



Д. Я. Доскач, В. М. Хавалко

Національний університет "Львівська політехніка", м. Львів, Україна

ВПЛИВ ПАРАМЕТРІВ ТЕКСТОВИХ ДОКУМЕНТІВ НА ТОЧНІСТЬ РОБОТИ МУЛЬТИМОДАЛЬНОЇ МОДЕЛІ GPT-4 VISION У ЗАВДАННЯХ РОЗПІЗНАВАННЯ ТЕКСТУ

Проаналізовано вплив параметрів текстових документів на ефективність мовної мультимодальної моделі GPT-4 Vision, яку активно застосовують для цифровізації документів зі складною структурою. Досліджено залежність точності роботи моделі від таких факторів, як розмір шрифту, фізичний розмір документа (кількість сторінок формату A4) та загальний обсяг тексту (кількість символів). Особливу увагу приділено проблемі генерування моделлю некоректної інформації ("галюцинацій"), яка залишається істотним недоліком цього підходу до цифровізації текстів. Розроблено п'ять експериментальних наборів PDF-документів із варіативними параметрами шрифту, обсягу тексту та кількості сторінок, що дало змогу оцінити вплив кожного з факторів окремо й у їхніх комбінаціях. Для кількісного оцінювання частоти появи галюцинацій застосовано метрику точності ROUGE-L. Встановлено, що частота виникнення галюцинацій значно зростає у документах, які містять понад 5000 символів тексту зі шрифтом в 14 pt або більш ніж дві сторінки формату A4. Наприклад, для чотиристорінкового документа з обсягом тексту близько 12000 символів значення метрики ROUGE-L становило 0,51, тоді як для аналогічного за обсягом тексту, компактно розташованого на одній сторінці за рахунок зменшення розміру шрифту до 6 pt, показник точності був значно вищим (ROUGE-L = 0,87). Це свідчить про значний вплив фізичних параметрів документа на точність роботи мультимодальних моделей. Розроблено практичні рекомендації щодо раціонального використання моделі GPT-4 Vision для цифровізації складних документів: мінімальний рекомендований розмір шрифту – 6 pt; максимальний обсяг тексту – до 5000 символів або не більше двох сторінок за одне оброблення; потреба мінімізації пустих ділянок шляхом видалення зайвих полів і порожніх сторінок. Отримані результати мають практичну цінність для подальшого вдосконалення алгоритмів цифровізації документів зі складною структурою та для підвищення надійності роботи мультимодальних моделей. Запропоновану методику оцінювання впливу параметрів текстових документів на точність роботи моделі GPT-4 Vision можна використати для наступного дослідження, спрямованого на зменшення частоти виникнення галюцинацій у завданнях автоматизованого розпізнавання тексту.

Ключові слова: частота появи галюцинацій; розпізнавання символів; цифровізація документів; великі мовні моделі.

Вступ / Introduction

Перетворення документів із складною структурою, таких як резюме, звіти, таблиці та форми, у текстовий формат є важливим завданням для сучасних інформаційних систем. Зазвичай документи містять нелінійну структуру з елементами, які важко інтерпретувати: колонки, зображення, графіки, таблиці та інші. Традиційні методи, такі як оптичне розпізнавання тексту OCR (англ. *Optical Character Recognition*), забезпечують базову цифровізацію, однак часто стикаються з проблемами збереження контексту та структури документа.

Останні досягнення у сфері штучного інтелекту, зокрема багатомодальні моделі, такі як GPT-4 Vision [2, 11], відкривають нові перспективи для оброблення складних документів. Завдяки здатності обробляти як текстові, так і візуальні елементи, модель GPT-4 Vision може інтерпретувати документи зі складною структурою та отримувати корисні дані. Ця здатність дає змогу моделі працювати з різними форматами документів, за-

безпечуючи високий рівень гнучкості та точності. Водночас модель стикається з низкою викликів, ключовим із яких є проблема появи галюцинацій [4] – некоректне інтерпретування або створення неправдивої інформації.

У сучасній науковій літературі значну увагу приділяють технологіям OCR та їхній інтеграції зі штучним інтелектом для підвищення точності оброблення текстів [13]. Застосування моделі GPT-4 Vision для перетворення складноструктурованих документів у текстовий формат є новою та актуальною сферою досліджень. Незважаючи на значний потенціал моделі, питання виникнення галюцинацій, особливо у контексті роботи з текстовими документами, залишається нерозкритим. Відсутність систематичного аналізу впливу параметрів документа, таких як розмір шрифту, фізичні розміри чи обсяг тексту, створює потребу у глибшому дослідженні цього явища для розроблення нових ефективних методів вирішення згаданого завдання.

Завданням цього дослідження є проведення аналізу

Інформація про авторів:

Доскач Денис Ярославович, аспірант, кафедра систем штучного інтелекту.

Email: denys.y.doshchak@lpnu.ua; <https://orcid.org/0009-0000-7042-3541>

Хавалко Віктор Михайлович, канд. техн. наук, доцент, кафедра систем штучного інтелекту, заступник директора з науково-педагогічної роботи інституту комп'ютерних наук. Email: viktor.m.khavalko@lpnu.ua; <https://orcid.org/0000-0002-9585-3078>

Цитування за ДСТУ: Доскач Д. Я., Хавалко В. М. Вплив параметрів текстових документів на точність роботи мультимодальної моделі GPT-4 vision у завданнях розпізнавання тексту. Науковий вісник НЛТУ України. 2025, т. 35, № 2. С. 124–131.

Citation APA: Doshchak, D. Ya., & Khavalko, V. M. (2025). Impact of text document parameters on the accuracy of the multimodal GPT-4 vision model in text recognition tasks. *Scientific Bulletin of UNFU*, 35(2), 124–131. <https://doi.org/10.36930/40350214>

залежності частоти появи галюцинацій моделі GPT-4 Vision від таких параметрів документів, як розмір шрифту, фізичні розміри документа та обсяг тексту.

