Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту
(повна назва)
Кафедра Інформатики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти перший (бакалаврський)

РОЗРОБЛЕННЯ ІНТЕЛЕКТУАЛЬНОЇ МОБІЛЬНОЇ СИСТЕМИ ДЛЯ СУПРОВОДУ ВІДВІДУВАЧІВ МУЗЕЇВ НА ОСНОВІ ТЕХНОЛОГІЙ КОМП'ЮТЕРНОГО ЗОРУ
(тема)

Виконав:
здобувач 4 року навчання,
групи ІТІНФ-22-2
Камишан П. С.
(прізвище, ініціали)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
Освітня програма Інформатика
(повна назва освітньої програми)
Керівник ст. викл. Кобилін І. О.
(посада, прізвище, ініціали)

Допускається до захисту

Завідувач кафедри інформатики _____ Кобилін О. А.
(підпис) (прізвище, ініціали)

2026 р.

Харківський національний університет радіоелектроніки

Факультет Інформаційно-аналітичних технологій та менеджменту

Кафедра Інформатики

Рівень вищої освіти перший (бакалаврський)

Спеціальність 122 Комп’ютерні науки
(код і повна назва)

Тип програми освітньо-професійна

Освітня програма Інформатика
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2026 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Камишан Поліні Сергіївні
(прізвище, ім’я, по батькові)

1. Тема роботи Розроблення інтелектуальної мобільної системи для супроводу відвідувачів музеїв на основі технологій комп’ютерного зору

затверджена наказом університету від 26 травня 2026 року № 435Ст
2. Термін подання здобувачем роботи до екзаменаційної комісії 20 червня 2026 р.
3. Вихідні дані до роботи науково-методична та науково-технічна література з питань комп'ютерного зору, методів виявлення та ідентифікації об’єктів, мультимодальних великих лінгвістичних моделей, нейромережевого синтезу мовлення та інтелектуальних систем музейного супроводу; матеріали міжнародних конференцій та наукових публікацій з тематики мультимодального аналізу зображень, порівняльного оцінювання великих лінгвістичних моделей та застосування штучного інтелекту в культурно-освітній сфері; дані інтернет-мережі; мультимодальні моделі Gemini 2.5 Flash та Gemini 3.1 Pro від Google DeepMind; мовна модель GPT-5.5 від OpenAI; платформа порівняльного рейтингування моделей LMArena Vision та академічний бенчмарк MMMU-Pro, Gradio.

4. Перелік питань, що потрібно опрацювати в роботі
1. Аналіз загального стану питання розпізнавання музейних експонатів.
2. Огляд існуючих систем і застосунків для музейного супроводу.
3. Розробка способу розпізнавання музейних експонатів.
4. Проєктування вебзастосунку музейного гіда.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) актуальність проблеми інтелектуального музейного супроводу та відсутності комплексних рішень, існуючі методи комп'ютерного зору та огляд наявних систем музейного супроводу, постановка задачі, порівняльний аналіз мультимодальних великих лінгвістичних моделей для задачі ідентифікації музейних експонатів, сформований власний еталонний набір даних, тестові експерименти розпізнавання експонатів із використанням спеціалізованого промпту та структурованого виведення, проведення експертного опитування та висновок щодо обрання моделі Gemini 2.5 Flash, схема алгоритму та діаграми взаємодії компонентів застосунку, ілюстрації роботи системи MuseGuide AI у різних сценаріях, аналіз результатів роботи застосунку.

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	13.04.2026	
2	Аналіз завдання, підбір літератури	15.04.26-17.04.26	
3	Аналіз літератури з проблеми, що вирішується	18.04.26-27.04.26	
4	Аналіз існуючих методів розпізнавання	28.04.26-29.04.26	
5	Формування кроків вирішення проблеми	30.04.26-02.05.26	
6	Програмна реалізація	03.05.26-20.05.26	
7	Аналіз отриманих результатів	21.05.26-04.06.26	
8	Оформлення пояснювальної записки	05.06.26-09.06.26	
9	Перевірка на нормоконтроль	10.06.26-19.06.26	
10	Перевірка на плагіат	11.06.26-30.06.26	
11	Рецензування	20.06.26-30.06.26	
12	Підготовка презентації та доповіді	20.06.26-30.06.26	
13	Занесення роботи в електронний архів	30.06.26-09.07.26	
14	Попередній захист кваліфікаційної роботи	30.06.26-09.07.26	

Дата видачі завдання 13 квітня 2026 р.

Здобувач _____
(підпис)

Керівник роботи _____ ст. викл. Кобилін І. О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ/ABSTRACT

Пояснювальна записка до кваліфікаційної роботи: 72 с., 8 табл., 10 рис., 30 джерел.

ВЕБЗАСТОСУНОК, ГЕНЕРАЦІЯ, РОЗПІЗНАВАННЯ, ШТУЧНИЙ ІНТЕЛЕКТ, GEMINI 2.5 FLASH, PYTHON.

Об’єктом роботи є процес автоматизованої ідентифікації музейних експонатів та формування персоналізованого екскурсійного контенту на основі мультимодального аналізу зображень.

Метою роботи є розроблення вебзастосунку MuseGuide AI для персоналізованого супроводу відвідувачів музеїв із використанням Google Gemini Vision API та синтезу мовлення засобами Microsoft Edge TTS.

Під час роботи використано: методи мультимодального аналізу зображень та порівняльного аналізу великих лінгвістичних моделей, методи нейромережевого синтезу мовлення та методи веброзробки.

Практична цінність застосунку полягає в автоматизації інформаційного супроводу відвідувачів музеїв без необхідності попередньої підготовки контенту кураторами.

Розроблено вебзастосунок, що забезпечує автоматичну ідентифікацію музейних експонатів за фотографією, генерування структурованого екскурсійного опису та озвучення результату синтезованим голосом українською мовою.

ARTIFICIAL INTELLIGENCE, GEMINI 2.5 FLASH, GENERATION, PYTHON, RECOGNITION, WEB APPLICATION.

The objective of this work is the automated identification of museum exhibits and the formation of personalized excursion content based on multimodal image analysis.

The goal of this work is to develop the MuseGuide AI web application for personalized museum visitor guidance using Google Gemini Vision API and Microsoft Edge TTS speech synthesis.

The following were used during the work: multimodal image analysis and comparative analysis of large language models, neural speech synthesis methods and web development methods.

The practical value of the application lies in automating the informational guidance of museum visitors without the need for curators to prepare content in advance.

