УДК 004.75
DOI https://doi.org/10.32782/2521-6643-2025-2-70.9

Божко О. Ю., аспірант кафедри інформаційних управляючих систем
Харківського національного університету радіоелектроніки
ORCID: 0009-0004-6820-1228

МЕТОД КЛАСИФІКАЦІЇ ДОКУМЕНТІВ ЗА СКЛАДНІСТЮ ЕКСТРАКЦІЇ ДАНИХ ВЕЛИКИМИ МОВНИМИ МОДЕЛЯМИ

У статті розглянуто актуальну проблему оптимізації автоматизованої екстракції даних з ділових документів за допомогою великих мовних моделей (LLM). Якість обробки документів суттєво варіюється в залежності від їх структурних та семантичних характеристик. Відсутність методів попереднього прогнозування якості екстракції призводить до неефективного використання ресурсів. Існуючі дослідження в галузі класифікації документів орієнтовані на тематичну категоризацію, а не на оцінку технічної складності отримання даних.

Для розв'язання даної проблеми запропоновано метод класифікації документів за складністю їх обробки з використанням великих мовних моделей. Метод базується на розмітці документів за бінарними ознаками структурно-семантичної складності. Для кожного документа виконується автоматична екстракція даних трьома мовними моделями у режимі без прикладів з обчисленням інтегральної метрики якості екстракції через гармонійне середнє точності та повноти. На основі метрик формуються класи складності, після чого будуються класифікатори з використанням багатокласової логістичної регресії. Валідація здійснюється через стратифіковану перехресну перевірку.

Ключовою особливістю методу є можливість автоматичного визначення очікуваної якості обробки документа на основі його формалізованих характеристик. Експериментальну перевірку працездатності методу здійснено на корпусі синтетичних документів з варійованими характеристиками складності. Для трьох LLM побудовано класифікатори триступеневої складності. Аналіз вагових коефіцієнтів виявив критичні фактори складності, які демонструють найбільший негативний вплив на якість екстракції.

Запропоноване рішення має як теоретичне так і практичне значення. Наукова новизна полягає у створенні оригінального емпірично обґрунтованого методу класифікації документів, де цільовою змінною виступає очікувана якість екстракції даних мовними моделями. Практична цінність розробленого методу полягає у можливості автоматизованого прийняття рішень про стратегію обробки документів в організаційних системах. Отримані результати створюють підґрунтя для розвитку інтелектуальних систем обробки документів та оптимізації використання обчислювальних ресурсів.

Ключові слова: *велика мовна модель, екстракція даних, класифікація документів, бінарна ознака складності, логістична регресія, метрика якості, промпт.*

Bozhko O. Yu. Method for classifying documents according to the complexity of data extraction by large language models

The article addresses the urgent problem of optimizing automated data extraction from business documents using large language models (LLM). The quality of processing documents varies significantly depending on their structural and semantic characteristics. The absence of methods for predicting extraction quality leads to inefficient resource utilization. Existing document classification research focuses on thematic categorization rather than assessing the technical complexity of data extraction.

To address this problem, a method for classifying documents based on the complexity of their processing by language models is proposed. The method is based on document markup using binary features of structural-semantic complexity. For each document, automatic data extraction is performed by three language models in zero-shot mode, with calculation of an integral quality metric through the harmonic mean of precision and recall. Based on these metrics, complexity classes are formed, followed by construction of classifiers using multiclass logistic regression. Validation is performed through stratified cross-validation.

The key feature of the method is the ability to automatically determine the expected quality of document processing based on its formalized characteristics. The method was tested on a corpus of synthetic documents with varied complexity characteristics. For three LLMs, three-level complexity classifiers were built. Analysis of weight coefficients revealed critical complexity factors that demonstrate the greatest negative impact on extraction quality.

The proposed solution has both theoretical and practical significance. The scientific novelty lies in creating the first empirically validated method for classifying documents where the target variable is the expected quality of data extraction by language models. The practical value is the possibility of automated decision-making regarding processing strategy in production systems. The results create a methodological foundation for developing intelligent document processing systems and optimizing computational resource utilization.

Key words: *large language models, data extraction, document classification, binary complexity features, logistic regression, quality metrics, prompt.*

Постановка проблеми. Автоматизація обробки ділових документів із використанням великих мовних моделей (LLM) стає поширеною практикою в корпоративному середовищі, охоплюючи процеси від розпізнавання інвойсів до витягування даних з контрактів та актів. Однак якість екстракції слабоструктурованих даних істотно варіюється в залежності від характеристик конкретного документа. Одні із високою точністю оброблюються простим підходом без прикладів в промптах (zero-shot), в той час, як інші вимагають певної кількості прикладів (few-shot) для забезпечення відповідної якості екстракції. Відсутність попереднього прогнозування призводить до невиправданого використання дороговартісних методів обробки або, навпаки, до неприйнятно низької точності.

