарастырылады. Зерттеу барысында құқықтық деректерді жинау, құрылымдау, алдын ала өңдеу және аннотациялау кезеңдері жүйелі түрде ұйымдастырылды. Сонымен қатар заңнамалық мәтіндерге арналған QA dataset қалыптастыру, LLM модельдерін fine-tuning арқылы бейімдеу және RAG архитектурасын интеграциялау тәсілдері зерттелді. Ұсынылған әдістер қазақ тілінің агглютинативті ерекшеліктерін, құқықтық терминологияны және low-resource language мәселелерін ескере отырып, дәл және сенімді интеллектуалды заңнамалық сұрақ-жауап жүйесін құруға бағытталған.

2.1 Қазақстан Республикасының құқықтық деректер кеңістігін талдау

Бұл бөлімде Қазақстан Республикасындағы құқықтық ақпараттық кеңістіктің құрылымы, негізгі цифрлық дереккөздері және заңнамалық мәтіндердің ерекшеліктері талданады. Зерттеу барысында мемлекеттік құқықтық порталдар, нормативтік-құқықтық актілер базалары және электрондық құқықтық ресурстар қарастырылып, олардың деректер көлемі, құрылымы және заңнамалық сұрақ-жауап жүйелерінде қолданылу мүмкіндіктері бағаланды. Сонымен қатар құқықтық мәтіндердің лингвистикалық, терминологиялық және семантикалық ерекшеліктері зерттеліп, қазақ тіліндегі заңнамалық деректерді өңдеу кезінде туындайтын негізгі мәселелер анықталды. Бұл талдау кейінгі кезеңдерде мамандандырылған құқықтық корпус пен QA dataset қалыптастырудың теориялық және практикалық негізі ретінде қолданылады.

2.1.1 Құқықтық ақпараттық ресурстар және цифрлық инфрақұрылым

Заңнамалық мәтіндер өз табиғаты бойынша күрделі және ақпараттық болып табылады. Құқықтық акт - бұл республикалық референдумда немесе уәкілетті орган қабылдаған құқық нормаларын қамтитын белгіленген нысандағы жазбаша ресми құжат немесе жеке билік құқықтық бұйрықты қамтитын заңмен белгіленген нысандағы шешім [44]. 2.1-суретте байқауға болатындай, Қазақстан Республикасының «Құқықтық актілер туралы» Заңына сәйкес, құқықтық актілер нормативтік және нормативтік емес болып бөлінеді. Нормативтік құқықтық актілер негізгі және туынды болып бөлінеді.



Сурет 2.1 – құқықтық актілердің жіктелуі

«Заң нысаны» бөлімінен жиналған деректер жиынтығы [44] Конституция, Кодекстер, Жарлықтар, Бұйрықтар, Декларациялар және т.б. сияқты әртүрлі актілер туралы ақпаратты қамтитын 9575 заңнамалық мәтіннен тұрады.

Қазақстанда Қазақстан Республикасы Әділет министрлігі қабылдаған электрондық форматтағы барлық құжаттар «Әділет» деп аталатын Қазақстан Республикасының Нормативтік құқықтық актілерінің ақпараттық-құқықтық жүйесінде жинақталған. Онда 1947 жыл мен 2026 жылдар аралығында жиналған көптеген құқықтық құжаттар бар. Бұл құжаттар ішінара үш тілге аударылған. Осы құқықтық құжаттардың ең көп саны, атап айтқанда 15 359 құжат, 2023 жылы жүктелген. Құжаттардың нормативтік және нормативтік емес [32] болып бөлінуін ескере отырып, біз барлық құжаттардың кейбір бөліктерін көре аламыз.

Кесте 2.1 - Құжаттар жинағының динамикасы (2026 жылғы наурыз)

Көрсеткіш	2026 жылғы 6 наурыз	2026 жылғы 10 наурыз	2026 жылғы 14 наурыз	2026 жылғы 20 наурыз	2026 жылғы 26 наурыз
Барлық құжаттар	447 131	447 385	447 648	448 149	448 599
Қазақ тілінде	222 260	222 392	222 535	222 804	223 017
Орыс тілінде	221 716	221 838	221 958	222 190	222 427
Ағылшын тілінде	2 202	2 202	2 202	2 202	2 202

Қазақстан Республикасының заңнамасы құқықтық қатынастардың кең ауқымын қамтиды және заңнамалық құжаттардың саны күн санап артып келеді; бұған мысал 2.1 кестесінде келтірілген.

Кесте 2.2 - Қазақстан Республикасының заңнамалық тармақтары

Акт нысаны	**Акт қабылдаған орган**	**Құқықтық қатынастар саласы**	**Қабылданған жылы**
Конституция (1)	133 000 000 000 (10)	Ауыл шаруашылығы (60)	2024 (16)
Конституциялық заң (48)	134 000 000 000 (1)	Азаматтық құқық (78)	2023 (10)
Кодекс (31)	135 000 000 000 (17)	Тәуелсіз Мемлекеттер Достастығы (4)	2022 (8)
Заң (353)	144 000 000 000 (1)	Байланыс (23)	2021 (177)
Бұйрық (52)	Халықаралық органдар мен ұйымдар (82)	Жануарлар дүниесін қорғау және пайдалану (16)	2020 (330)
Жарлық (631)	ҚР республикалық бюджеттің атқарылуын бақылау жөніндегі есеп комитеті (5)	Орман қорын қорғау және пайдалану (8)	2019 (290)
Өкім (963)	ҚР Мемлекеттік қызмет істері жөніндегі агенттігі (1)	Жерді қорғау және пайдалану (18)	2018 (275)
ҚНТ (35)	ҚР Орталық сайлау комиссиясы (ОСК)	Суды қорғау және пайдалану (8)	2017 (128)
ҚНТ (1)	-	Конституциялық құрылым және мемлекеттік басқару негіздері (671)	2016 (203)
Келісім (17)	-	Құрылыс және сәулет (22)	2015 (94)
Өзгерістер (13)	-	ЫНтымақтастық (4)	2014 (49)

Қазақстан Республикасы заңнамалық актілерінің түрі, бекітуші органы, құқықтық саласы және қабылданған жылы бойынша бөлінуін түсіну үшін 2.2 кестесі берілген. Бұл ақпарат ұлттық заңнаманың күрделілігі мен тақырыптық әртүрлілігін көрсетеді, бұл QA жүйелерін құрудың маңызды негізі болып табылады.

Қазақстан Республикасының заңнамасы саласындағы QA жүйелерін зерттеуден практикалық нәтижелер алу үшін мемлекеттік тілдегі деректер жиынтығы дайындалды. QA жүйесін құру бірнеше негізгі кезеңдерді қамтиды:

деректерді дайындау, модельді таңдау, қосымша оқыту және сапаны бағалау. Бұл процесс модельді нақты қажеттіліктерге сәйкес бейімдеуге мүмкіндік береді, осылайша жауаптардың дәлдігі мен өзектілігін арттырады. Деректер ашық заңнамалық көздерден жиналды: «Әділет» (https://adilet.zan.kz/kaz/), «Zqai» (https://www.zqai.kz/ru/questions?, 2026 жылғы 25 сәуірде қол жеткізілді) және «Gov» (https://www.gov.kz/, 2026 жылғы 25 сәуірде қол жеткізілді). Бұл электрондық порталдардың құрылымы әртүрлі, бірақ олар құзыретті, заңды түрде ресімделген білім базасын білдіреді.

2.1.2 Заңнамалық мәтіндер корпусын және сұрақ-жауап деректер жиынтығын қалыптастыру

Бұл бөлімде Қазақстан Республикасының заңнамасы саласындағы сұрақ-жауап жүйесіне арналған деректер жиынтығын дайындау процесі ұсынылған. Заңдарды, құқықтар мен міндеттерді білу кез келген елдің әрбір азаматы үшін, оның әлеуметтік мәртебесіне қарамастан, өз құқықтарын түсінуі үшін қажет. Қазақстан Республикасының мемлекеттік тілі – қазақ тілі. Қазақ тілі оқытуға қажетті деректердің жеткіліксіз көлемімен сипатталатын ресурстары шектеулі тілдер тобына жатады.

Бұл мәселені шешу үшін мәтіндік деректердің үлкен көлемі қажет. Осы зерттеу аясында заңнамалық актілер мәтіндерінің корпустары мен сұрақ-жауаптары бар мәтіндер қажет болды. Мәтіндік деректер Қазақстан Республикасы Әділет министрлігінің «Қазақстан Республикасы нормативтік құқықтық актілерінің ақпараттық-құқықтық жүйесі» [32] және «Қазақстан Республикасының заңнама және құқықтық ақпарат институты» [34] веб-сайттарынан веб-скрейпинг әдісімен жиналды.

Инструменттер ретінде Content Downloader және Selenium WebDriver қолданылды. Жиналған мәліметтер тақырыптарды, күндерді, мәтіндерді, сондай-ақ сұрақ-жауап блоктарын қамтиды. Бұл жүйені енгізу Қазақстан Республикасының заңнамасы саласындағы маңызды қадам болып, миллиондаған азаматтарға көмектеседі.

Қазақстан Республикасының нормативтік құқықтық актілерінің ақпараттық-құқықтық жүйесі «Әділет» веб-іздеу жүйесі және нормативтік құқықтық актілерінің толық электронды жиынтығы болып табылады. Егер біз бұл жүйенің негізгі артықшылықтарын тізіп өтсек, ең бастысы - ол тегін. Сондай-ақ, олардың дерекқорының мазмұны ресми дереккөздерден алынған деректерден тұрады. Жаңадан қабылданған заңдар, мерзімі өткен заңдар немесе қолданыстағы заңдардағы өзгерістер олардың дерекқорында көрсетіледі; яғни жүйе үнемі жаңартылып отырады, бұл да маңызды мәселе. Сондай-ақ, барлық актілердің мәтіндері қазақ және орыс тілдерінде ұсынылған. Веб-сайтта түсінікті және пайдаланушыға ыңғайлы интерфейс бар. Пайдаланушылар іздеу терезесіндегі негізгі беттен қажетті құжаттарды, мысалы, құжаттың атауын енгізу арқылы іздей бастай алады. Веб-сайтта 2024 жылдың мамыр айындағы деректерге сәйкес, дерекқорда қазақ, орыс және ағылшын тілдерінде 400 мыңға жуық құжат бар. Бұл жүйе пайдаланушыға көптеген әртүрлі функцияларды

ұсынады, мысалы, құқықтық актілердің мәтіндерін қарау, олардың өзгерістер тарихын және алдыңғы нұсқаларын қарау, интерфейс тілін және құжаттарды ауыстыру мүмкіндігі.

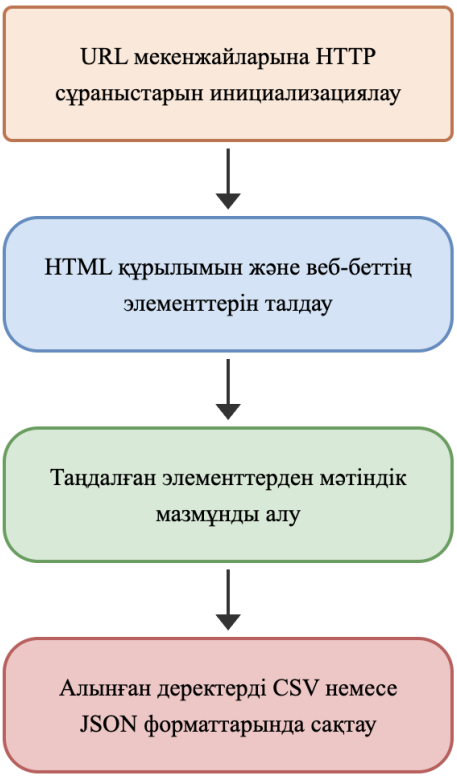
Қазақстан Республикасы Әділет министрлігінің «ҚР Заңнама және құқықтық ақпарат институты» 25 жыл бойы мемлекеттің ақпараттық-құқықтық саясатын жүргізуші және Қазақстандағы алғашқы «Заң» дерекқорын әзірлеуші болып келеді [32]. Дерекқордағы деректер тек ресми дереккөздерден ғана жасалады, бұл қолжетімді материалдардың сенімділігін қамтамасыз етеді. Ақпарат күн сайын жаңартылып отырады.

2.1.3 Мәтіндерді жинау, алдын ала өңдеу және аннотациялау әдістері

Веб-скрапинг процесі, әдетте, веб-сайттардан деректерді алудың маңызды процесі болып табылады, оған келесі қадамдар кіреді: HTTP сұрауларын мақсатты URL мекенжайына инициализациялау, веб-сайт құрылымын талдау (талдау қажет HTML құрылымдары мен нақты деректер элементтерін анықтау), таңдалған элементтерден мәтіндік мазмұнды алу және алынған деректерді CSV немесе JSON сияқты қолайлы форматта сақтау.

Python парсерімен деректерді скрапинг [45] веб-сайттардан ақпарат алуды қамтиды. Тікелей HTTP сұрауы және Selenium WebDriver сияқты құралдарды веб-беттермен өзара әрекеттесу және олардан деректерді автоматты түрде алу үшін пайдалануға болады. Тікелей сұраулар деректерді алудың жылдам процесінде тиімді болғанымен, Selenium WebDriver күрделірек механизмді іске асыруға мүмкіндік береді. Тиісті драйверді пайдаланып күрделі веб-мазмұнды алуда өте тиімді. Веб-драйвер [46] веб-беттің мазмұнын шығарған кезде, BeautifulSoup Python кітапханасы веб-беттерді терең талдау үшін талдау ағашын жасайды. Бұл жұмыста Қазақстан Республикасының сапалы заңнамалық деректерін ([32] және [34]) қамтитын екі веб-сайт талданды. Бірінші веб-сайтта көптеген сілтемелері бар барлық ресми заң құжаттары мен жарлықтар бар. Бастапқыда талданатын сілтемелер тізімін толтыру қажет болды. Бұл процедураны жеделдету үшін, барынша қамтуды қамтамасыз ету үшін өте пайдалы бағдарламалық жасақтама енгізілді. Бұл бағдарлама әртүрлі веб-сайттардан сілтемелер мен мазмұнды жинауға арналған Content Downloader бағдарламасы деп аталады. Процесс веб-сайттың URL мекенжайын, деректерді жинау жылдамдығын басқару үшін ағындар санын және тиісті сілтемелерді алу үшін сүзгілеу параметрлерін орнатуды талап етті. URL параметрлері сүзгілеу терезесінің мекенжайларында анықталды, іздеу қазақ тіліндегі құжаттарға сілтемелер үшін екенін көрсетті – kaz/docs/. Құжаттарды қамтымайтын сілтемелер search/, docs/search, index/docs сияқты үлгіден алынып тасталды.

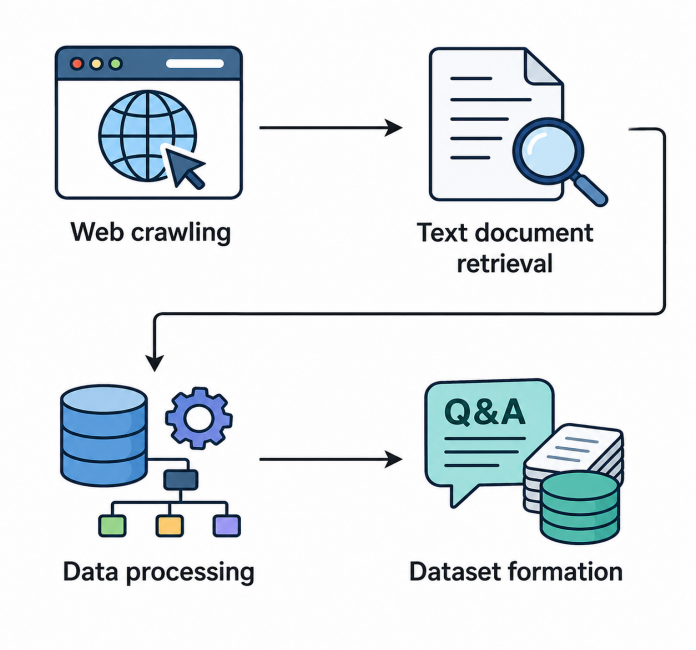
Веб-талдау арқылы деректер жиынтығын қалыптастыру қадамдары 2.2 суретте көрсетілген.



2-сурет. *Деректер жиынтығын қалыптастыру кезеңдері*

Сурет 2.2 – Деректер жиынтығын қалыстастыру кезеңдері

URL мекенжайлары жиналған кезде, веб-беттердің мазмұнын алу үшін Selenium WebDriver пайдаланылды, ал BeautifulSoup статикалық HTML мазмұнын талдауда ерекше болды [47]. Алдымен, қайталанатын және бос сілтемелер жойылды. Содан кейін, '#' таңбасын қамтитын айна сілтемелері де сүзгіден өткізілді. Сілтемелер тізімі жасалған кезде, скрапинг қадамы орындалды. Selenium WebDriver веб-беттермен өзара әрекеттесу үшін, URL мекенжайлары бойынша итерациялау үшін іске қосылды. HTML мазмұны BeautifulSoup көмегімен алынып, талданды. Заңнамалық мәтіндердің негізгі мазмұны 'div' элементіне 'con-'container_alpha slogan' класымен енгізілді. Егер контейнер элементі табылса, скрипт тақырып ретінде бірінші 'h1' элементін және күн ретінде бірінші 'p' элементін шығарып алды. Негізгі мәтін мен 'article' элементін қамтитын 'div' элементі табылды, ал 'article' элементіндегі барлық 'a' тегтері мәтінде тазартылды. Осы үш маңызды элементтің барлығы дайындалған кезде, тақырып, күн, тазартылған мәтін және бастапқы URL мекенжайы бар жаңа сөздік жасалды. Содан кейін WebDriver әрекеттері аяқталды.



Сурет 2.3 – Датасет құрудың этаптары

Екінші веб-сайттың әр бетінде сұрақ-жауап элементі бар [48] қарапайым құрылымы бар. Бұл веб-сайтта әр беттің URL түрі қарапайым болғандықтан, URL индексін өзгерту сілтемелер тізімін құрады. Содан кейін WebDriver инициализацияланады және URL мекенжайларының тізімін қайталайды. Ол «div» элементінен «views-field views-field-title» класы бар «санды» алып тастайды және айналасындағы бос орындарды жояды. Келесі қадамдарда «date-display-single» класының «span» элементінен «date-display-single» класынан «spain» элементінен «сұрақ» және «answer» деректері сәйкесінше «views-field views-field-field-question» және «views-field views-field-field-answer» кластарының «div» элементтерінен алынды. Барлық алынған деректер сөздікке қосылды. WebDriver-дің әрекеттері де сол кезде аяқталды.

Алынған нәтижелерді қорытындылай келе, жиналған барлық деректердің метадеректерін сипаттау маңызды болып табылады. Жиналған әрбір құжат белгілі бір метадеректермен байланыстырылған.

	header	date	text	source
0	2020 – 2022 жылдарға арналған Ауыл ауылдық о...	2020 - 23 қараша	"Қазақстан Республикасының Бюджет кодексі" Қаз...	https://adilet.zan.kz/kaz/docs/V20JF007809
1	Асыл тұқымды мал шаруашылығын дамытуға, мал ша...	2017 - 14 қараша	Есептеу. Құсі жөніндегі - Атырау облысы әкімдігі...	https://adilet.zan.kz/kaz/docs/V17T0003589

Сурет 2.4 – Құжаттар датасетінің көрінісі

Сурет 2.4-те веб-сайты деректері үшін міндетті метадеректер көрсетілген:
- HEADER - заңнамалық актінің атауын қамтиды.
- DATE - заңнаманың қабылданған күнін қамтиды.
- TEXT - заң немесе актінің толық мәтінін қамтиды.

- SOURCE - деректер алынған бастапқы веб-сайтқа сілтемені қамтиды.

NUMBER	DATE	QUESTION	ANSWER
1	2024-05-07	Жеке сот орындаушысының әрекетіне шағымды қалай беруге болады?	Жеке сот орындаушысының әрекетіне (немесе әрекетсіздігіне) Қазақстан Республикасы Әділет министрлігіне немесе сотқа шағым беруге болады.
2	2024-05-07	Алимент мөлшерін өзгерту үшін қайда жүгіну керек?	Алимент мөлшерін өзгерту туралы талап арызды сотқа (тұрғылықты жеріңіз бойынша аудандық сотқа) беру қажет.
3	2024-05-06	Жылжымайтын мүлікті сатып алу-сату шартына нотариалдық куәландыру қажет пе?	Жылжымайтын мүлікті сатып алу-сату шарты міндетті түрде нотариалды түрде куәландырылуға жатады.
4	2024-05-06	Қандай жағдайда жеке тұлға банкрот деп танылуы мүмкін?	ҚР «Төлем қабілеттілікті қалпына келтіру және банкроттық туралы» Заңына сәйкес, борышкер өзінің төлем қабілетсіздігін сот тәртібімен банкрот деп танылуы мүмкін.
5	2024-05-05	Жұмыс беруші өтемақы төлемей жұмыстан шығаруға құқылы ма?	Жұмыс беруші қызметкерді тек заңда көрсетілген негіздер бойынша ғана өтемақы төлемей жұмыстан шығара алады (мысалы, тәртіптік теріс қылық).

Сурет 2.5 – Сұрақ жауап датасет көрінісі

Сурет 2.5-те веб-сайты деректері үшін міндетті метадеректер көрсетілген:
- NUMBER - қойылған сұрақтардың бірегей нөмірін қамтиды.
- DATE - сұрақтың сайтқа орналастырылған күнін қамтиды.
- QUESTION - пайдаланушы қойған сұрақтың мәтінін тікелей қамтиды.
- ANSWER - қойылған сұраққа берілген жауаптың мәтінін қамтиды.

Екі веб-сайттан деректерді жинау процесі екі құрылымдық деректер жиынтығын қалыптастырды.

Бірінші веб-сайт тақырыпты, күнді, мәтінді және дереккөзді қамтитын жазбалар сериясын ұсынды. Әрбір URL-дің HTML элементтерін талдау нәтижесінде деректер жиынтығы қалыптасты. Бірінші деректер жиынтығы туралы ақпарат 2.3 кестеде берілген.

Кесте 2.3 - Заңнамалық мәтіндер деректер жиынтығы

Веб-сайт	**Сөйлемдер саны**	**Құжаттар саны**	**Көлемі**
adilet.zan.kz	5066360	115150	2.6Gb
zan.gov.kz	5147608	119850	2.7Gb
Жалпы саны:	10213968	235000	5.3Gb

«Әділет» порталында заңнамалық құжаттар реттелген мәтін түрінде берілген. Сондықтан сұрақтар мен жауаптар олардан қолмен құрастырылды.

«Zqai» және «Gov» порталдарында деректер сұрақ-жауап блоктары түрінде құрылымдалған, бұл жүйені оқыту үшін құжаттарды қалыптастыруды жеңілдетті. Сонымен қатар, заңнама саласындағы әртүрлі тақырыптар бойынша түрлі сұрақтар мен жауаптардан тұратын синтетикалық деректер жиынтығы қолмен жасалды. Бұл синтетикалық деректер жиынтығының ерекшелігі - оған салыстырмалы түрде қысқа сұрақтар мен жауаптар кіреді, бұл үлгілердің көп санын қажет етеді.

«Әділет» және «Zqai» жүйелері құқықтық ережелерді ұсынады, бірақ пайдаланушы сұрауларының санаттық талдауын жарияламайды. Дегенмен, құжаттардың көлемі мен тақырыптардың танымалдылығы туралы мәліметтерге сүйене отырып, біздің синтетикалық корпусымыздың (Қазақстан Республикасы заңнамасының 14 тақырыбы бойынша) тақырыпқа қаншалықты сәйкес келетінін және ұлттық құқықтық ақпараттық жүйелерде өзекті екенін салыстыруға және тексеруге болады. Қазақстан Республикасының заңнамасы бойынша құқықтық сұрақтар мен жауаптардың синтетикалық корпусын қалыптастыру барысында оның тақырыптық жағынан толық қамтылуын қамтамасыз етуге әрекет жасалды.

Алынған әрбір жазба тұтастықты қамтамасыз ету мақсатында тазартылып, пішімделді. Жазылған сұрақтар мен жауаптардың жоғары сапасына қарамастан, олардың бір бөлігі орыс тілінде жазылған болатын. Сондықтан оларды қазақ тіліне аудару үшін Google аударма жүйесін пайдалану қажет болды. Зерттеу аясында жиналған деректердің өзектілігін анықтау мақсатында олардың тақырыптық мазмұнына талдау жүргізілді. adilet.zan.kz порталының статистикасы пайдаланушылар тарапынан азаматтық, еңбек, қылмыстық, әкімшілік, қаржылық және жер құқығы салаларына, сондай-ақ тиісті кодекстерге деген сұраныстың жоғары екенін көрсетеді. Осыған байланысты, әзірленген синтетикалық деректер жиынтығы аталған базалық салаларды, соның ішінде мұрагерлік, корпоративтік құқық, келісімшарттар мен жауапкершілік сияқты кіші санаттарды толық қамтиды.

Кесте 2.4 - Сұрақ-жауап деректер жиынтығы

Вебсайт	**Сұрақ-жауап жұптарының саны**	**Сөйлемдер саны**	**Көлемі**
Adilet	1012	59402	912 Kb
Zqai	602	219125	756 Kb
Gov	738	127565	1.89 Mb
Synthetic	3764	67926	328 Kb
Manual	1884	57134	939 Kb
Жалпы саны	8000	531152	5.125 Mb

Сонымен қатар, Zqai ғылыми материалдарына негізделген деректер конституциялық, халықаралық, ақпараттық-сандық және қоршаған ортаны қорғау құқығы сияқты тереңірек аналитикалық тақырыптарды қамти отырып,

корпустың білім беру құндылығын кеңейтті. Салыстырмалы талдау нәтижесінде қазіргі синтетикалық корпус заңнаманың негізгі бағыттарын қамтығанымен, болашақта көші-қон, зейнетақы жүйесі, зияткерлік меншік және сыбайлас жемқорлыққа қарсы саясат сияқты арнайы тақырыптармен толықтыру қажеттілігі анықталды. Екінші деректер жиынтығы туралы ақпарат 2.4 кестеде берілген.

Нәтижесінде, екі веб-сайттан жиналған деректер жиынтықтары мәтіндік ақпаратты құрылымдық форматта қамтып, сұрақ-жауап жүйесін құруға арналған одан әрі дайындыққа толықтай дайын болды.

2.2 Үлкен тілдік модельдерді зерттеу және бейімдеу әдістері

Бұл бөлімде LLM модельдердің архитектуралық ерекшеліктері, олардың құқықтық мәтіндерді өңдеудегі мүмкіндіктері және заңнамалық доменге бейімдеу әдістері қарастырылады. Осы жұмыста заңнамалық сұрақ-жауап жүйесін құру үшін жеті LLM модель пайдаланылды: GPT, Gemini, KazLLM, LLaMA, Phi, Qwen және Mistral. Аталған модельдердің әрқайсысы өзіндік конфигурациялық сипаттамаларымен және үйрету механизмдерімен ерекшеленеді.

Қазақ тілінің агглютинативті табиғаты универсалды модельдерді заңнамалық мәтіндерге бейімдеу кезінде арнайы токенизация, морфологиялық нормализация және домендік дайындық стратегияларын талап етеді. Төменде осы бөлімде қарастырылатын модельдердің негізгі архитектуралық ерекшеліктері, таңдалған конфигурациялар және бейімдеу тәсілдері сипатталады.

2.2.1 Базалық LLM модельдерінің салыстырмалы талдауы

Зерттеу шеңберінде ҚР заңнамасы саласындағы сұрақ-жауап жүйесін құру үшін үлкен тіл моделі таңдалды. Модельдерді таңдаудың негізгі өлшемдері ретінде мыналар анықталды: мультитілдік деректерде алдын ала оқыту деңгейі; ашық кодтың немесе API қолжетімділігінің болуы; архитектуралық тиімділік; сондай-ақ аглютинативті тілдермен жұмыс тәжірибесі.

GPT архитектурасы трансформер декодер құрылымына негізделген және маскіленген көп басты өздік-назар механизмі (masked multi-head self-attention), кері қосылым желісі (FFN), қабат нормализациясы (Layer Normalization) және қалдықты байланыстар (residual connections) секілді бірдей қабаттардың стекінен тұрады [49]. Маскіленген өздік-назар механизмі модельдің әрбір токенді тек алдыңғы токендер негізінде болжауын қамтамасыз етіп, авторегрессиялық генерацияны мүмкін етеді. Соңғы шығыс қабаты жасырын күйлерді сөздік векторға проекциялап, келесі токен softmax функциясы арқылы болжанады.

GPT модельдері ішінде ең жетілдірілгендері — GPT-3.5-turbo, GPT-4 және GPT-4o mini. GPT-3.5-Turbo ұзақ контексттік сұрауларды қолдайды, бұл оған кеңейтілген пайдаланушы диалогтары мен құжат деректерін өңдеуге мүмкіндік береді. Нұсқалардың дамуымен GPT-3.5-Turbo үлгісі o4-mini-мен

алмастырылды, өйткені бұл модель арзанырақ, функционалдырақ және жылдамырақ болып табылады. Ол мәтін мен кескін деректерін қабылдайды, бұл оны шағын деректер жиынтықтарын дайындауға тиімді етеді. GPT-4o — кең ауқымды мәселелерді шешуге арналған әмбебап бас модель, алайда ол қымбатырақ. Заңнамалық саладағы сұрақ-жауап жүйесін қайта үйрету кезінде GPT-4 мен салыстырғанда жылдам, жоғары функционалды және экономикалық тиімді болғандықтан GPT-4o mini моделі таңдалды. Қайта үйрету процесі деректерді алдын ала дайындауды, деректерді OpenAI сервисіне жүктеуді және үйрету процесін бастауды қамтыды.

Gemini - Google DeepMind әзірлеген multimodal LLM модельдер отбасы. Модель мәтін, сурет, аудио және бейне сияқты бірнеше ақпарат түрлерін бір уақытта өңдеуге қабілетті және transformer архитектурасына негізделген. Gemini табиғи тілді өңдеу, мәтін генерациясы, reasoning және диалогтық тапсырмалар үшін оңтайландырылған. Алғашқы нұсқалары Gemini Ultra, Gemini Pro және Gemini Nano конфигурацияларында ұсынылды [50].

Gemini архитектурасы үлкен контексті өңдеуге, multimodal input қолдауға және күрделі reasoning міндеттерін орындауға бағытталған. Модельде transformer-based attention механизмдері, масштабталатын encoder-decoder компоненттері және тиімді inference тәсілдері қолданылады. Сонымен қатар Gemini Google экожүйесімен және cloud инфрақұрылымымен интеграцияланып, API және enterprise шешімдерінде кеңінен пайдаланылуда.

Заңнамалық сұрақ-жауап жүйесін әзірлеу барысында Gemini моделі пайдаланушы сұрауларын семантикалық тұрғыдан талдау, құқықтық контексті сақтау және логикалық жауап генерациялау мүмкіндіктеріне байланысты зерттеу объектілерінің бірі ретінде қарастырылды. Модельдің multimodal және long-context мүмкіндіктері үлкен көлемдегі құқықтық мәтіндермен жұмыс істеу үшін маңызды артықшылық болып табылады.

KazLLM — қазақ тіліне бағытталған мамандандырылған тілдік модель. Ол мәтінді түсіну және тапсырмаларды генерациялауға арналған. KazLLM BERT, RoBERTa және LLaMA отбасы архитектуралары, сондай-ақ басқа трансформерлер негізінде жасалған. Модель RoPE, RMSNorm және SwiGLU активациялары сияқты жетілдірілген архитектуралық мүмкіндіктерді қолданады [51]. Токенизатор 128 000 сөздіктен тұрады.

KazLLM 150 млрд токеннен тұратын ауқымды мультилингвалды корпуста үйретілген, оған сүзгіден өткен веб-деректер, Уикипедия және жаңалықтар мен білім беру мазмұны сияқты әртүрлі қазақ тіліне тән дереккөздер кіреді. Модель контекст ұзындығы 8192 токенді қолдайды, Grouped-Query Attention және FlashAttention v2 механизмдерін интеграциялайды. KazLLM тілдің морфологиясын, синтаксисін және семантикасын ескеріп, тереңдетілген қазақ тілі білімін қажет ететін міндеттерде мультилингвалды аналогтарды басып озады. Тіл бейімделуіне байланысты KazLLM күрделі грамматикалық құрылымдарды өңдеумен жақсы жұмыс жасайды, бұл Қазақстандағы заңнамалық салада жүйені құру үшін ерекше маңызды. QA жүйесін құру үшін 8 млрд параметрлі модель таңдалды.

LLaMA — Meta әзірлеген және мәтін генерациясы, мазмұндама жасау және сұрақ-жауап жүйелері сияқты табиғи тілді өңдеудің кең ауқымды міндеттеріне арналған тілдік модельдер сериясы [52]. LLaMA архитектурасы GPT-ге ұқсас тек декодерден тұратын трансформер модельдеріне жатады. Осы архитектурада:

- әрбір трансформер ішкі қабатының кірісі RMSNorm функциясымен нормализацияланады;
- SwiGLU активация функциясы ReLU нелинейлігінің орнына қолданылады;
- абсолютті позициялық ендірулер RoPE ендірулерімен алмастырылды.

Ашық архитектурасы және Hugging Face Transformers мен DeepSpeed кітапханаларымен жоғары үйлесімділігі арқасында LLaMA заманауи ML конвейерлеріне жақсы интеграцияланады және LoRA, QLoRA және басқа да тиімді техникаларды пайдаланып қайта үйретуді қолдайды. Заңнамалық деректерде үйрету барысында жоғары дәлдігімен, жетілдірілген интеграциясымен, масштабтылығымен және қайта үйрету мүмкіндігімен ерекшеленетін LLaMA 3 нұсқасы қолданылды. Тәжірибелерде QA жүйесін үйрету үшін 3 млрд параметрлі LLaMA пайдаланылды.

Phi — Microsoft Research табиғи тілді өңдеу саласындағы мәселелерді тиімді шешу үшін жасаған тілдік модель [53]. Phi моделі контекст терезесі бойынша толық назар аударуды (full attention) пайдаланады. Ол көп агентті нұсқау беру, нұсқауды кері айналдыру және өзін-өзі түзету сияқты техникаларды қамтитын жоғары сапалы синтетикалық деректерге баса мән беретін оқу жоспарымен үйретілген. Үйретуден кейінгі кезең Supervised Fine-Tuning (SFT) және Direct Preference Optimization (DPO) екі раундын қамтиды.

Phi-1, Phi-1.5 және Phi-2 нұсқалары код генерациясына, логикалық ойлауды жақсартуға және тілді түсінуді кеңейтуге бағытталды. Phi-3 модельдері жоғары білім тығыздығын қамтамасыз ететін синтетикалық және веб деректердің кең ауқымын қосқанда 3,3–4,8 трлн токенде үйретілді. Қазақстан Республикасының заңнамалық саласындағы сұрақ-жауап жүйесін жасау үшін 14 млрд параметрлі модель пайдаланылды.

Qwen - Alibaba Cloud 2023 жылы жасаған тілдік модель. Содан бері бірнеше буын модельдер шығарылды: Qwen 2, Qwen 2.5 және Qwen 3, олардың әрқайсысы өнімділік, көптілді қолдау және ойлау қабілеті жағынан жетілдірілді [54]. Qwen архитектурасы ашық код үздік LLM ретінде кеңінен қолданылатын LLaMA моделіне негізделген. Архитектуралық өзгерістерге мыналар кіреді: жоғары дәлдік үшін FP32 нақтылығын пайдаланатын RoPE ендірулері;

- дәстүрлі қабат нормализациясын алмастырған алдын ала нормализация (RMSNorm);
- Swish және Gated Linear Unit (GLU) негізіндегі SwiGLU активация функциясы.

Qwen 2.5 2025 жылдың қаңтарында таныстырылды және 18 трлн токенде үйретілді, бұл оның бағдарламалау мен математика бойынша білімін айтарлықтай жақсартты. 2025 жылдың сәуірінде шыққан Qwen 3 отбасының соңғы моделі болып табылады және тығыз (dense) пен сирек (sparse)

архитектураларды қамтиды. 36 трлн токенде үйретілген. Заңнамалық сұрақ-жауап жүйесін жасау үшін 7 млрд параметрлі Qwen 2.5 моделі пайдаланылды.

Mistral - Mistral AI 2023 жылы жасаған LLM отбасы. Осы модельдерді жасаудың негізгі мақсаты - OpenAI нәтижелерімен теңесетін немесе оларды басып озатын жоғары тиімді және қолжетімді ашық бастапқы код модельдерін жасау. Модельдер Grouped-Query Attention (GQA) және Sliding Window Attention (SWA) механизмдерін пайдаланатын трансформер архитектурасына негізделген [55].

GQA кілт-мән басыларының санын (8) сұрау басыларының санын (32) сақтай отырып азайтады. Бұл дизайн декодтау кезінде жадты пайдалануды азайтып, ерекше ауқымды орналастыруларда инференс жылдамдығын арттырады. SWA-да әрбір токен тек алдыңғы токендердің белгіленген терезесіне (W = 4096) назар аударады, бұл k қабат үшін k × W токенге дейін ұзақ диапазонды контекстті тиімді сақтауға мүмкіндік береді. Rolling Buffer Cache механизмі ең ескі жазбаларды дөңгелекті тәсілмен қайта жазу арқылы жады талаптарын сапаны төмендетпей айтарлықтай қысқартады. Mistral 7B LLaMA-мен салыстырылатын өнімділікпен (LLaMA 2 13B және LLaMA 1 34B-дан да асып түсетін) нәтижелерді шағынырақ өлшем мен жоғары өткізу қабілетімен ұштастырады. Зерттеуде 7 млрд параметрлі Mistral моделі пайдаланылды.

Кесте 2.5 - LLM модельдерінің архитектуралық сипаттамалары мен конфигурациялары

Модель	**Параметр саны**	**Архитектура**	**Токенизатор**	**Негізгі ерекшелік**
GPT-4o mini	Жасырын	Decoder-only Transformer	Byte-level BPE	Мәтін+кескін, OpenAI fine-tuning
Gemini	9 млрд	Transformer (RoPE, GeGLU, RMSNorm)	SentencePiece	Multi-Query Attention, LoRA/SFT
KazLLM	8 млрд	BERT+LLaMA (RoPE, SwiGLU, GQA)	SentencePiece 128k	Қазақ тілі, 150 млрд токен корпус
LLaMA 3	3 млрд	Decoder-only (RoPE, SwiGLU, RMSNorm)	SentencePiece	LoRA/QLoRA, ашық архитектура
Phi-3	14 млрд	Transformer (Full Attention, SFT+DPO)	SentencePiece	Синтетикалық деректер, логикалық ойлау
Qwen 2.5	7 млрд	LLaMA-based (RoPE FP32, SwiGLU)	Tiktoken/ SentencePiece	18 трлн токен, көп тілді
Mistral 7B	7 млрд	Transformer (GQA, SWA, Rolling Cache)	SentencePiece	Жылдам инференс, тиімді жады

Baseline бағалаудың мақсаты - fine-tuning жүргізілмес бұрын жеті модельдің де ҚР заңнамалық мәтіндеріне қатысты бастапқы қабілеттерін анықтау болды. Барлық модельдер бірдей тест жиынына нөлдік реттік (zero-shot) режимде бағаланды. 2.5-кестеде зерттеуде қолданылған модельдердің архитектуралық сипаттамалары берілген.

Кесте деректерінен көрініп тұрғандай, таңдалған модельдер архитектуралық тәсілдер жағынан айтарлықтай ерекшеленеді. GPT-4o mini мен Gemini диалогтық тапсырмаларға арналған decoder-only архитектурасын пайдаланады. KazLLM қазақ тілінде мамандандырылған жалғыз модель болып табылады - оның токенизаторы 128 000 токеннен тұратын сөздік қорды қолдайды, бұл аглютинативті морфологияны өңдеу тиімділігін арттырады. LLaMA және Mistral жеңіл пайдалануға қолайлы ашық архитектуралар ұсынады, ал Phi мен Qwen Microsoft пен Alibaba компанияларының арнайы оңтайландыруларын қамтиды.

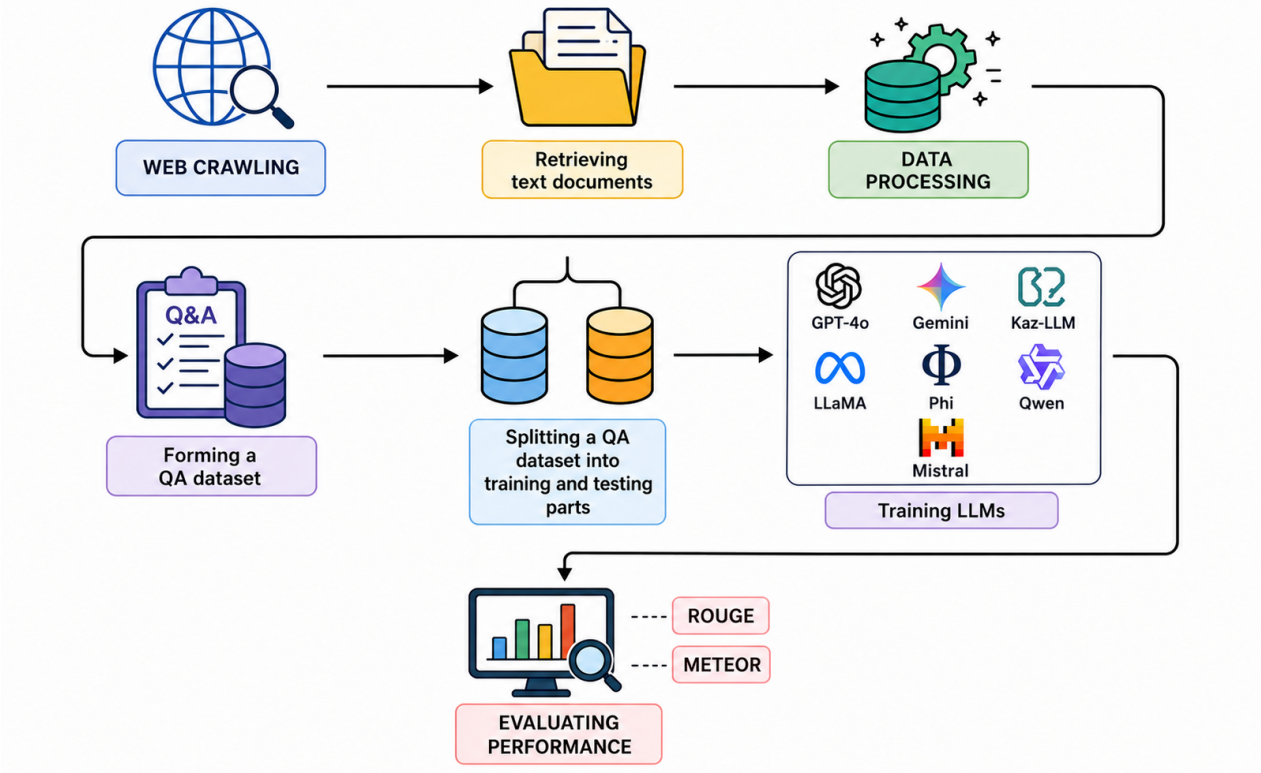
2.2.2 Fine-tuning негізінде модельдерді домендік бейімдеу

Казақ тілінің аглютинативті табиғаты токенизация тапсырмасын индоеуропалық тілдермен салыстырғанда едәуір күрделі етеді. Полислабтық сөз формалары мен көптеген жұрнақтардың болуы сөз формаларының жоғары вариативтілігіне алып келеді, бұл жиілік модельдері мен тіркелген сөздіктері бар трансформерлер үшін қиындық туғызады. GPT-4o mini байт деңгейіндегі BPE токенизаторын қолданады - ол тіл-бейтарап болғанымен, морфемалық шекараларды ескермейді. Ал Gemini, KazLLM, LLaMA, Phi және Mistral моделдерінде қолданылатын SentencePiece статистикалық BPE немесе unigram моделіне сүйенеді, алайда аглютинативті тілдерде сөз формаларын интерпретацияланбайтын фрагменттерге бөлу мәселесі туындауы мүмкін.

Осы ерекшеліктерді ескеру үшін алдын ала өңдеу кезеңінде мынадай шаралар қолданылды: мәтіндерді шуылдар мен артефактілерден тазалау; мәтіндерді бірыңғай емлеге келтіру; диалект формаларын нормализациялау. Бұл шаралар Gemini, LLaMA және Qwen сияқты жалпы мақсаттағы модельдерді қазақ грамматикасының ерекшеліктеріне бейімдеуге мүмкіндік берді.

