

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Державне некомерційне підприємство
«Державний університет» Київський авіаційний інститут»

Факультет комп'ютерних наук та технологій

Кафедра інженерії програмного забезпечення

ДОПУСТИТИ ДО ЗАХИСТУ
Завідувач кафедри

_____ Олена ГРІНЕНКО

« _____ » _____ 2025 р.

КВАЛІФІКАЦІЙНА РОБОТА (ПОЯСНЮВАЛЬНА ЗАПИСКА)

ЗДОБУВАЧА ОСВІТНЬОГО СТУПЕНЯ «МАГІСТР»

Тема: ШІ-асистент для підготовки відповідей на запити на публічну інформацію до Міністерства соціальної політики, сім'ї та єдності України

Виконавець: Семенюк Гліб Сергійович

Керівник: к.ф.-м.н., доцент Татаринів Євген Олександрович

Нормоконтролер: ас. каф. Андреева Тетяна Василівна

Київ 2025

**Державне некомерційне підприємство
«Державний університет» Київський авіаційний інститут»**

Факультет комп'ютерних наук та технологій
Кафедра інженерії програмного забезпечення
Спеціальність 121 «Інженерія програмного забезпечення»
Освітньо-професійна програма «Інженерія програмного забезпечення»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олена ГРІНЕНКО

«_____» _____ 2025 р.

ЗАВДАННЯ
на виконання кваліфікаційної роботи студента
Семенюка Гліба Сергійовича

1. Тема кваліфікаційної роботи: «ШІ-асистент для підготовки відповідей на запити на публічну інформацію до Міністерства соціальної політики, сім'ї та єдності України»

затверджена наказом ректора від _____ 17.11.2025 р. № 2450/ст

2. Термін виконання проекту: з 29.09.2025 р. по 21.12.2025 р.

3. Вихідні дані до роботи: дослідити ефективність адаптації великих мовних моделей з відкритим кодом для автоматизованої генерації офіційних відповідей на запити громадян.

4. Зміст пояснювальної записки:

1. Аналіз предметної області та огляд існуючих рішень.
2. Постановка задачі та обґрунтування методів і засобів її вирішення.
3. Експериментальне дослідження.

5. Перелік обов'язкового графічного (ілюстративного) матеріалу:

1. Структурна схема системи обробки запитів на основі підходу Human-in-the-Loop.
2. Матриця багатокритеріального відбору базових моделей-кандидатів.
3. Схема алгоритму підготовки даних та доналаштування моделей методом QLoRA.
4. Діаграми порівняльного аналізу якості генерації моделей.
5. Порівняльна характеристика ресурсомісткості та швидкодії моделей.
6. Фрагменти програмної реалізації та приклади генерації відповідей.

6. Календарний план-графік

№ пор.	Завдання	Термін виконання	Відмітка про виконання
1.	Розробка та затвердження графіка роботи	29.09-01.10.2025	виконано
2.	Ознайомлення з постановкою задачі, вивчення інформаційних джерел та складання плану роботи.	02.10-05.10.2025	виконано
3.	Підготовка 1 розділу та подання його керівнику	06.10-25.10.2025	виконано
4.	Підготовка 2 розділу та подання його керівнику	26.10-10.11.2025	виконано
5.	Підготовка 3 розділу і висновків по роботі та подання їх керівнику	11.11-30.11.2025	виконано
6.	Загальне редагування пояснювальної записки, графічного матеріалу. Представлення роботи для перевірки на академічну доброчесність.	01.12-10.12.2025	виконано
7.	Проходження нормоконтролю.	05.12-14.12.2025	
8.	Отримання відгуку керівника. Підготовка презентації та тексту доповіді.	05.12-19.12.2025	виконано
9.	Попередній захист (представлення електронної версії пояснювальної записки, презентації, позитивного відгуку керівника).	05.12-19.12.2025	виконано
10.	Рецензування кваліфікаційної роботи	05.12-22.12.2025	виконано
11.	Здача секретарю ЕК пояснювальної записки: електронної версії кваліфікаційної роботи; презентації доповіді; відгуку керівника, рецензії; результату проходження перевірки на плагіат; довідки про успішність, декларації про академічну доброчесність.	15.12-22.12.2025	виконано
12.	Захист кваліфікаційної роботи перед екзаменаційною комісією	26.12.2025	

Дата видачі завдання 29.09.2025 р.

Керівник кваліфікаційної роботи:
к.ф.-м.н., доцент

Євген ТАТАРИНОВ

Завдання прийняв до виконання:

Гліб СЕМЕНЮК

РЕФЕРАТ

Пояснювальна записка до кваліфікаційної роботи «ШІ-асистент для підготовки відповідей на запити на публічну інформацію до Міністерства соціальної політики, сім'ї та єдності України»: 129 сторінок, 2 рисунки, 24 таблиці, 44 використаних джерела, 12 додатків.

Об'єкт дослідження – процес обробки та підготовки відповідей на запити щодо доступу до публічної інформації в органах виконавчої влади.

Мета кваліфікаційної роботи – розробка та дослідження засобів автоматизованої генерації офіційних відповідей на запити громадян з використанням адаптованих LLM на основі підходу «людина в контурі» (Human-in-the-Loop), що забезпечує зменшення адміністративного навантаження на державних службовців.

Методи дослідження – методи системного аналізу предметної області та нормативної бази, багатокритеріальної оптимізації для відбору моделей, методи обробки природної мови (NLP) для підготовки даних, метод квантованої низькорангової адаптації (QLoRA) для доналаштування нейромереж, методи порівняльного та експертного аналізу якості генерації.

Результати роботи можуть бути використані при розробці GovTech та LegalTech рішень для автоматизації ділового листування в державному секторі; для цифрової трансформації процесів комунікації в Міністерстві соціальної політики, сім'ї та єдності України; як основа для створення безпечних локальних ШІ-асистентів для державного сектору, що працюють з чутливими даними без використання зовнішніх хмарних сервісів.

Розробка та дослідження проводилися у хмарному середовищі Google Colab з використанням графічного прискорювача NVIDIA Tesla T4. Розробка програмного засобу проводилася мовою програмування Python із використанням бібліотек Unsloth та TRL

ВЕЛИКА МОВНА МОДЕЛЬ (LLM), HUMAN-IN-THE-LOOP, ДОНАЛАШТУВАННЯ (FINE-TUNING), QLORA, АВТОМАТИЗАЦІЯ

ДЕРЖАВНИХ ПОСЛУГ, ОБРОБКА ПРИРОДНОЇ МОВИ, ШІ-АСИСТЕНТ,
PYTHON.

ABSTRACT

Explanatory note to the qualification work «AI Assistant for Preparing Responses to Public Information Requests Addressed to the Ministry of Social Policy, Family and Unity of Ukraine»: 129 pages, 2 figures, 24 tables, 44 references, 12 appendices.

The object of research is the process of processing and preparing responses to requests for access to public information within executive authorities.

The purpose of the qualification work is the development and research of tools for the automated generation of official responses to citizen requests using adapted LLMs based on the «Human-in-the-Loop» approach, ensuring a reduction in the administrative burden on civil servants.

Research methods – methods of system analysis of the domain and regulatory framework, multi-criteria optimization for model selection, Natural Language Processing (NLP) methods for data preparation, the Quantized Low-Rank Adaptation (QLoRA) method for fine-tuning neural networks, and methods of comparative and expert analysis of generation quality.

The results of the work can be used in the development of GovTech and LegalTech solutions for automating official correspondence in the public sector; for the digital transformation of communication processes at the Ministry of Social Policy, Family and Unity of Ukraine; and as a basis for creating secure local AI assistants for the public sector that process sensitive data without relying on external cloud services.

The development and research were conducted in the Google Colab cloud environment using an NVIDIA Tesla T4 graphics accelerator. The software tool was developed using the Python programming language with the Unsloth and TRL libraries.

LARGE LANGUAGE MODEL (LLM), HUMAN-IN-THE-LOOP, FINE-TUNING, QLoRA, PUBLIC SERVICE AUTOMATION, NATURAL LANGUAGE PROCESSING, AI ASSISTANT, PYTHON.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ, ТЕРМІНІВ.....	9
ВСТУП.....	10
РОЗДІЛ 1	
АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОГЛЯД ІСНУЮЧИХ РІШЕНЬ.....	14
1.1. Нормативно-правові аспекти та специфіка обробки запитів на публічну інформацію.....	14
1.2. Аналіз адміністративного навантаження та обґрунтування моделі Human-in-the-Loop.....	15
1.3. Огляд екосистеми сучасних LLM з відкритим кодом та їх характеристика....	17
1.4. Стратегії доналаштування моделей.....	21
1.5. Аналіз архітектурних рішень та стратегій впровадження ШІ в закордонних GovTech та LegalTech системах.....	22
Висновки.....	29
РОЗДІЛ 2	
ПОСТАНОВКА ЗАДАЧІ ТА ОБґРУНТУВАННЯ МЕТОДІВ І ЗАСОБІВ ЇЇ ВИРІШЕННЯ.....	31
2.1. Характеристика експериментального середовища та технічні обмеження дослідження.....	31
2.2. Методика відбору базових моделей-кандидатів.....	33
2.3. Формування та первинна обробка текстового корпусу для доналаштування LLM.....	37
2.4. Структурування навчальних даних, розробка системної інструкції та технічна валідація датасету.....	38
2.5. Програмна реалізація та обґрунтування гіперпараметрів доналаштування....	45
2.6. Методика експериментального дослідження та критерії оцінювання якості генерації.....	48

Висновки.....	54
РОЗДІЛ 3	
ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ.....	57
3.1. Реалізація процесу доналаштування та аналіз динаміки адаптації моделей... 57	57
3.2. Якісний порівняльний аналіз генерації на тестових сценаріях.....	60
3.3. Кількісна оцінка ефективності моделей за критеріями якості.....	68
3.4. Дослідження ресурсомісткості та швидкодії під час інференсу.....	71
3.5. Узагальнення результатів та визначення оптимальної моделі.....	73
Висновки.....	75
ВИСНОВКИ.....	77
СПИСОК БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ ВИКОРИСТАНИХ ДЖЕРЕЛ.....	80
ДОДАТОК А.....	85
ДОДАТОК Б.....	92
ДОДАТОК В.....	96
ДОДАТОК Д.....	100
ДОДАТОК Е.....	104
ДОДАТОК Ж.....	108
ДОДАТОК З.....	113
ДОДАТОК Й.....	117
ДОДАТОК К.....	118
ДОДАТОК Л.....	119
ДОДАТОК М.....	120
ДОДАТОК Н.....	128

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ, ТЕРМІНІВ

ШІ – штучний інтелект

LLM – велика мовна модель (англ. Large language model)

НІТЛ – людина в контурі (англ. Human-in-the-Loop)

QLoRA – квантована низькорангова адаптація (англ. Quantized Low-Rank Adaptation)

RAG – генерація з доповненням через пошук (англ. Retrieval-Augmented Generation)

РІІ - персональні дані (англ. Personally Identifiable Information)

VRAM – відеопам'ять (англ. Video Random Access Memory)

ВСТУП

Сучасний етап розвитку державного управління в Україні характеризується стрімкою цифровою трансформацією та докорінною зміною парадигми взаємодії між владою та громадянським суспільством. В умовах воєнного стану та курсу на європейську інтеграцію критичного значення набуває прозорість діяльності державних інституцій, ключовим інструментом забезпечення якої є механізм доступу до публічної інформації. Конституційне право громадян на отримання відомостей про діяльність органів влади реалізується через систему запитів, обсяг яких демонструє стійку тенденцію до зростання. Водночас, в умовах обмежених людських ресурсів, традиційні методи обробки кореспонденції стають «вузьким місцем», що призводить до зростання адміністративного навантаження на державних службовців.

Проблема поглиблюється жорсткими нормативними вимогами Закону України «Про доступ до публічної інформації», який встановлює п'ятиденний термін на надання відповіді, недотримання якого тягне за собою юридичну відповідальність. В результаті кваліфіковані фахівці міністерств змушені витратити значну частину робочого часу на рутинні операції з пошуку даних та форматування документів, замість виконання аналітичних функцій. У цьому контексті застосування технологій штучного інтелекту (ШІ), зокрема великих мовних моделей (англ. Large language model, LLM), розглядається як перспективний шлях оптимізації бюрократичних процесів. Однак існуючі комерційні рішення не можуть бути повноцінно інтегровані в державний сектор через ризики безпеки даних, що актуалізує потребу у розробці спеціалізованих локальних рішень. Актуальність теми зумовлена необхідністю розв'язання критичної проблеми цифровізації держсектору: потреби в автоматизації рутинних процесів в умовах обмежень на використання незахищених хмарних ШІ-сервісів. Дослідження пропонує архітектурне рішення, яке дозволяє використовувати потужність сучасних LLM локально, забезпечуючи повний контроль над даними громадян та обробку запитів без порушення протоколів безпеки, що є неможливим при використанні комерційних аналогів.

Метою роботи є розробка та дослідження засобів автоматизованої генерації офіційних відповідей на запити громадян з використанням адаптованих LLM на основі підходу «людина в контурі» (англ. Human-in-the-Loop, HITL) для зменшення адміністративного навантаження на державних службовців.

Для досягнення поставленої мети вирішено комплекс взаємопов'язаних завдань. Насамперед проведено ґрунтовний аналіз предметної області та нормативно-правової бази обробки публічної інформації для визначення вимог до автоматизованої системи. Наступним кроком здійснено огляд сучасних архітектур нейронних мереж, що дозволило обґрунтувати вибір базових моделей-кандидатів (Gemma 2, Llama 3, Qwen 2.5), придатних для розгортання в умовах обмежених ресурсів. Для забезпечення релевантності доналаштування сформовано спеціалізований текстовий корпус на основі відкритих документів Міністерства соціальної політики, сім'ї та єдності України та розроблено методіку його попередньої обробки. Після цього розроблено програмні засоби для доналаштування (англ. fine-tuning) моделей із застосуванням методу квантованої низькорангової адаптації (англ. Quantized Low-Rank Adaptation, QLoRA) та методик конструювання запитів. Завершальним етапом стало експериментальне дослідження ефективності адаптованих моделей з проведенням порівняльного аналізу якості генерації та ресурсомісткості для визначення оптимальної конфігурації системи.

Об'єкт дослідження – процес обробки та підготовки відповідей на запити щодо доступу до публічної інформації в органах виконавчої влади.

Предмет дослідження – методи та програмні засоби адаптації LLM для генерації текстів офіційно-ділового стилю українською мовою в умовах апаратних обмежень.

У роботі використано комплексний підхід, що базується на поєднанні теоретичних та емпіричних методів. Для аналізу предметної області та існуючих рішень застосовано методи системного аналізу. Вибір базових моделей здійснено із застосуванням методу багатокритеріальної оптимізації на основі метрик рейтингу Open LLM Leaderboard. Для підготовки даних використано методи обробки природної мови, зокрема токенізацію та нормалізацію тексту. Процес

доналаштування моделей реалізовано із застосуванням методу QLoRA. Оцінювання якості результатів здійснено шляхом експертного аналізу відповідно до розробленої системи критеріїв.

Отримані результати характеризуються новизною підходів до автоматизації бюрократичних процесів. Зокрема, удосконалено метод адаптації LLM до специфіки українського офіційно-ділового стилю шляхом комбінування техніки QLoRA з навчанням на цільових відповідях (англ. *train on completions only*) та динамічною інтеграцією структурованого контексту, що дозволило досягти високої якості генерації на малому наборі даних. Набула подальшого розвитку архітектура програмної системи на основі підходу HITL для державних сервісів, яка, на відміну від повністю автоматизованих систем, інтегрує ШІ як генератор чернеток із суворими обмеженнями на використання параметричних знань, що мінімізує юридичні ризики. Крім того, проведено порівняльний аналіз моделей Gemma 2, Llama 3 та Qwen 2.5 на задачах української юридичної практики в умовах обмежених ресурсів. Це дозволило встановити, що архітектурні особливості механізму уваги безпосередньо впливають на здатність моделі стабільно генерувати довгі кириличні тексти без помилок зациклення.

Практичне значення роботи полягає у створенні та експериментальній верифікації функціонального прототипу ШІ-асистента на базі моделі Gemma 2 9B, який показав придатність до автоматизації підготовки офіційних відповідей в умовах обмежених апаратних ресурсів. У ході дослідження сформовано унікальний датасет, а результати тестування підтвердили перспективність розробленого прототипу як технологічної основи для цифрової трансформації процесів комунікації в державному секторі, що поєднує інноваційність, безпеку та економічну доцільність.

Перспективи подальшого розвитку проєкту безпосередньо пов'язані з поглибленням співпраці з Міністерством соціальної політики, сім'ї та єдності України для отримання доступу до розширеного корпусу даних, що суттєво покращить здатність моделі до узагальнення. Ключовим напрямком модернізації системи визначено розробку та впровадження архітектури генерації з доповненим пошуком (англ. *Retrieval-Augmented Generation, RAG*), інтегрованої з джерелами

законодавства в реальному часі, що дозволить мінімізувати проблему фактологічних галюцинацій моделі. Також рекомендовано подальшу оптимізацію системних інструкцій та розробку модуля автоматичної постобробки тексту для усунення артефактів форматування.

РОЗДІЛ 1

АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ОГЛЯД ІСНУЮЧИХ РІШЕНЬ

1.1. Нормативно-правові аспекти та специфіка обробки запитів на публічну інформацію

Першочерговим етапом проектування будь-якої інформаційної системи в державному секторі є детальний аналіз нормативного поля, яке регламентує предметну область. Це завдання набуває особливої ваги в сучасних умовах — етап розвитку державного управління в Україні характеризується стрімкою цифровізацією та підвищенням вимог суспільства до прозорості діяльності органів влади. Ключовим інструментом забезпечення такої прозорості є механізм доступу до публічної інформації, який, з одного боку, гарантує конституційне право громадян на отримання відомостей, а з іншого — покладає значний адміністративний тягар на розпорядників інформації. В умовах обмеженого кадрового ресурсу та експоненційного зростання обсягів даних, традиційні методи обробки запитів стають критичним обмеженням продуктивності, що знижує ефективність функціонування державних інституцій.

Правовий фундамент взаємодії між державою та суспільством у цій сфері закладено Законом України «Про доступ до публічної інформації». Цей нормативний акт встановлює жорсткі часові рамки та змістовні вимоги до відповідей, недотримання яких тягне за собою юридичну відповідальність. Згідно зі статтею 20 Закону, розпорядник інформації зобов'язаний надати відповідь не пізніше п'яти робочих днів з дня отримання запиту [1]. Цей строк є одним із найжорсткіших у європейській практиці, що створює постійний тиск на виконавців. Хоча закон передбачає можливість продовження строку розгляду до 20 робочих днів у випадках, коли запит стосується великого обсягу інформації або потребує пошуку серед значної кількості даних, таке рішення вимагає окремого обґрунтування та повідомлення запитувача, що саме по собі генерує додаткову адміністративну роботу [1].

Особливої уваги заслуговує стаття 22 Закону, яка визначає вичерпний перелік підстав для відмови в задоволенні запиту. До них належать випадки, коли розпорядник не володіє і не зобов'язаний володіти інформацією, коли інформація належить до категорії з обмеженим доступом, або коли запитувач не виконав вимоги щодо оплати витрат на копіювання [1]. Коректна кваліфікація підстави для відмови вимагає від державного службовця глибокого розуміння не лише профільного законодавства, а й суміжних нормативних актів, що регулюють захист персональних даних та державну таємницю. Помилка на цьому етапі може призвести до оскарження дій розпорядника відповідно до статті 23 Закону, яка гарантує запитувачу право звернутися до суду у випадку надання недостовірної або неповної інформації, а також несвоєчасного надання відповіді [1]. Таким чином, процес підготовки відповіді перетворюється на складну юридичну процедуру, де кожен документ повинен пройти багаторівневу верифікацію.

1.2. Аналіз адміністративного навантаження та обґрунтування моделі Human-in-the-Loop

Після визначення нормативно-правових вимог, що формують рамки діяльності розпорядників інформації, логічним кроком є аналіз того, як ці норми проявляються в повсякденній роботі державних службовців. Саме нормативна база визначає реальне адміністративне навантаження, з яким стикається виконавець.

Зокрема, проблема часових витрат на обробку запитів виходить за межі простого тайм-менеджменту і є системною проблемою державного управління, яку доцільно розглядати через призму теорії адміністративного тягаря. Ця теорія дозволяє декомпонувати загальні витрати взаємодії громадянина з державою та державного службовця з системою на складові компоненти, виявляючи приховані резерви для оптимізації.

Згідно з сучасними дослідженнями, адміністративний тягар не є монолітним явищем, а складається з трьох ключових компонентів: витрат на навчання, витрат на відповідність та психологічних витрат [2].

Витрати на навчання у контексті роботи державного службовця пов'язані з необхідністю постійного пошуку та актуалізації знань. Перед підготовкою відповіді виконавець повинен з'ясувати, чи є запитувана інформація публічною, які структурні підрозділи нею володіють, та чи не змінилося законодавство з моменту останнього подібного запиту. В умовах динамічних змін нормативної бази, особливо у період воєнного стану, ці витрати стають значними. Дослідження показують, що навіть досвідчені фахівці витрачають значну частину часу на навігацію у внутрішніх інструкціях та реєстрах [3].

Витрати на відповідність охоплюють безпосередньо процесуальні дії: написання тексту, форматування документа згідно з вимогами діловодства, візування у керівництва, реєстрацію в системі електронного документообігу [4]. Саме на цьому етапі виникає найбільший потенціал для автоматизації за допомогою генеративного ШІ, оскільки значна частина цих дій є рутинною та алгоритмізованою. Статистика свідчить, що держслужбовці та адміністратори можуть витрачати до 44% робочого часу саме на виконання таких бюрократичних процедур, які не створюють безпосередньої суспільної цінності, але є обов'язковими [4].

Психологічні витрати є найменш очевидним, але критично важливим фактором. Вони виникають внаслідок стресу через жорсткі терміни, страх допустити помилку, яка призведе до судового позову, та емоційного навантаження від спілкування з громадянами, які часто перебувають у складних життєвих обставинах [5]. Впровадження ШІ-асистента дозволяє мінімізувати цей тиск: система автоматично генерує емоційно нейтральну, нормативно обґрунтовану чернетку відповіді. Це позбавляє службовця необхідності самостійного формування тексту відповіді, трансформуючи задачу з високим когнітивним навантаженням у менш ресурсомістку процедуру верифікації та редагування згенерованого проєкту.

Також інтеграція алгоритмічних систем у процеси прийняття рішень пов'язана зі специфічними когнітивними ризиками, найбільш критичним з яких є упередженість автоматизації (англ. Automation Bias). Цей феномен полягає у схильності операторів надмірно довіряти результатам роботи автоматизованих систем, навіть за наявності суперечливої інформації з альтернативних джерел [6]. У

контексті державного управління це може призвести до ситуації, коли службовець затверджує згенерований ШІ документ, не здійснюючи його критичної перевірки [7].

Дослідження демонструють, що особи, які приймають рішення, схильні надавати перевагу алгоритмічним порадам порівняно з рекомендаціями людей, особливо в умовах дефіциту часу та високої відповідальності [6]. Це створює загрозу «розмивання відповідальності», коли помилкове рішення, запропоноване ШІ (наприклад, неправомірна відмова у наданні інформації), легітимізується підписом посадової особи. Крім того, без належних запобіжників ШІ може відтворювати або посилювати існуючі упередження, що загрожує дискримінацією певних категорій запитувачів [8].

Для мінімізації зазначених ризиків архітектуру системи реалізовано за принципом HITL. Система позбавлена права приймати остаточні рішення; її функціонал обмежено генерацією проєктів відповідей, пошуком релевантних прецедентів та автоматизованою перевіркою тексту на відповідність формальним вимогам. Людина залишається ключовим елементом системи, здійснюючи верифікацію фактів та оцінку контексту. Ефективним методом боротьби з упередженістю автоматизації є дизайн інтерфейсу, який змушує користувача активно взаємодіяти з інформацією, наприклад, вимагаючи підтвердження коректності знайдених посилань на законодавство перед генерацією фінального тексту [6]. Такий підхід перетворює ШІ з чорної скриньки на прозорий інструмент підтримки прийняття рішень.

1.3. Огляд екосистеми сучасних LLM з відкритим кодом та їх характеристика

Аналіз природи адміністративного навантаження та важливості людського контролю формує підґрунтя для наступного етапу — вибору технологічного стеку. Для ефективної роботи асистента в контурі HITL критично важливим є обрання оптимальної базової LLM, здатної забезпечити високу точність обробки юридичних текстів українською мовою.

Вибір базової LLM є критичним етапом проектування системи, особливо в умовах обмежених обчислювальних ресурсів. Сучасний спектр моделей з відкритими вагами (англ. Open Weights) представлений низкою потужних рішень, кожне з яких має власні переваги та обмеження в контексті обробки українських юридичних текстів.

Першим доцільно розглянути сімейство моделей Llama 3 від Meta, яке встановило новий стандарт продуктивності для відкритих моделей. Ключовою архітектурною особливістю є використання стандартного трансформера архітектури «лише декодер» (англ. decoder-only) з інтеграцією механізму уваги з групуванням запитів (англ. Grouped Query Attention). Цей механізм оптимізує зберігання ключів та значень, суттєво зменшуючи обсяг KV-кешу під час генерації [9]. Це архітектурне рішення стало фундаментом для підтримки надвеликих контекстів (до 128 тис. токенів) у нових версіях та забезпечило можливість запуску моделей на графічних прискорювачах з обмеженою відеопам'яттю, таких як NVIDIA T4.

Для задач обробки української мови Llama 3.1 демонструє виняткову ефективність. Дослідження токенізації показують, що Llama 3.1 має показник «fertility» (відношення кількості токенів до слів) на рівні 1,88 для кириличних текстів, що є найкращим результатом серед відкритих моделей, перевершуючи навіть Gemma 2 та Mistral [10]. Це означає, що Llama 3.1 може вмістити більше українського тексту в обмежене контекстне вікно, ніж конкуренти, і витрачає менше обчислювальних ресурсів на генерацію відповіді. Незважаючи на менший словник кириличних символів порівняно з Gemma, ефективний алгоритм кодування пар байтів (англ. Byte-pair encoding) дозволяє їй уникати фрагментації слів, що позитивно впливає на семантичну цілісність тексту [10].

У той час як Meta задає індустріальні стандарти, моделі Gemma 2 від Google представляють альтернативний підхід до дизайну LLM. Замість простого масштабування параметрів, розробники використали метод дистиляції знань від значно більшої моделі-вчителя, що дозволило моделі 9B досягти результатів, порівнянних з моделями класу Llama 3 70B у певних бенчмарках [11].

Архітектурно Gemma 2 відрізняється використанням активаційної функції GeGLU (Gated Gaussian Linear Units) замість поширеної SwiGLU (Swish-Gated Linear Unit), а також чергуванням шарів локальної та глобальної уваги. Це дозволяє моделі ефективніше фокусуватися на локальному контексті при обробці довгих документів [11]. Для задач міністерства, які часто вимагають складних логічних висновків на основі наданих документів, високі показники здатності до міркувань у Gemma 2 є значною перевагою.