Наявні дослідження в цій галузі переважно зосереджуються на підвищенні точності екстракції post-factum (тонке налаштування промптів, постобробка результатів), а не на попередньому оцінюванні очікуваного результату. На теперішній час відсутні формалізовані підходи до кількісної оцінки складності документів для LLM-обробки, що вибір оптимальної стратегії екстракції. Існуючі методи класифікації документів орієнтовані переважно на тематичну категоризацію, а не на оцінку технічної складності автоматичного витягування структурованих даних. Крім того, відсутня методологія, що дозволила б масштабувати процес оцінки складності на великі корпуси документів без залучення експертів-аналітиків.

Тому актуальним є завдання розробки автоматичного методу класифікації документів за складністю екстракції, який би дозволив швидко відфільтровувати прості документи, та направляти складні документи на поглиблене вивчення факторів для визначення оптимальної стратегії їх обробки. Такий метод має забезпечити раціональне використання обчислювальних ресурсів. Він має поєднувати формалізовану систему оцінку складності документів, швидкий ML-класифікатор для їх стратифікації та селективний детальний аналіз факторів складності виключно для проблемних випадків.

Аналіз останніх досліджень і публікацій. Сучасні ділові процеси потребують обробки значних обсягів цифрових документів, які містять слабоструктуровані дані: інвойси, накладні, контракти, специфікації [1]. Зазвичай, такі документи створюються програмно, але їх автоматична обробка залишається складною задачею через варіативність макетів, семантичну неоднозначність та відсутність єдиних стандартів форматування. Застосування мультімодальних LLM на базі трансформерної архітектури відкрило нові можливості для екстракції даних завдяки здатності до контекстуального розуміння без попереднього навчання на конкретних шаблонах [2]. Однак питання оптимального вибору методу обробки залежно від характеристик конкретного документа залишається малодослідженим.

Мовні моделі ефективно використовувалися для екстракції даних з ділових документів ще до так званої «LLM революції». Так ще в 2020 році була продемонстрована можливість аналізу попередньо навченими моделями просторової інформації документів для підтвердження точності екстракції [3].

Після того, як великі мовні моделі стали масово доступними, численні дослідження підтвердили якісну перевагу систем, що використовують LLM-підходи, над традиційними системами, заснованими на правилах [4].

Водночас виявляються і критичні обмеження LLM-екстракції. Наприклад, істотна залежність якості екстракції від структурної складності документів [5] та проблеми з обробкою великих документів через обмеження контекстного вікна [6]. Були зафіксовані помилки в математичних обчисленнях та випадки генерації неіснуючих даних [7]. Крім того, в сучасних дослідженнях підкреслюється необхідність валідації результатів у зв'язку з ризиком генерації правдоподібних, але неточних значень [8].

Разом з цим значною прогалиною є відсутність методів попереднього оцінювання якості екстракції. Наявні дослідження фокусуються на підвищенні точності через налаштування промптів або постобробку, але не пропонують способів прогнозування ефективності перед застосуванням LLM-процедур.

Класифікація документів традиційно вирішує задачу категоризації за типом або тематичною приналежністю. У таких дослідженнях предикторами виступають структурні, текстові та візуальні характеристики документів, а цільовою змінною – належність до певного класу (наприклад, рахунок-фактура, контракт, квитанція, медичний запис).

Так в [4] для класифікації медичних документів дослідники використовували в якості предикторів частоту термінів, довжину тексту, наявність специфічних ключових слів, а метою класифікації було визначення типу клінічного документа. В [9] для класифікації документів цифрового розвитку використовуються TF-IDF вектори, лексичне різноманіття та структурні метрики в якості предикторів для віднесення документів до тематичних категорій.

Конструювання предикторів є критичним етапом побудови класифікаторів. Структурні ознаки включають кількість таблиць, рядків, колонок, щільність тексту, наявність вкладених структур. Текстові предиктори охоплюють частотні характеристики термінів, довжину документа, мовне різноманіття. Візуальні ознаки враховують розташування елементів на сторінці, типографічні параметри, геометричні характеристики текстових блоків. Дослідження класифікації біомедичних документів [10, 11] та документів за структурними ознаками [12] демонструють ефективність комбінування різнотипних предикторів для підвищення точності визначення типу документа.

Недостатня кількість досліджень, де цільовою змінною класифікації документів є очікувана якість екстракції даних суттєво ускладнює попереднє прогнозування ефективності обробки та раціональний вибір стратегії екстракції.

Мета дослідження полягає у розробці метода класифікації документів за ознаками складності екстракції слабоструктурованих даних великими мовними моделями, який би забезпечив оптимізацію використання обчислювальних ресурсів через поєднання швидкого ML-класифікатора та селективного детального аналізу факторів складності.

Виклад основного матеріалу. Для розв'язання задачі класифікації документів за складністю екстракції даних великими мовними моделями в роботі пропонується поетапний метод. Цей метод включає реалізацію п'яти послідовних етапів.