Об'єкт дослідження – перетворення документів у текстовий формат.

Предмет дослідження – методи і засоби встановлення впливу параметрів текстових документів на точність роботи мультимодальної моделі у завданнях розпізнавання тексту, що дасть змогу вдосконалити використання моделі GPT-4 Vision для цифровізації складноструктурованих документів та зменшити ймовірність генерування некоректної інформації.

Мета роботи – проаналізувати вплив параметрів документа на ефективність роботи моделі GPT-4 Vision, що дасть змогу розробити рекомендації з мінімізації появи галюцинацій та підвищення точності цифровізації складноструктурованих документів.

Для досягнення зазначеної мети визначено такі основні завдання дослідження:

- проаналізувати наявні підходи до цифровізації документів, що дасть змогу визначити їхні переваги, недоліки та потенційні напрями вдосконалення;
- провести експерименти для встановлення залежності частоти появи галюцинацій від параметрів документів, що дасть можливість оцінити вплив ключових характеристик тексту на якість генерування документів та запропонувати способи мінімізації помилок;
- розробити практичні рекомендації для покращення моделей, що уможливить підвищення точності оброблення документів і зменшення ризику появи некоректних або спотворених даних.

Аналіз останніх досліджень та публікацій. Модель GPT-4 Vision є перспективним інструментом для вирішення OCR-завдань, об'єднуючи здатність аналізувати як текстові, так і візуальні елементи документів. У роботі [14] оцінено його можливості за такими завданнями, як розпізнавання тексту на зображеннях, рукописних текстах, математичних виразів, структур таблиць та вилучення інформації зі складних документів. Результати демонструють ефективність моделі GPT-4V у роботі з латинським текстом та високоякісними зображеннями, однак виявляють обмеження у мультимовних сценаріях та рукописному тексті. Зокрема, відзначено схильність моделі GPT-4V до галюцинацій, особливо під час розпізнавання китайських рукописних текстів, де модель створювала змістовно правильний, але відсутній у документі текст.

У роботі [3] розглянуто ефективність моделі GPT-4 Vision у завданнях відповіді на запитання стосовно змісту документів із зображеннями. Результати показують, що мультимодальні моделі, такі як модель GPT-4 Vision, досягають кращих показників у завданнях, які потребують поєднання текстових даних із візуальними характеристиками документа, аніж традиційні OCR-підходи. Однак ефективність знижується для великих документів. Дослідження також ілюструє наявність "упередженості до початку" (primacy bias), коли модель краще виконує завдання з інформацією, розташованою на початку документа. Такі висновки підкреслюють потребу в подальшій адаптації моделей до роботи з багатосторінковими документами.

Схожі результати отримали й автори в роботі [17]. Розроблена модель, яка не використовує OCR і побудована на підставі мультимодальних великих мовних моделей MLLM (англ. *Multimodal Large Language Models*)

з використанням модуля адаптивного вирізання для збереження високої роздільної здатності тексту та просторових характеристик, демонструє значні досягнення у завданнях візуального розуміння мови, охоплюючи п'ять областей: документи, таблиці, діаграми, природні зображення та вебсторінки. Модель випереджає класичні підходи у восьми з десяти завдань, значно знижуючи обчислювальні витрати на навчання порівняно з традиційними OCR-залежними моделями. Однак виявлено і певні обмеження, такі як складнощі оброблення багатосторінкових документів та проблема появи галюцинацій під час генерування тексту. Це можна простежити у випадках, коли модель намагається інтерпретувати складні елементи документа або надає некоректну інформацію.

Доволі новим і важливим напрямом дослідження є інтеграція технологій OCR та Generative AI. У роботі [1] висвітлено різноманітні особливості їхнього спільного використання, враховуючи оцифрування текстів, класифікацію документів, автоматичне виправлення помилок, витягування даних, аналіз налаштувань, переклад і адаптацію текстів. Поєднання технологій OCR для розпізнавання тексту та технології Generative AI для контекстуального аналізу забезпечує значне підвищення точності та ефективності оброблення даних. Зокрема, зазначено, що технологія Generative AI дає змогу проводити уточнення та виправлення помилок OCR, що сприяє зменшенню кількості помилок. Завдяки аналізу початкового тексту за допомогою мовних моделей, технологія Generative AI автоматично визначає аномалії та пропонує виправлення, що є важливим у сферах, де потрібна висока точність, як-от юридичні або медичні документи. Ця робота вказує на важливість поєднання технологій OCR та Generative AI у цифровій трансформації документів, проте відсутній глибокий аналіз впливу параметрів документа на точність моделей, що залишається перспективним напрямом для подальшого дослідження.

У роботі [19] досліджено ефективність моделей під час виконання завдань розпізнавання тексту, таблиць та вилучення ключової інформації. Результати показують, що модель GPT-4V демонструє високу точність у передбаченні текстових відповідей, перевершуючи інші моделі за такими метриками, як Exact Match (EM) і F1-score. Важливим висновком є те, що мультимодальні моделі, зокрема GPT-4V, здатні інтегрувати текстові й візуальні елементи документів, забезпечуючи високу якість розуміння тексту навіть у складних макетах.

На підставі проведеного аналізу можна зробити висновки, що сучасні підходи, які містять як традиційні OCR-системи, так і мультимодальні моделі, такі як модель GPT-4 Vision, значно розширюють можливості оброблення документів зі складною структурою. Проте ці методи мають низку обмежень, зокрема труднощі з обробленням багатосторінкових і мультимовних документів, а також схильність до генерування галюцинацій – тексту, що не відповідає вхідним даним.