Baseline бағалау нәтижелері негізінде жеті модельдің барлығы арнайы заңнамалық деректер жиынтықтарында fine-tuning жүргізілді. Бейімдеудің мақсаты - модельдерді ҚР заңнамасының терминологиясы, логикалық құрылымы мен мәтіндік стилін игеруге үйрету. Сурет 2.6-да көрсетілгендей, fine-tuning үдерісі бірнеше кезеңнен тұрады: деректерді дайындау, промпт форматын конфигурациялау, оқыту параметрлерін баптау және сапаны бағалау.

Барлық деректер жиынтықтары - Adilet, Zqai, Gov және синтетикалық - оқыту 80%, валидация 10% мен тест 10% бөліктеріне бөлінді. Zqai мен Gov жиынтықтары олардың жалпы құрылымдық ұқсастығын ескере отырып, оқыту кезінде біріктірілді, бұл топтың жалпы көлемін арттырды.



Сурет 2.6 – QA модельдерін Finetuning жасау қадамдары

Кванттандырылған жуық рангілі бейімдеу (QLoRA). Gemini, KazLLM, LLaMA, Phi, Qwen және Mistral модельдері Quantized Low-Rank Adaptation (QLoRA) тәсілімен 4-биттік кванттандыру (nf4 типі, қосарлы кванттандырумен) арқылы бейімделді. LoRA адаптер параметрлері: rank r = 16, lora_alpha = 16, dropout = 0. Мақсатты проекциялар: q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj, down_proj. Оқыту AdamW (8-bit) оңтайландырушысымен жүргізілді: оқыту жылдамдығы lr = 2×10^{-4}, партия өлшемі = 2, градиент жинақтауы = 4 қадам, жалпы қадам саны = 60. 2.6-кестеде зерттеуде қолданылған fine-tuning конфигурациялары мен промпт форматтары берілген.

Кесте 2.6 - Fine-tuning конфигурациялары мен промпт форматтары

Модель	**Бейімдеу әдісі**	**LoRA r**	**α**	**Кванттау**	**Промпт форматы**	**Ерекшеліктер**
GPT-4o mini	API (OpenAI)	-	-	-	Chat messages (user/assistant)	Ресми API форматы; zero/few-shot; промпт инженериясы
Gemini 9B	SFT + LoRA	r=16	α=16	nf4	<start_of_turn> user/model	CLM; SentencePiece + EOS; диалог стилі

Кесте 2.6 жалғасы

KazLLM 8B	SFT + DPO	r=16	α=16	nf4	### Instruction ### Input ### Response	Alpaca-стиль; next-token prediction
LLaMA 3B	QLoRA 4-bit	r=16	α=16	nf4	ChatML (system/user/assistant)	Meta токен форматы; CLM loss
Phi 14B	SFT + DPO	r=16	α=16	nf4	<|im_start|> system/user/asst	ChatML-lite; компактты; CLM + EOS
Qwen 2.5 7B	QLoRA 4-bit	r=16	α=16	nf4	### Instruction ### Response	LLaMA-негізді архитектура; RoPE FP32
Mistral 7B	CLM + QLoRA	r=16	α=16	nf4	<|im_start|> system/user/asst	GQA + SWA; Rolling Buffer Cache

GPT моделін оқыту барысында GPT-4o mini моделі OpenAI компаниясының ресми API интерфейсі арқылы бапталды және пайдаланушы мен ассистент арасындағы чаттық өзара әрекеттесу схемасына сәйкес хабарламалар форматы қолданылды. Әрбір сұрақ-жауап (QA) жұбы екі бірізді элементтен тұратын хабарлама құрылымы ретінде ұйымдастырылды: “user” рөліндегі пайдаланушы сұрағы және “assistant” рөліндегі модель жауабы. Жазба форматы келесідей түрде берілді:

```
QA.append({
“messages”: [
{“role”: “user”, “content”: question},
        {“role”: “assistant”, “content”: answer}]
})
```

Нәтижесінде алынған QA тізімі OpenAI API жүйесінің fine-tuning талаптарына сәйкес келетін JSON форматына түрлендірілді. Файлдың әрбір жолы модельдің бір толыққанды оқыту сессиясын сипаттайтын дербес құрылым ретінде ұсынылды.

```
with open(“QA_finetune_adilet.jsonl”, “rb”) as f:
response = client.files.create(
file=f,
purpose= “fine-tune”
)
fine_tune_response = client.fine_tuning.jobs.create(
training_file=file_id,
```

model="gpt-4o-mini-2024-07-18"
)

GPT-4o mini моделін fine-tuning процесінде баптау OpenAI компаниясының ресми API форматына сәйкес жүргізілді. Бұл тәсілде әрбір QA жұбы "user" рөліндегі пайдаланушы сұрағынан кейін "assistant" рөліндегі модель жауабынан тұратын құрылымдалған чаттық хабарламалар схемасы арқылы ұйымдастырылды. Мұндай формат API-мен үйлесімділікті қамтамасыз етіп, нақты диалогтық өзара әрекеттесуді модельдегенімен, fine-tuning тиімділігі prompt engineering стратегиясына, әсіресе пайдаланушы сұрақтарының қалай құрастырылатынына тікелей тәуелді болады. Әдепкі тәсіл ретінде zero-shot prompting қолданылды, мұнда prompt-тар қысқа әрі нақты сұрақтар түрінде берілді. Мысалы: "Қазақстанда апелляция беру тәртібі қандай?" немесе "Сот шешіміне қалай шағымдануға болады?". Мұндай құрылым модельдің қосымша контекстік нұсқаусыз қажетті жауап форматы мен стилін өздігінен анықтай алатынына негізделеді.

Модельдің жалпылау қабілеті мен жауап сапасын арттыру мақсатында альтернативті prompt стратегияларын қолдануға болады. Солардың бірі - few-shot prompting тәсілі, мұнда негізгі сұрақтың алдында үлгі мысал беріледі. Мысалы: "Азаматтық талап қалай беріледі? Талапкер тиісті сотқа өтініш беруі тиіс… Енді: Апелляциялық шағым қалай рәсімделеді?". Бұл құрылым модельге күтілетін жауап форматын алдын ала көрсетеді. Тағы бір тиімді тәсіл - chain-of-thought prompting, мұнда сұрақ модельді кезең-кезеңімен reasoning жасауға бағыттайды. Мысалы: "Апелляция беру үшін қандай қадамдарды орындау қажет? Әр қадамды түсіндіріп беріңіз". Бұл әдіс модельдің процесті логикалық түрде талдау қабілетін күшейтеді.

Сонымен қатар instructional prompting тәсілі арқылы жауап стилі мен аудиторияға бейімделуі басқарылуы мүмкін. Мысалы: "Апелляциялық процесті бірінші рет естіп тұрған адамға түсіндіретін заңгер ретінде жауап беріңіз". Болашақтағы fine-tuning итерацияларында пайдаланушылардың әртүрлі сұраныстарына модельді жақсырақ бейімдеу үшін бірнеше prompt стильдерінің комбинациясын қолдану ұсынылады. Бұған формалды, бейресми немесе шұғыл тондар; студент, азамат немесе заңгер сияқты әртүрлі аудиториялар; сондай-ақ анық емес немесе күрделі сұрақ құрылымдары жатады. Prompt engineering-дегі мұндай әртүрлілік модельдің тұрақтылығын арттырып қана қоймай, оны нақты құқықтық сұрақ-жауап жүйелерінде қолданудың тиімділігін де жоғарылатады.

Gemini моделі instruction-tuning негізінде fine-tuning процесінен өткізілді. Оқыту деректер жиынтығы құрылымдалған prompt–response жұптарынан тұрды. Әрбір өзара әрекеттесу модель архитектурасының диалогтық табиғатына сәйкес келуі үшін (<start_of_turn>user, <start_of_turn>model) форматындағы арнайы тегтермен белгіленді. Модель autoregressive text generation мақсаты негізінде көптеген мысалдар арқылы оқытылды, ал әрбір диалогтың аяқталуын белгілеу үшін EOS token қолданылды. Tokenizer ретінде Gemini моделінің бастапқы pre-training кезеңімен үйлесімді SentencePiece пайдаланылды. Fine-tuning процесі

фактілік дәлдікті, пәндік саламен байланысты контекстті (domain grounding) және қазақ тіліндегі жауаптардың тілдік еркіндігін жақсартуға бағытталды.

Бұл модель диалогтық ортада нақты құқықтық сілтемелерді, процедуралық түсіндірмелерді және пәндік салаға тән жауаптарды генерациялау қабілетін көрсетті. Модель жергілікті сервер/API арқылы іске қосылып, кідірісті азайту мақсатында quantized weights қолдану арқылы GPU acceleration үшін оңтайландырылды.

Конфигурация үлгісі келесідей түрде берілді:

“““<start_of_turn>user
{}\n{}<end_of_turn>
<start_of_turn>model
{}<end_of_turn>””” + EOS_TOKEN

Құрылымдық бірізділік сақталғанымен, prompt мәтіндерінің құрастырылуы мен prompt engineering тәсілдері модельдің әртүрлі типтегі сұраныстарға тиімді жауап беруінде маңызды рөл атқарады.

Мысалы, базалық prompt келесідей түрде құрастырылуы мүмкін:

<start_of_turn>user
Құқықтық көмек көрсету үшін қайда жүгінуге болады? (Where can I turn for legal assistance?) <end_of_turn>
<start_of_turn>model
Сіз заң көмегін алу үшін тұрғылықты жеріңіздегі адвокаттар алқасына немесе мемлекеттік құқықтық ақпарат порталдарына жүгіне аласыз (You can contact your local bar association or state legal information portals for legal assistance) <end_of_turn>

Бұл zero-shot prompting стратегиясын сипаттайды, мұнда модель қосымша бағыттаушы мысалдарсыз-ақ диалог құрылымына сүйене отырып өз мінез-құлқын анықтайды. Алайда күрделі немесе мағынасы екіұшты жағдайларда альтернативті prompt стратегиялары модельдің өнімділігін едәуір жақсарта алады. Мысалы, chain-of-thought prompting тәсілі модельге reasoning процесін кезең-кезеңімен логикалық түрде орындауға мүмкіндік береді:

<start_of_turn>user
Сот шешіміне шағым беру процесін кезең-кезеңімен түсіндіріп беріңіз (Explain the process of appealing a court decision step by step) <end_of_turn>

KazLLM және Qwen модельдерін fine-tuning процесі қазақ тіліндегі құқықтық және әкімшілік QA деректері негізінде supervised instruction-tuning тәсілі арқылы жүзеге асырылды. Деректер жиынтығы instruction, input және күтілетін response бөліктерін бөлек көрсететін құрылымдалған prompt шаблоны арқылы форматталды. Мұндай құрылым модельдердің әртүрлі тапсырма түрлері мен енгізу формаларын жалпылау қабілетін арттыруға мүмкіндік берді. Сонымен қатар instruction форматы тапсырмаларды жалпылауды жақсартып, жауаптардың контекстке негізделуін (response grounding) күшейтті және пәндік сала мамандары үшін интерпретациялауды жеңілдетті.

Модельдер EOS token шекараларын қолданатын CLM loss функциясы арқылы оқытылды, ал inference процесі single-turn және multi-turn completion режимдерін қолдауға бейімделді.

Конфигурация шаблоны келесідей түрде берілді:

Instruction:
{}
Input:
{}
Response:
{}””” + *EOS_TOKEN*

Құрылымдалған формат әртүрлі пайдаланушы мақсаттарына бейімделген өнімділікті оңтайландыру үшін prompt engineering стратегияларының бірнеше түрін қолдануға мүмкіндік берді. Қарапайым zero-shot instruction мысалы келесідей түрде беріледі:

Instruction:
Сұраққа нақты, қысқаша жауап беріңіз (Answer the question clearly and concisely)
Input:
Сот шешіміне апелляцияны қай мерзімде беру керек? (When should an appeal against a court decision be filed?)
Response:
Апелляция сот шешімі шыққан күннен бастап 15 күн ішінде берілуі тиіс (The appeal must be filed within 15 days from the date of the court decision)

Күрделірек reasoning процесін қолдау үшін chain prompting тәсілін қолдануға болады, мұнда модельден кезең-кезеңімен түсіндіру нақты түрде сұралады:

Instruction:
Сот ісін қайта қарау процесін кезең-кезеңімен түсіндіріңіз (Explain the judicial review process step by step)
Input:
Қайта қарау бастамасын қалай көтеруге болады? (How to raise the initiative for reconsideration?)
Response:
1. Ең алдымен, шешімнің күшіне енгеніне қарамастан, жаңа мән-жайлар анықталуы керек (First of all, new circumstances must be identified, regardless of whether the decision has entered into force)
2. Содан кейін өтініш сотқа жазбаша түрде беріледі (The application is then submitted to the court in writing)

Бұл құрылымдалған тәсіл модельдің тапсырма семантикасы негізінде тілдік вариацияларды анықтау қабілетін күшейтіп, prompt әртүрлілігін қолдауға және қазақ тіліндегі құқықтық тапсырмалар бойынша анағұрлым тұрақты жалпылауды қамтамасыз етуге мүмкіндік береді.

LLaMA негізіндегі QA моделі чаттық нұсқауларды орындауға бейімделген 3 миллиард параметрлі модель негізінде құрылды. Fine-tuning процесі Meta компаниясының ChatML prompt құрылымы арқылы жүзеге асырылды, мұнда <|begin_of_text|>, <|start_header_id|> және <|eot_id|> сияқты арнайы токендер қолданылды. Бұл формат system, user және assistant рөлдерін нақты бөліп көрсетуге мүмкіндік беріп, модельдің бірнеше диалогтық айналым барысында coherency сақтауына көмектесті. Деректер жиынтығы қазақ тіліндегі құқықтық QA жазбаларынан тұрды, мұнда контекст байланыстырылған құқықтық құжаттар арқылы беріліп, жауаптар ресми стильдік нормаларға сәйкес қалыптастырылды.

Бұл конфигурация ықшамдылықты, төмен есептеу ресурстарын қажет етуді және GPU-light deployment орталарында жоғары responsiveness қамтамасыз етті.

LLaMA моделінің конфигурациясы келесідей түрде көрсетілді:

“““<|begin_of_text|><|start_header_id|>system
{}<|end_header_id|>{}<|eot_id|>
<|start_header_id|>user
{}<|end_header_id|>{}<|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
{}<|eot_id|>””” + EOS_TOKEN

Prompt құрылымы модельдің тиімді жұмыс істеуінде шешуші рөл атқарды және instruction мен пайдаланушы сұрақтарының қалай құрастырылғанына тікелей тәуелді болды. Мысалы, system рөлін нақтылайтын қарапайым zero-shot prompt келесідей түрде берілуі мүмкін:

<|begin_of_text|><|start_header_id|>system
Сіз Қазақстан заңнамасы бойынша сұрақтарға ресми және нақты жауап беретін көмекші боласыз (You will be an assistant who will provide official and accurate answers to questions on Kazakhstani legislation)
<|end_header_id|><|eot_id|>
<|start_header_id|>user
Жер теліміне меншік құқығын қалай алуға болады? (How to obtain ownership of a land plot?)
<|end_header_id|><|eot_id|>
<|start_header_id|>assistant<|end_header_id|>
Жеке меншікке жер телімін алу үшін сіз жергілікті атқарушы органға өтініш беруіңіз керек. Сонымен қатар, жер кадастрынан қажетті құжаттарды рәсімдеу қажет. (To obtain a land plot for private ownership, you must apply to the local executive body. In addition, you must obtain the necessary documents from the land cadastre)<|eot_id|>

Phi негізіндегі QA моделі Microsoft компаниясының ChatML-lite форматы арқылы конфигурацияланды, мұнда рөлдердің шекараларын белгілеу үшін <|im_start|>, <|im_sep|> және <|im_end|> сияқты арнайы тегтер қолданылды. Шамамен 1.3 миллиард параметрге ие Phi-2 моделіне негізделген бұл модель

қазақ тіліндегі құқықтық сұраныстар үшін instruction-tuning тәсілімен оқытылды. System prompt ассистенттің мінез-құлқын анықтап отырды, ал пайдаланушы сұрақтары мен модель жауаптары жеңіл бөлгіштер арқылы нақты ажыратылды. Fine-tuning процесі EOS token арқылы аяқталатын құқықтық корпус негізінде жүзеге асырылды.

Ықшам өлшемі мен тиімді tokenization тәсілінің арқасында Phi моделі edge-device deployment және жеңіл chatbot интеграциялары үшін тиімді шешім екенін көрсетті, сонымен қатар қазақ тіліндегі қысқа әрі фактілік жауаптарда жоғары дәлдікті сақтап қалды.

Phi моделінің конфигурациясы келесідей түрде берілді:

“““<|im_start|>system<|im_sep|>
{}<|im_end|>
<|im_start|>user<|im_sep|>
{}<|im_end|>
<|im_start|>assistant<|im_sep|>
{}<|im_end|>””” + EOS_TOKEN

Prompt құрылымы модельдің нақты әрі ресми жауаптарды сақтауын қамтамасыз етудегі негізгі факторлардың бірі болды. Стандартты prompt келесідей түрде құрастырылды:

<|im_start|>system<|im_sep|>
Сіз Қазақстан заңнамасы бойынша сұрақтарға қысқа әрі нақты жауап беретін заң кеңесшісісіз (You are a legal advisor who provides concise and clear answers to questions on Kazakhstani legislation)<|im_end|>
<|im_start|>user<|im_sep|>
Неке қию үшін қандай құжаттар қажет? (What documents are needed to get married?) <|im_end|>
<|im_start|>assistant<|im_sep|>
Неке қию үшін жеке куәлік, неке қию туралы өтініш және мемлекеттік баж салығын төлегені туралы түбіртек қажет (To get married, you need an identity card, a marriage application, and a receipt for paying the state fee)<|im_end|>

Chain response құрылымында пайдаланушы prompt-ы модельден құрылымдалған жауаптарды нақты түрде талап ете алады:

<|im_start|>user<|im_sep|>
Апелляциялық шағым беру қадамдарын ретімен сипаттаңыз (Describe the steps in order of filing an appeal)<|im_end|>
<|im_start|>assistant<|im_sep|>
1. Сот шешімімен келіспеген жағдайда апелляция жазылады (If you disagree with the court’s decision, you can file an appeal)

2. Шағым шешім шыққан күннен бастап 15 күн ішінде берілуі керек (The appeal must be filed within 15 days from the date of the decision) <|im_end|>

Әртүрлі prompt түрлерін қолдану арқылы бұл конфигурация Phi моделінің шектеулі есептеу орталарында да тұрақты жұмыс істеуіне мүмкіндік берді, сонымен қатар құқықтық дәлдік пен пайдаланушыға түсініктілікті сақтап қалды.

7 миллиард параметрлі Mistral моделі ChatML форматынан шабыттанған prompt құрылымы арқылы fine-tuning процесінен өтті. Бұл құрылым <|im_start|>, <|im_end|> және EOS_TOKEN арнайы токендері негізінде ұйымдастырылды. Әрбір оқыту мысалы нақты анықталған system prompt, пайдаланушы сұрағы және assistant жауабынан тұрды. Мұндай формат модельдің дұрыс диалогтық құрылымды сақтауына және рөлдердің нақты бөлінуіне мүмкіндік беріп, instruction талаптарына сәйкестікті күшейтті.

Оқыту мақсаты ретінде CLM (causal language modeling) қолданылды, ал fine-tuning барысында Mistral моделінің бастапқы tokenizer және positional encoding механизмдері сақталды. Бұл конфигурация ресми қазақ тілінде еркін әрі заңдық тұрғыдан дұрыс жауаптар бере алатын жылдам, ықшам және дәл ассистент құруға мүмкіндік берді.

Mistral моделінің конфигурациясы келесідей түрде көрсетілді:

<|im_start|>system
{}<|im_end|>
<|im_start|>user
{}<|im_end|>
<|im_start|>assistant
{}<|im_end|> + EOS_TOKEN

Mistral моделіне арналған prompt құрылымы модельдің мінез-құлқын пәндік салаға тән талаптармен сәйкестендіруде маңызды рөл атқарды. Стандартты prompt келесідей түрде ұйымдастырылды:

<|im_start|>system
Сіз Қазақстан заңдарына сүйеніп, заңды сұрақтарға нақты және ресми түрде жауап беретін кеңесші боласыз (You will be a consultant who will provide clear and official answers to legal questions based on the laws of Kazakhstan) <|im_end|>
<|im_start|>user
Жеке кәсіпкерді тіркеу үшін не істеу керек? (What do I need to do to register as a sole proprietor?) <|im_end|>
<|im_start|>assistant
Жеке кәсіпкер ретінде тіркелу үшін eGov порталында өтініш толтырып, ЭЦҚ арқылы қол қою қажет. Сонымен қатар, салық органдарына тіркелу жүргізіледі (To register as an individual entrepreneur, you need to fill out an application on the eGov portal and sign it using an electronic

signature. In addition, registration with the tax authorities is carried out) <|im_end|>

Chain response мысалы келесідей түрде беріледі:

<|im_start|>user

Неке бұзу процесін кезең-кезеңімен түсіндіріп беріңіз (Explain the divorce process step by step) <|im_end|>

<|im_start|>assistant

1. Алдымен некені бұзу туралы өтініш АХАТ бөліміне немесе сотқа беріледі (First, an application for divorce is filed with the Civil Registry Office or the court)

2. Егер екі тарап та келіссе, процесс жеңілдетілген түрде өтеді (If both parties agree, the process will be simplified) <|im_end|>

Осы құрылым арқылы Mistral моделі басқарылу деңгейінің жоғарылығына және жауаптардың айқындылығына қол жеткізді, бұл оны ресми немесе консультативтік ортада сенімді түрде қолдану үшін маңызды фактор болып табылады.

Модельдер шектеулі есептеу ресурстарында тиімді оқытуды қамтамасыз ету мақсатында 4-bit quantization негізіндегі Quantized Low-Rank Adaptation (QLoRA) тәсілі арқылы fine-tuning процесінен өткізілді. Мұнда quantization type ретінде nf4 қолданылып, double quantization және float16 есептеулері пайдаланылды.

Модель архитектурасы adaptive LoRA қабаттарымен толықтырылды, мұнда rank параметрі $r = 16$, ал $lora\_alpha = 16$ ретінде орнатылды және dropout мәні нөлге тең болды. Оқыту барысында $q\_proj, k\_proj, v\_proj, o\_proj, gate\_proj, up\_proj$ және $down\_proj$ проекцияларына ерекше назар аударылды.

Деректерді дайындау кезеңі tokenization және EOS token қолдану арқылы Alpaca prompt стилінде форматтауды қамтыды.

Есептелген ROUGE және METEOR метрикаларынан бөлек, генерацияланған жауаптар синтаксистік, семантикалық және морфологиялық қателерді, сондай-ақ hallucination жағдайларын анықтау мақсатында қосымша талданды. Gemini және Qwen модельдері құқықтық мәтіндерде қосымша оқытусыз жеткілікті деңгейде семантикалық дәлдік көрсете алмады. Атап айтқанда, Gemini моделі агглютинативті тіл құрылымдарына бейімделмеген SentencePiece tokenizer қолдануына байланысты морфологиялық қателер мен hallucination жағдайларын жиі тудырды. Бұл модельдің жауаптарында сөздер мен сөз тіркестерінің қайталануы байқалды. Кейбір жауаптар шамадан тыс ұзын немесе өте қысқа болды. Мұндай жауаптарды не кеңейту, не қысқарту арқылы сұрақтың негізгі аспектілерін нақты көрсету қажеттілігі туындады.

Qwen моделі қазақ тіліндегі құқықтық контекске жеткілікті дәрежеде бейімделмегенін көрсетті. Оның жауаптарында елеулі hallucination жағдайлары көп кездеспегенімен, кейбір жауаптар тым ұзақ болды. Сонымен қатар, кейбір жағдайларда жауаптар сұрақ тақырыбына мүлде сәйкес келмей, hallucination түрінде көрінді.

LLaMA моделінің жауаптарында <|end_of_text|> сияқты техникалық элементтердің сақталуы байқалды, алайда мұндай артефактілерді regular expression көмегімен оңай жоюға болады. Кейбір жауаптар да шамадан тыс ұзын болып, қысқартуды талап етті.

Mistral моделі өз жауаптарында <|im_start|>system\nCategory: Тікелей сұрақтар\n<|im_start|>user\n сияқты instruction markup элементтерін сақтап қалды. Сонымен қатар кейбір жауаптарда сөздер мен сөз тіркестерінің қайталануы кездесіп, жауаптың түсініктілігін төмендетті.

Phi моделінің жауаптарында hallucination жағдайлары көп байқалмады. Оның жауаптары анық әрі түсінікті болғанымен, кейбіреулері басқа модельдердің орташа жауап ұзындығымен салыстырғанда шамадан тыс ұзын немесе қысқа болды.

KazLLM моделі нақты әрі грамматикалық тұрғыдан дұрыс жауаптар құрастыруда ерекше нәтиже көрсетті. Дегенмен KazLLM моделінің fine-tuning кезеңі шешуші рөл атқарды, себебі базалық модель бастапқыда қатесіз жауап бере алмады. Ең жиі кездескен қателердің бірі - нақты жауап берудің орнына сұрақты қайталау болды. Кейбір жауаптарда сөздердің көп рет қайталануы немесе жауаптардың тым қысқа болуы да байқалды.

Ұсынылған модельдердің ішінде GPT-4o mini моделі ең жоғары сапалы нәтижелер көрсетті. Оның жауаптарында елеулі қателер байқалмады. Сонымен қатар модель мүмкіндігінше толық әрі жан-жақты жауап беруге ұмтылды, бұл оның басқа модельдермен салыстырғандағы тиімділігін арттырды.

Кесте 2.7 - Әртүрлі модельдер мен деректер жиынтықтарында сұрақтарға жауап беру қателерін анықтау.

Датасет	Сұрақ	Predictions_Gemini	Predictions_LLaMA	Predictions_Qwen
Adilet	Ұйым халықаралық ұйым ретінде толыққанды заңды тұлға мәртебесін пайдаланатын болады ма?	Жоқ, ұйым заңды тұлға мәртебесін алмауға шешім қабылдады.	Қазақстан Республикасының заңдары мен халықаралық халықаралық ұйымдарға қатысуға құқықтық жағдайларды белгілейтін заңдарында көрсетілген құқықтық жағдайлармен айқындалады.	Ұйым халықаралық ұйым ретінде толыққанды заңды тұлға мәртебесін пайдаланатын болады.

Тәжірибелердің тек KazLLM моделінде емес, бірнеше түрлі модельдерде жүргізілуінің негізгі себептері төмендегідей. Gemini, Qwen, Mistral және Phi сияқты жалпы мультитілді LLM модельдері үлкен көлемдегі мультитілді

корпустарда алдын ала оқытылған және ашық салмақтар мен тиімді inference мүмкіндіктерімен кеңінен қолжетімді болды. Бұл оларды қазақ тілі үшін бастапқы кезеңде ауқымды инфрақұрылым құрмай-ақ прототиптеу және benchmarking жүргізуге ыңғайлы құрал етті. Зерттеу жүргізілген кезеңде қазақ тіліндегі әртүрлі мәтін корпустарында оқытылған жоғары сапалы LLM жүйелерінің саны шектеулі болды және кең қолжетімді емес еді.

Керісінше, жалпы мультитілді модельдер pre-training барысында қазақ тілін қоса алғанда көптеген тілдік деректерді көргендіктен, әртүрлі тақырыптарды қамту қабілетіне ие болды. Сонымен қатар мұндай модельдер feasibility тексеру, baseline performance бағалау және қазақ тіліне тән лингвистикалық қиындықтарды анықтау үшін негіз ретінде қолданылды. Бұл кейіннен KazLLM сияқты қазақ тіліне арнайы оңтайландырылған модельдерді дамыту мен жетілдіруге бағытталған жұмыстарға негіз қалады. Сондықтан жалпы LLM модельдері бастапқыда шектеулі тақырыптарда оқытылған KazLLM моделіне қарағанда анағұрлым әртүрлі output деректерін ұсынды.

Пайдаланылған модельдер мен датасеттерге қатысты қателер мысалдары Сурет 2.7-де көрсетілген.

```
df.iloc[[28, 37]]
```

№	input	instruction	Predictions_Mistral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_KazLLM	Predictions_Phi
28 29	Category: Тікелей сұрақтар	Зейнетақы жинақтарын аудару туралы өтінішпен қ...	<|im_start|>system\rInCategory: Тікелей сұрақтар...	Зейнетақы жинақтарын аудару туралы өтінішпен х...	Жүйелік кірістер кезінде жеке тұлға анықталған...	Зейнетақы жинақтарын аудару үшін Қазақстан Рес...	Құрметті Мәдени, біз сізге Қазақстан Республика...	Зейнетақы жинақтарын аудару үшін өтініш Елік К...
37 38	Category: Тікелей сұрақтар	ХҚБ жобасына қалай өтініш беруге болады?	<|im_start|>system\rInCategory: Тікелей сұрақтар...	ХҚБ-да білім беру, жұмысқа кіруін алу және қы...	«Қазақстан Республикасының Құқықтық академиясы...	ХҚБ жобасына қалай өтініш беруге болады? ХҚБ ж...	Қазақстан Республикасындағы Құқықтық ақпаратты...	Әділет министрлігінің қаулысына сәйкес қалай ...

```
df[df['instruction'] == 'Зейнетақы жинақтарын аудару туралы өтінішпен қайда жүгінуге болады']
```

№	input	instruction	Predictions_Mistral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_KazLLM	Predictions_Phi
28 29	Category: Тікелей сұрақтар	Зейнетақы жинақтарын аудару туралы өтінішпен қ...	<|im_start|>system\rInCategory: Тікелей сұрақтар...	Зейнетақы жинақтарын аудару туралы өтінішпен х...	Жүйелік кірістер кезінде жеке тұлға анықталған...	Зейнетақы жинақтарын аудару үшін Қазақстан Рес...	Құрметті Мәдени, біз сізге Қазақстан Республика...	Зейнетақы жинақтарын аудару үшін өтініш Елік К...

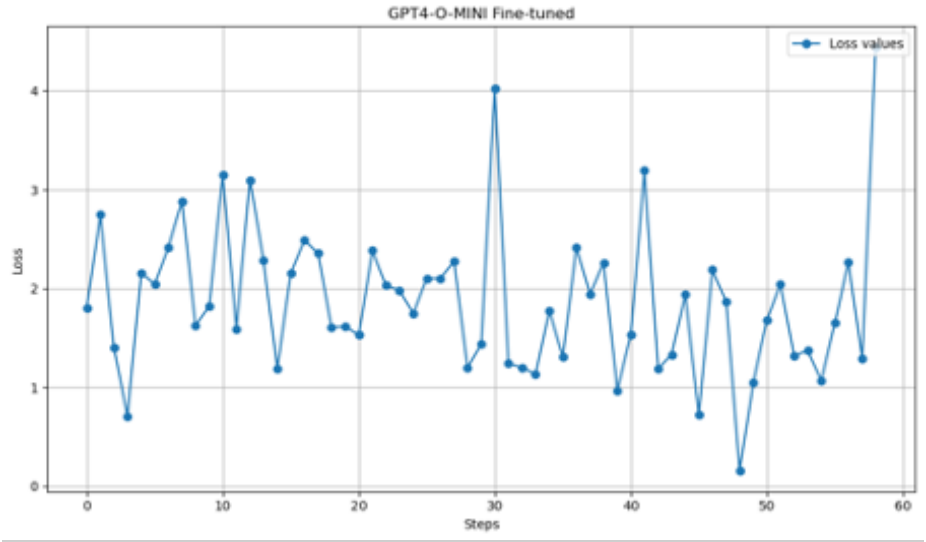
```
df[df['instruction'] == 'ХҚБ жобасына қалай өтініш беруге болады?']
```

№	input	instruction	Predictions_Mistral	Predictions_Qwen	Predictions_LLaMA	Predictions_Gemma	Predictions_KazLLM	Predictions_Phi
37 38	Category: Тікелей сұрақтар	ХҚБ жобасына қалай өтініш беруге болады?	<|im_start|>system\rInCategory: Тікелей сұрақтар...	ХҚБ-да білім беру, жұмысқа кіруін алу және қы...	«Қазақстан Республикасының Құқықтық академиясы...	ХҚБ жобасына қалай өтініш беруге болады? ХҚБ ж...	Қазақстан Республикасындағы Құқықтық ақпаратты...	Әділет министрлігінің қаулысына сәйкес қалай ...

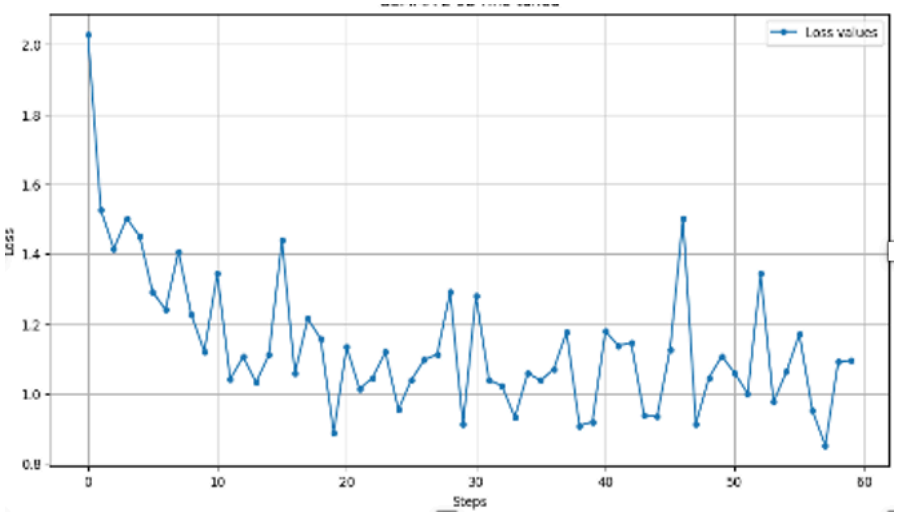
2.7-сурет. Модельдің генерациялаған жауаптарының қателіктері

Бұл метрикалық үрдістерді жоғалту графиктері (2.8-2.14 суреттер) одан әрі растайды; аталған графиктер модельдердің жинақталу динамикасын көрсетеді.

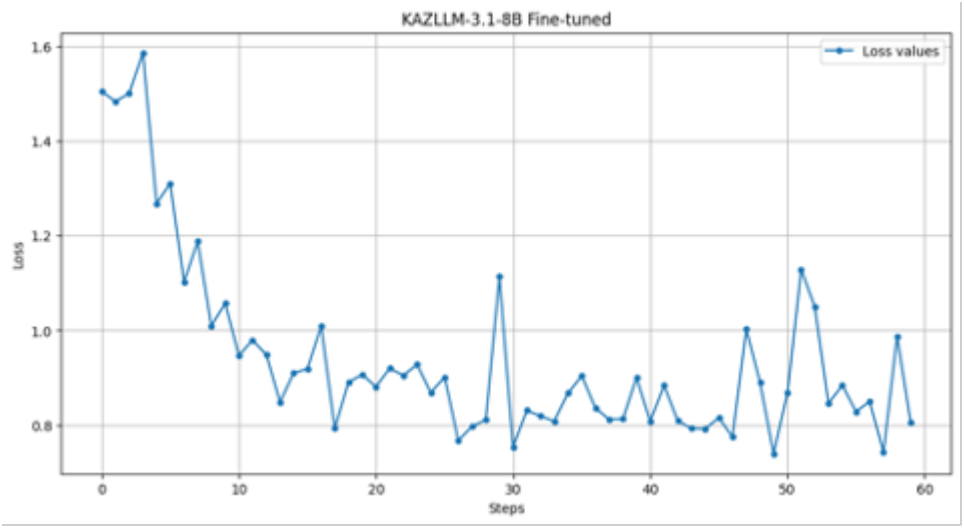
Әрбір модель үшін алынған loss function графиктері де сәйкесінше 2.8–2.14 суреттерде берілген.



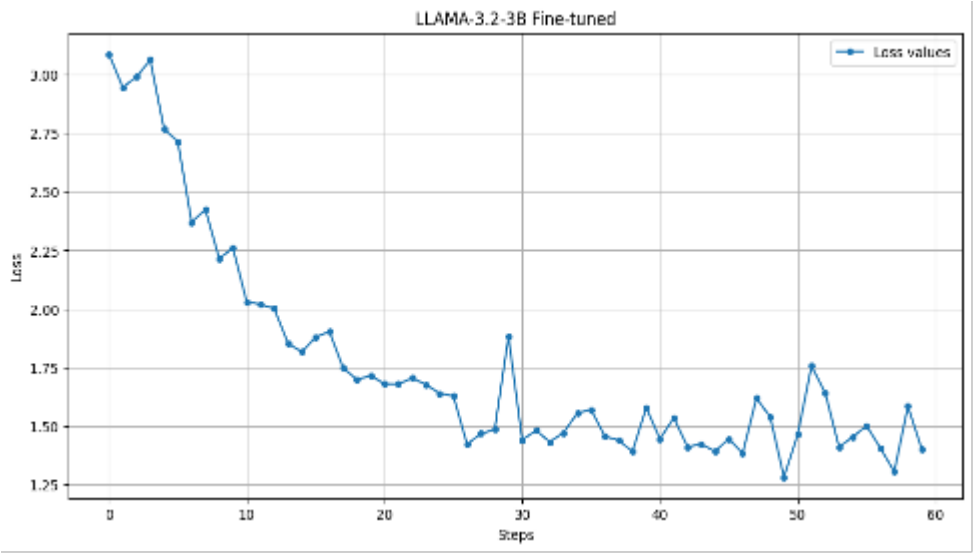
Сурет 2.8 - GPT-4o mini моделінің жоғалту графигі



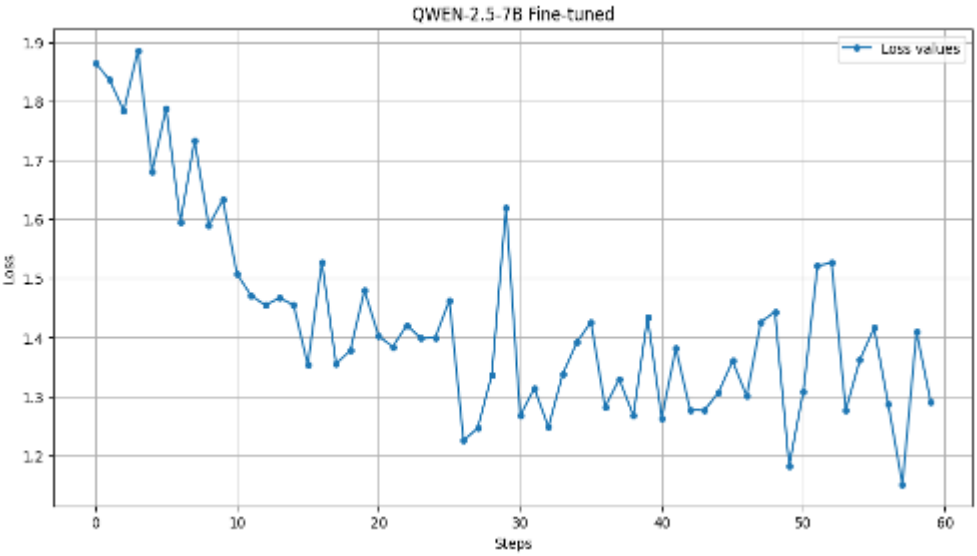
Сурет 2.9 - Gemini моделінің жоғалту графигі



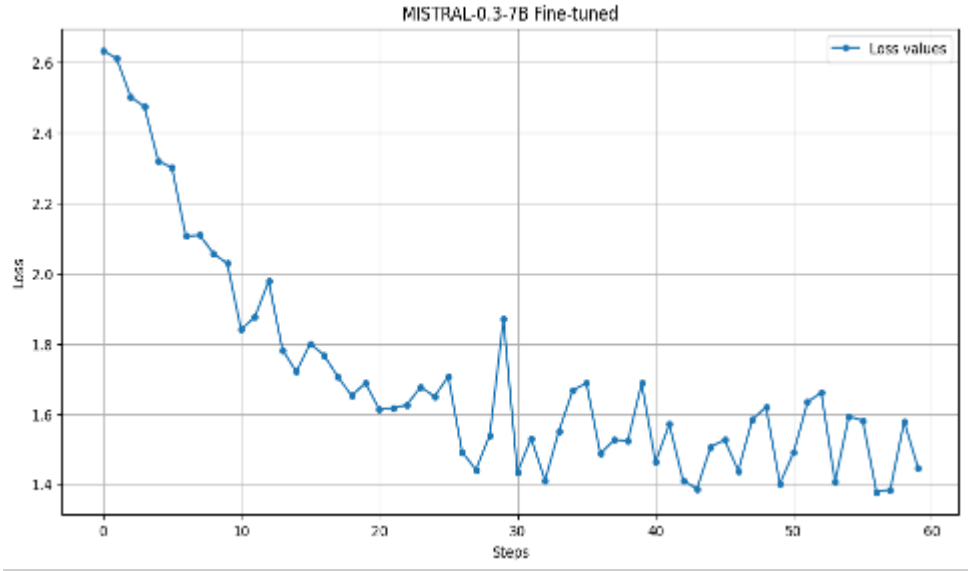
Сурет 2.10 - KazLLM моделінің жоғалту графигі



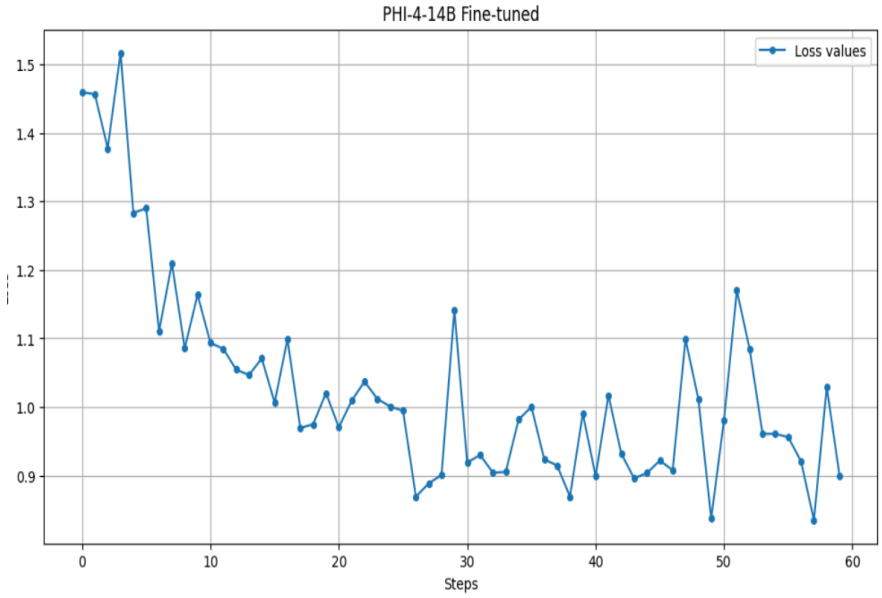
Сурет 2.11 - Llama моделінің жоғалту графигі



Сурет 2.12 - QWEN моделінің жоғалту графигі



Сурет 2.13 - QWEN моделінің жоғалту графигі



Сурет 2.13 - Phi-4 моделінің жоғалту графигі

GPT-4o mini төмен амплитудалы тербелістермен тегіс және біркелкі төмендеу көрсетеді, ұзақ тоқырау кезеңдері байқалмайды - бұл градиенттердің тұрақтылығын, үйрену үдерісінің соңына дейін жалғасуын және жақсы жалпылау қабілетін білдіреді. Gemini мен KazLLM, керісінше, тоқыраулар мен тербелістермен сипатталатын тұрақсыз жинақталу үлгісін көрсетті - бұл асқын үйренуді немесе оқыту жылдамдығы кестесіне сезімталдықты білдіруі мүмкін. Gemini алғашқы бірнеше қадамда күрт төмендеп, содан кейін кейде шыңдармен 1,0–1,2 мәндері аралығында тұрақтайды; KazLLM болса алғашқы 10 қадамда 1,5-тен 0,8-ге дейін жылдам төмендеп, одан кейін тек қысқа мерзімді ауытқулармен ең төмен және тұрақты қисықты сақтайды. LLaMA 3,1-ден 1,5-ке дейін үлкен бастапқы құлдырау көрсетіп, содан кейін 1,4–1,6 аралығында жеңіл шуылды

тоқыраумен жалғасады. Mistral мен Qwen модельдері орташа тұрақтылық танытты, алайда олардың қисықтары ерте қанығуды және одан арғы оңтайландыру мүмкіндіктерінің шектеулілігін аңғартты. Qwen 20-қадамға дейін 1,9-дан 1,3-ке дейін біртіндеп төмендеп, үйрену ортасында бірнеше шыңдармен салыстырмалы түрде тұрақты деңгейде қалады. Mistral 15 қадам ішінде 2,6-дан 1,6-ға дейін тез төмендеп, содан кейін кейде бөтен мәндермен 1,45–1,6 белдеуінде тұрақтанады. Барлық модельдер градиенттік жинақтаумен және QLoRA негізіндегі оңтайландырумен үйлестірілген кросс-энтропияға негізделген CLM жоғалту функциясы арқылы үйретілді. Бұл әдістер шектеулі ресурстар жағдайында үйрету тұрақтылығы мен тиімділігін арттырғанымен, нәтижелер жоғалтуды азайтудың өзі заң саласындағы домендік генерация сапасын бағалау үшін жеткіліксіз екенін көрсетеді.