1. Формування корпусу документів з маркуванням за бінарними критеріями складності.
2. Проведення екстракції за допомогою LLM у контрольованому режимі без прикладів (zero-shot) з розрахунком еталонних метрик якості.
3. Формування класів складності на основі емпіричного розподілу якості екстракції.
4. Побудова ML-класифікаторів для автоматичного визначення класу складності документа.
5. Валідація класифікаторів та оцінка економічної ефективності підходу.

Розглянемо детальний опис кожного з етапів.

Етап 1. В ході реалізації цього етапу створюється навчальний корпус документів з систематизованим описом їх характеристик складності та еталонними мітками якості екстракції.

Для формалізованого опису властивостей документів, що впливають на якість автоматичної екстракції, потрібно спочатку сформувати систему бінарних ознак їх складності. Кожна з ознак складності представляє собою індикатор, який фіксує наявність конкретної характеристики, яка ускладнює обробку документів.

Формування переліку ознак базується на аналізі збоїв у системах екстракції даних, а також на системному аналізі літературних джерел. Їх детальний опис наведено в (табл. 1).

Таблиця 1

Система формальних ознак складності документів для LLM-обробки

№	Ознака складності c_{ij}	Опис	Механізм впливу на LLM
1.	dense_listing	Документ містить понад 50 табличних записів	Когнітивне перевантаження, проблеми з контекстним вікном
2.	has_multilingual_content	Наявність інформації, що надана двома або більше мовами	Міжмовна інтерференція
3.	has_ambiguous_fields	Скорочення без розшифрування	Неоднозначність інтерпретації
4.	has_unlabeled_data	Відсутність явних маркерів типу даних	Складність визначення семантичної ролі
5.	has_compound_cells	Множинні значення в одному полі	Порушення атомарності даних
6.	has_multiline_records	Табличні записи займають декілька рядків	Порушення очікування «один рядок = один запис»
7.	has_spanning_data	Дані перетинають границі колонок/рядків	Порушення табличної структури
8.	has_header_data_mismatch	Невідповідність заголовків та даних	Протиріччя між очікуваннями та реальністю
9.	requires_calculations	Потреба в арифметичних обчисленнях	Не всі LLM однаково точні в математиці
10.	has_contextual_dependencies	Дані з заголовка застосовуються до всіх записів	Вимагає розуміння ієрархії документа
11.	has_implicit_structure	Логічна організація без явних маркерів	Потреба в виведенні структури з контексту

Розмітка цих 11 ознак складності здійснюється на рівні типу документа через розробку узгоджених правил присвоєння бінарних значень. Результатом є «профіль складності документу *i*» – вектор характеристик складності $C_i = (c_{i1}, c_{i2} \ldots, c_{i11})$, де $c_{ij} \in \{0,1\}$, $i = \overline{1,n}$, $j = \overline{1,11}$.

Формування навчального корпусу здійснюється за принципом забезпечення репрезентативності покриття комбінацій ознак складності при збереженні реалістичності структури та змісту документів. Заповнення шаблонів здійснюється з використанням реалістичних, але деперсоналізованих даних, що дозволяє зберегти автентичність структурно-семантичних характеристик документів при дотриманні вимог конфіденційності. Паралельно з генерацією документів створюються еталонні файли розмітки.

Формально корпус документів $i=\overline{1,n}$ описується матрицею ознак складності $C=\left(c_{ij}\right)_{n\times 11}$, де рядки відповідають документам, а стовпці – бінарним ознакам складності $j=\overline{1,11}$.

Контроль якості сформованого корпусу включає верифікацію покриття простору ознак, аналіз збалансованості розподілу характеристик складності та перевірку відсутності надлишкових кореляцій між предикторами. Сформований корпус використовується як навчальна і валідаційна база для побудови та оцінки регресійних моделей прогнозування якості екстракції даних на рівні документа.

Етап 2. Для забезпечення коректності порівняння застосовується єдиний протокол обробки всіх документів. Екстракція виконується в zero-shot режимі, що дозволяє оцінити базову здатність моделей до обробки документів різного рівня складності без додаткового налаштування під конкретні формати.

Для екстракції використовуються три LLM з різною архітектурою: GPT-4.1-nano, GPT-4o-mini та Gemini-1.5-flash з фіксованими параметрами декодування для кожної з моделей. Результат екстракції формується у вигляді структурованих даних відповідно до попередньо визначеної схеми, що специфікує назви полів, типи даних та кардинальності записів.

Для кожного документа d_i, $i=\overline{1,115}$, кожна модель $m=\overline{1,3}$ генерує набір екстрагованих записів E_{im}, який зберігається для подальшого порівняння з еталонною розміткою R_i.

Якість екстракції оцінюється зіставленням еталонної розмітки документа з результатами екстракції. Для кожної комбінації (документ d_i, модель m) виконується процедура порівняння екстрагованих записів з еталонними на основі подібності значень полів. На основі цього для кожного спостереження визначаються:

TP_{im} (true positives) – кількість коректно екстрагованих записів, що мають відповідний еталонний запис;

FP_{im} (false positives) – кількість помилково екстрагованих записів, що не мають відповідності в еталоні;

FN_{im} (false negatives) – кількість пропущених записів, що присутні в еталоні, але не екстраговані моделлю.