A web application has been developed that ensures automatic museum exhibit identification from photographs, structured excursion description generation, and audio narration in Ukrainian using a synthesized voice.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

МВММ – мультимодальна велика мовна модель

млрд – мільярд

млн – мільйон

мс – мілісекунда

пкс – піксель

ШІ – штучний інтелект

API – Application Programming Interface (інтерфейс програмного застосунку для обміну даними між клієнтом і сервером)

CNN – Convolutional Neural Network (згорткова нейронна мережа)

CRNN – Convolutional Recurrent Neural Network (згорткова рекурентна нейронна мережа для розпізнавання послідовностей)

CSS – Cascading Style Sheets (каскадні таблиці стилів для оформлення вебсторінок)

JSON – JavaScript Object Notation (відкритий формат структурованого обміну даними)

K – Kilo (тисячі)

LMArena – краудсорсингова платформа порівняльного рейтингування великих мовних моделей

M – Million (мільйон)

MMMU-Pro – Massive Multidiscipline Multimodal Understanding Pro (стандартизований академічний бенчмарк оцінювання мультимодальних моделей)

MP3 – Moving Picture Experts Group Audio Layer III (формат стиснення аудіофайлу зі збереженням якості)

OCR – Optical Character Recognition (оптичне розпізнавання тексту)

RPD – Requests Per Day (запитів на день)

SDK – Software Development Kit (комплект засобів розроблення програмного забезпечення)

TTS – Text-to-Speech (синтез мовлення з тексту)

UUID – Universally Unique Identifier (уніфікований унікальний ідентифікатор)

YOLO – You Only Look Once (архітектура нейронної мережі для одноетапного виявлення об’єктів у реальному часі)

$ – долар США

% – відсоток

ВСТУП

Сучасні технології штучного інтелекту розділяються за трьома основними напрямами, якими є комп’ютерний зір, оброблення природної мови та мультимодальний аналіз.

Комп’ютерний зір являє собою вилучення структурованої інформації на основі цифрових зображень. Приміром, це може бути розпізнавання облич, виявлення транспортних засобів, аналіз медичних знімків або ідентифікація унікальних артефактів у музейних просторах.

Основне завдання мультимодальних великих мовних моделей полягає в одержанні глибокого семантичного розуміння зображених об’єктів та генерації відповідного контенту. Мета такого аналізу може бути різна, що включає як класифікацію окремих витворів мистецтва, так і формування повноцінних екскурсійних розповідей. У певному сенсі завдання мультимодального аналізу є комплексним розширенням базового розпізнавання образів. Області застосування охоплюють розроблення віртуальних асистентів, створення розумних систем навігації та проєктування вебзастосунків для персоналізованого обслуговування туристів.

Актуальність роботи полягає у необхідності створення інноваційного вебзастосунку для автоматизованого інформаційного супроводу відвідувачів культурних установ. Впровадження системи на основі нейромережевого розпізнавання дозволить швидко ідентифікувати музейні експонати за фотографіями та миттєво генерувати адаптивні екскурсійні описи із природним аудіосупроводом. Реалізація цього підходу забезпечить подолання мовних бар’єрів та суттєво підвищить загальну якість культурного досвіду користувачів.

1 АНАЛІЗ СУЧАСНОГО СТАНУ ПИТАННЯ РОЗПІЗНАВАННЯ МУЗЕЙНИХ ЕКСПОНАТІВ

1.1 Музейна культура у світі

Музейна галузь належить до тих інституційних середовищ, де традиційні підходи до організації взаємодії відвідувача з експозицією зазнають суттєвої трансформації під впливом сучасних інформаційних технологій. Протягом останніх десятиліть музеї поступово еволюціонували від пасивних сховищ культурних артефактів до динамічних освітніх та культурних просторів, що активно впроваджують цифрові інструменти для підвищення якості відвідування. Ця трансформація зумовлена як зростаючими вимогами аудиторії до персоналізованого досвіду, так і стрімким розвитком технологій штучного інтелекту, комп’ютерного зору та мобільних обчислювальних систем.

Проблема навігації та інформаційного супроводу у великих музейних комплексах є багатогранною. Більшість відвідувачів не мають достатнього часу для повноцінного ознайомлення з усією експозицією, а традиційні аудіогіди та паперові карти не забезпечують динамічної адаптації до поведінки відвідувача в реальному часі. Мовний бар’єр залишається серйозним обмеженням для міжнародних туристів, якщо система супроводу не підтримує багатомовний режим.

Зазначені проблеми сформували науковий та прикладний інтерес до розроблення інтелектуальних систем музейного супроводу, які поєднують методи комп’ютерного зору, розпізнавання об’єктів та оптичного розпізнавання тексту. Відповідні підходи до побудови таких систем розглядаються у роботах [1–3].

Актуальність роботи визначається, передусім, відсутністю комплексних рішень, що ефективно поєднують методи комп’ютерного зору та оптичного розпізнавання тексту у єдину навігаційну систему. Незважаючи

на значний прогрес у кожній із зазначених технологічних областей окремо, їх ефективне поєднання у контексті музейного супроводу залишається відкритим науковим питанням [4, 5]. Більшість наявних рішень або обмежуються виключно візуальним розпізнаванням об’єктів без урахування текстової семантики, або покладаються на статичні бази даних без можливості динамічного оновлення інформації у реальному часі.

Окремої уваги заслуговує проблема ефективності оброблення даних на мобільних пристроях. Музейний застосунок функціонує в умовах обмежених обчислювальних ресурсів порівняно зі стаціонарними системами, що висуває специфічні вимоги до архітектурних рішень та алгоритмічної складності застосовуваних методів. Використання легковагових архітектур нейронних мереж, квантизація моделей та оптимізований мобільний інференс є обов’язковими умовами для досягнення прийнятної продуктивності на сучасних смартфонах середнього класу.

Середовище музейної зали характеризується складними умовами зйомки, а саме неоднорідним і часто штучним освітленням, наявністю відблисків на захисному склі вітрин, частковим перекриттям об’єктів іншими відвідувачами та значною варіативністю кутів зйомки. На відміну від контрольованого лабораторного середовища, де моделі комп’ютерного зору демонструють близькі до ідеальних показники точності, реальні умови музею суттєво знижують ефективність стандартних підходів. Це вимагає окремої уваги до методів аугментації даних при навчанні моделей та застосування технік підвищення стійкості до спотворень вхідних зображень.

Слід також зазначити, що музейна галузь характеризується специфічною природою об’єктів розпізнавання. На відміну від типових задач комп’ютерного зору, до яких належить виявлення автомобілів, людей або побутових предметів музейні експонати часто є унікальними або малосерійними об’єктами, для яких практично неможливо зібрати репрезентативний навчальний набір даних у традиційному розумінні. Картина в єдиному екземплярі, рідкісний артефакт або скульптура

невідомого майстра не мають численних аналогів у відкритих базах зображень. Саме тому підходи, що ґрунтуються на навчанні без прикладів або мультимодальному зіставленні з текстовими описами, є принципово важливими для цієї предметної галузі.

З огляду на це, загальний стан питання характеризується наявністю значного науково-практичного потенціалу у галузі інтелектуальних систем музейного супроводу при одночасній відсутності комплексних рішень. Варто також зазначити, що впровадження інтелектуальних систем супроводу відповідає загальносвітовій стратегії цифрової трансформації культурно-освітніх інституцій. За даними Міжнародної ради музеїв, культурні заклади, що активно впроваджують цифрові технології взаємодії з відвідувачами, демонструють суттєве зростання відвідуваності та підвищення рівня задоволеності аудиторії [6].