Барлық модельдер реттіліктегі нақты келесі токендерге қарсы болжанған токендердің теріс логарифмдік ықтималдығын өлшейтін кросс-энтропияға негізделген стандартты CLM жоғалту функциясын қолданды. Қарапайымдылығына қарамастан, CLM авторегрессиялық тілдік модельдерді үйрету үшін, әсіресе нұсқаулықпен реттелген жағдайларда, тиімді болып қала береді. Үйрету үдерісі сондай-ақ градиенттік жинақтаудан және 4-биттік кванттауы бар QLoRA негізіндегі оңтайландырудан пайда көрді; бұл модель сапасын сақтай отырып, шектеулі есептеу ресурстарын тиімді пайдалануға мүмкіндік берді.

Үйрету жоғалтуының төмендеуі әрқашан жоғары шығыс сапасын кепілдендірмейтінін атап өткен жөн. Мысалы, LLaMA моделі тек орташа жоғалту азаюын көрсетті, бірақ семантикалық бағалауда бәсекеге қабілетті нәтиже берді - жеңіл модельдер арасында ең жоғары METEOR көрсеткішіне қол жеткізді. Бұл жоғалту қисықтарын кейінгі кезең бағалау метрикаларымен бірлесіп талдаудың қажеттілігін - әсіресе тілдік дәлдік пен контекстік дұрыстықтың маңызды болатын заңдық сұрақ-жауап сияқты домендік міндеттерде - айқын көрсетеді.

Болашақ жетілдірулерге баламалы жоғалту функцияларымен эксперименттер жүргізу кіруі мүмкін, мысалы адамның кері байланысынан күшейтілген үйрену (RLHF) немесе контрастивті жоғалту функциялары - олар токен деңгейіндегі дәлдіктен тыс ұзақ мерзімді тәуелділіктерді, семантикалық келісімділікті және заңдық түсіндірмелілікті жақсырақ қамти алады.

Қорытындылай келе, GPT-4o mini моделі үйрету тиімділігі де, шығыс сапасы да тұрғысынан айқын басымдылық танытты. GPT-4o mini-мен қатар, LLaMA және Gemini модельдері негізгі бағалау метрикалары бойынша жоғары нәтижелері негізінде қосымша сараптамалық бағалауға іріктелді. GPT-4o mini өзін ең күшті модель ретінде дәлелдегенімен, оны кеңінен өндірістік енгізуге назар аударуды қажет ететін бірқатар аспектілер байланысты. Алаңдаушылық тудыратын мәселелердің бірі - бұл модельді орналастыру кезеңінің маңызды бөлігі болып табылатын интернет байланысының тұрақтылығы. Бұлтқа негізделген API өзара іс-қимылы үшін тұрақты және жоғары жылдамдықты байланыс міндетті шарт болып табылады. Өкінішке орай, Қазақстанның кейбір

өңірлері мықты желілік инфрақұрылыммен мақтана алмайды, бұл жүйесіз интеграцияны немесе кеңінен таралуды шектеуі мүмкін. Тағы бір маңызды фактор - тұрақты пайдалану шығындарын тудыратын коммерциялық API қолжетімділігімен байланысты қаржылық аспект. Мемлекеттік сектордың бюджеттік шектеулері көптеген компаниялар мен ұйымдар үшін орнықсыз болуы мүмкін, ал оны пайдалану басымдығы ірі мемлекеттік кәсіпорындарға тиесілі болып қалады. Дегенмен, бұл зерттеу негізінен сұрақ-жауап жүйелерін құрудағы модельдер мүмкіндіктерін салыстыруға бағытталған; GPT-4o mini пайдаланудың қарастырылған мүмкіндіктері төмен бюджетті пайдаланушылар мен ұйымдар үшін де ондаған және жүздеген сұраныстарда салыстырмалы түрде қолжетімді болып қала береді. Сондықтан сұрақ-жауап жүйесінің толық коммерциялық енгізілуіне дейін ұсынылған модельдердің кез келгенін пайдалану барлығына дерлік қолайлы болып табылады.

Fine-tuning процесінің толық нәтижелері, барлық жеті модель бойынша салыстырмалы талдау және эксперименттік бағалау деректері 4-тарауда толық баяндалады. Онда zero-shot және fine-tuning режимдеріндегі модельдердің ROUGE-1, ROUGE-2, ROUGE-L, METEOR метрикалары бойынша сандық салыстырмасы, сондай-ақ ExpertScore мен F1 Score нәтижелері кешенді түрде көрсетіледі.

2.3 RAG архитектурасын заңнамалық сұрақ-жауап жүйесіне интеграциялау

Алдыңғы бөлімдерде қарастырылған LLM модельдерінің fine-tuning нәтижелері жауап сапасын айтарлықтай жақсартқанымен, жүйелік қателерді талдау бірқатар шешілмеген мәселелерді айқындады: қажетті заңнамалық құжатты дәл таба алмау, бейресми сұрақтардың нашар өңделуі, морфологиялық вариативтіліктен туындайтын іздеу дәлдігінің төмендеуі және hallucination жағдайлары. Бұл мәселелерді шешу үшін RAG архитектурасы жүйеге интеграцияланды.

RAG технологиясы ақпараттық іздеу мен генеративті модельдерді біріктіретін гибридті архитектура болып табылады. Lewis et al. [42] ұсынған retrieve-then-generate принципіне сәйкес, жүйе алдымен сыртқы білім қорынан релевантты құжаттарды іздейді, кейін сол контекст негізінде тілдік модель жауап генерациялайды. Бұл тәсіл генеративті модельдің параметрлік білімін нақты дереккөздермен толықтырып, жауаптардың дәлдігін, түсіндірілетіндігін және сенімділігін арттыруға мүмкіндік береді. Жүйе алдымен пайдаланушы сұрағына семантикалық тұрғыдан жақын заңнамалық үзінділерді іздейді, содан кейін осы үзінділер контекст ретінде LLM-ге берілетін промптқа қосылады да, модель осы негізде жауап генерациялайды.

Заңнамалық доменде RAG архитектурасының маңызы ерекше - жауаптың нақты нормативтік-құқықтық актіге, бапқа немесе тармаққа сүйенуін қамтамасыз ету және hallucination деңгейін азайту мүмкіндігі береді. Тек генеративті модельдерді қолдану hallucination қаупін арттырады, яғни модель нақты заңнамада жоқ құқықтық ақпаратты сенімді түрде генерациялауы мүмкін. Сондықтан retrieval компоненті құқықтық QA жүйелерінде жауаптарды нақты

құқықтық дереккөздермен grounding жасау үшін шешуші рөл атқарады. Осыған байланысты RAG архитектурасын құқықтық сұрақ-жауап жүйесіне интеграциялау және retrieval сапасын жетілдіру диссертациялық жұмыстың негізгі зерттеу бағыттарының бірі болып табылады.

Халықаралық зерттеулер де құқықтық RAG жүйелерінің тиімділігін дәлелдейді. Pipitone және Alami ұсынған LegalBench-RAG benchmark жұмысы retrieval компонентін заңнамалық доменге арнайы бейімдеп бағалауға бағытталған [56]. Зерттеу нәтижелері құқықтық QA жүйелерінде жалпы мәтіндік retrieval тәсілдері жеткіліксіз екенін көрсетті: заңнамалық жүйелер аз көлемдегі, бірақ жоғары релевантты құқықтық үзінділерді қажет етеді. Сонымен қатар Barron et al. зерттеуі [57] құқықтық retrieval жүйелерінде vector database, knowledge graph және hierarchical indexing тәсілдерін біріктіру қажеттігін көрсетті. Сонымен қатар зерттеулер GPT-4 сияқты үлкен тілдік модельдердің өздігінен жұмыс істеген кезде заңнамалық тапсырмаларда 49% немесе одан да жоғары уақытта hallucination жіберетінін көрсетеді - бұл заңнамалық AI жүйелерінде RAG компонентінің маңыздылығын растайды.

Агглютинативті тілдер үшін retrieval мәселесі одан да күрделі сипатқа ие. Түрік тіліндегі құқықтық RAG жүйелерін зерттеген Köse S. K. et al. [58] жұмысы морфологиялық ерекшеліктерді ескеретін Hybrid RAG архитектурасын ұсынды. Зерттеу нәтижелері morphology-aware retrieval тәсілдерінің агглютинативті тілдерде MRR және ROUGE-L метрикаларын айтарлықтай жақсартатынын көрсетті. Бұл қорытындылар қазақ тілі үшін де маңызды, себебі қазақ тілі де жоғары морфологиялық вариативтілікпен сипатталады.

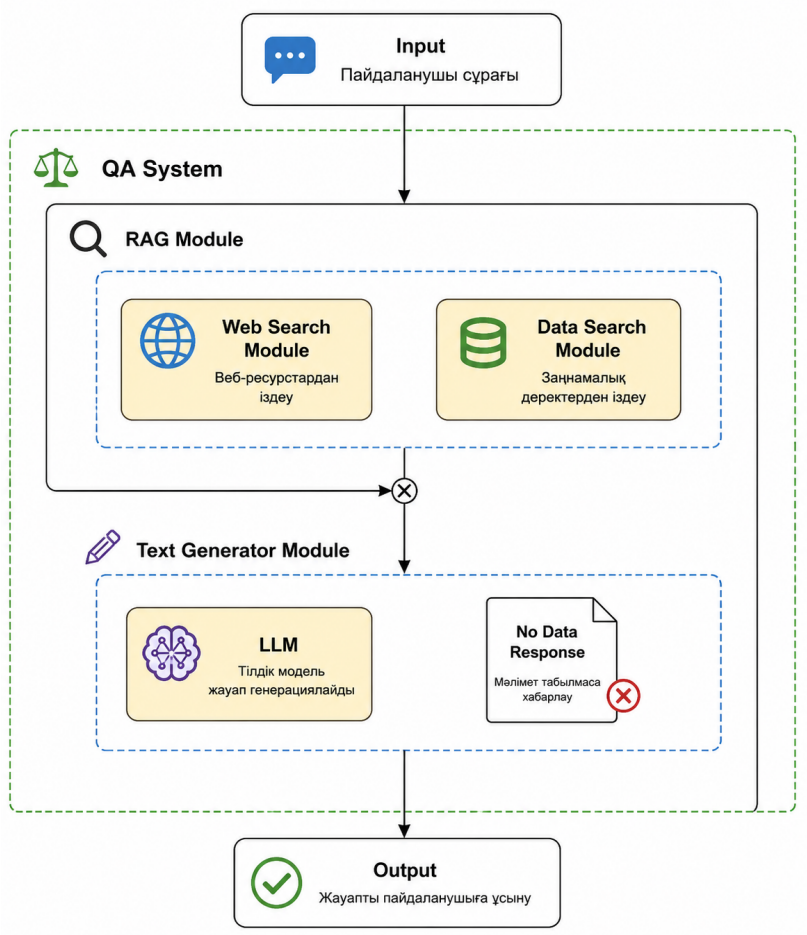
Сонымен қатар түрік тіліндегі зерттеулер заңнамалық мәтіндер үшін арнайы chunking стратегиясының қажеттігін көрсетті. Құқықтық құжаттар иерархиялық құрылымнан тұрады: бөлім → тарау → бап → тармақ → тармақша. Егер retrieval жүйесі бұл құрылымды ескермесе, қажетті құқықтық контексттің тек фрагменті ғана табылып, жауаптың толықтығы мен дәлдігі төмендеуі мүмкін. Осы тәжірибеге сүйене отырып, ұсынылған жүйеде заңнамалық иерархияны ескеретін chunking және indexing стратегиялары қолданылды.

Қазақстандағы құқықтық AI зерттеулерінде де retrieval мәселесінің өзектілігі байқалады. «Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation» жұмысында [4] қазақ тіліндегі құқықтық QA жүйесіне арналған деректер жиыны құрылып, бірнеше LLM модельдері бағаланды. Алайда зерттеу retrieval сапасының тұрақсыздығын, бейресми сұрақтарды заңнамалық стильге автоматты түрде түрлендірудің жеткіліксіздігін және morphology-aware preprocessing кезеңінің болмауын негізгі шектеулер ретінде көрсетті.

2.3.1 RAG жүйесіндегі іздеу модулінің құрылымы

Сұрақ-жауап жүйесі пайдаланушы сұрауларын өңдеудің интеллектуалды көпкезеңді кешені болып табылады және берілген деректер мен тілдік модельдер негізінде жауаптарды автоматты түрде алуды, интерпретациялауды және

генерациялауды қамтамасыз етеді. 2.14-суретте QAS жүйесінің жалпыланған жұмыс схемасы берілген.



Сурет 2.14 - Сұрақ-жауап жүйесінің жалпыланған жұмыс схемасы

Сурет 2.14-тен көрініп тұрғандай, QAS жүйесі келесі тізбекті процесс бойынша жұмыс істейді. Сұрақ (Input) жүйеге түскеннен кейін ол RAG модуліне жіберіледі. RAG модулі екі параллель іздеу каналынан тұрады: Web Search Module - веб-ресурстардан, атап айтқанда adilet.zan.kz және басқа ресми порталдардан іздеу жүргізеді; Data Search Module - ішкі заңнамалық дерекқордан деректерді шығарып алады. Іздеу нәтижесі Text Generator Module-ге жіберіледі: деректер табылған жағдайда LLM модулі жауап генерациялайды, деректер табылмаған жағдайда «No Data Response» хабарламасы қайтарылады. Соңғы нәтиже (Output) пайдаланушыға ұсынылады.

Retrieval модулі RAG жүйесінің бірінші және іргелі кезеңі болып табылады. Оның негізгі міндеті - пайдаланушының сұрағына семантикалық және лексикалық тұрғыдан ең релевантты заңнамалық үзінділерді корпустан тауып, оларды жауап генерациясы үшін контекст ретінде ұсыну. Retrieval модулінің сапасы жалпы жүйенің тиімділігін тікелей анықтайды: жауап генерациясы қаншалықты жетілдірілген болса да, іздеу сапасыз үзінділерді қайтарған жағдайда соңғы жауаптың дұрыстығына кепілдік беру мүмкін емес.

Заңнамалық корпусты индекстеу және бөлшектеу. adilet.zan.kz және zan.gov.kz порталдарынан жиналған 235 000 құжат (5,3 ГБ) retrieval жүйесіне енгізілді. Заңнамалық мәтіндер заңнамалық иерархияны ескере отырып бірнеше деңгейде бөлшектелді: тарау деңгейі, бап деңгейі және тармақ деңгейі. Толық заң мәтінін бір chunk ретінде берудің орнына нақты баптар мен тармақтар жеке индексациялық бірліктер ретінде өңделді. Мұндай тәсіл retrieval нәтижелерінің нақтылығын арттырып, LLM-ге берілетін контекстің ықшам және мазмұнды болуын қамтамасыз етеді. Заңнамалық іздеу жүйесі Қазақстан Республикасы заңнамасының 34 саласын қамтитын тақырыптық каталог негізінде жұмыс істейді, бұл пайдаланушы сұрауын тиісті құқықтық саладағы құжаттармен байланыстырады.

Гибридті іздеу механизмі. Іздеу сапасын арттыру мақсатында гибридті іздеу (hybrid retrieval) тәсілі қолданылды - лексикалық іздеу мен семантикалық іздеудің комбинациясы. Іріктеу үшін келесі формула қолданылды:

$$C = \alpha \cdot BM25(Qlegal, V) + \beta \cdot cos(vq, vi) \qquad (2.1)$$

мұнда α және β лексикалық және семантикалық іздеу салмақтарын реттейтін гиперпараметрлер;

$Qlegal -$ түрлендірілген сұрақ;

$v_q -$ сұрақтың векторлық ұсынымы;

$v_i - i -$ ші заңнамалық үзіндінің векторы.

Лексикалық іздеу (BM25). Сұрақ пен заңнамалық мәтін арасындағы кілт сөздердің сәйкестігін өлшейді. Нақты терминдердің, бап нөмірлерінің, нормативтік актілер атауларының дәлме-дәл сәйкестігі маңызды болатын жағдайларда тиімді [60]. Зерттеу деректері бойынша тек BM25-ке негізделген Adilet жүйесінің MRR=0,32, Paragraph жүйесінің MRR=0,35 деңгейінде болды - бұл лексикалық іздеудің өздігінен жеткіліксіз екенін растайды.

Семантикалық іздеу (Dense Retrieval). Сұрақ пен заңнамалық мәтіндерді жоғары өлшемді векторлық кеңістікке проекциялап, косинустық ұқсастықты өлшейді. Лексикалық сәйкестік болмаса да семантикалық жағынан жақын мәтіндерді таба алады [61]. Мысалы, пайдаланушы «жұмыстан шығарып жіберді» деп сұраса, семантикалық іздеу «еңбек шартын бұзу» немесе «қызметкерді жұмыстан босату» секілді ресми заңнамалық тіркестерді де тауып бере алады. Сұрақ векторы мынадай формуламен есептелді:

$$vq = Enc(Qlegal) \in Rd \qquad (2.2)$$

Векторлық дерекқор. Заңнамалық үзінділердің векторлық ұсынымдарын сақтау және жылдам іздеу үшін FAISS кітапханасы қолданылды. Vectorstore индексіндегі әрбір вектор акт атауы, бап нөмірі және URL-сілтемесі бар метадеректермен бірге сақталды [62].

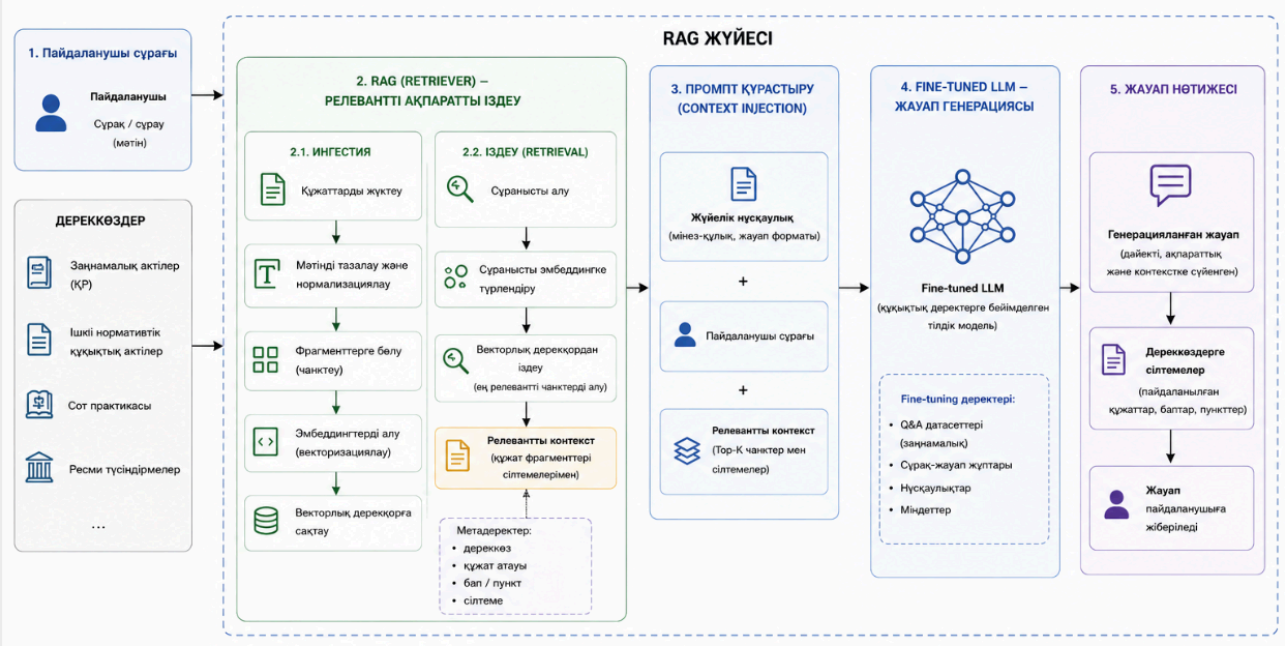
Іріктеу механизмі және threshold шегі. Retrieval модулі векторлық ұқсастық шегін (τ) қолданады. Егер sim(vq,vi)≥τ болса - ішкі НҚА базасынан деректер алынады; төмен болса - веб-іздеу модуліне (adilet.zan.kz) жүгінеді; веб-іздеу де нәтиже таппаса - «Мағлұмат табылмады. Норма базада жоқ» хабарламасы

қайтарылады. Top-k үзінділер іріктелгеннен кейін контекст жиыны C={d1,d2,…,dk} жауап генерациясы модуліне жіберіледі.

Іздеу модулінің бағалауы. Алайда іздеу модулінің қателіктерін талдау мынаны анықтады: бейресми сұрақтар нашар өңделеді, қазақ тілінің морфологиялық ерекшеліктері іздеу дәлдігіне теріс әсер етеді, ал кілт сөздердің заңнамалық стильге сәйкес келмеуі retrieval нәтижесінің сапасын төмендетеді. Бұл тұжырым іргелі қорытындыға алып келеді: retrieval сапасы тек архитектурамен ғана емес, сұрақтың сапасымен де тікелей байланысты. Дұрыс түрлендірілген, заңнамалық терминологияға сәйкестендірілген сұрақ RAG жүйесіне дұрыс заңнамалық үзінділерді табуға мүмкіндік береді, ал бұл өз кезегінде жауап генерациясының дәлдігін қамтамасыз етеді.

2.3.2 Генеративті жауап қалыптастыру моделі

Іздеу модулі заңнамалық корпустан релевантты үзінділерді іріктегеннен кейін бұл деректер генеративті жауап қалыптастыру модуліне жіберіледі. Бұл модуль интеллектуалды жүйенің негізгі компоненті болып табылады және пайдаланушы сұрағы мен іздеу нәтижелері контексі негізінде мағыналы, логикалық тұрғыдан байланысты және грамматикалық жағынан дұрыс жауап қалыптастыруды қамтамасыз етеді. 2.16-суретте мәтін генерациясы модулінің архитектурасының жалпыланған схемасы берілген.



Сурет 2.15 - Мәтін генерациясы модулінің архитектурасы

Сурет 2.15-тен көрінетіндей, модуль жұмысы келесі тізбек бойынша жүзеге асырылады. Text Generator Module-ден Input түскеннен кейін жүйе «If output exists» шартын тексереді. Шарт орындалған жағдайда (Yes) RAG Module-ден алынған RAG Output (Title, Text, Link) және пайдаланушы сұрағы Prompt-қа біріктіріледі: Prompt Instruction + User Input және Input + RAG Output Text

өрістерінен тұрады. Бұл біріктірілген промпт LLM модуліне жіберіліп, жауап генерацияланады. Генерацияланған жауап RAG Output деректерімен (Title, Text, Link) біріктіріліп, соңғы Output қалыптастырылады. Шарт орындалмаған жағдайда (No) - «No Data Response» хабарламасы қайтарылады.

Модуль жұмысының процесі кіріс деректерді қабылдаудан басталады. Кіріс деректер пайдаланушының мәтіндік сұрағын (Qlegal), retrieval модулі қайтарған заңнамалық үзінділерді (C) және жүйелік промптты қамтиды. Бірінші кезеңде деректер нормализация, құрылымдау және контекстуалдық толықтыру орындалатын кіріс ақпаратты дайындау модуліне түседі. Бұл операциялар мағыналық екіұштылықтарды жоюға және кіріс деректердің бірыңғай форматын қамтамасыз етуге бағытталған.

Өңделген сұрақ алдын ала оқытылған тілдік модельге негізделген генерация модуліне жіберіледі. Генерация модулі бірнеше негізгі функцияны орындайды: кіріс мәтінді талдайды және ықтимал жауаптардың ықтималдық таралуын қалыптастырады; жауаптың ұзындығы, стилі мен форматы сияқты генерация параметрлерін ескереді; сондай-ақ қалыптастырылатын жауаптың семантикалық келісімділігі мен логикалық байланыстылығын қамтамасыз етеді.

Промпт құрылымы. Генерация процесінде LLM-ге берілетін промпт жүйелік нұсқаулықтан, пайдаланушы сұрағынан (Qlegal) және retrieval модулі қайтарған контекстен (C) тұрды [63]. Жауап генерациясы келесі формуламен сипатталды:

$$A = LLM(Qlegal,\ C\ ;\ \varphi) \qquad (2.3)$$

Мұнда A - генерацияланған жауап; Qlegal – түрленген заңнамалық сұрақ; C - контекст жиыны; φ - LLM модель параметрлері.

Генерация режимдері. Мәтін генерациясы екі режимде жүзеге асырылды.

Шаблондық режим (Template-based Generation) - қатаң түрленген жауаптар қажет болған жағдайда, яғни нақты бапқа сілтеме жасай отырып стандартты жауап беру кезінде қолданылады.

Еркін генерация режимі (Free Generation) - ықтималдық әдістерге негізделген; жүйе логикалық құрылым мен мағыналық тұтастықты сақтай отырып, бірегей, контекстке негізделген жауап қалыптастырады. Пайдаланушының бейресми тілде қойған күрделі сұрақтарына толық түсіндірмелі жауап беру үшін қолданылады [64].

Заңнамалық верификация. Генерацияланған жауаптың заңнамалық дереккөздермен сәйкестігін тексеру үшін верификация кезеңі енгізілді:

$$V(A)=sim(A,C)\geq\tau \qquad (2.4)$$

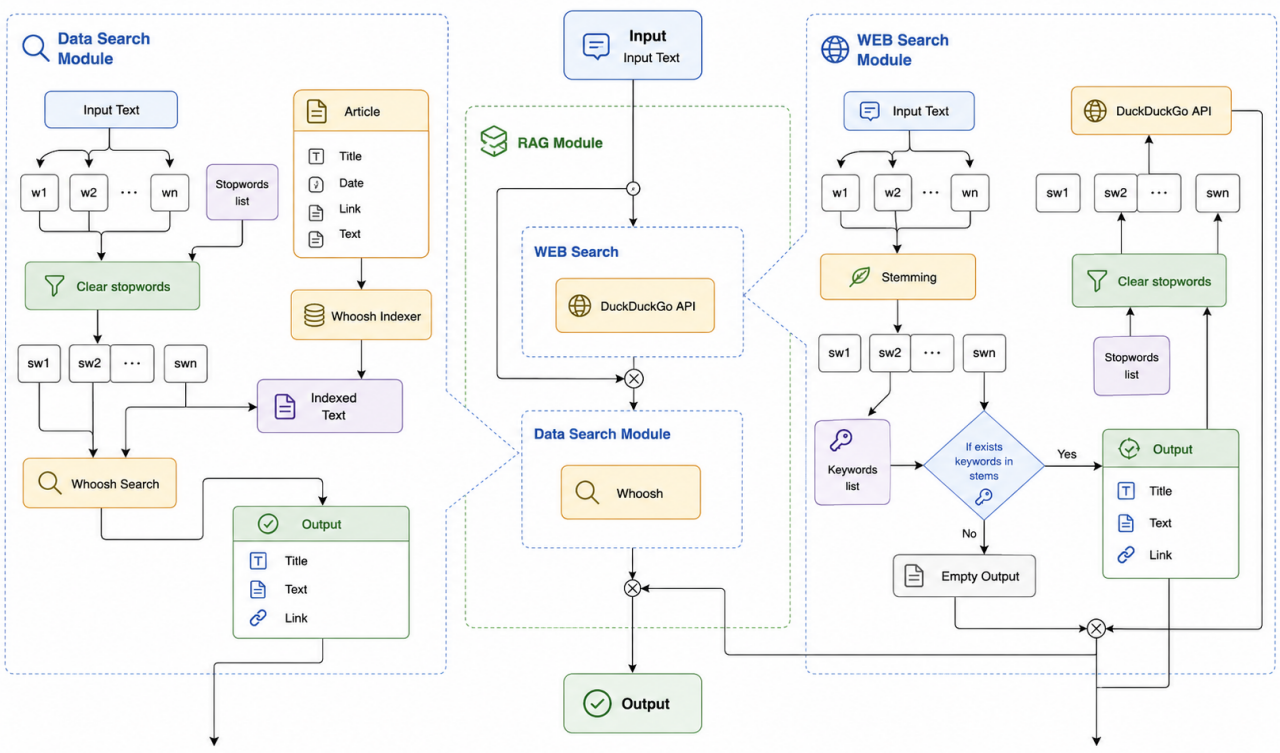
Мұнда sim(A,C) - генерацияланған жауап пен контекст арасындағы семантикалық ұқсастық; τ - сәйкестік шегі [65]. Егер шарт орындалса, жауап пайдаланушыға ұсынылады; орындалмаса, жүйе қайта іздеу жасайды немесе деректер жеткіліксіздігі туралы хабарлама береді. Schwarcz et al. жүргізген рандомизацияланған бақыланатын сынақта [66] RAG-мен жұмыс істейтін заңнамалық AI алты заңнамалық тапсырманың төртеуінде статистикалық

тұрғыдан айтарлықтай жақсару көрсетті; бұл нәтиже верификация механизмінің маңыздылығын эмпирикалық тұрғыдан растайды.

Осылайша, мәтін генерациясы модулінің архитектурасы семантикалық синтездің көпдеңгейлі процесін жүзеге асырады. Алайда бұл процестің тиімділігі тікелей RAG модулі ұсынған контексттің сапасына байланысты. Егер RAG дұрыс заңнамалық үзінділерді таба алса, LLM дәл және дереккөзге негізделген жауап генерациялайды; дұрыс үзінділер табылмаса, жауаптың сапасы да төмендейді. Бұл байланыс retrieval кезеңін жетілдірудің маңыздылығын қайтадан растайды.

2.3.3 Интеллектуалды заңнамалық сұрақ-жауап жүйесінің RAG архитектурасы

Іздеу модулі мен генеративті жауап қалыптастыру моделі бірлесіп RAG жүйесінің толық архитектурасын құрайды. Жалпы архитектура бес негізгі модульден тұрады: пайдаланушы сұрағын қабылдау және алдын ала өңдеу, деректерді іздеу (Retrieval), контекстті қалыптастыру, жауап генерациясы және верификациямен бірге постөңдеу.



Сурет 2.16 – Сұрақ жауап жүйесіне арналған RAG архитектурасы

Жалпы жүйенің жұмыс тізбегі. Пайдаланушы табиғи тілде сұрақ (q - raw text) береді. Жүйе сұрақтың заңнамаға қатыстылығын тексереді: қатысты болса - препроцессинг модуліне жіберіледі; болмаса - «Заңға қатысты сұрақ қойыңыз» хабарламасымен қайтарылады. Препроцессинг модулінде сұрақ өңделіп, RAG модуліне жіберіледі. Векторлық іздеу жүргізіледі; шарт орындалса - ішкі НҚА

базасынан BM25 + косинустық ұқсастық арқылы іріктеу жасалады; орындалмаса - веб-іздеуге жүгінеді. Контекст жиыны С қалыптасып, мәтін генераторы - LLM модулінде жауап генерацияланады. Заңнамалық верификациядан өткен соң пайдаланушыға тақырып, мәтін, бап үзіндісі және НҚА сілтемесін қамтитын толық жауап ұсынылады (2.1 формула).

Архитектураның негізгі компоненттері. Жалпы RAG архитектурасында төрт негізгі компонент өзара тығыз байланыста жұмыс істейді.

Деректерді алу модулі (Retriever) - заңнамалық актілер базасынан релевантты құжаттарды векторлық іздеу мен семантикалық салыстыру арқылы табады.

Сұранысты өңдеу модулі - мәтінді нормализациялайды; кілт сөздерді, заңдық терминдерді және нысандарды (баптар, тармақтар, кодекстер) бөліп алады.

Генеративті модуль (Generator) - ізделіп алынған үзінділер негізінде семантикалық келісімділік пен заңдық дұрыстықты қамтамасыз ете отырып байланысты жауап қалыптастырады. Постөңдеу модулі - тілдік түзету, жауапты құрылымдау жүргізеді және пайдаланылған нормативтік-құқықтық актілерге сілтемелер қосады.

Кесте 2.8 – Fine-tuning LLM және RAG интеграциясындағы қателіктер

№	Қателік түрі	Сипаттама
1	Іздеу қателігі	**Қажетті құжат табылмайды** *Кілт сөздер сәйкес келмейді; векторлық ұқсастық шегінен (τ) төмен*
2	Сұрақ тәуелділігі	**Сұрақ формасына тәуелді** *Бейресми / диалект сұрақтар нашар өңделеді; бір мағынадағы сұрақ түрлі нәтиже береді*
3	Морфологиялық күрделілік	**Сөз формалары ескерілмейді** *Қазақ тілінің жалғаулары іздеу дәлдігіне әсер етеді*
4	Заңнамалық құрылым олқылығы	**Заң логикасы толық түсінілмейді** *Субъект, әрекет, шарт ажыратылмайды; заң иерархиясы сақталмайды*

Retrieval кезеңіндегі негізгі шектеулі фактор - пайдаланушы сұрағы мен заңнамалық мәтін арасындағы семантикалық алшақтық. Бейресми тілде берілген сұрақтар заңнамалық терминологиямен бірден сәйкес келмейді, бұл іздеу нәтижесінің сапасын төмендетеді [67,68]. Демек, RAG жүйесінің тиімділігі тікелей retrieval алдындағы сұрақ өңдеу сапасына байланысты. Егер сұрақ заңнамалық тілге дұрыс сәйкестендірілсе, RAG дұрыс заңнамалық үзінділерді табады; дұрыс контекст негізінде LLM дәл жауап генерациялайды.

Осы бөлімде сипатталған RAG архитектурасының теориялық негіздері, retrieval компонентін заңнамалық доменге бейімдеу стратегиялары және

верификация механизмі Law-AI жүйесінің іске асырылуының концептуалдық базасын құрайды. SemanticPrep препроцессинг модулін RAG архитектурасымен интеграциялаудың нақты техникалық шешімдері, гибридті іздеу алгоритмдерінің параметрлері және жүйенің толық архитектуралық схемасы 3-тарауда баяндалады. RAG компонентінің заңнамалық QA жүйесіндегі тиімділігін бағалауға арналған эксперименттік нәтижелер - BERTScore, MRR, Recall@5 және Expert Score метрикалары бойынша салыстырмалы талдау - 4 тарауда толық көлемде ұсынылады.

3 ЗАҢНАМАЛЫҚ СҰРАҚ-ЖАУАП ЖҮЙЕСІ ҮШІН СЕМАНТИКАЛЫҚ ПРЕПРОЦЕССИНГ (SEMANTICPREP) ТӘСІЛІН ӘЗІРЛЕУ

RAG архитектурасына негізделген интеллектуалды сұрақ-жауап жүйелерінің тиімділігі пайдаланушы сұранысының семантикалық сапасына және заңнамалық корпустан релевантты ақпаратты іздеу дәлдігіне тікелей тәуелді. 2-тарауда жүргізілген талдау нәтижелері қазақ тіліндегі заңнамалық мәтіндерді өңдеу барысында retrieval сапасын төмендететін бірқатар жүйелік мәселелердің бар екенін көрсетті. Олардың қатарына қазақ тілінің агглютинативті морфологиялық құрылымынан туындайтын лексикалық вариативтілік, пайдаланушының табиғи тілдегі бейресми сұраныстары мен заңнамалық мәтіндердің ресми құқықтық стилі арасындағы семантикалық алшақтық, сондай-ақ құқықтық терминологияның әртүрлі тілдік формада берілуі жатады.

Сонымен қатар жүргізілген эксперименттік зерттеулер көрсеткендей, тек LLM модельдерін заңнамалық деректер негізінде fine-tuning арқылы бейімдеу немесе тек RAG архитектурасын енгізу құқықтық сұрақ-жауап жүйесінің сапасын белгілі бір деңгейде арттырғанымен, retrieval кезеңіндегі семантикалық сәйкессіздік мәселесін толық шеше алмайды. Атап айтқанда:

- fine-tuned LLM модельдері құқықтық терминологияны жақсырақ түсінгенімен, пайдаланушының бейресми сұраныстарын әрдайым дұрыс интерпретациялай алмайды;
- RAG архитектурасы нақты құқықтық құжаттарды пайдалануға мүмкіндік бергенімен, retrieval сапасы пайдаланушы сұранысының формасына жоғары деңгейде тәуелді болып қалады;
- морфологиялық вариативтілік пен құқықтық терминдердің көптүрлілігі релевантты нормативтік фрагменттердің толық анықталуына кедергі келтіреді;
- retrieval кезеңіндегі қателіктер генеративті жауаптың дәлдігіне, groundedness қасиетіне және құқықтық сенімділігіне тікелей әсер етеді;
- low-resource тілдер жағдайында embedding және semantic matching механизмдерінің тиімділігі қосымша семантикалық нормализациясыз шектеулі болады.

Аталған факторлар retrieval кезеңінде релевантты құқықтық фрагменттердің толық және дәл анықталуына кедергі келтіреді, нәтижесінде генеративті модельдің жауап сапасы мен нормативтік негізділігі төмендейді. Әсіресе ресурстары шектеулі және морфологиялық күрделі тілдер үшін бұл мәселе RAG жүйелерінің негізгі шектеуші факторларының бірі болып табылады.

Осы ғылыми және қолданбалы мәселелерді шешу мақсатында SemanticPrep деп аталатын екі кезеңді семантикалық алдын ала өңдеу тәсілі ұсынылады. Ұсынылған тәсілдің негізгі ғылыми идеясы табиғи тілдегі пайдаланушы сұранысын retrieval кезеңіне дейін құқықтық контекске семантикалық тұрғыдан бейімделген құрылымға түрлендіру арқылы заңнамалық ақпаратты іздеу сапасын арттыруға негізделген. Сонымен қатар SemanticPrep тәсілі генеративті модельдің нақты нормативтік құжаттарға сүйену деңгейін жоғарылатып,

құқықтық жауаптардың семантикалық дәлдігі мен groundedness қасиетін күшейтуге бағытталған.

3.1 SemanticPrep тәсілінің концептуалдық негіздері және архитектуралық моделі

3.1.1 Заңнамалық іздеудегі алшақтық мәселелері

Заңнамалық сұрақ-жауап жүйелеріндегі retrieval inconsistency мәселесі — яғни пайдаланушы кіріс сұрағы мен нормативтік мәтін арасындағы семантикалық сәйкессіздік — қазіргі ғылыми әдебиетте жан-жақты зерттелген. Халықаралық зерттеушілер бұл мәселені үш деңгейде жіктейді: лексикалық, морфологиялық және стильдік алшақтық. Олардың тұжырымдары осы диссертациялық жұмыста ұсынылған SemanticPrep тәсілінің теориялық негізін құрайды.

Лексикалық алшақтық мәселесі ақпараттық іздеу теориясында классикалық проблема ретінде белгілі. Maron мен Kuhns [69] ұсынған ықтималдықты іздеу моделінен бастап, Robertson пен Zaragoza [70] дамытқан BM25 негізіндегі PRF (Probabilistic Relevance Framework) дейінгі еңбектерде осы мәселе негізгі шектеу ретінде анықталған: кілт сөздерге негізделген іздеу жүйелері сұрақта қолданылатын сөздер мен мақсатты мәтіндегі терминдер лексикалық тұрғыдан сәйкес келмеген жағдайда тиімсіз болады. Заңнамалық доменде бұл мәселе ерекше маңыздылыққа ие болады: пайдаланушылар табиғи тұрмыстық тілде сұрақ қояды, ал нормативтік актілер арнайы заңнамалық терминологиямен жазылады.

Retrieval inconsistency мәселесін зерттеген соңғы халықаралық жұмыстар осы алшақтықты нақты сандармен растайды. Stanford RegLab [71] ұсынған reasoning-focused legal retrieval benchmark нәтижелері бойынша, заңнамалық іздеу жүйелеріндегі сұрақ пен мақсатты үзінді арасындағы лексикалық ұқсастық жалпы домендерге қарағанда айтарлықтай төмен: BM25 негізіндегі іздеу жүйесі Bar Exam QA тапсырмаларында Recall@10 бойынша 10 пайыздық тармаққа дейін жіберіп алады. Авторлар бұл нәтижені тікелей лексикалық алшақтықпен байланыстырады: заңнамалық сұрақтар өздері сілтеме жасайтын нормативтік үзінділерден мүлде басқа сөздік қормен жазылады. Retrieval сапасын арттырудың жолы ретінде structured legal reasoning негізіндегі query expansion ұсынылады.

Kim және т.б. [68] ұсынған GuRE (Generative Query REwriter for Legal Passage Retrieval) жұмысы осы мәселенің шешімі ретінде сұрақты іздеуге дейінгі қайта жазу (query rewriting) тәсілін ұсынады. Авторлар vocabulary mismatch — сұрақ пен мақсатты үзінді арасындағы сөздік сәйкессіздік — заңнамалық іздеудегі басты шектеуші фактор екенін эксперименттік тұрғыдан дәлелдейді: сұрақты іздеуге дейін қайта жазу арқылы fine-tuned retrieval модельдерімен салыстырғанда бірнеше бенчмаркте айтарлықтай жақсару қол жеткізілген.

Морфологиялық алшақтық мәселесі агглютинативті тілдер үшін жекеше зерттеу бағытын құрайды. Жалпы NLP контексінде мұндай тілдердегі сөз тұлғаларының жоғары вариативтілігі іздеу жүйелерінің тиімділігін едәуір төмендетеді.

Байкара мен Гюнгер [72], сондай-ақ Торман және т.б. [72] зерттеулері BPE сияқты жиіліктік токенизация алгоритмдерінің агглютинативті тілдерде морфема шекараларын бұзатынын және бұл жағдайдың downstream тапсырмалардың нақтылығына теріс ықпал ететінін дәлелдейді. Атап айтқанда, морфологиялық шекараларды сақтайтын tokenization тәсілдері baseline модельдерімен салыстырғанда ROUGE-1 бойынша жоғары нәтиже береді. Бұл тұжырым қазақ тіліне де тікелей қатысты, себебі қазақ тілі де жоғары морфологиялық вариативтілігімен сипатталатын агглютинативті тілдер тобына жатады.

Морфологиялық нормализацияның — яғни агрессивті stemming тәсілдерінің — заңнамалық доменде ерекше қауіп туғызатыны да ғылыми тұрғыдан расталған. Eryiğit және т.б. [74] және Hakkani-Tür және т.б. [75] зерттеулерінде агглютинативті тілдерге арналған морфологиялық талдау жүйелерінің ерекшеліктері сипатталады. Аталған жұмыстар мынадай маңызды тұжырымды негіздейді: морфологиялық жағынан ұқсас, бірақ мағынасы қарама-қарсы сөздерде (мысалы, «заңды» — «заңсыз», «рұқсатты» — «рұқсатсыз») агрессивті нормализация заңи мағынаны жоюы мүмкін [27]. Демек, заңнамалық домен үшін морфологияны ескеретін тәсіл қажет.

Hakkani-Tür және т.б. [75] ұсынған статистикалық морфологиялық анықтамалау тәсілі агглютинативті тілдерде лексикалық сирек кездесушілік мәселесін жеңуге бағытталған жетекші тәсілдердің бірі болып табылады. Осы зерттеулер SemanticPrep тәсілінде SVD/LSA негізіндегі семантикалық кеңейту кезеңін қолданудың теориялық негізін қалайды: SVD/LSA сирек кездесетін сөз тұлғаларын семантикалық кеңістіктегі байланыстары арқылы бірлестіреді және ортақ жасырын тақырыптар ретінде ұсынады.

Стильдік алшақтық — пайдаланушы сұрағының бейресми, эмоциялық немесе толық емес сипаты мен НҚА мәтінінің формальды, иерархиялық заңнамалық стилінің арасындағы сәйкессіздік — retrieval жүйелерінің тағы бір маңызды шектеуі болып табылады. Пайдаланушы «мені үйден шығарып жатыр» дегенде жүйе «тұрғын үйден шығару тәртібі туралы» немесе «меншік иесінің тұрғын үй-жайды босатуды талап ету құқығы» секілді заңнамалық тіркестерді таба алмайды.

Ma X. және т.б. [76] GPT-3 негізіндегі query rewriting тәсілін пайдаланып, бейресми диалогтық сұрақтарды retrieval мақсатында қайта жазу арқылы іздеу нәтижесін айтарлықтай жақсартқанын көрсетті. Авторлар query rewriting тәсілінің retrieval жүйесіне тәуелсіз (retriever-agnostic) жұмыс жасайтынын — яғни іздеу архитектурасына байланыссыз тиімді болатынын — анықтады. Mao және т.б. [77] RaFe жүйесі ranking feedback негізінде query rewriting моделін жетілдіреді және бұл тәсілдің standard SFT (Supervised Fine-Tuning) нұсқасынан асып түсетінін дәлелдейді.