Матеріали та методи дослідження. Для досягнення поставлених завдань у дослідженні використано експериментальні методи аналізу впливу параметрів текстових документів на частоту появи галюцинацій у моделі GPT-4 Vision. Створено п'ять наборів PDF-документів із варіацією розміру шрифту, фізичного розміру документа та обсягу тексту. Для оцінювання результатів

застосовано метрику ROUGE-L, яка дає змогу кількісно визначити відхилення між початковим текстом документа та текстом, згенерованим моделлю. Аналіз проводили порівнянням отриманих оцінок для кожного експериментального набору, що дало змогу виявити закономірності впливу різних параметрів на точність роботи моделі GPT-4 Vision під час опрацювання текстових документів.

Результати дослідження та їх обговорення / Research results and their discussion

Постановка завдання. Потрібно проаналізувати вплив параметрів текстових документів на частоту виникнення галюцинацій під час їхнього аналізу за допомогою моделі GPT-4 Vision, яку створено OpenAI [11]. Мультиmodalна модель GPT-4 Vision об'єднує можливості аналізу тексту та зображень, тобто дає змогу їй працювати з документами у форматах PDF, які поєднують текстові й графічні елементи. Ця модель розроблена для завдань візуального розпізнавання об'єктів, опису зображень та аналізу комплексного візуального наповнення, тому її використання для розпізнавання текстових документів є перспективним та актуальним нап'ямом цього дослідження.

Важливість проведеного дослідження полягає в оцінюванні точності роботи моделі в разі зміни таких параметрів документів, як розмір шрифту, фізичний розмір документа та обсяг тексту, що дає змогу моделювати реальні сценарії використання моделі GPT-4 Vision для роботи з документами складної структури. Особливу увагу приділено здатності моделі зберігати контекст у багатосторінкових документах із високою щільністю тексту та оцінювання впливу параметрів на частоту виникнення помилок.

Проблема появи галюцинацій. Галюцинації є однією з ключових проблем мультиmodalних моделей, таких як модель GPT-4 Vision. Вони проявляються як результат генерування тексту, що не відповідає змісту початкового документа, або містить недоречну інформацію [6]. У завданнях перетворення тексту це може призводити до істотних проблем, коли отримана інформація є неточною або спотвореною.

Питання залежності частоти появи галюцинацій від параметрів документів залишається мало вивченим. Наприклад, зменшення шрифту або збільшення щільності тексту може призводити до збільшення кількості помилок. Аналогічно, більший фізичний розмір документа може спричинити втрату контексту під час аналізу багатосторінкових документів. Галюцинації також можуть виникати, коли модель створює текст на підставі передбачуваних, але не реальних, даних.

Методика проведення дослідження. Для аналізу впливу параметрів текстових документів на частоту появи галюцинацій у моделі GPT-4 Vision створено п'ять експериментальних наборів документів, кожен з яких мав на меті встановлення одного або кількох ключових параметрів: розміру шрифту, фізичного розміру документа або обсягу тексту. Усі документи генерували у форматі PDF, що давало змогу дослідити сценарії реального використання моделі в завданнях автоматизації. У всіх випадках за базовий формат сторінки обирали стандартну сторінку формату A4.

Перший набір документів спрямований на дослідження впливу розміру шрифту. Документи містили од-

наковий текстовий вміст, а шрифт змінювався. При цьому фізичні розміри сторінок і загальний обсяг тексту залишалися незмінними. Це дало можливість оцінити, чи впливає зміна розміру шрифту на здатність моделі коректно інтерпретувати текст без втрати його змісту.

Другий експериментальний набір спрямований на встановлення впливу фізичного розміру документа, де розміри тексту та його обсяг залишалися постійними. Для цього до текстового документа додавали порожні сторінки, збільшуючи кількість сторінок документа. Це дало можливість перевірити, чи впливають багатосторінкові формати на здатність моделі зберігати контекст і коректно працювати з великими документами.

У третьому наборі вивчали альтернативний спосіб зміни фізичного розміру документа шляхом збільшення міжрядкових відступів. Цей підхід давав змогу створити документ із більшою кількістю сторінок за рахунок розтягнення тексту, не змінюючи його змісту чи обсягу. Використання такого підходу дало змогу оцінити, чи залежить частота появи галюцинацій від візуального компоненту та структури тексту, коли кількість текстових блоків стала.

Четвертий набір документів спрямований на дослідження впливу обсягу тексту шляхом додавання нових сторінок із текстом до документа. На кожній сторінці був унікальний текст, який семантично пов'язаний із текстом на інших сторінках, утворюючи цілісний зміст. Такий підхід дав змогу оцінити, чи впливає збільшення обсягу тексту в документі на здатність моделі GPT-4 Vision зберігати контекст і підтримувати точність інтерпретації у складних сценаріях, де інформація розподілена на багатьох сторінках.

П'ятий набір використовували для дослідження залежності від обсягу тексту іншим способом. У цьому разі текст розташовували на тій самій кількості сторінок, але розмір шрифту зменшували, що давало змогу розмістити більше тексту на одній сторінці. Такий підхід забезпечував оцінку того, як щільність тексту впливає на частоту виникнення галюцинацій.

Кожен набір документів розроблено. Отже, щоб забезпечити контрольовану зміну одного чи кількох параметрів, таких як розмір шрифту, фізичний розмір документа або обсяг тексту, залежно від досліджуваного сценарію, можна оцінити вплив окремих характеристик документа або їх комбінацій на частоту появи галюцинацій, забезпечуючи базу для подальшого аналізу взаємозв'язків між параметрами документа і точністю роботи мультиmodalних моделей.

Методика оцінювання частоти появи галюцинацій. Для оцінювання частоти появи галюцинацій, що можуть виникати під час використання моделі GPT-4 Vision, обрано підхід порівняння тексту, згенерованого моделлю, з референтним текстом документа. У цьому дослідженні референтний текст є початковим текстом, використаним для створення PDF-документів. Порівняння виконувалося автоматизовано з використанням метрики ROUGE-L (англ. *Recall-Oriented Understudy for Gisting Evaluation*), яка враховує структурну й семантичну відповідність текстів.