1.2 Розвиток галузі музейного супроводу

1.2.1 Прогрес у сфері технологій комп’ютерного зору

Розвиток методів глибокого навчання протягом останнього десятиліття докорінно змінив підходи до розв’язання задач комп’ютерного зору. Фундаментальним прискорювачем цього прогресу стала поява згорткових нейронних мереж, які продемонстрували здатність автоматично виявляти ієрархічні візуальні ознаки без необхідності ручного конструювання дескрипторів. Детальний огляд методів класифікації у задачах комп’ютерного зору, включаючи згорткові нейронні мережі та їх модифікації, наведено у роботі [7].

Подальший розвиток архітектур виявлення об’єктів призвів до появи двох концептуально різних парадигм, порівняльна схема яких наведена на рисунку 1.1. Одноетапні детектори, найвідомішим представником яких є архітектура YOLO, реалізують парадигму єдиного прямого проходу крізь

мережу, що одночасно генерує координати обмежувальних рамок та їх класові ймовірності [4]. Ця особливість забезпечує значно вищу швидкість оброблення при прийнятному рівні точності. Двоетапні детектори сімейства R-CNN та Mask R-CNN, розширили задачу виявлення до рівня сегментації екземплярів [5].

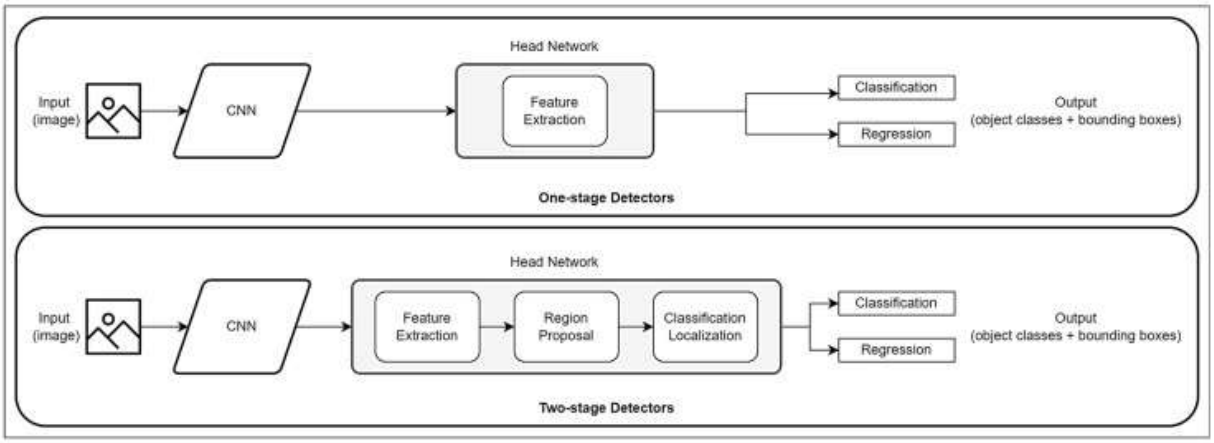

Рисунок 1.1 – Порівняння архітектур одноетапних та двоетапних детекторів об’єктів

Паралельно з розвитком методів виявлення відбувалася еволюція систем оптичного розпізнавання тексту. Принциповим кроком вперед стала архітектура CRNN, яка поєднує згорткові шари для вилучення візуальних ознак із рекурентними шарами для моделювання послідовнісних залежностей між символами [8]. Механізм тимчасової класифікації з’єднань усуває необхідність посимвольного сегментування рядка, що суттєво підвищує стійкість до варіативності шрифтів та просторових спотворень.

Революційним кроком у розвитку штучного інтелекту стала поява трансформерної архітектури та заснованих на ній великих мовних моделей. Механізм уваги дозволяє моделі динамічно зосереджуватися на найбільш релевантних частинах вхідних даних. Порівняльний аналіз сучасних трансформерних архітектур у задачах оброблення природної мови наведено у роботі [9].

Логічним продовженням розвитку великих мовних моделей стала поява мультимодальних моделей, здатних одночасно обробляти інформацію різної природи – зображення, текст, аудіо. Такі системи формують спільний латентний простір для візуальних та текстових представлень (рис. 1.2), що уможливлює семантичну оцінку подібності між зображеннями і текстовими запитами.

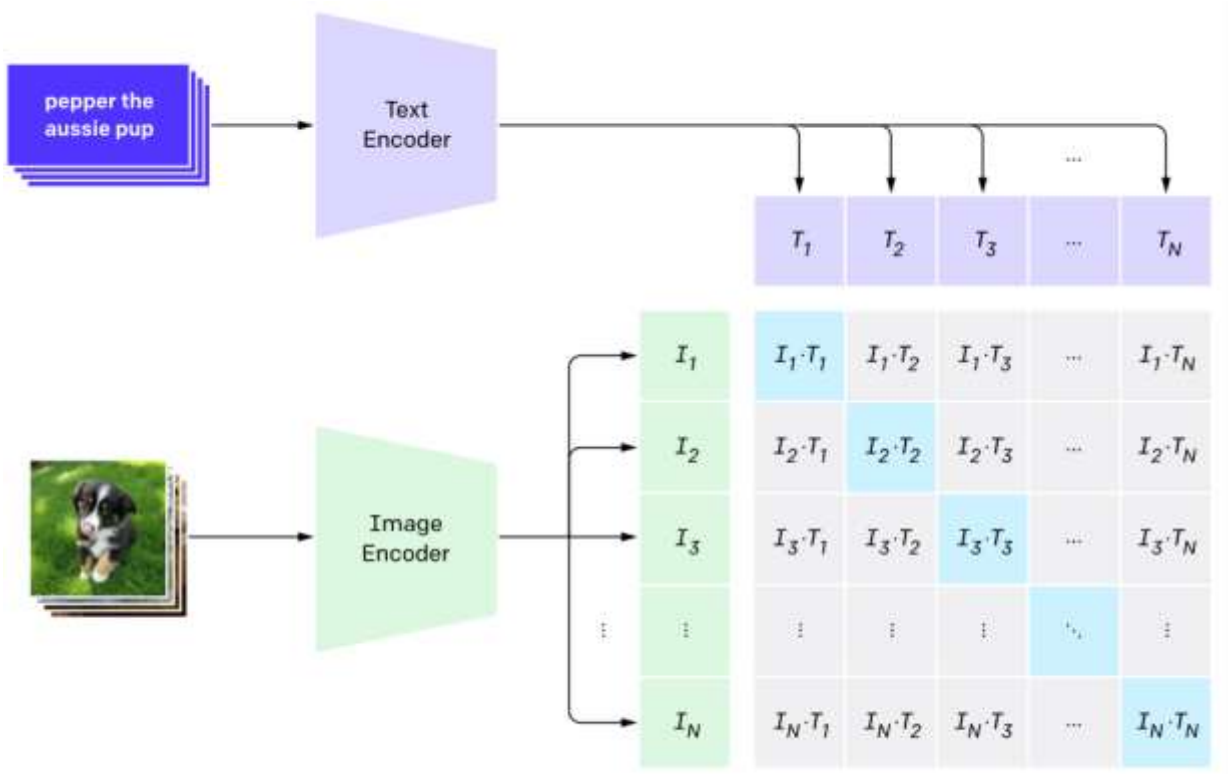

Рисунок 1.2 – Принцип роботи мультимодальної моделі на основі контрастивного навчання