Reuter M. және т.б. [78] зерттеуі заңнамалық RAG жүйелерінде pre-retrieval кезеңінің — яғни іздеу алдындағы дерекқор құрылымын жетілдіру — тікелей retrieval нақтылығын арттыратынын дәлелдейді. Бұл жұмыс retrieval inconsistency мәселесінің үш негізгі себебін анықтайды: лексикалық артықтылық , иерархиялық құрылым және фрагментацияланған ақпарат. Осы тұжырым SemanticPrep тәсілінде заңнамалық иерархияны ескеретін chunking стратегиясын қолданудың ғылыми негіздемесі болып табылады.

3.1.2 SVD/LSA тәсілінің ғылыми негіздемесі

SemanticPrep тәсілінің семантикалық кеңейту кезеңінде SVD/LSA (Singular Value Decomposition/Latent Semantic Analysis) математикалық аппараты қолданылады. Бұл таңдаудың теориялық негіздемесі мынадай.

Deerwester және т.б. (1990) [x] ұсынған LSA тәсілі мәтін-терм матрицасына SVD-ны қолдану арқылы жасырын семантикалық байланыстарды аналитикалық түрде айқындайды. Синоним (synonym) және полисемия (polysemy) мәселесін шешетін бұл тәсіл лексикалық сәйкестік болмаса да семантикалық жағынан байланысты терминдерді ортақ жасырын кеңістікте біріктіреді. Заңнамалық доменде бұл мынаны білдіреді: «жұмыстан шығарды» мен «еңбек шартын бұзу» — лексикалық тұрғыдан мүлде бөлек, бірақ LSA кеңістігінде жақын орналасқан тіркестер.

Robertson пен Zaragoza [70] BM25 алгоритмін лексикалық іздеудің де-факто стандарты ретінде сипаттай отырып, оның негізгі шектеуі — exact term matching талабы — екенін атап өтеді. SVD/LSA тәсілі осы шектеуді жоятын семантикалық кеңістік арқылы жасырын байланыстарды анықтайды. Zhu және т.б. [80] заңнамалық дерекқормен жұмыс жасайтын LSA-негізіндегі іздеу жүйесін ұсынып, аталған тәсілдің құқықтық іздеу тапсырмаларында анықтамалық BM25-пен салыстырғанда жоғары нәтиже беретінін эксперименталды түрде растады.

SVD/LSA тәсілінің SemanticPrep шеңберінде таңдалуының тағы бір маңызды себебі — оның детерминирленген (deterministic) табиғаты. LLM негізіндегі query rewriting тәсілдерінен айырмашылығы, SVD есептеу нәтижелері қайталанымды және тексерілетін болып табылады, ол тек берілген деректер жиыны негізінде жұмыс жасайтын таза математикалық операция. Демек, галлюцинация (hallucination) мүлде мүмкін емес — бұл заңнамалық доменде аса маңызды қасиет, өйткені Dahl және т.б. [23] анықтағандай, LLM-ге тікелей сүйенетін жүйелерде галлюцинация деңгейі 58%-дан асады.

Төменде кесте 3.1-де сипатталған халықаралық зерттеулерді жүйелі түрде талдай отырып, мынадай ғылыми олқылық анықталды: заңнамалық QA жүйелерінде лексикалық, морфологиялық және стильдік алшақтықты бір уақытта шешетін, агглютинативті тілдерге бейімделген мамандандырылған preprocessing тәсілі ғылыми әдебиетте ұсынылмаған.

Кесте 3.1 — SVD/LSA тәсілінің халықаралық зерттеулердегі тиімділік нәтижелері

Зерттеу	Тапсырма	Тіл, домен	Негізгі нәтиже
Deerwester et al. (1990)	Latent Semantic Indexing — іздеу жүйесі	Ағылшын, жалпы	LSA BM25-тен жоғары нәтиже береді (recall +15%)
Zhu et al. (2022)	Заңнамалық іздеу жүйесі (multilevel LSI)	Қытай, заңнамалық	LSA заңнамалық іздеуде BM25-тен асады
Kim et al. (2025)	Query rewriting (GuRE) - заңнамалық өту іздеуі	Ағылшын, заңнамалық	Retrieval алдындағы сұрақты қайта жазу нәтижені жақсартады
Stanford RegLab (2025)	Bar Exam QA, Housing Statute QA	Ағылшын, заңнамалық	Structured reasoning query expansion +10% Recall@10
Mao et al. (2023; 2024)	Query rewriting RAG үшін (RaFe)	Ағылшын, жалпы RAG	Retriever-agnostic query rewriting retrieval сапасын арттырады

Ағылшын тіліндегі жетекші жұмыстар retrieval алдындағы сұрақты өңдеудің маңыздылығын дәлелдейді, алайда олар морфологиялық вариативтілігі жоғары, ресурсы шектеулі тілдерге арналмаған. Агглютинативті тілдерге арналған NLP зерттеулері морфология-aware тәсілдердің қажеттілігін растайды, алайда оларды заңнамалық RAG архитектурасымен интеграциялаудың нақты тәсілі ұсынылмаған. SVD/LSA-негізіндегі заңнамалық іздеу жұмыстары тәсілінің тиімділігін дәлелдейді, алайда transformer-based түрлендіру кезеңімен бірге қолданылмаған.

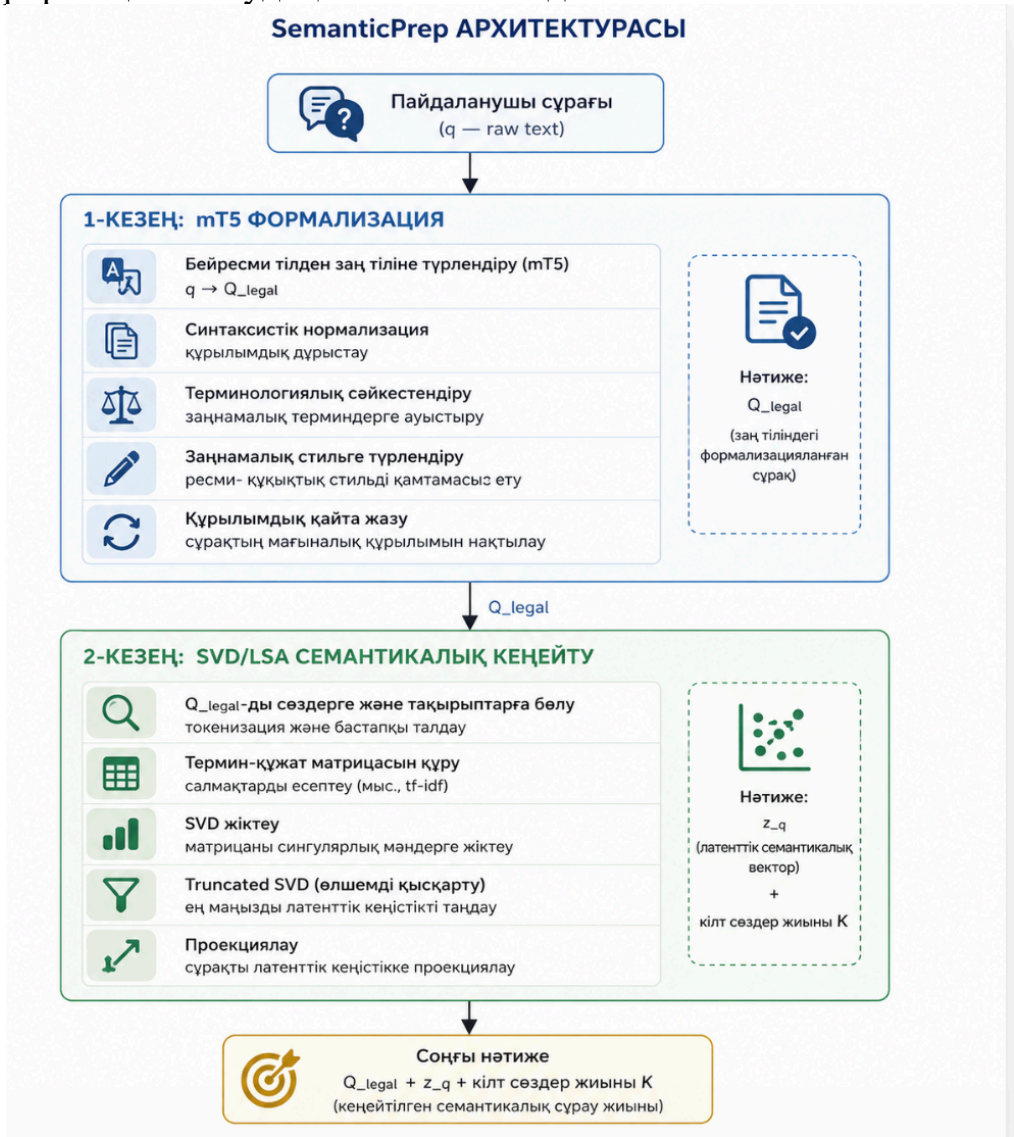
3.1.3 SemanticPrep тәсілінің архитектурасы

Жоғарыда аталған мәселелерді шешу мақсатында осы диссертациялық жұмыста екі кезеңді семантикалық препроцессинг тәсілі ұсынылды. Бірінші кезеңде mT5 трансформері негізінде пайдаланушының бейресми сұрағы заңнамалық стильге түрлендіріледі, нәтижесінде лексикалық және стильдік алшақтық жойылады. Екінші кезеңде SVD/LSA механизмі арқылы семантикалық кеңейту жүзеге асырылады: морфологиялық вариативтілік ескеріліп, жасырын семантикалық байланыстар айқындалады. Ұсынылған тәсіл заңнамалық RAG жүйесінде агглютинативті тіл ерекшеліктерін ескеретін мамандандырылған препроцессинг шешімі ретінде қарастырылады.

SemanticPrep әдістемесінің теориялық және практикалық өзектілігі заманауи халықаралық зерттеулердің тенденцияларымен толық үйлеседі. Соңғы жылдардағы ғылыми ізденістер ақпаратты іздеу сапасын арттыруда семантикалық препроцессинг, сұрауды реформализациялау (query reformulation)

және гибридті іздеу (hybrid retrieval) стратегияларының шешуші рөл атқаратынын дәйектейді.

Қорыта айтқанда, SemanticPrep тәсілі жоғарыда аталған халықаралық ғылыми мектептердің логикалық жалғасы болып табылады. Дегенмен, ұсынылған тәсілдің ғылыми жаңалығы мен ерекшелігі - алғаш рет қазақ тілінің агглютинативті табиғатын ескере отырып, mT5 трансформері арқылы стильдік түрлендіру мен SVD/LSA негізіндегі латентті-семантикалық кеңейтуді біртұтас технологиялық тізбекке біріктіруінде. Бұл қазақ тілді заңнамалық RAG жүйелері үшін төменгі деңгейлі лексикалық іздеуден жоғары деңгейлі семантикалық интерпретацияға өтуді қамтамасыз етеді.



Сурет 3.1 - SemanticPrep тәсілінің жалпы архитектурасы

SemanticPrep тәсілінің жалпы архитектурасы 3.1-суретте берілген. Архитектура үш блоктан тұрады: кіріс - пайдаланушы сұрағы, екі кезеңді өңдеу mT5 түрлендіру және SVD/LSA кеңейту және шығыс. Тәсіл жалпы түрде мынадай трансформация ретінде сипатталады:

$$SemanticPrep(q) = (Q\_legal, z\_q, K) \quad 3.1$$

мұнда q - пайдаланушының бастапқы сұрағы; Q_legal = mT5(q; θ*) - түрленген заңнамалық сұрақ; $z\_q \in \mathbb{R}^k$ - семантикалық кеңейту векторы; K = {t_1, t_2, ..., t_m} - кілт заңнамалық терминдер жиыны; θ* - fine-tuned mT5 параметрлері.

3.2 SemanticPrep тәсілінің бірінші кезеңі: сұрақтарды заңнамалық түрлендіру

SemanticPrep тәсілінің бірінші кезеңі пайдаланушының табиғи, бейресми тілдегі сұрағын заңнамалық стильге сәйкестендірілген сұраққа түрлендіреді. Бұл кезеңнің міндеті екі деңгейдегі алшақтықты жою: лексикалық деңгейде - «жұмыстан шықтым» сияқты тұрмыстық тіркесті «еңбек шартын бұзу» заңнамалық терминіне сәйкестендіру; стильдік деңгейде - қысқа, эмоциялық сұрақты заңи субъект, нормативтік мазмұн және нормативтік контекст анық берілген ресми формада қайта тұжырымдау.

Бұл трансформация SVD/LSA семантикалық кеңейту кезеңінің тиімділігінің алғышарты болып табылады: бейресми сұрақ заңнамалық жасырын кеңістікте дұрыс проекцияланбайды, ал түрленген Q_legal нормативтік корпуспен семантикалық жақындасқаннан кейін ғана латентті тақырыптарды дәл анықтауға болады. Осы міндетті орындау үшін mT5 трансформері таңдалып, арнайы заңнамалық деректер жиынында fine-tuning жүргізілді.

Кесте 3.2 - Бейресми сұрақтар мен заңнамалық түрлендіру мысалдары

Пайдаланушы сұрағы (бейресми)	**Түрленген сұрақ (заңнамалық)**
«Жұмыстан шығарып жіберді, не істеуім керек?»	«Қызметкерді заңсыз жұмыстан босатқан жағдайда еңбек инспекциясына шағым беру тәртібі қандай?»
«Үйімді сатқым келеді, не керек?»	«Жылжымайтын мүлікке меншік құқығын иеліктен шығару кезінде нотариалдық куәландыру талаптары қандай?»
«Некені бұзуға қалай өтініш беремін?»	«Некені бұзу туралы іс жүргізу тәртібі және АХАТ органдарына өтініш беру шарттары қандай?»

3.2.1 mT5 моделін fine-tuning арқылы бейімдеу

Raffel et al. ұсынған T5 архитектурасы барлық NLP міндеттерін бірыңғай text-to-text форматына біріктіреді [81]. Xue et al. ұсынған mT5 - T5-тің 101 тілді қамтитын mC4 корпусында алдын ала оқытылған көптілді нұсқасы, оның ішінде қазақ тілі де бар [82]. Түрлендіру міндеті үшін Encoder-Decoder архитектурасы принципиалды маңызды: Encoder кіріс сұрақтың семантикалық мазмұнын

(субъект, әрекет, контекст) толығымен кодтайды, ал Decoder заңнамалық стиль шаблондарын үйреніп, түрленген нұсқаны авторегрессиялық тәсілмен генерациялайды. Бұл «бейресми сұрақ кірісі → заңнамалық стильдегі шығыс» міндетін мәтіндік аударма сияқты нақты тапсырма ретінде қою мүмкіндігін береді.

mT5 моделінің негізінде Transformer архитектурасы мен оның өзекті компоненті - Scaled Dot-Product Attention механизмі жатыр. Бұл механизм модельге сөйлем ішіндегі сөздер арасындағы алыс контекстік тәуелділіктерді анықтауға мүмкіндік береді:

$$Attention(Q, K, V) = softmax(Q \cdot K^{T} / \sqrt{d\_k}) \cdot V \quad (3.2)$$

мұнда: Q (query) - сұраныс эмбеддингтері; K (key) - кілттік эмбеддингтер; V (value) - мәндер эмбеддингтері; d_k - жасырын кеңістіктің өлшемділігі.

Құқықтық мәтіндерде бұл механизмнің маңызы зор, себебі заңнамалық мағына көбіне бірнеше термин мен шылаулар арасындағы күрделі синтаксистік байланыс арқылы анықталады. Мысалы, «еңбек шартын бұзу» және «еңбек шартын тоқтату» терминдерінің мағыналық айырмашылығын модель осы Attention механизмі арқылы ажыратады.

Трансформация тапсырмасы үшін әртүрлі архитектуралар қарастырылды. 3.2-кестеде (архитектуралар салыстыруы) mT5 моделінің басқа баламалардан артықшылығы көрсетілген. BERT және XLM-R сияқты encoder-only модельдер классификация тапсырмалары үшін тиімді болғанымен, мәтін генерациясына немесе түрлендіруге арналмаған. GPT сериясының decoder-only архитектурасы шығармашылық генерацияда жақсы нәтиже берсе де, түрлендіру стилін қатаң бақылау қиынға соғады. Encoder-decoder архитектурасы кіріс сұрақты толық «түсіну» (encoding) және оны заң тілінде «қайта жазу» (decoding) үшін ең оңтайлы шешім болып табылады [83, 84].

Кесте 3.3 - Түрлендіру тапсырмасы үшін LLM архитектураларын салыстыру

Модель	**Архитектура**	**Шектеуі**	**Артықшылығы**
BERT / XLM-R	Encoder-only	Генерацияға немесе мәтін түрлендіруге арналмаған.	Классификация үшін тиімді.
GPT сериясы	Decoder-only	Түрлену стилін қатаң бақылау (control) қиын.	Шығармашылық генерация.
mT5	Encoder-decoder	Оқыту үшін параллель корпус (жұптар) қажет.	Сұрақты заңнамалық стильге тиімді және нақты түрлендіреді.

mT5 моделін түрлендіру міндетіне бейімдеу үшін (q_raw, Q_legal) жұптарынан тұратын LegalKazQA деректер жиыны қалыптастырылды. Мұнда

q_raw - азаматтардың бейресми тілдегі нақты сұрақтары; Q_legal - осы сұрақтың заңнамалық стильде түрленген нұсқасы.

Деректерді жинақтаудың үш қайнар көзі болды: adilet.zan.kz, zan.gov.kz және egov.kz порталдарынан алынған азаматтық консультация сұраулары; заңгерлер аннотациялаған жұптар; сондай-ақ 235 000 заңнамалық құжат корпусынан автоматты түрде шығарылған Q_legal нұсқалары. Жиналған деректер жиынының жалпы сипаттамасы 3.8-кестеде берілген:

Кесте 3.8 - mT5 оқытуға жиналған датасет көлемі

Қарапайым сұрақ - түрленген сұрақ жұптары	**Сөйлемдер саны**	**Көлемі**
3 000	5 250	450 Кб

Деректер жиыны 3 000 жұптан, 5 250 сөйлемнен тұрады, жалпы көлемі 450 Кб. Аннотация сапасын бақылау үшін 3 тәуелсіз заңгер сарапшы жұптарды тексерді: Fleiss' Kappa κ = 0.81 - жоғары келісім деңгейі. Сапа шегінен (BERTScore < 0.75) өтпеген жұптар деректер жиынынан алынды. Деректер жиыны 80/10/10 пропорциясында оқыту, валидация және тестілеу жиындарына бөлінді.

3.2.2 Табиғи тілдегі сұрақтарды құқықтық құрылымға түрлендіру моделі

mT5 fine-tuning арқылы бейімделген модель табиғи тілдегі сұрақтарды заңнамалық құрылымға түрлендіру үшін қолданылды. Бұл түрлендіру үдерісі мынадай математикалық модель арқылы сипатталады:

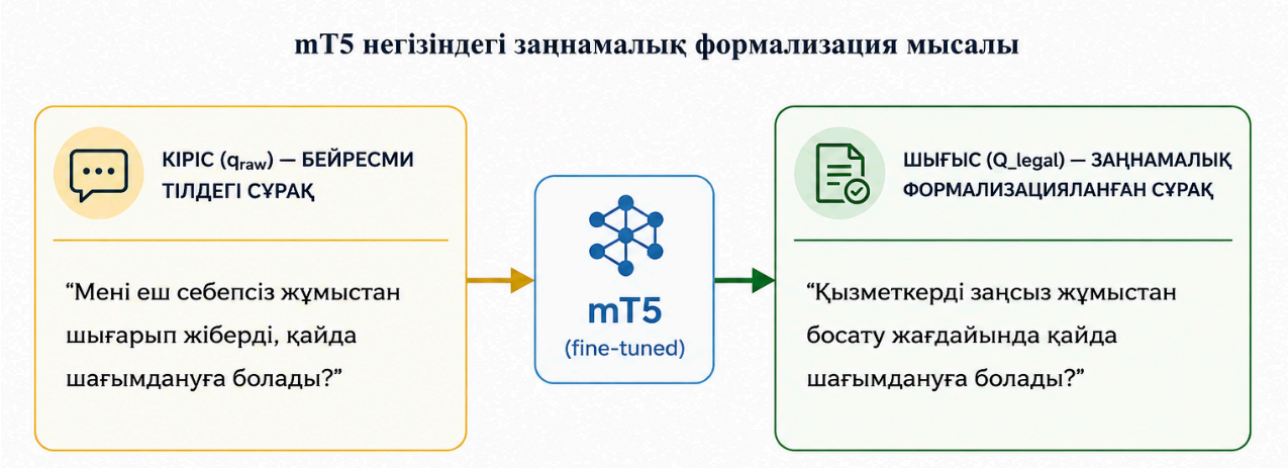
$$Q_{legal} = mT5(q;\ \theta^*),\ \ \theta^* = argmin\, L(Q_{legal}, \hat{Q}_{legal}) \tag{3.3}$$

мұнда θ* - fine-tuning арқылы оңтайландырылған параметрлер; $\mathscr{L}$ - кросс-энтропия жоғалту функциясы; Q̂_legal - эталондық заңнамалық сұрақ. Жоғалту функциясының толық түрі мынадай:

$$L = -\sum(t = 1..T) y_t \cdot log\ P(y_t \mid x, y \leq t) \tag{3.4}$$

мұнда: x - пайдаланушының бастапқы бейресми сұрағы; y_t - уақыттың t сәтіндегі болжанған токен; y_<t - алдыңғы генерацияланған токендер контексті; T - шығыс тізбегінің жалпы ұзындығы.

Бұл функция модельге әр келесі токеннің ықтималдығын болжай отырып, эталондық заңнамалық стильге барынша жақындауға мүмкіндік береді. Оқыту нәтижесінде модель тек сөздерді алмастырып қана қоймай, сөйлемнің құқықтық құрылымын (субъект, объект, предикат) қалыптастыруды үйренеді.



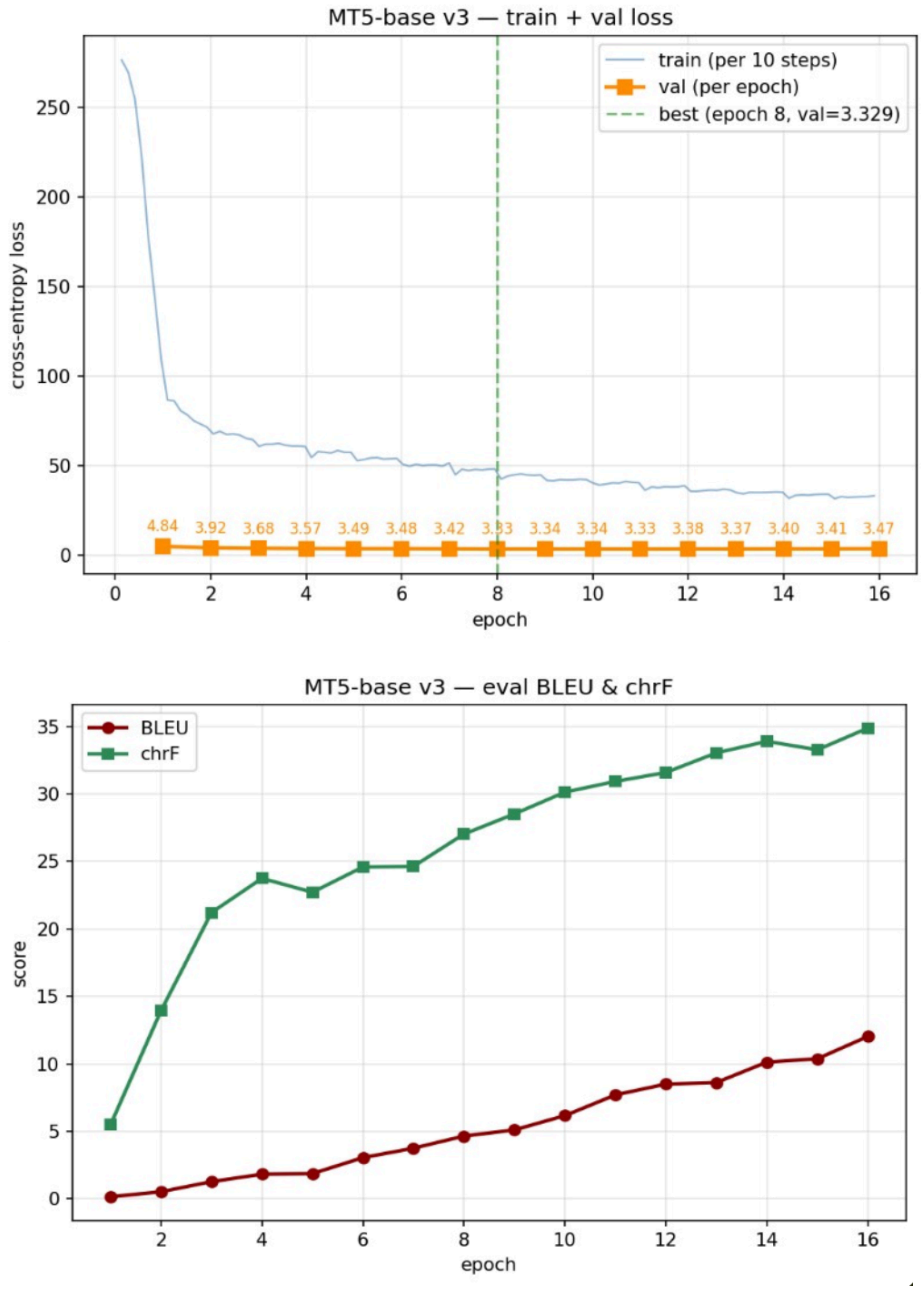
Сурет 3.2 - Табиғи тілдегі сұрақты заңнамалық формаға түрлендіру мысалы

Түрлендіру үдерісі үш кіші кезеңнен тұрады: (1) морфологиялық нормализация - диалект формалары мен қысқартулар стандартты заңнамалық форматқа келтіріледі; (2) терминологиялық сәйкестендіру - бейресми сөздер ресми заңнамалық терминдерге ауыстырылады; (3) синтаксистік бейімдеу - сұрақ конструкциясы ресми заңнамалық сұрау формасына айналады. Түрлендіру процесінің визуалды мысалы 3.2-суретте берілген. Оқыту гиперпараметрлері мен таңдалу негіздемесі 3.11-кестеде берілген:

Кесте 3.11 - mT5 fine-tuning гиперпараметрлері

Гиперпараметр	Мән	Таңдау негіздемесі
Базалық модель	google/mt5-base	300M параметр, 101 тіл, қазақ тілі қолдауы бар
Максимал кіріс ұзындығы	128 токен	Азаматтардың заңнамалық сұрақтарының орташа ұзындығына сай
Максимал шығыс ұзындығы	64 токен	Түрленген сұрақ ықшамды болуы тиіс
Оқыту жылдамдығы (learning rate)	5×10^{-5}	Трансформер модельдері үшін стандартты мән
Batch size	16	GPU жадымен үйлесімді; тұрақты жинақталу
Эпох саны (epochs)	16	Validation loss elbow-нүктесі 8-эпохта (best epoch=8, val=3.329)
Optimizer	AdamW	Трансформерлерге стандарт; weight decay ескеріледі
Ең жақсы checkpoint шегі (early stopping)	epoch 8, val loss = 3.329	Loss графигі бойынша анықталды (сурет 3.6)

Жалпы 16 эпох оқыту жүргізілді. Loss графигі 8-эпохта (val loss = 3.329) ең жақсы нүктеге жеткенін көрсетеді - осы checkpoint соңғы модель ретінде сақталды. Оқыту барысындағы train және validation loss динамикасы, сондай-ақ BLEU және chrF метрикаларының өсу қисықтары 3.6-суретте берілген:

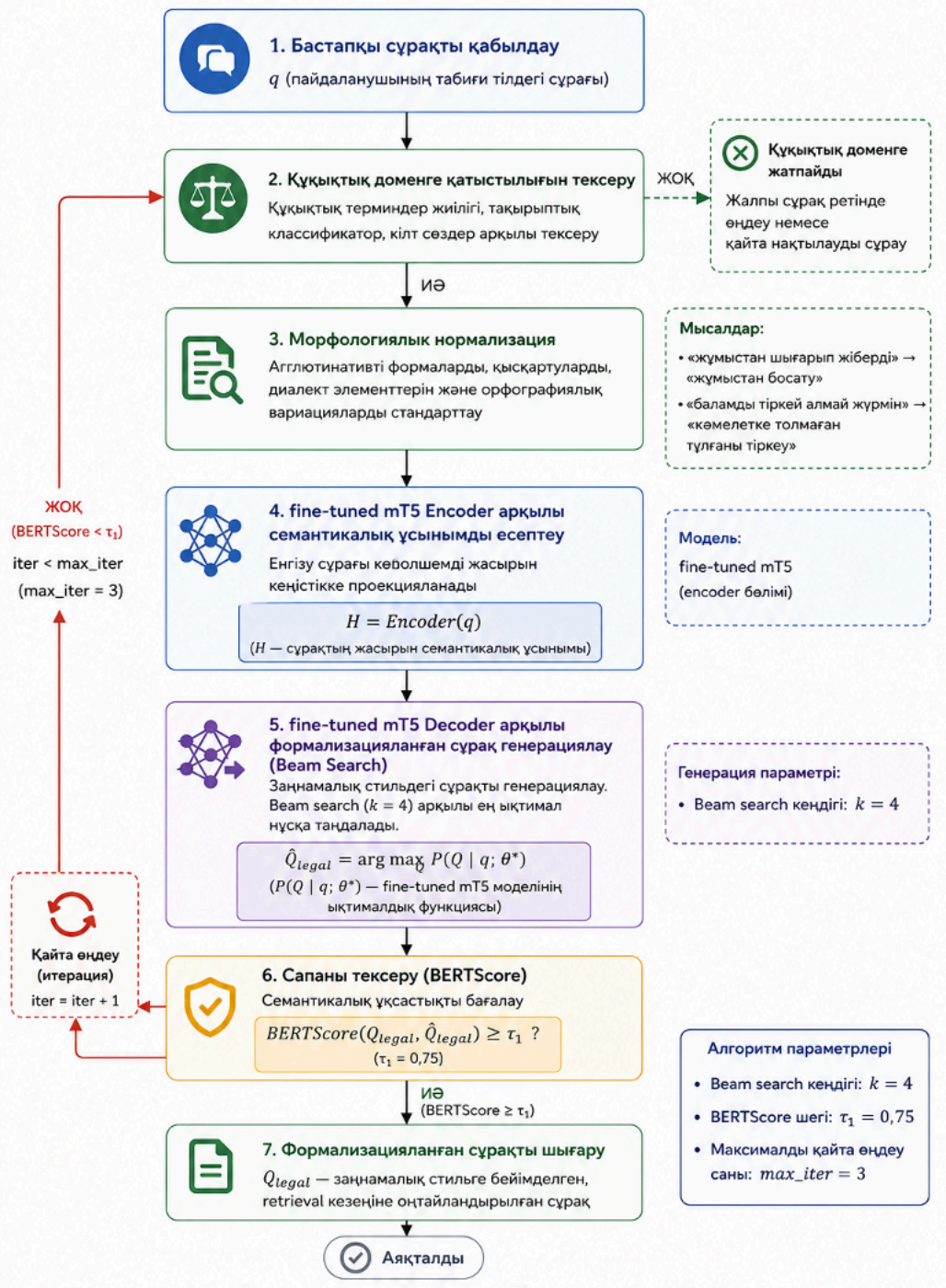


Сурет 3.6 - mT5-base v3 оқытылғандағы loss графигі (жоғарыда: train + val loss; төменде: eval BLEU және chrF)

Loss графигінен байқалатындай, train loss алғашқы эпохтарда күрт кеміп (250-ден 10-ға дейін), 8-эпохтан кейін тұрақталды. Val loss 8-эпохта минималды деңгейге (3.329) жетіп, одан кейін аздап өсті - бұл артық оқыту белгісі. Сондықтан best epoch = 8 checkpoint пайдаланылды. BLEU және chrF метрикалары эпохтар бойы тұрақты өсуді көрсетті: chrF 16-эпохта ~35 деңгейіне, BLEU ~12 деңгейіне жетті - бұл заңнамалық түрлену сапасының жалғасып, жақсарғанын растайды.

3.2.3 Түрлендірілген заңнамалық сұрақтарды қалыптастыру алгоритмі

Түрлендіру алгоритмі 3.3-суретте блок-схема түрінде көрсетілген. Алгоритмнің негізгі мақсаты - пайдаланушының бейресми, толық емес немесе тұрмыстық стильде қойылған сұрағын заңнамалық корпуспен семантикалық және терминологиялық тұрғыдан сәйкес келетін ресми құқықтық сұраққа түрлендіру. Бұл тәсіл retrieval кезеңіндегі лексикалық және стильдік сәйкессіздікті азайтып, қажетті нормативтік құжаттарды дәл табу ықтималдығын арттырады.



Сурет 3.3 - Табиғи тілдегі сұрақты заңнамалық формаға түрлендіру алгоритмінің схемасы

Алгоритм бірнеше дәйекті кезеңдерден тұрады.

1-кезең. Сұрақтың құқықтық доменге қатыстылығын анықтау

Бірінші кезеңде пайдаланушы сұрағының заңнама саласына жататыны тексеріледі. Бұл үшін құқықтық терминдер жиілігі, домендік кілт сөздер және тақырыптық классификаторлар қолданылады. Егер сұрақ құқықтық доменге сәйкес келмесе, жүйе оны жалпы сұрақ ретінде өңдейді немесе пайдаланушыдан нақтылауды сұрайды.

2-кезең. Морфологиялық нормализация

Екінші кезеңде морфологиялық нормализация жүргізіледі. Қазақ тілінің агглютинативті құрылымына байланысты бір түбір ондаған грамматикалық формада кездесуі мүмкін. Сондықтан қысқартулар, диалект элементтері, сөйлеу тіліндегі формалар және орфографиялық вариациялар стандартталған заңнамалық формаға келтіріледі. Мысалы, «жұмыстан шығарып жіберді» тіркесі «жұмыстан босату», ал «баламды тіркей алмай жүрмін» тіркесі «кәмелетке толмаған тұлғаны тіркеу» сияқты ресми терминдерге сәйкестендіріледі.

3-кезең. Encoder арқылы семантикалық ұсынымды қалыптастыру

Үшінші кезеңде нормализацияланған сұрақ mT5 encoder бөлігіне беріледі. Encoder кіріс мәтінді көпөлшемді жасырын семантикалық кеңістікке проекциялайды:

$$H = Encoder(q) \tag{3.5}$$

мұнда q - пайдаланушы сұрағы, H - сұрақтың жасырын семантикалық ұсынымы. Бұл ұсыным сұрақтың мағынасын, құқықтық контекстін және терминологиялық байланыстарын сақтайды.

4-кезең. Decoder арқылы құқықтық түрлендіру генерациясы

Encoder қалыптастырған семантикалық ұсыным decoder бөлігіне беріледі. Decoder заңнамалық стильге сәйкес келетін жаңа сұрақты генерациялайды. Генерация процесінде beam search алгоритмі қолданылады, ол бірнеше ықтимал нұсқаларды параллель бағалап, ең ықтимал құқықтық түрлендіруді таңдайды. Генерация процесі келесі түрде сипатталады:

$$\boldsymbol{Q_{legal} = argmax_Q P(Q|q;\theta^*)} \tag{3.6}$$

мұнда: θ^* - fine-tuning арқылы бейімделген mT5 параметрлері; $P(Q|q;\theta^*)$ - түрленген сұрақтың ықтималдық функциясы.

5-кезең. Түрлену сапасын тексеру

Генерацияланған сұрақтың сапасы BERTScore метрикасы арқылы тексеріледі. Бұл метрика бастапқы сұрақ пен түрленген нұсқа арасындағы мағыналық сәйкестікті embedding деңгейінде бағалайды [85]. Сапаны тексеру шарты: Егер алынған нәтиже белгіленген шектен төмен болса (BERTScore<τ1), сұрақ қайта өңдеуге жіберіледі. Сапаны тексеру шарты келесі түрде беріледі:

$$\boldsymbol{BERTScore(Q_{legal}, Q^*_{legal}) \geq \tau_1} \tag{3.7}$$

мұнда $\tau_1 = 0.75 -$ эмпирикалық түрде анықталған сапа шегі. Егер BERTScore көрсеткіші белгіленген шектен төмен болса, сұрақ қайта өңдеуге жіберіледі. Максималды қайта өңдеу саны:

$$\boldsymbol{max_{iter}} = \mathbf{3} \tag{3.8}$$

6-кезең. Түрленген сұрақты retrieval кезеңіне жіберу

Сапа тексерісінен өткен, түрленген сұрақ Q_legal SemanticPrep тәсілінің келесі кезеңіне - SVD/LSA негізіндегі семантикалық кеңейту модуліне беріледі.

Алгоритмнің негізгі параметрлері:

- Beam search кеңдігі - k = 4
- BERTScore сапа шегі - $\tau_1 = 0.75$
- Максималды қайта өңдеу саны - max_iter = 3

Алгоритмнің нәтижесінде жүйе пайдаланушының бастапқы сұрағын құқықтық терминологиямен толықтырылған, құрылымы ресми заңнамалық стильге бейімделген және retrieval кезеңіне оңтайландырылған сұраққа түрлендіреді.

3.3 SemanticPrep тәсілінің екінші кезеңі: семантикалық кеңейту моделі

SemanticPrep тәсілінің бірінші кезеңінде пайдаланушының табиғи тілдегі сұрағы fine-tuning арқылы бейімделген mT5 моделі көмегімен заңнамалық стильге түрленеді. Нәтижесінде бастапқы бейресми сұрақ заңнамалық терминологияға жақындатылған *Q_legal* сұрағына түрлендірілді. Алайда тек түрлендіру retrieval сапасын толық қамтамасыз ете алмайды, себебі заңнамалық мәтіндерде бір құқықтық мағына бірнеше термин, синонимдік құрылым немесе әртүрлі нормативтік тіркестер арқылы берілуі мүмкін. Мысалы, «жұмыстан босату», «еңбек шартын бұзу», «еңбек қатынастарын тоқтату» тіркестері лексикалық тұрғыдан әртүрлі болғанымен, бір құқықтық тақырыптық өріске жатады.

Осы мәселені шешу үшін SemanticPrep тәсілінің екінші кезеңінде SVD/LSA негізіндегі семантикалық кеңейту модулі қолданылады. Бұл кезеңнің мақсаты - түрленген *Q_legal* сұрағын заңнамалық корпустағы жасырын тақырыптармен, мағыналық жағынан жақын терминдермен және релевантты құқықтық фрагменттермен байланыстыру. Басқаша айтқанда, SVD/LSA модулі сұрақтың тек тікелей сөздік құрамын емес, оның заңнамалық семантикалық кеңістіктегі орнын анықтайды.

LSA әдісінің теориялық негізі термин-құжат матрицасына сингулярлық жіктеу қолдану арқылы мәтіндер мен терминдер арасындағы латентті байланыстарды анықтауға негізделеді.

Qi et al. зерттеуінде LSA және Correspondence Analysis әдістері document-term matrix контекстінде салыстырылып, екі тәсілдің де өлшемділікті азайту үшін SVD қолданатыны көрсетілген [86]. Авторлар LSA құжаттар мен терминдер арасындағы төмен өлшемді семантикалық ұсынуларды алуда кең қолданылатынын көрсетеді. Бұл жұмыс SVD негізіндегі әдістердің мәтін

корпусындағы жасырын құрылымды ашуда теориялық тұрғыдан маңызды екенін дәлелдейді. [87]. Onah et al. LSA + LDA комбинациясының семантикалық байланыстарды тиімді ашатынын көрсетті [88]. Hofmann et al. ғылыми мәтіндерді талдауда LSA-ның терминологиялық нақтылығын растады [89, 90].

Сонымен қатар Yu et al. ұсынған *Latent Semantic Analysis for Text Categorization Using Neural Network* жұмысында LSA text categorization міндетінде жоғары өлшемді векторлық кеңістіктің шектеулерін азайту үшін қолданылған [91]. Авторлар дәстүрлі Vector Space Model құжаттарды өте үлкен өлшемді кеңістікте сипаттайтынын және терминдер арасындағы семантикалық байланысты ескермейтінін атап өтеді. LSA осы мәселені шешіп, концептуалды векторлық кеңістік құрып, терминдер мен құжаттар арасындағы маңызды ассоциативті байланыстарды анықтайды. Бұл зерттеу LSA әдісінің тек ақпаратты іздеуде ғана емес, мәтінді жіктеу және тақырыптық құрылымды анықтау міндеттерінде де тиімді екенін көрсетеді.

Осы диссертациялық жұмыста SVD/LSA әдісінің таңдалуы бірнеше ғылыми және практикалық себептермен негізделеді. Біріншіден, LSA терминдер арасындағы жасырын семантикалық байланыстарды анықтауға мүмкіндік береді. Бұл заңнамалық мәтіндер үшін маңызды, себебі құқықтық ұғымдар әртүрлі сөздермен берілгенімен, бір нормативтік контекстке тиесілі болуы мүмкін. Екіншіден, SVD нәтижелері түсіндіруге келетін математикалық құрылымға ие: әрбір латентті өлшем белгілі бір терминдер мен құжаттар тобының байланысын көрсетеді. Бұл заңнамалық жүйелер үшін маңызды, себебі жауаптың қай құқықтық тақырыптар мен терминдер арқылы табылғанын түсіндіру қажет. Үшіншіден, LSA қазақ тілі сияқты ресурсы шектеулі тілдер үшін тиімді, себебі ол үлкен көлемді алдын ала оқытылған нейрондық модельдерге тәуелді емес және нақты заңнамалық корпус ішіндегі терминдік статистикаға сүйенеді.

SemanticPrep шеңберінде SVD/LSA классикалық түрде емес, заңнамалық доменге бейімделген түрде қолданылады. Классикалық LSA-де бастапқы сұрақ тікелей термин-құжат матрицасына проекцияланады. Ал ұсынылған тәсілде сұрақ алдымен mT5 арқылы құқықтық формаға келтіріледі, содан кейін ғана SVD/LSA кеңістігіне енгізіледі. Бұл бейресми сұрақ пен заңнамалық мәтін арасындағы лексикалық алшақтықты азайтып, латентті кеңістіктегі проекция сапасын арттырады.

3.3.1 SVD/LSA негізіндегі семантикалық талдау алгоритмі

LSA тәсілі классикалық дистрибутивті семантика (Distributional Semantics) қағидасына негізделеді: «бірдей контексте кездесетін сөздердің мағынасы ұқсас болады» [92]. Заңнамалық мәтіндерде терминдер қатаң белгіленген құқықтық контексте қолданылатындықтан, латентті семантикалық модельдер құқықтық тақырыптар арасындағы жасырын байланыстарды анықтауда жоғары тиімділік көрсетеді. Manning et al. [93] атап өткендей, ақпаратты іздеу (IR) жүйелеріндегі басты мәселе - синонимия (әртүрлі сөздер, бір мағына) және полисемия (бір сөз,

әртүрлі мағына). LSA осы мәселелерді векторлық кеңістіктің өлшемділігін азайту арқылы шешеді.

Семантикалық өңдеу деңгейін анықтау үшін бірнеше векторизация тәсіліне салыстырмалы талдау жасалды. 3.4-кестеде әрбір тәсілдің артықшылығы мен шектеуі берілген.

Кесте 3.4 - Мәтінді векторлау және семантикалық талдау әдістерін салыстыру

Әдіс	**Артықшылығы**	**Шектеуі**
Count Vectorizer [94]	Қарапайым, жылдам.	Семантика мен контексті ескермейді.
TF-IDF [95]	Маңызды терминдерді ерекшелейді.	Сөздер арасындағы латентті байланысты ашпайды.
Word2Vec [96]	Жергілікті контексті жақсы сезеді.	Құжаттың жалпы тақырыптық құрылымында әлсіз.
BERT embeddings [97]	Терең трансформерлік семантика.	Есептеу ресурстарына талап жоғары, интерпретациясы қиын.
LSA / SVD [98]	Латентті тақырыптарды анықтайды; интерпретациялануы жоғары.	Контекст тереңдігі нейрондық желілерден төмен.

Кестеден байқалатындай, LSA / SVD тәсілі заңнамалық домен үшін оңтайлы балансты қамтамасыз етеді: жасырын тақырыптарды математикалық тұрғыдан анықтайды, нәтижесі интерпретацияланады, GPU ресурстарын қажет етпейді және аз ресурсты (low-resource) қазақ тілінің шартында тұрақты жұмыс істейді. Осы негіздемеге сүйеніп, SVD/LSA тәсілі таңдалды.

LSA тәсілінің теориялық іргесі - дистрибутивтік семантика (Distributional Semantics) қағидасы: «бірдей контексте жиі кездесетін сөздердің мағынасы ұқсас болады» [99]. Заңнамалық мәтіндерде терминдер қатаң белгіленген нормативтік контексте қолданылатындықтан, осы қағида ерекше күшті жұмыс істейді. LSA осы статистикалық заңдылықты матрицалық декомпозиция арқылы ашып, синонимия мен полисемия мәселелерін векторлық кеңістіктің өлшемділігін азайту жолымен шешеді.