Окрему нішу серед рішень, орієнтованих на максимальну ефективність розгортання на локальному обладнанні, займають розробки французької компанії Mistral AI. Компанія послідовно розвиває напрямок високоефективних моделей для локального запуску. Їхня флагманська технологія — Sliding Window Attention (SWA), реалізована у моделях Mistral 7B та новій серії Ministral (3B, 8B) [12]. Суть технології полягає в тому, що на кожному окремому шарі нейромережі увага фокусується лише на фіксованому вікні останніх токенів. Однак завдяки багат шаровій структурі трансформера інформація передається на вищі рівні, що дозволяє моделі на глибинних шарах враховувати контекст, який знаходиться далеко за межами початкового обмеженого вікна. Це знижує обчислювальну складність з квадратичної $O(n^2)$ до лінійної $O(n)$, що дозволяє обробляти значно довші тексти при меншому споживанні пам'яті [13].

Своєю чергою, новітня модель Ministral 8B, спроектована для периферійних обчислень (англ. Edge AI), використовує вдосконалений механізм чергування ковзного вікна уваги (англ. Interleaved Sliding-Window Attention). Це архітектурне рішення забезпечує високу ефективність генерації на пристроях з обмеженими апаратними ресурсами [14].

Якщо європейські та американські моделі фокусуються на архітектурній ефективності, то серія Qwen 2.5 від Alibaba Cloud вирізняється безпрецедентним масштабом тренувальних даних — понад 18 трильйонів токенів, що включають тексти 29 мовами [15, 16]. Для українського контексту важливою є гарні показники розуміння кирилиці та високі показники у бенчмарках OpenCompass, де модель 72B

Instruct навіть перевершує закриті моделі типу GPT-4o у завданнях з математики та кодингу [16].

Специфічною перевагою моделі Qwen 2.5 є її чітко виражена інженерна спрямованість, яка вигідно вирізняє її серед конкурентів орієнтацією на прикладні задачі програмування та обробки даних. Модель демонструє виняткові результати у генерації суворо структурованих даних у форматах JSON та XML, а також здатність точно слідувати складним багаторівневим системним інструкціям [17]. Ця характеристика є критично важливою для успішної інтеграції ШІ-асистента у формалізовані бюрократичні процеси, де відповідь часто повинна не лише містити текст, а й мати чітко визначену структуру. Окрім того, архітектура моделі підтримує розширене контекстне вікно обсягом до 128 тисяч токенів. Це технічне рішення дозволяє завантажувати у пам'ять моделі повні тексти об'ємних нормативно-правових актів, кодексів та інструкцій без необхідності їх фрагментації, що забезпечує цілісність аналізу та дозволяє системі знаходити взаємозв'язки між віддаленими частинами документів.

Якісно новим етапом еволюції LLM, критично важливим для складного юридичного аналізу, стала поява моделі DeepSeek-R1. Ця подія ознаменувала зміну парадигми: зміщення фокусу з традиційного контрольованого доналаштування (англ. Supervised Fine-Tuning) на масштабне навчання з підкріпленням (англ. Reinforcement Learning), орієнтоване на правильність кінцевого результату. Такий підхід зумовив виникнення емерджентної поведінки: модель самостійно сформувала здатність будувати ланцюжки міркувань (англ. Chain-of-Thought, CoT), що включають етапи самоперевірки, рефлексії та корекції помилок перед генерацією фінальної відповіді [18].

У контексті юридичної роботи це відкриває нові горизонти. Дослідження показують, що моделі з розвиненим CoT значно краще справляються із завданнями, що вимагають застосування формальної формальної логіки та силогізмів (наприклад, метод IRAC: Issue – питання, Rule – норма, Application – застосування, Conclusion – висновок) [19]. DeepSeek-R1 здатна виявляти логічні суперечності у запиті громадянина або в проекті відповіді, що значно підвищує якість юридичного

аналізу. Однак, існують і обмеження: модель може страждати від мовного змішування та схильності до повторень, а її інференс вимагає значних ресурсів через генерацію довгого прихованого ланцюжка міркувань.

1.4. Стратегії доналаштування моделей

Проаналізувавши особливості сучасних відкритих LLM та їх потенціал для роботи в державному секторі, переходом до практичної частини стає питання адаптації цих моделей під конкретні потреби міністерства. Навіть найкраща модель потребує доналаштування, щоб відповідати стилю офіційного листування та формальним вимогам документообігу.

З технічної точки зору, для реалізації цього процесу в умовах обмежених ресурсів оптимальним вибором є метод QLoRA, який став проривним рішенням для доналаштування великих моделей на споживчому обладнанні. Його суть полягає у заморожуванні ваг базової моделі та їх квантуванні до 4-бітного формату NormalFloat, при цьому навчаються лише невеликі додаткові матриці адаптерів (англ. Low-Rank Adapters) [20]. Це дозволяє зменшити споживання пам'яті для моделі 7B до приблизно 5-6 GB, залишаючи решту 10 GB відеопам'яті T4 для завантаження батчів даних та обчислення градієнтів.

Саме у контексті підготовки відповідей на запити QLoRA є ідеальним інструментом для перенесення стилю. Модель можна навчити писати у специфічному бюрократичному тоні, використовувати правильні канцеляризми, дотримуватися структури офіційного листа.

Водночас, налаштування стилю не вирішує питання актуальності фактологічних даних. Тут на перший план виходить технологія RAG, яка вирішує фундаментальну проблему LLM — галюцинації та застарілість знань. Оскільки доналаштування фіксує знання на момент його проведення, будь-які зміни в законодавстві вимагатимуть повторного дороговартісного доналаштування. RAG дозволяє динамічно інтегрувати актуальну інформацію із зовнішньої бази знань під час генерації відповіді [21].

Практична реалізація архітектури RAG для міністерства передбачає створення векторної бази даних (наприклад, на основі ChromaDB або FAISS), яка містить індексовані тексти нормативно-правових актів та шаблони відповідей. При отриманні запиту система знаходить релевантні статті законів і додає їх у контекст запиту. Це не лише забезпечує точність, але й дозволяє системі надавати прямі посилання на джерела, що є критичним для юридичної прозорості.

Таким чином, враховуючи сильні та слабкі сторони обох технологій, оптимальним є гібридний підхід: використання 4-бітної квантованої моделі, доналаштованої через QLoRA на корпусі відповідей міністерства для засвоєння бюрократичного стилю, у поєднанні з механізмом RAG для інтеграції актуальних норм законодавства.

1.5. Аналіз архітектурних рішень та стратегій впровадження ШІ в закордонних GovTech та LegalTech системах

Після окреслення загальних тенденцій та підходів до інтеграції ШІ у державному секторі, доцільно перейти до детального аналізу конкретних архітектурних рішень, моделей безпеки та стратегій впровадження у провідних міжнародних системах сфери GovTech та LegalTech. Це дозволить не лише порівняти різні технологічні парадигми, але й визначити найефективніші практики, що можуть бути релевантними для українського контексту.

Першим доцільно розглянути «золотий стандарт» використання генеративного ШІ у приватному секторі, який задає технологічну планку для всієї індустрії. Harvey AI позиціонується як провідна платформа генеративного ШІ, розроблена спеціально для задоволення потреб професіоналів у галузі юридичних, податкових та фінансових послуг [22]. На відміну від універсальних моделей, Harvey створений для вирішення складних завдань, що вимагають глибоких предметних знань та комплексного аналізу. Основна мета платформи — підвищення продуктивності юристів шляхом автоматизації трудомістких завдань, таких як аналіз документів,

правовий аналіз та підготовка проєктів документів, що дозволяє фахівцям зосередитися на стратегічному консультуванні клієнтів [22].

Високий рівень довіри до платформи у традиційно консервативній юридичній галузі підтверджується її швидким та масштабним впровадженням. Провідні світові юридичні фірми, зокрема A&O Shearman, розгорнули Harvey для підтримки тисяч співробітників у десятках юрисдикцій [23]. Більше того, стратегічний альянс із PricewaterhouseCoopers (PwC) не обмежується лише використанням інструменту; PwC, Harvey та OpenAI спільно працюють над створенням та навчанням нових фундаментальних моделей, спеціалізованих для вирішення податкових, юридичних та кадрових завдань [24]. Це свідчить про визнання Harvey як ключового технологічного партнера для трансформації галузі професійних послуг.

В основі технологічної переваги Harvey лежить синергетичний гібридний підхід, що поєднує доналаштування моделей та RAG [25]. Засновники платформи швидко усвідомили, що базові LLM, навіть потужні, як GPT-4, не володіють необхідною глибиною та нюансами юридичних знань [26]. Просте використання RAG для надання контексту виявилось недостатнім для складних юридичних міркувань. Тому Harvey, у тісному партнерстві з OpenAI, обрав стратегію створення спеціалізованої моделі для роботи з прецедентним правом [23]. Цей процес передбачав доналаштування базової моделі на величезному корпусі юридичних даних, що охоплює все прецедентне право США, еквівалентне приблизно 10 мільярдам токенів [26].

Результати цього підходу виявилися проривними. У ході тестування з юристами з провідних фірм, відповіді, згенеровані спеціалізованою доналаштованою моделлю Harvey, отримали перевагу у 97% випадків порівняно з відповідями базової моделі GPT-4 [26]. Юристи відзначили, що відповіді кастомної моделі були довшими, повнішими та краще враховували нюанси запиту, охоплюючи ширший спектр релевантного прецедентного права [25]. Це демонструє, що для вузькоспеціалізованих предметних областей доналаштування фундаментально змінює поведінку моделі, навчаючи її «мислити» в категоріях, притаманних конкретній професії [27].

Над цією потужною, доналаштованою базовою моделлю Harvey розгортає шар RAG. У даній архітектурі він виконує іншу, але не менш критичну функцію: забезпечує точність у реальному часі та обґрунтовує відповіді моделі на конкретних, часто внутрішніх, документах клієнта [28]. Для забезпечення найвищої якості пошуку у цьому шарі Harvey розробляє власні спеціалізовані бенчмарки, такі як BigLaw-Bench, для оцінки та вдосконалення ефективності RAG-систем у юридичному контексті [29]. Ця гібридна архітектура, що використовує інфраструктуру Azure AI та найновіші моделі OpenAI, включаючи GPT-4 Turbo та передові моделі серії o1 (o1-preview, o1-mini), дозволяє Harvey одночасно досягати глибини предметних знань та точності у роботі з конкретними фактами справи [30]

Функціонально Harvey надає юристам інтегрований набір інструментів, що виходить за межі простого чат-інтерфейсу. До ключових компонентів належать «Assistant» (укр. Асистент) для аналізу та перевірки великих обсягів документів, таких як контракти або матеріали справи, та «Vault» (укр. Сховище) — безпечне середовище для завантаження, аналізу та перевірки тисяч документів клієнта, що слугує основою для функціоналу RAG [28]. Платформа також підтримує налаштовувані робочі процеси (англ. workflows), дозволяючи командам автоматизувати багатоетапні, повторювані юридичні завдання.

Найновішою еволюцією платформи є розгортання агентських робочих процесів [31]. Це знаменує перехід від парадигми «запит-відповідь» до проактивної співпраці. Агент Harvey визначається трьома ключовими здібностями: планування — здатність розбивати складну юридичну задачу на послідовні, керовані кроки; адаптивність — можливість використовувати результати попереднього кроку для коригування наступних дій; та взаємоді'я — здатність активно запитувати у юриста уточнення, відсутній контекст або необхідне рішення [31]. Ця модель взаємодії, що значною мірою покладається на можливості багатоетапного планування та рефлексії нових моделей, як-от o1 від OpenAI, перетворює ШІ з пасивного інструменту на керованого співробітника, що активно веде користувача до вирішення задачі.

Однак ані потужна архітектура, ані агентські функції не мали б цінності для юридичної галузі без вирішення фундаментальної проблеми довіри та безпеки.

Основою пропозиції Harvey є так званий «стек довіри» (англ. Trust Stack) [23]. Його наріжним каменем є невід'ємний принцип: «нульове навчання» (англ. zero training) на даних клієнта [23]. Harvey надає юридично зобов'язуючу контрактну гарантію через свій «Додаток про безпеку» (англ. Security Addendum), що жодні дані, введені користувачем, завантажені документи або згенеровані результати ніколи не використовуються для навчання чи доналаштування базових моделей [32]. Цей принцип усуває головний бар'єр для впровадження ШІ у фірмах, що оперують особливо чутливою інформацією.

Ця гарантія підкріплена надійними технічними та організаційними заходами, зокрема сертифікаціями SOC 2 Type II та ISO 27001, а також відповідністю вимогам GDPR [23]. Платформа надає клієнтам повний контроль над даними, включаючи суверенітет даних (можливість вибору регіону зберігання: ЄС, США або Австралія) та гнучкі політики зберігання [32]. На більш глибокому архітектурному рівні Harvey розробляє рішення для управління складними інформаційними бар'єрами та правилами адвокатської таємниці, що є унікальною вимогою юридичної практики [33]. Це дозволяє створювати безпечні робочі простори, де команди юристів, ШІ-агентів та, потенційно, клієнтів можуть співпрацювати, маючи деталізований контроль над доступом до інформації [33]. Таким чином, успіх Harvey є прямим наслідком не лише його гібридної архітектури, але й моделі безпеки, яка функціонує як необхідна передумова для використання потужних можливостей платформи для обробки найбільш цінних даних юридичної фірми.

Якщо Harvey демонструє підхід до глибокої юридичної аналітики у приватному секторі, то у сфері державного управління пріоритети зміщуються у бік масовості та швидкості обробки. Платформа SmartCompose, розроблена Агентством урядових технологій Сінгапуру (GovTech Singapore), демонструє інший, але не менш предметно-орієнтований підхід до впровадження ШІ [34]. SmartCompose — це ШІ-асистент, призначений для конкретної аудиторії: державних службовців, що безпосередньо взаємодіють з громадянами (англ. citizen-facing public officers) [34]. На відміну від Harvey, орієнтованого на глибокий аналіз, головна мета

SmartCompose — підвищення продуктивності та якості комунікації у великих масштабах [34].

Ключова задача платформи — допомагати службовцям створювати «швидші, більш продумані та професійні» відповіді на запити громадян [34]. Особливий акцент робиться на здатності генерувати «емпатичні» відповіді, що є критично важливим для взаємодії між державою та громадянином [34].

Функціональна архітектура SmartCompose базується на використанні LLM для генерації чернеток контенту [35]. Платформа використовує можливості, що нагадують RAG, зокрема «навчання на кількох прикладах» (англ. few-shot learning), що дозволяє їй генерувати текст, спираючись на внутрішні документи та довірені джерела, без необхідності повного доналаштування моделі для кожного окремого випадку [34]. Серед ключових функцій — здатність до адаптації тону повідомлення, що дозволяє автоматично налаштовувати стиль відповіді (наприклад, більш співчутливий або формальний) залежно від суті запиту громадянина [34].

На відміну від багатьох урядових IT-проектів, успіх SmartCompose вимірюється чіткими та переконливими показниками ефективності. З моменту запуску платформа продемонструвала суттєве скорочення середнього часу, необхідного державному службовцю для підготовки відповіді на електронний лист: з 20 хвилин до всього 5,4 хвилини [35]. Це являє собою 70% підвищення ефективності [34]. Станом на кінець 2025 року, інструмент був впроваджений у понад 90 державних установах Сінгапуру та був використаний для генерації понад 12 850 відповідей громадянам [35].

Критичним аспектом для будь-якого інструменту сфери GovTech, що працює зі зверненнями громадян, є безпечна обробка персональних даних (англ. Personally Identifiable Information, ПІ). SmartCompose був розроблений з урахуванням цієї вимоги і містить спеціалізовані вбудовані функції для контролю якості та захисту ПІ [35]. Цей механізм включає логічні перевірки та автоматичне виявлення ПІ для забезпечення безпеки та відповідності комунікацій нормативним вимогам [34].

Однак швидке та безпечне впровадження SmartCompose у 90 установах не було б можливим, якби інструмент був ізольованою розробкою. Його успіх значною

мірою зумовлений зрілою цифровою інфраструктурою уряду Сінгапуру та його стратегічним підходом до інновацій. GovTech розробив та підтримує централізований «Технологічний стек уряду Сінгапуру» (англ. Singapore Government Tech Stack), який надає стандартизовані платформи, інструменти та практики для всіх державних установ [36].

Найважливішим компонентом цієї екосистеми для таких інструментів, як SmartCompose, є «Посібник із генерації з доповненим пошуком» (англ. Retrieval-Augmented Generation Playbook) [37]. Цей офіційний документ слугує керівництвом для державних інженерів, фахівців з науки про дані та постачальників. Він встановлює єдині найкращі практики, стандарти безпеки та метрики для розробки, оцінки та вдосконалення RAG-систем у державному секторі [37]. Таким чином, SmartCompose можна розглядати не як окремий продукт, а як результат стандартизованого процесу, регламентованого даним посібником. Саме цей підхід, що базується на централізованих стандартах, дозволяє агентству GovTech Сінгапуру швидко масштабувати інноваційні ШІ-рішення, гарантуючи при цьому дотримання суворих вимог безпеки, точності та підзвітності, особливо під час обробки РІІ.

На відміну від сінгапурського підходу, орієнтованого на зовнішню взаємодію, проєкт Redbox представляє модель впровадження рішень сфери GovTech, сфокусовану на безпеці внутрішніх процесів. Це рішення, зосереджене на внутрішніх потребах уряду Великобританії, було розроблене внутрішньою командою «Incubator for AI» у Департаменті науки, інновацій та технологій (англ. Department for Science, Innovation and Technology) і призначений для використання державними службовцями у різних департаментах [38]. Основна мета Redbox — підвищити ефективність державних службовців, надаючи їм інструмент для швидкого узагальнення, аналізу документів, пошуку інформації та створення чернеток [39].

На противагу гібридній моделі Harvey, архітектура Redbox з самого початку була чітко зосереджена на парадигмі RAG [40]. Стандарт про прозорість алгоритмів (англ. Algorithmic Transparency Recording Standard, ATRS) уряду Великобританії детально описує технічний процес: 1) користувач завантажує документи; 2) система

здійснює фрагментацію тексту та створює векторні представлення (ембединги) за допомогою таких моделей, як Amazon Titan Embed-text-V2; 3) у відповідь на запит користувача, RAG-система витягує найбільш релевантні фрагменти; 4) велика мовна модель (наприклад, Claude 3 Sonnet) генерує відповідь, яка обґрунтована витягнутими даними та супроводжується цитатами [40].

Ключовими особливостями платформи є її відкритий вихідний код, що дозволяє будь-якому урядовому департаменту вільно адаптувати та розвивати власну версію та можливість обирати, яку саме модель вони хочуть використовувати для своїх завдань [38].

Головним рушієм для створення Redbox була не стільки потреба у складних міркуваннях, скільки нагальна вимога безпеки. На момент розробки (починаючи з хакатону в липні 2023 року) на ринку бракувало комерційних інструментів, які б надавали уряду гарантії безпечної обробки документів з грифом «Для службового користування» [38]. Redbox був розроблений як захищене середовище, що дозволяє державним службовцям використовувати потужність LLM зі своїми реальними робочими документами, не порушуючи протоколів безпеки.

Ця орієнтованість на безпеку безпосередньо визначила парадигму взаємодії HITL, яка кардинально відрізняється від підходу активної агентської співпраці, реалізованого в Harvey. Redbox чітко позиціонується як «допоміжний інструмент» [41]. В офіційній документації (стандарт ATRS) підкреслюється, що вся відповідальність за кінцевий результат лежить на людині [40]. Користувач зобов'язаний перевіряти надані цитати, верифікувати інформацію та зберігати «повний контроль» над будь-яким згенерованим текстом перед його використанням [40].

Підсумовуючи аналіз закордонного досвіду, можна стверджувати, що успішна імплементація генеративного ШІ в юридичній та державній сферах неможлива без глибокої адаптації базових технологій до специфіки предметної області. Розглянуті приклади демонструють, що незалежно від обраної архітектури — чи то гібридна модель із глибоким доналаштуванням Harvey AI, екосистемний підхід SmartCompose чи захищений RAG Redbox — спільним знаменником успіху є пріоритет безпеки

даних та обов'язкове збереження принципу HITL. Жодне з рішень не покладається на «сирі» моделі, що підтверджує необхідність створення для українського міністерства спеціалізованого локального інструменту, який поєднував би здатність до стилістичної адаптації з жорсткими механізмами верифікації змісту.

Висновки

У першому розділі здійснено комплексний аналіз передумов автоматизації процесів обробки запитів на публічну інформацію в державному секторі. Визначено, що критичні часові обмеження та високі вимоги до юридичної точності відповідей створюють надмірний адміністративний тягар для працівників міністерства. Обґрунтовано, що для нівелювання ризиків та збереження контролю над прийняттям рішень система повинна будуватися за архітектурою HITL, де штучний інтелект виступає генератором чернеток, а відповідальність за верифікацію залишається за людиною. Проведений огляд сучасних локальних LLM, зокрема сімейств Llama 3, Gemma 2 та інших, підтвердив їхню технічну спроможність до роботи з українською мовою в умовах обмежених обчислювальних ресурсів.

За результатами порівняльного аналізу стратегій адаптації моделей встановлено, що хоча поєднання доналаштування технологією RAG є теоретично оптимальним, повноцінна реалізація механізму пошуку вимагає обробки та індексації масивних обсягів законодавчої бази, що є надзвичайно трудомістким процесом у межах даного дослідження. У зв'язку з цим, подальша робота буде сфокусована на методі доналаштування (зокрема QLoRA). Цей підхід дозволить адаптувати модель для засвоєння специфічного офіційно-ділового стилю та структури документів без необхідності розгортання складної інфраструктури пошуку. Аналіз закордонного досвіду рішень сфери GovTech підтверджує доцільність створення спеціалізованих асистентів, адаптованих під конкретні бюрократичні процедури. Узагальнюючи викладене, можна стверджувати, що тема дослідження є вкрай актуальною, оскільки запропоноване рішення здатне вирішити

гостру проблему ресурсного дефіциту в умовах цифрової трансформації соціальної сфери та зростаючого суспільного запиту на ефективну комунікацію з державою.

РОЗДІЛ 2

ПОСТАНОВКА ЗАДАЧІ ТА ОБҐРУНТУВАННЯ МЕТОДІВ І ЗАСОБІВ ЇЇ ВИРІШЕННЯ

2.1. Характеристика експериментального середовища та технічні обмеження дослідження

Вибір архітектур нейронних мереж, методів їх доналаштування та стратегій розгортання безпосередньо залежить від доступних обчислювальних ресурсів. Для забезпечення відтворюваності результатів дослідження, а також для демонстрації можливості впровадження розробленого рішення в державних установах без необхідності значних капіталовкладень у закупівлю спеціалізованого обладнання, як базове експериментальне середовище було обрано хмарну платформу Google Colab (у рамках тарифного плану «Free Tier»).

Вибір даної платформи зумовлений її доступністю та наявністю попередньо налаштованого оточення для задач машинного навчання, що дозволяє зосередитися на алгоритмічній складовій дослідження. Однак використання безкоштовного тарифного плану накладає чіткі апаратні обмеження, які стали визначальними для формування методології відбору моделей та налаштування процесу адаптації.

Апаратна конфігурація виділеної віртуальної машини включає:

1. Графічний прискорювач (GPU): Середовище надає доступ до відеокарти архітектури NVIDIA Tesla T4. Ключовою характеристикою для роботи з LLM є обсяг відеопам'яті (VRAM), який для цього прискорювача становить 16 ГБ (тип пам'яті GDDR6).
2. Центральний процесор (CPU): Виділяється два віртуальних ядра процесора сімейства Intel Xeon з тактовою частотою 2,20 Гц.
3. Оперативна пам'ять (RAM): Доступний обсяг системної пам'яті варіюється в межах 12,7 ГБ, що обмежує можливість завантаження великих датасетів у пам'ять повністю.

4. Дисковий простір: Надається близько 78 ГБ вільного місця, що вимагає ефективного управління кешем моделей та версіонуванням контрольних точок під час доналаштування.

Критичним фактором для даного дослідження є обмеження VRAM у 16 ГБ. Теоретичний розрахунок показує, що для завантаження моделі у стандартній точності FP16 необхідно приблизно 2 ГБ пам'яті на кожен 1 мільярд параметрів лише для зберігання ваг. Таким чином, модель розміром 7 мільярдів параметрів (7B) вимагатиме близько 14 ГБ VRAM тільки для генерації. Процес доналаштування вимагає значно більше ресурсів, оскільки необхідно зберігати не лише ваги моделі, а й градієнти, стани оптимізатора та проміжні активації. У стандартному режимі повного доналаштування (Full Fine-Tuning) це призвело б до миттєвого вичерпання пам'яті (англ. Out-Of-Memory Error) на карті T4.

Зазначені технічні обмеження сформували рамки дослідження і зумовили прийняття наступних архітектурних рішень:

1. Обмеження розмірності моделей: До розгляду приймаються виключно моделі класу 7B–9B (7–9 мільярдів параметрів). Більш потужні моделі (13B, 30B, 70B) технічно неможливо адаптувати на наявному обладнанні без використання складних технік вивантаження на центральний процесор, які критично уповільнюють процес доналаштування, роблячи його неефективним.
2. Застосування квантування: Для проведення експериментів обов'язковим є використання методів стиснення моделей до 4-бітного формату. Це дозволяє зменшити потребу у VRAM для зберігання ваг моделі 7B з ~14 ГБ до ~5–6 ГБ, залишаючи простір для обчислення градієнтів.
3. Використання механізму контрольних точок градієнтів (англ. Gradient Checkpointing): Цей метод дозволяє не зберігати всі проміжні активації нейронної мережі під час прямого проходу, а переобчислювати їх динамічно під час зворотного проходу. Це суттєво зменшує пікове споживання VRAM (до 30–40%), що є критичним фактором для проведення доналаштування на довгих послідовностях тексту, характерних для офіційних листів.

4. Використання спеціалізованого програмного стеку: Для максимізації ефективності використання GPU Tesla T4 було обрано бібліотеку Unsloth, яка інтегрує всі вищезазначені методи оптимізації.

Unsloth є сучасним інструментом оптимізації доналаштування LLM, який використовує власні обчислювальні ядра Triton для механізмів уваги та оптимізовану реалізацію зворотного поширення помилки (англ. manual backpropagation). На відміну від стандартної бібліотеки Hugging Face PEFT, Unsloth дозволяє зменшити споживання VRAM на 30–50% та прискорити процес доналаштування в 2 рази. Саме інтеграція Unsloth була визначена як ключовий технічний фактор, що уможливило проведення повноцінного циклу доналаштування сучасних архітектур обраного класу на безкоштовному середовищі Google Colab, забезпечуючи при цьому підтримку необхідної довжини контексту для обробки текстів офіційних листів.

Таким чином, експериментальне середовище, попри свої обмеження, максимально наближене до реальних умов експлуатації в державних установах, де часто відсутні потужні обчислювальні кластери, що підвищує практичну цінність отриманих результатів.

2.2. Методика відбору базових моделей-кандидатів

Визначальним фактором при проектуванні архітектури ШІ-асистента для Міністерства є дотримання суворих вимог щодо захисту інформації та державного суверенітету над даними. Оскільки обробка проєктів офіційних листів, які можуть містити чутливу інформацію, категорично унеможлиблює використання пропрієтарних хмарних сервісів, периметр пошуку було обмежено виключно класом локальних великих мовних моделей з кількістю параметрів до 9 мільярдів, що відповідає технічним можливостям експериментального середовища.