На основі цих величин обчислюються метрики точності (Precision) та повноти (Recall) екстракції:

$$\mathrm{Precision}_{im}=\frac{\mathrm{TP}_{im}}{\mathrm{TP}_{im}+\mathrm{FP}_{im}} \tag{1}$$

$$\mathrm{Recall}_{im}=\frac{\mathrm{TP}_{im}}{\mathrm{TP}_{im}+\mathrm{FN}_{im}} \tag{2}$$

Інтегральна оцінка якості екстракції обчислюється як гармонічне середнє точності та повноти – F1-міра:

$$F1_{im}=2\cdot\frac{\mathrm{Precision}_{im}\cdot\mathrm{Recall}_{im}}{\mathrm{Precision}_{im}+\mathrm{Recall}_{im}},\quad i=\overline{1,115},\quad m=\overline{1,3} \tag{3}$$

Результатом виконання другого етапу є три навчальні вибірки по 115 документів для розв'язання задачі класифікації. В контексті задачі предикторами виступають 11 бінарних характеристик складності документа: $C_i=\left(c_{i1},c_{i2},...,c_{i,11}\right)$. Відповідно, цільовою змінною є якість екстракції $F1_{im}\in[0,1]$.

Етап 3. Залежність якості екстракції від моделі стає аргументом на користь вибору модель-специфічної стратифікації за складністю. Для кожної моделі визначаються індивідуальні пороги. Такий підхід забезпечує можливість застосування методу до нових моделей на тому самому корпусі документів.

Пороги класів встановлюються виходячи з практичного призначення кожного класу в контексті вибору стратегії обробки документів. Розглядаються три категорії складності, що відповідають різним підходам до екстракції даних:

– **easy** – документи для яких базовий zero-shot підхід забезпечує стабільно високу якість екстракції;

– **hard** – документи які потребують застосування підсилених стратегій або, у крайніх випадках, ручної обробки;

– **medium** – документи де базовий підхід може бути недостатнім, але додаткові зусилля (наприклад, налаштування промпту) здатні підвищити якість до прийнятного рівня.

На основі аналізу практичних застосувань екстракції даних та врахування стандартів якості в задачах машинного навчання встановлено наступні порогові значення: $T_{\text{easy}}=0.85, T_{\text{hard}}=0.70$.

Для кожної моделі m виконується класифікація 115 результатів екстракції за встановленими порогами:

$$\mathrm{class}_m\left(F1_{im}\right)=\begin{cases}\text{Easy, якщо } \mathrm{F1_{im}}>0.85\\ \text{Medium, якщо } 0.70<\mathrm{F1_{im}}\le 0.85\\ \text{Hard, якщо } \mathrm{F1_{im}}\le 0.70\end{cases} \tag{4}$$

Результатом виконання третього етапу є формування трьох навчальних вибірок для задачі класифікації – по одній для кожної моделі. Кожна вибірка складається зі 115 спостережень, де предикторами виступають 11

бінарних ознак складності документа C_i, а цільовою змінною – клас складності $\text{class}_m \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$, визначений на основі практично обґрунтованих порогів.

Етап 4. Для розв'язання задачі багатокласової класифікації обрано метод логістичної регресії з регуляризацією. Вибір зумовлений наступними чинниками: інтерпретовність моделі через коефіцієнти вагових внесків ознак, стабільність роботи на обмежених вибірках, відсутність схильності до перенавчання при невеликій кількості предикторів, детермінованість результатів.

Класифікація виконується у два послідовних етапи: спочатку проводиться стандартизація вхідних ознак, після чого застосовується багатокласова логістична регресія. Стандартизація застосовується без центрування, оскільки бінарні ознаки мають фіксований діапазон [0,1]. Логістична регресія налаштована з параметрами: схема "один проти решти" (multi_class='ovr'), збалансовані вагові коефіцієнти класів (class_weight='balanced') для компенсації нерівномірного розподілу спостережень, обрана максимальна кількість ітерацій 2000 для гарантованої збіжності оптимізації.

Для об'єктивної оцінки якості класифікації застосовано стратифіковану перехресну валідацію (stratified k-fold cross-validation). Кількість фолдів визначається адаптивно: $k = \min(5, n_{\text{Hard}})$, де n_{Hard} – обсяг найменшого класу у вибірці моделі.

Для оцінювання якості класифікації використовуються наступні метрики: загальна точність (accuracy), F1-міра для кожного класу окремо, макроусереднена F1-міра (macro-averaged F1), збалансована точність (balanced accuracy), що враховує нерівномірність класів через усереднення частки правильних класифікацій у кожному класі.