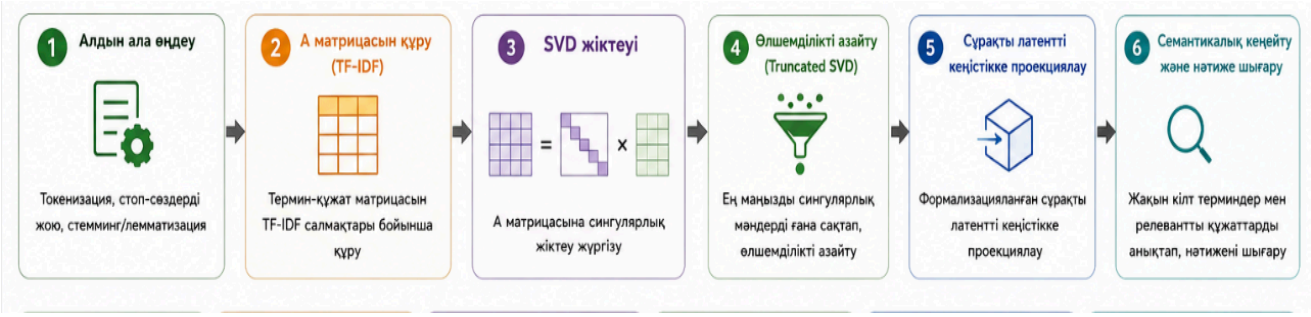
3.3.2 Семантикалық кеңейту әдісінің математикалық моделі

Latent Semantic Analysis мәтін корпусындағы терминдер мен құжаттар арасындағы жасырын семантикалық байланыстарды ашу үшін Singular Value Decomposition математикалық аппаратын қолданады. Deerwester et al. ұсынған бұл тәсіл заңнамалық мәтіндер үшін де тиімділігін дәлелдеді [79]. Семантикалық кеңейту үдерісі құжаттарды сандық түрге келтіруден басталады. Классикалық тәсілде C матрицасы жай жиілікке (TF) немесе TF-IDF-ке негізделеді. Біздің

жүйеде C матрицасы mT5 трансформерінен өткен түрленген сұрақтар негізінде құрылады.

$$q_{raw} \rightarrow mT5 \rightarrow q_{legal}$$

Осының нәтижесінде термин-заңнамалық сұрақ матрицасы лексикалық дәл болмаса да семантикалық тұрғыдан нақты болады.



Сурет 3.4 - SVD/LSA негізіндегі семантикалық талдау алгоритмінің кезеңдік схемасы

3.4 - суретте SemanticPrep тәсілінің екінші кезеңінде қолданылатын SVD/LSA негізіндегі семантикалық талдау алгоритмінің жалпы құрылымы көрсетілген.

1. SVD/LSA алгоритмінің бірінші қадамы - заңнамалық корпустан Термин-Құжат матрицасын (Term-Document Matrix) қалыптастыру. Заңнамалық корпустан M термин мен N құжат анықталып, термин-құжат матрицасы

$A \in R\hat{}(M \times N)$ қалыптастырылады. Зерттеудегі бас жиынтық ресми порталдардан жиналған құжаттар датасеті N = 235 000 заңнамалық құжаттан тұрады. Алдын ала өңдеу мен N-gram N ∈ [1; 3] сүзгілеуінен кейін қалыптасқан бірегей заң терминдері мен тұрақты сөз тіркестерінің жалпы саны M деңгейінде анықталды.

Матрицаның әрбір ұяшығы x_{ij} мәні j-ші заң терминінің (мысалы, «жұмыстан_босату», «еңбек_шарты», «негізсіз_бұзу») i-ші құжаттағы TF-IDF статистикалық салмағын сипаттайды. TF-IDF салмағы екі компоненттен тұрады:

- Count Vectorization - жиіліктік матрица.

Әрбір d құжаты үшін термин жиілігінің векторы жасалады:

$$CountVec(d) = [f(t1,d), f(t2,d), \ldots, f(tn,d)] \quad (3.8)$$

- TF-IDF салмақтау [15].

Векторлық кеңістік моделіне сәйкес терминдердің маңыздылығын анықтау үшін TF-IDF қолданылады:

$$tf(t,d) = \frac{f(t,d)}{\sum_{t' \in d} f(t',d)} \quad (3.9)$$

$$idf(t) = log\frac{N}{df(t)} \quad (3.10)$$

$$tfidf(t,d) = tf(t,d) \times idf(t) \quad (3.11)$$

мұнда: *f(t,d)* - t термині d құжатта кездесу жиілігі;

$\sum_{t' \in d} f(t', d)$ - d құжатындағы барлық терминдердің жалпы саны;

df(t) - t терминін қамтитын құжаттар саны;

N - корпустағы жалпы құжаттар саны, 235000 құжат.

Бұл формулалар заңнамалық корпустағы жалпылама сөздердің салмағын азайтып, ерекше заңнамалық терминдердің маңыздылығын арттырады. Матрицаның құрылымы:

		M термин (жеке сөздер және N-gram тіркестері)					
		T_1	T_2	⋯	T_j	⋯	T_M
235 000 құжат	D_1	x_{11}	x_{12}	⋯	x_{1j}	⋯	x_{1M}
	D_2	x_{21}	x_{22}	⋯	x_{2j}	⋯	x_{2M}
	⋮	⋮	⋮	⋱	⋮		⋮
	D_i	x_{i1}	x_{i2}	⋯	x_{ij}	⋯	x_{iM}
	⋮	⋮	⋮	⋱	⋮		⋮
	D_{235000}	$x_{235000,1}$	$x_{235000,2}$	⋯	$x_{235000,j}$	⋯	$x_{235000,M}$

мұнда: D_i - i-ші заңнамалық құжат; T_j - j-ші термин немесе токен.

Терминдер униграмма: «еңбек», биграмма: «еңбек_шарты», триграмма: «еңбек_шартын_бұзу» түрінде берілуі мүмкін.

Бұл матрица жоғары өлшемді сирек матрица (sparse matrix) болып табылады, себебі әрбір заңнамалық құжатта барлық терминдердің тек аз бөлігі ғана кездеседі. Қалған элементтердің көпшілігі нөлдік мәнге ие болады:

$$x_{ij} = 0 \quad for\ most\ (i,j) \tag{3.11}$$

Сондықтан есептеу жадысын тиімді пайдалану мақсатында матрица scipy.sparse форматында сақталады.

2. SVD/LSA алгоритмінің екінші қадамы - толық SVD жіктеу.

А матрицасы сингулярлық мән жіктеуі арқылы үш матрицаның көбейтіндісі ретінде жазылады:

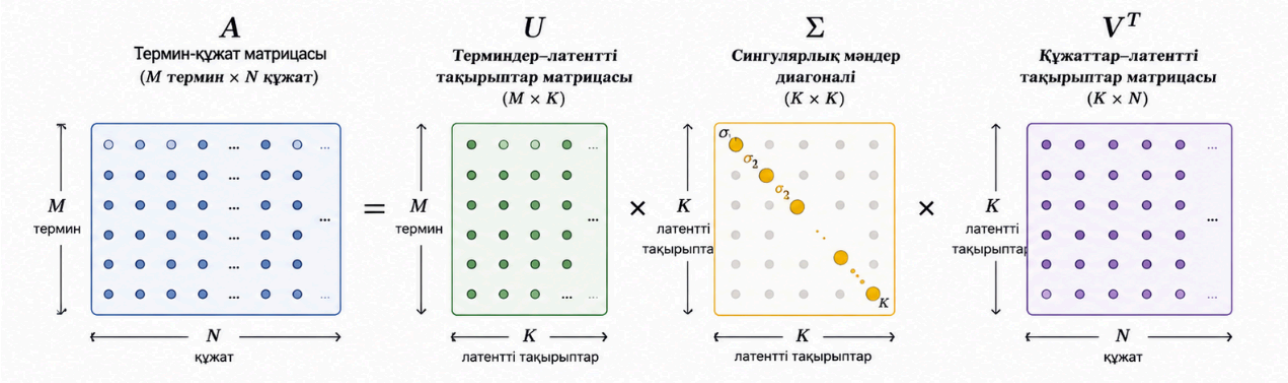
$$A = U \Sigma V^T \tag{3.12}$$

мұнда:

$U \in \mathbb{R}^{M \times M}$ – терминдер мен латентті тақырыптар арасындығы байланыс матрицасы;

$\Sigma \in \mathbb{R}^{M \times N}$ – сингулярлық мәндердің диагональды матрицасы ($\sigma^1 \geq \sigma^2 \geq \ldots \geq \sigma_r \geq 0$);

$V^T \in \mathbb{R}^{N \times N}$- құжаттар мен латентті тақырыптар арасындағы байланыс матрицасы.



3.5-сурет – Сингулярлық жіктеудің (SVD) иллюстрациясы.

3. SVD/LSA алгоритмінің үшінші қадамы - өлшемділікті азайту (Truncated SVD).

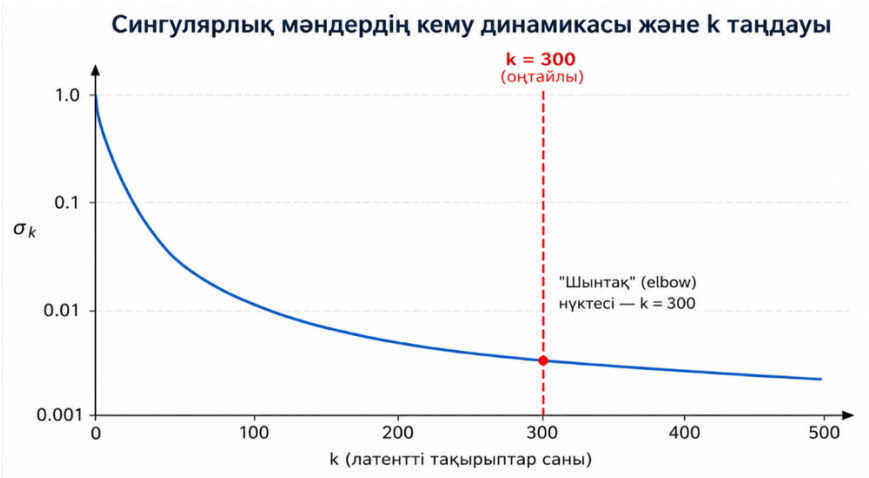
235 000 × МММ өлшеміндегі термин-құжат матрицасын толық SVD арқылы өңдеу есептеу жадысы мен уақыт шығынын едәуір арттырады. Сонымен қатар толық SVD барлық сингулярлық мәндерді, соның ішінде терминологиялық шу мен кездейсоқ морфологиялық ауытқуларды да сақтайды. Сондықтан заңнамалық корпустың негізгі семантикалық құрылымын сақтау үшін өлшемділікті қысқарту, яғни Truncated SVD қолдану қажет.

Эккарт–Янг–Мирский теоремасына сәйкес, Фробениус нормасы бойынша бастапқы A матрицасына ең жақын rank-k жуықтама алғашқы k ең үлкен сингулярлық мәнді сақтау арқылы алынады [100, 101]:

$$A_k = U_k \Sigma_k V_k^T \qquad (3.13)$$

$$\min\_rank(Z) \le k \parallel A - Z \parallel_F = \parallel A - A_k \parallel_F = \sigma_{k+1} \qquad (3.14)$$

Бұл жерде A_k - бастапқы матрицаның k дәрежелі ең жақсы жуықтамасы, ал σ_{k+1} жуықтау қатесін сипаттайтын келесі сингулярлық мән. Практикалық тұрғыдан бұл сингулярлық мәндердің кему қисығында «шынтақ нүктесін» таңдауға мүмкіндік береді: осы нүктеден кейінгі мәндер семантикалық ақпаратқа аз үлес қосады және көбіне терминологиялық шу мен кездейсоқ морфологиялық вариацияларды сипаттайды.

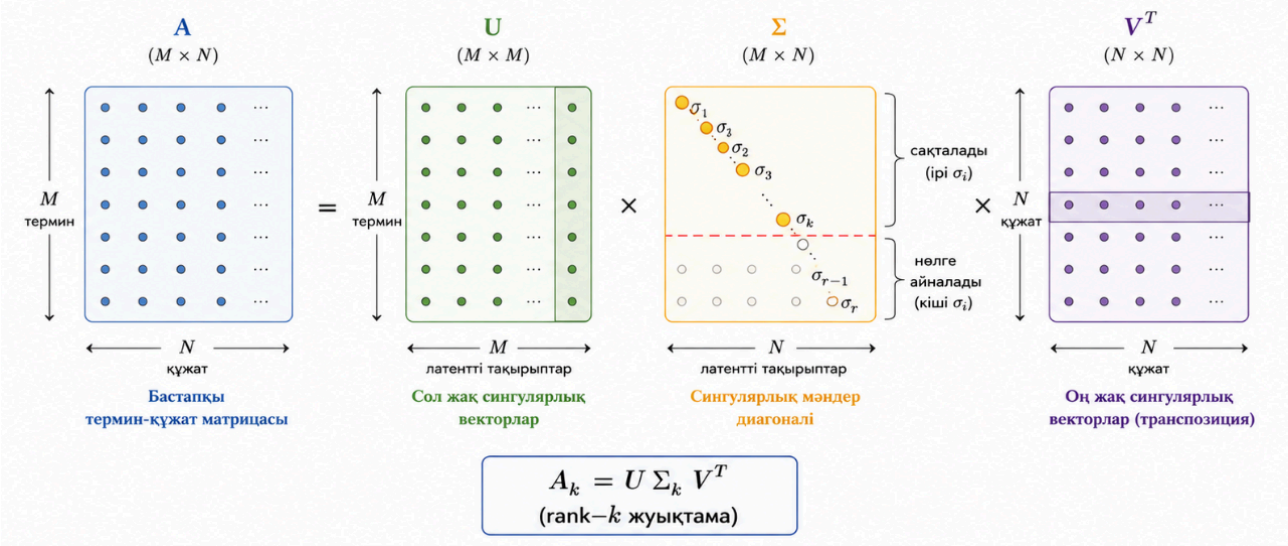


Сурет 3.7 - Сингулярлық мәндердің кему динамикасы және k = 300 таңдауы

Зерттеу жұмысында жүргізілген эксперименттер сингулярлық мәндер кему қисығын талдап, k = 300 кезінде корпустың жиынтық дисперсиясының 95%-нан астамы қамтылатынын растады:

$$\frac{\sum_{i=1}^{300} \sigma_i^2}{\sum_{i=1}^{r} \sigma_i^2} \geq 0.95 \qquad (3.15)$$

мұндағы σi - сингулярлық мәндер; r - матрицаның толық рангі;
алым - алғашқы 300 латентті компонент сақтайтын энергия;
бөлім - барлық сингулярлық мәндердің толық энергиясы.



Сурет 3.6 - ең кіші сингулярлық мәндерді нөлге айналдыру арқылы кіші дәрежелі жуықтама

k = 300 параметрі кездейсоқ таңдалмаған. Бұл гиперпараметр 50-ден 500-ге дейінгі аралықта тәжірибелік түрде тексерілді. Эксперимент нәтижелері бойынша k = 300 кезінде ҚР заңнамасының 50 ресми құқықтық саласы және олардың ішкі тақырыптары ең тұрақты түрде бөлінетіні анықталды.

Кесте 3.5 – k коэффициентін таңдау экспериментінің көрсеткіші

k мәні	R(k) дисперсиясы	Артықшылығы	Кемшілігі
k = 50	~71%	Есептеу жылдамдығы жоғары	Тақырыптар бірігіп кетеді
k = 100	~82%	Негізгі салалар бөлінеді	Тар салалар ажыратылмайды
k = 150	~88%	Орта деңгейлі тақырыптар ашылады	34 сала толық бөлінбейді
k = 300 ✓	~95%	34 сала және ішкі тармақтар нақты бөлінеді	Есептеу шығыны жоғары
k = 500	~98%	Ең толық қамту	Артық бөлшектену байқалады

Truncated SVD жуықтамасы:

$$\hat{A}_{235000\times M} \approx U_{235000\times 300} \cdot \Sigma_{300\times 300} \cdot V^{T}_{300\times M} \qquad 3.16$$

U - құжаттардың латентті тақырыптар кеңістігіндегі координаттары;
Σ - тақырыптардың маңыздылық деңгейін сипаттайтын сингулярлық мәндер диагоналі;
V^T - N-грамм терминдерінің жасырын тақырыптарға қатыстылық салмақтары.



Сурет 3.7 - Truncated SVD нәтижесінде алынған матрицалардың алгебралық құрылымы

Суретте Truncated SVD әдісі арқылы алынған үш негізгі матрицаның құрылымы көрсетілген: $U_{235000\times 300}$, $\Sigma_{300\times 300}$, $V^T_{300\times M}$ матрицасы латентті тақырыптардың маңыздылық деңгейін сипаттайтын сингулярлық мәндер диагоналінен тұрады. Үлкен сингулярлық мәндер заңнамалық корпуста жиі кездесетін және семантикалық тұрғыдан маңызды құқықтық тақырыптарды көрсетеді.

U матрицасы әрбір заңнамалық құжаттың 300 өлшемді жасырын тақырыптар кеңістігіндегі координаттарын сипаттайды. Бұл матрица арқылы құжаттардың қай құқықтық тақырыптарға жақын екені анықталады. Ал V^T матрицасы униграммалар, биграммалар және триграммалар сияқты N-грамм токендерінің жасырын тақырыптарға қатыстылық салмағын көрсетеді.

SemanticPrep тәсілінде V^Tматрицасы негізгі семантикалық оператор ретінде қолданылады. Дәл осы матрица мағыналық жағынан жақын құқықтық терминдерді бір латентті тақырып кеңістігінде байланыстырып, retrieval кезеңінде терминологиялық вариациялар арасындағы семантикалық сәйкестікті

анықтауға мүмкіндік береді.

4. SVD/LSA алгоритмінің төртінші қадамы - латентті кеңістікке проекция. Формализацияланған құқықтық сұрақ төмен өлшемді латентті тақырыптар кеңістігіне проекцияланады. Бұл кезеңнің негізгі мақсаты - пайдаланушы сұрағын заңнамалық корпуспен бір семантикалық кеңістікте көрсету және мағыналық жағынан жақын құқықтық құжаттарды анықтау.

Классикалық LSA тәсілінде $q \in R^M$ сұрақ векторы k-өлшемді латентті кеңістікке келесі түрде проекцияланады:

$$q^* = \Sigma_k^{-1} U_k^T q \quad (3.17)$$

мұнда:

$q \in R^M$ - сұрақтың TF-IDF векторы;

U_k^T - бастапқы M - өлшемді кеңістікті k - өлшемді латентті тақырыптар кеңістігіне проекциялау матрицасы;

Σ_k^{-1}- сингулярлық мәндер бойынша масштабтау матрицасы;

$q^* \in R^M$- сұрақтың латентті семантикалық векторы.

Бұл проекция нәтижесінде бастапқы жоғары өлшемді сұрақ ықшам латентті тақырыптар кеңістігінде ұсынылады. Ұсынылған SemanticPrep тәсілінде классикалық LSA схемасы заңнамалық доменге бейімделген түрде қолданылады. Бастапқы пайдаланушы сұрағы q_{raw} алдымен fine-tuning арқылы бейімделген mT5 трансформері арқылы заңнамалық стильге формализацияланады:

$$q_{legal} = G_{mT5}(q_{raw}) \quad (3.18)$$

мұнда: G_{mT5} - fine-tuned mT5 трансформерінің генерация функциясы;

q_legal - формализацияланған заңнамалық сұрақ векторы.

Бұл адаптацияның негізгі артықшылығы - сұрақ пен заңнамалық корпус арасындағы лексикалық алшақтықты азайту.

Формализацияланған сұрақ Q_{legal} алдымен TF-IDF векторына айналдырылады $q \in R^M$, содан кейін латентті тақырып кеңістігіне проекцияланады:

$$q_{adapted} = \Sigma_k^{-1} U_k^T q_{legal} \quad (3.19)$$

$$D_k = \Sigma_k V_k^T \quad (3.20)$$

мұнда U_k^T - сұрақты тақырып кеңістігіне аударады (M өлшемнен k өлшемге); Σ_k^{-1} - сингулярлық мәндер бойынша масштабтайды; $q_k \in R^k$ - сұрақтың латентті тақырыптық векторы (семантикалық кеңейту нәтижесі), D_k - заңнамалық құжаттардың латентті тақырыптық ұсынымдары. Осылайша, сұрақ пен құжаттар бірдей латентті кеңістікте көрсетіледі, бұл олардың семантикалық ұқсастығын салыстыруға мүмкіндік береді.

SVD/LSA алгоритмінің бесінші қадамы - семантикалық ұқсастық.

Сұрақ пен заңнамалық құжаттар арасындағы семантикалық ұқсастық косинустық ұқсастықпен өлшенеді. Бұл метрика вектор ұзындығына тәуелді емес, тек олардың «бағытын» салыстырады, сондықтан мәтін ұзындығы кең

ауқымда ерекшеленетін заңнамалық корпус үшін оңтайлы таңдау болып табылады:

$$sim(q_k, d_{j,k}) = \frac{q_k \cdot d_{j,k}}{||q_k|| \cdot ||d_{j,k}||} \quad (3.21)$$

мұнда $d_{j,k} \in \mathbb{R}^k$ - j-ші заңнамалық құжаттың латентті тақырыптық векторы. Косинустық ұқсастық мәні 1-ге жақындаған сайын сұрақ пен құжат бірдей заңнамалық тақырып кластерін ортақ бөледі; 0-ге жақын мән - мазмұндық байланыстың жоқтығын білдіреді. Жасырын кеңістіктегі z_q векторы негізінде жүйе сұраққа семантикалық тұрғыдан ең жақын N-грамм терминдерді автоматты түрде таңдайды. Кілт терминдер жиыны келесі өрнекпен анықталады:

$$K = Top_{m \leq 10}\{t_j \epsilon V_{ngram} : Sim(z_q, v_j) \geq \tau_2\} \quad (3.22)$$

мұнда τ_2 - кілт термин таңдау шегі;

Vngram - униграммалар, биграммалар және триграммалар сөздігі;

vj - j-ші N-грамм токенінің латентті кеңістіктегі векторлық бейнесі;

Topm≤10 - ең жоғары ұқсастыққа ие терминдерді іріктейтін оператор.

Практикалық тұрғыдан K жиыны сұрақтың мағынасына ең жақын құқықтық терминдерді автоматты түрде анықтайды. Кілт терминдер саны retrieval сапасына тікелей әсер етеді. Сондықтан m параметрі 3-тен 20-ға дейінгі аралықта тәжірибелік түрде зерттелді. Нәтижелер көрсеткендей, m шамадан тыс ұлғайған кезде Recall@5 аздап өссе де, Precision@5 төмендей бастайды. Бұл retrieval жүйелеріндегі классикалық *precision–recall trade-off* құбылысын көрсетеді.

Кесте 3.5 - Кілт сөздер санының (m) retrieval сапасына әсері

m (кілт сөз саны)	**Precision@5**	**Recall@5**	**Бағалау**
m = 3	0.72	0.58	Тым аз - маңызды тақырыптар жіберіліп қалады
m = 5	0.81	0.74	Жақсы, бірақ іргелес тақырыптар жетіспейді
m = 7	0.87	0.82	Оңтайлы: семантикалық қамту жоғары, шу жоқ
m = 10 ✓	0.88	0.86	Оңтайлы максимум: RAG retrieval үшін ең тиімді аралық
m = 15	0.83	0.87	Шектен тыс: іздеу шуы артады, Precision төмендейді
m = 20	0.74	0.88	Тиімсіз: нерелевантты терминдер RAG іздеу фокусын бұзады

Эксперимент нәтижелері бойынша m = 7–10 аралығы Precision@5 пен Recall@5 арасындағы ең тұрақты тепе-теңдікті қамтамасыз ететіні анықталды. m мәні 10-нан асқан кезде Recall@5 аздап өссе де, Precision@5 0.88-ден 0.74-ке дейін төмендейді - бұл retrieval жүйелеріндегі классикалық precision–recall trade-

off құбылысын білдіреді: тым көп кілт сөз берілсе, RAG модулі нерелевантты заңнамалық үзінділерді де іздеп, фокусын жоғалтады. Сондықтан ұсынылған жүйеде $m \leq 10$ шектеуі қолданылды.

3.4 SemanticPrep тәсілін басқа LLM модельдермен салыстырмалы талдау

SemanticPrep тәсілінің тиімділігін объективті бағалау мақсатында ұсынылған архитектура қазіргі заманауи сұрақты түрлендіру және retrieval-ді жақсарту тәсілдерімен салыстырмалы түрде талданды. Салыстыру барысында тек метрикалық нәтижелер ғана емес, сонымен қатар тәсілдердің архитектуралық ерекшеліктері, құқықтық доменге бейімделуі, семантикалық кеңейту қабілеті, retrieval жүйесіне ықпалы және агглютинативті қазақ тілімен жұмыс істеу ерекшеліктері талданды. Тәжірибелік өлшеу нәтижелері 4-тарауда баяндалады, осы бөлімде тәсілдердің теориялық-архитектуралық сипаттамасы беріледі.

Соңғы жылдары RAG жүйелерінде сұрақтарды алдын ала кеңейту және семантикалық түрлендіру retrieval сапасын арттырудың негізгі бағыттарының біріне айналды. Классикалық query expansion тәсілдері кілт сөздер мен синонимдерді қосуға негізделсе, қазіргі нейрондық тәсілдер гипотетикалық құжат генерациясы, dense retrieval үшін псевдо-құжат құру немесе LLM арқылы сұрақты қайта жазу әдістерін қолданады. Алайда құқықтық доменде бұл тәсілдердің бірқатар шектеулері бар: терминологиялық нақтылықтың төмендеуі, hallucination тәуекелі, API тәуелділігі және заңнамалық мәтіндердің құрылымдық ерекшеліктерін ескермеуі.

Салыстыру үшін таңдалған тәсілдер query transformation саласының үш түрлі архитектуралық класын өкілдендіреді: статистикалық кеңейту Query Expansion, генеративті псевдо-құжат тәсілдері HyDE, Query2Doc және нейрондық қайта жазу тәсілдері XLM-R, GPT-4o-mini, Gemini 1.5 Pro. Ұсынылған SemanticPrep тәсілі осы кластардың ешқайсысына толық жатпайды: ол нейрондық формализация мен статистикалық семантикалық кеңейтуді бір конвейерге біріктіреді.

Тәсілдердің концептуалды салыстырмасы. Метрикалық салыстырудан бұрын тәсілдердің негізгі идеялары мен артықшылықтары қарастырылды.

Кесте 3.6 - Сұрақ кеңейту тәсілдерін концептуалды салыстыру

Тәсіл	**Негізгі идея**	**Артықшылығы**	**Шектеуі**	**Заңнамалық домен үшін тиімділігі**
Query Expansion [93]	Кілт сөздер мен синонимдер қосу	Жылдам, ресурс сыйымдылығы төмен	Семантикалық тереңдігі әлсіз; лексикалық сәйкессіздік сақталады	Төмен

Кесте 3.6 жалғасы

HyDE [102]	LLM арқылы гипотетикалық құжат генерациясы	Dense retrieval сапасын арттырады	Hallucination қаупі жоғары; API тәуелді	Орта (hallucination тәуекеліменен)
Query2Doc [103]	Псевдо-құжат генерациясы арқылы контекстті кеңейту	Dense retrieval контекстін молайтады	Үлкен LLM ресурстарын қажет етеді; терминологиялық нақтылық төмен	Орта
GPT-4o-mini zero-shot rewrite	LLM арқылы сұрақты сілтеусіз қайта жазу	Жоғары генерация сапасы; табиғи тіл стилі	API тәуелді; hallucination; retrieval alignment жоқ	Орта
Gemini 1.5 Pro zero-shot rewrite	Ұзын контекст терезесімен сұрақты қайта жазу	Күрделі трансформацияларды орындайды; қазақ тілін қолдайды	API тәуелді; заңнамалық домен бейімделуі шектеулі	Орта
SemanticPrep - ұсынылған тәсіл	*Заңнамалық формализация және латентті семантикалық кеңейту*	*Домен бейімделген; API тәуелсіз; интерпретацияланады; retrieval alignment қамтамасыз ету*	*Алдын ала fine-tuning және SVD индекстеуін қажет етеді*	*Жоғары*

Query Expansion - ақпараттық іздеу жүйелеріндегі ең ерте кең таралған тәсілдердің бірі. Бұл әдісте бастапқы сұраққа қосымша кілт сөздер, синонимдер немесе морфологиялық варианттар қосылады [93]. Тәсіл recall көрсеткішін арттырғанымен, заңнамалық мәтіндерде терминдердің нақты мағынасы маңызды болғандықтан, keyword expansion кей жағдайда семантикалық тұрғыдан қатысы жоқ нормативтік құжаттарды қайтаруы мүмкін. Мысалы, «шартты бұзу» тіркесі еңбек шартына да, мүліктік жалдау шартына да бірдей қатысты болуы ықтимал, ал keyword matching осы айырмашылықты ажырата алмайды.

HyDE (Hypothetical Document Embeddings) тәсілін Gao et al. ұсынды [102]. Бұл әдісте LLM бастапқы сұраққа сәйкес гипотетикалық жауап немесе құжат генерациялайды; retrieval осы генерацияланған мәтіннің embedding-і негізінде орындалады. HyDE dense retrieval сапасын жақсартқанымен, заңнамалық жүйелер үшін маңызды тәуекел - hallucination мәселесі болып қалады. Модель заңнамада нақты жоқ терминдер немесе баптарды генерациялауы мүмкін, бұл пайдаланушыға жалған заңнамалық ақпарат ұсыну қаупін тудырады.

Query2Doc тәсілін Wang et al. ұсынды [103]. Бұл әдіс бастапқы сұрақтан псевдо-құжат генерациялап, retrieval жүйесіне кеңейтілген контекст береді. Dense retrieval үшін тиімді болғанымен, тәсіл үлкен LLM ресурстарын қажет етеді, тұрақты API инфрақұрылымына тәуелді, сонымен қатар заңнамалық мәтіндердегі терминологиялық нақтылықты толық қамтамасыз ете алмайды.

XLM-R (paraphrase adaptation) - Conneau et al. ұсынған көптілді encoder-only трансформер моделі [104]. Қазақ тілін қолдайды, бірақ заңнамалық домен үшін арнайы оқытылмаған. Encoder-only архитектурасы мәтінді кодтауға бейімделген болғандықтан, күрделі заңнамалық сұрақтарды генеративті түрде формализациялауда шектеулерге ие; paraphrase adaptation жалпы перифраза генерациясын орындайды, заңнамалық стильге бейімдемейді.

GPT-4o-mini (zero-shot rewrite) - OpenAI коммерциялық decoder-based LLM моделі. Zero-shot режимінде сұрақтарды қайта жаза алады, алайда тұрақты API инфрақұрылымын қажет етеді, hallucination тәуекелі сақталады және retrieval кезеңіне арнайы бейімделмеген. Яғни модель мәтінді стилистикалық тұрғыдан өзгертуі мүмкін, бірақ мұндай трансформацияның retrieval жүйесіне пайдасы тұрақсыз болады.

Gemini 1.5 Pro (zero-shot rewrite) - Google DeepMind компаниясының мультимодальды LLM моделі; 1 миллион токенге дейінгі контекст терезесімен ерекшеленеді [50]. Қазақ тілін салыстырмалы түрде жақсы қолдайды және күрделі мәтіндік трансформацияларды орындай алады. Дегенмен заңнамалық доменге арнайы бейімделмегендіктен, кей жағдайда заңнамалық терминдерді жалпылама немесе бейресми стильде генерациялауы мүмкін. GPT-4o-mini сияқты, коммерциялық API-ге тәуелді болуы мемлекеттік ақпараттық жүйелер үшін деректер құпиялылығы тәуекелін тудырады.

SemanticPrep - ұсынылған толық екі кезеңді тәсіл. Бірінші кезеңде mT5 моделі пайдаланушы сұрағын заңнамалық стильге формализациялайды, ал екінші кезеңде SVD/LSA арқылы латентті тақырыптар мен құқықтық терминдер анықталады. Тәсілдің маңызды артықшылығы - жергілікті орналастыру мүмкіндігі және API тәуелсіздігі. Бұл мемлекеттік және құқықтық ақпараттық жүйелер үшін маңызды, себебі сыртқы коммерциялық сервистерге тәуелділікті азайтады және деректер құпиялылығын сақтауға мүмкіндік береді.

Сонымен қатар ұсынылған тәсіл retrieval жүйесіне тек кеңейтілген мәтінді ғана емес, семантикалық құрылымдалған құқықтық ұсынымды береді. Нәтижесінде retrieval модулі сұрақтың лексикалық формасына емес, оның құқықтық мағынасына сүйене отырып заңнамалық құжаттарды табады. Бұл әсіресе агглютинативті тілдер мен құқықтық терминологиясы күрделі домендер үшін маңызды артықшылық болып табылады.

Жоғарыдағы архитектуралық талдау негізінде SemanticPrep тәсілі классикалық query expansion, HyDE және Query2Doc архитектураларынан принципиалды айырмашылығы - заңнамалық сұрақтарды тек мәтіндік деңгейде кеңейтпей, оларды заңнамалық стильге формализациялап, кейін жасырын семантикалық кеңістікке проекциялайды. Бұл retrieval жүйесіне тек кеңейтілген мәтінді ғана емес, семантикалық құрылымдалған заңнамалық ұсынымды береді.

Нәтижесінде retrieval модулі сұрақтың лексикалық формасына емес, оның заңнамалық мағынасына сүйене отырып нормативтік құжаттарды табады.

Ұсынылған тәсілдің теориялық артықшылығы - бейресми тілде жазылған азаматтық сұрақты заңнамалық корпустың статистикалық заңдылықтарымен сәйкестендіру механизмінің болуы. Бұл механизм GPT-4o-mini немесе Gemini 1.5 Pro тәсілдерінде жоқ: олар мәтінді стилистикалық тұрғыдан жақсартуы мүмкін, бірақ retrieval кезеңімен семантикалық alignment қамтамасыз ете алмайды. Аталған гипотезаны тексеру 4-тарауда эксперименттік деректермен расталады.

4 ҰСЫНЫЛҒАН ТӘСІЛДІ ЭКСПЕРИМЕНТТІК БАҒАЛАУ ЖӘНЕ ЗАҢНАМАЛЫҚ СҰРАҚ-ЖАУАП ЖҮЙЕСІНІҢ НӘТИЖЕЛЕРІН ТАЛДАУ

Бұл бөлімде ұсынылған SemanticPrep-RAG-LLM негізіндегі заңнамалық сұрақ-жауап жүйесінің эксперименттік бағалау нәтижелері талданады. Зерттеу барысында базалық және fine-tuning арқылы бейімделген LLM модельдерінің тиімділігі салыстырылып, retrieval сапасы мен жауап генерациясының дәлдігі сандық және сапалық метрикалар арқылы бағаланды. Сонымен қатар ұсынылған семантикалық препроцессинг тәсілінің құқықтық сұрақтарды өңдеу, заңнамалық фрагменттерді іздеу және галлюцинацияны азайтуға әсері зерттелді. Алынған нәтижелер ROUGE, BERTScore, METEOR, Recall@5, MRR және эксперттік бағалау көрсеткіштері негізінде талданып, жүйенің артықшылықтары, шектеулері және практикалық қолдану мүмкіндіктері анықталды.

4.1 Эксперименттік деректер сипаттамасы және бағалау метрикалары

Ұсынылған тәсілді эксперименттік тексеру барысында заңнамалық сұрақ-жауап деректер жиыны машиналық оқытудағы кең таралған тәжірибеге сәйкес оқыту, валидация және тестілеу кезеңдеріне 80% / 10% / 10% пропорциясында бөлінді. Мұндай бөлу стратегиясы модельдің оқыту деректеріне шамадан тыс бейімделуін (overfitting) азайтып, оның бұрын көрмеген сұрақтарға жалпылау қабілетін объективті бағалауға мүмкіндік береді. Эксперимент барысында төрт дереккөз пайдаланылды: Adilet, Zqai and Gov, Synthetic және Manual. Adilet жиыны Қазақстан Республикасының ресми заңнамалық порталы негізінде қалыптастырылып, нақты нормативтік мәтіндер мен ресми стильдегі сұрақ-жауап жұптарынан тұрады. Zqai and Gov жиыны мемлекеттік және сараптамалық порталдардан алынған, күрделі логикалық конструкциялар мен еркін синтаксиспен сипатталады. Synthetic жиыны үлкен тілдік модельдер арқылы генерацияланған әртүрлі форматтағы жұптардан тұрады. Manual жиыны қолмен құрастырылып, сараптамалық тексеруден өткен ең жоғары сапалы деректерді қамтиды. Зерттеу 2-тарауда (2.1 бөлімі) сипатталған деректер жиыны негізінде жүргізілді: 235 000 заңнамалық құжат 5.3 ГБ пен 8 000 сұрақ-жауап жұбынан тұратын арнайы датасет.

Кесте 4.1 - Деректер жиынының оқыту, валидация және тест бөліктеріне бөлінуі

Дереккөз	Train	Validation	Test	Барлығы	%
Adilet	810	101	101	1 012	12.7
Zqai and Gov	1 072	134	134	1 340	16.8
Synthetic	3 011	376	376	3 763	47.1
Manual	1 500	190	190	1 880	23.5
Жалпы	6 393	801	801	7 995	100

Жеке датасеттерде жеке оқыту стратегиясы таңдалды: бұл тәсіл форматтары мен күрделілік деңгейі ерекшеленетін дереккөздер бойынша әрбір модельдің

күшті және осал жақтарын тәуелсіз анықтауға мүмкіндік берді. Деректер жиынының толық бөліну статистикасы 4.1-кестеде берілген.

Жалпы деректер жиыны 7 995 сұрақ-жауап жұбынан тұрады: оқыту үшін 6 393 жұп (80%), валидация үшін 801 жұп (10%), тестілеу үшін 801 жұп (10%) бөлінді. Synthetic жиыны жалпы деректердің 47.1%-ін құрайды; бұл жиынның үлкен болуы LLM арқылы генерацияланған деректердің заңнамалық тақырыптардың кең ауқымын қамтуымен байланысты.

Деректер жиынын осындай құрылымда пайдалану төрт дәйекті эксперименттік сатыны жүргізуге негіз қалыптастырды. Эксперименттің жалпы дизайны 4.2-кестеде берілген.

Кесте 4.2 - Эксперименттің төрт сатылы дизайны мен бағалау метрикалары

Саты	**Атауы**	**Орындалған операциялар**	**Бағаланатын метрикалар**
I	Zero-shot бағалау	7 LLM модель fine-tuning қолданылмаған күйінде тестіленді	ROUGE-1, ROUGE-2, ROUGE-L, METEOR
II	Fine-tuning бейімдеу	7 модель заңнамалық QA датасетінде LoRA/QLoRA арқылы оқытылды; оқыту/тест нәтижелері салыстырылды	ROUGE-1, ROUGE-2, ROUGE-L, METEOR, F1-score, Expert Score
III	Fine-tuned LLM + RAG интеграциясы	Fine-tuning өткен LLM модельдері RAG архитектурасымен (FAISS + гибридтік retrieval) біріктірілді	BERTScore, ROUGE-L, MRR, Recall@5, Expert Score
IV	SemanticPrep–RAG–LLM толық пайплайны	mT5 формализация + SVD/LSA кеңейту + RAG + Fine-tuned LLM - толық пайплайн бағаланды	BERTScore, ROUGE-L, MRR, Recall@5, Expert Score

Бірінші сатыда жеті LLM модель - GPT-4o-mini, Gemini-2.0, KazLLM-3.1-8B, LLaMA-3.2-3B, Phi-4-14B, Qwen-2.5-7B және Mistral - fine-tuning қолданылмаған zero-shot режимінде тестіленді. Бұл сатының мақсаты бастапқы нәтиже деңгейін анықтап, модельдердің заңнамалық қазақ тілін өңдеудегі негізгі қателік түрлерін талдау болды. Нәтижелер ROUGE-1, ROUGE-2, ROUGE-L және METEOR метрикалары арқылы өлшенді.

Екінші сатыда барлық жеті модель 8 000 заңнамалық QA-датасетінде LoRA/QLoRA адаптер тәсілі арқылы fine-tuning өтті. Датасет оқытуға 80%, валидацияға 10% және тестке 10% бөлінді. Оқыту жылдамдығы $lr = 5\times10^{-5}$, batch

size = 16, epoch саны = 3–5 деп белгіленді. Fine-tuning алдындағы және кейінгі нәтижелер ROUGE-1, ROUGE-2, ROUGE-L, METEOR метрикаларымен салыстырылды. Сонымен қатар F1-score мен сарапшылық бағалау (Expert Score) арқылы генерация сапасы кешенді түрде бағаланды.

Үшінші сатыда fine-tuning өткен LLM модельдері RAG архитектурасымен интеграцияланды. Іздеу механизмі гибридтік формула - α·BM25 + β·cos(v_q, v_i) - негізінде жұмыс істеп, 235 000 заңнамалық фрагмент FAISS векторлық индексінде сақталды. Бұл сатыда retrieval сапасы мен генерация дәлдігі BERTScore, ROUGE-L, MRR, Recall@5 және Expert Score метрикалары арқылы бірлесе өлшенді.

Төртінші сатыда ұсынылған SemanticPrep–RAG–LLM толық пайплайнының тиімділігі бағаланды. Алдыңғы сатымен (Fine-tuned LLM + RAG) салыстыра отырып, SemanticPrep препроцессинг модулінің - mT5 түрлендіру мен SVD/LSA семантикалық кеңейтудің - retrieval және генерация сапасына қосатын үлесі анықталды. Нәтижелер үшінші сатымен бірдей метрикалар бойынша - BERTScore, ROUGE-L, MRR, Recall@5 және Expert Score - өлшенді, бұл тікелей салыстырмалы талдауға мүмкіндік берді.

Бағалау метрикаларының таңдалу негіздемесі мынадай: ROUGE метрикалары лексикалық сәйкестікті, BERTScore семантикалық жақындықты, MRR мен Recall@5 retrieval сапасын, ал Expert Score мен Fleiss Kappa сарапшылық тексеруді өлшейді. Метрикалардың анықтамалары мен есептеу формулалары төменде толық баяндалады. Генерация сапасын өлшеу үшін мәтіндік ұқсастық метрикалары пайдаланылды:

$$ROUGE-1 = \frac{\sum_{w \epsilon Reference} Count_{match}(w)}{\sum_{w \epsilon Reference} Count(w)} \quad (4.1)$$

– ROUGE-1, ROUGE-2, ROUGE-L - генерацияланған жауаптың эталон мәтінмен лексикалық ұқсастығын бағалайды; n-gram қабыстырмасын есептейді;

$$ROUGE-2 = \frac{\sum_{bg \epsilon Reference} Count_{match}(bg)}{\sum_{bg \epsilon Reference} Count(bg)} \quad (4.2)$$

– BERTScore - жауаптардың семантикалық сәйкестігін бағалайды; bert-base-multilingual моделі негізінде есептеледі;

$$ROUGE-L = \frac{LCS(X;Y)}{length(X)} \quad (3.3)$$

$$ROUGE-L = \frac{LCS(X;Y)}{length(Y)} \quad (4.4)$$

$$ROUGE-L_{F1} = \frac{(1+\beta)^2 \times Precision \times Recall}{Recall + \beta^2 \times Precision} \quad (4.5)$$

– METEOR - морфологиялық және семантикалық жақындықты өлшейді; қазақ тіліндегі сөз формаларына сезімтал.

$$METEOR = F_{mean}(1 - Penalty) \quad (4.6)$$

$$Penalty = \gamma \times \left(\frac{ch}{m}\right)^{\theta} \quad (4.7)$$

Ақпаратты іздеу (retrieval) сапасын бағалау үшін:

– Recall@5 - қажетті заңнамалық құжаттардың алғашқы бес нәтиже ішінде табылу ықтималдығы;

$$Recall@5 = \frac{|Relevant\ Documents \cap Retrieved@5|}{|Relevant\ Documents|} \quad (4.8)$$

мұнда:

Relevant Documents - эталон бойынша релевантты құжаттар жиыны;

Retrieved@5 - жүйе қайтарған алғашқы 5 құжат;

Recall@5 ∈ [0,1].

– MRR (Mean Reciprocal Rank) - қажетті құжаттың іздеу нәтижесіндегі орнын бағалайды; 1/rank формуласымен есептеледі.

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i} \quad (4.9)$$

мұнда:

N - сұрақтар саны;

$rank_i$ – i-ші сұрақ үшін алғашқы дұрыс құжаттың рангі;

егер дұрыс жауап бірінші орында болса, онда $\frac{1}{rank} = 1$;

MRR мәні 0 мен 1 аралығында өзгереді, жоғары мән retrieval жүйесінің тиімділігін көрсетеді.