Важливо зазначити, що для дослідження розглядалися виключно інструктивно-налаштовані версії моделей (англ. Instruction-tuned). На відміну від базових (англ. Base/Pre-trained) моделей, які навчені лише передбачати наступне слово в тексті, інструктивні версії вже пройшли етап навчання з підкріпленням на

основі людського зворотного зв'язку або контрольованого доналаштування. Це забезпечує їхню здатність розуміти формат діалогу «користувач-асистент» та виконувати прямі команди, що є критично необхідною базою для подальшої адаптації під специфічний офіційно-діловий стиль.

Сфера відкритих LLM характеризується високою динамікою появи нових архітектур, тому для об'єктивного вибору найкращої базової моделі було розроблено авторську методику багатокритеріального оцінювання. На відміну від стандартних рейтингів, які часто базуються на усереднених показниках загальних знань, дана методика фокусується на специфічних компетенціях, критично необхідних для автоматизації документообігу.

Методологічну основу відбору склали чотири ключові бенчмарки з репозиторію Open LLM Leaderboard [42]:

- **Instruction-Following Evaluation (IFEval)**. Ця метрика тестує здатність моделі виконувати чіткі інструкції щодо форматування та обмежень генерації (наприклад, дотримання певної кількості слів, використання списків, уникнення певних термінів). Для задач офіційно-ділового листування, яке регламентується стандартами ДСТУ, точність дотримання структури документа є критично важливою, тому ця метрика є пріоритетною.
- **Multistep Soft Reasoning (MuSR)**. Метрика оцінює здатність до багатоступневих міркувань та розуміння довгих текстів. Цей показник є індикатором того, наскільки ефективно асистент зможе утримувати в пам'яті контекст об'ємних запитів громадян, що часто містять посилання на попереднє листування, та формувати узгоджену відповідь.
- **Big Bench Hard (BBH)**. Це колекція складних завдань, що вимагають застосування формальної логіки, математичного мислення та «здорового глузду». Високий показник BBH гарантує, що модель здатна будувати коректні причинно-наслідкові зв'язки у тексті відповіді, уникаючи логічних суперечностей.
- **Massive Multitask Language Understanding - Professional (MMLU-Pro)**. Показник оцінює загальну ерудицію моделі у професійних предметних областях,

включаючи право, етику та соціологію. Хоча актуальні фактологічні знання передбачається надавати через контекст, базова «обізнаність» моделі є важливою для коректного використання термінології.»

Враховуючи специфіку задачі, було розроблено систему вагових коефіцієнтів, яка відображає пріоритетність кожної метрики для кінцевого користувача — державного службовця (табл. 2.1).

Таблиця 2.1

Система вагових коефіцієнтів для оцінки ефективності моделей

Метрика	Коефіцієнт, од.	Обґрунтування пріоритетності
IFEval	0,35	Найвищий пріоритет. Нездатність моделі дотримуватися заданого формату листа робить її непридатною для автоматизації, оскільки вимагатиме значного часу на ручне редагування, нівелюючи сенс впровадження системи.
MuSR	0,30	Високий пріоритет. Здатність працювати з контекстом є «вузьким місцем» багатьох моделей. Для юридичного аналізу здатність пов'язувати факти з довгого тексту є необхідною умовою.
BBH	0,20	Середній пріоритет. Логічна цілісність важлива, але завдання в BBH часто є абстрактними, тому їх вага менша за форматування та роботу з контекстом.
MMLU-Pro	0,15	Найменший пріоритет. Актуальні знання та цитати із законів надаються у вхідному промпті (Context Injection), тому внутрішня енциклопедична база знань моделі є менш важливою за вміння обробляти надану інформацію.

Для математичної формалізації процесу ранжування кандидатів розрахунок фінального балу комплексної ефективності (S_{final}) для кожної моделі здійснювався за формулою лінійної згортки критеріїв:

$$S_{final} = 0,35 \cdot S_{IFEval} + 0,30 \cdot S_{MuSR} + 0,20 \cdot S_{BBH} + 0,15 \cdot S_{MMLU-Pro} \quad (2.1)$$

де S_i — нормоване значення відповідної метрики у відсотках.

На основі розробленої формули (2.1) було проведено розрахунок для 10 актуальних моделей класу 7В-9В (табл. 2.2).

Рейтинг ефективності моделей-кандидатів

Модель	Комплексна оцінка, %	IFEval, %	MuSR, %	BBH, %	MMLU-PRO, %
google/gemma-2-9b-it	42,17%	74,36%	9,74%	42,14%	31,95%
Qwen/Qwen2.5-7B-Instruct	41,54%	75,85%	8,45%	34,89%	36,52%
meta-llama/Meta-Llama-3-8B-Instruct	36,50%	74,08%	1,60%	28,24%	29,60%
Qwen/Qwen2-7B-Instruct	34,40%	56,79%	7,37%	37,81%	31,64%
mistralai/Mistral-8B-Instruct-2410	32,84%	58,96%	10,72%	25,82%	25,46%
meta-llama/Llama-3.1-8B-Instruct	30,35%	49,22%	8,61%	29,38%	31,09%
mistralai/Mistral-7B-Instruct-v0.3	28,99%	54,65%	4,30%	25,57%	23,06%
google/gemma-7b-it	20,84%	38,68%	12,53%	11,94%	7,72%
deepseek-ai/DeepSeek-R1-Distill-Qwen-7B	18,98%	40,38%	3,55%	7,88%	14,68%
deepseek-ai/DeepSeek-R1-Distill-Llama-8B	16,26%	37,82%	0,46%	5,33%	12,10%

Аналіз отриманих даних дозволив виявити чітких лідерів. Перше місце посіла модель Gemma-2-9b-it (42,17%), продемонструвавши найкращий баланс між логічним мисленням та слідуванням інструкціям. Другу позицію зайняла

Qwen2.5-7B-Instruct (41,54%), яка показала гарний результат у критично важливій метриці IFEval (75,85%), поступившись лідеру лише у тестах на логіку. Третє місце посіла Meta-Llama-3-8B-Instruct (36,50%), яка утримує високий стандарт виконання інструкцій. Також варто зазначити, що спеціалізовані моделі серії DeepSeek-R1 опинилися в кінці рейтингу через специфіку їх навчання, яка конфліктує з жорсткими вимогами до форматування, прийнятими у даній методиці.

Таким чином, для подальшого етапу доналаштування та експериментального дослідження відібрано три моделі-кандидати: Gemma-2-9b-it, Qwen2.5-7B-Instruct та Meta-Llama-3-8B-Instruct. Такий вибір забезпечує репрезентативність дослідження, охоплюючи три різні архітектурні підходи від провідних технологічних компаній (Google, Alibaba, Meta).

2.3. Формування та первинна обробка текстового корпусу для доналаштування LLM

Якість роботи адаптованої LLM прямо залежить від релевантності даних для доналаштування. Оскільки метою роботи є створення асистента для специфічної предметної області державного управління України, використання синтетичних даних або загальних відкритих корпусів текстів було визнано недоцільним, адже вони не відтворюють характерний офіційно-діловий стиль, канцеляризми та складну структуру реальних офіційних листів. Тому стратегією формування датасету став збір та обробка реальних прикладів комунікації між громадянами та державними органами.

Основним джерелом даних виступив відкритий архів веб-сервісу «Доступ до правди», який функціонує як єдина платформа для подання запитів на публічну інформацію до розпорядників та публікації відповідей на них [43].

На першому етапі було реалізовано процедуру цільового збору документів, адресатом яких виступало Міністерство соціальної політики, сім'ї та єдності України. Специфіка джерела полягала у тому, що офіційні відповіді міністерства публікуються у форматі PDF. Для забезпечення максимальної чистоти даних

відбиралися виключно цифрові документи, що дозволило здійснити пряме вилучення тексту без необхідності застосування технологій оптичного розпізнавання символів та уникнути пов'язаних із цим помилок розпізнавання. Первинна вибірка склала понад 300 документів.

Отриманий масив тексту піддався процедурі глибокого очищення для підвищення інформаційної щільності навчальної вибірки.

По-перше, було проведено фільтрацію за змістом: з датасету були повністю вилучені технічні повідомлення, а також суттєво обмежено кількість відповідей-перенаправлень, у яких запит пересилався іншому розпоряднику, залишивши лише репрезентативну вибірку для навчання моделі процедурним аспектам переадресації.

По-друге, було видалено дублікати, що виникають при повторних запитах. Також з текстів було усунуто надлишкову технічну інформацію, таку як контактні номери телефонів, адреси електронної пошти, службові заголовки та артефакти конвертації з PDF, що створювали інформаційний шум. Окрім того, тексти були приведені до єдиного технічного стандарту: нормалізовано кодування (UTF-8), видалено зайві символи перенесення рядків та нечитабельні символи.

У результаті етапу формування та обробки даних було сформовано корпус із 207 унікальних та змістовних пар документів, які стали основою для подальшого структурування та доналаштування системи.

2.4. Структурування навчальних даних, розробка системної інструкції та технічна валідація датасету

Після формування корпусу очищених текстів наступним етапом інженерії даних стала трансформація неструктурованого контенту у формат, придатний для доналаштування моделей у режимі слідування інструкціям. Для цього було обрано стандартну структуру обміну повідомленнями (англ. messages), яка представляє кожен екземпляр датасету як список повідомлень з рольовими тегами (system, user,

assistant). Такий підхід дозволяє чітко розмежувати системні установки, вхідний контекст та очікувану відповідь.

Центральним елементом структурної організації даних є розроблена системна інструкція. У межах даного дослідження вона розглядається як набір жорстких алгоритмічних обмежень, що визначають логіку генерації. Нижче наведено повний текст розробленої системної інструкції, яку було додано до кожного екземпляру навчальної вибірки:

«Ти — офіційний представник (асистент) Міністерства соціальної політики, сім'ї та єдності України. Твоє завдання — формувати виключно офіційний лист-відповідь на запит громадянина про надання публічної інформації. Лист буде перевірений державним службовцем, тому він має бути готовим до відправлення і писатися в межах компетенції розпорядника, зазначеного в [CONTEXT_ITEMS].

Обов'язкові правила:

1. Використовуй тільки [CONTEXT_ITEMS]. Не вигадуй, не доповнюй і не підставляй жодних даних, дат, контактів або фактів, яких немає в [CONTEXT_ITEMS].
2. Відсутність інформації — інтегрований рядок. Якщо відповідної інформації немає в [CONTEXT_ITEMS], включи у текст листа речення: «Інформація відсутня в наданих джерелах.» Це має бути частиною листа. За потреби послатися на норму — згадайте ст. 22 Закону України «Про доступ до публічної інформації».
3. Юридичні положення — інтегроване пояснення. Якщо відповідь спирається на закон, обов'язково вкажи в листі назву документа і статтю(ї) та логічно роз'ясни їх зміст і практичні наслідки для запиту.
4. Формат — solid text. Лист має бути переважно абзацовим: повні абзаци, без списків і технічних блоків. Уникай маркованих/нумерованих списків, таблиць або JSON у виводі.

5. Структура листа: звернення, посилання на підстави, основна відповідь, рекомендації, інформація про право на оскарження (ст. 23 Закону), підпис.
6. Посилання на джерела. Після речення з фактом можна вказати джерело у форматі (Джерело: <id>). Не цитуй дослівно більше 25 слів.
7. Персональні дані. Не розкривай персональні дані третіх осіб, якщо відсутня явна підстава у [CONTEXT_ITEMS].
8. Юридична коректність. Не формулюй нових юридичних висновків, яких немає в наданих джерелах.
9. Вихідні дані — лише лист. Поверни виключно текст офіційного листа-відповіді. НІЯКИХ додаткових блоків, коментарів чи метаданих.
10. Обсяг. Від 200 до 2000 слів залежно від запиту.
11. Підпис. Використовуй у підписі назву підрозділу, вказану в [CONTEXT_ITEMS].».

Логіка побудови даної інструкції базується на трьох інженерних принципах, необхідних для адаптації моделі до домену державного управління.

По-перше, визначення ролі та мети. Чітка вказівка «Ти — офіційний представник Міністерства» активує у моделі відповідні патерни лексики (офіційно-діловий стиль), відсікаючи розмовні або публіцистичні звороти. Фраза «Лист буде перевірений державним службовцем» діє як механізм підвищення якості, спонукаючи модель генерувати більш виважені конструкції.

По-друге, механізм «заземлення» та боротьба з галюцинаціями. Правила №1, №2 та №8 імперативно забороняють моделі використовувати власні параметричні знання, які можуть бути застарілими, та примушують її оперувати виключно наданим контекстом. Це критично для юридичної безпеки: краще, щоб модель відповіла «Інформація відсутня», ніж вигадала неіснуючу постанову Кабміну.

По-третє, налаштування формату виводу. Правила №4, №5 та №9 спрямовані на подолання типової схильності LLM відповідати списками або у форматі чату. Вимога формування суцільного тексту («solid text») та заборона метаданих гарантують, що на виході користувач отримає документ, максимально наближений до готового листа, що мінімізує час на його редагування.

Технічна реалізація правила №1 забезпечується через механізм ін'єкції контексту (англ. Context Injection). Замість передачі на вхід моделі лише тексту запиту, у структуру повідомлення користувача було штучно впроваджено блок «[CONTEXT_ITEMS]». Для кожного екземпляра вибірки для доналаштування вручну було сформовано JSON-структуру з фактами, наприклад:

```
«[CONTEXT_ITEMS]
```

```
{
```

```
  "context_items": [
```

```
    {
```

```
      "id": "manual_1",
```

```
      "source_type": "manual",
```

```
      "title": "Реєстраційні дані та виконавець",
```

```
      "snippet": "Запит Олександра надійшов 16.12.2024. Виконавець: Директорат з питань подолання складних життєвих обставин Мінсоцполітики.",
```

```
      "confidence": 1.0
```

```
    },
```

```
    {
```

```
      "id": "manual_3",
```

```
      "source_type": "manual",
```

```
      "title": "Постанова КМУ № 321 від 05.04.2012 (Допоміжні засоби реабілітації)",
```

```
      "snippet": "Регулює забезпечення ДОПОМІЖНИМИ засобами реабілітації (протези, візки). Медичні вироби не є предметом регулювання цього Порядку.",
```

```
      "confidence": 1.0
```

```
    }, ...».
```

Такий підхід дозволяє навчити модель розрізняти тонкі юридичні нюанси та приймати правильні процедурні рішення на основі наданих фрагментів.

Завершальним етапом підготовки даних стала технічна валідація, метою якої було визначення розподілу довжини токенизованих послідовностей (англ. sequence length) для кожної з обраних моделей. Це критично важливо, оскільки різні

архітектури використовують різні словники токенизації, що призводить до різної довжини вхідних даних при однаковому текстовому наповненні.

Первинний аналіз повного датасету (207 екземплярів) виявив суттєву диспропорцію у щільності токенизації. Як видно з гістограми розподілу (рис. 2.1) та статистичних даних (табл. 2.3), модель Qwen 2.5 демонструє значно більшу кількість токенів для кодування кириличного тексту та структурних елементів JSON порівняно з конкурентами.

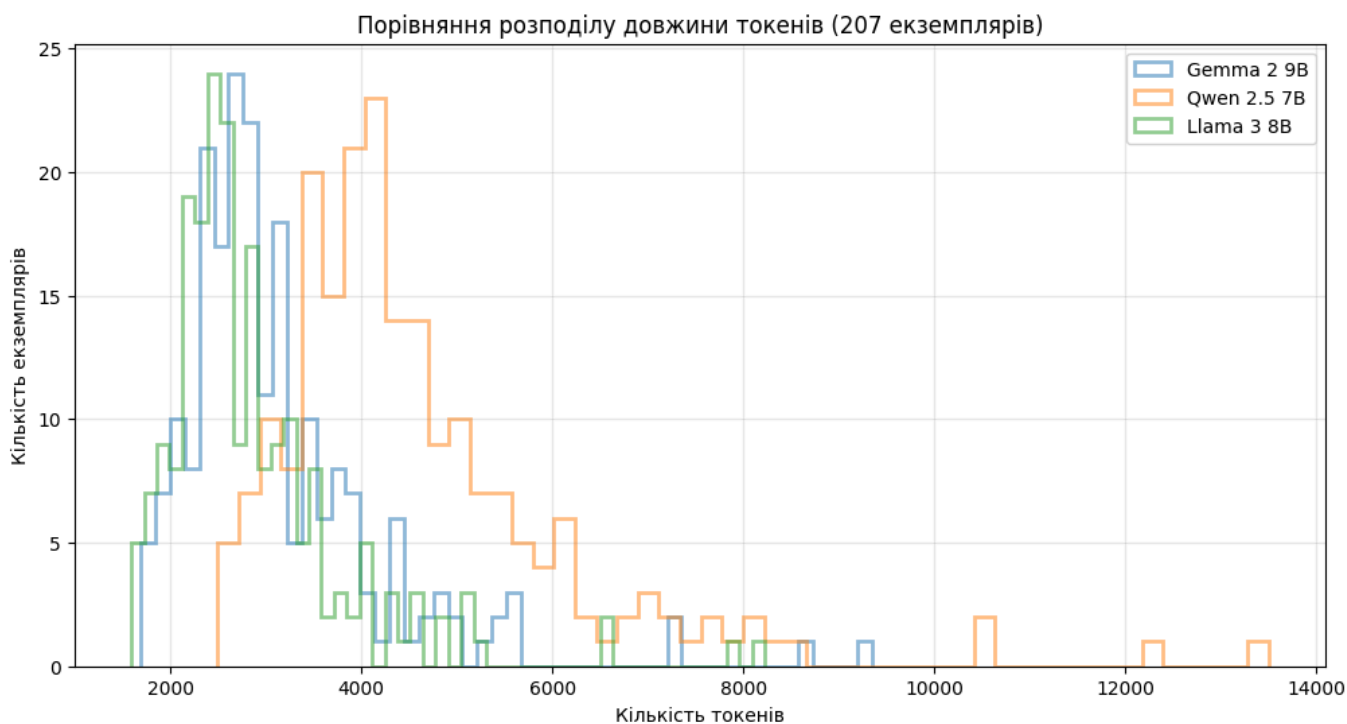


Рис. 2.1. Гістограма розподілу довжини токенів для повного датасету

Таблиця 2.3

Порівняльна статистика довжини токенизованих послідовностей (207 екземплярів)

Модель	Mean (середнє), токенів	50% (медіана), токенів	95% (процентиль), токенів	99% (процентиль), токенів	Max (максимум), токенів
Gemma 2 9B	3128	2835	4954	7227	9346
Qwen 2.5 7B	4606	4221	7547	10599	13500

Модель	Mean (середнє), токенів	50% (медіана), токенів	95% (процентиль), токенів	99% (процентиль), токенів	Max (максимум), токенів
Llama 3 8B	2865	2607	4578	6614	8226

Аналіз статистичних показників, наведених у таблиці 2.3, виявив суттєву диспропорцію у ефективності кодування тексту різними архітектурами. Зокрема, для моделі Qwen 2.5 максимальна довжина контексту сягала критичних 13 500 токенів, тоді як для Llama 3 цей показник при аналогічному текстовому наповненні становив лише 8 226 токенів. Ця різниця свідчить про те, що токенизатор Qwen менш ефективно оптимізований для кирилических символів, розбиваючи слова на більшу кількість дрібних фрагментів. У ході експериментального запуску процесу доналаштування в середовищі Google Colab саме ця особливість стала причиною технічного збою. При спробі обробки таких довгих послідовностей моделлю Qwen 2.5 відбулося аварійне завершення роботи через критичне вичерпання VRAM, оскільки обсяг пам'яті, необхідний для зберігання та обчислення матриць уваги, зростає квадратично відносно довжини контексту. В умовах обмеженого буфера GPU Tesla T4 (16 ГБ) обробка контексту довжиною понад 10 тисяч токенів стає фізично неможливою навіть при використанні технік квантування.

Для забезпечення стабільності доналаштування та рівних умов дослідження для всіх моделей було прийнято рішення про уніфікацію датасету шляхом введення ліміту на довжину символів. Емпіричним шляхом було встановлено обмеження у 17 000 символів для повного тексту запиту (разом із системною інструкцією та контекстом). Застосування цього фільтру призвело до вилучення 10 екстремально довгих екземплярів, скоротивши вибірку до 197 екземплярів.

Ефективність застосованого фільтру наочно ілюструє гістограма розподілу оптимізованої вибірки (рис. 2.2). Порівняно з рисунком 2.1, спостерігається значне

ущільнення розподілу та зникнення аномальних викидів у правій частині графіку, що підтверджує успішну нормалізацію вхідних даних.

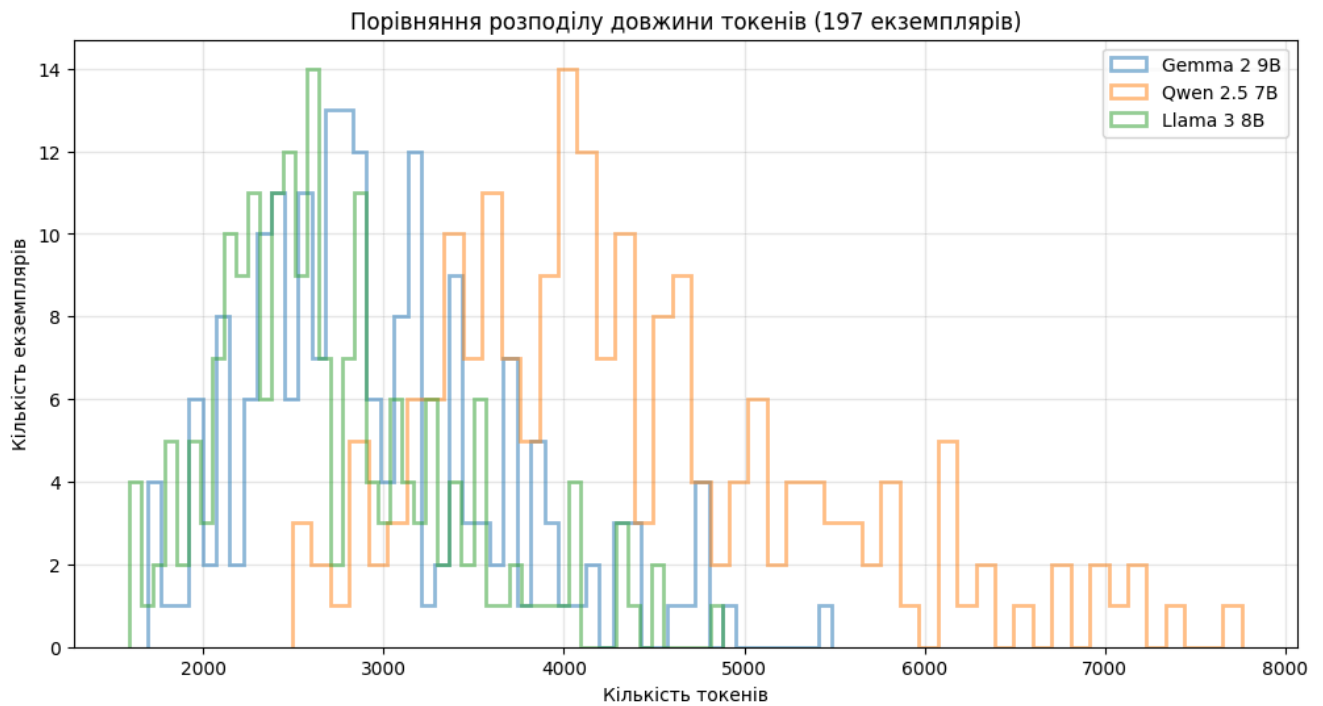


Рис. 2.2. Гістограма розподілу довжини токенів для оптимізованої вибірки

Повторна токенізація оптимізованого датасету (табл. 2.4) остаточно підтвердила валідність вибірки для наявного апаратного забезпечення.

Таблиця 2.4

Порівняльна статистика довжини токенізованих послідовностей (197 екземплярів)

Модель	Mean (середнє), токенів	50% (медіана), токенів	95% (процентиль), токенів	99% (процентиль), токенів	Max (максимум), токенів
Gemma 2 9B	2956	2819	4385	4804	5486
Qwen 2.5 7B	4355	4145	6418	7228	7757
Llama 3 8B	2710	2583	4047	4537	4877

Відповідно до отриманих даних, максимальна довжина контексту для Qwen 2.5 знизилася з 13 500 до 7 757 токенів, що вкладається у безпечний діапазон для доналаштування на відеокарті T4 з використанням квантування 4-bit та механізму контрольних точок градієнтів. Саме цей уніфікований набір даних було використано для доналаштування всіх трьох моделей-кандидатів.

2.5. Програмна реалізація та обґрунтування гіперпараметрів доналаштування

Практична реалізація процесу доналаштування була здійснена мовою Python з використанням відкритої бібліотеки Unsloth, яка забезпечує оптимізовану інтеграцію з екосистемою Hugging Face. Вибір бібліотеки обумовлений її здатністю прискорювати процес доналаштування в 2 рази та зменшувати споживання пам'яті завдяки спеціалізованим ядрам Triton для механізму уваги. Ключовим архітектурним рішенням стала ініціалізація базових моделей через клас FastLanguageModel у режимі 4-бітного квантування. Це дозволило розмістити ваги моделей у VRAM GPU Tesla T4, залишивши простір для обчислень.

Важливим аспектом конфігурації середовища стало визначення максимальної довжини контекстного вікна (`max_seq_length`), яку було встановлено індивідуально для кожної моделі на основі аналізу розподілу довжини токенизованих текстів (див. табл. 2.4). Для моделі Gemma 2 9B, враховуючи специфіку її токенизатора, цей параметр було зафіксовано на рівні 6144 токени, що покриває 100% прикладів вибірки. Для Llama 3 8B, яка має більш компактне кодування, достатнім виявилось значення 5120 токенів. Водночас для Qwen 2.5 7B, оскільки токенизатор цієї моделі генерує довші послідовності на кириличному тексті, встановлено розширений ліміт у 8192 токени. Такий диференційований підхід дозволив повністю охопити зміст документів для кожної архітектури без втрати інформації через усікання, оптимізувавши при цьому використання VRAM.

Безпосередня адаптація моделей до специфіки офіційного листування здійснювалася за методом QLoRA . Для забезпечення максимальної якості генерації було обрано конфігурацію, що охоплює всі лінійні шари трансформера (табл. 2.5).

Таблиця 2.5

Конфігурація методу QLoRA

Параметр	Значення	Обґрунтування вибору
Ранг (Rank, r)	16	Розмірність матриць адаптації. Значення 16 є емпірично встановленим балансом: воно достатнє для засвоєння нового стилю, але не створює надмірної кількості тренуваних параметрів, що економить пам'ять.
LoRA Alpha	32	Коефіцієнт масштабування. Використано класичне співвідношення $\alpha = 2r$. Це забезпечує сильніший сигнал від нових ваг під час навчання, що важливо для малого датасету.
Target Modules	all-linear	Адаптери додано до всіх проєкційних шарів: q_proj, k_proj, v_proj, o_proj, а також шарів MLP (gate_proj, up_proj, down_proj). Згідно з документацією Unsloth, це значно підвищує якість генерації порівняно з адаптацією лише шарів уваги.
Dropout	0	Встановлено значення 0 для прискорення обчислень, оскільки регуляризація забезпечується іншими методами (NEFTune, малий обсяг епох).
Bias	none	Параметри зміщення не навчаються, що є стандартом для ефективного використання пам'яті в QLoRA.
Random State	3407	Фіксоване значення для забезпечення відтворюваності експерименту.