Результатом виконання четвертого етапу є навчені класифікатори – по одному для кожної великої мовної моделі – що дозволяють за вектором 11 бінарних ознак документа автоматично визначати очікуваний клас складності його обробки. Ці класифікатори інтегруються в систему маршрутизації документопотоків, де на основі передбаченого класу приймається рішення про застосування базової (zero-shot), підсиленої (few-shot) або альтернативної стратегії екстракції даних.

Для кожної з трьох досліджених мовних моделей (GPT-4.1-nano, GPT-4o-mini, Gemini-1.5-flash) побудовано окремий класифікатор на основі багатокласової логістичної регресії. Класифікатор представлено набором числових коефіцієнтів, що визначають вагові внески кожної з 11 бінарних ознак складності документа у визначення класу.

Нехай $\mathbf{x} = (x_1, x_2, \ldots, x_{11})$ – вектор бінарних ознак складності документа, де $x_j \in \{0,1\}$ відповідає наявності або відсутності j-ї характеристики складності. Для кожного класу $k \in \{\text{Easy}, \text{Medium}, \text{Hard}\}$ обчислюється лінійна комбінація ознак:

$$z_k(\mathbf{x}) = w_{k0} + \sum_{j=1}^{11} w_{kj} \cdot x_j \tag{5}$$

де w_{k0} – вільний член для класу k, а w_{kj} – ваговий коефіцієнт j-ї ознаки для класу k.

Ймовірність приналежності документа до класу k визначається логістичною функцією (сигмоїдою):

$$p_k(x) = \sigma(z_k(x)) = \frac{1}{1 + exp\{-z_k(x)\}} \tag{6}$$

Передбачений клас складності визначається як:

$$\hat{y}(x) = \arg\max_k p_k(x) \tag{7}$$

Етап 5. Для документа класу c з істинною складністю c_{true} та передбаченою складністю c_{pred} математичне сподівання втрат від помилкової класифікації обчислюється за формулою:

$$E[Cost] = \sum_{c_{true}} \sum_{c_{pred}} P(c_{pred} c_{true}) \cdot L(c_{true}, c_{pred}), \tag{8}$$

де $L(c_{true}, c_{pred})$ – функція втрат, що визначає критичність помилкової класифікації. Конкретні значення цієї функції визначаються вимогами конкретної прикладної задачі.

У подальшому приймаємо консервативне припущення: пропуск помилки коштує суттєво дорожче за зайву перевірку, тому документи класу Easy обробляємо автоматично, а Medium і Hard – підлягають експертній верифікації.

На основі матриці невідповідностей, що отримана на етапі 4, розраховуються частота критичних помилок та частота надлишкових перевірок:

$$P_{critical} = \frac{FN_{Easy}}{N_{Medium} + N_{Hard}} \tag{9}$$

$$P_{redundant} = \frac{FP_{Easy}}{N_{Easy}} \tag{10}$$

Аналіз та обговорення результатів дослідження. Результатом дослідження є розроблений метод автоматичної класифікації документів за складністю екстракції даних великими мовними моделями. Цей метод реалізовано з використанням математичної моделі, що приймає на вхід профіль складності документа та повертає передбачуваний клас складності обробки. Нижче наведено аналіз результатів, що отримані на проміжних етапах та в ході експериментального дослідження.

Після завершення другого етапу було проведено оцінку зв'язку між 11 бінарними критеріями складності та метрикою $F1_{doc}$ за допомогою пойнт-бісеріальних коефіцієнтів кореляції. Загалом сила зв'язку оцінюється як помірна. Вплив критеріїв складності на якість екстракції наведено в (табл. 2).

Таблиця 2

Вплив критеріїв складності на якість екстракції

Критерій	GPT-4o-mini	GPT-4.1-nano	Gemini 1.5 Flash	Середнє |r|	Ранг (за середнім |r|)
contextual_dependencies	-0.553	-0.546	-0.581	0.560	1
unlabeled_data	-0.563	-0.476	-0.598	0.546	2
spanning_data	-0.430	-0.462	-0.449	0.447	3
compound_cells	-0.335	-0.397	-0.391	0.374	4
header_data_mismatch	-0.364	-0.265	-0.463	0.364	5
implicit_structure	-0.429	-0.260	-0.357	0.349	6
ambiguous_fields	-0.258	-0.297	-0.469	0.341	7
requires_calculations	-0.131	-0.330	-0.277	0.246	8
dense_listing	-0.196	-0.282	-0.192	0.223	9
multilingual_content	0.000	-0.164	-0.283	0.149	10
multiline_records	-0.089	-0.134	-0.170	0.131	11

У (табл. 3) наведено емпіричний розподіл документів за класами складності для кожної великої мовної моделі. Класи сформовано за фіксованими порогами (згідно з (4)). Розподіли відображають реальну різницю у поведінці моделей на одному й тому самому корпусі.