Модельдердің жауаптары F1-score метрикасы арқылы да бағаланды; бұл метрика дәлдік (precision) пен толықтық (recall) арасындағы тепе-теңдікті көрсетеді. F1-score метрикасы шығарылған мәтін мен эталондық жауап арасындағы сәйкестік деңгейін бағалау үшін қолданылады. Ол лексикалық және ішінара семантикалық ұқсастықты өлшейді әрі дәлдік пен толықтықтың гармоникалық ортасы ретінде есептеледі: мұндағы дәлдік - шығарылған барлық токендер ішіндегі релевантты токендердің үлесі, ал толықтық - шығарылған мәтінде эталондық токендердің қамтылу дәрежесін көрсетеді. F1-score мәні ішінара дұрыс сәйкестіктерге де мүмкіндік береді, яғни табылған фрагмент эталоннан ауытқуы немесе оны тек жартылай қамтуы мүмкін; осылайша ол қатаң сәйкестік метрикаларымен салыстырғанда шығару сапасын анағұрлым икемді әрі ақпаратты бағалауды қамтамасыз етеді. F1-score (14) формуласымен есептеледі:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4.10)$$

Сонымен қатар сарапшылық бағалау (Expert Score) жүргізілді. Ғылыми әдебиетте сарапшылық бағалаудың бірнеше тәсілі қарастырылады. Олардың ішіндегі ең кең таралғаны - көп аспектілі критерийлер бойынша сапалық бағалау тәсілі.

Жұптық салыстыру әдісінде сарапшылар әртүрлі модельдерден алынған жауаптарды жұп-жұппен салыстырып, сапасы жоғары нұсқаны таңдайды. Рейтинг әдісінде сарапшыларға барлық жауаптарды сапа шкаласы бойынша (мысалы, 0-ден 5-ке дейін) рейтингілеу тапсырылады. Дельфи әдісі - сарапшылар арасындағы итерациялық анонимдік сауалнама болып табылады; консенсусқа жету үшін нәтижелер талқыланады. Бұл әдіс болжау мен күрделі жүйелерді бағалауға байланысты зерттеулерде кеңінен қолданылады.

Қазақ тіліндегі заңнамалық сұрақ-жауап жүйесін сарапшылық бағалау үшін көп аспектілі критерийлер бойынша сапалық бағалау тәсілі таңдалды. Таңдаудың негіздемесі төмендегідей. Біріншіден, бұл тәсіл стандартты жауаппен сәйкестікті ғана емес, пайдаланушы тұрғысынан маңызды қосымша аспектілерді - баяндалу анықтығын, логикалық дәлелдемелерді, жауаптың толықтығы мен релеванттылығын - бағалауға мүмкіндік береді. Екіншіден, қазақ тілімен және заңдық контекстпен жұмыс кезінде заңдық ақпараттың қаншалықты нақты және анық жеткізілгенін бағалау принципиалды маңызға ие. Үшіншіден, критерийлерді жүйенің нақты талаптарына бейімдеу икемділігі - тек фактілік дәлдік ғана емес, сонымен қатар жауаптың құрылымы мен толықтығы маңызды болатын күрделі заңдық нормаларға қатысты сұрақтар үшін ерекше қажет. Бірнеше критерий бойынша бағалауларды жиынтықтау жеке бағалаулардағы субъективтіліктің әсерін азайтады.

Сарапшылық бағалау бес критерий бойынша бес балдық шкала (1-ден 5-ке дейін) бойынша жүргізілді. Критерийлердің толық сипаттамасы 4.3-кестеде берілген.

Кесте 4.3 - Заңнамалық сұрақ-жауап жүйесін сарапшылық бағалаудың критерийлері

Критерий	**1 (Нашар)**	**2 (Орта деңгейден төмен)**	**3 (Орта)**	**4 (Жақсы)**	**5 (Өте жақсы)**
Заңнамалық дәлдік	Елеулі заңдық қателер мен жаңылыс түсіндірмелер бар.	Ішінара дұрыс, бірақ маңызды қателіктер бар.	Жалпы дұрыс, бірақ кейбір заңдық аспектілер анық емес.	Дәл, тек шамалы ауытқулар бар.	Толық дәл, ҚР заңнамасын толық көрсетеді.

Кесте 4.3 жалғасы

Заңнамалық толықтық	Мәселені түсіну үшін маңызды аспектілер жіберіліп кеткен.	Кейбір аспектілер қамтылған, бірақ маңызды заңдық ақпарат жоқ.	Негізгі заңдық тармақтар қамтылған, толықтық жетіспейді.	Барлық дерлік негізгі заңдық аспектілер толық қарастырылған.	Жан-жақты, заңдық аспектілерді толық ашады.
Анықтық және логикалық құрылым	Анық емес, құрылымы нашар.	Тұрақсыз құрылым, логикасы түсініксіз.	Негізінен анық, бірақ шамалы логикалық кемшіліктер бар.	Жақсы құрылымдалған, негізінен анық.	Мінсіз құрылымдалған, түсінуге оңай.
Релевантылық және контекстілік	Релевантты емес, қазақ заңдық контекстімен байланысы жоқ.	Ішінара релевантты, контекстуализация әлсіз.	Негізінен релевантты, контекстен шамалы ауытқулар бар.	Релевантты, шамалы контекстік мәселелер бар.	Толық релевантты, қазақ контекстіне бейімделген.
Заңдық дұрыстық және кәсіби стиль	Кәсіби емес стиль, терминология мен грамматикада көптеген қателер.	Нашар стиль, байқалатын қателер бар.	Қолайлы стиль, шамалы терминологиялық мәселелер бар.	Кәсіби, дұрыс терминология, аз қателер.	Мінсіз заңдық стиль мен терминология.

Ұсынылған заңнамалық сұрақ-жауап жүйесінің сапасын бағалау үшін құқықтық және тілдік сараптама салаларындағы мамандардан тұратын эксперттік топ құрылды. Эксперттік сарапшылар құрамына 3 эксперт: заң ғылымдары саласының өкілі, мемлекеттік құқықтық қызмет маманы және азаматтарға заңдық қызметтер бойынша мемлекеттік заң кеңесшісі маманы енгізілді.

Бағалаудың объективтілігі мен тәуелсіздігін қамтамасыз ету мақсатында сараптама анонимді түрде жүргізілді. Сарапшыларға жүйені әзірлеуші туралы, қолданылған модель архитектурасы туралы және басқа эксперттердің бағалаулары туралы ақпарат берілген жоқ. Әрбір эксперт жауаптардың

заңнамалық дәлдігін, құқықтық сілтемелердің дұрыстығын, мазмұнның толықтығын және тілдік сапасын бір-бірінен тәуелсіз түрде бағалады.

Анонимді сараптама принципі сыртқы факторлардың бағалау нәтижелеріне ықпал ету ықтималдығын төмендетуге мүмкіндік берді және алынған нәтижелердің ғылыми объективтілігі мен сенімділігін арттырды.

Орташа баллды есептеу мынадай дәйектілікпен жүргізілді. Әрбір сарапшы әрбір критерий бойынша және әрбір модель үшін бес балдық шкала бойынша баллдар қойды. Әрбір критерий бойынша орташа балл барлық сарапшылардың қойған баллдарының арифметикалық ортасы ретінде есептелді. Модельдің жалпы баллы барлық критерийлер бойынша барлық баллдардың арифметикалық ортасы ретінде анықталды.

Орташа сарапшылық балл мынадай формуламен есептеледі:

$$C_i = \frac{\sum_{j=1}^{n} O_{ij}}{n} \quad (4.11)$$

мұнда:

O_{ij} - j-ші сарапшының і-ші критерий бойынша қойған бағасы;

n - сарапшылар саны.

Модельдің жалпы баллы мынадай формуламен анықталады:

$$Total_{score} = \frac{\sum_{k=1}^{m} C_k}{M} \quad (4.12)$$

мұнда:

C_k - m критерийдің әрқайсысы бойынша орташа балл;

m - критерийлер саны.

Қолмен тексеру нәтижелеріндегі үш бағалаушы арасындағы келісімді сандық тұрғыдан өлшеу үшін Fleiss Kappa коэффициенті қолданылды. Бұл статистика дискреттік шкала пайдаланылған кезде екіден артық аннотаторлар арасындағы келісім дәрежесін өлшейді.

Fleiss коэффициентін есептеу қадамдары мынадай.

1-қадам. Бақылау матрицасын қалыптастыру. Әрбір сұрақтың әрбір критерийі бойынша үш сарапшының бағалаулары тіркелді. Мысалы, «заңнамалық дәлдік» критерийі бойынша бірінші сұраққа баллдар 4, 4 және 5 болса, сәйкес бақылау жолы [0, 0, 0, 2, 1] түрінде берілді, мұнда мәндер 1-ден 5-ке дейінгі бағалар үшін дауыс санын білдіреді.

2-қадам. Әрбір бақылау үшін келісімді есептеу:

$$P_i = \frac{1}{n(n-1)} \times \sum n_{ik}(n_{ik} - 1) \quad (4.13)$$

мұнда n = 3 - сарапшылар саны; n_{ik} - k категориясын таңдаған сарапшылар саны.

3-қадам. Орташа байқалған келісімді қалыптастыру:

$$\bar{P} = \frac{1}{N} \times \sum P_i, \quad (4.14)$$

мұнда N - критерийлердің жалпы саны (сұрақтар саны × 5).

4-қадам. Барлық категориялар бойынша дауыс үлестері негізінде күтілетін келісімді есептеу:

$$P_e = \Sigma (p_k^2) \quad (4.15)$$

мұнда p_k - k категориясына (1-ден 5-ке) берілген барлық бағалардың үлесі.

5-қадам. Түпкілікті формула:

$$k = \frac{\bar{P} - P_e}{1 - P_e} \quad (4.16)$$

Лэндис және Кох ұсынған классикалық шкаладан айырмашылығы - осы зерттеуде к мәндерін интерпретациялауда қатаңырақ тәсіл қолданылды. Бұл заңнамалық сұрақ-жауап жүйелерін бағалаудағы күрделі категориялық тапсырмалардың ерекшеліктері мен деректер теңсіздігіне сезімталдықпен байланысты. Заманауи зерттеулер де Kappa коэффициентін интерпретациялаудың тапсырма контексті мен деректер теңсіздігіне сезімталдықты ескеруі керектігін атап өтеді; әсіресе заңнамалық сұрақ-жауап жүйелеріндегі жауаптарды бағалау сияқты күрделі категориялық тапсырмаларда к мәндері 0.2–0.4 аралығы ресми түрде «қанағаттанарлық келісім» деп белгіленгенімен, тәжірибеде аннотаторлар арасындағы айтарлықтай айырмашылықтарды жасырып кетуі мүмкін [105-108]. Осыған байланысты осы зерттеуде «айтарлықтай келісім» шегі к > 0.60 деп белгіленді (4.4-кесте).

Кесте 4.4 - Fleiss Kappa мәндерін интерпретациялау (осы зерттеуде)

к мәні	**Келісім деңгейі**
к ≤ 0.00	Келісім жоқ (кездейсоқ деңгейінен төмен)
0.01–0.20	Әлсіз келісім
0.21–0.40	Шектеулі немесе тұрақсыз келісім
0.41–0.60	Орташа келісім
0.61–0.80	Айтарлықтай келісім (осы зерттеудегі шек)
0.81–1.00	Толыққанды немесе толық келісім

Жоғары сарапшылар арасындағы келісімнің тұрақты түрде тіркелуі бағалау критерийлерінің тұрақтылығын растайды және модель нәтижелерінің тәуелсіз сарапшылар тарапынан дәйекті бағалауға жеткілікті дәрежеде түсінікті және құрылымдалған екенін дәлелдейді. Бұл жалпы сарапшылық баллдарды талдаудың берік негізін қалыптастырады, әсіресе тілдік нюанстар бағалаушылар арасындағы интерпретациялық вариацияны күшейте алатын заң сияқты жоғары тәуекелді салаларда.

4.2 Базалық LLM модельдерінің эксперименттік салыстырмалы талдауы

Осы бөлімде қазақ тіліндегі заңнамалық сұрақ-жауап міндетінде қолданылған LLM тиімділігі эксперименттік тұрғыда салыстырмалы түрде бағаланады. Негізгі мақсат - әртүрлі архитектураларға негізделген модельдердің бастапқы (zero-shot) және бейімделген (fine-tuning) күйдегі генерация сапасын анықтау, сондай-ақ қазақ тіліндегі заңнамалық мәтіндерге тән қателіктердің құрылымын талдау.

Эксперимент бірыңғай әдістемелік негізде ұйымдастырылды. Барлық модельдер бірдей шарттарда бағаланды: 8 000 заңнамалық сұрақ-жауап жұбынан бөлінген тест бөлігі қолданылды, бірдей промпт құрылымы пайдаланылды және нәтижелерді бағалау бірыңғай метрикалық конвейер (ROUGE-1, ROUGE-2, ROUGE-L, METEOR, F1 Score, Expert Score) арқылы жүзеге асырылды. Мұндай тәсіл әртүрлі модельдердің нәтижелерін объективті және репродуктивті түрде салыстыруға мүмкіндік береді.

4.2.1 Базалық LLM модельдерінің zero-shot режиміндегі бағалануы

Бұл бөлімде қазақ тіліндегі заңнамалық сұрақтарға жауап беру сапасын анықтау мақсатында бірнеше үлкен тілдік модельге салыстырмалы эксперимент жүргізілді. Эксперименттің негізгі мақсаты - retrieval жүйелерін, қосымша семантикалық препроцессинг әдістерін және сыртқы құқықтық дерекқорларды қолданбай, модельдердің бастапқы генеративті мүмкіндіктерін бағалау болды. Бұл кезеңде алынған нәтижелер кейінгі fine-tuning, RAG интеграциясы және SemanticPrep препроцессинг әдістерінің тиімділігін салыстыру үшін бастапқы эталон (baseline) ретінде қарастырылады.

Зерттеу барысында GPT-4o-mini, Gemini, KazLLM, Llama, Phi, Qwen және Mistral модельдері қарастырылды. Модельдерді таңдау олардың архитектуралық ерекшеліктеріне, көптілді қолдауына, ашық (open-source) және жабық (proprietary) модельдер арасындағы айырмашылықтарына, сондай-ақ қазақ тіліндегі мәтіндерді өңдеу мүмкіндіктеріне негізделді. Мұндай таңдау заманауи LLM модельдерінің әртүрлі архитектуралық парадигмаларын қазақ тіліндегі заңнамалық доменде объективті түрде салыстыруға мүмкіндік берді.

Эксперимент closed-book режимінде орындалды, яғни модельдер тек өздерінің алдын ала оқытылған параметрлік білімдеріне сүйене отырып жауап генерациялады. Пайдаланушылардың құқықтық сипаттағы сұрақтары модельдерге тікелей беріліп, алынған жауаптар алдын ала дайындалған эталондық жауаптармен салыстырылды. Бұл тәсіл модельдердің заңнамалық контекстті қосымша ақпараттық қолдаусыз, өздігінен түсіну қабілетін бағалауға мүмкіндік береді.

Жауап сапасын бағалау үшін мәтін генерациясында кеңінен қолданылатын ROUGE-1, ROUGE-2 және METEOR метрикалары пайдаланылды. ROUGE-1 көрсеткіші модель жауабы мен эталондық жауап арасындағы жеке сөздердің (униграммалардың) сәйкес келу деңгейін анықтайды. ROUGE-2 екі сөзден тұратын тіркестердің (биграммалардың) сәйкестігін бағалап, жауаптың

құрылымдық байланысын сипаттайды. METEOR метрикасы сөздердің морфологиялық және семантикалық ұқсастығын ескеретін кешенді бағалау тәсілі болып табылады. Қазақ тілінің агглютинативті ерекшеліктерін ескергенде, METEOR метрикасы жауаптардың мағыналық сапасын анықтауда айрықша рөл атқарады, себебі ол сөздердің түбір-жалғаулық вариацияларын да ескереді.

Жеті LLM моделінің zero-shot режиміндегі сандық нәтижелері Кесте 4.4-те келтірілген.

Кесте 4.4 - Базалық LLM модельдерінің zero-shot режиміндегі нәтижелері

Модель	**ROUGE-1**	**ROUGE-2**	**METEOR**
GPT-4o-mini	0,118	0,039	0,105
Gemini	0,057	0,021	0,045
KazLLM	0,052	0,015	0,039
Llama	0,045	0,022	0,046
Phi	0,047	0,019	0,038
Qwen	0,058	0,028	0,040
Mistral	0,056	0,018	0,042

Эксперимент нәтижелері бойынша GPT-4o-mini моделі барлық метрикаларда ең жоғары көрсеткіштерге қол жеткізді: ROUGE-1 - 0,118, ROUGE-2 - 0,039 және METEOR - 0,105. Бұл модель заңнамалық терминологияны салыстырмалы түрде дәл өңдеп, эталондық жауаптарға мағыналық жағынан жақын әрі логикалық байланысы жоғары мәтіндер генерациялай алды. Сонымен қатар жауап құрылымының толықтығы мен семантикалық релеванттылығы басқа модельдерге қарағанда жоғары болды. Кестеден көрініп тұрғандай, GPT-4o-mini моделінің көрсеткіштері басқа модельдерден екі есеге жуық жоғары - бұл оның көптілді және доменаралық генерациялау қабілетінің салыстырмалы артықшылығын дәлелдейді.

Gemini (ROUGE-1 = 0,057) және Qwen (ROUGE-1 = 0,058) модельдері орташа нәтижелер көрсетті. Аталған модельдер жалпы құқықтық контексті түсінгенімен, кейбір жағдайларда нақты заңнамалық нормаларды немесе құқықтық әрекеттерді толық көрсете алмады; жауаптарда жалпылама түсіндірулер жиі байқалды. Mistral (ROUGE-1 = 0,056) де осы орташа топқа жақын нәтиже берді, алайда оның ROUGE-2 көрсеткіші (0,018) салыстырмалы төмен болды, бұл жауаптардағы тіркестік байланыстың әлсіздігін көрсетеді.

KazLLM моделінің нәтижелері (ROUGE-1 = 0,052) қазақ тіліне арнайы бағытталған модель болғанына қарамастан төмен деңгейде қалыптасты. Бұл құқықтық доменге арналған арнайы оқыту деректерінің жеткіліксіздігімен және заңнамалық контексті қамтитын retrieval-механизмдерінің болмауымен түсіндіріледі. Кейбір жауаптарда құқықтық терминдердің дұрыс қолданылмауы және логикалық толық еместігі байқалды. Бұл нәтиже тілдік бейімделудің өзі құқықтық доменге автоматты түрде бейімделуді қамтамасыз етпейтінін, доменге тән арнайы оқыту қажеттігін көрсетеді.

Llama (ROUGE-1 = 0,045), Phi (ROUGE-1 = 0,047) және Mistral модельдері де ұқсас төмен нәтижелер көрсетті. Әсіресе күрделі құқықтық сұрақтарда жауаптардың толық болмауы, терминологиялық сәйкессіздік және құқықтық нормаларды нақты интерпретациялай алмау мәселелері анықталды. Llama моделінде ROUGE-1 көрсеткіші ең төмен (0,045) болғанымен, METEOR мәні (0,046) салыстырмалы жоғары болуы - модельдің лексикалық дәл сәйкестіктен гөрі семантикалық жуықтауға бейім екенін көрсетеді.

Жалпы алғанда, эксперимент нәтижелері үлкен тілдік модельдердің retrieval және семантикалық препроцессинг әдістерінсіз қазақ тіліндегі заңнамалық сұрақтарға дәл және толық жауап беру мүмкіндігі шектеулі екенін көрсетті. ROUGE-1 көрсеткіштерінің барлық модельде 0,045–0,118 аралығында болуы базалық генеративті сапаның жалпы төмен деңгейде екенін айқындайды. Көптеген модельдер жауаптың жалпы мазмұнын сақтағанымен, нақты құқықтық нормаларды, заң баптарын және ресми құқықтық терминологияны тұрақты түрде қолдана алмады. Бұл шектеулер құқықтық доменге бейімделген арнайы семантикалық тәсілдерді - fine-tuning, RAG интеграциясы және семантикалық препроцессинг - қолданудың қажеттілігін дәлелдейді.

4.2.2 LLM модельдерін fine-tuning арқылы бейімдеу нәтижелерін бағалау

Эксперименттің екінші кезеңінде zero-shot режимінде анықталған шектеулерді жою мақсатында бірнеше үлкен тілдік модель fine-tuning арқылы бейімделді. Fine-tuning кезеңінің негізгі мақсаты - жалпы мақсаттағы көптілді модельдерді Қазақстан Республикасының заңнамалық доменіне адаптациялау және олардың құқықтық терминологияны, ресми стильді және заңнамалық логиканы өңдеу сапасын жақсарту болды.

Кесте 4.5 - Fine-tuning кезеңінде қолданылған деректер жиынтықтары

Датасет	Сипаттамасы	Ерекшелігі
Adilet	Ресми нормативтік-құқықтық құжаттардан құралған	Құрылымдалған заң стилі
Zqai + Gov	Азаматтардың нақты сұрақтарына жақын мәтіндер	Еркін синтаксис, күрделі логика
Synthetic	Жасанды генерацияланған сұрақ-жауап жұптары	Формулировкалардың әртүрлілігі

Зерттеу барысында GPT-4o mini, Gemini, KazLLM, LLaMA, Phi, Qwen және Mistral модельдері пайдаланылды. Модельдерді бейімдеу үшін Adilet, Zqai + Gov және synthetic типіндегі бірнеше заңнамалық деректер жиынтығы қолданылды. Adilet датасеті ресми нормативтік-құқықтық құжаттардан құралғандықтан, нақты заңнамалық стиль мен құрылымдалған құқықтық нормаларды қамтиды. Zqai + Gov деректер жиынтығында азаматтардың нақты сұрақтарына жақын,

еркін синтаксисті және күрделі логикалық құрылымды мәтіндер кездеседі. Synthetic датасеті әртүрлі формулировкаларды қамту мақсатында жасанды түрде генерацияланған сұрақ-жауап жұптарынан құралды. Деректер жиынтықтарының сипаттамасы Кесте 4.5-те берілген.

Барлық модельдер шектеулі есептеу ресурстарында тиімді оқыту үшін Quantized Low-Rank Adaptation (QLoRA) тәсілі арқылы 4-биттік кванттаумен (nf4 түрі, қос кванттау және float16 есептеулер) бейімделді. Модель архитектурасы адаптивті LoRA қабаттарымен толықтырылды: ранг параметрі r = 16, lora_alpha коэффициенті = 16, dropout = 0, ал бейімдеу q_proj, k_proj, v_proj, o_proj, gate_proj, up_proj және down_proj проекцияларына бағытталды. Оқыту 60 қадам бойы, batch size = 2 және төрт итерация бойынша градиент жинақтаумен жүргізілді; бастапқы оқыту жылдамдығы 2×10^{-4}, оптимизатор ретінде AdamW (8-bit) қолданылды. Толық гиперпараметрлер жиыны Кесте 4.6-да келтірілген.

Кесте 4.6 - Fine-tuning (QLoRA) гиперпараметрлері

Параметр	**Мәні**	**Сипаттамасы**
Бейімдеу тәсілі	QLoRA (4-bit)	nf4 кванттау, қос кванттау, float16 есептеу
LoRA рангі (r)	16	Адаптер матрицаларының өлшемі
lora_alpha	16	Масштабтау коэффициенті
Dropout	0	Регуляризациясыз
Мақсатты проекциялар	q, k, v, o, gate, up, down_proj	Бейімделетін қабаттар
Оқыту жылдамдығы (lr)	2×10^{-4}	Бастапқы learning rate
Batch size	2	Градиент жинақтау = 4
Қадам саны	60	Оқыту step саны
Оптимизатор	AdamW (8-bit)	Жадыны үнемдейтін нұсқа
Шығын функциясы	CLM (Cross-Entropy)	Авторегрессивті тіл моделі

Модельдердің сапасын бағалау үшін ROUGE-1, ROUGE-2, ROUGE-L және METEOR метрикалары қолданылды. ROUGE-1 бірлік сөздердің сәйкестігін, ROUGE-2 сөз тіркестерінің сәйкестігін, ал ROUGE-L жауап құрылымының толықтығын бағалайды. METEOR метрикасы морфологиялық және семантикалық ұқсастықтарды ескеріп, қазақ тілінің агглютинативті

ерекшеліктерін бағалауға мүмкіндік береді. Модельдердің барлық датасет бойынша орташа нәтижелері Кесте 4.7-де келтірілген.

Кесте 4.7 - Fine-tuning кейінгі модельдердің орташа ROUGE және METEOR нәтижелері

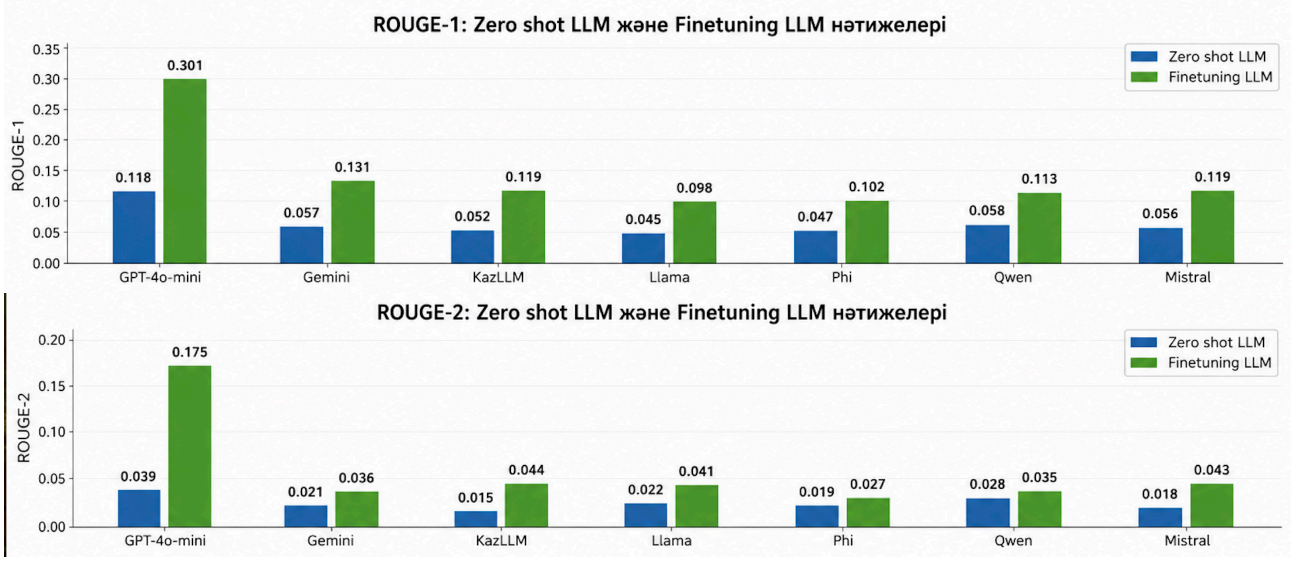
Модель	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**	**METEOR**
GPT-4o mini	0,205	0,121	0,182	0,205
Gemini	0,131	0,036	0,124	0,108
KazLLM	0,119	0,044	0,113	0,075
LLaMA	0,098	0,041	0,095	0,110
Phi	0,102	0,027	0,096	0,093
Qwen	0,113	0,035	0,109	0,092
Mistral	0,119	0,043	0,115	0,098

Эксперимент нәтижелері бойынша GPT-4o mini барлық датасеттерде ең жоғары нәтижелер көрсетті. Оның орташа нәтижелері ROUGE-1 = 0,205, ROUGE-2 = 0,121, ROUGE-L = 0,182 және METEOR = 0,205 болды. Бұл модельдің құқықтық контекстi жақсы түсініп, эталондық жауаптарға жақын мәтін генерациялай алатынын көрсетеді. Модельдердің датасеттер бойынша барлық көрсеткіштері Кесте 4.8-де берілген.

Кесте 4.8 - Модельдердің датасеттер бойынша нәтижелері (ROUGE-1 / METEOR)

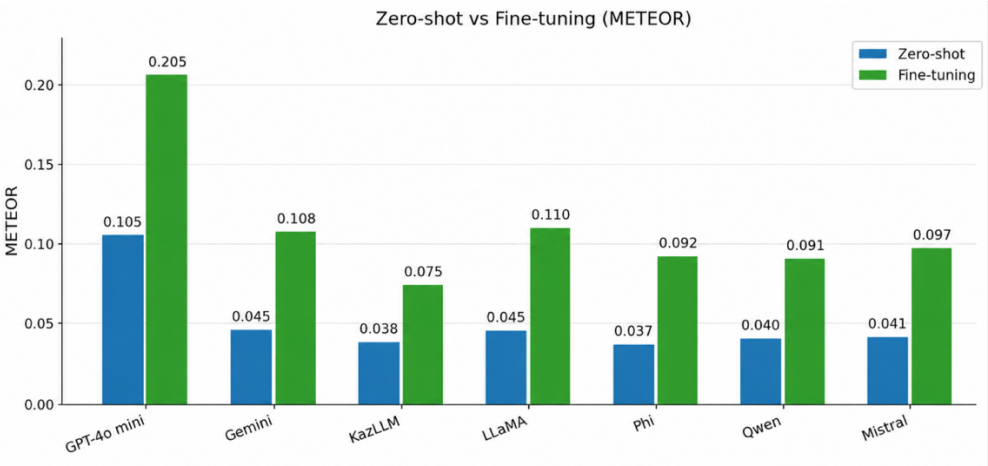
Модель	**Adilet**	**Zqai + Gov**	**Synthetic**
GPT-4o mini	R1: 0,242 M: 0,237	R1: 0,139 M: 0,122	R1: 0,235 M: 0,256
Gemini	R1: 0,081 M: 0,071	R1: 0,105 M: 0,068	R1: 0,208 M: 0,186
KazLLM	R1: 0,094 M: 0,066	R1: 0,106 M: 0,066	R1: 0,157 M: 0,093
LLaMA	R1: 0,069 M: 0,119	R1: 0,075 M: 0,047	R1: 0,150 M: 0,165
Phi	R1: 0,081 M: 0,083	R1: 0,079 M: 0,043	R1: 0,146 M: 0,152
Qwen	R1: 0,065 M: 0,055	R1: 0,091 M: 0,050	R1: 0,183 M: 0,172
Mistral	R1: 0,067 M: 0,048	R1: 0,114 M: 0,078	R1: 0,176 M: 0,168

Zero-shot және fine-tuning режимдеріндегі нәтижелерді салыстыру бейімделген оқытудың әсерін айқын көрсетеді. Сурет 4.3-те модельдердің ROUGE-1 метрикасы бойынша, ал Сурет 4.4-те METEOR метрикасы бойынша екі режимдегі көрсеткіштері салыстырмалы түрде берілген.



Сурет 4.1 - Модельдердің ROUGE-1 және ROUGE-2 бойынша zero-shot және fine-tuning салыстырмасы

Сурет 4.1-тен көрініп тұрғандай, fine-tuning барлық модельдің ROUGE-1 көрсеткішін екі-үш есеге арттырды. Ең жоғары өсімді GPT-4o mini көрсетті: zero-shot режиміндегі 0,118 мәні fine-tuning кейін 0,205-ке дейін өсті. Қалған модельдерде де тұрақты жақсару байқалады - мысалы, Gemini 0,057-ден 0,131-ге, KazLLM 0,052-ден 0,119-ға, Mistral 0,056-дан 0,119-ға дейін көтерілді. Бұл лексикалық сәйкестіктің доменге бейімделгеннен кейін айтарлықтай артқанын дәлелдейді.



Сурет 4.4 - Модельдердің METEOR бойынша zero-shot және fine-tuning салыстырмасы

METEOR метрикасы сурет 4.4-те берілген бойынша динамика да ұқсас үрдісті көрсетеді. GPT-4o mini моделінің METEOR мәні 0,105-тен 0,205-ке дейін өсіп, ең жоғары семантикалық сәйкестікке қол жеткізді. Назар аударарлық тұс - LLaMA моделі ROUGE-1 бойынша салыстырмалы төмен нәтиже бергенімен (0,098), METEOR бойынша GPT-тен басқа модельдер арасында ең жоғары мәнге (0,110) ие болды; бұл оның лексикалық дәл сәйкестіктен гөрі семантикалық жуықтауға бейім екенін көрсетеді. Жалпы алғанда, METEOR көрсеткішінің барлық модельде тұрақты артуы fine-tuning кезеңінің қазақ тілінің агглютинативті морфологиясын меңгеруге оң әсер ететінін растайды.

Gemini моделі орташа деңгейдегі нәтижелер көрсетті (орташа ROUGE-1 = 0,131, METEOR = 0,108). Модель жалпы құқықтық контексті сақтағанымен, кейбір жауаптарда нақты құқықтық аргументация жеткіліксіз болды, қайталанатын сөз тіркестері мен артық ұзын жауаптар байқалды. KazLLM моделі қазақ тіліне бейімделген жүйе болғанына қарамастан, GPT-4o mini деңгейіне жете алмады (ROUGE-1 = 0,119, METEOR = 0,075); оның басты мәселелері - сұрақты қайталау, қысқа жауап беру және терминологиялық дәлсіздік. Дегенмен fine-tuning кейін модельдің грамматикалық дұрыстығы мен жауап құрылымы айтарлықтай жақсарды.

LLaMA, Phi, Qwen және Mistral модельдері де салыстырмалы төмен нәтижелер берді. Әсіресе құқықтық синтаксис күрделі болған жағдайда жауап сапасы төмендеді: кейбір жауаптарда hallucination құбылыстары, қайталанатын фразалар және техникалық маркерлердің (<|end_of_text|>, <|im_start|>) сақталуы байқалды. Бұл жағдай аталған модельдердің құқықтық доменге толық бейімделмегенін көрсетті.

Автоматты метрикалардан бөлек, модельдердің жауаптары сарапшылар тарапынан да бағаланды. Сарапшылардың бағалауы қаншалықты бір-бірімен сәйкес келетінін анықтау үшін Fleiss’ Kappa коэффициентін есептеу қажет. Бұл көрсеткіш эксперттік бағалаудың сенімділігі мен тұрақтылығын дәлелдейді, яғни модель жауаптарын бағалау кездейсоқ емес, сарапшылар арасында ортақ келісім бар екенін көрсетеді.

Кесте 4.9 - Модельдер мен датасеттер бойынша Fleiss' Kappa коэффициенттері

Модель	Adilet	Zqai + Gov	Synthetic	Келісім деңгейі
GPT-4o mini	0,75	0,72	0,73	Елеулі келісім
Gemini	0,62	0,67	0,69	Елеулі келісім
LLaMA-3.2	0,61	0,64	0,62	Елеулі келісім

Кесте 4.9-да көрсетілген Fleiss' Kappa коэффициенттері сарапшылар арасындағы бағалау тұрақтылығының жеткілікті жоғары деңгейін көрсетеді. Бұл коэффициент бірнеше сарапшының бірдей жауаптарды бағалау кезіндегі келісім дәрежесін анықтау үшін қолданылды. Kappa мәндерінің 0,61–0,75 аралығында болуы бағалаушылар арасында substantial agreement, яғни елеулі келісім бар екенін білдіреді.

GPT-4o mini моделі үшін ең жоғары келісім деңгейі Adilet датасетінде байқалды (κ = 0,75). Бұл сарапшылардың модель жауаптарын біркелкі және тұрақты бағалағанын көрсетеді. Аталған модельдің жауаптары құрылымдық жағынан нақты, құқықтық терминологияны дұрыс қолданатын және логикалық

байланысы жоғары болғандықтан, сарапшылар арасында бағалау айырмашылықтары аз болды.

LLaMA-3.2 моделінде Kappa коэффициенттері 0,61–0,64 аралығында болды. Бұл көрсеткіштер сарапшылар арасында ортадан жоғары келісім бар екенін көрсеткенімен, GPT-4o mini нәтижелерінен төмен. Мұндай айырмашылық модель жауаптарында кездескен құрылымдық қателермен, толық емес жауаптармен және құқықтық аргументацияның әлсіздігімен түсіндіріледі.

Gemini моделінде Zqai and Gov және Synthetic датасеттерінде салыстырмалы түрде жоғары келісім байқалды (к = 0,67 және к = 0,69). Бұл модельдің кейбір еркін синтаксисті сұрақтарға тұрақты жауап бергенін көрсетеді. Дегенмен Adilet корпусында келісім деңгейі төменірек болды (к = 0,62), себебі ресми құқықтық стильді сақтау барысында сарапшылар арасында пікір айырмашылықтары кездескен.

Fleiss' Kappa нәтижелері қолданылған эксперттік бағалау әдісінің сенімділігін растады. Сарапшылар арасындағы келісімнің жоғары болуы бағалау критерийлерінің тұрақты екенін және модельдердің жауаптары интерпретациялауға жеткілікті деңгейде түсінікті болғанын көрсетеді. Бұл заңнамалық сұрақ-жауап жүйелерін бағалауда тек автоматты метрикалар ғана емес, сонымен қатар сарапшылық талдаудың да маңызды екенін дәлелдейді.

Модельдердің F1-score және эксперттік баллдарының салыстырмасы Кесте 4.10-де, ал қазақ тіліндегі бағаланған сұрақ-жауап мысалдары ағылшын тіліндегі аудармаларымен бірге А4 кестесінде келтірілген.

Кесте 4.10 - Модельдердің эксперттік бағалауы

Модель	**F1 Synthetic**	**F1 Adilet**	**F1 Zqai+ Gov**	**F1 Жалпы**	**Эксп. Synthetic**	**Эксп. Adilet**	**Эксп. Zqai +Gov**	**Эксп. Жалпы**
GPT-4o mini	0,171	0,114	0,096	0,127	2,8	3,697	2,893	3,13
LLaMA	0,077	0,082	0,066	0,075	1,06	1,33	1,11	1,17
Gemini	0,118	0,086	0,049	0,084	2,07	3,16	2,67	2,63

Ұсынылған F1 және эксперттік баллдар келесідей интерпретацияланды. GPT-4o mini моделінің F1-score көрсеткіштері LLaMA мен Gemini модельдерінен айтарлықтай асып түсті: жалпы ең жоғары балл 0,127 және синтетикалық датасеттегі ең үздік нәтиже 0,171. Gemini мен LLaMA модельдері айтарлықтай төмен F1-score көрсеткіштерін берді - сәйкесінше 0,084 және 0,075. Бұл нәтижелер GPT-4o mini моделінің мәтінді жіктеу мен ақпаратты іздеу міндеттеріндегі, әсіресе синтетикалық датасеттердегі басым қабілетін көрсетеді.

Дегенмен барлық модельдің Zqai + Gov датасетінде өнімділігі төмендегенін атап өткен жөн.

Эксперттік баллдар GPT-4o mini моделінің үстемдігін одан әрі нығайтады: ол ең жоғары жалпы эксперттік бағаға - 3,13 ие болып, Gemini мен LLaMA модельдерінің көрсеткіштерінен айқын асып түсті. GPT-4o mini барлық датасет бойынша тұрақты түрде жоғары бағаланды, әсіресе Adilet датасетінде 3,697 баллға жетті. Бұл бағалауда Gemini екінші орынды иеленіп, әр доменде салыстырмалы тұрақты баллдарды сақтады, ал LLaMA, әсіресе синтетикалық және Zqai + Gov датасеттерінде, тұрақты түрде артта қалды. Жалпы алғанда, GPT-4o mini жоғары өнімділік көрсеткіштерін көрсетіп, әртүрлі датасеттерді өңдеудегі тиімділігін дәлелдеді.

Эксперттік баға автоматты метрикалардың заңнамалық QA жүйелерін бағалауға жеткіліксіз екенін айқындайды: сарапшылық тексеру сандық көрсеткіштерде көрінбейтін аспектілерді (терминологияның дәлдігі, логикалық байланыстылық, құқықтық сілтемелердің дұрыстығы) ашады. GPT-4o mini автоматты метрикаларда да, сарапшылық бағалауда да сенімді көшбасшылық танытады, әсіресе нақты заңнамалық тұжырымдармен жұмыс істеуді талап ететін міндеттерде. Сирек жағдайлардағы кемшілігі - артық перифраздау (қайта тұжырымдау), бұл заң нормасын дәл дәйектеу дәлдігін төмендетуі мүмкін; әсіресе синтетикалық датасетте байқалады. Орташа эксперттік баға - 3,13, бұл модельдің заңнамалық контекстке бейімделу қабілетін растайды әрі мазмұны мен стилі жағынан сапалы жауап беретінін көрсетеді.

LLaMA сарапшылық бағалауда әлсіз нәтижелер көрсетті. Модель күрделі заңнамалық синтаксисті өңдеуде қиналады және ақпараты аз жауаптар береді. Мысалы, Zqai + Gov датасетінің үлгілерінде ол жауаптың логикалық құрылымын сақтамайды әрі стилистикалық қателер жібереді. Синтетикалық датасеттегі ең төмен балл (1,06) толық емес немесе тым жалпылама жауаптардың жоғары жиілігімен және негізгі бөлшектердің жоғалуымен байланысты. Оң тұсы - модель кейде қарапайым әрі қысқа сұрақтармен сәтті жұмыс істейді.

Gemini LLaMA-ға қарағанда тұрақтырақ нәтиже көрсетеді, бірақ сапалық және сандық метрикалар бойынша GPT-4o mini моделінен төмен. Айырмашылық, әсіресе еркін синтаксис пен күрделі формулировкаларды қамтитын корпустармен (Zqai + Gov) жұмыс істегенде айқын байқалады. Adilet корпусында модель нормативтік актілерді дұрыс дәйектеп, ресми стильді сақтайды. Ал синтетикалық датасетте сирек немесе стандартты емес сұрақтарда дәлдіктің төмендеуі, жауаптың жеткілікті бөлшектенбей оңайлатылып берілуі байқалды, бұл 2,07 деген төмен эксперттік баллға әкелді.

Қорыта айтқанда, біріктірілген талдау GPT-4o mini моделінің қазіргі таңда қазақ тіліндегі заңнамалық QA міндеттері үшін ең дәл, интерпретацияланатын және құқықтық тұрғыдан жарамды жауаптар беретінін растайды. Автоматты және сарапшылық бағалаудың үйлесімі сенімді нәтижелер берді әрі заң сияқты жоғары тәуекелді салаларда доменге тән жасанды интеллект қолданбаларын валидациялауда сарапшылық пайымның маңыздылығын айқындады.

Автоматты метрикалар мен сараптамалық бағалаудан кейін модельдердің жауаптарына сапалық талдау жүргізілді. Бұл талдаудың мақсаты - fine-tuning кезеңінен өткен LLM модельдерінің заңнамалық сұрақ-жауап міндетінде жиі жіберетін қателіктерін жүйелі түрде анықтау және оларды категорияларға жіктеу болды. Кесте 4.11-де сараптамалық тексеру барысында модель жауаптарында сандық көрсеткіштерде толық көрінбейтін, бірақ заңнамалық доменде шешуші маңызға ие үш негізгі қателік түрі айқындалды.

Кесте 4.11 - Fine-tuning кейінгі LLM модельдерінің жиі жіберетін қателіктері

№	Қателік түрі	Сипаттамасы және салдары
1	Галлюцинация (мәтіндік және мағыналық ауытқулар)	Модель нақты дереккөзге сүйенбей, сенімді көрінетін, бірақ заңнамалық шындыққа сәйкес келмейтін ақпарат генерациялайды. Бұл жауаптың мәтіндік құрылымы мен мағыналық мазмұнының эталоннан ауытқуына әкеледі.
2	Заңгерлік дәлдіктің жеткіліксіздігі (заңға сілтеме)	Жауаптар нақты нормативтік-құқықтық акт, бап немесе тармаққа сілтеме жасамайды. Модель жалпы мазмұнды дұрыс берсе де, құқықтық негіздемені (заңға сілтемені) көрсетпейтіндіктен, жауаптың заңгерлік жарамдылығы төмендейді.
3	Деректер қорындағы ақпараттың толық еместігі	Оқыту деректер қорында уақытқа тәуелді мәліметтердің, нақты мекенжайлар мен байланыс нөмірлерінің болмауы. Нәтижесінде модель өзекті әрі нақты практикалық ақпаратты бере алмайды.

Бірінші қателік түрі - галлюцинация, яғни мәтіндік және мағыналық ауытқулар. Модель нақты заңнамалық дереккөзге сүйенбестен, сенімді әрі дұрыс көрінетін, алайда шын мәнінде қолданыстағы құқықтық нормаларға сәйкес келмейтін ақпарат генерациялайды. Мұндай қателіктер заңнамалық QA жүйелерінде ерекше қауіпті, өйткені пайдаланушы жалған ақпаратты дұрыс деп қабылдап, нақты құқықтық шешім қабылдау кезінде қателесуі мүмкін.