Варто відзначити і прогрес у методах навчання моделей комп’ютерного зору. Підхід перенесеного навчання, що передбачає донавчання попередньо натренованих моделей на специфічних задачах, суттєво знизив вимоги до обсягу розміченого навчального матеріалу та обчислювальних ресурсів. Це є надзвичайно важливим для музейного контексту, де колекції постійно поповнюються новими унікальними об’єктами, а підготовка масштабного

розміченого набору даних для кожного конкретного музею є практично нереалістичним завданням.

Суттєвим кроком у цьому напрямі стала концепція навчання з нульовою кількістю прикладів, яка дозволяє розпізнавати об’єкти класів, що були відсутні у навчальній вибірці. Мультимодальні моделі, навчені на великих наборах пар «зображення-текстовий опис», здатні зіставляти нове зображення з текстовим запитом і знаходити семантично близькі відповідності без жодного прикладу цільового класу. Це відкриває перспективу застосування у музейному контексті без необхідності розроблення спеціалізованих класифікаторів для конкретної установи. Практична значущість такого підходу підтверджується тим, що провідні хмарні платформи – зокрема Google Vertex AI та Microsoft Azure AI Vision вже інтегрують подібні можливості у свої комерційні продукти, орієнтовані зокрема на культурну спадщину.

Суттєвих змін зазнали і методи структурованого виведення у великих мовних моделях. Якщо ранні генеративні моделі повертали виключно довільний текст, то сучасні підходи дозволяють специфікувати точну схему відповіді у вигляді типізованого JSON-об’єкта. Це означає, що модель не лише генерує опис, а й автоматично заповнює визначені поля: назву об’єкта, автора, часовий період, матеріал тощо – з гарантованою відповідністю очікуваному формату. Такий підхід усуває необхідність у крихкому парсингу вільного тексту та суттєво спрощує інтеграцію мовних моделей у практичні застосунки. Зокрема, Google Gemini API підтримує механізм types.Schema, що дозволяє задати структуру відповіді безпосередньо у параметрах запиту. Завдяки цьому розробник отримує не неструктурований текстовий відгук, а готовий словник Python із заздалегідь відомими ключами, що значно підвищує надійність системи у виробничому середовищі.

1.2.2 Застосування технологій штучного інтелекту в музейній сфері

Проникнення технологій штучного інтелекту у музейну галузь відбувається по кількох взаємопов'язаних напрямах. Першим і найбільш розвиненим напрямом є системи автоматичної ідентифікації та каталогізації об'єктів. Провідні музеї світу, серед яких Метрополітен-музей у Нью-Йорку, Рейксмюзеум в Амстердамі та Національна галерея в Лондоні, впроваджують системи комп'ютерного зору для автоматичного визначення авторства, датування та стилістичної класифікації творів мистецтва. Такі системи навчаються на масштабних наборах даних, що містять мільйони зображень художніх творів із верифікованими метаданими, і здатні виявляти стилістичні подібності між творами різних авторів та епох.

Другим напрямом є системи персоналізованих рекомендацій та адаптивного контенту. Застосовуючи методи колаборативної та контентної фільтрації, музейні цифрові платформи формують персоналізовані рекомендації щодо черговості перегляду об'єктів та додаткових матеріалів для ознайомлення відповідно до інтересів конкретного відвідувача. Зростання популярності великих мовних моделей відкрило новий вектор у цьому напрямі – використання ВММ для відповіді на запитання відвідувачів у форматі природного діалогу безпосередньо під час огляду експозиції.

Третім значущим напрямом є системи доповненої реальності, які накладають інформаційні шари безпосередньо поверх зображення реального експоната через екран мобільного пристрою. Це дозволяє демонструвати анімовані реконструкції артефактів, порівнювати поточний стан об'єкта з його первісним виглядом, а також показувати приховані деталі, недоступні при безпосередньому огляді. Четвертим напрямом є аналітика поведінки відвідувачів та оптимізація музейного простору: системи комп'ютерного зору в анонімізованому режимі відстежують траєкторії руху відвідувачів та оцінюють середній час перебування перед кожним експонатом.

Окремим перспективним напрямом є застосування синтезу мовлення для генерування персоналізованого аудіосупроводу в реальному часі. На відміну від традиційних аудіогідів, що містять наперед записані фрагменти мовлення з фіксованим текстом, системи на основі нейромережевого синтезу здатні озвучувати довільний текст, сформований безпосередньо під час роботи застосунку. Це відкриває можливість адаптувати не лише зміст, а й стиль подання матеріалу: лаконічний для поспішаючого відвідувача, розгорнутий і деталізований для зацікавленого глядача. Сучасні системи синтезу мовлення на основі нейронних мереж, такі як Microsoft Neural TTS, забезпечують природне звучання та правильну інтонацію навіть для спеціалізованих термінів у галузі мистецтва та культури.

Водночас аналіз наукової літератури виявляє суттєвий методологічний розрив: переважна більшість систем вирішує лише одну з зазначених задач, не забезпечуючи їх органічної інтеграції. Актуальним залишається завдання побудови комплексної системи, що поєднує виявлення та розпізнавання об’єктів із генеруванням персоналізованого контенту [10].

1.3 Огляд існуючих систем музейного супроводу

Аналіз існуючих систем музейного супроводу дозволяє виявити незаповнену нішу для комплексного рішення. Bloomberg Connects – застосунок, розгорнутий у понад 60 провідних культурних установах світу – базується на статичному контенті і не має механізмів динамічної персоналізації.

Платформа Smartify реалізує функцію розпізнавання художніх творів за допомогою камери смартфона. Принциповим обмеженням платформи є відсутність механізмів навігації та маршрутизації: система відповідає на запит, але не розв’язує задачу генерування персоналізованого контенту [11].

Застосунок izi.TRAVEL є відкритою платформою для створення аудіогідів. Попри широке розповсюдження, він залишається системою з фіксованими маршрутами без можливості адаптації до персональних інтересів відвідувача та без використання комп'ютерного зору.

Академічна система PEACH реалізує концепцію адаптивного музейного гіда, що будує модель інтересів відвідувача. Проте вона вимагає значної апаратної інфраструктури та не використовує комп'ютерний зір для безпосереднього розпізнавання експонатів.