Таблиця 3

Розподіл за класами складності та характеристики якості

Модель	Клас	N	%	Діапазон F1	Середнє	Std
GPT-4.1-nano	Easy	49	42.6%	0.856–0.965	0.903	0.036
	Medium	54	47.0%	0.714–0.849	0.797	0.039
	Hard	12	10.4%	0.504–0.681	0.633	0.061
GPT-4o-mini	Easy	62	53.9%	0.854–0.965	0.910	0.036
	Medium	43	37.4%	0.705–0.850	0.784	0.036
	Hard	10	8.7%	0.495–0.690	0.627	0.073
Gemini-1.5-flash	Easy	88	76.5%	0.855–0.993	0.921	0.036
	Medium	25	21.7%	0.774–0.850	0.816	0.019
	Hard	2	1.7%	0.673–0.680	0.677	0.004

Розподіл документів по класах суттєво відрізняється між моделями при застосуванні однакових порогових значень, що підтверджує різницю в їх можливостях обробки документів. Модель Gemini демонструє найбільшу частку Easy-документів (76.5%), що вказує на її високу ефективність для даного корпусу. Натомість GPT-4.1-nano класифікує лише 42.6% документів як Easy, при цьому майже половина документів (47.0%) потрапляє до класу Medium, що свідчить про більш широкий діапазон складності обробки для цієї моделі.

Особливу увагу привертає клас Hard для моделі Gemini: лише 2 документи (1.7%) демонструють якість екстракції нижче порогу 0.70. Така мала вибірка обмежує статистичну надійність характеристик цього класу для даної моделі (std=0.004 на основі двох спостережень), проте підтверджує загальну високу ефективність Gemini для обробки документів представленого корпусу.

У (табл. 4) наведено розподіл спостережень за класами складності.

Нерівномірний розподіл класів виникає внаслідок застосування фіксованих порогів до емпіричних розподілів якості екстракції, що суттєво відрізняються між моделями. Для моделі Gemini клас Hard представлений лише двома спостереженнями, що обмежує статистичну надійність оцінок якості класифікації для цього класу. Метрики якості класифікації наведені в (табл. 5).

Таблиця 4

Розподіл спостережень за класами складності

Модель	Easy	Medium	Hard	Кількість фолдів
GPT-4.1-nano	49	54	12	5
GPT-4o-mini	62	43	10	5
Gemini-1.5-flash	88	25	2	2

Таблиця 5

Метрики якості класифікації (stratified k-fold CV)

Модель	Accuracy	F1 (Easy)	F1 (Medium)	F1 (Hard)	F1 (macro)	Balanced Accuracy
GPT-4.1-nano	0.704	0.800	0.704	0.370	0.625	0.632
GPT-4o-mini	0.696	0.779	0.667	0.444	0.630	0.669
Gemini-1.5-flash	0.661	0.781	0.453	0.000	0.411	0.402

Класифікатори демонструють задовільну загальну точність (66 – 70%), проте якість розпізнавання класу Hard суттєво нижча порівняно з Easy та Medium. Для моделі Gemini F1-міра класу Hard дорівнює нулю, що відображає неможливість надійного навчання на двох спостереженнях. Макроусереднена F1-міра варіює від 0.411 (Gemini) до 0.630 (GPT-4o-mini), що вказує на помірну якість класифікації за умов обмеженого обсягу навчальних даних та нерівномірного розподілу класів.

Детальний аналіз помилок класифікації для моделі GPT-4o-mini надано у (табл. 6) (матриця невідповідностей, confusion matrix).

Таблиця 6

Матриця невідповідностей для GPT-4o-mini (OOF передбачення)

	Передбачено Easy	Передбачено Medium	Передбачено Hard
Фактичний Easy	44	14	4
Фактичний Medium	6	30	7
Фактичний Hard	1	3	6

Аналіз матриці свідчить про типовий патерн помилок. Класифікатор схильний недооцінювати складність (помилкова класифікація Medium як Easy зустрічається частіше, ніж переоцінка). Для класу Hard спостерігається значна частка помилкових віднесень до сусідніх класів через малий обсяг навчальної вибірки цього класу.

Коефіцієнти логістичної регресії дозволяють кількісно оцінити внесок кожної ознаки у класифікацію. Порівняння коефіцієнтів між моделями показує узгодженість у визначенні критичних факторів складності. Ознаки compound_cells, dense_listing та contextual_dependencies стабільно демонструють високі абсолютні значення коефіцієнтів для всіх трьох архітектур LLM.

Отримані класифікатори забезпечують базовий рівень автоматизації визначення складності документів, достатній для практичного впровадження диференційованих стратегій обробки. Водночас слід враховувати наступні обмеження методу.

По-перше, обмежений обсяг вибірки класу Hard (особливо для моделі Gemini: 2 спостереження) призводить до низької статистичної надійності метрик якості для цього класу. Оцінки F1-міри для Hard характеризуються високою дисперсією та можуть не відображати справжньої здатності класифікатора розпізнавати складні документи на ширших корпусах.