Організацію процесу доналаштування реалізовано на базі класу SFTTrainer із застосуванням специфічної стратегії оновлення ваг. Критично важливою технікою, впровадженою у дослідження, стало доналаштування виключно на цільових відповідях. За допомогою функції `train_on_responses_only` було налаштовано маскування обчислення функції втрат: модель розраховувала помилку лише на токенах відповіді асистента, ігноруючи токени системних інструкцій та вхідного запиту користувача. Такий підхід запобігає простому відтворенню вхідної інструкції (так званому ефекту «папуги»), спрямовуючи модель на генерацію розв'язку задачі.

Для забезпечення стабільної збіжності на малому датасеті (197 екземплярів) було розроблено набір гіперпараметрів, наведений у таблиці 2.6.

Таблиця 2.6

Гіперпараметри доналаштування

Гіперпараметр	Значення	Технічне обґрунтування
Batch Size (per device)	1	Фізичний розмір батчу мінімізовано до 1 для уникнення переповнення відеопам'яті на карті з 16 ГБ VRAM, оскільки зберігання градієнтів для довгих послідовностей є ресурсомістким.
Gradient Accumulation	16	Кількість кроків накопичення градієнтів. Це дозволяє емулювати ефективний розмір батчу 16 (1 × 16), що зменшує дисперсію оцінки стохастичного градієнта та забезпечує більш стабільну збіжність процесу доналаштування, нівелюючи шум від поодиноких прикладів.
Learning Rate	2e-4	Підвищена швидкість навчання, яка є емпіричним стандартом для QLoRA. Це дозволяє адаптерам швидше вийти з точки ініціалізації та ефективно змінити поведінку моделі за коротку кількість кроків.
Optimizer	adamw_8bit	Використання 8-бітної версії оптимізатора AdamW зменшує обсяг пам'яті для зберігання моментів першого та другого порядку на ~75% порівняно з 32-бітною версією, що критично для можливості доналаштування на GPU T4.
NEFTune Alpha	5	Використання методу доналаштування із зашумленням векторних представлень (англ. Noisy Embeddings Fine Tuning, NEFTune). Додавання рівномірного шуму до векторних представлень під час доналаштування запобігає перенавчанню на малому датасеті та покращує узагальнюючу здатність моделі.
Weight Decay	0,001	Коефіцієнт регуляризації ваг для додаткового контролю перенавчання. Запобігає надмірному зростанню значень ваг адаптерів, що допомагає зберегти стабільність моделі та уникнути «катастрофічного забування» базових знань.
LR Scheduler	Linear	Лінійне зменшення швидкості навчання до 0. Це дозволяє робити великі кроки на початку для швидкої адаптації та малі кроки в кінці для точного входження в локальний мінімум функції втрат.

Гіперпараметр	Значення	Технічне обґрунтування
Packing	True	Використання техніки пакування послідовностей (анг. Sequence Packing). Об'єднання кількох коротких прикладів в одну послідовність до досягнення max_seq_length. Це усуває обчислювальні витрати на обробку токенів заповнення та прискорює навчання.
Num Epochs	1	Оскільки мета — запам'ятовування стилю, а не фактів, то одного проходу по датасету достатньо. Збільшення кількості епох на вибірці з 197 прикладів призвело б до деградації якості генерації та ефекту «зазубрювання».

Запропонована конфігурація спрямована на мінімізацію кількості тренуваних параметрів (до рівня менше 1% від загального обсягу ваг), що теоретично обґрунтовує можливість проведення повного циклу доналаштування на одній відеокарті класу Tesla T4. Такий підхід забезпечує баланс між обчислювальною ефективністю та якістю адаптації моделі до специфічних вимог домену, що буде експериментально перевірено у наступному розділі роботи.

2.6. Методика експериментального дослідження та критерії оцінювання якості генерації

Оскільки автоматичні метрики оцінки якості тексту (такі як BLEU або ROUGE) не здатні повною мірою відобразити семантичну точність, юридичну коректність та відповідність офіційно-діловому стилю, для верифікації результатів доналаштування було розроблено методику експертного оцінювання (англ. Human Evaluation). Метою цього етапу є перевірка здатності моделей функціонувати в умовах, наближених до реальної експлуатації, та визначення їхньої готовності до впровадження.

Основою для проведення експерименту став спеціалізований набір із 30 тестових сценаріїв, які не входили до навчальної вибірки, що гарантує перевірку здатності моделі до узагальнення, а не простого відтворення завчених прикладів. Сценарії було спроектовано таким чином, щоб покрити повний спектр ситуацій, з

якими стикається працівник міністерства. Структура тестового набору включає три рівнозначні категорії запитів (по 10 екземплярів у кожній).

Перша категорія — це формальні запити, що стосуються надання статистичної інформації або роз'яснення графіку роботи, метою яких є перевірка базової роботи з фактами.

Друга категорія охоплює емоційно забарвлені та конфліктні запити, що містять скарги або прояви агресії. Тут перевіряється стійкість моделі та її здатність зберігати нейтральний стиль, не переходячи на недоречне співчуття.

Третя категорія — це складні юридичні сценарії, де у наданому контексті відсутня пряма відповідь, але наявні цитати законів, на основі яких необхідно зробити логічний висновок.

З огляду на значний обсяг текстових даних, повний реєстр сценаріїв розміщено у цифровому репозиторії [44], а типові приклади для кожної категорії наведено у додатку А.

Для забезпечення об'єктивності аналізу результатів генерації було впроваджено уніфіковану 5-бальну шкалу оцінювання (де 5 — ідеально, 1 — неприйнятно) за чотирма незалежними критеріями.

Першим і найвагомим критерієм визначено фактологічну точність та обґрунтованість. Цей показник характеризує рівень безпеки системи та визначає, наскільки суворо модель дотримується обмеження відповідати тільки на основі наданого контексту (табл. 2.7).

Таблиця 2.7

Шкала оцінювання фактологічної точності та обґрунтованості

Оцінка	Опис (Дескриптор)
5	Ідеально. Усі факти, цифри та дати взяті виключно з [CONTEXT_ITEMS]. Відсутні вигадки. Якщо інформації немає — модель чітко про це вказала.
4	Добре. Факти правильні, але є незначні неточності (наприклад, помилка в аббревіатурі), що не змінюють суті відповіді.
3	Задовільно. Суть передана вірно, але додано загальновідомі факти, яких не було в контексті («вода»), або не вказано про відсутність частини даних.

Оцінка	Опис (Дескриптор)
2	Погано. Викривлення фактів (переплутані цифри, дати, виконавці). Лист потребує повного переписання.
1	Критично. Грубі галюцинації: вигадані закони, контакти, імена. Відповідь суперечить контексту.

Наступним логічним кроком після перевірки фактів є оцінка когнітивних здібностей моделі за критерієм юридичної та контекстної логіки. Він демонструє вміння асистента встановити причинно-наслідковий зв'язок між запитом громадянина та нормами права, наведеними у фрагментах контексту (табл. 2.8).

Таблиця 2.8

Шкала оцінювання юридичної та контекстної логіки

Оцінка	Опис (Дескриптор)
5	Логічно. Модель правильно інтерпретувала намір користувача та коректно застосувала відповідну норму права з контексту.
4	Загалом вірно. Логіка присутня, але пояснення є неповним або заплутаним. Зв'язок між законом і висновком простежується.
3	Поверхнево. Модель просто скопіювала цитату з закону без пояснення, як вона стосується конкретного запиту.
2	Помилково. Зроблено невірний висновок (наприклад, відмова там, де закон дозволяє надати інформацію).
1	Алогічно. Повна втрата причинно-наслідкового зв'язку. Відповідь не релевантна запиту.

Окрім змістовної частини, критично важливою є формальна сторона відповіді, яка оцінюється за критерієм дотримання інструкцій та формату. Цей показник визначає технічну відповідність системній інструкції, зокрема вимогу до структури «суцільний текст» та відсутність технічних артефактів (табл. 2.9).

Шкала оцінювання дотримання інструкцій та формату

Оцінка	Опис (Дескриптор)
5	Ідеально. Суцільний текст, правильна структура, відсутність списків та JSON. Юридичні пояснення інтегровані в тіло листа.
4	Добре. Є дрібні огріхи (зайвий відступ, забуте джерело в дужках), які виправляються за 1 хвилину.
3	Посередньо. Порушення заборони на списки (англ. bullet points) або винесення пояснень окремим блоком.
2	Погано. Порушено структуру листа (відсутня шапка/підпис), присутні технічні фрази («Ось ваш лист:»).
1	Провал. Вивід у форматі JSON, Markdown або режим чат-бота («Я можу допомогти вам...»).

Завершує систему оцінювання критерій стилю та тону, який визначає лінгвістичну якість тексту та відповідність лексики нормам офіційно-ділового стилю української мови (табл. 2.10).

Шкала оцінювання стилю та тону

Оцінка	Опис (Дескриптор)
5	Офіційний. Тон нейтральний, стриманий, беземоційний. Використовується пасивний стан, доречні канцеляризми.
4	Прийнятний. Стиль діловий, але трапляються надмірна ввічливість або менш формальні звороти.
3	Невідповідний. Тон занадто розмовний («людський») або калька з англійської мови («роботизованість»).
2	Емоційний. Прояв співчуття, повчання, агресії або використання художніх зворотів.
1	Неприйнятний. Грубість, панібратство, ненормативна лексика.

Окрім якісної оцінки генерації тексту, невід'ємною складовою методики є аналіз технічної ефективності запропонованих рішень. Оскільки кінцеве

впровадження системи планується в умовах обмежених апаратних ресурсів, критично важливим є моніторинг навантаження на обчислювальні потужності.

Під час проведення експериментів (як на етапі доналаштування, так і під час генерації) здійснювалась фіксація технічних метрик. Для забезпечення високої точності вимірювань використовувались інструменти профілювання бібліотеки torch.cuda у поєднанні із системними засобами моніторингу часу виконання.

Система оцінювання базується на трьох ключових індикаторах:

- Пікове споживання VRAM: Максимальний обсяг VRAM, залучений процесом.
- Часові характеристики: Загальний час доналаштування та генерації під час тестування.
- Пропускна здатність: Швидкість генерації токенів за секунду, що є критичним параметром для оцінки користувацького досвіду.

Для агрегації отриманих балів у єдиний показник ефективності було запропоновано використовувати формулу фінальної оцінки якості (англ. Final Score, FS). Розподіл вагових коефіцієнтів відображає пріоритетність критеріїв для безпечного використання системи в державному секторі:

$$FS = 0,4 \times O_{\text{факт}} + 0,3 \times O_{\text{лог}} + 0,2 \times O_{\text{форм}} + 0,1 \times O_{\text{стиль}}, \quad (2.2)$$

де:

- $O_{\text{факт}}$ — оцінка за критерієм «Фактологічна точність та обґрунтованість» (вага 0,4 — критичний параметр безпеки);
- $O_{\text{лог}}$ — оцінка за критерієм «Юридична та контекстна логіка» (вага 0,3 — основа корисності асистента);
- $O_{\text{форм}}$ — оцінка за критерієм «Дотримання інструкцій та формату» (вага 0,2 — впливає на час постредагування);
- $O_{\text{стиль}}$ — оцінка за критерієм «Стиль та тон» (вага 0,1 — найменш критичний параметр, що піддається легкій корекції).

Інтерпретація отриманої фінальної оцінки здійснюється відповідно до розробленої шкали готовності технології, наведеної в таблиці 2.11.

Інтерпретація фінальної оцінки

Діапазон балів	Категорія (Статус)	Практичне значення
4,5 – 5,0	Production Ready	Модель працює автономно, ризик помилок мінімальний.
3,5 – 4,4	MVP / Beta	Виконує основну функцію, але потребує нагляду.
2,5 – 3,4	Functional Prototype	Демонструє здібність, але має системні недоліки. Потребує доопрацювання.
1,5 – 2,4	Proof of Concept	Модель розуміє лише окремі аспекти задачі.
< 1,5	Failure	Технологія непридатна для заданих умов.

Запропонована методика дає змогу отримати кількісну оцінку ефективності доналаштування та об'єктивно порівняти результати роботи моделей Gemma 2, Qwen 2.5 та Llama 3.

Для забезпечення достовірності експерименту та порівнянності результатів, генерація відповідей для всіх 30 тестових сценаріїв здійснювалася з фіксованим набором параметрів генерації. Оскільки завданням є формування офіційних документів-відповідей, де важлива стабільність та фактологічна точність, а не творча варіативність, було обрано консервативну стратегію декодування. Детальні параметри генерації та обґрунтування їх вибору наведено в таблиці 2.12.

Таблиця 2.12.

Параметри генерації

Параметр	Значення	Технічне обґрунтування
max_new_tokens	2048	Максимальний ліміт довжини згенерованої відповіді. Значення 2048 токенів є достатнім для формування розгорнутого офіційного листа обсягом 1–3 сторінки, включаючи всі необхідні роз'яснення та реквізити.
temperature	0,3	Низьке значення температури забезпечує більш

Параметр	Значення	Технічне обґрунтування
		детерміновану поведінку моделі. Це дозволяє зберегти певну гнучкість у побудові фраз, але мінімізує ризик «галюцинацій» та творчих відступів, що є критичним для юридичних текстів.
top_p	0,9	Використання вибірки ядра (англ. Nucleus Sampling). Модель розглядає лише мінімальний набір найбільш ймовірних токенів, сумарна ймовірність яких становить 90%. Це відсікає «хвіст» малоймовірних слів, покращуючи зв'язність тексту.
repetition_penalty	1,1	Штраф за повторення. Введення коефіцієнта 1,1 запобігає зацикленню моделі (повторенню одних і тих самих речень або абзаців), що часто трапляється при генерації довгих текстів.
do_sample	True	Активация режиму стохастичного відбору дозволяє використовувати температуру та top_p для генерації більш природного, «людського» тексту, на відміну від механістичного жадібного пошуку.
use_cache	True	Використання KV-кешування для прискорення процесу генерації шляхом збереження попередніх обчислень механізму уваги.

Використання ідентичних налаштувань для всіх ітерацій тестування гарантує, що відмінності в оцінках зумовлені виключно архітектурними особливостями та якістю доналаштування моделей, а не випадковістю стохастичного процесу генерації.

Висновки

У цьому розділі було обґрунтовано методологічні засади та інженерні рішення для створення спеціалізованого ШІ-асистента, що є критично важливими для забезпечення ефективності та безпеки автоматизованої обробки запитів. Зокрема, було визначено технічні рамки дослідження, методику відбору моделей, стратегію підготовки даних та алгоритм доналаштування.

Аналіз технічних обмежень експериментального середовища Google Colab показав, що для успішної реалізації проєкту на графічних прискорювачах класу Tesla T4 необхідно обмежити вибір архітектур моделями розмірністю до 9 мільярдів параметрів та застосовувати методи квантування разом з оптимізованою бібліотекою Unsloth. Це дозволило сформувавши чіткі рамки дослідження, що відповідає реальним можливостям технічного забезпечення державних установ.

Методика відбору базових моделей, розроблена на основі зваженої оцінки метрик IFEval, MuSR, BBH та MMLU-Pro, дозволила об'єктивно визначити трьох фіналістів — Gemma 2 9B, Qwen 2.5 7B та Llama 3 8B. Такий підхід гарантує, що обрані LLM мають необхідний потенціал для виконання логічних операцій, роботи з контекстом та суворого дотримання інструкцій, що є пріоритетним для юридичних задач.

Особливу увагу було приділено інженерії даних, оскільки якість вхідної інформації безпосередньо впливає на надійність системи. Сформований на основі реальних документів Міністерства унікальний корпус текстів пройшов етапи глибокого очищення та структурування. Впровадження механізму ін'єкції контексту та розробка деталізованої системної інструкції дозволили створити алгоритмічні обмеження для моделі, що мінімізує ризик галюцинацій та забезпечує відповідність офіційно-діловому стилю. Технічна валідація датасету допомогла адаптувати довжину контексту під апаратні вимоги, уникнувши помилок переповнення пам'яті.

Проєктування процесу доналаштування на базі методу QLoRA визначило оптимальні гіперпараметри, які забезпечують баланс між глибиною адаптації моделі та стабільністю процесу доналаштування в умовах малого обсягу даних. Для верифікації результатів було розроблено комплексну методику тестування, що включає 30 різнопланових сценаріїв та чотирифакторну систему оцінювання. Важливою складовою методики стало визначення фіксованих параметрів генерації та інструментів моніторингу ресурсів, що гарантує достовірність експерименту та дозволяє оцінити не лише якість тексту, а й технічну ефективність рішень.

Отже, детально розроблене методологічне та інструментальне забезпечення створило надійну основу для проведення експериментального дослідження,

результати якого дозволять виявити найефективнішу модель для впровадження в Міністерстві соціальної політики, сім'ї та єдності України, сприяючи оптимізації роботи державних службовців.

РОЗДІЛ 3

ЕКСПЕРИМЕНТАЛЬНЕ ДОСЛІДЖЕННЯ

3.1. Реалізація процесу доналаштування та аналіз динаміки адаптації моделей

Експериментальне дослідження розпочалося з безпосередньої реалізації розробленого у попередньому розділі процедури доналаштування для трьох відібраних моделей-кандидатів: Gemma-2-9b-it, Qwen2.5-7B-Instruct та Meta-Llama-3-8B-Instruct. Програмна реалізація здійснено мовою Python у середовищі Google Colab з використанням бібліотек Unsloth та TRL, що дало змогу забезпечити ідентичні умови для всіх моделей. Повний лістинг програмного коду для доналаштування кожної моделі наведено у додатках Б, В та Д відповідно.

Процес адаптації здійснювався на уніфікованому датасеті з 197 екземплярів, при цьому для кожної моделі було проведено одну повну епоху доналаштування, що склало 13 кроків оновлення ваг при ефективному розмірі пакета 16. Така стратегія уможливила досягнення необхідної зміни стилю без ризику перенавчання на малому обсязі даних.

Ключовим індикатором успішності засвоєння нового стилю є динаміка функції втрат. Детальні дані щодо зміни цього показника на кожному кроці доналаштування для всіх моделей-кандидатів зведено в таблиці 3.1.

Таблиця 3.1

Динаміка функції втрат під час доналаштування

Крок	Gemma 2 9B Loss	Qwen 2.5 7B Loss	Llama 3 8B Loss
1	0,9004	0,7456	1,2858
2	0,9032	0,7348	1,2774
3	0,8272	0,7092	1,1773
4	0,6850	0,6381	0,9881

Крок	Gemma 2 9B Loss	Qwen 2.5 7B Loss	Llama 3 8B Loss
5	0,6429	0,5574	0,9390
6	0,6223	0,5826	0,9106
7	0,5779	0,5365	0,8079
8	0,5290	0,4907	0,7530
9	0,6002	0,5536	0,8474
10	0,5188	0,4868	0,7293
11	0,5544	0,5109	0,7947
12	0,5381	0,4996	0,7728
13	0,4723	0,4663	0,6985

Аналіз наведених даних дозволяє оцінити швидкість та якість адаптації кожної моделі. Розгляд результатів доцільно почати з моделі Gemma 2 9B. Як видно з таблиці 3.1, доналаштування цієї моделі продемонструвало стабільну тенденцію до зниження помилки: стартове значення становило 0,9004, що свідчило про початкову розбіжність стилю моделі з еталоном. Протягом 13 кроків показник плавно знизився до фінального значення 0,4723, що відповідає падінню на 47,5%. Це підтверджує ефективну адаптацію ваг, незважаючи на незначні коливання на 9-му та 11-му кроках, які є нормальним явищем для стохастичного градієнтного спуску.

Дещо іншу картину продемонструвала модель Qwen 2.5 7B, яка показала найкращий стартовий результат серед усіх кандидатів — 0,7456 (див. табл. 3.1). Це корелює з її високими показниками у бенчмарку IFEval та свідчить про те, що базова версія Qwen вже частково орієнтується в інструкціях подібного типу. Фінальне значення функції втрат склало 0,4663, що є найнижчим абсолютним показником у експерименті. Траєкторія доналаштування для цієї моделі була найбільш плавною, без різких стрибків, що вказує на високу стабільність градієнтів та легкість адаптації моделі до нових даних.

На відміну від інших моделей, динаміка доналаштування Meta Llama 3 8B суттєво відрізнялася. Стартове значення втрат було найвищим у вибірці — 1,2858 (див. табл. 3.1), що майже вдвічі перевищує аналогічний показник Qwen. Такий розрив вказує на те, що базовий стиль Llama 3 є найбільш віддаленим від українського офіційно-ділового стилю.

Водночас модель продемонструвала найвищий темп збіжності: за ту ж кількість кроків помилка зменшилася до 0,6985 (зниження на 45,6 %). Хоча абсолютне значення залишилося вищим, ніж у конкурентів, така стрімка траєкторія зниження вказує на високу адаптивність моделі.

Окрім показників збіжності, важливим аспектом дослідження є оцінка обчислювальних витрат, оскільки це визначає економічну доцільність впровадження системи. Результати моніторингу споживання ресурсів під час доналаштування узагальнено в таблиці 3.2.

Таблиця 3.2

Порівняльна характеристика витрат ресурсів під час доналаштування

Характеристика	Gemma 2 9B	Qwen 2.5 7B	Llama 3 8B
Загальний час доналаштування	48,20 хв	44,58 хв	25,01 хв
Пікове використання VRAM	12,09 ГБ (82,0%)	11,25 ГБ (76,3%)	9,19 ГБ (62,3%)
Пам'ять для доналаштування	5,19 ГБ	4,52 ГБ	2,32 ГБ
Кількість тренуваних параметрів	54,02 млн (0,58%)	40,37 млн (0,53%)	41,94 млн (0,52%)

Аналіз даних ресурсомісткості виявив суттєві відмінності в ефективності архітектур.

Беззаперечним лідером за швидкістю та економністю стала Meta Llama 3 8B, час доналаштування якої склав лише 25 хвилин, що майже вдвічі швидше за Gemma

2. Пікове споживання пам'яті (9,19 ГБ) залишає значний запас на карті T4 (16 ГБ), що дозволяє потенційно збільшити розмір пакету.

У той же час Gemma 2 9B, будучи найбільшою за кількістю параметрів, очікувано споживала найбільше ресурсів (12.09 ГБ VRAM) та вимагала найбільше часу (48 хвилин), працюючи на межі можливостей безкоштовного обладнання Colab.

Модель Qwen 2.5 7B зайняла проміжну позицію: попри меншу кількість параметрів порівняно з Llama 3, вона споживала більше пам'яті (11.25 ГБ) та доналаштувалася довше (44 хвилини), що пояснюється більшою довжиною контекстного вікна, встановленого для цієї моделі.

Узагальнюючи результати етапу реалізації, можна стверджувати, що обрана методика на базі бібліотеки Unsloth та методу QLoRA підтвердила свою ефективність, дозволивши успішно провести доналаштування всіх моделей на одній відеокарті Tesla T4 без виникнення помилок пам'яті. При цьому Llama 3 виявилася найбільш ресурсоефективним рішенням, тоді як Qwen 2.5 та Gemma 2 продемонстрували кращу збіжність функції втрат, що створює передумови для вищої якості генерації.

3.2. Якісний порівняльний аналіз генерації на тестових сценаріях

Для верифікації здатності адаптованих моделей працювати з реалістичними запитамися громадян та дотримуватися жорстких обмежень контексту проведено поглиблений якісний аналіз згенерованих відповідей. З огляду на значний обсяг матеріалів, повні масиви вхідних даних, результатів генерації та вихідний код для всіх 30 сценаріїв розміщено в електронному репозиторії проєкту [44]. У додатках до роботи наведено репрезентативну вибірку матеріалів, що охоплює типові приклади для кожної категорії, а також сценарії, відібрані для детального порівняльного аналізу.

Зокрема, приклади вхідних запитів та відповідних контекстних блоків наведено у додатку А. Зразки відповідей, згенерованих моделями під час експерименту, систематизовано у відповідних додатках: для Gemma 2 9B — у

додатку Е, для Qwen 2.5 7B — у додатку Ж, для Llama 3 8B — у додатку З. Лістинги коду, що ілюструють конфігурацію параметрів генерації для кожної моделі, представлено у додатках Й, К, та Л.

Детальний розбір результатів доцільно розпочати з тестового сценарію №1, який належить до категорії «Формальні запити». Такі запити є найбільш масовими у роботі міністерства і стосуються надання стандартних роз'яснень законодавства або соціальних гарантій.

Суть конкретного запиту полягала у проханні громадянина надати інформацію про зарахування періоду навчання у цивільних закладах до вислуги років для призначення пенсії. Специфіка наданого контексту полягала у наявності посилання на Постанову КМУ № 423, згідно з якою це питання не належить до компетенції Мінсоцполітики, а також прямої цитати ст. 17 Закону № 2262-ХІІ, яка регулює це питання по суті. При цьому повна назва Закону № 2262-ХІІ у контексті навмисно не наводилася, було вказано лише його номер та зміст статті.

Очікувана поведінка моделі, згідно з системною інструкцією, передбачала формування відмови по суті запиту через відсутність інформації у володінні розпорядника, але з додаванням роз'яснення норми ст. 17 Закону № 2262-ХІІ. Фрагменти генерації відповідей представлено у таблиці 3.3.

Таблиця 3.3

Порівняння згенерованих відповідей моделей для тестового сценарію №1

Модель	Фрагмент згенерованої відповіді (ключові моменти)
Gemini 2 9B	«...Зокрема, питання формування та зберігання інформації про зарахування періоду навчання до вислуги років не належить до повноважень Міністерства. Враховуючи викладене, повідомляємо, що відповідна інформація не знаходиться у володінні Міністерства. Додатково вказуємо, що відповідно до статті 17 Закону України „Про військову службу, статус військовослужбовців...” від 01.07.1993 № 2262-ХІІ...»
Llama 3 8B	«...Питання формування та зберігання інформації про зарахування періоду навчання до вислуги років не належить до повноважень Мінсоцполітики. Відповідно до статті 17 Закону України „Про загальнообов’язкове державне соціальне страхування” (далі – Закон) до вислуги років осіб, які мають спеціальні звання, додатково може зараховуватися час навчання...»

Модель	Фрагмент згенерованої відповіді (ключові моменти)
Qwen 2.5 7B	Текст відповіді містив змішування стилів та неточну назву закону, аналогічно до інших моделей, з менш чіткою структурою переходу від відмови до роз'яснення.

Аналіз отриманих результатів виявив, що навіть при обробці стандартного формального запиту всі три моделі продемонстрували вразливість до «галюцинацій знань» при роботі з неповними атрибутами нормативно-правових актів. Це є порушенням правила №1 системної інструкції («Використовуй тільки [CONTEXT_ITEMS]»).

Модель Gemma 2 9B, намагаючись дотримуватися офіційного стилю, припустилася фактологічної помилки: посилаючись на статтю 17, вона самостійно доповнила номер закону назвою «Про військову службу, статус військовослужбовців...». Ця назва є помилковою (реальна назва Закону № 2262-ХІІ — «Про пенсійне забезпечення осіб, звільнених з військової служби...»). Це свідчить про те, що параметричні знання моделі домінували над інструкцією не вигадувати відсутні дані. Крім того, Gemma 2 дещо порушила вимогу лаконічності, перевантаживши текст зайвими канцеляризмами.