З практичної точки зору важливим є також досвід застосунків, що реалізують концепцію «розумного аудіогіда» на основі розпізнавання мовлення та діалогових інтерфейсів. Такі системи дозволяють відвідувачу ставити запитання про експонат у довільній формі та отримувати відповідь у режимі природного діалогу. Однак більшість реалізованих прототипів покладаються на попередньо підготовлену базу знань і не здатні генерувати описи для об'єктів, відсутніх у базі.

У роботах, присвячених автоматичній класифікації художніх творів, досягається точність розпізнавання понад 85% на тестових вибірках [12]. Принципово важливим є досвід гібридних систем, що поєднують комп'ютерний зір та оптичне розпізнавання тексту. У роботі [1] представлено порівняльний аналіз: окремий модуль оптичного розпізнавання тексту демонстрував точність лише 61,5%, тоді як гібридний підхід забезпечував точність 100%. В іншій роботі [2] запропонована концепція багатокритеріального ранжування релевантності інформаційних блоків у мобільних системах супроводу. Архітектура системи персоналізованої навігації з семантичною розміткою об'єктів розглянута у роботі [3].

Загалом, аналіз існуючих комерційних та академічних рішень демонструє, що попри значні успіхи в окремих технологічних напрямах, досі бракує універсальної платформи. Більшість наявних розробок зосереджені або виключно на візуальному розпізнаванні, або на наданні статичної довідкової інформації без можливості динамічного супроводу.

Порівняльна характеристика розглянутих систем наведена у таблиці 1.1.

Таблиця 1.1 – Порівняльна характеристика існуючих систем музейного супроводу

Система	Ключові можливості та обмеження
Bloomberg Connects	Статичний аудіо- та відеоконтент, структурований за залами. Відсутня персоналізація та компоненти комп’ютерного зору.
Smartify	Розпізнавання художніх творів за зображенням. Відсутнє генерування персоналізованого контенту та синтез мовлення.
izi.TRAVEL	Геолоковані аудіогіди, фіксовані маршрути. Без персоналізації та комп’ютерного зору.
PEACH (академічна)	Адаптивна персоналізація. Вимагає апаратної інфраструктури, без комп’ютерного зору.

1.4 Технологічні можливості для вирішення задачі

Вибір технологічного стека є одним із визначальних рішень при проєктуванні системи з компонентами машинного навчання. Для інтелектуальної системи музейного супроводу цей вибір детермінується необхідністю оброблення зображень у реальному часі, обмеженнями мобільного обчислювального середовища та простотою розгортання для кінцевого користувача.

У галузі фреймворків для розроблення вебзастосунків домінуючими підходами є нативна розробка для платформ iOS та Android, а також кросплатформенні рішення – Flutter від Google та React Native від Meta. Flutter забезпечує продуктивність, близьку до нативної, та широкі можливості для інтеграції з камерою пристрою. React Native має велику екосистему бібліотек та підтримку гарячого перезавантаження коду.

Вибір між нативним та кросплатформенним підходом суттєво впливає на доступність кінцевого продукту. Нативна розробка для iOS (Swift) та Android (Kotlin) забезпечує найвищу продуктивність та повний доступ до апаратних можливостей пристрою, однак вимагає підтримки двох незалежних кодових баз та відповідно вдвічі більших ресурсів на розроблення та тестування. Кросплатформенні рішення, навпаки, дозволяють охопити обидві платформи з єдиної кодової бази, що значно скорочує час виходу продукту на ринок. З огляду на прототипний характер системи MuseGuide AI та необхідність максимально широкого охоплення аудиторії, кросплатформенний підхід є обґрунтованим вибором. Варто підкреслити, що використання сучасних кросплатформенних фреймворків дозволяє ефективно реалізувати асинхронне отримання відеопотоку на рівні абстракції операційної системи. Це мінімізує затримки під час передачі кадрів до модулів комп’ютерного зору, що є критично важливим для мобільних пристроїв середнього цінового сегмента, які мають обмежені ресурси локального графічного прискорення.

У контексті задачі виявлення об’єктів архітектура YOLOv8, розроблена компанією Ultralytics, є фактичним стандартом для систем, що вимагають балансу між точністю та швидкістю оброблення в реальному часі [4]. YOLOv8 побудована на концепції виявлення без опорних точок, що спрощує процес навчання та підвищує точність локалізації об’єктів порівняно з попередніми версіями архітектури. Уніфікований програмний інтерфейс бібліотеки Ultralytics дозволяє виконувати навчання, оцінювання та експорт моделі в єдиному середовищі, а підтримка формату ONNX забезпечує

сумісність із середовищами розгортання TensorFlow Lite та CoreML для мобільних платформ. YOLOv8 доступна у кількох варіантах розміру – від YOLOv8n до YOLOv8x; порівняльні характеристики варіантів наведені у таблиці 1.2.

Таблиця 1.2 – Порівняльні характеристики варіантів архітектури YOLOv8 на наборі даних COCO [13]

Модель	Розмір (пкс)	mAP val 50-95	Speed CPU ONNX (мс)	Speed A100 TensorRT (мс)	Params (млн)	FLOPs (млрд)
YOLOv8n	640	37,3	80,4	0,99	3,2	8,7
YOLOv8s	640	44,9	128,4	1,20	11,2	28,6
YOLOv8m	640	50,2	234,7	1,83	25,9	78,9
YOLOv8l	640	52,9	375,2	2,39	43,7	165,2
YOLOv8x	640	53,9	479,1	3,53	68,2	257,8

Аналіз обчислювальних метрик, наведених у таблиці 1.2, дозволяє зробити висновок щодо недоцільності прямого використання важких моделей класу YOLOv8x або YOLOv8l безпосередньо на мобільних процесорах (CPU) через критичне зростання часу інференсу (понад 375 мс). Для забезпечення роботи інтелектуальної системи супроводу в реальному часі найефективнішим компромісом є квантизація легковагових моделей типу YOLOv8n або YOLOv8s у 8-бітний формат (INT8). Такий підхід дає змогу суттєво скоротити обсяг оперативної пам’яті, що споживається мобільним пристроєм, та оптимізувати витрату заряду акумулятора під час тривалого використання програми в експозиційних залах.

Для оптичного розпізнавання тексту основним інструментом залишається Tesseract – відкритий рушій, що підтримується компанією

Google та демонструє високу точність на структурованому тексті. Альтернативним рішенням є нейромережеві моделі на основі архітектури CRNN, яка перевершує Tesseract у сценаріях з нестандартними шрифтами та неоднорідним фоном [8]. Вибір на користь комбінованих методів розпізнавання інформаційних експлікацій біля експонатів зумовлений нестабільністю світлотехнічних умов у реальних музейних просторах. Інтеграція алгоритмів попереднього цифрового оброблення візуальних даних, зокрема бінаризації, усунення оптичних відблисків та геометричної корекції перспективних спотворень, дозволяє мінімізувати деструктивний вплив зовнішніх чинників і підвищити точність розпізнавання семантичного вмісту.