По-друге, використання бінарних ознак спрощує процедуру розмітки документів, проте не дозволяє враховувати градації прояву характеристик складності. Документи з однаковими бінарними профілями можуть суттєво відрізнятися за фактичною складністю обробки.

По-третє, застосування фіксованих порогів (0.70 та 0.85) для формування класів призводить до нерівномірного розподілу спостережень, що обмежує ефективність навчання класифікаторів на міноритарних класах.

Визначення 11 бінарних ознак складності документа може бути **автоматизоване** шляхом застосування алгоритмів структурного аналізу документів. **На практиці** для класифікації нового документу необхідно визначити вектор його бінарних ознак складності, застосувати формули вище з відповідним набором коефіцієнтів та отримати передбачений клас і рекомендацію щодо стратегії обробки.

Висновки та перспективи подальших досліджень. В роботі вперше отримано емпірично обґрунтовану систему класифікації документів за складністю їх автоматичної обробки великими мовними моделями, де цільовою змінною є очікувана якість екстракції, виміряна через метрику F1.

Удосконалено методику оцінювання якості екстракції слабоструктурованих даних через розробку композитної метрики на рівні документа, що враховує якість обробки різних семантичних компонентів з відповідними ваговими коефіцієнтами.

Розроблений метод створює основу для автоматизованого прийняття рішень про вибір стратегії обробки документів у виробничих системах. Класифікатор дозволяє на етапі надходження документа передбачити очікувану якість його автоматичної обробки та обрати оптимальну стратегію: повна автоматизація для документів класу Easy, експертна перевірка для класів Medium та Hard.

Метод апробовано на корпусі з 115 синтетичних ділових документів з систематично варійованими характеристиками складності. Результати експериментів підтверджують статистично значущий зв'язок між формалізованими ознаками складності та якістю екстракції даних для трьох сучасних архітектур мовних моделей.

Перспективи подальших досліджень пов'язані з кількома напрямами розвитку запропонованого методу. Інтеграція неперервних кількісних характеристик документів поряд з бінарними ознаками може підвищити точність класифікації. До таких характеристик належать: щільність текстової інформації на сторінці, кількість табличних елементів, співвідношення структурованих та неструктурованих фрагментів.

Валідація методу на документах інших класів дозволить оцінити універсальність запропонованої системи ознак та виявити специфічні фактори складності для різних предметних областей.

Перевірка класифікаторів на нових архітектурах мовних моделей, що з'являться після завершення цього дослідження, забезпечить оцінку темпоральної стабільності результатів та виявить потребу в періодичному переналаштуванні моделей.

Побудова окремих класифікаторів для підмножин документів з домінуванням певних типів складності (наприклад, багатомовні документи, документи з матричними структурами) може забезпечити вищу точність передбачень у порівнянні з єдиною універсальною моделлю.

Список використаних джерел:

1. Божко О. Ю. Використання великих мовних моделей для розпізнавання інформації в нумізматичних описах. *Наука і техніка сьогодні*. 2024. № 1(29). С. 615–625. DOI: 10.52058/2786-6025-2024-1(29)-615-625.

2. Божко О. Ю. Розробка ітеративного методу екстракції даних з неструктурованих документів на основі використання великих мовних моделей. *Вісник Кременчуцького національного університету імені Михайла Остроградського*. 2025. № 1. С. 119–124. DOI: 10.32782/1995-0519.2025.1.15.

3. Xu Y., Li M., Cui L., Huang S., Wei F., Zhou M. LayoutLM: pre-training of text and layout for document image understanding. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '20). 2020. С. 1192–1200. DOI: 10.1145/3394486.3403172.

4. Rijcken E., Zervanou K., Mosteiro P., Scheepers F., Spruit M., Kaymak U. Machine learning vs. rule-based methods for document classification of electronic health records within mental health care: a systematic literature review. *Natural Language Processing Journal*. 2025. Т. 10. Стаття 100129. DOI: 10.1016/j.nlp.2025.100129.

5. Li B., та ін. AID-Agent: an LLM-Agent for advanced extraction and integration of documents. Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025). 2025. С. 80–88. DOI: 10.18653/v1/2025.realm-1.6.

6. Li H., та ін. Extracting financial data from unstructured sources: leveraging large language models. *SSRN Electronic Journal*. 2023. DOI: 10.2139/ssrn.4567607. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4567607 (Дата звернення: 15.10.2025).

7. Almeida F. C., Caminha C. Evaluation of entry-level open-source large language models for information extraction from digitized documents. Symposium on Knowledge Discovery, Mining and Learning (KDMiLe). 2024. С. 25–32. DOI: 10.5753/kdmile.2024.243859. (Офіц. стор.: sol.sbc.org.br).

8. Tito R., Karatzas D., Valveny E. Hierarchical multimodal transformers for multi-page DocVQA. arXiv:2212.05935. 2023. DOI: 10.48550/arXiv.2212.05935. URL: https://arxiv.org/abs/2212.05935 (Дата звернення: 15.10.2025).