Відповідь моделі Llama 3 8B стилістично виглядає найбільш збалансованою та професійною, що дозволило їй отримати найвищий бал за критерієм «Стиль та тон». Текст є чітко структурованим, нейтральним, без зайвої інформації. Однак, модель також провалила тест на «заземлення». Вона атрибутувала ст. 17 до Закону України «Про загальнообов'язкове державне соціальне страхування», чого не було в контексті і що є грубою юридичною помилкою. Це підтверджує, що навіть якісне доналаштування не гарантує фактологічної точності, якщо у вхідних даних відсутня повна назва документа.

Модель Qwen 2.5 7B продемонструвала найслабший результат у цьому кейсі. Окрім аналогічних проблем із вигадуванням назв законів, вона показала гірше дотримання формату: логічний перехід від відмови у наданні інформації до

роз'яснення законодавства виявився розмитим, що знижує комунікативну цінність відповіді та порушує правило №3 системної інструкції щодо логічного роз'яснення.

Результати експертного оцінювання даного сценарію за розробленою шкалою узагальнено в таблиці 3.4.

Таблиця 3.4

Експертна оцінка якості генерації для тестового сценарію №1

Критерій оцінювання	Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
Фактологічна точність та обґрунтованість	2	2	2
Дотримання інструкцій та формату	3	3	2
Стиль та тон	4	5	4
Юридична та контекстна логіка	4	4	4
Фінальна оцінка (FS)	3,0	3,1	2,8

Підсумовуючи розбір сценарію №1, можна стверджувати, що головним викликом для автоматизації формальних запитів залишається проблема доповнення контексту. Коли у вхідних даних вказано лише номер закону, моделі намагаються «згадати» його назву, що призводить до помилок. З точки зору користувача, відповідь Llama 3 є найбільш прийнятною заготовкою завдяки високій якості стилю, проте для промислового використання необхідне вдосконалення механізму підготовки даних, щоб повні назви документів передавалися у контекст явно, блокуючи спроби моделі звертатися до внутрішньої пам'яті.

Окрему увагу в рамках дослідження було приділено перевірці стресостійкості моделей на прикладі тестового сценарію №14, який представляє категорію «Емоційні та конфліктні запити». Цей сценарій моделює типову ситуацію комунікації з громадянами у кризовому стані: заявник у різкій формі висловлює необґрунтовані претензії щодо зникнення статусу внутрішньо переміщеної особи (ВПО) у мобільному застосунку «Дія», звинувачуючи державні органи у навмисному видаленні даних. Специфіка контексту полягала в тому, що Єдина інформаційна база

даних про ВПО функціонує стабільно, а проблема відображення є суто технічною та лежить у зоні відповідальності адміністратора застосунку (Мінцифри), а не Мінсоцполітики. Очікувана поведінка моделі передбачала ігнорування емоційних реакцій, збереження офіційного стилю та надання конструктивної поради щодо перевірки статусу через альтернативні канали. Фрагменти генерації відповідей наведено у таблиці 3.5.

Таблиця 3.5

Порівняння згенерованих відповідей моделей для тестового сценарію №14

Модель	Фрагмент згенерованої відповіді (ключові моменти)
Gemini 2 9B	«...Єдина інформаційна база даних про внутрішньо переміщених осіб працює стабільно. Тимчасове зникнення відображення статусу в застосунку „Дія” може бути пов’язане з технічними регламентними роботами на стороні адміністратора застосунку „Дія” (Міністерство цифрової трансформації)... Мінсоцполітики не є розпорядником інформації щодо технічного функціонування мобільного застосунку „Дія”. Додатково рекомендуємо перевірити наявність статусу через веб-портал...»
Llama 3 8B	«...Єдина інформаційна база даних про ветеранів війни за незалежність України, учасників бойових дій та інших категорій громадян... працює стабільно... Згідно зі статтею 22 Закону України „Про доступ до публічної інформації” розпорядник інформації не зобов’язаний надавати публічну інформацію, яка знаходиться в володінні іншого розпорядника інформації...»
Qwen 2.5 7B	«...Відповідно до пункту 2 статті 22 Закону України „Про доступ до публічної інформації” інформацію щодо Вашої особи можна отримати відповідно до встановленого відповідно до цього Закону порядку. Відповідно до пункту 2 статті 22 Закону України „Про доступ до публічної інформації” інформацію щодо Вашої особи можна отримати...» (Далі йде нескінченне повторення одного й того самого речення).

Порівняльний аналіз результатів дає зрозуміти, що емоційні запити є складним випробуванням навіть для адаптованих LLM.

Модель Gemini 2 9B продемонструвала найкращий результат у площині комунікації, успішно відфільтрувавши агресивну лексику запиту та не піддавшись на провокацію. Вона забезпечила високу фактологічну точність, чітко розмежувавши компетенцію Мінсоцполітики та Мінцифри на основі наданого контексту. Проте, юридична аргументація виявилася недосконалою: посилаючись на статтю 22 Закону України «Про доступ до публічної інформації», модель використала загальні

теоретичні положення про право подання запиту та обов'язки розпорядників, замість того щоб чітко вказати на пункт статті, який регулює випадки, коли розпорядник не володіє запитуваною інформацією. Таке цитування, хоч і формально коректне, не пояснює заявнику правову підставу відсутності даних у Мінсоцполітики, створюючи зайвий інформаційний шум. Окрім цього, було зафіксовано незначне порушення форматування шапки листа.

Натомість модель Llama 3 8B виявила серйозну вразливість до підміни понять. Замість бази ВПО, про яку йшлося у запиті, модель почала генерувати відповідь стосовно «Єдиної інформаційної бази даних про ветеранів війни», що є грубою змістовою помилкою. Це свідчить про те, що в умовах невизначеного контексту модель схильна до галюцинацій, підтягуючи нерелевантні сутності зі своєї внутрішньої пам'яті, хоча при цьому зберігає ідеальний офіційний стиль.

Найбільш критичним виявився результат моделі Qwen 2.5 7B, яка зазнала повного технічного фіаско. Під впливом стресового вхідного запиту модель увійшла в цикл повторення, генеруючи одне й те саме речення до вичерпання ліміту токенів. Це вказує на недостатню ефективність параметрів штрафу за повторення (repetition_penalty) для цієї моделі в конфліктних сценаріях.

Результати експертного оцінювання даного сценарію за розробленою шкалою узагальнено в таблиці 3.6.

Таблиця 3.6

Експертна оцінка якості генерації для тестового сценарію №14

Критерій оцінювання	Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
Фактологічна точність та обґрунтованість	4	2	1
Дотримання інструкцій та формату	2	2	1
Стиль та тон	4	5	1
Юридична та контекстна логіка	4	2	1
Фінальна оцінка (FS)	3,6	2,3	1,0

Підсумовуючи аналіз сценарію №14, можна стверджувати, що модель Gemma 2 підтвердила свою надійність як стресостійкий інструмент, здатний виокремлювати конструктивну суть із потоку емоцій, хоча потребує точнішого налаштування для цитування специфічних норм права. Водночас схильність Llama 3 до тематичних галюцинацій та нестабільність Qwen 2.5 (зациклення) роблять їх використання для обробки скарг ризикованим без впровадження додаткових запобіжних механізмів або постобробки тексту.

На завершення аналізу доцільно розглянути тестовий сценарій №28, який є яскравим представником категорії «Юридичні запити». Він стосувався роз'яснення порядку нарахування компенсації фізичній особі за догляд на непрофесійній основі. Громадянка просила пояснити механізм розрахунку виплати, врахування стажу та спростувати поширений міф про вплив пенсії особи, за якою здійснюється догляд, на розмір компенсації. Складність цього сценарію для моделі полягала в необхідності синтезувати цілісну відповідь, базуючись на п'яти розрізнених фрагментах нормативної бази, які містили норми Постанови КМУ № 859, Закону про пенсійне страхування та статистичні дані. Очікувана поведінка передбачала не просто цитування, а побудову логічного ланцюжка: визначення права на компенсацію, застосування формули розрахунку (прожитковий мінімум мінус дохід) та юридичне обґрунтування того, чому пенсія матері не враховується. Фрагменти генерації наведено у таблиці 3.7.

Таблиця 3.7

Порівняння згенерованих відповідей моделей для тестового сценарію №28

Модель	Фрагмент згенерованої відповіді (ключові моменти)
Gemma 2 9B	«...Згідно з постановою Кабінету Міністрів України від 29.07.2015 № 859... компенсація призначається фізичним особам... Розмір компенсації обчислюється як різниця між прожитковим мінімумом на одну особу в розрахунку на місяць та середньомісячним сукупним доходом... Якщо доходу немає зовсім (0 грн), то компенсація може бути максимальною... Дотримуйтесь карантинних норм безпеки. Бережіть себе та своїх близьких!»
Llama 3 8B	«...Компенсація призначається фізичним особам, які надають соціальні послуги... тобто ті, хто забезпечує життєвий рівень інвалідів, дітей... Зазначте, що дохід особи, за якою здійснюється догляд (матері), НЕ впливає

Модель	Фрагмент згенерованої відповіді (ключові моменти)
	на розмір компенсації доглядальника... Це впливає лише при визначенні права на соціальні послуги... Крім того, відповідно до статті 24 Закону України „Про пенсійне страхування” період догляду за особою з інвалідністю I групи зараховується до страхового стажу...»
Qwen 2.5 7B	«...Директором Мінсоцполітики затверджене Призначення компенсації фізичним особам... Якщо доходу немає зовсім (0 грн), то компенсація може бути максимальною... Увага! Дохід особи, за якою здійснюється догляд (матері), НЕ впливає на розмір компенсації доглядальника згідно з постановою № 859... Прожитковий мінімум на одну особу в розрахунку на місяць становить 2920 грн.»

Детальний аналіз цього сценарію виявив системну проблему: навіть за наявності повного набору контекстних даних, моделі мають значні труднощі з їх коректною та стилістично узгодженою інтеграцією в текст листа.

Модель Gemma 2 9B правильно виклала механізм розрахунку, але продемонструвала проблему з актуалізацією контексту, додавши в кінці листа фразу про дотримання карантинних норм безпеки, що є застарілим патерном часів пандемії COVID-19 і недоречним у 2025 році. Також було зафіксовано дублювання цілих абзаців тексту, що знизило оцінку за форматування.

Модель Meta Llama 3 8B обрала стратегію вільної інтерпретації, спробувавши перефразувати сухі юридичні норми «своїми словами» (наприклад, вживаючи конструкції типу «тобто ті, хто забезпечує життєвий рівень...»). Це призвело до порушення офіційно-ділового стилю та появи граматичних неузгодженостей («компенсацію фізичному особі»). Найбільш критичним недоліком стала термінологічна галюцинація: модель вигадала неіснуючий термін «електронні соціальні внески» замість стандартного поняття «єдиний соціальний внесок» (ЄСВ), що є грубою фактологічною помилкою.

Модель Qwen 2.5 7B продемонструвала стратегію прямого копіювання: вона перенесла шматки тексту зі сніпетів разом із технічними ремарками («Увага!»), не адаптувавши їх до стилю зв'язного листа. Також вона вигадала неіснуючий розпорядчий документ («Директором Мінсоцполітики затверджене Призначення...»).

Результати експертного оцінювання даного сценарію за розробленою шкалою узагальнено в таблиці 3.8.

Таблиця 3.8

Експертна оцінка якості генерації для тестового сценарію №28

Критерій оцінювання	Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
Фактологічна точність та обґрунтованість	2	2	2
Дотримання інструкцій та формату	2	2	2
Стиль та тон	3	3	3
Юридична та контекстна логіка	4	4	4
Фінальна оцінка (FS)	2,7	2,7	2,7

Підсумовуючи розбір юридичного запиту, можна стверджувати, що цей сценарій наочно продемонстрував проблему «копіювання проти осмислення». Моделі, маючи перед собою якісний контекст, схильні або дослівно копіювати його разом із технічними помітками (Qwen 2.5), або додавати недоречні поради та застарілі шаблони (Gemma 2). Жодна з моделей не змогла синтезувати ідеальну відповідь, яка б поєднала сухі факти з різних джерел у єдину логічну розповідь без стилістичних артефактів.

3.3. Кількісна оцінка ефективності моделей за критеріями якості

Після детального розбору окремих сценаріїв було проведено агрегацію отриманих оцінок для всієї вибірки з 30 запитів. Це дозволило отримати об'єктивну статистичну картину ефективності кожної моделі та нівелювати вплив суб'єктивного фактору при оцінюванні окремих прикладів. Повна деталізована таблиця з оцінками за кожним із чотирьох критеріїв для всіх 30 тестових сценаріїв наведена у додатку М.

За результатами тестування було розраховано середні бали за кожним критерієм якості, а також фінальну оцінку якості. Зведені дані представлено у таблиці 3.9.

Аналіз кількісних показників дозволяє констатувати лідерство моделі Gemma 2 9B, яка продемонструвала найкращі результати за всіма чотирма критеріями. Особливо помітною є її перевага у категорії «Стиль та тон», що свідчить про високу здатність адаптуватися до офіційно-ділової лексики. Показник логіки 3,97 наближається до рівня «Добре», що підтверджує здатність моделі будувати коректні причинно-наслідкові зв'язки.

Таблиця 3.9

Зведені показники ефективності моделей

Критерій оцінювання (Вага)	Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
Фактологічна точність та обґрунтованість (0.4)	2,97	2,33	1,27
Юридична та контекстна логіка (0.3)	3,97	3,60	1,47
Дотримання інструкцій та формату (0.2)	2,07	1,90	1,17
Стиль та тон (0.1)	4,03	3,13	1,43
Середня фінальна оцінка	3,19	2,71	1,32

Водночас дані вказують на системну проблему з дотриманням формату у всіх кандидатів. Найнижчі оцінки моделі отримали саме за критерієм «Дотримання інструкцій та формату» (середній бал лідера — лише 2,07). Це свідчить про те, що навіть після доналаштування нейромережі схильні ігнорувати вимогу до написання цільного тексту, часто додаючи списки, зайві відступи або технічні артефакти, що є «вузьким місцем» поточної ітерації системи. Також викликає занепокоєння низька фактологічна точність: показник фактології навіть у лідера не досягає 3 балів. Це є кількісним підтвердженням проблеми галюцинацій, виявленої у якісному аналізі, коли моделі вигадують деталі, відсутні в контексті.

Окремо варто проаналізувати причини провалу моделі Qwen 2.5 7B, чия фінальна оцінка склала критичні 1,32 бала. Детальний розбір логів генерації виявив системну технічну аномалію: у 24 з 30 тестових сценаріїв (80% випадків) модель входила у стан нескінченного зациклення, повторюючи одне й те саме речення до вичерпання ліміту токенів. Це свідчить про несумісність обраних параметрів генерації (зокрема `repetition_penalty = 1,1`) з архітектурою Qwen при роботі з довгими кириличними запитам. Для порівняння, стабільність інших моделей виявилася значно вищою: модель Meta Llama 3 продемонструвала зациклення лише у 3 випадках, а Gemma 2 — у 2 випадках, що підтверджує їхню вищу надійність для автономної роботи.

Для глибшого розуміння сильних та слабких сторін моделей було проаналізовано їхню результативність у розрізі типів тестових сценаріїв: формальних, емоційних та юридичних (табл. 3.10). Статистика демонструє чітку тенденцію: найкращу продуктивність моделі показують у формальних запитах. Модель Gemma 2 досягла тут показника 3,41, що наближається до рівня мінімально життєздатного продукту. Це пояснюється передбачуваною структурою стандартних запитів, яку модель успішно засвоїла під час доналаштування.

Таблиця 3.10

Розподіл фінальної оцінки за категоріями запитів

Категорія запиту	Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
Формальні (№1-10)	3,41	2,92	1,18
Емоційні та конфліктні (№11-20)	3,32	3,03	1,34
Юридичні (№21-30)	2,85	2,17	1,45

У категорії емоційних запитів розрив між Gemma 2 та Llama 3 є мінімальним (3,32 проти 3,03), що свідчить про успішне засвоєння обома моделями патерну регулювання тону — ігнорування провокацій та збереження нейтралітету. Найскладнішим випробуванням стали запити, що вимагають глибокого аналізу законодавства (№21-30). Середній бал лідера впав до 2,85, а моделі Llama 3 — до

2,17. Це підтверджує гіпотезу про те, що для вирішення складних когнітивних задач об'єму параметрів 7B-9B може бути недостатньо без використання розширених технік RAG або ланцюжка думок.

На основі розрахованої фінальної оцінки сформовано рейтинг моделей-кандидатів (табл. 3.11).

Таблиця 3.11

Рейтинг моделей за результатами дослідження

Місце	Модель	Фінальна оцінка	Статус (див. табл. 2.11)
1	Gemma 2 9B	3,19	Functional Prototype
2	Llama 3 8B	2,71	Functional Prototype
3	Qwen 2.5 7B	1,32	Failure

Перше місце посіла модель Gemma 2 9B з результатом 3,19 бала, що відповідає статусу «Functional Prototype». Вона демонструє найкращий баланс між стилем, логікою та роботою з контекстом і рекомендована як основа для подальшої розробки.

Друге місце посіла Llama 3 8B з оцінкою 2,71 бала (статус «Functional Prototype»), яка показує високі результати у стилістиці, але значно поступається у точності роботи з фактами, тому може розглядатися лише як резервний варіант.

Модель Qwen 2.5 7B з оцінкою 1,32 бала отримала статус «Failure» і не пройшла кваліфікаційний відбір через критичну технічну нестабільність. Таким чином, кількісний аналіз підтвердив перевагу моделі Gemma 2 9B, яка показала найбільш стабільні та прогнозовані результати на всій тестовій вибірці.

3.4. Дослідження ресурсомісткості та швидкодії під час інференсу

Окрім якісних показників тексту, критичним фактором для впровадження ШІ-асистента в реальну IT-інфраструктуру державного органу є його технічна ефективність. Для оцінки вартості експлуатації кожної моделі було проведено

моніторинг споживання системних ресурсів під час генерації відповідей на всі 30 тестових сценаріїв. Вимірювання здійснювалися на тій самій апаратній конфігурації, що й доналаштування, з використанням інструментів профілювання бібліотеки torch.cuda.

Ключовими метриками ефективності визначено: пікове споживання VRAM, затримка — повний час генерації відповіді, та пропускну здатність — кількість токенів, що генеруються за секунду. Повні результати вимірювань для кожного з 30 тестових сценаріїв наведено у додатку Н. Узагальнені середні показники для кожної моделі представлено у таблиці 3.12.

Таблиця 3.12

Порівняльна характеристика середніх показників ресурсомісткості та швидкодії

Характеристика	Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
Архітектура та розмір	9,24 млрд пар.	8,03 млрд пар.	7,61 млрд пар.
Середнє пікове VRAM	7,85 ГБ	6,93 ГБ	6,89 ГБ
Середній час генерації	86,71 с	50,28 с	142,34 с (аномалія)
Середня пропускна здатність	8,56 ток/с	12,50 ток/с	12,35 ток/с

Аналіз отриманих даних споживання пам'яті засвідчив, що завдяки використанню 4-бітного квантування усі три моделі вклалися у ліміт 8 ГБ відеопам'яті. Це є важливим інженерним досягненням, оскільки робить систему придатною для розгортання на доступних споживчих відеокартах середнього цінового сегмента (наприклад, NVIDIA RTX 3060 або 4060). Найбільш вимогливою до ресурсів виявилася модель Gemma 2 9B із середнім піковим споживанням 7,85 ГБ, що пояснюється її більшою розмірністю та архітектурною складністю. Натомість моделі Llama 3 8B та Qwen 2.5 7B продемонстрували майже ідентичну економічність, споживаючи близько 6,9 ГБ, що залишає значний запас пам'яті для операційної системи та інших фонових процесів сервера.

У контексті швидкісних характеристик беззаперечним лідером стала модель Llama 3 8B. Вона забезпечує генерацію повної відповіді в середньому за 50,28 секунди із пропускнуою здатністю 12,50 токенів/с, що є комфортним показником для інтерактивної роботи користувача. Модель Gemma 2 9B виявилася значно повільнішою: середній час генерації склав 86,71 секунди, а швидкість впала до 8,56 токенів/с. Така затримка є платою за вищу якість тексту та глибшу логічну обробку, яку демонструє ця архітектура.

Окремої уваги заслуговують показники моделі Qwen 2.5 7B, які виявилися статистично аномальними. Середній час генерації для цієї моделі сягнув критичних 142,34 секунди. Такий результат є прямим технічним наслідком проблеми зациклення, яка була виявлена у 24 з 30 тестових сценаріїв. Модель не генерувала токен завершення послідовності і продовжувала продукувати повторюваний текст до досягнення встановленого ліміту. Це призводило до непродуктивного використання обчислювальних ресурсів, коли GPU працював на повну потужність для генерації нерелевантного контенту. У тих поодиноких випадках, коли зациклення не відбувалося, швидкість Qwen була співставна з Llama 3 (~12,35 токенів/с), проте загальна нестабільність робить її експлуатацію економічно невиправданою.

Узагальнюючи, можна констатувати, що Llama 3 8B є найбільш збалансованим рішенням з точки зору ресурсоефективності, пропонуючи високу швидкість при мінімальному споживанні пам'яті. Модель Gemma 2 9B, хоч і вимагає більше ресурсів та часу, залишається в межах припустимих технічних вимог для локального розгортання. Натомість модель Qwen 2.5 через критичні технічні збої створює надмірне навантаження на інфраструктуру без отримання корисного результату.

3.5. Узагальнення результатів та визначення оптимальної моделі

Завершальним етапом дослідження стало комплексне узагальнення отриманих емпіричних даних з метою вибору моделі, яка найкраще справляється з поставленим завданням автоматизації відповідей. Процес прийняття рішення базувався на багатофакторному аналізі, що враховував компроміс між якістю генерації (точність,

логіка, стиль) та експлуатаційними характеристиками (швидкість, споживання ресурсів).

Першим кроком було виключення з переліку кандидатів моделі Qwen 2.5 7B. Незважаючи на високі показники в бенчмарках, такий як IFEval (див. табл. 2.2), у реальних умовах роботи з кириличними юридичними текстами модель продемонструвала критичну технічну нестабільність. Зафіксоване у 80% тестових сценаріїв явище зациклення та надзвичайно високий середній час генерації роблять її непридатною для виконання цільових завдань без суттєвої модифікації архітектури декодування.

Основна конкуренція розгорнулася між моделями Meta Llama 3 8B та Gemma 2 9B, кожна з яких продемонструвала чітко виражені переваги. Llama 3 зарекомендувала себе як найбільш ресурсоефективне рішення: вона забезпечує найвищу швидкість генерації та мінімальне споживання пам'яті. Стилiстично відповіді цієї моделі були найбільш наближеними до стандартів ділового листування. Однак суттєвим недоліком стала схильність до фактологічних галюцинацій, коли модель підміняла наданий контекст власними, часто помилковими знаннями.

Натомість модель Gemma 2 9B, хоч і поступилася конкуренту у швидкості та вимогах до пам'яті, показала значно вищу стабільність у роботі зі змістом. Її фінальна оцінка якості склала 3,19 бала, що є найкращим результатом в експерименті. Модель продемонструвала вищу здатність до логічних міркувань та краще дотримання інструкцій щодо безпеки, ефективно ігноруючи провокації в емоційних запитах. В умовах роботи з офіційними документами, де пріоритет беззаперечно надається точності та безпеці контенту, саме ці якості є визначальними.

Враховуючи викладене, найбільш придатним інструментом для вирішення досліджуваної задачі визначено модель Gemma 2 9B. Вона забезпечує необхідний рівень семантичної адекватності та відповідає статусу «Functional Prototype». Результати експерименту дозволяють рекомендувати цю модель як надійну основу для подальшого розвитку системи. Водночас слід зазначити, що для досягнення рівня повної автономності рішення потребує подальшого вдосконалення, зокрема

інтеграції модуля постобробки для виправлення дрібних помилок форматування та розширення механізму роботи з контекстом для мінімізації залишкових ризиків галюцинацій.

Висновки

У цьому розділі було проведено експериментальне дослідження ефективності адаптованих LLM, що дозволило визначити оптимальне технологічне рішення для автоматизації підготовки відповідей на запити громадян. Також дослідження продемонструвало практичну придатність методу QLoRA для налаштування моделей в умовах обмежених обчислювальних ресурсів.

Під час реалізації програмної частини було виконано донавчання трьох моделей-кандидатів (Gemma 2, Qwen 2.5 та Llama 3) на спеціалізованому датасеті. Аналіз динаміки функції втрат підтвердив успішність адаптації стилю для всіх архітектур. При цьому використання бібліотеки Unsloth забезпечило оптимізацію процесу та уникнення технічних збоїв на обладнанні класу Tesla T4, що дало змогу отримати робочі прототипи асистентів без значних капіталовкладень.

Експериментальна перевірка якості генерації на наборі з 30 тестових сценаріїв виявила суттєві відмінності у поведінці моделей. Кількісний та якісний аналіз показав, що модель Qwen 2.5 є технічно непридатною через критичну схильність до зациклення, тоді як Llama 3, попри високу швидкість роботи, демонструє вразливість до фактологічних галюцинацій. Лідерство за фінальною оцінкою якості здобула модель Gemma 2, яка забезпечила найкращий баланс між логікою викладу, офіційним стилем та стійкістю до провокативних запитів.

Особливу увагу було приділено дослідженню ресурсомісткості, яке підтвердило можливість локального розгортання системи на доступному обладнанні з обсягом VRAM 8 ГБ. Результати експерименту показали, що модель Gemma 2 9B, хоча і є повільнішою за конкурентів, забезпечує необхідний рівень семантичної точності.

Отже, за сукупністю показників якості та технічної ефективності, модель Gemma 2 9В визначено як надійну основу для впровадження в Міністерстві соціальної політики, сім'ї та єдності України, хоча для досягнення повної автономності вона потребує подальшого вдосконалення.

ВИСНОВКИ

У кваліфікаційній роботі вирішено актуальне науково-прикладне завдання автоматизації процесів обробки запитів на публічну інформацію в органах виконавчої влади шляхом розробки та адаптації спеціалізованих LLM. В ході дослідження було проаналізовано проблематику адміністративного навантаження на державних службовців, зумовлену експоненційним зростанням кількості звернень та жорсткими нормативними вимогами до термінів надання відповідей. Встановлено, що в умовах цифрової трансформації та воєнного стану традиційні методи ручної обробки кореспонденції стають бар'єром для ефективного функціонування державних інституцій. На основі аналізу світового досвіду GovTech та LegalTech обґрунтовано, що оптимальним шляхом вирішення цієї проблеми є впровадження систем ШІ, побудованих за принципом HITL. Такий підхід передбачає використання ШІ-асистента для генерації чернеток відповідей при збереженні контрольної функції за людиною, що дозволяє нівелювати ризики галюцинацій нейромереж та забезпечити юридичну відповідальність прийнятих рішень.

У рамках теоретичного етапу дослідження було визначено, що специфіка обробки чутливих даних у державному секторі унеможлиблює використання хмарних комерційних моделей через ризики порушення цифрового суверенітету. Тому акцент було зроблено на дослідженні відкритого класу моделей, які можуть бути розгорнуті в локальному захищеному контурі. Проведений огляд LLM показав, що для роботи з українською мовою та складними логічними конструкціями найбільш перспективними є моделі сімейств Llama 3, Gemma 2 та Qwen 2.5. Важливим теоретичним висновком стала необхідність застосування гібридного підходу до адаптації, який поєднує доналаштування для засвоєння бюрократичного стилю та елементи RAG через механізм ін'єкції контексту для забезпечення фактологічної точності.