Бібліотека edge-tts забезпечує доступ до нейромережевих голосів Microsoft Azure без придбання платної підписки і підтримує широкий спектр мов, зокрема українську. Ключовою перевагою є асинхронне виконання на основі asyncio, що дозволяє генерувати аудіофайл паралельно з іншими операціями, не блокуючи основний потік обробки. З огляду на вимогу безкоштовного розгортання системи та необхідність природного звучання українською мовою, edge-tts є оптимальним вибором для системи MuseGuide AI.

Платформа Hugging Face Spaces надає безкоштовне середовище для хмарного розгортання застосунків на основі Gradio та Streamlit з підтримкою GPU-прискорення для платних тарифів. Це дозволяє зробити систему загальнодоступною без потреби у власній серверній інфраструктурі. Альтернативою є локальне розгортання на музейних стаціонарних пристроях або планшетах, де застосунок запускається у режимі kiosk безпосередньо у приміщенні музею.

Порівняльна характеристика ключових технологічних компонентів системи наведена у таблиці 1.3.

Таблиця 1.3 – Порівняльна характеристика ключових технологічних компонентів системи

Компонент системи	Розглянуті варіанти та особливості вибору
Мобільна розробка	Flutter (Dart) – кросплатформенність, висока продуктивність, підтримка нативної камери. React Native – велика екосистема бібліотек.
Виявлення об’єктів	YOLOv8 – оптимальний баланс точності та швидкості, підтримка ONNX-експорту. Mask R-CNN – вища точність сегментації, але більша затримка.
Розпізнавання тексту	CRNN – стійкість до складних умов зйомки. Tesseract – висока точність на структурованому тексті. Хмарні API – найвища точність із мережевою затримкою.
Синтез мовлення	edge-tts – нейромережеві голоси без платної підписки, асинхронне виконання. Google TTS – простота інтеграції.
Мультимодальна модель	Google Gemini Vision API – структуроване виведення через types.Schema, обробка зображень та тексту в єдиному запиті. Підтримка SDK для Python.

1.5 Постановка задачі

Автоматизована ідентифікація музейних експонатів та формування персоналізованого екскурсійного контенту є актуальним завданням у контексті стрімкого розвитку мультимодальних великих лінгвістичних моделей. Наявні системи музейного супроводу або обмежуються виключно

візуальним розпізнаванням без генерування пояснень, або покладаються на статичні бази знань. Тому ставиться задача розробити вебзастосунок, що дозволить ідентифікувати музейні експонати на фотографії, генерувати структурований екскурсійний опис та озвучувати результат синтезованим голосом без жодної попередньої підготовки контенту кураторами.

Об’єктом роботи є процес автоматизованої ідентифікації музейних експонатів та формування персоналізованого екскурсійного контенту на основі мультимодального аналізу зображень.

Метою роботи є розроблення вебзастосунку MuseGuide AI для персоналізованого супроводу відвідувачів музеїв із використанням Google Gemini Vision API та синтезу мовлення засобами Microsoft Edge TTS.

Для досягнення мети сформульовано такі завдання:

– провести аналіз сучасного стану та тенденцій розроблення інтелектуальних систем для музейного супроводу;

– здійснити огляд існуючих систем і застосунків та виявити незаповнену нішу для комплексного рішення;

– провести порівняльний аналіз провідних мультимодальних великих лінгвістичних моделей для задачі ідентифікації музейних експонатів;

– сформувати власний еталонний набір даних та оцінити якість генерованого контенту методом експертного оцінювання;

– спроєктувати архітектуру системи;

– реалізувати та верифікувати вебзастосунок.

2 РОЗРОБКА СПОСОБУ РОЗПІЗНАВАННЯ МУЗЕЙНИХ ЕКСПОНАТІВ

2.1 Огляд МВММ для вирішення задачі

Мультимодальні великі мовні моделі є сучасним класом нейромережевих систем, які успішно поєднують опрацювання природної мови з можливістю сприйняття та глибокої інтерпретації інших типів даних, зокрема зображень, аудіовізуальних потоків та просторової інформації. На відміну від класичних мовних моделей, які обмежені виключно текстовим вводом, МВММ здатні одночасно отримувати візуальні матеріали та текстові запити. Це дозволяє їм виконувати надзвичайно складні комплексні задачі. До таких задач належать контекстне підписування зображень, формування розгорнутих відповідей на запитання за візуальним вмістом, точна просторова ідентифікація об’єктів, а також генерування високоструктурованих аналітичних описів мультимедійного контенту. Еволюція цих систем відображає концептуальний перехід від вузькоспеціалізованих алгоритмів комп’ютерного зору до універсальних когнітивних архітектур, здатних до комплексного семантичного аналізу [14, 15].

З архітектурного погляду фундаментальною особливістю МВММ є механізм семантичного вирівнювання модальностей. У цьому процесі модуль візуального енкодера, який переважно будується на базі архітектури Vision Transformer, перетворює вхідне зображення у послідовність щільних ознакових векторів. Далі ці вектори проєктуються у спільний багатовимірний латентний простір разом із закодованими текстовими токенами. Завдяки такому математичному перетворенню лінгвістичний декодер трансформерного типу отримує здатність одночасно опрацьовувати як візуальні, так і текстові ознаки через механізми перехресної уваги. Узгодження різних типів представлень у єдиному векторному просторі

забезпечує високий рівень релевантності та семантичної точності згенерованих відповідей [16]. Розширення базової архітектури додатковими проєкційними шарами дозволяє сучасним моделям ефективно долати семантичний розрив між низькорівневими піксельними даними та високорівневими лінгвістичними концептами.

Для об’єктивної кількісної оцінки аналітичних можливостей MBMM застосовуються стандартизовані бенчмарки. Найбільш авторитетним академічним інструментом на сьогодні є набір масштабного мультидисциплінарного тестування мультимодального розуміння під назвою MMMU-Pro. Цей комплексний бенчмарк охоплює понад одинадцять з половиною тисяч питань університетського рівня складності з тридцяти різних наукових дисциплін. Окремий блок цього тестування присвячений глибокому мистецькому аналізу, природничим наукам та інтерпретації багатих на деталі технічних схем. Паралельно із суворим академічним оцінюванням активно функціонує краудсорсингова платформа LMArena. У червні 2024 року ця платформа запустила окремий глобальний рейтинг Arena Vision, який формується на основі мільйонів сліпих попарних порівнянь відповідей нейромереж реальними користувачами. Саме синергетичне поєднання суворих академічних результатів MMMU-Pro та користувацьких оцінок Arena Vision формує найбільш об’єктивний орієнтир для вибору оптимальної моделі під конкретні прикладні задачі [17].