Метрика ROUGE-L оцінює текст способом обчислення довжини найдовшої спільної підпоследовності слів LCS (англ. *Longest Common Subsequence*) між референтним текстом і текстом, згенерованим моделлю. Ця підпоследовність враховує порядок слів, що є критич-

ним для завдань, де важливо зберігати логічний і контекстний зв'язки тексту. Значення метрики ROUGE-L базується на таких компонентах [9]:

1. Recall (повнота). Цей показник демонструє, яка частка слів референтного тексту була відтворена у згенерованому тексті.

$$Recall = \frac{LCS}{\text{кількість слів у референтному тексті}}$$

2. Precision (точність). Цей показник визначає, наскільки добре модель дотримується структури референтного тексту.

$$Precision = \frac{LCS}{\text{кількість слів у згенерованому тексті}}$$

3. F1-міра. F1-міра збалансовує Precision і Recall, забезпечуючи узагальнений показник відповідності текстів.

$$F1 = 2 \cdot \frac{Precision + Recall}{Precision \cdot Recall}$$

Використання метрики ROUGE-L є доцільним у нашому завданні, оскільки воно враховує кілька важливих особливостей [9, 10]:

- збереження структури тексту. У разі, коли модель змінює порядок слів або пропускає важливі частини тексту, метрика ROUGE-L знижує оцінку, що дає змогу ідентифікувати галюцинації;
- гнучкість для різних розмірів тексту. Метрика ефективно працює як для коротких фрагментів тексту, так і для великих обсягів даних;
- семантичну узгодженість. Завдяки орієнтації на послідовності слів метрика ROUGE-L дає змогу враховувати контекстний зв'язок, що є критично важливим для аналізу текстів із складною структурою.

Процедура оцінювання частоти появи галюцинацій передбачає розрахунок метрики ROUGE-L для кожного документа в експериментальних наборах. У разі, коли значення метрики знижувалося, це інтерпретувалося як свідчення потенційних галюцинацій, спричинених змінами параметрів документа. Якщо значення метрики становить 1, це інтерпретувалося як відсутність галюцинацій.

Аналіз результатів дослідження. Як уже зазначено вище, для проведення дослідження підготовлено п'ять експериментальних наборів документів у форматі PDF. Далі, усі документи було перетворено у зображення способом вертикального об'єднання сторінок. Ці зображення використовували для подальшого їхнього аналізу за допомогою моделі GPT-4 Vision.

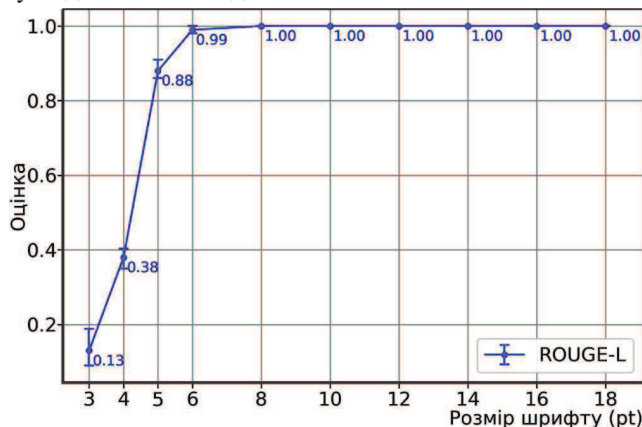


Рис. 1. Оцінювання метрикою ROUGE-L коректності оброблення набору документів зі сталим текстом та його розміром за змінного розміру шрифту / Evaluation of the correctness of processing a set of documents with fixed text and its size with variable font size using the ROUGE-L metric

Під час дослідження було отримано результати, які демонструють вплив різних параметрів текстових документів на частоту появи галюцинацій у моделі GPT-4 Vision.

Для набору документів із варіацією розміру шрифту виявлено, що модель коректно обробляє текст із розміром шрифту в межах 8-18 pt (рис. 1). У разі, коли розмір шрифту зменшувався до 6 pt і менше, частота появи галюцинацій різко зростала. Це може бути пов'язано зі складністю розпізнавання малих символів, однак для реальних сценаріїв такий розмір тексту рідко використовують через його незручність для читання людьми.

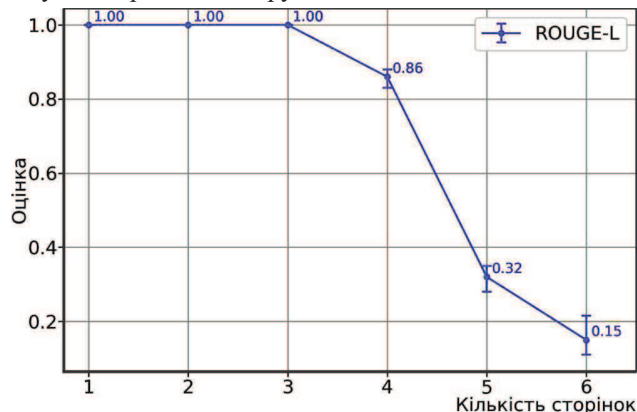


Рис. 2. Оцінювання метрикою ROUGE-L коректності оброблення набору документів зі сталим текстом та його розміром за рахунок міжрядкового інтервалу / Evaluation of the correctness of processing a set of documents with fixed text and its size by the ROUGE-L metric due to the interline interval

В експерименті з набором, де розмір документа змінювався внаслідок збільшення міжрядкових відступів, модель демонструвала стабільну роботу для документів обсягом до трьох сторінок (рис. 2). Починаючи з чотирьох сторінок, частота появи галюцинацій почала стрімко зростати, що свідчить про втрату контексту під час оброблення великих документів.