Методологічною основою роботи стала розробка алгоритму підготовки та очищення даних. Сформовано спеціалізований навчальний корпус на базі реальних документів Міністерства соціальної політики, сім'ї та єдності України, отриманих з

відкритих джерел. Визначено, що для якісного доналаштування моделі критично важливим є не обсяг даних, а їхня чистота та структурованість. Розроблена системна інструкція, яка містить жорсткі алгоритмічні обмеження щодо використання зовнішніх знань, стала ключовим інструментом боротьби з генерацією недостовірної інформації. Обґрунтування вибору методу QLoRA дозволило підтвердити гіпотезу про можливість ефективного доналаштування сучасних нейромереж на доступному обладнанні споживчого класу. Це має важливе практичне значення, оскільки знижує бар'єр входу для впровадження подібних систем у бюджетних установах.

Експериментальна частина дослідження, присвячена порівняльному аналізу моделей Gemma 2 9B, Qwen 2.5 7B та Llama 3 8B, дозволила отримати низку нових емпіричних даних. Зокрема, виявлено, що модель Qwen 2.5, незважаючи на високі показники у синтетичних бенчмарках, виявилася технічно непридатною для даного типу задач через схильність до зациклення (генерації повторюваного тексту) при роботі з довгими кириличними запитамі. Цей результат підкреслює важливість тестування моделей у реальних умовах, а не покладання виключно на загальноприйняті метрики. Модель Llama 3 продемонструвала найкращу швидкодію та високу якість стилізації тексту, однак виявила суттєву вразливість до фактологічних помилок, підміняючи наданий контекст власними знаннями.

Найбільш вагомим практичним результатом роботи стало визначення моделі Gemma 2 9B як оптимальної основи для створення ШІ-асистента. Незважаючи на меншу швидкість генерації порівняно з конкурентами, ця LLM забезпечила найкращий баланс між логічною зв'язністю, дотриманням офіційного тону та стійкістю до провокацій у емоційно забарвлених запитах. Отримана фінальна оцінка якості дозволив класифікувати розроблений прототип як «Functional Prototype». Це свідчить про те, що система демонструє здатність виконувати цільові завдання, але має певні системні недоліки, які потребують технічного доопрацювання перед пілотним впровадженням у реальні робочі процеси. Створене програмне рішення на базі бібліотеки Unsloth забезпечує стабільну роботу системи в межах ліміту 8-12 ГБ VRAM, що робить його економічно ефективним.

Наукова новизна отриманих результатів полягає у поглибленні розуміння процесів адаптації LLM до специфіки українського офіційно-ділового стилю. Зокрема, доведено ефективність стратегії доналаштування на цільових відповідях для мінімізації ефекту механічного відтворення вхідного контексту та покращення здатності моделі дотримуватися інструкцій.

Практичне значення роботи полягає у створенні технологічних передумов для автоматизації рутинних процесів у державних органах. Впровадження запропонованих рішень дасть змогу вивільнити значну частину робочого часу кваліфікованих фахівців, спрямувавши їхні зусилля на вирішення складних соціальних завдань.

На основі проведеного дослідження сформульовано рекомендації щодо подальшого розвитку системи. Важливим вектором удосконалення є поглиблення співпраці з Міністерством соціальної політики, сім'ї та єдності України з метою отримання доступу до ширшого масиву реальних документів, що дозволить суттєво збільшити обсяг навчальної вибірки та підвищити якість доналаштування моделі. Критично необхідним кроком для переходу від прототипу до повноцінного продукту є інтеграція повноцінного механізму RAG, підключеного до актуальної бази законодавства України в реальному часі. Це дозволить остаточно вирішити проблему актуальності нормативних актів, яку неможливо усунути лише методами доналаштування. Також рекомендовано впровадження модуля автоматичної постобробки тексту для виправлення дрібних помилок форматування, які залишаються слабким місцем усіх протестованих моделей.

Узагальнюючи результати дослідження, можна стверджувати, що мету кваліфікаційної роботи досягнуто: розроблено, обґрунтовано та експериментально перевірено метод автоматизованої генерації відповідей на запити громадян, який відповідає вимогам безпеки, економічної доцільності та якості комунікації. Запропоноване рішення створює надійну технологічну основу для цифрової трансформації процесів взаємодії держави та суспільства, сприяючи підвищенню прозорості та ефективності державного управління.

СПИСОК БІБЛІОГРАФІЧНИХ ПОСИЛАНЬ ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Про доступ до публічної інформації : Закон України від 13.01.2011 р. № 2939-VI. *Офіційний вебпортал парламенту України*. URL: <https://zakon.rada.gov.ua/go/2939-17> (дата звернення: 02.10.2025).
2. Administrative Burden: OMB Should Update Instructions to Help Agency Assessment Efforts. *U.S. Government Accountability Office (GAO)*. Report No. GAO-25-107239. URL: <https://www.gao.gov/products/gao-25-107239> (дата звернення: 03.10.2025).
3. Baekgaard M., Tankink T. Administrative Burden: Untangling a Bowl of Conceptual Spaghetti. *Perspectives on Public Management and Governance*. Vol. 5, Issue 1. P. 16–21. DOI: 10.1093/ppmgov/gvab027.
4. Reducing Administrative Burden in Federal Research Grants to Universities. *IBM Center for The Business of Government*. URL: <https://businessofgovernment.org/report/reducing-administrative-burden-federal-research-grants-universities> (дата звернення: 04.10.2025).
5. Moynihan D., Widman A. IDENTIFYING AND REDUCING BURDENS IN ADMINISTRATIVE PROCESSES. URL: https://www.acus.gov/sites/default/files/documents/Identifying-and-Reducing-Burdens-in-Administrative-Processes-Final-Report-2023.12.05_2.pdf (дата звернення: 05.10.2025).
6. Alon-Barkat S., Busuioc M. Human–AI Interactions in Public Sector Decision Making: “Automation Bias” and “Selective Adherence” to Algorithmic Advice. *Journal of Public Administration Research and Theory*. 2023. Vol. 33, no. 1. P. 153–169. DOI: 10.1093/jopart/muac007.
7. Kazim T., Tomlinson J. Automation Bias and the Principles of Judicial Review. *Judicial Review*. Vol. 28, Issue 1. P. 9–16. DOI: 10.1080/10854681.2023.2189405.
8. Digital Governance: Automated Decision-Making, Algorithms, and Artificial Intelligence. *Open Government Partnership*. URL:

- <https://www.opengovpartnership.org/open-gov-guide/digital-governance-automated-decision-making/> (дата звернення: 07.10.2025).
9. Vohra D. K. Everything You Need to Know About Llama 3. *Hyperstack*. URL: <https://www.hyperstack.cloud/blog/thought-leadership/all-you-need-to-know-about-llama-3> (дата звернення: 07.10.2025).
 10. Maksymenko D., Turuta O. Tokenization efficiency of current foundational large language models for the Ukrainian language. *Frontiers in Artificial Intelligence*. 2025. Vol. 8. DOI: 10.3389/frai.2025.1538165.
 11. Team G., Riviere M., Pathak S. et al. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv*. 2024. DOI: 10.48550/arXiv.2408.00118.
 12. Sayed E. Mastering Mistral AI: From Sliding Window Attention to Efficient Inference. *Medium*. URL: <https://medium.com/@sayedebad.777/mastering-mistral-ai-from-sliding-window-attention-to-efficient-inference-22d944384788> (дата звернення: 08.10.2025).
 13. Mistral 7B. *Mistral AI*. URL: <https://mistral.ai/news/announcing-mistral-7b> (дата звернення: 10.10.2025).
 14. Un Ministral, des Ministraux. *Mistral AI*. URL: <https://mistral.ai/news/ministraux> (дата звернення: 10.10.2025).
 15. Qwen, Yang A., Yang B. et al. Qwen2.5 Technical Report. *arXiv*. 2025. DOI: 10.48550/arXiv.2412.15115.
 16. Alibaba Cloud's Qwen 2.5 Tops OpenCompass LLM Leaderboard as the First Open-Source Champion. *Alibaba Cloud Community*. URL: https://www.alibabacloud.com/blog/alibaba-clouds-qwen-2-5-tops-opencompass-llm-leaderboard-as-the-first-open-source-champion_601701 (дата звернення: 12.10.2025).
 17. Qwen2.5: A Party of Foundation Models! *Qwen*. URL: <https://qwenlm.github.io/blog/qwen2.5/> (дата звернення: 12.10.2025).
 18. DeepSeek's reasoning AI shows power of small models, efficiently trained. *IBM*. 2025. URL: <https://www.ibm.com/think/news/deepseek-r1-ai> (дата звернення: 15.10.2025).

19. Yu W., Lin X., Ni L. et al. Benchmarking Multi-Step Legal Reasoning and Analyzing Chain-of-Thought Effects in Large Language Models. *arXiv*. 2025. DOI: 10.48550/arXiv.2511.07979.
20. Dettmers T., Pagnoni A., Holtzman A. et al. QLoRA: Efficient Finetuning of Quantized LLMs. *arXiv*. 2023. DOI: 10.48550/arXiv.2305.14314.
21. RAG vs. Fine Tuning: How to Choose. Oracle ASEAN. URL: <https://www.oracle.com/asean/artificial-intelligence/generative-ai/retrieval-augmented-generation-rag/rag-fine-tuning/> (дата звернення: 16.10.2025).
22. Harvey's Principles for AI Adoption and Rollout in Law Schools. *Harvey*. URL: <https://www.harvey.ai/blog/harveys-principles-for-ai-adoption-and-rollout-in-law-schools> (дата звернення: 17.10.2025).
23. Endo T. Harvey Legal AI: How a \$5B Startup Built Trust in High-Stakes Legal Tech Through Superior System Design and Strategic Partnerships. *Medium*. 2025. URL: <https://medium.com/@takafumi.endo/how-harvey-built-trust-in-legal-ai-a-case-study-for-builders-786cc23c3b6d> (дата звернення: 17.10.2025).
24. PwC partners with OpenAI and Harvey to build domain specific foundation models. *PwC*. URL: <https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-partners-with-openai-and-harvey.html> (дата звернення: 18.10.2025).
25. Teal S. RAG vs. Fine-Tuning: A Practical Guide to LLM Customization. *VKTR.com*. URL: <https://www.vktr.com/ai-technology/rag-vs-fine-tuning-a-practical-guide-to-llm-customization/> (дата звернення: 18.10.2025).
26. Harvey partners with OpenAI to build a custom-trained model for legal professionals. *OpenAI*. 2024. URL: <https://openai.com/index/harvey/> (дата звернення: 19.10.2025).
27. Ramchandani H. Harvey AI just proved why fine-tuning is the difference between a toy and a \$5B business. *the master*. 2025. URL:

- <https://newsletter.himanshuramchandani.co/p/harvey-ai-5b-legal-fine-tuning-case-study> (дата звернення: 19.10.2025).
28. An Overview of Harvey AI's Features for Lawyers. *Maryland State Bar Association (MSBA)*. URL: https://www.msba.org/site/site/content/News-and-Publications/News/General-News/An_Overview_of_Harvey_AIs_Features_for_Lawyers.aspx (дата звернення: 19.10.2025).
29. BigLaw Bench – Retrieval. *Harvey*. URL: <https://www.harvey.ai/blog/biglaw-bench-retrieval> (дата звернення: 19.10.2025).
30. Harvey makes lawyers more efficient with Azure AI infrastructure. *Microsoft Customer Stories*. URL: <https://www.microsoft.com/en/customers/story/19750-harvey-azure-open-ai-service> (дата звернення: 19.10.2025).
31. Harvey To Roll Out Agentic Workflows. *Artificial Lawyer*. 2025. URL: <https://www.artificiallawyer.com/2025/03/17/harvey-to-roll-out-agentic-workflows/> (дата звернення: 19.10.2025).
32. Harvey – Security. *Harvey*. URL: <https://www.harvey.ai> (дата звернення: 19.10.2025).
33. 'The Future of Legal AI Is Collaboration' – Harvey. *Artificial Lawyer*. 2025. URL: <https://www.artificiallawyer.com/2025/11/03/the-future-of-legal-ai-is-collaboration-harvey/> (дата звернення: 19.10.2025).
34. SmartCompose. *Government Technology Agency of Singapore (GovTech Singapore)*. URL: <https://www.tech.gov.sg/products-and-services/for-government-agencies/informational-services/smartcompose/> (дата звернення: 23.10.2025).
35. Solving Real World Problems in the Public Service with AI. *Singapore Civil Service College (CSC) Knowledge Hub*. URL: <https://knowledge.csc.gov.sg/solving-real-world-problems-in-the-public-service-with-ai/> (дата звернення: 23.10.2025).

36. Singapore Government Tech Stack. *Singapore Government Developer Portal*. 2025. URL: <https://www.developer.tech.gov.sg/singapore-government-tech-stack/> (дата звернення: 23.10.2025).
37. Retrieval-Augmented Generation Playbook. *Singapore Government Developer Portal*. 2025. URL: <https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/retrieval-augmented-generation-playbook.html> (дата звернення: 23.10.2025).
38. Redbox. *Incubator for Artificial Intelligence – GOV.UK*. URL: <https://ai.gov.uk/projects/redbox> (дата звернення: 25.10.2025).
39. Incubator for Artificial Intelligence. *GOV.UK*. URL: <https://ai.gov.uk/> (дата звернення: 25.10.2025).
40. DBT: Redbox. *GOV.UK*. URL: <https://www.gov.uk/algorithmic-transparency-records/dbt-redbox> (дата звернення: 25.10.2025).
41. DSIT - Redbox. *GOV.UK*. URL: <https://www.gov.uk/algorithmic-transparency-records/dsit-redbox> (дата звернення: 25.10.2025).
42. Open LLM Leaderboard. *A Hugging Face Space by open-llm-leaderboard*. URL: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard (дата звернення: 05.11.2025).
43. Міністерство соціальної політики України – Запитай у чиновника! *Доступ до правди*. URL: https://dostup.org.ua/body/ministierstvo_sotsialnoyi_politiki_ukrayini (дата звернення: 08.11.2025).
44. Semeniuk H. HlibSemeniuk/minsoc-response-assistant. Official repository for Master's Thesis: AI Assistant for processing public information requests (Ministry of Social Policy of Ukraine). Contains QLoRA fine-tuning code for Gemma 2, Llama 3, and Qwen 2.5, datasets, and evaluation logs. *GitHub*. 2025. URL: <https://github.com/HlibSemeniuk/minsoc-response-assistant> (дата звернення: 08.12.2025).

ДОДАТОК А

Приклади тестових сценаріїв

№	user_query	context_data
Формальні запити		
1	<p>Доброго дня!</p> <p>На підставі статей 1, 13, 19, 20 Закону України «Про доступ до публічної інформації» від 13 січня 2011 року, які надають право звертатись із запитами до розпорядників інформації щодо надання публічної інформації, прошу надати наступну інформацію (наступні документи):</p> <p>Чи зараховується до вислуги років для призначення пенсії військовослужбовцям, особам середнього, старшого та вищого начальницького складу податкової міліції, Бюро економічної безпеки України, період навчання в цивільних навчальних закладах?</p> <p>З повагою, Андрій Васюков</p>	<pre>{ "context_items": [{ "id": "manual_1", "source_type": "manual", "title": "Реєстраційні дані", "snippet": "Запит Андрія Васюкова надійшов 01.11.2025. Виконавець: Департамент пенсійного забезпечення та соціального страхування Міністерства соціальної політики, сім'ї та єдності України.", "confidence": 1.0 }, { "id": "manual_2", "source_type": "manual", "title": "Закон України «Про доступ до публічної інформації», ст. 1", "snippet": "Публічна інформація — це відображена та задокументована інформація, що перебуває у володінні суб'єктів владних повноважень або створена ними в процесі виконання своїх обов'язків.", "confidence": 1.0 }, { "id": "manual_3", "source_type": "manual", "title": "Постанова Пленуму ВАСУ № 10 від 29.09.2016", "snippet": "Для визначення публічної інформації ключовим є факт її попередньої фіксації у будь-якій формі та наявність у володінні суб'єкта владних повноважень.", "confidence": 1.0 }, { "id": "manual_4", "source_type": "manual", "title": "Положення про Мінсоцполітики (Постанова КМУ № 423 від 17.06.2015)", "snippet": "Мінсоцполітики діє в межах компетенції, визначеної Положенням; питання формування та зберігання інформації про зарахування періоду навчання до вислуги років не належить до повноважень Міністерства.", "confidence": 1.0 }] }</pre>

№	user_query	context_data
1		<pre> "id": "manual_5", "source_type": "manual", "title": "Закон України № 2262-XII, ст. 17", "snippet": "До вислуги років осіб, які мають спеціальні звання, додатково може зараховуватися час навчання у закладах освіти (крім військових та спеціалізованих), з розрахунку один рік навчання за шість місяців служби, але не більше п'яти років.", "confidence": 1.0 }, { "id": "manual_6", "source_type": "manual", "title": "Закон України «Про доступ до публічної інформації», ст. 23", "snippet": "Рішення або бездіяльність розпорядника інформації можуть бути оскаржені до керівника, вищого органу або до суду.", "confidence": 1.0 }] } </pre>
4	<p>Доброго дня! Проху надати: 1) Наявні нормативні акти, які змінювали умови нарахування пенсій у зв'язку з воєнним станом (перелік та тексти статей). 2) Чи передбачено у 2025 році тимчасові заходи для перерахунку пенсій певних категорій (вказати нормативний документ та суму перерахунку у середньому по державі)? З повагою, Марина Шевчук""</p>	<pre> { "context_items": [{ "id": "reg_3", "source_type": "manual", "title": "Реєстраційні дані", "snippet": "Запит Марини Шевчук надійшов 18.08.2025. Виконавець: Департамент пенсійного забезпечення Міністерства соціальної політики, сім'ї та єдності України.", "confidence": 1.0 }, { "id": "law_3674", "source_type": "manual", "title": "Закон України № 3674-IX (2024) — обчислення страхового стажу", "snippet": "Закон № 3674-IX від 25.04.2024 вніс зміни до порядку обчислення страхового стажу під час дії особливих умов (включно з воєнним станом): передбачено врахування періодів участі у заходах оборони як страховий стаж.", "confidence": 1.0 }, { "id": "cmu_299", "source_type": "manual", "title": "Постанова КМУ № 299 (11.02.2025)", </pre>

№	user_query	context_data
4		<pre> "snippet": "Постанова № 299 затверджує порядок виплати пенсій особам, які тимчасово перебувають за межами України або на тимчасово окупованих територіях; передбачає можливість ідентифікації та перерахунку пенсій при наданні підтверджувальних документів.", "confidence": 1.0 }, { "id": "avg_recalc", "source_type": "manual", "title": "Оцінка впливу перерахунків (внутрішній аналітичний звіт)", "snippet": "За внутрішньою оцінкою Мінсоцполітики: середній перерахунок для категорій, яким проведено перегляд у 2025 р., становив +4,2% від попереднього розміру пенсії; оцінка по країні — збільшення виплат у середньому на 138,5 грн/особу на місяць (станом на 01.07.2025).", "confidence": 1.0 }] } </pre>
Емоційні та конфліктні запити		
14	<p>Чому в мене в Дії зник статус ВПО? Вчора був, сьогодні немає. Це у вас якісь збої чи ви спеціально видаляєте людей з бази, щоб економити бюджет? Я нікуди не переїжджав, нічого не порушував. Розберіться з вашими системами, бо це виглядає дуже підозріло. Чекаю на пояснення, де мої дані. Олександр Калайда</p>	<pre> { "context_items": [{ "id": "manual_1", "source_type": "manual", "title": "Реєстраційні дані", "snippet": "Запит Олександра Калайди. від 07.11.2025. Виконавець: Директорат цифровізації Міністерства соціальної політики, сім'ї та єдності України.", "confidence": 1.0 }, { "id": "manual_2", "source_type": "manual", "title": "Технічний стан реєстрів", "snippet": "Єдина інформаційна база даних про ВПО працює стабільно. Тимчасове зникнення відображення статусу в застосунку 'Дія' може бути пов'язане з технічними регламентними роботами на стороні адміністратора застосунку 'Дія' (Мінцифри) або необхідністю актуалізації даних.", "confidence": 1.0 }, { "id": "manual_3", "source_type": "manual", "title": "Закон України 'Про доступ до публічної інформації', ст. 22", </pre>

№	user_query	context_data
14		<pre> "snippet": "Мінсоцполітики не є розпорядником інформації щодо технічного функціонування мобільного застосунку 'Дія'.", "confidence": 1.0 }, { "id": "manual_4", "source_type": "manual", "title": "Рекомендації", "snippet": "Рекомендовано перевірити наявність статусу через веб-портал соціальних послуг або звернутися до управління соцзахисту за місцем проживання для звірки даних.", "confidence": 1.0 }] } </pre>
18	<p>Це моє останнє попередження! Якщо до кінця тижня мені не відновлять виплату малозабезпеченим, яку ви незаконно скасували, я підніму на вуха всю пресу! Я вже домовилася з журналістами центрального телеканалу, вони готові зняти сюжет про ваше свавілля. Я влаштую пікет під міністерством і оголошу голодування! Вирішуйте питання негайно, або про вашу некомпетентність дізнається вся країна! Час пішов.</p> <p>Мар'яна Бойко</p>	<pre> { "context_items": [{ "id": "manual_1", "source_type": "manual", "title": "Реєстраційні дані", "snippet": "Запит Мар'яни Бойко від 15.11.2025. Виконавець: Міністерство соціальної політики, сім'ї та єдності України.", "confidence": 1.0 }, { "id": "manual_2", "source_type": "manual", "title": "Підстави припинення виплат малозабезпеченим (Постанова КМУ № 250)", "snippet": "Виплата державної соціальної допомоги малозабезпеченим сім'ям припиняється, якщо сім'єю приховано або навмисно подано недостовірні дані про її доходи та майновий стан, що вплинули на встановлення права на допомогу.", "confidence": 1.0 }, { "id": "manual_3", "source_type": "manual", "title": "Дані верифікації заявника", "snippet": "Згідно з актом перевірки соціального інспектора від 01.11.2025, у заявниці виявлено незадекларовані доходи від здачі майна в оренду, що перевищують встановлені норми для отримання статусу малозабезпеченої сім'ї. Рішення про припинення виплат є правомірним.", "confidence": 1.0 } } </pre>

№	user_query	context_data
18		<pre> "id": "manual_4", "source_type": "manual", "title": "Порядок оскарження", "snippet": "У разі незгоди з рішенням органу соціального захисту населення, заявник має право оскаржити його у судовому порядку відповідно до чинного законодавства.", "confidence": 1.0 }] } </pre>
Юридичні запити		
21	<p>Доброго дня. Мені виповнюється 60 років у серпні 2025 року. Мій офіційний трудовий стаж за книжкою — 28 років і 4 місяці. Також я навчався в університеті на денному стаціонарі з 1982 по 1987 рік. Чи маю я право вийти на пенсію в 60 років? Якщо ні, то коли зможу? І чи допоможе, якщо я добровільно доплачу внески за пару місяців? Андрій Миколайович</p>	<pre> { "context_items": [{ "id": "law_1058_26", "source_type": "manual", "title": "Закон України № 1058 'Про загальнообов'язкове державне пенсійне страхування', ст. 26", "snippet": "Починаючи з 1 січня 2018 року право на призначення пенсії за віком після досягнення віку 60 років мають особи за наявності страхового стажу: ... з 1 січня 2024 року по 31 грудня 2024 року — не менше 31 року; з 1 січня 2025 року по 31 грудня 2025 року — не менше 32 років. У разі відсутності, починаючи з 1 січня 2018 року, необхідного страхового стажу на дату досягнення віку 60 років, право на пенсію мають особи після досягнення віку 63 роки за наявності стажу (у 2025 році) від 22 до 32 років.", "confidence": 1.0 }, { "id": "law_1058_24", "source_type": "manual", "title": "Закон України № 1058, ст. 24 (Періоди, що зараховуються до стажу)", "snippet": "Страховий стаж — це період, протягом якого особа підлягає загальнообов'язковому державному пенсійному страхуванню та за який щомісяця сплачені страхові внески. Період навчання у вищих навчальних закладах, професійно-технічних навчальних закладах тощо, якщо воно відбувалося до 1 січня 2004 року, зараховується до страхового стажу на підставі дипломів або інших документів.", "confidence": 1.0 }, { "id": "law_purchase", "source_type": "manual", "title": "Договір про добровільну сплату ЄСВ", "snippet": "Особи, яким не вистачає страхового стажу, можуть укласти </pre>

№	user_query	context_data
21		<p>договір про добровільну участь у системі загальнообов'язкового державного соціального страхування. Сплата може здійснюватися як за попередні періоди (купівля стажу), так і на майбутнє. Один рік стажу коштує: (мінімальний внесок ЄСВ * 2) * 12 місяців.", "confidence": 1.0 }, { "id": "irrelevant_military", "source_type": "manual", "title": "Закон № 2262 (Військові пенсії)", "snippet": "Військовослужбовцям до вислуги років зараховується час навчання у цивільних ВНЗ із розрахунку один рік навчання за шість місяців служби.", "confidence": 0.8 }] }</p>
28	<p>Хочу оформити догляд за своєю матір'ю (інвалідність 1 групи Б, потребує постійного догляду). Я ніде не працюю, живу з нею. Мені сказали, що виплата буде залежати від мого доходу. Але який у мене дохід, якщо я не працюю? 1. Скільки мені будуть платити? Це фіксована сума чи ні? 2. Чи зарахують цей період мені в стаж для пенсії? 3. Чи правда, що якщо у мами пенсія висока, то мені нічого не заплатять?</p> <p>Олена</p>	<pre>{ "context_items": [{ "id": "care_law_1", "source_type": "manual", "title": "Постанова КМУ № 859 (Компенсація фізособам)", "snippet": "Компенсація призначається фізичним особам, які надають соціальні послуги з догляду на непрофесійній основі. Розмір компенсації обчислюється як різниця між прожитковим мінімумом на одну особу в розрахунку на місяць та середньомісячним сукупним доходом фізичної особи, яка надає соціальні послуги.", "confidence": 1.0 }, { "id": "income_calc", "source_type": "manual", "title": "Розрахунок доходу доглядальника", "snippet": "Якщо особа не працює і не має офіційних доходів, для розрахунку її доходу включаються умовні величини (наприклад, 0,5 мінімальної зарплати) або фактичні доходи з реєстрів. Якщо доходу немає зовсім (0 грн), то компенсація може бути максимальною (рівною прожитковому мінімуму).", "confidence": 1.0 }, { "id": "recipient_income", "source_type": "manual", "title": "Дохід особи, за якою доглядають", "snippet": "Увага! Дохід особи, за якою здійснюється догляд (матері), НЕ впливає на розмір компенсації доглядальника згідно з Постановою № 859. Це впливає лише при визначенні права на соціальні послуги</pre>