Лінійка моделей Google Gemini була первинно представлена наприкінці 2023 року та пройшла серію суттєвих архітектурних оновлень протягом наступних років. На сьогодні вона утворює найбільш розгалужений та технологічно довершений мультимодальний стек серед усіх глобальних хмарних провайдерів. Флагманська версія Gemini 3.1 Pro була випущена у лютому 2026 року і продемонструвала безпрецедентні результати розпізнавання. Модель досягла показника у 83,6 відсотка на бенчмарку MMMU-Pro та посіла беззаперечне перше місце у рейтингу Arena Vision з показником Elo близько 1490, очоливши більшість ключових світових

метрик [18, 19]. Водночас для цілей поточного прикладного дослідження принципове значення відіграє версія Gemini 2.5 Flash. Ця архітектура спеціально оптимізована для забезпечення найкращого балансу між високою швидкістю генерації, точністю розпізнавання та економічною доцільністю використання. Важливою перевагою цієї моделі є підтримка надвеликого контекстного вікна обсягом один мільйон токенів та здатність до нативного опрацювання різних форматів вводу. Крім того, система забезпечує надійний механізм примусового структурованого виведення даних через спеціалізовані схеми валідації, що є критично важливою вимогою для програмної інтеграції і часто реалізується зі значними похибками у продуктах конкурентів [18].

Провідним технологічним конкурентом у сфері глибокого мультимодального аналізу виступає розробка GPT-5.5 від компанії OpenAI, яка стала доступною для розробників у квітні 2026 року. Ця модель вирізняється винятковою здатністю до детального аналізу складного мистецького та культурно-історичного контенту. Такі високі показники обумовлені цілеспрямованим включенням величезних масивів мистецтвознавчих даних у корпуси для попереднього тренування нейромережі. Архітектура GPT-5.5 також забезпечує стабільну підтримку режиму жорсткого структурування вихідних даних та володіє розширеним контекстним вікном обсягом понад мільйон токенів [20]. Проте значно вища вартість використання програмного інтерфейсу суттєво обмежує можливості для широкого комерційного масштабування прикладних рішень, особливо при порівнянні з економічно оптимізованими аналогами від інших технологічних компаній.

Вагому позицію на ринку займає серія моделей Claude від лабораторії Anthropic, зокрема версії Sonnet 4.6 та Opus 4.6. Ці нейромережі відомі виключно точним дотриманням багаторівневих системних інструкцій та демонструють лідируючі позиції у специфічних задачах з аналізу складної документації. За даними тестувань початку 2026 року старша модель цієї лінійки стабільно утримує високі показники точності розпізнавання [21].

Представлена серія виступає оптимальним технічним вибором для програмних сценаріїв з абсолютним пріоритетом мінімізації ризику генерації хибних фактів. Водночас цінова політика провайдера робить масове використання цих інструментів для обробки великих потоків візуальних даних менш рентабельним порівняно з іншими доступними альтернативами.

У стрімко зростаючому сегменті відкритих моделей з публічно доступними вагами домінують кілька ключових розробок. Особливе місце серед них посідає архітектура LLaMA 3.2 Vision від корпорації Meta, яка представлена у кількох варіантах розмірності параметрів і поширюється за ліцензією з дозволом на комерційну інтеграцію. На сьогодні ця модель залишається повноцінним рішенням для проєктів із максимально суворими вимогами до захисту конфіденційних даних або для систем, які змушені функціонувати в умовах повної ізоляції від глобальної мережі. Така автономність досягається завдяки можливості локального розгортання нейромережі безпосередньо на апаратних потужностях замовника. Ще однією помітною відкритою розробкою є Qwen2-VL від Alibaba Cloud, яка демонструє неабияку ефективність у процесах багатомовного оптичного розпізнавання символів. Однак у вузькоспеціалізованому контексті атрибуції класичних європейських музейних об’єктів ці специфічні лінгвістичні переваги не відіграють вирішальної ролі [22].

Процес автоматизованої ідентифікації музейних експонатів за допомогою комп’ютерного зору є комплексною проблемою, яка складається з низки самостійних етапів. Насамперед система повинна здійснити точну класифікацію загального типу культурного об’єкта. Після успішної первинної категоризації алгоритм переходить до встановлення авторства твору, визначення історичного періоду його створення та розпізнавання специфічних матеріалів або технік виконання. Завершальним етапом цієї аналітичної послідовності є генерація зв’язного та інформативного текстового опису, витриманого у відповідному академічному чи екскурсійному стилі. Така багатовимірна сукупність вимог формує дуже

жорсткий набір критеріїв до вибору базової нейромережевої платформи. Обрана модель повинна мати глибоке мистецтвознавче розуміння контексту та бездоганно підтримувати структуроване виведення даних у форматі JSON. Останній фактор виступає фундаментальною вимогою для забезпечення надійної програмної інтеграції інтелектуального модуля у серверну архітектуру застосунку. Крім того, не менш важливими параметрами залишаються мінімальна затримка відповіді сервера та економічно обґрунтована вартість обробки кожного запиту. Комплексне зіставлення характеристик провідних світових систем за всіма переліченими критеріями наведено у таблиці 2.1.

Таблиця 2.1 – Порівняльна оцінка MBMM для задачі ідентифікації об’єктів (дані MMMU-Pro та LMArena Vision, 2025–2026 рр.)

Модель	Провайдер	MMMU-Pro, %	Arena Vision Elo	JSON-схема
Gemini 3.1 Pro	Google	83,6	~1 490	Так (types.Schema)
Gemini 3 Flash	Google	80,0	1 284	Так (types.Schema)
Gemini 2.5 Flash	Google	~72	–	Так (types.Schema)
GPT-5.5	OpenAI	н/д	Топ-3 (2026)	Так (JSON mode)
GPT-4o	OpenAI	69,1	Топ-3 (2024)	Так (JSON mode)
Claude Opus 4.5	Anthropic	74,0	–	Частково
LLaMA 3.2 Vision 90B	Meta (open)	~65	–	Обмежено

Важливою складовою вибору технологічного стека є вартість використання API. Більшість хмарних MBMM тарифікуються за кількістю

оброблених токенів: вхідних і вихідних. Зведену порівняльну таблицю цін на основні МВММ (станом на 2025–2026 рр.) наведено у таблиці 2.2.

Таблиця 2.2 – Порівняльна вартість токенів провідних МВММ (станом на 2025–2026 рр.)

Модель	Вхідні, $/1M	Вихідні, $/1M	Безкоштовний рівень
Gemini 2.5 Flash	$0,30	$2,50	15 RPM / 1500 RPD
Gemini 3.1 Pro (≤200K)	$2,00	$12,00	2 RPM / 50 RPD
GPT-5.5	$5,00	$30,00	Ні
GPT-4o (legacy)	$2,50	$10,00	Ні
Claude Sonnet 4.6	$3,00	$15,00	Ні
Claude Opus 4.6	$5,00	$25,00	Ні
LLaMA 4 Maverick (API)	$0,15	$0,60	Так (обмежено)

З-поміж розглянутих рішень для аналізу обрано такі моделі, як Gemini 2.5 Flash, Gemini 3.1 Pro та GPT-5.5. Визначальним критерієм відбору стала доступність безкоштовного режиму використання: Gemini 2.5 Flash і Gemini 3.1 Pro доступні через Google AI Studio в межах безкоштовного рівня, тоді як GPT-5.5 доступна через інтерфейс ChatGPT без необхідності оплати API-викликів, що дозволяє проводити тестування без додаткових фінансових витрат. Не менш важливим є підтверджена придатність усіх трьох моделей до аналізу зображень: за даними бенчмарків MMMU-Pro та рейтингу LMArena Vision вони входять до числа найрезультативніших МВММ у задачах мультимодального розуміння, підтримують структуроване виведення та демонструють достатній рівень мистецтвознавчої компетентності для надійної ідентифікації музейних об’єктів. Залучення двох моделей

платформи Google і однієї моделі OpenAI додатково уможливлює крос-вендорне порівняння та дозволяє оцінити стабільність результатів розпізнавання на різних програмних стеках.