Екінші қателік түрі - заңгерлік дәлдіктің жеткіліксіздігі, атап айтқанда нақты заңға сілтеменің болмауы. Сараптамалық бағалау барысында модельдер көп жағдайда жауаптың жалпы мазмұнын дұрыс жеткізгенімен, оны нақты нормативтік-құқықтық актіге, бапқа немесе тармаққа сілтеме жасау арқылы негіздей алмайтыны байқалды. Құқықтық негіздеменің болмауы жауаптың заңгерлік тұрғыдан тексерілуі мен жарамдылығын айтарлықтай төмендетеді.

Үшінші қателік түрі - оқыту деректер қорындағы ақпараттың толық еместігімен байланысты. Деректер қорында уақытқа тәуелді мәліметтердің,

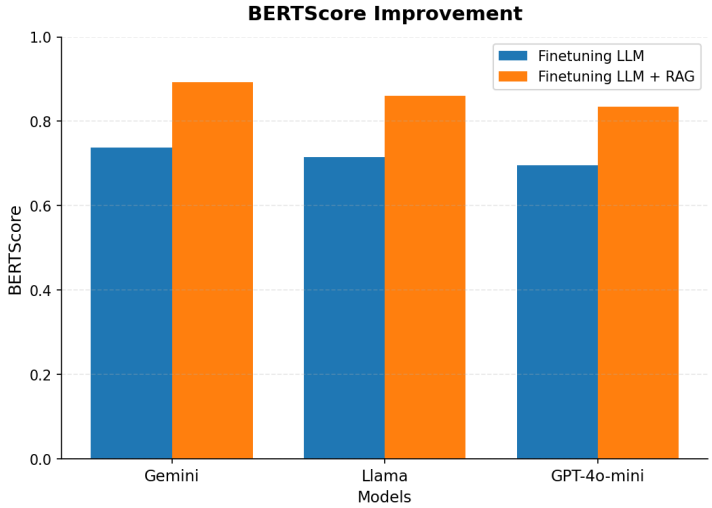
нақты мекенжайлар мен байланыс нөмірлерінің болмауы салдарынан модель өзекті әрі практикалық тұрғыдан пайдалы ақпаратты бере алмайды. Бұл шектеу әсіресе мемлекеттік қызметтер мен азаматтарға бағытталған нақты анықтамалық сұрақтарда айқын көрінеді.

Жалпы алғанда, fine-tuning нәтижелері құқықтық доменге бейімдеу үлкен тілдік модельдердің сапасын айтарлықтай жақсартатынын көрсетті. Анықталған бұл қателіктер fine-tuning тәсілінің жалғыз өзі заңнамалық сұрақ-жауап міндетін толық шеше алмайтынын көрсетеді. Галлюцинация мен заңгерлік дәлдіктің жеткіліксіздігі модельдің параметрлік білімін нақты әрі тексерілетін заңнамалық дереккөздермен байланыстыруды талап етеді, ал деректер қорындағы ақпараттың толық еместігі сыртқы, үнемі жаңартылатын дерекқорды қажет етеді. Осы себептерге байланысты келесі кезеңде модель жауаптарын нақты заңнамалық құжаттармен толықтыратын RAG тәсілін енгізу қажеттілігі негізделді.

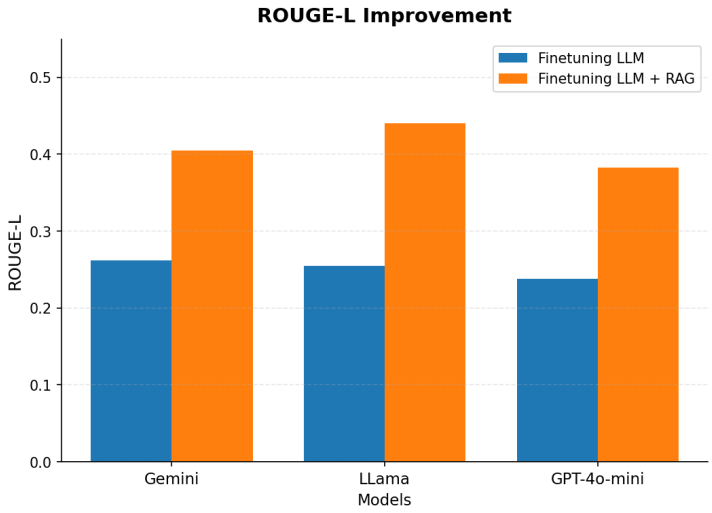
4.3 Finetuning LLM және RAG интеграциясы негізіндегі жүйенің нәтижелерін бағалау және талдау

Эксперименттің үшінші кезеңінде fine-tuning арқылы бейімделген модельдер RAG тәсілімен интеграцияланды. Бұл интеграцияның мақсаты - алдыңғы кезеңде анықталған шектеулерді (галлюцинация, заңға сілтеменің болмауы, ақпараттың ескіруі) жою және модель жауаптарын нақты заңнамалық дереккөздермен байланыстыру болды. RAG архитектурасында пайдаланушы сұрағы алдымен векторлық дерекқордан семантикалық тұрғыдан ең жақын құжаттарды іздеуге жіберіледі, содан кейін табылған контекст fine-tuning моделіне беріліп, нақты дереккөзге негізделген жауап генерацияланады.

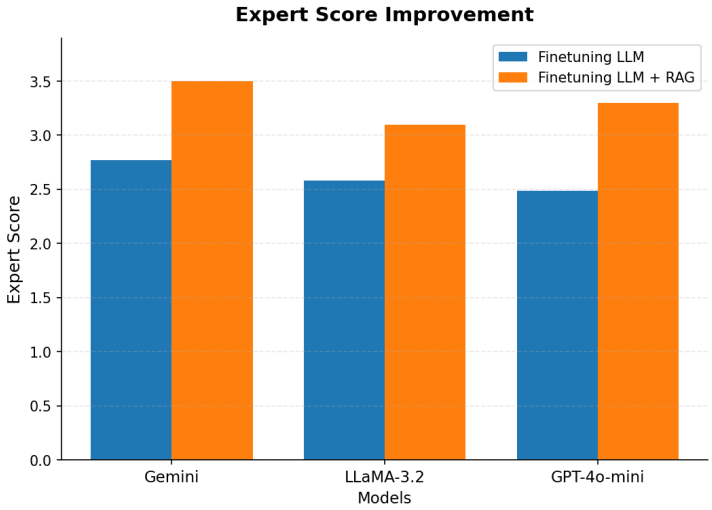
Интеграцияланған жүйенің сапасы кеңейтілген метрикалар жиынымен бағаланды: лексикалық және құрылымдық сәйкестік үшін ROUGE-L, семантикалық сәйкестік үшін BERTScore, іздеу сапасын бағалау үшін MRR (Mean Reciprocal Rank) және Recall@5, сондай-ақ сапаны кешенді бағалау үшін эксперттік баға (Expert Score). Fine-tuning режимі мен fine-tuning + RAG режимінің салыстырмалы нәтижелері Сурет 4.5, 4.6 және 4.7-де берілген.



Сурет 4.5 - RAG қолдану арқылы BERTScore көрсеткішінің өзгерісі



Сурет 4.6 - RAG қолдану арқылы ROUGE-L көрсеткішінің өзгерісі



Сурет 4.7 - RAG қолдану арқылы Expert Score көрсеткішінің өзгерісі

Сурет 4.5–4.7-ден көрініп тұрғандай, RAG интеграциясы барлық модельдің және барлық метриканың көрсеткіштерін айтарлықтай арттырды. BERTScore метрикасы бойынша Gemini моделі ең жоғары семантикалық сәйкестікке қол жеткізді (0,738-ден 0,893-ке дейін), Llama 0,715-тен 0,872-ге, ал GPT-4o-mini 0,695-тен 0,835-ке дейін өсті. ROUGE-L бойынша ең жоғары өсімді Llama моделі көрсетті (0,255-тен 0,440-қа дейін). Эксперттік бағалауда да барлық модель жақсарды: Gemini 2,77-ден 3,5-ке, GPT-4o-mini 2,49-дан 3,3-ке, Llama 2,58-ден 3,1-ге дейін көтерілді. Бұл нәтижелер RAG тәсілінің модель жауаптарын нақты заңнамалық контекстпен байланыстыру арқылы сапаны жүйелі түрде арттыратынын дәлелдейді.

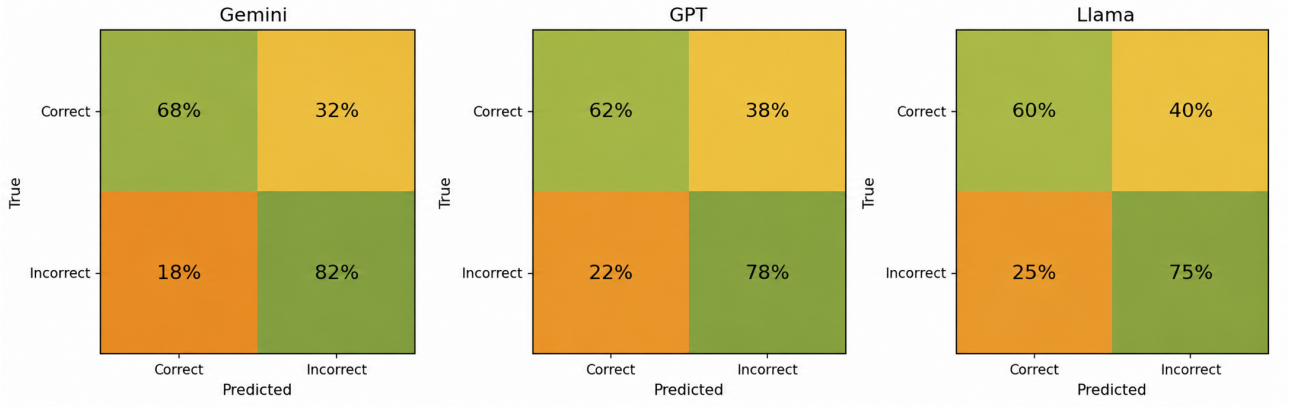
Интеграцияланған жүйенің толық сандық көрсеткіштері Кесте 4.13-те жинақталған. Кестеде BERTScore, ROUGE-L, MRR, Recall@5 және Expert Score метрикалары бойынша үш модельдің нәтижелері келтірілген.

Кесте 4.12 - RAG интеграциясымен fine-tuning LLM модельдерінің сапасын бағалау

Model	**BERTScore** *max:1*	**ROUGE-L** *max:1*	**MRR** *max:1*	**Recall@5** *max:1*	**Expert Score** *max:5*
Gemini-2.0	0,893	0,405	0,68	0,82	3,5
Llama-3.2	0,872	0,440	0,52	0,68	3,1
GPT-4o-mini	0,835	0,383	0,60	0,75	3,3

Кесте 4.12 бойынша Gemini-2.0 моделі BERTScore (0,893), MRR (0,68), Recall@5 (0,82) және Expert Score (3,5) метрикаларында ең жоғары нәтиже көрсетті, бұл оның семантикалық сәйкестік пен іздеу сапасы бойынша көшбасшы екенін білдіреді. Llama-3.2 моделі ROUGE-L метрикасында ең жоғары мәнге (0,440) ие болды, алайда MRR (0,52) мен Recall@5 (0,68) бойынша салыстырмалы төмен нәтиже берді - бұл оның мәтін генерациясы күшті болғанымен, релевантты құжаттарды іздеу дәлдігінің төменірек екенін көрсетеді. GPT-4o-mini барлық метрика бойынша тұрақты, теңгерімді нәтижелер берді.

Интеграцияланған жүйенің нәтижелеріне сапалық талдау жүргізу мақсатында әрбір модельдің жауаптары қателік матрицасы (confusion matrix) арқылы бағаланды. Матрица модельдердің дұрыс және қате жауаптарының үлесін, сондай-ақ қателіктердің түрлерін (false positive және false negative) айқындауға мүмкіндік береді. Үш модельдің қателік құрылымы Сурет 4.8-де берілген. Матрицада Gemini моделінің дұрыс жауаптарды тану дәлдігі ең жоғары (68%), одан кейін GPT-4o-mini (62%) және Llama (60%) орналасты.



Сурет 4.8 - Модельдердің жауаптарын талдау матрицасы (confusion matrix)

Модельдердің қателік құрылымы сурет 4.8-де көрсетілген. Барлық модельде дұрыс жауаптардың үлесі шамамен 60–68% аралығында болғанымен, қателіктердің айтарлықтай деңгейде сақталып отырғаны байқалады. Дұрыс жауаптардың үлесі бойынша Gemini моделі ең жоғары нәтиже көрсетті - шамамен 68%, одан кейін GPT-4o-mini (62%) және Llama-3.2 (60%) орналасты. Бұл көрсеткіштер модельдердің жалпы дәлдік деңгейін сипаттайды. Негізгі сандық көрсеткіштер Кесте 4.13-те жинақталған.

Кесте 4.13 - Модельдердің қателік матрицасы бойынша негізгі көрсеткіштері

Модель	Дұрыс жауап (TP)	False Positive	False Negative
Gemini	68%	18%	32%
GPT-4o-mini	62%	22%	38%
Llama-3.2	60%	25%	40%

Қателік матрицасын талдау екі негізгі қателік түрін айқындады. Біріншісі - false positive (жалған оң), яғни қате жауаптардың дұрыс деп болжануы. Бұл көрсеткіш Gemini моделінде ең төмен (18%), GPT-4o-mini моделінде 22%, ал Llama-3.2 моделінде ең жоғары (25%) болды. False positive деңгейінің жоғары болуы модельдердің кей жағдайда қате немесе нақты дереккөзге сүйенбеген ақпаратты дұрыс жауап ретінде ұсыну тенденциясын көрсетеді. Заңнамалық сұрақ-жауап жүйелерінде мұндай қателіктер ерекше қауіпті, өйткені пайдаланушы жалған ақпаратты дұрыс деп қабылдап, қате құқықтық шешім қабылдауы мүмкін.

Екінші қателік түрі - false negative (жалған теріс), яғни дұрыс жауаптардың анықталмай қалуы. Бұл көрсеткіш барлық модельде 32–40% аралығында болды: Gemini - 32%, GPT-4o-mini - 38%, Llama-3.2 - 40%. False negative деңгейінің салыстырмалы жоғары болуы модельдердің сұраққа қатысты толық әрі релевантты ақпаратты толығымен қамти алмайтынын білдіреді. Бұл әсіресе бірнеше нормативтік дереккөзді біріктіруді немесе кешенді құқықтық жауапты талап ететін сұрақтарда айқын көрінеді.

Жалпы алғанда, қателік матрицасының талдауы Gemini моделінің false positive (18%) және false negative (32%) көрсеткіштері бойынша ең теңгерімді әрі сенімді нәтиже беретінін дәлелдеді. Дегенмен барлық модельде сақталған қателік деңгейі заңнамалық QA жүйелерін одан әрі жетілдіру қажеттілігін көрсетеді: false positive қателіктерін азайту үшін генерацияланған жауаптарды нақты дереккөзбен тексеру механизмін күшейту, ал false negative қателіктерін төмендету үшін іздеу (retrieval) кезеңінің толықтығы мен қамтуын арттыру қажет.

RAG интеграциясы жалпы сапаны айтарлықтай арттырғанымен, сапалық талдау барысында жүйеде сақталған бірқатар қателік түрлері анықталды. Бұл қателіктер негізінен генерация кезеңінде емес, іздеу кезеңінде туындайтыны байқалды. Анықталған қателіктер Кесте 4.14-те төрт негізгі категорияға жіктелді.

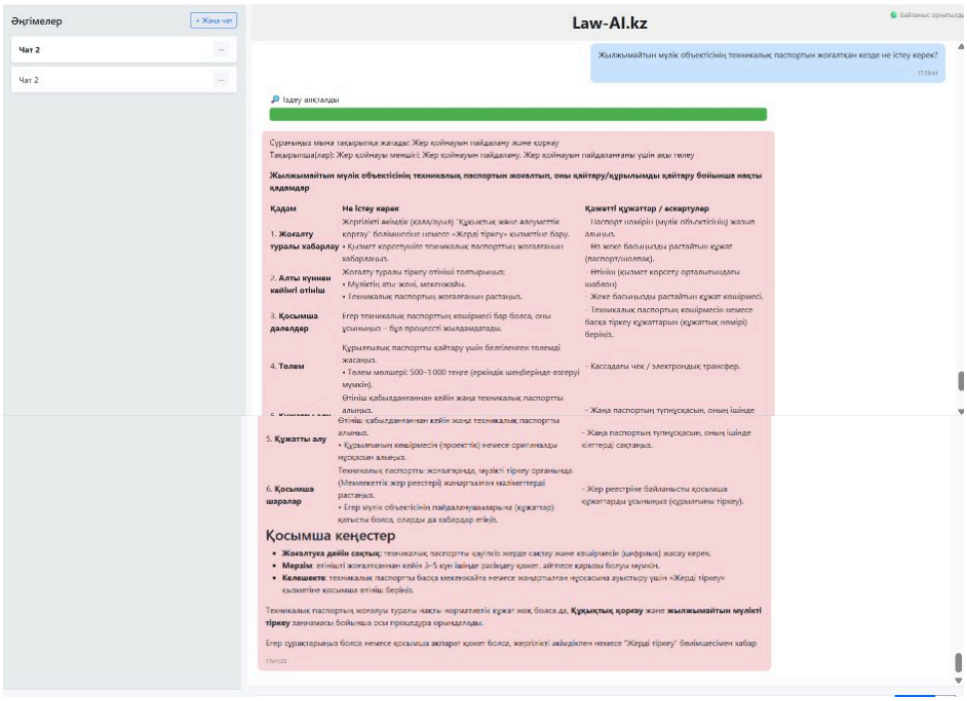
Кесте 4.14 - RAG интеграциясымен fine-tuning LLM модельдерінің жіберетін қателік түрлері

№	Қателік түрі	Сипаттама
1	Іздеу қателігі	Қажетті құжат табылмайды Кілт сөздер сәйкес келмейді; векторлық ұқсастық шегінен (τ) төмен
2	Сұрақ тәуелділігі	Сұрақ формасына тәуелді Бейресми / диалект сұрақтар нашар өңделеді; бір мағынадағы сұрақ түрлі нәтиже береді
3	Морфологиялық күрделілік	Сөз формалары ескерілмейді Қазақ тілінің жалғаулары іздеу дәлдігіне әсер етеді
4	Заңнамалық құрылым олқылығы	Заң логикасы толық түсінілмейді Субъект, әрекет, шарт ажыратылмайды; заң иерархиясы сақталмайды

Бірінші қателік түрі - іздеу қателігі, мұнда қажетті құжат табылмайды: пайдаланушы сұрағындағы кілт сөздер дерекқордағы құжаттармен сәйкес келмейді немесе векторлық ұқсастық белгіленген шектен (τ) төмен болғандықтан, релевантты үзінді іздеу нәтижесіне енбей қалады. Екінші қателік түрі - сұрақ тәуелділігі: жүйе сұрақтың тұжырымдалу формасына сезімтал болып, бейресми немесе диалект элементтері бар сұрақтарды нашар өңдейді, нәтижесінде бір мағынадағы әртүрлі тұжырымдалған сұрақтар түрлі нәтиже береді.

Үшінші қателік түрі - морфологиялық күрделілік: жүйе сөз формаларын толық ескермейді. Қазақ тілінің агглютинативті құрылымы мен жалғаулар жүйесі іздеу дәлдігіне теріс әсер етеді, өйткені бір сөздің әртүрлі тұлғалары векторлық кеңістікте әрқалай орналасуы мүмкін. Төртінші қателік түрі - заңнамалық құрылым олқылығы: модель заң логикасын толық түсінбейді, субъект, әрекет және шарт сияқты құқықтық норманың құрылымдық элементтерін бір-бірінен ажырата алмайды әрі заң иерархиясын (кодекс - бап - тармақ) сақтамайды.

Жоғарыда сипатталған қателік түрлерін нақты мысалмен көрнекі түрде көрсету мақсатында әзірленген жүйенің (Law-AI.kz прототипі) жауаптарының біріне жеке кейс-талдау жүргізілді. Мысалы, пайдаланушы «Жылжымайтын мүлік объектісінің техникалық паспортын жоғалтқан кезде не істеу керек?» деген сұрақ қойды. Жүйенің берген жауабының экрандық көрінісі Сурет 4.9-да келтірілген.



Сурет 4.9 - Law-AI.kz жүйесінің техникалық паспорт туралы сұраққа берген жауабы

Бірінші қарағанда жауап құрылымдық әрі толық көрінеді: ол кезең-кезеңмен берілген қадамдарды, қажетті құжаттар тізімін және қосымша кеңестерді қамтиды. Алайда жауапты заңнамалық дәлдік тұрғысынан сараптамалық тексеру оның нақты регламентке сәйкес келмейтінін және бірқатар елеулі қателіктерді қамтитынын көрсетті. Анықталған қателіктер Кесте 4.15-да жүйеленген.

Кесте 4.15 - Law-AI.kz жүйесінің жауабындағы анықталған қателіктерді талдау

№	Анықталған қателік	Неліктен бұл сыни маңызды
1	Тақырыпты қате жіктеу: сұрақ «жер қойнауын пайдалану» тақырыбына жатқызылды	Тұрғын үйдің техникалық паспорты жер қойнауын пайдаланумен байланысты емес. Бұл - классификация (тақырыпты анықтау) қателігі.
2	Процедура ЦОН / «Азаматтарға арналған үкімет» регламентіне сәйкес сипатталмаған	Іс жүзінде техникалық паспортты қайта алу тек «Азаматтарға арналған үкімет» (ЦОН) арқылы жүзеге асырылады, жауапта бұл орган көрсетілмеген.

Кесте 4.15 жалғасы

3	Техникалық түгендеу бюросы (ТТБ / БТИ) - негізгі орган - мүлдем аталмаған	Техникалық паспортты дәл осы ТТБ береді әрі қайта жасайды, жергілікті әкімдік емес. Негізгі жауапты органның көрсетілмеуі жауапты практикалық тұрғыдан жарамсыз етеді.
4	Төлем мөлшері «500–1 000 теңге» деп қате көрсетілген	Заң бойынша бұл - мемлекеттік баж, 0,5 АЕК шамасында (2026 жылы 1 АЕК = 4 325 теңге, демек ≈ 2 163 теңге). Көрсетілген сома нақты заңнамалық мөлшерге сәйкес келмейді.
5	Онлайн (eGov) қызметінің бар екені айтылмаған	Қазіргі таңда мемлекет цифрлық форматты белсенді ұсынады; қызметті eGov арқылы алу мүмкіндігі жауапта ескерілмеген.
6	«3–5 күн ішінде рәсімдеу қажет, әйтпесе қарыз болуы мүмкін» - толық ойдан шығарылған	Техникалық паспортты жоғалтқаны үшін ешқандай айыппұл немесе борыш заңнамада қарастырылмаған. Бұл - айқын галлюцинация үлгісі.

Кесте 4.15-те келтірілген талдау бір ғана жауаптың аясында бұған дейін аталған қателік түрлерінің бірнешеуі қатар кездесетінін көрсетеді. Атап айтқанда, сұрақты «жер қойнауын пайдалану» тақырыбына жатқызу - тақырыпты қате жіктеу (классификация) қателігі; нақты жауапты органның (ТТБ / ЦОН) аталмауы мен заңға сілтеменің болмауы - заңгерлік дәлдіктің жеткіліксіздігі; ал «3–5 күн ішінде рәсімдемесе қарыз болады» деген тұжырым - таза галлюцинация үлгісі, өйткені мұндай норма заңнамада мүлдем жоқ. Сонымен қатар төлем мөлшерінің қате көрсетілуі деректер қорындағы ақпараттың өзекті еместігін айғақтайды.

Қорытынды баға бойынша, бұл жауап ішінара пайдалы болғанымен, ЦОН/ТТБ нақты регламентіне сәйкес келмейді әрі төлем мөлшері, өтініш беретін орган және салдары жөнінде қате мәліметтерді қамтиды. Мұндай мәтінді ресми анықтама ретінде жариялауға болмайды. Бұл нақты мысал автоматты метрикалардың (ROUGE, BERTScore) жеткіліксіздігін көрнекі түрде дәлелдейді: жауап лексикалық және құрылымдық тұрғыдан эталонға жақын көрінгенімен, оның құқықтық мазмұны жарамсыз болуы мүмкін.

Жүргізілген талдау RAG интеграциясы жүйенің сапасын барлық метрика бойынша айтарлықтай арттырғанымен, сақталған қателіктердің басым бөлігі генерация емес, дәл іздеу (retrieval) кезеңіне қатысты екенін көрсетті. Атап айтқанда, бұл қателіктер сұрақтың бейресми тұжырымдалуынан, қазақ тілінің морфологиялық күрделілігінен және заңнамалық құрылымның толық ескерілмеуінен туындады. Демек, жүйенің сапасын одан әрі арттыру үшін негізгі назар генерация кезеңіне емес, іздеу алдындағы сұрақты өңдеу (pre-retrieval

processing) сатысына аударылуы қажет. Осы тұжырым ұсынылып отырған SemanticPrep тәсілінің - сұрақты заңнамалық стильге формализациялап, кілт терминдерді семантикалық деңгейде анықтайтын алдын ала өңдеу әдісінің - қажеттілігін негіздейді әрі оның тиімділігі келесі бөлімде эксперименттік тұрғыда қарастырылады.

4.4 SemanticPrep негізіндегі ұсынылған тәсілдің тиімділігін бағалау

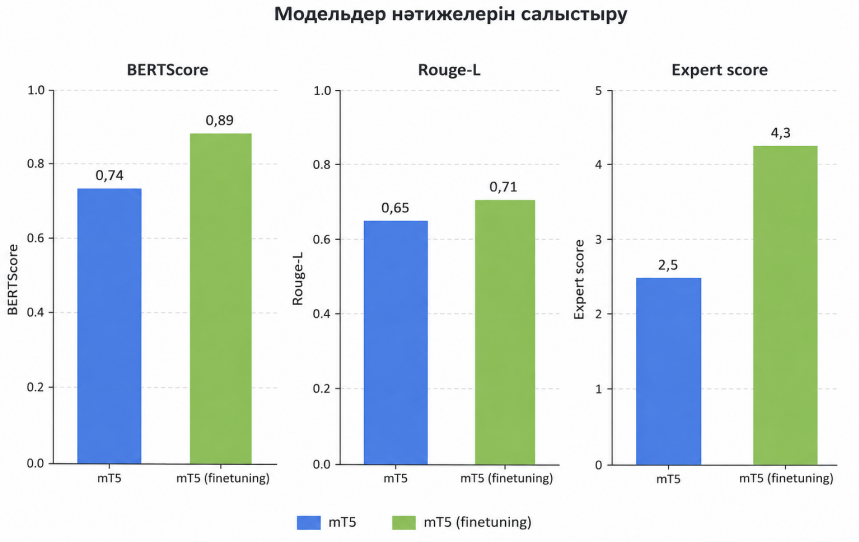
SemanticPrep тәсілінің бірінші кезеңінде пайдаланушының бейресми, ауызекі тілде қойған сұрағы заңнамалық (ресми) стильге түрлендіріледі. Бұл міндет mT5 (multilingual T5) трансформер моделін арнайы жиналған параллель датасетте fine-tuning жасау арқылы шешілді.

Оқытуға дейінгі және оқытудан кейінгі mT5 көрсеткіштерінің салыстырмасы Кесте 4.16-да келтірілген.

Кесте 4.16 - Оқытуға дейінгі және оқытудан кейінгі mT5 көрсеткіші

Модель	**BERTScore**	**ROUGE-L**	**Expert Score**
mT5	0,74	0,65	2,5
mT5 (fine-tuning)	0,89	0,71	4,3

Fine-tuning нәтижесінде mT5 моделінің барлық көрсеткіші айтарлықтай артты: BERTScore 0,74-тен 0,89-ға, ROUGE-L 0,65-тен 0,71-ге, ал эксперттік баға 2,5-тен 4,3-ке дейін өсті. Бұл модельдің бейресми сұрақтарды заңнамалық стильге дәл әрі сапалы түрлендіру қабілетінің едәуір жоғарылағанын дәлелдейді.



Сурет 4.10 - mT5 моделінің метрика салыстыру көрсеткіштері

SemanticPrep тәсілінің екінші кезеңінде формализацияланған сұрақ SVD/LSA (Singular Value Decomposition / Latent Semantic Analysis) алгоритмі арқылы өңделеді. Бұл кезеңде сұрақтан заңнамалық тақырып пен кілт терминдер

анықталып, семантикалық тұрғыдан жақын құжаттар айқындалады. SVD/LSA кезеңінің сапасы Кесте 4.17-да келтірілген.

Кесте 4.17 - SVD/LSA тақырып пен кілт сөздерді дұрыс табу көрсеткіші

Көрсеткіш	**SVD/LSA**
Keyword F1	0,79
Topic Accuracy	0,84
Expert Score	4,2

Кесте 4.17 бойынша SVD/LSA алгоритмі кілт сөздерді анықтауда F1 = 0,79, ал тақырыпты анықтауда дәлдік (Topic Accuracy) = 0,84 көрсетті; эксперттік баға 4,2 болды. Бұл нәтижелер алгоритмнің заңнамалық сұрақтардан мағыналық тұрғыдан маңызды элементтерді сенімді түрде бөліп ала алатынын дәлелдейді. Әзірленген семантикалық препроцессинг тәсілінің нақты қадамдарының мысалы Кесте 4.18-да көрсетілген.

Кесте 4.18 - Әзірленген семантикалық препроцессинг тәсілінің қадамдарының мысалы

Қадам	**Мазмұны**
Пайдаланушы сұрағы	Мені еш себепсіз жұмыстан шығарып жіберді, еңбек кітапшамды бермей жатыр, қайда шағымданамын?
mT5 кірісі (біріктіру)	Сұрақ: Мені еш себепсіз жұмыстан шығарып жіберді... Контекст: заңнамалық формализация үшін қарапайым тілдегі сұрақ берілді
mT5 шығысы (формализацияланған сұрақ - Q_legal)	Қызметкерді заңсыз жұмыстан босату және еңбек құжаттарын бермеу жағдайында еңбек инспекциясына немесе сотқа шағым беру тәртібі қандай?
SVD/LSA (семантикалық кеңейту)	Кілт терминдер: жұмыстан босату, қызметкер, еңбек шарты, еңбек инспекциясы, шағым, өтемақы, еңбек кодексі, жұмысқа қайта алу, моральдық зиян. Заңнамалық тақырып: Еңбек құқығы. Семантикалық жақын үзінділер: D002, D003, D010
Келесі кезең	Шығыс: Q_legal + тақырып + кілт сөздер

Ұсынылған SemanticPrep тәсілінің тиімділігін объективті бағалау үшін ол басқа заманауи query transformation (сұрақты түрлендіру) әдістерімен - XLM-R (paraphrase adaptation), GPT-4o-mini (zero-shot rewrite) және Gemini 1.5 Pro (zero-shot rewrite) - салыстырылды. Салыстырмалы нәтижелер Кесте 4.19-да келтірілген.

Кесте 4.19 - SemanticPrep тәсілінің басқа query transformation әдістерімен салыстырмасы

Тәсіл	ROUGE-L	BERTScore	Legal Term Precision	Expert Score
XLM-R (paraphrase adaptation)	0,51	0,77	0,63	3,3
GPT-4o-mini (zero-shot rewrite)	0,59	0,83	0,72	3,8
Gemini 1.5 Pro (zero-shot rewrite)	0,63	0,85	0,75	4,0
SemanticPrep (mT5 + SVD/LSA)	0,75	0,91	0,88	4,4

Кесте 4.19 бойынша ұсынылған SemanticPrep (mT5 + SVD/LSA) тәсілі барлық метрика бойынша салыстырылған әдістерден асып түсті: ROUGE-L = 0,75, BERTScore = 0,91, Legal Term Precision = 0,88 және Expert Score = 4,4. Әсіресе заң терминдерінің дәлдігі (Legal Term Precision) көрсеткішінде SemanticPrep ең жақын бәсекелес Gemini 1.5 Pro-дан (0,75) айтарлықтай жоғары нәтиже берді. Бұл екі сатылы мамандандырылған препроцессингтің жалпы мақсаттағы модельдердің zero-shot rewrite тәсілінен заңнамалық доменде едәуір тиімді екенін дәлелдейді.

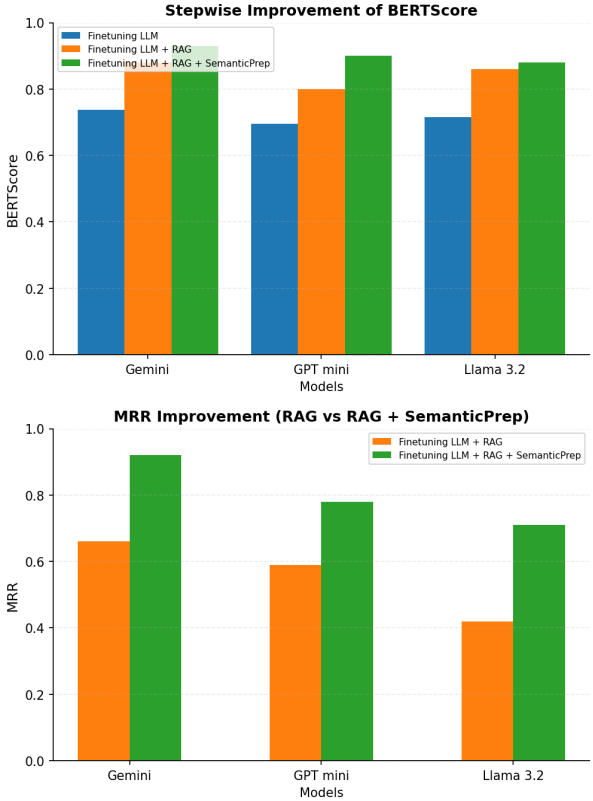
Кесте 4.20 - SemanticPrep әдісінің әртүрлі LLM модельдерінің сапасына әсері

Модель	Pipeline	BERTScore	ROUGE-L	MRR	Recall@5	Expert Score
Gemini 2.0	RAG + LLM	0,893	0,405	0,68	0,82	3,6
Gemini 2.0	SemanticPrep RAG + LLM	0,93	0,63	0,84	0,89	4,5
	Өзгеріс	*+3,7%*	*+22,5%*	*+16,0%*	*+7,0%*	*+25,0%*
GPT-mini 4.0	RAG + LLM	0,872	0,440	0,60	0,75	3,2
GPT-mini 4.0	SemanticPrep RAG + LLM	0,90	0,59	0,78	0,86	4,2
	Өзгеріс	*+2,8%*	*+15,0%*	*+18,0%*	*+11,0%*	*+31,0%*
Llama 3.2	RAG + LLM	0,835	0,383	0,42	0,68	3,1
Llama 3.2	SemanticPrep RAG + LLM	0,88	0,56	0,69	0,83	4,0
	Өзгеріс	*+4,5%*	*+17,7%*	*+27,0%*	*+15,0%*	*+29,0%*

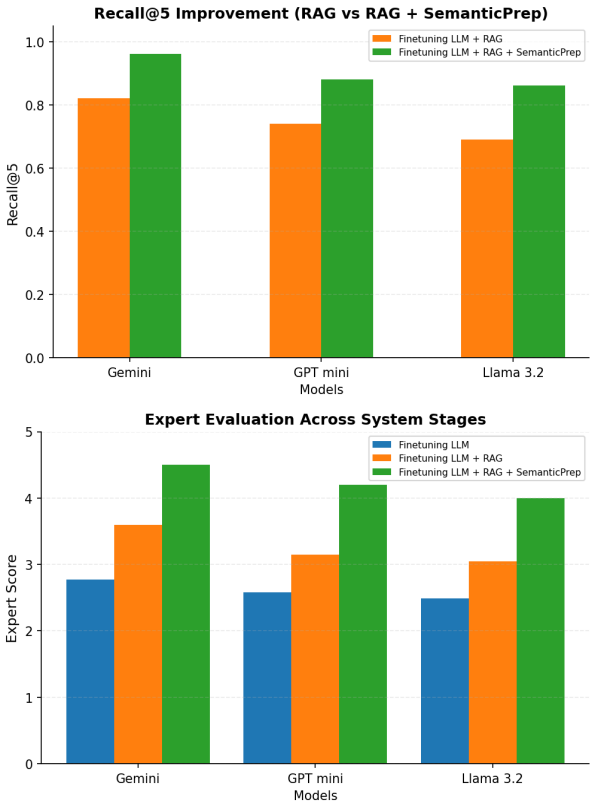
Ұсынылған тәсілдің негізгі ерекшелігі - оның модельге тәуелсіздігі (model-agnostic сипаты). SemanticPrep retrieval алдындағы сұрақты өңдеу сатысында жұмыс істегендіктен, ол қандай LLM генерация моделімен біріктірілсе де, жүйе сапасын арттырады. Осы тұжырымды тексеру үшін тәсіл үш түрлі архитектурадағы модельге (Gemini 2.0, GPT-4o-mini, Llama 3.2) қолданылып, RAG + LLM режимі мен SemanticPrep + RAG + LLM режимі салыстырылды. Нәтижелер Кесте 4.20-да келтірілген.

Кесте 4.20 нәтижелері ұсынылған тәсілдің негізгі идеясын дәлелдейді: SemanticPrep кезеңін қосу барлық үш модельдің де барлық метрикасын біркелкі арттырды. Атап айтқанда, эксперттік баға бойынша өсім Gemini 2.0 үшін +25%, GPT-mini үшін +31%, ал Llama 3.2 үшін +29% болды. Іздеу сапасын сипаттайтын MRR метрикасында да барлық модельде елеулі жақсару тіркелді (+16%-дан +27%-ға дейін). Бұл көрсеткіштер тәсілдің жекелеген бір модельге ғана емес, әртүрлі архитектурадағы LLM жүйелеріне әмбебап түрде оң әсер ететінін айқындайды.

Метрикалардың кезең-кезеңдік (Finetuning → RAG → SemanticPrep) өзгерісі Сурет 4.11–4.12-де көрнекі түрде берілген.



Сурет 4.11 - BERTScore және MRR метрикасының өзгерісі



Сурет 4.12 - Recall@5 және Expert Score метрикасының өзгерісі

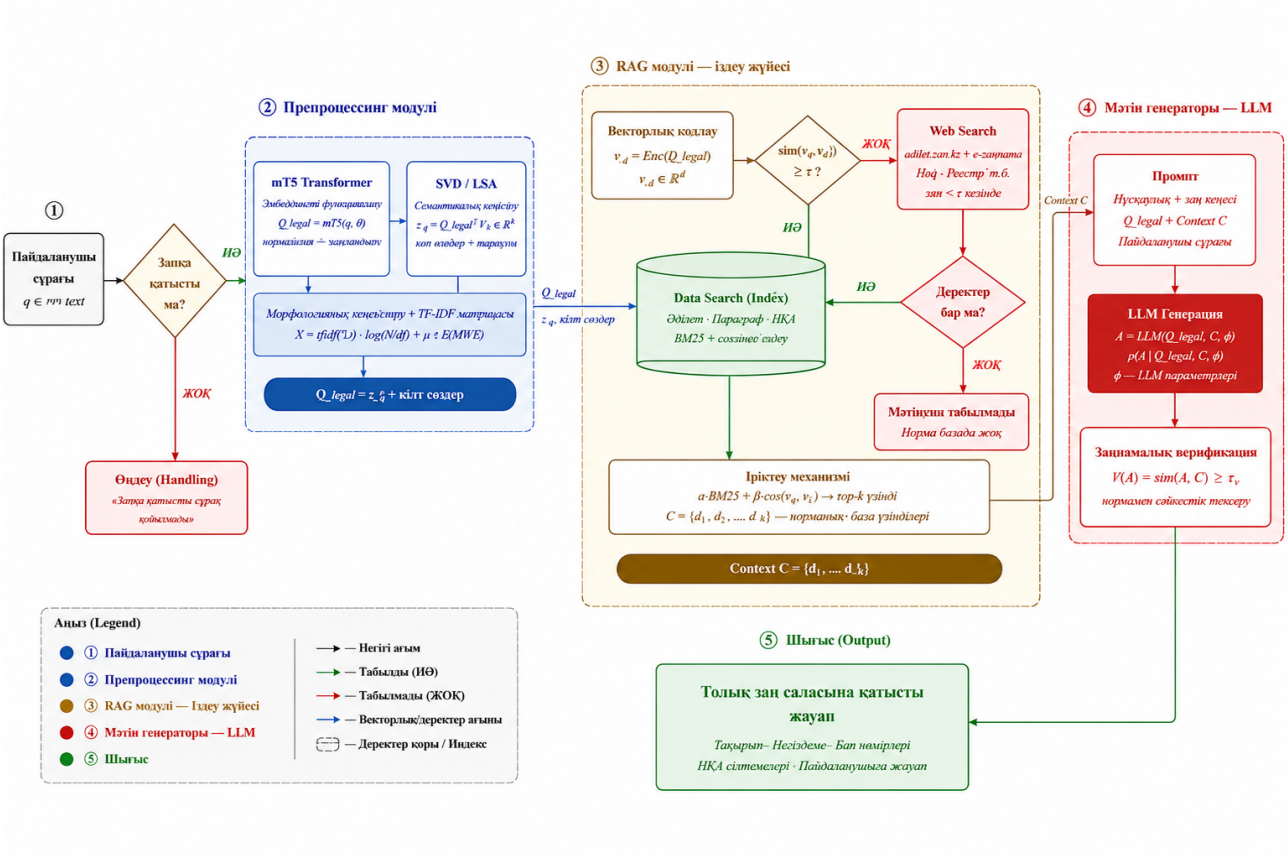
Сурет 4.11–4.12-ден көрініп тұрғандай, әрбір кезең (RAG, одан кейін SemanticPrep) метрика мәндерін дәйекті түрде арттырады. Әсіресе SemanticPrep кезеңі іздеу сапасын сипаттайтын MRR және Recall@5 метрикаларында ең үлкен өсімді қамтамасыз етті - бұл тәсілдің дәл retrieval кезеңінің әлсіздігін жоюға бағытталғанын және осы мақсатқа толық қол жеткізгенін растайды.

Жүргізілген эксперименттер ұсынылған SemanticPrep тәсілінің қазақ тіліндегі заңнамалық сұрақ-жауап жүйесінің сапасын зерттеудің барлық кезеңінде (zero-shot → fine-tuning → RAG → SemanticPrep) дәйекті түрде арттыратынын дәлелдеді. Ең бастысы, тәсілдің модельге тәуелсіз (model-agnostic) сипаты эксперименттік тұрғыдан расталды: ол қандай LLM генерация моделіне қолданылса да, көрсеткіштердің артуын қамтамасыз етті.

Тәсілдің бұл қасиеті оның практикалық құндылығын едәуір арттырады. SemanticPrep генерация моделінен тәуелсіз, retrieval алдындағы дербес модуль ретінде жұмыс істейтіндіктен, болашақта қазақ тілін тереңірек түсінетін, неғұрлым жетілдірілген LLM модельдері пайда болған жағдайда, жүйенің базалық генерация моделін қайта құрылымдаусыз, тікелей сол жаңа модельге ауыстыруға болады. Бұл ұсынылған архитектураның ұзақ мерзімді тұрақтылығын және тілдік модельдердің қарқынды дамуына бейімделу икемділігін қамтамасыз етеді. Демек, SemanticPrep тек ағымдағы модельдер үшін ғана емес, болашақта пайда болатын модельдер үшін де сапаны арттыратын әмбебап препроцессинг қабаты ретінде қарастырылады.

4.5 Әзірленген заңнамалық сұрақ-жауап жүйесінің архитектурасы

Ұсынылған SemanticPrep - RAG - LLM пайплайны Law-AI.kz платформасы ретінде практикалық апробациядан өтті. Жүйе https://law-ai.kz/ мекенжайы бойынша қолжетімді. Law-AI жүйесі бес функционалдық модульден тұрады: пайдаланушы сұрағын қабылдайтын кіріс блогы, заңға қатыстылығын анықтайтын сұраққа баға беру бөлімі, SemanticPrep негізіндегі препроцессинг модулі, RAG іздеу жүйесі және LLM мәтін генераторы. Толық архитектурасы 4.13 суретте төменде көрсетілген.