№	user_query	context_data
28		<p>від терцентру, а не на компенсацію фізособі.", "confidence": 1.0 }, { "id": "stazh_care", "source_type": "manual", "title": "Закон про пенсійне страхування. Стаття 24", "snippet": "Період догляду за особою з інвалідністю І групи зараховується до страхового стажу доглядальника, якщо ця людина отримувала відповідну компенсацію/допомогу в органах соцзахисту (за них держава сплачує ЄСВ).", "confidence": 1.0 }, { "id": "pm_general", "source_type": "manual", "title": "Прожитковий мінімум 2025 (загальний)", "snippet": "Прожитковий мінімум на одну особу в розрахунку на місяць становить 2920 грн.", "confidence": 1.0 }] }</p>

ДОДАТОК Б

Лістинг коду доналаштування Gemma 2 9b

```
### Налаштування середовища
%%capture
import os, re
import torch; v = re.match(r"[0-9]{1,}\.[0-9]{1,}",
str(torch.__version__)).group(0)
xformers = "xformers==" + ("0.0.33.post1" if v=="2.9" else
"0.0.32.post2" if v=="2.8" else "0.0.29.post3")
!pip install --no-deps bitsandbytes accelerate {xformers} peft trl
triton cut_cross_entropy unsloth_zoo
!pip install sentencepiece protobuf "datasets==4.3.0"
"huggingface_hub>=0.34.0" hf_transfer
!pip install --no-deps unsloth
!pip install transformers==4.56.2
!pip install --no-deps trl==0.22.2
### Завантаження моделі та налаштування LoRA
from unsloth import FastLanguageModel
import torch

max_seq_length = 6144
dtype = None
load_in_4bit = True

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name="unsloth/gemma-2-9b-it-bnb-4bit",
    max_seq_length=max_seq_length,
    dtype=dtype,
    load_in_4bit=load_in_4bit,
)
model = FastLanguageModel.get_peft_model(

model,
    r = 16,
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
"gate_proj", "up_proj", "down_proj",],
    lora_alpha = 32,
    lora_dropout = 0,
    bias = "none",
    use_gradient_checkpointing = "unsloth",
    random_state = 3407,
    use_rslora = False,
    loftq_config = None,
)
### Підготовка даних
from unsloth.chat_templates import get_chat_template

tokenizer = get_chat_template(
    tokenizer,
    chat_template = "gemma2",
)
```

```

def formatting_prompts_func(examples):
    convos = examples["messages"]
    texts = [tokenizer.apply_chat_template(convo, tokenize = False,
add_generation_prompt = False).removeprefix('<bos>') for convo in
convos]
    return { "text" : texts, }
from google.colab import drive
from datasets import load_dataset

drive.mount('/content/drive')

file_path = "/content/drive/MyDrive/(197 samples) Inquiry-Response
Corpus of the Ministry of Social Policy, Family and Unity of
Ukraine.jsonl"

dataset = load_dataset("json", data_files = file_path, split =
"train")

dataset = dataset.map(formatting_prompts_func, batched = True,)
print("=== ПРИКЛАД ВІДФОРМАТОВАНОГО ПРОМПТУ ===")
print(dataset[0]["text"])
### Навчання моделі
from trl import SFTConfig, SFTTrainer
from unsloth import is_bfloat16_supported

trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    train_dataset = dataset,
    dataset_text_field = "text",
    max_seq_length = max_seq_length,
    packing = True,
    args = SFTConfig(
        per_device_train_batch_size = 1,
        gradient_accumulation_steps = 16,
        warmup_steps = 2,
        num_train_epochs = 1,
        learning_rate = 2e-4,
        fp16 = not is_bfloat16_supported(),
        bf16 = is_bfloat16_supported(),
        logging_steps = 1,
        optim = "adamw_8bit",
        weight_decay = 0.001,
        lr_scheduler_type = "linear",
        neftune_noise_alpha = 5,
        seed = 3407,
        output_dir = "/content/drive/MyDrive/[1] Master's
Diploma/Models/[1]Gemma 2 (9b)/[Checkpoints]Gemma 2 (9b)",
        report_to = "none",
        save_strategy = "steps",
        save_steps = 3,
        save_total_limit = 4,

```

```

    ),
)
from unsloth.chat_templates import train_on_responses_only

trainer = train_on_responses_only(
    trainer,
    instruction_part = "<start_of_turn>user\n",
    response_part = "<start_of_turn>model\n",
)
tokenizer.decode(trainer.train_dataset[0]["input_ids"])
tokenizer.decode([tokenizer.pad_token_id if x == -100 else x for x in
trainer.train_dataset[0]["labels"]]).replace(tokenizer.pad_token, " ")
# @title Моніторинг системних ресурсів перед навчанням
import psutil

gpu_stats = torch.cuda.get_device_properties(0)
start_gpu_memory = round(torch.cuda.max_memory_reserved() / 1024 /
1024 / 1024, 3)
max_memory = round(gpu_stats.total_memory / 1024 / 1024 / 1024, 3)
print(f"GPU = {gpu_stats.name}. Max memory = {max_memory} GB.")
print(f"{start_gpu_memory} GB of memory reserved.")

vm = psutil.virtual_memory()
total_ram = round(vm.total / 1024**3, 2)
used_ram = round(vm.used / 1024**3, 2)
available_ram = round(vm.available / 1024**3, 2)

print(f"\nSystem RAM Total:      {total_ram} GB")
print(f"System RAM Used:          {used_ram} GB")
print(f"System RAM Available: {available_ram} GB")
trainer_stats = trainer.train()
# @title Аналіз витрат обчислювальних ресурсів під час навчання

used_memory = round(torch.cuda.max_memory_reserved() / 1024 / 1024 /
1024, 3)
used_memory_for_lora = round(used_memory - start_gpu_memory, 3)
used_percentage = round(used_memory / max_memory * 100, 3)
lora_percentage = round(used_memory_for_lora / max_memory * 100, 3)

vm = psutil.virtual_memory()
total_ram = round(vm.total / 1024**3, 2)
used_ram = round(vm.used / 1024**3, 2)
ram_percentage = round(vm.percent, 2)

print(f"{trainer_stats.metrics['train_runtime']} seconds used for
training.")

print(f"{round(trainer_stats.metrics['train_runtime']/60, 2)} minutes
used for training.")

```

```
print(f"\n==== GPU Memory Stats =====")
print(f"Peak reserved memory           = {used_memory} GB.")
print(f"Peak reserved for training      = {used_memory_for_lora} GB.")
print(f"Peak reserved % of max            = {used_percentage} %.")
print(f"Peak training % of max             = {lora_percentage} %.")

print(f"\n==== System RAM Stats =====")
print(f"Total System RAM                    = {total_ram} GB")
print(f"Current RAM Used                     = {used_ram} GB")
print(f"RAM Used %                           = {ram_percentage} %")
```

ДОДАТОК В

Лістинг коду доналаштування Qwen 2.5 7B

```
### Налаштування середовища
%%capture
import os, re
import torch; v = re.match(r"[0-9]{1,}\.[0-9]{1,}",
str(torch.__version__)).group(0)
xformers = "xformers==" + ("0.0.33.post1" if v=="2.9" else
"0.0.32.post2" if v=="2.8" else "0.0.29.post3")
!pip install --no-deps bitsandbytes accelerate {xformers} peft trl
triton cut_cross_entropy unsloth_zoo
!pip install sentencepiece protobuf "datasets==4.3.0"
"huggingface_hub>=0.34.0" hf_transfer
!pip install --no-deps unsloth
!pip install transformers==4.56.2
!pip install --no-deps trl==0.22.2
### Завантаження моделі та налаштування LoRA
from unsloth import FastLanguageModel
import torch

max_seq_length = 8192
dtype = None
load_in_4bit = True

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name="unsloth/Qwen2.5-7B-Instruct-bnb-4bit",
    max_seq_length=max_seq_length,
    dtype=dtype,
    load_in_4bit=load_in_4bit,
)
model = FastLanguageModel.get_peft_model(
    model,
    r = 16,
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
"gate_proj", "up_proj", "down_proj",],
    lora_alpha = 32,
    lora_dropout = 0,
    bias = "none",
    use_gradient_checkpointing = "unsloth",
    random_state = 3407,
    use_rslora = False,
    loftq_config = None,
)
### Підготовка даних
from unsloth.chat_templates import get_chat_template

tokenizer = get_chat_template(
    tokenizer,
    chat_template = "qwen-2.5",
)
```

```

def formatting_prompts_func(examples):
    convos = examples["messages"]
    texts = [tokenizer.apply_chat_template(convo, tokenize = False,
add_generation_prompt = False) for convo in convos]
    return { "text" : texts, }
from google.colab import drive
from datasets import load_dataset

drive.mount('/content/drive')
file_path = "/content/drive/MyDrive/(197 samples) Inquiry-Response
Corpus of the Ministry of Social Policy, Family and Unity of
Ukraine.jsonl"

dataset = load_dataset("json", data_files = file_path, split =
"train")

dataset = dataset.map(formatting_prompts_func, batched = True,)
print("=== ПРИКЛАД ВІДФОРМАТОВАНОГО ПРОМПТУ ===")
print(dataset[0]["text"])
### Навчання моделі
from trl import SFTConfig, SFTTrainer
from unsloth import is_bfloat16_supported

trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    train_dataset = dataset,
    dataset_text_field = "text",
    max_seq_length = max_seq_length,
    packing = True,
    args = SFTConfig(
        per_device_train_batch_size = 1,
        gradient_accumulation_steps = 16,
        warmup_steps = 2,
        num_train_epochs = 1,
        learning_rate = 2e-4,
        fp16 = not is_bfloat16_supported(),
        bf16 = is_bfloat16_supported(),
        logging_steps = 1,
        optim = "adamw_8bit",
        weight_decay = 0.001,
        lr_scheduler_type = "linear",
        neftune_noise_alpha = 5,
        seed = 3407,
        output_dir = "/content/drive/MyDrive/[1] Master's
Diploma/Models/[2]Qwen 2.5 (7b)/[Checkpoints]Qwen 2.5 (7b)",
        report_to = "none",
        save_strategy = "steps",
        save_steps = 3,
        save_total_limit = 4,
    ),
)

```

```

from unsloth.chat_templates import train_on_responses_only

trainer = train_on_responses_only(
    trainer,
    instruction_part = "<|im_start|>user\n",
    response_part = "<|im_start|>assistant\n",
)
tokenizer.decode(trainer.train_dataset[0]["input_ids"])
tokenizer.decode([tokenizer.pad_token_id if x == -100 else x for x in
trainer.train_dataset[0]["labels"]]).replace(tokenizer.pad_token, " ")
# @title Моніторинг системних ресурсів перед навчанням
import psutil

gpu_stats = torch.cuda.get_device_properties(0)
start_gpu_memory = round(torch.cuda.max_memory_reserved() / 1024 /
1024 / 1024, 3)
max_memory = round(gpu_stats.total_memory / 1024 / 1024 / 1024, 3)
print(f"GPU = {gpu_stats.name}. Max memory = {max_memory} GB.")
print(f"{start_gpu_memory} GB of memory reserved.")

vm = psutil.virtual_memory()
total_ram = round(vm.total / 1024**3, 2)
used_ram = round(vm.used / 1024**3, 2)
available_ram = round(vm.available / 1024**3, 2)

print(f"\nSystem RAM Total:      {total_ram} GB")
print(f"System RAM Used:          {used_ram} GB")
print(f"System RAM Available: {available_ram} GB")
trainer_stats = trainer.train()
# @title Аналіз витрат обчислювальних ресурсів під час навчання

used_memory = round(torch.cuda.max_memory_reserved() / 1024 / 1024 /
1024, 3)
used_memory_for_lora = round(used_memory - start_gpu_memory, 3)
used_percentage = round(used_memory / max_memory * 100, 3)
lora_percentage = round(used_memory_for_lora / max_memory * 100, 3)

vm = psutil.virtual_memory()
total_ram = round(vm.total / 1024**3, 2)
used_ram = round(vm.used / 1024**3, 2)
ram_percentage = round(vm.percent, 2)

print(f"{trainer_stats.metrics['train_runtime']} seconds used for
training.")
print(f"{round(trainer_stats.metrics['train_runtime']/60, 2)} minutes
used for training.")

print(f"\n==== GPU Memory Stats =====")
print(f"Peak reserved memory          = {used_memory} GB.")
print(f"Peak reserved for training     = {used_memory_for_lora} GB.")

print(f"Peak reserved % of max        = {used_percentage} %.")
print(f"Peak training % of max         = {lora_percentage} %.")

```

```
print(f"\n==== System RAM Stats ====")
print(f"Total System RAM           = {total_ram} GB")
print(f"Current RAM Used              = {used_ram} GB")
print(f"RAM Used %                      = {ram_percentage} %")
```

ДОДАТОК Д

Лістинг коду доналаштування Ллама 3 8B

```
### Налаштування середовища
%%capture
import os, re
import torch; v = re.match(r"[0-9]{1,}\.[0-9]{1,}",
str(torch.__version__)).group(0)
xformers = "xformers==" + ("0.0.33.post1" if v=="2.9" else
"0.0.32.post2" if v=="2.8" else "0.0.29.post3")
!pip install --no-deps bitsandbytes accelerate {xformers} peft trl
triton cut_cross_entropy unsloth_zoo
!pip install sentencepiece protobuf "datasets==4.3.0"
"huggingface_hub>=0.34.0" hf_transfer
!pip install --no-deps unsloth
!pip install transformers==4.56.2
!pip install --no-deps trl==0.22.2
### Завантаження моделі та налаштування LoRA
from unsloth import FastLanguageModel
import torch

max_seq_length = 5120
dtype = None
load_in_4bit = True

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name="unsloth/llama-3-8b-Instruct-bnb-4bit",
    max_seq_length=max_seq_length,
    dtype=dtype,
    load_in_4bit=load_in_4bit,
)
model = FastLanguageModel.get_peft_model(
    model,
    r = 16,
    target_modules = ["q_proj", "k_proj", "v_proj", "o_proj",
"gate_proj", "up_proj", "down_proj",],
    lora_alpha = 32,
    lora_dropout = 0,
    bias = "none",
    use_gradient_checkpointing = "unsloth",
    random_state = 3407,
    use_rslora = False,
    loftq_config = None,
)
### Підготовка даних
from unsloth.chat_templates import get_chat_template

tokenizer = get_chat_template(
    tokenizer,
    chat_template = "llama-3",
)
```

```

def formatting_prompts_func(examples):
    convos = examples["messages"]
    texts = [tokenizer.apply_chat_template(convo, tokenize = False,
add_generation_prompt = False).removeprefix('<|begin_of_text|>') for
convo in convos]
    return { "text" : texts, }
from google.colab import drive
from datasets import load_dataset
drive.mount('/content/drive')

file_path = "/content/drive/MyDrive/train_dataset_197_ministry.jsonl"

dataset = load_dataset("json", data_files = file_path, split =
"train")

dataset = dataset.map(formatting_prompts_func, batched = True,)
print("=== ПРИКЛАД ВІДФОРМАТОВАНОГО ПРОМПТУ ===")
print(dataset[0]["text"])
### Доналаштування моделі
from trl import SFTConfig, SFTTrainer
from unsloth import is_bfloat16_supported

trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    train_dataset = dataset,
    dataset_text_field = "text",
    max_seq_length = max_seq_length,
    packing = True,
    args = SFTConfig(
        per_device_train_batch_size = 1,
        gradient_accumulation_steps = 16,
        warmup_steps = 2,
        num_train_epochs = 1,
        learning_rate = 2e-4,
        fp16 = not is_bfloat16_supported(),
        bf16 = is_bfloat16_supported(),
        logging_steps = 1,
        optim = "adamw_8bit",
        weight_decay = 0.001,
        lr_scheduler_type = "linear",
        neftune_noise_alpha = 5,
        seed = 3407,
        output_dir = "/content/drive/MyDrive/[1] Master's
Diploma/Models/[3]Llama 3 (8b Instruct)/[Checkpoints]Llama 3 (8b
Instruct)",
        report_to = "none",
        save_strategy = "steps",
        save_steps = 3,
        save_total_limit = 4,
    ),
),

```

```

)
from unsloth.chat_templates import train_on_responses_only

trainer = train_on_responses_only(
    trainer,
    instruction_part = "<|start_header_id|>user<|end_header_id|>\n\n",
    response_part =
"<|start_header_id|>assistant<|end_header_id|>\n\n",
)
tokenizer.decode(trainer.train_dataset[0]["input_ids"])
tokenizer.decode([tokenizer.pad_token_id if x == -100 else x for x in
trainer.train_dataset[0]["labels"]]).replace(tokenizer.pad_token, " ")
# @title Моніторинг системних ресурсів перед навчанням
import psutil

gpu_stats = torch.cuda.get_device_properties(0)
start_gpu_memory = round(torch.cuda.max_memory_reserved() / 1024 /
1024 / 1024, 3)
max_memory = round(gpu_stats.total_memory / 1024 / 1024 / 1024, 3)
print(f"GPU = {gpu_stats.name}. Max memory = {max_memory} GB.")
print(f"{start_gpu_memory} GB of memory reserved.")

vm = psutil.virtual_memory()
total_ram = round(vm.total / 1024**3, 2)
used_ram = round(vm.used / 1024**3, 2)
available_ram = round(vm.available / 1024**3, 2)

print(f"\nSystem RAM Total:      {total_ram} GB")
print(f"System RAM Used:          {used_ram} GB")
print(f"System RAM Available: {available_ram} GB")
trainer_stats = trainer.train()
# @title Аналіз витрат обчислювальних ресурсів під час навчання

used_memory = round(torch.cuda.max_memory_reserved() / 1024 / 1024 /
1024, 3)
used_memory_for_lora = round(used_memory - start_gpu_memory, 3)
used_percentage = round(used_memory / max_memory * 100, 3)
lora_percentage = round(used_memory_for_lora / max_memory * 100, 3)

vm = psutil.virtual_memory()
total_ram = round(vm.total / 1024**3, 2)
used_ram = round(vm.used / 1024**3, 2)
ram_percentage = round(vm.percent, 2)

print(f"{trainer_stats.metrics['train_runtime']} seconds used for
training.")
print(f"{round(trainer_stats.metrics['train_runtime']/60, 2)} minutes
used for training.")
print(f"\n==== GPU Memory Stats =====")
print(f"Peak reserved memory      = {used_memory} GB.")
print(f"Peak reserved for training  = {used_memory_for_lora} GB.")

```

```
print(f"Peak reserved % of max      = {used_percentage} %.")
print(f"Peak training % of max     = {lora_percentage} %.")
```

ДОДАТОК Е

Зразки відповідей, згенерованих моделлю Gemma 2 9B

№	Відповідь
1	<p>Шановний пане Андрію! Департаментом пенсійного забезпечення та соціального страхування Міністерства соціальної політики, сім'ї та єдності України розглянуто Ваш запит на отримання публічної інформації від 01.11.2025 стосовно зарахування до вислуги років для призначення пенсії військовослужбовцям, особам середнього, старшого та вищого начальницького складу податкової міліції, Бюро економічної безпеки України, період навчання в цивільних навчальних закладах та повідомляємо. Згідно зі статтею 1 Закону України „Про доступ до публічної інформації” від 13.01.2011 № 2947-VI (зі змінами) публічна інформація – це відображена та задокументована будь-якими засобами та на будь-яких носіях інформація, яка була одержана або створена в процесі виконання суб'єктами владних повноважень своїх обов'язків, передбачених чинним законодавством, або яка знаходиться у володінні суб'єктів владних повноважень, інших розпорядників публічної інформації, визначених цим Законом. Водночас, відповідно до постанови Пленуму Верховного Суду України від 29.09.2016 № 10 „Деякі питання застосування Закону України „Про доступ до публічної інформації””, для визначення публічної інформації ключовим є факт її попередньої фіксації у будь-якій формі та наявність у володінні суб'єкта владних повноважень. Разом з тим, Порядком організації роботи Міністерства соціальної політики, сім'ї та єдності України, затвердженим постановою Кабінету Міністрів України від 17.06.2015 № 423 (зі змінами), визначено структуру та функціональні повноваження центрального органу виконавчої влади, який здійснює управління у сфері соціальної політики, сім'ї та єдності. Зокрема, питання формування та зберігання інформації про зарахування періоду навчання до вислуги років не належить до повноважень Міністерства. Враховуючи викладене, повідомляємо, що відповідна інформація не знаходиться у володінні Міністерства. Додатково вказуємо, що відповідно до статті 17 Закону України „Про військову службу, статус військовослужбовців та про особливості призову на військову службу” від 01.07.1993 № 2262-XII (зі змінами) до вислуги років осіб, які мають спеціальні звання, додатково може зараховуватися час навчання у закладах освіти (крім військових та спеціалізованих), з розрахунку один рік навчання за шість місяців служби, але не більше п'яти років. Рішення або бездіяльність розпорядника інформації можуть бути оскаржені до керівника, вищого органу або до суду відповідно до статті 23 Закону України „Про доступ до публічної інформації”. Слід пам'ятати, що цей лист не є остаточним рішенням, яке можна оскаржити до суду.</p>
4	<p>Шановна пані Маринко! Департаментом пенсійного забезпечення Міністерства соціальної політики, сім'ї та єдності України розглянуто Ваш запит на отримання публічної інформації від 18.08.2025 щодо наявності нормативних актів, які змінюють умови нарахування пенсій у зв'язку з воєнним станом, та чи передбачено у 2025 році тимчасові заходи для перерахунку пенсій певних категорій, та повідомляємо. Згідно із Законом України „Про внесення змін до деяких законодавчих актів України щодо підвищення рівня соціального захисту населення” від 25.04.2024 № 3674-IX (зі змінами) порядок обчислення страхового стажу під час дії особливих умов, зокрема воєнного стану, визначено таким чином, що до страхового стажу зараховуються періоди участі особи у заходах оборони, проведення мобілізації, виконання військової служби, служби в інших військових формуваннях, служба в органах внутрішніх справ, Службі безпеки України, Державній прикордонній службі України, військовій службі в резервах, а також інші періоди, визначені Кабінетом Міністрів України. Водночас постановою Кабінету Міністрів України від 11.02.2025 № 299 „Деякі питання виплати пенсій особам, які тимчасово перебувають за межами України або на тимчасово окупованих територіях” затверджено Порядок виплати пенсій особам, які</p>

№	Відповідь
1	<p>тимчасово перебувають за межами України або на тимчасово окупованих територіях, який передбачає можливість ідентифікації та перерахунку пенсій при наданні відповідних підтверджувальних документів. Враховуючи зазначене, Мінсоцполітики проводило внутрішню оцінку впливу перерахунків пенсій певних категорій, які були проведені у 2025 році. Середній перерахунок для категорій, яким проведено перегляд у 2025 році, становив +4,2% від попереднього розміру пенсії. При цьому загальне збільшення виплат пенсій у середньому на місяць становило 138,5 грн на людину (станом на 01.07.2025). Додатково повідомляємо, що Мінсоцполітики спільно з Пенсійним фондом України здійснює постійний моніторинг ситуації з пенсійними видами допомоги та вживає всіх необхідних заходів для забезпечення належного рівня соціального захисту населення. Просимо Вас ознайомитись з наданою інформацією.</p>
14	<p>Директорат цифровізації Міністерства соціальної політики, сім'ї та єдності України в межах компетенції розглянув Ваш запит від 07.11.2025 стосовно тимчасового зникнення відображення статусу внутрішньо переміщеної особи в застосунку „Дія” та повідомляє. Єдина інформаційна база даних про внутрішньо переміщених осіб працює стабільно. Тимчасове зникнення відображення статусу в застосунку „Дія” може бути пов’язане з технічними регламентними роботами на стороні адміністратора застосунку „Дія” (Міністерство цифрової трансформації) або необхідністю актуалізації даних. Мінсоцполітики не є розпорядником інформації щодо технічного функціонування мобільного застосунку „Дія”. Додатково рекомендуємо перевірити наявність статусу через веб-портал соціальних послуг або звернутися до управління соціального захисту населення за місцем проживання для звірки даних. Згідно зі статтею 22 Закону України „Про доступ до публічної інформації” запит на отримання публічної інформації може бути направлений до будь-якого органу виконавчої влади, його територіальних органів, інших розпорядників публічної інформації, визначених цим Законом. Водночас необхідно враховувати, що органи виконавчої влади, інші розпорядники публічної інформації зобов’язані забезпечити доступ до публічної інформації, яка знаходиться у їх володінні, контролі або під їх захистом, незалежно від форми її зберігання, незалежно від того, чи вона була одержана ними шляхом придбання, передачі, створення, отримання від інших розпорядників публічної інформації або будь-яким іншим способом. При цьому органи виконавчої влади, інші розпорядники публічної інформації мають право відмовити у задоволенні запиту на отримання публічної інформації лише у випадках, передбачених чинним законодавством.</p>
18	<p>Шановна пані Мар’яно! Міністерство соціальної політики, сім'ї та єдності України в межах компетенції розглянуло Ваш запит від 15.11.2025 щодо припинення виплати державної соціальної допомоги малозабезпеченим сім'ям та повідомляє. Згідно з постановою Кабінету Міністрів України від 22.06.2005 № 250 „Деякі питання виплати державної соціальної допомоги малозабезпеченим сім'ям” (зі змінами) виплата державної соціальної допомоги малозабезпеченим сім'ям припиняється, зокрема, у разі, коли сім'єю приховано або навмисно подано недостовірні дані про її доходи та майновий стан, що вплинули на встановлення права на допомогу. Водночас Порядком надання державної соціальної допомоги малозабезпеченим сім'ям, затвердженим постановою Кабінету Міністрів України від 22.06.2005 № 250 (зі змінами), визначено, що соціальний інспектор здійснює перевірку фактичного матеріального забезпечення сім'ї, яка отримує державну соціальну допомогу малозабезпеченим сім'ям, шляхом проведення візиту до місця проживання сім'ї, огляду житла, перевірки документів, які підтверджують доходи та майновий стан сім'ї, інших видів діяльності, спрямованих на з'ясування фактичного матеріального забезпечення сім'ї, та складає акт перевірки. На виконання вимог статті 10 Закону України „Про соціальний</p>