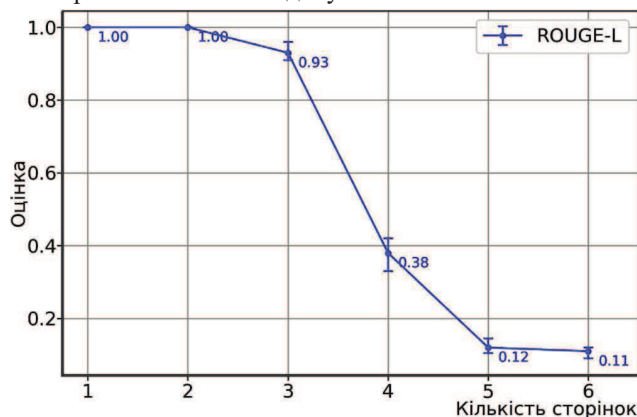


Рис. 3. Оцінювання метрикою ROUGE-L коректності оброблення набору документів зі сталим текстом та його розміром за рахунок порожніх сторінок / Evaluation of the correctness of processing a set of documents with fixed text and its size at the expense of blank pages by the ROUGE-L metric

Аналогічні результати спостерігалися у наборі документів, де розмір документа змінювався внаслідок додавання порожніх сторінок (рис. 3). Галюцинації відсутні під час оброблення документів до двох сторінок, але різко зростають зі збільшенням їх кількості. Вплив порожніх сторінок на частоту появи помилок був навіть більшим, ніж у разі із рівномірно заповненими сторінками тексту.

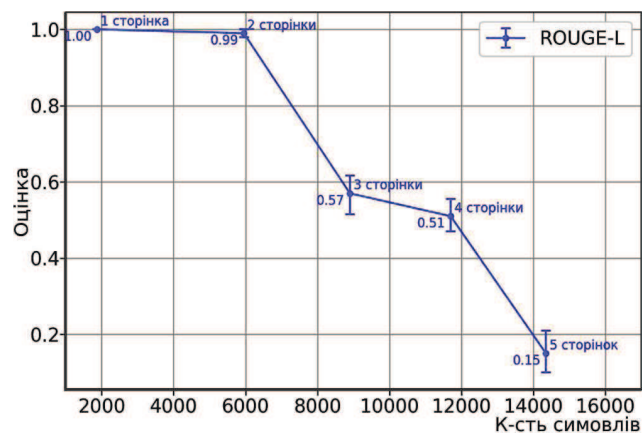


Рис. 4. Оцінювання метрикою ROUGE-L коректності оброблення набору документів зі змінним обсягом тексту та його розміром / Evaluation of the correctness of processing a set of documents with variable text volume and size using the ROUGE-L metric

Для набору документів, де змінювався і обсяг тексту, і розмір документа, перші галюцинації з'являлися за обсягу тексту близько 5000 символів на двох сторінках (рис. 4). Далі частота помилок стрімко зростала, що вказує на сильний вплив обох параметрів.

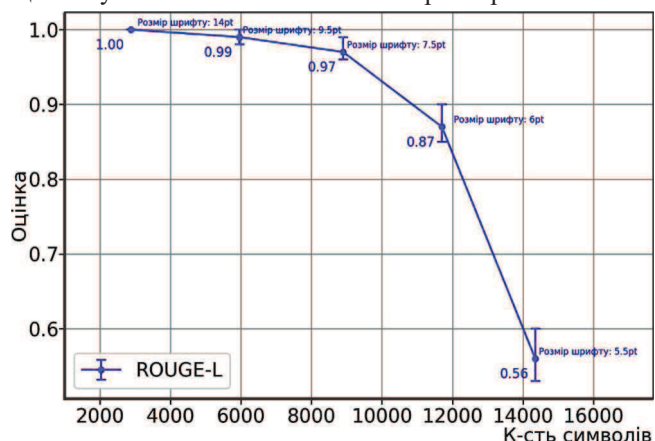


Рис. 5. Оцінювання метрикою ROUGE-L коректності оброблення набору документів зі змінним обсягом тексту та розміром шрифту / Evaluation of the correctness of processing a set of documents with variable text volume and font size using the ROUGE-L metric

Набір документів із варіацією текстового обсягу способом зменшення шрифту продемонстрував зростання галюцинацій у тексті розміром більше 5000 символів (9,5 pt) із подальшим різким зниженням метрики ROUGE-L до 0,56 у тексті розміром близько 14000 символів (5,5 pt).

Аналізуючи згадані вище результати, можемо зробити висновок, що для тексту з розміром шрифту 6 pt і більше спад оцінок метрики ROUGE-L пов'язаний переважно зі зростанням обсягу тексту. Цей висновок базується на результатах першого експериментального набору документів, де було продемонстровано, що розмір шрифту в межах 6-18 pt не спричиняє істотного зростання частоти появи галюцинацій. Однак для тексту з меншим розміром шрифту (менше 6 pt) спостерігалось значне зниження точності через комбінацію складнощів розпізнавання малих символів і впливу обсягу тексту.

Порівняння результатів експериментів демонструє, що як фізичний розмір документа (кількість сторінок), так і обсяг тексту істотно впливають на частоту появи

галюцинацій у моделі GPT-4 Vision. Для документа, що складається з чотирьох сторінок і містить текст обсягом близько 12000 символів, оцінка метрикою ROUGE-L коректності оброблення тексту становила 0,51. Це вказує на те, що збільшення обсягу тексту та кількості сторінок створює складність для моделі, що призводить до значного зростання частоти появи галюцинацій.

У разі, коли розмір документа збільшувався до чотирьох сторінок за рахунок міжрядкового інтервалу за незмінного текстового обсягу, оцінка метрикою ROUGE-L залишалася на рівні 0,86. Це вказує на те, що фізичний розмір документа також відіграє важливу роль, але його вплив є менш значним, ніж у комбінації з обсягом тексту.