9. Ranaweera U., Mawitagama B., Liyanage S., Keshan S., De Silva T., Hewawalpita S. Comparison of machine learning models to classify documents on digital development. у кн.: Data Science and Artificial Intelligence / ред. С. Anutariya, M. M. Bonsangue. Singapore: Springer, 2023. (CCIS, т. 1942). С. 59–73. DOI: 10.1007/978-981-99-7969-1_5.

10. Le D. X., Thoma G. R. Page layout classification technique for biomedical documents. Proceedings of the World Multiconference on Systems, Cybernetics and Informatics (SCI). 2000. Т. X. С. 348–352. URL: https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub2000015.pdf (Дата звернення: 15.10.2025).

11. Petrov K., Chalyi T. Situational model of a medical business process. *Bulletin of National Technical University "KhPI"*. Series: System Analysis, Control and Information Technologies. 2024. № 2(12). С. 42–45. DOI: 10.20998/2079-0023.2024.02.07.

12. Shin C., Doermann D., Rosenfeld A. Classification of document pages using structure-based features. *International Journal on Document Analysis and Recognition (IJDAR)*. 2001. Т. 3, № 4. С. 232–247. DOI: 10.1007/PL00013566.

References:

1. Bozhko, O.Yu. (2024). Vykorystannia velykykh movnykh modelei dlia rozpiznavannia informatsii v numizmatychnykh opysakh [Using large language models for information recognition in numismatic descriptions]. *Nauka i tekhnika sohodni*, 1(29), 615–625. https://doi.org/10.52058/2786-6025-2024-1(29)-615-625 [in Ukrainian].

2. Bozhko, O.Yu. (2025). Rozrobka iteratyvnoho metodu ekstraktsii danykh z nestrukturovanykh dokumentiv na osnovi vykorystannia velykykh movnykh modelei [Development of an iterative method for data extraction from unstructured documents based on the use of large language models]. *Visnyk Kremenchutskoho natsionalnoho universytetu imeni Mykhaila Ostrohradskoho.* 1, 119–124. https://doi.org/10.32782/1995-0519.2025.1.15 [in Ukrainian].

3. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., & Zhou, M. (2020). LayoutLM: Pre-training of text and layout for document image understanding. In Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 1192–1200). https://doi.org/10.1145/3394486.3403172

4. Rijcken, E., Zervanou, K., Mosteiro, P., Scheepers, F., Spruit, M., & Kaymak, U. (2025). Machine learning vs. rule-based methods for document classification of electronic health records within mental health care: A systematic literature review. *Natural Language Processing Journal*, 10, Article 100129. https://doi.org/10.1016/j.nlp.2025.100129

5. Li, B., et al. (2025). AID-Agent: An LLM-agent for advanced extraction and integration of documents. In Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025) (pp. 80–88). https://doi.org/10.18653/v1/2025.realm-1.6

6. Li, H., et al. (2023). Extracting financial data from unstructured sources: Leveraging large language models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4567607. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4567607

7. Almeida, F. C., & Caminha, C. (2024). Evaluation of entry-level open-source large language models for information extraction from digitized documents. In Symposium on Knowledge Discovery, Mining and Learning (KDMiLe 2024) (pp. 25–32). https://doi.org/10.5753/kdmile.2024.243859

8. Tito, R., Karatzas, D., & Valveny, E. (2023). Hierarchical multimodal transformers for multi-page DocVQA. arXiv:2212.05935. https://doi.org/10.48550/arXiv.2212.05935. Retrieved from: https://arxiv.org/abs/2212.05935

9. Ranaweera, U., Mawitagama, B., Liyanage, S., Keshan, S., De Silva, T., & Hewawalpita, S. (2023). Comparison of machine learning models to classify documents on digital development. In C. Anutariya & M. M. Bonsangue (Eds.), Data Science and Artificial Intelligence (CCIS, Vol. 1942, pp. 59–73). Springer. https://doi.org/10.1007/978-981-99-7969-1_5

10. Le, D. X., & Thoma, G. R. (2000). Page layout classification technique for biomedical documents. In Proceedings of the World Multiconference on Systems, Cybernetics and Informatics (SCI) (Vol. X, pp. 348–352). Retrieved from: https://lhncbc.nlm.nih.gov/LHC-publications/PDF/pub2000015.pdf

11. Petrov, K., & Chalyi, T. (2024). Situational model of a medical business process. *Bulletin of National Technical University "KhPI"*. Series: System Analysis, Control and Information Technologies, 2(12), 42–45. https://doi.org/10.20998/2079-0023.2024.02.07

12. Shin, C., Doermann, D., & Rosenfeld, A. (2001). Classification of document pages using structure-based features. *International Journal on Document Analysis and Recognition (IJDAR)*, 3(4), 232–247. https://doi.org/10.1007/PL00013566

Дата надходження статті: 17.10.202
Дата прийняття статті: 10.11.2025
Опубліковано: 30.12.2025