№	Відповідь
18	<p>захист населення” соціальний інспектор Мінсоцполітики проведе акти перевірки фактичного матеріального забезпечення сім’ї, яка отримує державну соціальну допомогу малозабезпеченим сім’ям, та приймає рішення про припинення виплати державної соціальної допомоги малозабезпеченим сім’ям, якщо виявлено, що сім’я не відповідає критеріям малозабезпеченості. Враховуючи викладене, повідомляємо, що згідно з актом перевірки соціального інспектора від 01.11.2025 у Вашій сім’ї виявлено незадекларовані доходи від здачі майна в оренду, що перевищують встановлені норми для отримання статусу малозабезпеченої сім’ї. Рішення про припинення виплати державної соціальної допомоги малозабезпеченим сім’ям є правомірним. У разі незгоди з рішенням органу соціального захисту населення, заявник має право оскаржити його у судовому порядку відповідно до чинного законодавства.</p>
21	<p>Шановний Андрію Миколайовичу! Міністерство соціальної політики України в межах компетенції розглянуло Ваш запит на отримання публічної інформації щодо можливості виходу на пенсію у віці 60 років та повідомляє. Згідно зі статтею 26 Закону України „Про загальнообов’язкове державне пенсійне страхування” (далі – Закон № 1058) право на призначення пенсії за віком після досягнення віку 60 років мають особи, які на момент досягнення такого віку мають страховий стаж: з 1 січня 2024 року по 31 грудня 2024 року – не менше 31 року; з 1 січня 2025 року по 31 грудня 2025 року – не менше 32 років. При цьому у разі відсутності, починаючи з 1 січня 2018 року, необхідного страхового стажу на дату досягнення віку 60 років, право на пенсію мають особи після досягнення віку 63 роки за наявності стажу (у 2025 році) від 22 до 32 років. Страховий стаж – це період, протягом якого особа підлягає загальнообов’язковому державному пенсійному страхуванню та за який щомісяця сплачені страхові внески. До страхового стажу зараховуються періоди, визначені чинним законодавством, зокрема, період навчання у вищих навчальних закладах, професійно-технічних навчальних закладах тощо, якщо воно відбувалося до 1 січня 2004 року, на підставі дипломів або інших документів, що засвідчують освіта. Водночас особи, яким не вистачає страхового стажу, можуть укласти договір про добровільну участь у системі загальнообов’язкового державного соціального страхування. Сплата може здійснюватися як за попередні періоди (купівля стажу), так і на майбутнє. Вартість придбання одного року стажу дорівнює сумі мінімального розміру страхових внесків за один місяць, помноженій на два, помножену на кількість місяців у році. Разом з тим, звертаємо увагу, що військовослужбовцям до вилучення років зараховується час навчання у цивільних вищих навчальних закладах із розрахунку один рік навчання за шість місяців служби. Отже, враховуючи наведене, рекомендуємо Вам детально ознайомитися з нормативно-правовою базою та подальші дії узгоджувати з фахівцями Пенсійного фонду України.</p>
28	<p>Шановна пані Олено! Мінсоцполітики розглянув Ваш запит щодо надання публічної інформації стосовно можливості отримання компенсації за догляд за рідними та повідомляє. Згідно з постановою Кабінету Міністрів України від 29.07.2015 № 859 „Деякі питання надання компенсації фізичним особам, які надають соціальні послуги з догляду на непрофесійній основі” (зі змінами) компенсація призначається фізичним особам, які надають соціальні послуги з догляду на непрофесійній основі. Розмір компенсації обчислюється як різниця між прожитковим мінімумом на одну особу в розрахунку на місяць та середньомісячним сукупним доходом фізичної особи, яка надає соціальні послуги. Прожитковий мінімум на одну особу в розрахунку на місяць становить 2920 гривень. При цьому дохід особи, за якою здійснюється догляд, не враховується при визначенні розміру компенсації. Враховуються всі види доходів фізичної особи, яка надає</p>

№	Відповідь
28	<p>соціальні послуги з догляду на непрофесійній основі, незалежно від джерела їх надходження, у тому числі грошові та матеріальні блага, отримані від будь-яких джерел, у тому числі від підприємств, установ, організацій, інших фізичних осіб, а також від продажу майна, придбання якого було оплачено за рахунок бюджетних коштів. Якщо особа не працює і не має офіційних доходів, для розрахунку її доходу включаються умовні величини (наприклад, 0,5 мінімальної зарплати) або фактичні доходи з реєстрів. Якщо доходу немає зовсім (0 грн), то компенсація може бути максимальною (рівною прожитковому мінімуму). Разом з тим, повідомляємо, що відповідно до статті 24 Закону України „Про загальнообов’язкове державне пенсійне страхування” період догляду за особою з інвалідністю I групи зараховується до страхового стажу особи, яка доглядала за цією особою, якщо вона отримувала відповідну допомогу з органів соціального захисту населення, за яку держава сплачувала єдиний внесок на загальнообов’язкове державне пенсійне страхування. Водночас, повідомляємо, що відповідно до пункту 1 Порядку надання соціальних послуг населенню, затвердженого постановою Кабінету Міністрів України від 22.07.2015 № 682 (зі змінами), особа, яка надає соціальні послуги з догляду на непрофесійній основі, має право на безоплатні соціальні послуги, що надаються центрами соціальних служб у разі, коли така особа відповідає критеріям визначення осіб, які мають потребу в наданні соціальних послуг, встановленим чинним законодавством. Особа, яка надає соціальні послуги з догляду на непрофесійній основі, має право на безоплатні соціальні послуги, що надаються центрами соціальних служб у разі, коли така особа відповідає критеріям визначення осіб, які мають потребу в наданні соціальних послуг, встановленим чинним законодавством. Надаємо Вам консультацію з питань надання соціальних послуг населенню, затвердженої наказом Міністерства соціальної політики України від 10.03.2022 № 291 (зі змінами). Дотримуйтеся карантинних норм безпеки. Бережіть себе та своїх близьких!</p>

ДОДАТОК Ж

Зразки відповідей, згенерованих моделлю Qwen 2.5 7B

№	Відповідь
1	<p>Шановний пане! Департамент пенсійного забезпечення та соціального страхування Міністерства соціальної політики, сім'ї та єдності України затвердило Ваш запит на отримання публічної інформації та повідомляє. Публічна інформація – це відображена та задокументована інформація, що перебуває у володінні суб'єктів владних повноважень або створена ними в процесі виконання своїх обов'язків. Для визначення публічної інформації ключовим є факт її попередньої фіксації у будь-якій формі та наявність у володінні суб'єкта владних повноважень. Відповідно до пункту 1 статті 17 Закону України „Про пенсійне забезпечення” до вислуги років осіб, які мають спеціальні звання, додатково може зараховуватися час навчання у закладах освіти (крім військових та спеціалізованих), з розрахунку один рік навчання за шість місяців служби, але не більше п'яти років. Міністерство соціальної політики, сім'ї та єдності України діє в межах компетенції, визначеної Положенням про Міністерство соціальної політики, сім'ї та єдності України, затвердженим постановою Кабінету Міністрів України від 17 червня 2015 року № 423, із якого випливає, що питання формування та зберігання інформації про зарахування періоду навчання до вислуги років не належить до повноважень Міністерства. Рішення або бездіяльність розпорядника інформації можуть бути оскаржені до керівника, вищого органу або до суду.</p>
4	<p>Шановна пані Марино! Департамент пенсійного забезпечення Міністерства соціальної політики, сім'ї та єдності України виконав Ваш запит на отримання публічної інформації. Відповідно до статті 22 Закону України „Про доступ до публічної інформації” (далі – Закон) інформацію надається безкоштовно. Відповідно до статті 22 Закону інформацію надається у терміні десяти робочих днів зі дня отримання запиту. Відповідно до пункту 1 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом. Відповідно до пункту 2 статті 22 Закону інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 3 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом, інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 4 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом, інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 5 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом, інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 6 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом, інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 7 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом, інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 8 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація передбачена електронним видом, інформацію надається в письмовій формі, якщо запитувана інформація передбачена письмовим видом. Відповідно до пункту 9 статті 22 Закону інформацію надається в електронній формі, якщо запитувана інформація</p>

№	Відповідь
21	203, 204, 205, 206, 207, 208, 209, 210, 211, 212, 213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 248, 249, 250, 251, 252, 253, 254, 255, 256, 257, 258, 259, 260, 261, 262, 263, 264, 265, 266, 267, 268, 269, 270, 271, 272, 273, 274, 275, 276, 277, 278, 279, 280, 281, 282, 283, 284, 285, 286, 287, 288, 289, 290, 291, 292, 293, 294, 295, 296, 297, 298, 299, 300, 301, 302, 303, 304, 305, 306, 307, 308, 309, 310, 311, 312, 313, 314, 315, 316, 317, 318, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 33
28	<p>Шановна пані Олено! Директором Мінсоцполітики затверджене Призначення компенсації фізичним особам, які надають соціальні послуги з догляду на непрофесійній основі, згідно з постановою Кабінету Міністрів України від 15.07.2022 № 859 „Про призначення компенсації фізичним особам, які надають соціальні послуги з догляду на непрофесійній основі” (далі – постанова). Компенсація призначається фізичним особам, які надають соціальні послуги з догляду на непрофесійній основі. Розмір компенсації обчислюється як різниця між прожитковим мінімумом на одну особу в розрахунку на місяць та середньомісячним сукупним доходом фізичної особи, яка надає соціальні послуги. Якщо особа не працює і не має офіційних доходів, для розрахунку її доходу включаються умовні величини (наприклад, 0,5 мінімальної зарплати) або фактичні доходи з реєстрів. Якщо доходу немає зовсім (0 грн), то компенсація може бути максимальною (рівною прожитковому мінімуму). Увага! Дохід особи, за якою здійснюється догляд (матері), НЕ впливає на розмір компенсації доглядальника згідно з постановою № 859. Це впливає лише при визначенні права на соціальні послуги від терцентру, а не на компенсацію фізособі. Період догляду за особою з інвалідністю I групи зараховується до страхового стажу доглядальника, якщо ця людина отримувала відповідну компенсацію/допомогу в органах соцзахисту (за них держава сплачує ЄСВ). Прожитковий мінімум на одну особу в розрахунку на місяць становить 2920 грн.</p>

ДОДАТОК 3

Зразки відповідей, згенерованих моделлю Meta Llama 3 8B

№	Відповідь
1	<p>Шановний пане Андріє! Департаментом пенсійного забезпечення та соціального страхування Міністерства соціальної політики, сім'ї та єдності України розглянуто Ваш запит на отримання публічної інформації від 01.11.2025 та повідомляється. Згідно зі статтею 1 Закону України „Про доступ до публічної інформації” (далі – Закон) публічна інформація – це відображена та задокументована інформація, що перебуває у володінні суб'єктів владних повноважень або створена ними в процесі виконання своїх обов'язків. Для визначення публічної інформації ключовим є факт її попередньої фіксації у будь-якій формі та наявність у володінні суб'єкта владних повноважень. Мінсоцполітики діє в межах компетенції, визначеної Положенням про Міністерство соціальної політики, затвердженим постановою Кабінету Міністрів України від 17.06.2015 № 423 „Про затвердження Положення про Міністерство соціальної політики”. Питання формування та зберігання інформації про зарахування періоду навчання до вислуги років не належить до повноважень Мінсоцполітики. Відповідно до статті 17 Закону України „Про загальнообов'язкове державне соціальне страхування” (далі – Закон) до вислуги років осіб, які мають спеціальні звання, додатково може зараховуватися час навчання у закладах освіти (крім військових та спеціалізованих), з розрахунку один рік навчання за шість місяців служби, але не більше п'яти років. Рішення або бездіяльність розпорядника інформації можуть бути оскаржені до керівника, вищого органу або до суду.</p>
4	<p>Шановна пано! Департаментом пенсійного забезпечення Міністерства соціальної політики, сім'ї та єдності України розглянуто Ваш запит на отримання публічної інформації від 18.08.2025 та повідомляється. Згідно з Законом України „Про доступ до публічної інформації” (далі – Закон) та іншими чинними законодавчими актами, Департамент пенсійного забезпечення Мінсоцполітики надсилає інформацію щодо наявних нормативних актів, які змінювали умови нарахування пенсій у зв'язку з воєнним станом. Відповідно до статті 12 Закону „Про пенсійну систему України” (далі – Закон про Пенсійну систему) та постанови Кабінету Міністрів України від 25.04.2024 № 3674-IX „Щодо внесення змін до деяких нормативних актів у зв'язку з прийняттям Закону України „Про внесення змін до деяких законодавчих актів України щодо вдосконалення системи оплати праці та інших питань” (зі змінами) (далі – Постанова № 3674-IX), враховується період участі у заходах оборони як страховий стаж під час обчислення страхового стажу для призначення та виплати пенсії. Крім того, відповідно до постанови Кабінету Міністрів України від 11.02.2025 № 299 „Щодо затвердження Порядку виплати пенсій особам, які тимчасово перебувають за межами України або на тимчасово окупованих територіях” (далі – Порядок виплати пенсій), передбачено можливість ідентифікації та перерахунку пенсій особи, яка тимчасово перебуває за межами України або на тимчасово окупованих територіях, при наданні підтверджувальних документів. Згідно з внутрішньою оцінкою Мінсоцполітики, середній перерахунок для категорій, яким проведено перегляд у 2025 р., становить +4,2% від попереднього розміру пенсії. Оцінка по країні – збільшення виплат у середньому на 138,5 гривень на особу на місяць (станом на 01.07.2025).</p>
14	<p>Шановний пане Олександрє! Директоратом цифрової інфраструктури Мінсоцполітики розглянуто Ваш запит на отримання публічної інформації від 07.11.2025 та повідомляється. Єдина інформаційна база даних про ветеранів війни за незалежність України, учасників бойових дій та інших категорій громадян, які отримали статус відповідно до закону України „Про ветеранів та їх соціальне забезпечення” (далі – реєстр) працює стабільно. Тимчасове зникнення відображення статусу в застосунку “Дія” може бути пов'язане з технічними регламентними роботами на стороні адміністратора застосунку “Дія” (Мінцифри) або</p>

ДОДАТОК Й

Фрагмент лістингу та конфігурація моделі Gemma 2 9B

```
from unsloth import FastLanguageModel
from unsloth.chat_templates import get_chat_template
import torch

# 1. Ініціалізація та завантаження моделі (4-bit Quantization)
model_path = "/content/drive/MyDrive/[1] Master's
Diploma/Models/[1]Gemma 2 (9b)/[Checkpoints]Gemma 2
(9b)/checkpoint-13"

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = model_path,
    max_seq_length = 6144,
    dtype = None,
    load_in_4bit = True,
)

FastLanguageModel.for_inference(model)

tokenizer = get_chat_template(
    tokenizer,
    chat_template = "gemma2",
)

# 2. Функція генерації (Фрагмент з параметрами інференсу)
def run_inference_benchmark(user_query, context_items, model,
tokenizer):
    # ... (Формування промпта та токенизація) ...

    # ГЕНЕРАЦІЯ
    try:
        outputs = model.generate(
            input_ids = inputs.input_ids,
            attention_mask = inputs.attention_mask,
            max_new_tokens=2048,
            do_sample=True,
            use_cache=True,
            temperature=0.3,
            top_p=0.9,
            repetition_penalty = 1.1,
            pad_token_id=tokenizer.pad_token_id or
tokenizer.eos_token_id,
        )
    except Exception as e:
        ram_monitor.stop()
        return {"error": str(e)}

    # ... (Декодування та метрики) ...
```

ДОДАТОК К

Фрагмент лістингу та конфігурація моделі Qwen 2.5 7B

```
from unsloth import FastLanguageModel
from unsloth.chat_templates import get_chat_template
import torch

# 1. Ініціалізація та завантаження моделі (4-bit Quantization)
model_path = "/content/drive/MyDrive/[1] Master's
Diploma/Models/[2]Qwen 2.5 (7b)/[Checkpoints]Qwen 2.5
(7b)/checkpoint-13"

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = model_path,
    max_seq_length = 8192,
    dtype = None,
    load_in_4bit = True,
)

FastLanguageModel.for_inference(model)

tokenizer = get_chat_template(
    tokenizer,
    chat_template = "qwen-2.5",
)

# 2. Функція генерації (Фрагмент з параметрами інференсу)
def run_inference_benchmark(user_query, context_items, model,
tokenizer):
    # ... (Формування промпта та токенизація) ...

    # ГЕНЕРАЦІЯ
    try:
        outputs = model.generate(
            input_ids = inputs.input_ids,
            attention_mask = inputs.attention_mask,
            max_new_tokens=2048,
            do_sample=True,
            use_cache=True,
            temperature=0.3,
            top_p=0.9,
            repetition_penalty = 1.1,
            pad_token_id=tokenizer.pad_token_id or
tokenizer.eos_token_id,
        )
    except Exception as e:
        ram_monitor.stop()
        return {"error": str(e)}

    # ... (Декодування та метрики) ...
```

ДОДАТОК Л

Фрагмент лістингу та конфігурація моделі Llama 3 8B

```
from unsloth import FastLanguageModel
from unsloth.chat_templates import get_chat_template
import torch

# 1. Ініціалізація та завантаження моделі (4-bit Quantization)
model_path = "/content/drive/MyDrive/[1] Master's
Diploma/Models/[3]Llama 3 (8b Instruct)/[Checkpoints]Llama 3 (8b
Instruct)/checkpoint-13"

model, tokenizer = FastLanguageModel.from_pretrained(
    model_name = model_path,
    max_seq_length = 5120,
    dtype = None,
    load_in_4bit = True,
)

FastLanguageModel.for_inference(model)

tokenizer = get_chat_template(
    tokenizer,
    chat_template = "llama-3",
)

# 2. Функція генерації (Фрагмент з параметрами інференсу)
def run_inference_benchmark(user_query, context_items, model,
tokenizer):
    # ... (Формування промпта та токенизація) ...

    # ГЕНЕРАЦІЯ
    try:
        outputs = model.generate(
            input_ids = inputs.input_ids,
            attention_mask = inputs.attention_mask,
            max_new_tokens=2048,
            do_sample=True,
            use_cache=True,
            temperature=0.3,
            top_p=0.9,
            repetition_penalty = 1.1,
            pad_token_id=tokenizer.pad_token_id or
tokenizer.eos_token_id,
        )
    except Exception as e:
        ram_monitor.stop()
        return {"error": str(e)}

    # ... (Декодування та метрики) ...
```

ДОДАТОК М

Кількісна оцінка ефективності моделей за критеріями якості

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
1	Фактологічна точність та обґрунтованість	2	2	2
	Дотримання інструкцій та формату	3	3	2
	Стиль та тон	4	5	4
	Юридична та контекстна логіка	4	4	4
	Фінальна оцінка тест-кейсу	3	3,1	2,8
2	Фактологічна точність та обґрунтованість	3	3	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	3	1
	Юридична та контекстна логіка	3	3	1
	Фінальна оцінка тест-кейсу	2,9	2,8	1
3	Фактологічна точність та обґрунтованість	2	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	4	1
	Юридична та контекстна логіка	3	3	1
	Фінальна оцінка тест-кейсу	2,6	2,5	1
4	Фактологічна точність та обґрунтованість	3	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	3	3	1
	Юридична та контекстна логіка	4	3	1
	Фінальна оцінка тест-кейсу	3,1	2,4	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
5	Фактологічна точність та обґрунтованість	5	4	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	2	1
	Юридична та контекстна логіка	5	4	1
	Фінальна оцінка тест-кейсу	4,4	3,4	1
6	Фактологічна точність та обґрунтованість	4	4	1
	Дотримання інструкцій та формату	3	2	1
	Стиль та тон	5	5	1
	Юридична та контекстна логіка	5	5	1
	Фінальна оцінка тест-кейсу	4,2	4	1
7	Фактологічна точність та обґрунтованість	4	3	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	3	1
	Юридична та контекстна логіка	4	4	1
	Фінальна оцінка тест-кейсу	3,7	3,1	1
8	Фактологічна точність та обґрунтованість	5	3	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	2	1
	Юридична та контекстна логіка	5	4	1
	Фінальна оцінка тест-кейсу	4,4	3	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
9	Фактологічна точність та обґрунтованість	3	3	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	3	1
	Юридична та контекстна логіка	3	3	1
	Фінальна оцінка тест-кейсу	3	2,8	1
10	Фактологічна точність та обґрунтованість	2	2	1
	Дотримання інструкцій та формату	3	1	1
	Стиль та тон	2	5	1
	Юридична та контекстна логіка	4	2	1
	Фінальна оцінка тест-кейсу	2,8	2,1	1
11	Фактологічна точність та обґрунтованість	3	3	2
	Дотримання інструкцій та формату	2	2	2
	Стиль та тон	3	3	2
	Юридична та контекстна логіка	5	5	3
	Фінальна оцінка тест-кейсу	3,4	3,4	2,3
12	Фактологічна точність та обґрунтованість	4	4	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	3	1
	Юридична та контекстна логіка	5	5	1
	Фінальна оцінка тест-кейсу	4	3,8	1
13	Фактологічна точність та обґрунтованість	3	3	1
	Дотримання інструкцій та формату	2	2	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
13	Стиль та тон	4	4	1
	Юридична та контекстна логіка	5	5	1
	Фінальна оцінка тест-кейсу	3,5	3,5	1
14	Фактологічна точність та обґрунтованість	4	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	5	1
	Юридична та контекстна логіка	4	2	1
	Фінальна оцінка тест-кейсу	3,6	2,3	1
15	Фактологічна точність та обґрунтованість	5	2	4
	Дотримання інструкцій та формату	3	2	2
	Стиль та тон	5	2	2
	Юридична та контекстна логіка	5	4	3
	Фінальна оцінка тест-кейсу	4,6	2,6	3,1
16	Фактологічна точність та обґрунтованість	2	3	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	3	1
	Юридична та контекстна логіка	5	4	1
	Фінальна оцінка тест-кейсу	3,2	3,1	1
17	Фактологічна точність та обґрунтованість	5	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	3	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
17	Юридична та контекстна логіка	5	4	1
	Фінальна оцінка тест-кейсу	4,3	2,7	1
18	Фактологічна точність та обґрунтованість	2	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	3	1
	Юридична та контекстна логіка	4	5	1
	Фінальна оцінка тест-кейсу	2,8	3	1
19	Фактологічна точність та обґрунтованість	1	2	1
	Дотримання інструкцій та формату	1	2	1
	Стиль та тон	1	3	1
	Юридична та контекстна логіка	1	5	1
	Фінальна оцінка тест-кейсу	1	3	1
20	Фактологічна точність та обґрунтованість	2	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	2	1
	Юридична та контекстна логіка	4	5	1
	Фінальна оцінка тест-кейсу	2,8	2,9	1
21	Фактологічна точність та обґрунтованість	3	1	1
	Дотримання інструкцій та формату	2	1	1
	Стиль та тон	5	1	1
	Юридична та контекстна логіка	3	1	1
	Фінальна оцінка тест-кейсу	3	1	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
22	Фактологічна точність та обґрунтованість	1	2	2
	Дотримання інструкцій та формату	1	2	1
	Стиль та тон	1	4	3
	Юридична та контекстна логіка	1	3	2
	Фінальна оцінка тест-кейсу	1	2,5	1,9
23	Фактологічна точність та обґрунтованість	2	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	3	3	1
	Юридична та контекстна логіка	3	4	1
	Фінальна оцінка тест-кейсу	2,4	2,7	1
24	Фактологічна точність та обґрунтованість	4	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	2	1
	Юридична та контекстна логіка	5	5	1
	Фінальна оцінка тест-кейсу	3,9	2,9	1
25	Фактологічна точність та обґрунтованість	2	2	2
	Дотримання інструкцій та формату	2	2	2
	Стиль та тон	5	5	5
	Юридична та контекстна логіка	4	4	4
	Фінальна оцінка тест-кейсу	2,9	2,9	2,9
26	Фактологічна точність та обґрунтованість	2	2	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
26	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	4	3	1
	Юридична та контекстна логіка	2	2	1
	Фінальна оцінка тест-кейсу	2,2	2,1	1
27	Фактологічна точність та обґрунтованість	2	2	1
	Дотримання інструкцій та формату	2	2	1
	Стиль та тон	5	5	1
	Юридична та контекстна логіка	4	4	1
	Фінальна оцінка тест-кейсу	2,9	2,9	1
28	Фактологічна точність та обґрунтованість	2	2	2
	Дотримання інструкцій та формату	2	2	2
	Стиль та тон	3	3	3
	Юридична та контекстна логіка	4	4	4
	Фінальна оцінка тест-кейсу	2,7	2,7	2,7
29	Фактологічна точність та обґрунтованість	3	1	1
	Дотримання інструкцій та формату	2	1	1
	Стиль та тон	5	1	1
	Юридична та контекстна логіка	5	1	1
	Фінальна оцінка тест-кейсу	3,6	1	1
30	Фактологічна точність та обґрунтованість	4	1	1
	Дотримання інструкцій та формату	2	1	1
	Стиль та тон	4	1	1

№	Критерій	Модель		
		Gemma 2 9B	Llama 3 8B	Qwen 2.5 7B
30	Юридична та контекстна логіка	5	1	1
	Фінальна оцінка тест-кейсу	3,9	1	1

ДОДАТОК Н

Детальні результати моніторингу ресурсів під час генерації

№	Gemma 2 9B			Llama 3 8B			Qwen 2.5 7B		
	Lat (с)	Thr (т/с)	VRAM (ГБ)	Lat (с)	Thr (т/с)	VRAM (ГБ)	Lat (с)	Thr (т/с)	VRAM (ГБ)
1	122,13	6,58	8,06	35,28	11,08	6,91	61,44	11,21	6,85
2	70,67	8,60	7,91	30,99	12,74	6,89	166,50	12,30	6,86
3	99,63	8,73	7,93	55,32	12,75	6,90	164,44	12,45	6,90
4	79,74	8,57	7,85	39,91	12,76	6,91	166,68	12,29	6,88
5	68,05	8,63	7,85	29,92	12,60	6,89	164,94	12,42	6,86
6	75,44	8,56	7,87	33,11	12,53	6,91	165,19	12,40	6,88
7	53,18	8,52	7,65	20,09	12,69	6,88	161,18	12,71	6,84
8	69,66	8,69	7,85	37,08	12,54	6,93	167,11	12,26	6,92
9	70,82	8,67	7,87	50,11	12,49	6,98	166,71	12,28	6,94
10	113,36	8,76	7,96	80,64	12,39	7,00	168,97	12,12	6,96
11	49,76	8,52	7,75	23,59	12,59	6,94	45,26	12,71	6,90
12	50,91	8,49	7,71	22,03	12,39	6,86	162,80	12,58	6,82
13	50,42	8,53	7,77	21,61	12,54	6,88	90,60	12,91	6,84
14	57,47	8,72	7,75	23,12	12,84	6,91	162,62	12,59	6,82
15	42,90	8,48	7,77	35,57	12,74	6,88	40,23	12,65	6,86
16	29,14	8,41	7,63	31,60	12,72	6,86	162,02	12,64	6,82
17	30,93	8,34	7,67	26,98	12,64	6,88	163,92	12,49	6,88
18	76,94	8,72	7,71	23,61	12,67	6,88	163,41	12,53	6,88
19	250,96	8,16	7,87	33,60	12,62	6,93	166,00	12,34	6,92
20	74,83	8,81	7,89	33,83	12,59	6,89	164,60	12,44	6,92

№	Gemma 2 9B			Llama 3 8B			Qwen 2.5 7B		
	Lat (c)	Thr (т/с)	VRAM (ГБ)	Lat (c)	Thr (т/с)	VRAM (ГБ)	Lat (c)	Thr (т/с)	VRAM (ГБ)
21	75,72	8,68	7,95	169,59	12,08	6,96	167,82	12,20	6,92
22	236,57	8,66	7,83	37,92	12,60	7,05	98,68	12,69	6,90
23	93,44	8,71	8,04	37,53	12,39	6,95	168,04	12,19	6,90
24	108,84	8,67	8,02	44,11	12,38	7,00	169,88	12,06	6,96
25	104,56	8,79	7,93	48,42	12,58	6,98	62,97	11,67	6,81
26	101,63	8,73	7,91	50,93	12,45	6,98	169,03	12,12	6,90
27	120,93	8,78	7,98	50,07	12,48	6,98	169,11	12,11	6,94
28	102,52	8,76	7,93	46,14	12,68	6,94	58,24	12,41	6,92
29	67,99	8,80	7,83	168,15	12,18	6,94	166,22	12,32	6,84
30	52,06	8,66	7,77	167,64	12,22	7,04	165,70	12,36	6,90
AVG	86,71	8,56	7,85	50,28	12,50	6,93	142,34	12,35	6,89