Водночас для документа з цим самим обсягом тексту (12000 символів), але розташованого на одній сторінці способом зменшення розміру шрифту до 6 pt, оцінка метрикою ROUGE-L коректності оброблення тексту становила 0,87. Це свідчить про те, що і фізичний розмір документа, і обсяг тексту мають незалежний вплив на частоту появи галюцинацій, але їхній комбінований ефект призводить до ще більшого зростання частоти помилок.

Отримані результати підтверджують, що частота появи галюцинацій у моделі GPT-4 Vision істотно залежить від таких параметрів документа, як обсяг тексту, фізичний розмір документа та меншою мірою – розмір шрифту. Аналіз виявив, що найбільший вплив спостерігається за умов комбінованого збільшення обсягу тексту та кількості сторінок, а також під час роботи з текстом надто малого розміру.

На підставі отриманих даних можна запропонувати такі рекомендації для ефективного використання моделі GPT-4 Vision:

1. Мінімальний розмір шрифту – 6 pt. Текст із розміром шрифту менше 6 pt спричиняє істотне зростання частоти появи галюцинацій через складність розпізнавання дрібних символів.
2. Оброблення не більше двох сторінок A4 за раз. Результати показали, що збільшення фізичного розміру документа понад два аркуші значно ускладнює роботу моделі через втрату контексту.
3. Мінімізація порожнього простору на етапі препроцесингу. Великі незаповнені ділянки у документах сприяють появі галюцинацій, особливо у багатосторінкових файлах. Під час препроцесингу рекомендують видаляти зайві поля та порожні сторінки.
4. Обмеження оброблення тексту до 5000 символів за раз. Збільшення обсягу тексту понад цей ліміт призводить до стрімкого зростання частоти помилок, зумовленого як перевантаженням моделі, так і її схильністю до втрати контексту.

Обговорення результатів дослідження. Дослідження можливостей моделі GPT-4 Vision у розпізнаванні тексту демонструє, що модель здатна до високоякісного розпізнавання символів. Автори роботи [5] вказують, що модель GPT-4 Vision може ефективно працювати з історичними документами та полегшувати їхню цифровізацію, що узгоджується з нашими висновками. Водночас, відзначено кілька ключових обмежень моделі, зокрема обмеження щодо розміру вхідного зображення та чутливість до макета документа. У запропонованому підході виявлено, що параметри тексту, такі як щільність розміщення символів і фізичний розмір документа, істотно впливають на точність роботи моделі,

що доповнює результати і дає змогу згенерувати практичні рекомендації з використання моделі і підвищити цим самим її точність.

Дослідження [15] використовує GPT-моделі для розпізнавання тексту та автоматичного формування історичних наративів. Автори відзначають, що застосування GPT для OCR забезпечує високу точність розпізнавання символів (F1-міра 93,5 %) та значно покращує доступність історичних документів. У роботі, як і в проведеному дослідженні, встановлено, що точність розпізнавання залежить від параметрів документа, зокрема якості початкового зображення та розміщення текстових блоків. Водночас, ця робота фокусує увагу на створенні рекомендацій на підставі архівних документів, а наше дослідження визначає вплив обсягу тексту та його структури на точність оброблення моделі GPT-4 Vision.

У роботі [16] оцінено здатність моделі GPT-4 до вирішення завдань візуального розпізнавання. Автори демонструють, що використання моделі GPT-4 покращує точність класифікації зображень і підвищує точність top-1 на 7 %. У нашому дослідженні також відзначено, що модель GPT-4 Vision демонструє високу ефективність під час роботи з текстовими документами, проте експериментально встановлено, що його точність істотно залежить від параметрів документа, таких як розмір шрифту та кількість сторінок. Висновки роботи та запропонованого підходу доповнюють одне одного та можна застосувати для покращення роботи моделі GPT-4 Vision у завданнях розпізнавання тексту.

У роботі [3] проаналізовано можливість застосування моделей сімейства GPT-4 до завдань розуміння документів. Автор зазначає, що традиційні текстові моделі GPT-4 демонструють недостатню точність у завданнях розпізнавання структурованих документів, тоді як модель GPT-4 Vision забезпечує значно кращі результати за умови подачі як тексту, отриманого за допомогою OCR, так і зображень документа. Водночас у роботі виявлено значне зниження продуктивності під час оброблення довгих документів, що узгоджується з нашими висновками. На відміну від цієї роботи, ми дослідили конкретні граничні значення, за яких частота появи галюцинацій істотно зростає, та запропонували рекомендації з обмеження кількості символів і сторінок для зменшення помилок.

Комплексне дослідження OCR-можливостей моделі GPT-4 Vision, представлене в роботі [14], демонструє, що модель добре справляється з розпізнаванням латинського тексту, проте має істотні обмеження в мультимовних сценаріях та складних завданнях, таких як розпізнавання рукописних математичних виразів, таблиць та семантичних зв'язків у документах. Водночас, ця робота аналізує загальну ефективність моделі в різних OCR-завданнях, тобто наше дослідження є більш сфокусованим на проблемі галюцинацій та встановленні взаємозв'язку між параметрами документа й точністю розпізнавання.

У дослідженні [7] розглянуто підхід до цифровізації документів способом комбінування технологій OCR та великих мовних моделей LLMs (англ. *Large Language Models*). Автори демонструють, що використання LLMs для післяоброблення OCR-даних дає змогу підвищити точність розпізнавання тексту, особливо у разі, коли початкові зображення містять шуми або викривлення. Це узгоджується з нашими висновками щодо чутливості

GPT-4 Vision до макета документа та його залежності від якісного вхідного тексту. Водночас, у роботі наголошено, що без додаткової корекції LLM може не тільки виправляти, а й спотворювати текст, що підтверджує важливість аналізу параметрів документа для мінімізації появи галюцинацій.