Сурет 4.13 - Law-ai.kz жүйесінің архитектурасы

Препроцессинг модулі - жүйенің негізгі инновациялық компоненті. Ол мынадай кезеңдерден тұрады:

mT5 Трансформер арқылы қазақ тіліндегі заңнамалық формализация: Q_legal = mT5(q; θ) - бастапқы сұрақты заңдық терминологияға сәйкестендіру;

SVD / LSA арқылы семантикалық кеңейту:

$z\_q = Q\_legal \cdot V\_k \in R\^k$ - тақырыптар мен кілт сөздерді анықтау;

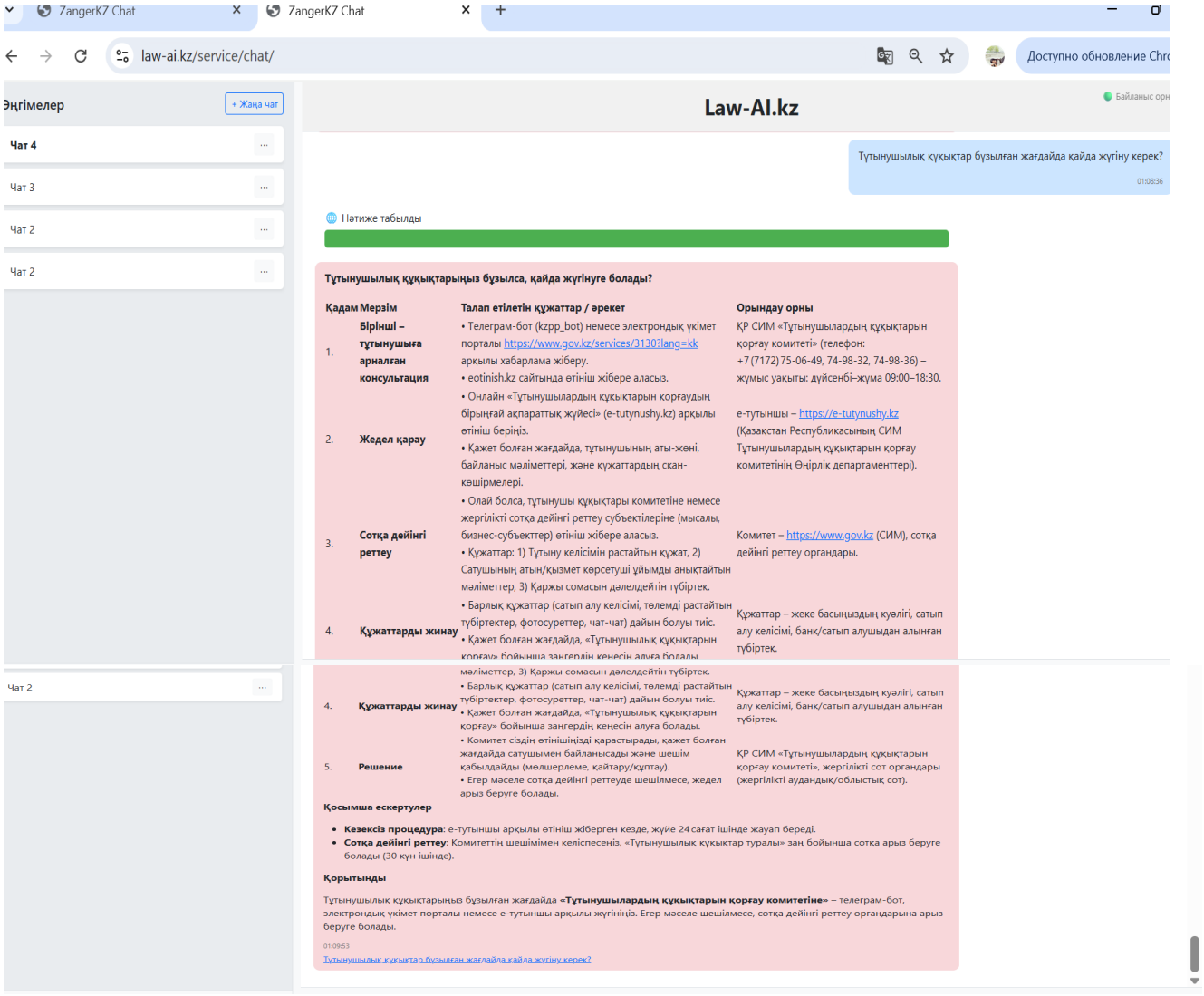
Морфологиялық салыстыру + TF-IDF матрицасы:

$X = tfidf(T\_d) \cdot log(N/df) + \mu \in R\^\{M \times N\}$;

Шығыс: Q_legal + z_q + кілт сөздер тіркесімі RAG модуліне берілді.

RAG іздеу жүйесі векторлық индекспен жұмыс істейді. Сұрақ enc(Q_legal) арқылы векторланады, косинустық ұқсастық sim(v_q, v_i) ≥ τ болса ішкі дерекқордан іздейді; болмаса adilet.zan.kz, «Реестр» тәрізді веб-ресурстарға сілтеме жасайды. Іріктеу кезінде α·BM25 + β·cos(v_q, v_i') формуласы бойынша top-k үзінділері таңдалып, Context C = {d_1, ..., d_K} қалыптасады.

LLM генераторы алынған контекст пен формализацияланған сұрақты бірге өңдейді: A = LLM(Q_legal, C; φ). Заңнамалық верификация сатысында V(A) = sim(A, C) ≥ τ_v шартымен нормаға сәйкестік тексеріледі. Шығыс ретінде тақырып, негіздеме, бап нөмірлері, НҚА сілтемелері мен пайдаланушыға арналған жауап ұсынылады.



Сурет 4.14 - Law-AI.kz жүйесінің интерфейсі

Law-ai.kz веб-порталы пайдаланушыға ыңғайлы, жауапты интерфейспен жабдықталғанын сурет 4.14-тен көруге болады. Платформаның негізгі беті үш аймаққа бөлінеді: сұрақ енгізу аймағы, жауап шығару аймағы және НҚА сілтемелері мен баптар тізімі.

Пайдаланушы қазақ тілінде еркін формада сұрақ жазады. Жүйе оны заңдық жанрға тиесілілігі тұрғысынан бағалайды. Заңға қатысты болса, SemanticPrep модулі арқылы өңдеп, RAG жүйесіне жіберіп, LLM жауабын генерациялайды. Заңдық сипаттамасы жоқ сұрақтарда: «Заңға қатысты сұрақ қойыңыз» деп хабарлайды.

Жауаптар мынадай бөліктерден тұрады: қысқа қорытынды жауап, заңнамалық негіздеме (нормативтік акт атауы, бап нөмірі), тікелей сілтеме

(adilet.zan.kz) және қосымша түсініктеме. Интерфейс мобильді және десктоп форматтарда жұмыс жасайды.

4.6 Әзірленген жүйенің практикалық апробациясы және қолданбалы нәтижелері

Апробация барысындағы нақты мысал: пайдаланушы «Тұтынушылық құқықтар бұзылған жағдайда қайда жүгіну керек?» деп сұрағанда, жүйе қадамдық жауап генерациялады:

1) Бірінші тұтынушыға арналған консультация (Telegram-бот, электрондық үкімет порталы);

2) Жедел қарау (қажетті деректер тізімі);

3) Сотқа дейінгі реттеу (нормативтік сілтемелермен);

4) Құжаттарды жинау.

Жауапта ҚР СИМ «Тұтынушылардың құқықтарын қорғау комитеті» телефондары мен жұмыс уақыты нақты берілді. Ұсынылған жүйенің тиімділігін бағалау үшін LawVision, GPT-4o, Gemini, Claude сынды заманауи жүйелермен салыстыру жүргізілді. BERTScore, ROUGE-L және Expert Score метрикалары қолданылды. 4.20-кестеде нәтижелер ұсынылған.

Кесте 4.20 – Law-AI жүйесін басқа жүйелермен салыстыру

Жүйе	BERTScore	ROUGE-L	Expert Score
LawVision	0.60	0.35	2.5
GPT-4o	0.67	0.58	2.8
Gemini	0.72	0.45	2.9
Claude	0.65	0.40	2.7
Law-AI	**0.87**	**0.63**	**3.8**

4.20-кестеден көрінетіндей, Law-AI жүйесі BERTScore бойынша 0.87 мәніне жетіп, бәсекелес жүйелерден орташа 20–27%-ға жоғары нәтиже көрсетті. ROUGE-L метрикасы бойынша 0.63 мән алынды, бұл GPT-4o-дан 8,6%-ға жоғары. Эксперттік баға 3.8/5 деңгейінде тіркелді.

Іздеу сапасын бағалау мақсатында MRR (Mean Reciprocal Rank) және Recall@5 метрикалары пайдаланылды. 4.21-кестеде Law-AI жүйесі Adilet және Параграф жүйелерімен салыстырылды.

Кесте 4.21 – Law-AI жүйесін Adilet/Параграф жүйелерімен іздеу сапасы бойынша салыстыру

Жүйе	Retrieval тәсілі	MRR	Recall@5
Adilet	BM25	0.32	0.25
Параграф	Elasticsearch	0.35	0.31
Law-AI	**Proposed Method**	**0.84**	**0.89**

4.21-кестеде көрсетілгендей, Law-AI жүйесінің MRR көрсеткіші 0.84 болып, Adilet жүйесінен 2,6 есе, Параграф жүйесінен 2,4 есе жоғары нәтиже

берді. Recall@5 метрикасы бойынша 0.89 мән алынды, бұл іздеу жүйесінің қоры ең маңызды 5 нормативтік үзіндінің 89%-ын дұрыс анықтайтынын дәлелдейді. Бұл нәтижелер SemanticPrep препроцессинг модулінің іздеу нақтылығына тікелей оң ықпал ететінін растайды.

Жұмыстың практикалық маңыздылығы жүйенің нақты ұйымдастырушылық процестерге апробация жасалумен дәлелденеді. 2025 жылдың қорытындысы бойынша «Алмалы корпорациясы» ЖШС-нің №1 «Арман» филиалына енгізу актісі ресімделді.

SemanticPrep тәсілі негізінде әзірленген заңнамалық сұрақ-жауап жүйесінің апробациясы 2026 жылы «ART SERVISE» ЖШС-де жүргізілді. Апробация барысында жүйенің заңнамалық сұрақтарға жауап беру дәлдігі, семантикалық сәйкестігі және құқықтық ақпаратты генерациялау сапасы тәжірибелік ортада тексерілді.

Law-AI жүйесі ұйымның қызметіне үш бағыт бойынша интеграцияланды:

- азаматтар мен қызметкерлерді Қазақстан Республикасының қолданыстағы заңнамасы бойынша қазақ тілінде консультациялау;
- нормативтік құқықтық актілерге қолжетімділікті оңтайландыру және өзекті құқықтық ақпаратты іздеу уақытын қысқарту;
- құқықтық ақпараттандыру тиімділігін арттыру және типтік сұрауларды өңдеу кезінде заң қызметкерлерінің жұмыс жүктемесін азайту.

Енгізу актісінде жобаның нәтижелері жоғары практикалық, ғылыми және әлеуметтік маңыздылыққа ие екені атап өтілген. Жүзеге асырылған жүйе цифрлық ортадағы мемлекеттік тіл саясатының дамуына ықпал етеді, заңтануда жасанды интеллект пен семантикалық технологиялардың мүмкіндіктерін көрсетеді, сондай-ақ халықтың цифрлық сауаттылығын арттыруды және кең пайдаланушылар тобы үшін заңнаманың қолжетімділігін қамтамасыз етеді.

Law-ai.kz платформасы мамандардың оң пікірлерін алды, нақты жүктеме жағдайында тұрақты жұмысын, архитектурасының ауқымдылығын және басқа ведомстволық және білім беру құрылымдарына одан әрі тарату дайындығын растады.

Law-AI заңдық анықтамалар алуды жылдамдатады, нормативтік актілерге қолжетімділікті кеңейтеді және мемлекеттік тілдегі цифрлық заңнамалық сервистердің дамуына ықпал етеді. Болашақта жүйені нотариат, азаматтарға қызмет көрсету орталықтары секілді мемлекеттік мекемелерде енгізу жоспарланған.

ҚОРЫТЫНДЫ

Осы диссертациялық жұмыста Қазақстан Республикасының заңнамасы саласында мемлекеттік тілге арналған интеллектуалды сұрақ-жауап жүйесін семантикалық тәсілдер негізінде зерттеу және әзірлеу мәселесі кешенді түрде қарастырылды. Жүргізілген ғылыми зерттеу нәтижесінде қойылған барлық міндеттер орындалып, диссертациялық жұмыстың мақсатына қол жеткізілді.

Зерттеудің бірінші бөлімінде заңнамалық сұрақ-жауап жүйелерінің теориялық және әдіснамалық негіздері талданды. Сұрақ-жауап жүйелерінің ережелерге негізделген қарапайым модельдерден бастап заманауи үлкен тілдік модельдер мен Retrieval-Augmented Generation архитектураларына дейінгі даму эволюциясы жүйелі түрде қарастырылды. Сонымен қатар қазақ тілінің агглютинативті морфологиялық құрылымы мен сөзтүрлендіру ерекшеліктерінің заңнамалық мәтіндерді автоматты өңдеу барысында қосымша күрделілік тудыратыны анықталды. Отандық және шетелдік құқықтық сұрақ-жауап жүйелерін салыстырмалы талдау нәтижесінде қолданыстағы жүйелердің басым бөлігі кілт сөздерге негізделген іздеумен шектелетіні және табиғи тілдегі сұраныстардың семантикалық мазмұнын толық деңгейде интерпретациялай алмайтыны көрсетілді.

Екінші бөлімде Қазақстан Республикасының заңнамалық деректер кеңістігін қалыптастыру және үлкен тілдік модельдерді құқықтық доменге бейімдеу мәселелері зерттелді. Зерттеу барысында ресми құқықтық порталдардан жинақталған 5.3 ГБ көлеміндегі заңнамалық құжаттар мен 8 000-нан астам сұрақ-жауап жұбынан тұратын мамандандырылған деректер жиыны құрылды. Қалыптастырылған ресурс қазақ тіліндегі заңнамалық NLP зерттеулеріне арналған кешенді деректер базасының бірі болып табылады. Сонымен қатар RAG архитектурасын заңнамалық сұрақ-жауап жүйесіне интеграциялаудың теориялық негіздері мен практикалық ұйымдастырылу принциптері қарастырылды. Жүргізілген эксперименттер тек LLM модельдерін fine-tuning арқылы бейімдеу немесе тек RAG архитектурасын енгізу retrieval сапасын белгілі бір деңгейде жақсартқанымен, пайдаланушының табиғи тілдегі сұраныстары мен заңнамалық терминология арасындағы семантикалық сәйкессіздік мәселесін толық шешпейтінін көрсетті.

Үшінші бөлімде диссертациялық жұмыстың негізгі ғылыми нәтижелерінің бірі болып табылатын SemanticPrep екі кезеңді семантикалық препроцессинг тәсілі ұсынылды. Ұсынылған тәсілдің ғылыми негізі табиғи тілдегі пайдаланушы сұраныстарын retrieval кезеңіне дейін құқықтық контекске бейімделген семантикалық құрылымға түрлендіруге бағытталды. Бірінші кезеңде mT5 трансформері негізінде табиғи тілдегі сұраныстарды заңнамалық формализацияланған құрылымға келтіру моделі әзірленді. Екінші кезеңде SVD/LSA негізіндегі семантикалық кеңейту алгоритмі іске асырылып, сұраныстың мағыналық кеңістігін кеңейту арқылы ақпараттық іздеу сапасын арттыру мүмкіндігі қамтамасыз етілді. Ұсынылған тәсілдің тиімділігі математикалық модельдеу, есептеу күрделілігін талдау және заманауи LLM

модельдерімен салыстырмалы эксперименттік бағалау арқылы теориялық және практикалық тұрғыдан негізделді.

Төртінші бөлімде ұсынылған тәсілдің эксперименттік бағалануы жүргізілді. Зерттеу аясында базалық LLM модельдерінің zero-shot режиміндегі нәтижелері, fine-tuning арқылы бейімделген модельдердің көрсеткіштері және SemanticPrep, RAG және LLM технологияларын біріктіретін кешенді пайплайнның тиімділігі салыстырмалы түрде зерттелді. Эксперименттік нәтижелер ұсынылған тәсілдің baseline жүйелермен салыстырғанда заңнамалық сұрақ-жауап міндеттерін орындауда жоғары тиімділікке ие екенін көрсетті. Атап айтқанда, ұсынылған жүйе ROUGE-L көрсеткішін 15-20%-ға, BERTScore мәнін 3–5%-ға жақсартты. Сонымен қатар retrieval сапасын сипаттайтын Recall@5 көрсеткіші 11–15%-ға, MRR метрикасы 16-18%-ға артты. Сараптамалық бағалау нәтижелері де жауаптардың семантикалық дәлдігі мен құқықтық негізділігінің айтарлықтай жоғарылағанын көрсетті.

Жүргізілген зерттеу нәтижесінде қазақ тілінің морфологиялық ерекшеліктері мен құқықтық терминологияның спецификасын ескеретін екі кезеңді семантикалық препроцессинг тәсілі әзірленді. Сонымен қатар семантикалық алдын ала өңдеу, ақпараттық іздеу және бейімделген үлкен тілдік модельдерді біріктіретін интеллектуалды заңнамалық сұрақ-жауап жүйесінің кешенді архитектурасы ұсынылып, оның тиімділігі халықаралық бағалау метрикалары негізінде эксперименттік түрде дәлелденді.

Диссертациялық жұмыстың ғылыми нәтижелері халықаралық және республикалық ғылыми конференцияларда баяндалып, ғылыми қауымдастық тарапынан апробациядан өтті. Зерттеу нәтижелері бойынша Web of Science және Scopus дерекқорларында индекстелетін халықаралық басылымдарда, сондай-ақ Қазақстан Республикасы Ғылым және жоғары білім министрлігінің ұсынған ғылыми журналдарында ғылыми мақалалар жарияланды. Атап айтқанда, диссертация нәтижелері негізінде Web of Science (SCIE, Q1) базасына енетін *Computers (MDPI)* журналында, Springer баспасының *Communications in Computer and Information Science (CCIS)* сериясында, ACIIDS 2025 және UBMK 2025 халықаралық конференцияларында, сондай-ақ ҚР ҒЖБМ ҒЖБССҚ ұсынған ғылыми журналдарында жарияланымдар жарық көрді. Бұл зерттеу нәтижелерінің халықаралық деңгейде ғылыми апробациядан өткенін және ұсынылған тәсілдердің ғылыми негізділігі мен практикалық маңыздылығын растайды.

Диссертациялық жұмыстың практикалық маңыздылығы ұсынылған интеллектуалды жүйені мемлекеттік цифрлық платформаларға, құқықтық ақпараттық сервистерге, электрондық үкімет жүйелеріне және заңнамалық консультациялық платформаларға интеграциялау мүмкіндігімен анықталады. Әзірленген Law-AI жүйесі , құқықтық анықтамалық платформалар, мемлекеттік онлайн-консультациялық сервистер және азаматтарға арналған цифрлық құқықтық көмек жүйелері аясында қолданылуы мүмкін. Сонымен қатар зерттеу нәтижелері мемлекеттік тілдегі құқықтық ақпаратқа қолжетімділікті кеңейтуге,

азаматтардың заңнамалық сауаттылығын арттыруға және қазақ тілінің цифрлық технологиялар кеңістігіндегі қолданылу аясын дамытуға бағытталған.

ПАЙДАЛАНЫЛҒАН ӘДЕБИЕТТЕР ТІЗІМІ

1. Rakhimova D., Karyukin V., Karibayeva A., Sarsenbayeva A.K. Developing advanced question-answering models for Legal Kazakh texts: A comparative study of modern approaches. ACIIDS 2025: 17th Asian Conference on Intelligent Information and Database Systems. DOI: 10.1007/978-981-96-5881-7_19 (Scopus, процентиль 39)
2. Сарсенбаева А.К., Рахимова Д.Р., Шормакова А.Н., Мансурова М. Агглютинативті тілдердің заңды мәтіндеріндегі мағыналық ұқсастықты анықтаудың контекстке негізделген тәсілі: қазақ тіліндегі жағдайды зерттеу. Қазақстан Республикасы Ұлттық ғылым академиясының Хабаршысы, - физика математикалық сериясы, № 1, Алматы 2026 – С. 354-372 https://doi.org/10.32014/2026.2518-1726.417
3. D.Rakhimova, A.Sarsenbayeva, E. Adali, A. Kulzhanova, Classification of questions in the intellectual question-answer system for the state language, Вестник НИА РК №4 Алматы, 2024, – С. 166-179 (КОКСНВО МНВО РК) https://doi.org/10.47533/2024.1606-146X.014
4. Rakhimova D., Karyukin V., Turarbek A., Sarsenbayeva A.K., Alieyev R. (2025). Legal AI in Low-Resource Languages: Building and Evaluating QA Systems for the Kazakh Legislation. Computers (MDPI). 14(9), 354. https://doi.org/10.3390/computers14090354 WoS (SCIE), Q1, 2025; 5-Year IF: 2.7; SJR: 0.492; Percentile: 84 (Computer Science, Applications)
5. Green Jr B. F. et al. Baseball: an automatic question-answerer //Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference. – 1961. – С. 219-224.
6. Weizenbaum J. ELIZA — A Computer Program for the Study of Natural Language Communication Between Man and Machine // Communications of the ACM. – 1966. – Vol. 9(1). – P. 36–45.
7. Dang H. T. et al. Overview of the TREC 2007 Question Answering Track //Trec. – 2007. – Т. 7. – С. 63.
8. Rajpurkar P. et al. Squad: 100,000+ questions for machine comprehension of text //Proceedings of the 2016 conference on empirical methods in natural language processing. – 2016. – С. 2383-2392.
9. Vaswani A. et al. Attention is all you need //Advances in neural information processing systems. – 2017. – Т. 30.
10. Devlin J. et al. Bert: Pre-training of deep bidirectional transformers for language understanding //Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). – 2019. – С. 4171-4186.
11. Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog. – 2019. – Т. 1. – №. 8. – С. 9.
12. Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //Journal of machine learning research. – 2020. – Т. 21. – №. 140. – С. 1-67.

13. Yang Z. et al. Xlnet: Generalized autoregressive pretraining for language understanding //Advances in neural information processing systems. – 2019. – T. 32.
14. Liu Y. et al. Roberta: A robustly optimized bert pretraining approach //arXiv preprint arXiv:1907.11692. – 2019.
15. Lan Z. et al. Albert: A lite bert for self-supervised learning of language representations //arXiv preprint arXiv:1909.11942. – 2019.
16. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in neural information processing systems. – 2020. – T. 33. – C. 9459-9474.
17. Gao Y. et al. Retrieval-augmented generation for large language models: A survey //arXiv preprint arXiv:2312.10997. – 2023. – T. 2. – №. 1. – C. 32.
18. Chalkidis I. et al. LexGLUE: A benchmark dataset for legal language understanding in English //Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2022. – C. 4310-4330.
19. Dai Y. et al. LAiW: A Chinese legal large language models benchmark //Proceedings of the 31st International conference on computational linguistics. – 2025. – C. 10738-10766.
20. Guha N. et al. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models //Advances in neural information processing systems. – 2023. – T. 36. – C. 44123-44279.
21. Fei Z. et al. Lawbench: Benchmarking legal knowledge of large language models //Proceedings of the 2024 conference on empirical methods in natural language processing. – 2024. – C. 7933-7962.
22. Louis A., Van Dijck G., Spanakis G. Interpretable long-form legal question answering with retrieval-augmented large language models //Proceedings of the AAAI conference on artificial intelligence. – 2024. – T. 38. – №. 20. – C. 22266-22275.
23. Dahl M. et al. Large legal fictions: Profiling legal hallucinations in large language models //Journal of Legal Analysis. – 2024. – T. 16. – №. 1. – C. 64-93.
24. Gao A. McCoy Family Center for Ethics in Society. At the Stanford Institute for Human-Centered Artificial Intelligence (HAI), he works at the intersection of AI, the social sector, and the Global South. His research interests include the political economy of emerging economies and its relationships with political philosophy and ethics. He holds a bachelor's degree in philosophy and economics from New York University.
25. Butler A. R., Butler U. Legal RAG Bench: an end-to-end benchmark for legal RAG //arXiv preprint arXiv:2603.01710. – 2026.
26. Niklaus J. et al. Multilegalpile: A 689gb multilingual legal corpus //Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2024. – C. 15077-15094.

27. Makhambetov O. et al. Data-driven morphological analysis and disambiguation for kazakh //International Conference on Intelligent Text Processing and Computational Linguistics. – Cham : Springer International Publishing, 2015. – С. 151-163.
28. Yessenbayev Z., Kozhirbayev Z., Makazhanov A. KazNLP: A pipeline for automated processing of texts written in Kazakh language //International Conference on Speech and Computer. – Cham : Springer International Publishing, 2020. – С. 657-666.
29. Kessikbayeva G., Cicekli I. Rule based morphological analyzer of Kazakh language //Proceedings of the 2014 Joint Meeting of SIGMORPHON and SIGFSM. – 2014. – С. 46-54.
30. UD Kazkh KTB. https://universaldependencies.org/treebanks/kk_ktb/ (25.02.2022)
31. Mansurova M. E., Rakhimova D. R. Morphological parsing of Kazakh texts with deep learning approaches //Journal of Mathematics, Mechanics and Computer Science. – 2024. – Т. 124. – №. 4. – С. 48-58.
32. Информационно-правовая систему «Әділет». https://adilet.zan.kz/kaz/ (Дата обращения: 26.09.2025)
33. Портал электронного правительства РК. https://egov.kz/cms/kk (Дата обращения: 01.10.2025)
34. Эталонный контрольный банк нормативных правовых актов. https://zan.gov.kz/client/# (Дата обращения: 01.10.2025)
35. Эталонный контрольный банк нормативных правовых актов. https://prg.kz/
36. Институт законодательства и правовой информации. https://www.zqai.kz/ru/questions? (Дата обращения: 01.10.2025)
37. Официальный сайт Президента РК. https://www.akorda.kz/kz (Дата обращения: 01.10.2025)
38. Rakhimova D., Kassymova D., Isabaeva D. Исследование и разработка системы вопрос-ответ на основе модели bert для казахского языка //Bulletin of Abai KazNPU. Series of Physical and Mathematical sciences. – 2021. – Т. 76. – №. 4. – С. 119-127.
39. Xiao C. et al. Lawformer: A pre-trained language model for chinese legal long documents //AI Open. – 2021. – Т. 2. – С. 79-84.
40. Faisal F., Yousaf U. Legal-uqa: A low-resource urdu-english dataset for legal question answering //arXiv preprint arXiv:2410.13013. – 2024.
41. Yeshpanov R. et al. KazQAD: Kazakh open-domain question answering dataset //Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). – 2024. – С. 9645-9656.
42. Lewis P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks //Advances in neural information processing systems. – 2020. – Т. 33. – С. 9459-9474.

43. Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //Journal of machine learning research. – 2020. – Т. 21. – №. 140. – С. 1-67.
44. О правовых актах. https://adilet.zan.kz/rus/docs/Z1600000480 (Дата обращения:01.10.2025)
45. Zhekova M., Yumer E. JavaScript Web Scraping Tool for Extraction Information from Agriculture Websites //BIO Web of Conferences. – EDP Sciences, 2024. – Т. 102. – С. 03008.
46. Volkov A. S., Chernenky M. V. Generalized algorithm for website parsing //2024 6th International Youth Conference on Radio Electronics, Electrical and Power Engineering (REEPE). – IEEE, 2024. – С. 1-5.
47. Bisht V. et al. Utilizing python for web scraping and incremental data extraction //2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS). – IEEE, 2023. – С. 1450-1455.
48. Sewunetie W. T., Kovács L. Automatic question generation based on sentence structure analysis //2023 24th International Carpathian Control Conference (ICCC). – IEEE, 2023. – С. 411-414.
49. Meng W. et al. Using the retrieval-augmented generation to improve the question-answering system in human health risk assessment: the development and application //Electronics. – 2025. – Т. 14. – №. 2. – С. 386.
50. Rakhimova D. et al. Collection and Preprocessing of Data for LLM in the Kazakh Language in the Field of Legislation //International Conference on Modelling and Simulation of Social-Behavioural Phenomena in Creative Societies. – Cham : Springer Nature Switzerland, 2024. – С. 129-139.
51. Рахимова Д. и др. Применение глубокого обучения для повышения точности ответов в мультиязычных вопросно ответных системах //Academic Scientific Journal of Computer Science. – 2025. – №. 2. – С. 196–211-196–211.
52. Touvron H., Lavril T., Izacard G. et al. *LLaMA: Open and Efficient Foundation Language Models* // arXiv preprint arXiv:2302.13971. – 2023.
53. Abdin M. et al. Phi-4 technical report //arXiv preprint arXiv:2412.08905. – 2024.
54. Bai J. et al. Qwen technical report. arXiv 2023 //arXiv preprint arXiv:2309.16609. – 2023.
55. Jiang Y. et al. 6G non-terrestrial networks enabled low-altitude economy: Opportunities and challenges //arXiv preprint arXiv:2311.09047. – 2023.
56. Pipitone N., Alami G. H. Legalbench-rag: A benchmark for retrieval-augmented generation in the legal domain //arXiv preprint arXiv:2408.10343. – 2024.
57. Barron R. C. et al. Bridging legal knowledge and AI: retrieval-augmented generation with vector stores, Knowledge graphs, and hierarchical non-negative matrix factorization //arXiv preprint arXiv:2502.20364. – 2025.
58. Köse S. K. et al. RAGTurk: Best Practices for Retrieval Augmented Generation in Turkish //Proceedings of the Second Workshop Natural Language Processing for Turkic Languages (SIGTURK 2026). – 2026. – С. 179-196.

59. Schwarcz D. et al. Ai-powered lawyering: Ai reasoning models, retrieval augmented generation, and the future of legal practice //Journal of Law & Empirical Analysis. – 2025. – С. 2755323X261427048.
60. Robertson S., Zaragoza H. The probabilistic relevance framework: BM25 and beyond. – Now Publishers Inc, 2009. – Т. 4.
61. Karpukhin V. et al. Dense passage retrieval for open-domain question answering //Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP). – 2020. – С. 6769-6781.
62. Douze M. et al. The faiss library //IEEE Transactions on Big Data. – 2025.
63. Brown T. et al. Language models are few-shot learners //Advances in neural information processing systems. – 2020. – Т. 33. – С. 1877-1901.
64. Garbacea C., Mei Q. Neural language generation: Formulation, methods, and evaluation //arXiv preprint arXiv:2007.15780. – 2020.
65. Gao T., Yao X., Chen D. Simcse: Simple contrastive learning of sentence embeddings //Proceedings of the 2021 conference on empirical methods in natural language processing. – 2021. – С. 6894-6910.
66. Schwarcz D. et al. Ai-powered lawyering: Ai reasoning models, retrieval augmented generation, and the future of legal practice //Journal of Law & Empirical Analysis. – 2025. – С. 2755323X261427048.
67. Reuter M. et al. Towards reliable retrieval in RAG systems for large legal datasets //Proceedings of the Natural Legal Language Processing Workshop 2025. – 2025. – С. 17-30.
68. Kim D. et al. GuRE: Generative Query REwriter for Legal Passage Retrieval //Proceedings of the Natural Legal Language Processing Workshop 2025. – 2025. – С. 424-438.
69. Maron M. E., Kuhns J. L. On relevance, probabilistic indexing and information retrieval //Journal of the ACM (JACM). – 1960. – Т. 7. – №. 3. – С. 216-244.
70. Robertson S., Zaragoza H. The probabilistic relevance framework: BM25 and beyond. – Now Publishers Inc, 2009. – Т. 4.
71. Zheng L. et al. A reasoning-focused legal retrieval benchmark //Proceedings of the 2025 Symposium on Computer Science and Law. – 2025. – С. 169-193.
72. Baykara B., Güngör T. Abstractive text summarization and new large-scale datasets for agglutinative languages Turkish and Hungarian //Language Resources and Evaluation. – 2022. – Т. 56. – №. 3. – С. 973-1007.
73. Sarsenbayeva A. et al. Multilingual Large Language Models in the Legal Domain: An Effectiveness Analysis in Kazakh, Turkish and English //2025 10th International Conference on Computer Science and Engineering (UBMK). – IEEE, 2025. – С. 410-415.
74. Eryiğit G. ITU Turkish NLP web service //Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics. – 2014. – С. 1-4.
75. Hakkani-Tür D. Z., Oflazer K., Tür G. Statistical morphological disambiguation for agglutinative languages //Computers and the Humanities. – 2002. – Т. 36. – №. 4. – С. 381-410.

76. Ma X. et al. Query rewriting in retrieval-augmented large language models //Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. – 2023. – С. 5303-5315.
77. Mao S. et al. RaFe: Ranking feedback improves query rewriting for RAG //Findings of the Association for Computational Linguistics: EMNLP 2024. – 2024. – С. 884-901.
78. Reuter M. et al. Towards reliable retrieval in RAG systems for large legal datasets //Proceedings of the Natural Legal Language Processing Workshop 2025. – 2025. – С. 17-30.
79. Deerwester S. et al. Indexing by latent semantic analysis //Journal of the American society for information science. – 1990. – Т. 41. – №. 6. – С. 391-407.
80. Zhu G. et al. Design of knowledge graph retrieval system for legal and regulatory framework of multilevel latent semantic indexing //Computational Intelligence and Neuroscience. – 2022. – Т. 2022. – №. 1. – С. 6781043.
81. Raffel C. et al. Exploring the limits of transfer learning with a unified text-to-text transformer //Journal of machine learning research. – 2020. – Т. 21. – №. 140. – С. 1-67.
82. Xue L. et al. mT5: A massively multilingual pre-trained text-to-text transformer //Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies. – 2021. – С. 483-498.
83. Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate //arXiv preprint arXiv:1409.0473. – 2014.
84. Cho K. et al. Learning phrase representations using RNN encoder–decoder for statistical machine translation //Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). – 2014. – С. 1724-1734.
85. Zhang T. et al. Bertscore: Evaluating text generation with bert //arXiv preprint arXiv:1904.09675. – 2019.
86. Qi Q. et al. A comparison of latent semantic analysis and correspondence analysis of document-term matrices //Natural Language Engineering. – 2024. – Т. 30. – №. 4. – С. 722-752.
87. Qi Q., Hessen D. J., Van Der Heijden P. G. M. Improving information retrieval through correspondence analysis instead of latent semantic analysis //Journal of Intelligent Information Systems. – 2024. – Т. 62. – №. 1. – С. 209-230.
88. Onah D. F. O., Pang E. L. L., El-Haj M. A data-driven latent semantic analysis for automatic text summarization using LDA topic modelling //2022 IEEE International Conference on Big Data (Big Data). – IEEE, 2022. – С. 2771-2780.
89. Hofmann T. Probabilistic latent semantic analysis //arXiv preprint arXiv:1301.6705. – 2013.
90. Landauer T. K., Dumais S. T. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge //Psychological review. – 1997. – Т. 104. – №. 2. – С. 211.
91. Yu B., Xu Z., Li C. Latent semantic analysis for text categorization using neural network //Knowledge-Based Systems. – 2008. – Т. 21. – №. 8. – С. 900-904.

92. Harris Z. S. Distributional structure //Word. – 1954. – Т. 10. – №. 2-3. – С. 146-162.
93. Manning C. D. Introduction to information retrieval. – Syngress Publishing,, 2008.
94. Kanapala A., Jannu S., Pamula R. Passage-based text summarization for legal information retrieval //Arabian Journal for Science and Engineering. – 2019. – Т. 44. – №. 11. – С. 9159-9169.
95. Sparck Jones K. A statistical interpretation of term specificity and its application in retrieval //Journal of documentation. – 1972. – Т. 28. – №. 1. – С. 11-21.
96. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. – 2013. – Т. 26.
97. Reimers N., Gurevych I. Sentence-bert: Sentence embeddings using siamese bert-networks //Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP). – 2019. – С. 3982-3992.
98. Cer D. et al. Universal sentence encoder //arXiv preprint arXiv:1803.11175. – 2018.
99. Turney P. D., Pantel P. From frequency to meaning: Vector space models of semantics //Journal of artificial intelligence research. – 2010. – Т. 37. – С. 141-188.
100. Eckart C., Young G. The approximation of one matrix by another of lower rank //Psychometrika. – 1936. – Т. 1. – №. 3. – С. 211-218.
101. Mirsky L. Symmetric gauge functions and unitarily invariant norms //The quarterly journal of mathematics. – 1960. – Т. 11. – №. 1. – С. 50-59.
102. Gao L. et al. Precise zero-shot dense retrieval without relevance labels //Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2023. – С. 1762-1777.
103. Wang L., Yang N., Wei F. Query2doc: Query expansion with large language models //Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. – 2023. – С. 9414-9423.
104. Conneau A. et al. Unsupervised cross-lingual representation learning at scale //Proceedings of the 58th annual meeting of the association for computational linguistics. – 2020. – С. 8440-8451.
105. Landis J. R., Koch G. G. The measurement of observer agreement for categorical data //biometrics. – 1977. – С. 159-174.
106. McHugh M. L. Interrater reliability: the kappa statistic //Biochemia medica. – 2012. – Т. 22. – №. 3. – С. 276-282.
107. Byrt T., Bishop J., Carlin J.B. Bias, Prevalence and Kappa // Journal of Clinical Epidemiology. – 1993. – Vol. 46, № 5. – P. 423–429. – DOI: 10.1016/0895-4356(93)90018-V
108. Opitz J. From Bias and Prevalence to Macro F1, Kappa, and MCC: A Structured Overview of Metrics for Multi-class Evaluation // Proceedings of the

29th International Conference on Computational Linguistics (COLING 2022). – 2022. – URL: https://aclanthology.org/2022.coling-1.40

ҚОСЫМША А Ендіру актісі

1.

ЖАУАПКЕРШІЛІГІ ШЕКТЕУЛІ СЕРІКТЕСТІК

ТОВАРИЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ

ART Servise

Юридический адрес: Республика Казахстан, г.Шымкент, район Каратау, мкр. Нуртас, 4582/3.
БИН 250 940 014 845. KZ4196516F0012624414 в АО "ForteBank"

№ 7 «15» сәуір 2026 ж.

ЕНДІРУ АКТІСІ

ҒЫЛЫМИ-ЗЕРТТЕУ ЖОБАСЫНЫҢ НӘТИЖЕЛЕРІН ЕНДІРУ ТУРАЛЫ

Осы акт «Қазақстан Республикасының заңнамасы саласындағы мемлекеттік тілге арналған семантикалық тәсілдер негізінде модельдерді зерттеу және интеллектуалды сұрақ-жауап жүйесін әзірлеу» тақырыбындағы ғылыми-зерттеу жұмысының нәтижелері «ART Servise» жауапкершілігі шектеулі серіктестігінің қызметіне ендірілгені туралы жасалды.

1. Жоба авторлары мен орындаушылары

Рахимова Диана Рамазановна — философия докторы (PhD), қауымдастырылған профессор, әл-Фараби атындағы Қазақ ұлттық университеті, Ақпараттық жүйелер кафедрасы, ғылыми жетекші;

Сарсенбаева Асия Калисеитовна — докторант (PhD), әл-Фараби атындағы Қазақ ұлттық университеті, 8D06101 – Ақпараттық жүйелер мамандығы, жоба орындаушысы.

2. Жобаның мақсаты

Жоба Қазақстан Республикасының заңнамасы саласында мемлекеттік тілде ақпаратты автоматты іздеуді, өңдеуді және ұсынуды қамтамасыз ететін интеллектуалды сұрақ-жауап жүйесін зерттеу және әзірлеуге бағытталған. Негізгі мақсат — жасанды интеллект технологияларын қолдана отырып, заңнамалық білімнің қолжетімділігін арттыру және мемлекеттік-құқықтық ақпаратты цифрландыру.

3. Ендіру мазмұны мен нәтижелері

Ғылыми-зерттеу жұмысының нәтижесінде заңнамалық сұрақ-жауап жүйесі (Law-AI.kz платформасы) әзірленіп, «ART Servise» ЖШС-нің практикалық қызметінде апробациядан өтті. Жүйе мынадай негізгі компоненттерді қамтиды:

1) SemanticPrep семантикалық препроцессинг модулі — пайдаланушының қарапайым тілдегі сұрағын mT5 трансформері арқылы заңнамалық стильге түрлендіреді, содан кейін SVD/LSA алгоритмі арқылы заңнамалық терминдермен семантикалық кеңейту жасайды;
2) RAG (Retrieval-Augmented Generation) іздеу модулі — Қазақстан Республикасының нормативтік-құқықтық актілер базасынан ең сәйкес заңнамалық үзінділерді іріктейді;
3) LLM жауап генерациясы модулі — іріктелген контекст пен формализацияланған сұрақ негізінде НҚА баптарына сілтемемен дәлелді заңнамалық жауап қалыптастырады;
4) Пайдаланушы интерфейсі — заңнамалық сұрақтарға жауап алуға арналған чат-бот платформасы (https://law-ai.kz).

Апробация барысында жүйе заңнамалық сұрақтарға дәлелді және дәл жауап беру қабілетін тәжірибе жүзінде дәлелдеді, ұйымның анықтамалық-кеңес беру жұмысының тиімділігін арттырды.

4. Практикалық қолдану

Ендірілген жүйе «ART Servise» ЖШС-нің қызметінде мынадай бағыттарда пайдаланылады: заңнама мәселелері бойынша пайдаланушылардың сұрақтарына автоматты жауап беру; нормативтік-құқықтық актілерге қолжетімділікті жеңілдету; ішкі құжат айналымы мен анықтамалық қызметтерді цифрландыру.

5. Ендірудің тиімділігі мен маңыздылығы

Жүйені ендіру заңнамалық процестерді автоматтандыру деңгейін арттыруға, мемлекеттік тілдің цифрлық ортада дамуына және лингвистикалық талдаудың жаңа құралдарын қалыптастыруға жәрдемдесті. Платформа іздеу және кеңес беру жұмысының дәлдігі мен тұрақтылығын қамтамасыз ететін тиімді құрал ретінде өзін көрсетті.

6. Ендіруді растау

Осымен расталады: **«Қазақстан Республикасының заңнамасы саласындағы мемлекеттік тілге арналған семантикалық тәсілдер негізінде модельдерді зерттеу және интеллектуалды сұрақ-жауап жүйесін әзірлеу»** ғылыми-зерттеу жұмысының нәтижелері «ART Servise» Жауапкершілігі шектеулі серіктестігінің қызметіне ендіріліп, ұйымның практикалық және ұйымдастыру-ақпараттық жұмысында қолданылуда.

«ART Servise» ЖШС директоры _____ **Б.А.Карабаев**

М.О.

2.

«КОРПОРАЦИЯ АЛМАЛЫ»
ЖАУАПКЕРШІЛІГІ ШЕКТЕУЛІ
СЕРІКТЕСТІГІ
«АРМАН»
№1 ФИЛИАЛЫ

ФИЛИАЛ №1 «АРМАН»
ТОВАРИЩЕСТВО
С ОГРАНИЧЕННОЙ
ОТВЕТСТВЕННОСТЬЮ
«КОРПОРАЦИЯ АЛМАЛЫ»

050000, Алматы қаласы, Желтоқсан к-сі. 118
050000, г.Алматы, ул.Желтоксан 118
тел.: +7 (727) 272-34-90

№ 047 «23» октября 2025г.

АКТ ВНЕДРЕНИЯ

РЕЗУЛЬТАТОВ НАУЧНО-ИССЛЕДОВАТЕЛЬСКОГО ПРОЕКТА

Настоящий акт составлен о том, что результаты научно-исследовательского проекта № AP19677835 «Исследование моделей и разработка интеллектуальной вопросно-ответной системы на основе семантических подходов для государственного языка в сфере законодательства Республики Казахстан», выполненного в рамках грантового финансирования, внедрены в деятельность Филиала №1 «Арман» ТОО «Корпорация Алмалы»

1. Автор и разработчик проекта

PhD, ассоциированный профессор Рахимова Диана Рамазановна, Казахский национальный университет имени аль-Фараби, является автором и разработчиком интеллектуальной вопросно-ответной системы.

2. Цель проекта

Проект направлен на разработку и исследование интеллектуальной вопросно-ответной системы, обеспечивающей автоматический поиск, обработку и предоставление правовой информации на казахском языке в сфере законодательства Республики Казахстан. Основная цель — повышение доступности юридических знаний и цифровизация государственной правовой информации с применением технологий искусственного интеллекта.

3. Содержание и результаты внедрения

В результате научно-исследовательской работы создана и внедрена интеллектуальная чат-бот система – онлайн-платформа https://law-ai.kz, предназначенная для предоставления пользователям быстрых и точных ответов на юридические вопросы на казахском языке. Система включает модули семантического поиска, обработки естественного языка, модель машинного обучения и интерфейс чат-бота.

4. Практическое применение

по вопросам законодательства, оптимизации доступа к нормативно-правовым актам, повышения эффективности внутреннего документооборота и цифровизации справочных сервисов.

5. Эффективность и значимость внедрения

Внедрение интеллектуальной системы Law-AI.kz способствует повышению уровня автоматизации юридических процессов, развитию применения государственного языка в цифровой среде, формированию новых инструментов лингвистического анализа и семантического поиска. Платформа зарекомендовала себя как эффективный инструмент, обеспечивающий точность и устойчивость поисковой и консультативной работы.

6. Подтверждение внедрения

Настоящим подтверждается, что результаты проекта № AP19677835 внедрены в деятельность Филиала №1 «Арман» ТОО «Корпорация Алмалы» и используются в практической и организационно-информационной работе учреждения.

Директор Филиала №1 «Арман» ТОО «Корпорация Алмалы» Давлетов Д.Е.