У роботі [18] досліджено можливість використання моделей GPT-4 для корекції помилок, що виникають після OCR-розпізнавання історичних документів. Автори відзначають, що модель демонструє значне покращення у виправленні текстових помилок, проте її ефективність залежить від початкової якості розпізнаного тексту. У нашому дослідженні також було виявлено, що точність роботи моделі GPT-4 Vision знижується під час оброблення текстів із високою щільністю символів або дрібним шрифтом, що ускладнює процес корекції. Отже, результати цієї роботи доповнюють наші висновки та підтверджують потребу в урахуванні структурних параметрів документа для підвищення точності цифровізації.

У дослідженні [12] наведено метод аналізу структури документів на основі технології DLA (англ. *Document Layout Analysis*). Автори показують, що розпізнавання тексту в документах значно залежить від правильного визначення їхніх структурних компонентів, таких як блоки тексту, заголовки та таблиці. Це доповнює наші результати, згідно з якими фізичний розмір документа та щільність тексту впливають на точність роботи GPT-4 Vision. Окрім цього, виявлена уразливість моделі до змін у розміщенні тексту узгоджується з нашими висновками щодо залежності частоти появи галюцинацій від параметрів документа.

У роботі [8] досліджено особливості застосування моделей великих мов у завданнях розуміння розмовної мови. Автори показують, що галузь розуміння розмовної мови SLU (англ. *Spoken Language Understanding*), піддомену оброблення природної мови NLP (англ. *Natural Language Processing*). SLU зосереджується на обробленні та розумінні усного або голосового введення, прагнучи витягти змістовну інформацію з висловлювань користувача. Ключовим компонентом SLU є автоматичне розпізнавання мовлення ASR (англ. *Automatic Speech Recognition*), яке перетворює розмовну мову в текстову форму. У роботі окреслюються цілі та значення SLU, вказуючи на його здатність давати змогу комп'ютерам розуміти та діяти відповідно до усних настанов. Інтегруючи ASR та наступні методи розуміння мови, системи SLU сприяють розвитку більш інтуїтивної та ефективної взаємодії людини з комп'ютером. У роботі наведено короткий аналіз ключових методів та компонент, що беруть участь у SLU, вказуючи на важливість точного розпізнавання мовлення та ефективного парсингу мови для досягнення бажаних результатів.

Отже, отримані результати підтверджують актуальність і новизну цього дослідження, оскільки раніше ніхто детально не аналізував вплив параметрів текстових документів на точність роботи GPT-4 Vision. Вперше визначено граничні значення обсягу тексту та кількості сторінок, за яких зростає частота появи галюцинацій. Ці висновки доповнюють попередні роботи та дають змогу сформулювати практичні рекомендації для підвищення точності цифровізації складноструктурованих документів.

Отже, внаслідок виконаної роботи можна сформулювати такі наукову новизну та практичну значущість результатів дослідження.

Наукова новизна отриманих результатів дослідження – розроблено методику оцінювання впливу параметрів текстових документів на точність роботи моделі GPT-4 Vision, що дало змогу встановити граничні значення розміру шрифту, фізичні розміри документа та обсягу тексту, а також визначити залежність між цими параметрами і частотою появи галюцинацій.

Практична значущість результатів дослідження – запропоновану методику можна застосувати для розроблення нових методів з покращенням точності автоматизованого розпізнавання тексту у мультимодальних моделях способом впорядкування вхідних документів відповідно до визначених у роботі параметрів, а також для розроблення рекомендацій щодо попередньої підготовки документів перед їх обробленням моделлю GPT-4 Vision, що дасть змогу зменшити частоту появи галюцинацій, підвищити ефективність опрацювання документів за допомогою систем штучного інтелекту.

Висновки / Conclusions

Проаналізовано вплив параметрів документа на ефективність роботи моделі GPT-4 Vision, що дало змогу розробити рекомендації з мінімізації появи галюцинацій та підвищення точності цифровізації складних документів. За результатами проведеного дослідження можна зробити такі основні висновки:

1. Досліджено вплив розміру шрифту на частоту появи галюцинацій і встановлено, що для текстів із шрифтом понад 6 pt точність оброблення моделі здебільшого залежить від обсягу тексту, тоді як для менших шрифтів додатково впливають текстова щільність і складність розпізнавання символів.
2. На підставі запропонованого підходу до опрацювання текстових документів встановлено, що значне збільшення обсягу тексту або фізичного розміру документа призводить до втрати контексту та зростання частоти появи галюцинацій. Наприклад, для документа з 12000 символами, розташованими на чотирьох сторінках, оцінка метрикою ROUGE-L знижувалася до 0,51, тоді як в разі розміщення того ж текстового обсягу на одній сторінці (шрифт 6 pt) оцінка підвищувалася до 0,87.
3. З'ясовано, що обсяг тексту та фізичний розмір документа є незалежними, але взаємопов'язаними факторами, які впливають на точність роботи моделі.
4. Запропоновано практичні рекомендації для мінімізації появи галюцинацій під час опрацювання текстових документів за допомогою моделі GPT-4 Vision, зокрема: використання тексту з розміром шрифту не менше 6 pt; оброблення не більше двох сторінок A4 або 5000 символів за раз; зменшення порожнього простору на етапі препроцесингу.
5. Обґрунтовано, що дотримання цих рекомендацій сприятиме покращенню точності розпізнавання текстів у складноструктурованих документах та збільшенню ефективності застосування моделі GPT-4 Vision у реальних сценаріях автоматизації процесу оброблення текстової інформації.

References

1. Abdelaziz, T. A. I., & Fazil, U. (2023). Applications of integration of AI-based Optical Character Recognition (OCR) and Generative AI in Document Understanding and Processing. *Applied Research in Artificial Intelligence and Cloud Computing*, 6(11), 1–16. URL: <https://researchberg.com/index.php/araic/article/view/171>
2. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., & McGrew, B. (2023). Gpt-4 technical report. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2303.08774>

3. Borchmann, L. (2024). Notes on applicability of gpt-4 to document understanding. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2405.18433>
4. Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., & Yao, H. (2023). Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2311.03287>
5. Ghiriti, A., Göderle, W., & Kern, R. (2024). Exploring the Capabilities of GPT-4 Vision as OCR Engine. In: Antonacopoulos, A., et al. *Linking Theory and Practice of Digital Libraries. TPDL 2024. Lecture Notes in Computer Science*, article ID 15178. Springer, Cham. https://doi.org/10.1007/978-3-031-72440-4_1
6. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), article ID 248. <https://doi.org/10.1145/3571730>
7. Lee, T., Kim, G., Ahn, H., Jeong, J., Jeong, M., & Song, J. (2024). Integrating OCR and LLMs for enhanced document digitization in ERP systems. *Proceedings of the 15th International Conference on Information and Communication Technology Convergence (ICTC)*, 1650–1653. <https://doi.org/10.1109/ICTC62082.2024.10827723>
8. Li, M. (2024). Exploring the Application of Large Language Models in Spoken Language Understanding Tasks. *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, Jinzhou, China, 2024, pp. 1537–1542. <https://doi.org/10.1109/ICSECE61636.2024.10729345>
9. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*, 74–81. Association for Computational Linguistics. URL: <https://aclanthology.org/W04-1013/>
10. Lin, C.-Y., & Och, F. J. (2004). Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. *Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.3115/1218955.1219032>
11. OpenAI. (2023). GPT-4V (ision) system card. URL: <https://openai.com/research/gpt-4-vision>
12. Peña, A., Morales, A., Fierrez, J., Ortega-Garcia, J., Puente, I., Cordova, J., & Cordova, G. (2024). Continuous document layout analysis: Human-in-the-loop AI-based data curation, database, and evaluation in the domain of public affairs. *Information Fusion*, 108, article ID 102398. <https://doi.org/10.1016/j.inffus.2024.102398>
13. Raj, A., Kumar, S., Singh, P., & Patel, A. (2023). Revolutionizing data entry: An in-depth study of optical character recognition technology and its future potential. *International Journal for Research in Applied Science and Engineering Technology*, 2, 645–653. <https://doi.org/10.22214/ijraset.2023.49108>
14. Shi, Y., Peng, D., Liao, W., Lin, Z., Chen, X., Liu, C., & Jin, L. (2023). Exploring ocr capabilities of gpt-4v (ision): A quantitative and in-depth evaluation. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2310.16809>
15. Ströbel, P. B., Clematide, S., Meyer, J., & Werner, P. (2024). New "ArchAlval" Practices: Using GPT for OCR and Historical Narration of Index Cards. In: Antonacopoulos, A., et al. *Linking Theory and Practice of Digital Libraries. TPDL 2024. Lecture Notes in Computer Science*, article ID 15178. Springer, Cham. https://doi.org/10.1007/978-3-031-72440-4_18
16. Wu, W., Yao, H., Zhang, M., Song, Y., Ouyang, W., & Wang, J. (2023). GPT4Vis: what can GPT-4 do for zero-shot visual recognition?. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2311.15732>
17. Ye, J., Hu, A., Xu, H., Ye, Q., Yan, M., Xu, G., & Huang, F. (2023). Ureader: Universal ocr-free visually-situated language understanding with multimodal large language model. *arXiv preprint*. <https://doi.org/10.48550/arXiv.2310.05126>
18. Zhang, J., Haverals, W., Naydan, M., & Kernighan, B. W. (2024). Post-OCR correction with OpenAI's GPT models on challenging English prosody texts. *Proceedings of the ACM Symposium on*

D. Ya. Doskach, V. M. Khavalko

Lviv Polytechnic National University, Lviv, Ukraine

IMPACT OF TEXT DOCUMENT PARAMETERS ON THE ACCURACY OF THE MULTIMODAL GPT-4 VISION MODEL IN TEXT RECOGNITION TASKS

The impact of textual document parameters on the effectiveness of the multimodal large language model GPT-4 Vision, widely applied for digitizing documents with complex structures, has been analyzed. The study investigates the dependency of the model accuracy on such factors as font size, physical document size (number of A4 pages), and total text volume (number of characters). Particular attention is given to the problem of the model generating incorrect information ("hallucinations"), which remains a significant drawback of this approach to text digitization. Five experimental datasets of PDF documents with varying font sizes, text volumes, and page counts were developed, enabling evaluation of the individual and combined impact of these factors. The ROUGE-L accuracy metric was used to quantitatively assess the frequency of hallucinations. It has been established that hallucination frequency significantly increases in documents exceeding 5000 characters or consisting of more than two A4 pages. For instance, a four-page document containing approximately 12,000 characters had a ROUGE-L accuracy of 0.51. In contrast, a document with the same amount of text compactly placed on a single page by reducing the font size to 6 pt showed considerably higher accuracy (ROUGE-L = 0.87). These results highlight the substantial influence of physical document parameters on the accuracy of multimodal models. Practical recommendations for the optimal use of GPT-4 Vision in digitizing complex documents have been formulated as follows: the recommended minimum font size is 6 pt; the maximum processed text volume should not exceed 5000 characters or two pages at a time; empty spaces should be minimized by removing unnecessary margins and blank pages. The obtained results are practically valuable for further improving digitization algorithms for structurally complex documents and enhancing the reliability of multimodal models. The proposed evaluation methodology can also serve as a foundation for future research aimed at reducing the frequency of hallucinations in automated text recognition tasks.

Keywords: hallucination frequency; character recognition; document digitization; large language models; multimodal AI systems.