2.2 Створення власного набору даних

Для оцінювання якості та точності відповідей трьох обраних моделей, якими є Gemini 2.5 Flash, Gemini 3.1 Pro та GPT-5.5, було сформовано власний еталонний набір даних. Головна мета створення цього масиву інформації полягає у забезпеченні дослідження абсолютно достовірними фактами. Якість роботи мультимодальних великих мовних моделей у задачі ідентифікації об’єктів неможливо оцінити суб’єктивно, оскільки цей процес вимагає суворого зіставлення з підтвердженими даними. Принцип побудови набору ґрунтується на тому, що для кожного музейного експоната заздалегідь зібрано та верифіковано повний перелік коректних метаданих. Надалі ці показники слугують точкою відліку при перевірці згенерованих машиною відповідей.

Кожен запис організовано у вигляді структурованого аркуша електронної таблиці формату XLSX. Запис включає офіційну назву твору або об’єкта, ім’я автора або назву культурної традиції, рік чи діапазон років створення, а також тип об’єкта. Додатково фіксується повна назва музею або місця зберігання, перелік матеріалів або технік виконання, тематичні цікаві факти про пам’ятку із зазначенням категорії та посилання на використане фото. Такий набір характеристик визначено невипадково, оскільки він дзеркально відображає ті параметри, які система MuseGuide AI повинна автоматично визначати у відповідь на системний запит. Відповідно кожне поле еталонного запису виступає окремою контрольною точкою при оцінюванні відповіді моделі.

Поточна версія набору містить п’ять записів. Стислий перелік об’єктів наведено у таблиці 2.3.

Таблиця 2.3 – Склад власного тестового набору даних

№	Назва	Автор	Рік	Місце зберігання
1	Катерина	Тарас Шевченко	1842	Нац. музей Тараса Шевченка (Київ)
2	Соняшники	Вінсент ван Гог	1888	Музей ван Гога, (Амстердам)
3	Бібліотекар	Джузеппе Арчимбольдо	1566	Замок Скоклостер (Швеція)
4	Циганка, яка спить	Анрі Руссо	1897	МоМА (Нью-Йорк, США)
5	Острів мертвих	Арнольд Бьоклін	1880–1886	Берлінська нац. галерея (Берлін)

Усі метадані набору надано за офіційними джерелами відповідних музейних установ ще до початку тестування. Це є принциповою вимогою, оскільки у разі наявності неточностей у еталонних даних саме оцінювання втрачає сенс. Саме тому під час підготовки масиву кожне поле перевірялося незалежно від загальновідомості конкретного твору.

Склад еталонного набору охоплює як широко відомі твори, зокрема картину «Катерина» Тараса Шевченка та «Бібліотекаря» Джузеппе Арчимбольдо, так і роботи з меншим рівнем впізнаваності за межами

вузькопрофесійної аудиторії. До останніх належить полотно «Циганка, яка спить» Анрі Руссо зі збірки нью-йоркського Музею сучасного мистецтва. Таке поєднання є навмисним, оскільки очікується, що моделі з навчанням на масштабних загальнодоступних корпусах краще впораються з ідентифікацією популярних у мережі об’єктів. Водночас саме на менш відомих творах можуть проявитися відмінності у глибині знань між різними платформами. Це дозволяє отримати диференційовану оцінку якості кожної моделі замість усередненого загального результату. У реальних умовах використання застосунку туристи часто фотографують не лише головні шедеври експозиції, але й вузькоспеціалізовані артефакти або твори місцевих митців. Здатність нейромережі успішно опрацьовувати подібні нетипові запити стає головним показником її надійності. Тому тестування на спеціально відібраних неочевидних зображеннях допомагає виявити справжні межі когнітивних можливостей кожної системи та гарантує стабільну роботу майбутнього програмного забезпечення у непередбачуваних ситуаціях.

2.3 Експериментальний аналіз роботи МВММ для задачі розпізнавання музейних експонатів

2.3.1 Створення спеціалізованого промпту

Розробка, структурування та детальна оптимізація системної інструкції для мультимодальної великої мовної моделі зумовлені специфікою інженерії підказок у рамках створення вебзастосунку «Музейний гід». Цей процес вимагає формування чіткого алгоритму взаємодії між користувачем та штучним інтелектом. Оскільки архітектура програмного забезпечення базується на автоматизованій обробці вхідних медіаданих на стороні сервера та подальшому динамічному відображенні згенерованого контенту в інтерфейсі користувача, виникає критична потреба у стандартизації

вихідного потоку даних. Наведена текстова інструкція безпосередньо вирішує це завдання шляхом накладання суворих обмежень на формат відповіді моделі. Згенерована інформація має бути представлена виключно у вигляді валідного об’єкта формату JSON згідно із заздалегідь заданою схемою. Таке архітектурне рішення повністю унеможливлює появу довільного супровідного або пояснювального тексту з боку нейромережі. Суворе структурування мінімізує ризики виникнення синтаксичних помилок та збоїв під час автоматичного синтаксичного аналізу даних програмними модулями застосунку. Крім того, перетворення неструктурованого виводу на чіткі програмні змінні дозволяє миттєво інтегрувати отриману інформацію у графічний інтерфейс додатка без додаткового оброблення тексту.

Важливим аспектом проєктування цього текстового запиту є забезпечення відмовостійкості системи та первинної валідації зображень за допомогою інтегрованого логічного перемикача класифікації об’єктів. Впровадження бінарного критерію ідентифікації дозволяє системі чітко розмежовувати релевантні культурні артефакти та випадкові побутові знімки. Сценарне розгалуження для випадків завантаження невідповідних зображень дозволяє коректно обробляти помилкові запити користувачів. У таких ситуаціях алгоритм повертає пусті значення у структурі даних замість генерації хибних вигадок з боку мультимодальної моделі. Відсіювання нерелевантного контенту на ранніх етапах обробки суттєво підвищує загальну надійність та стабільність роботи застосунку, а також заощаджує обчислювальні ресурси сервера. Поряд із цим інструкція містить чіткі директиви щодо адаптації атрибуції об’єктів залежно від їхньої природи. Система вимагає диференціації між конкретним авторством для класичного живопису та ширшим цивілізаційним чи культурним контекстом для археологічних знахідок. Гнучкий підхід до класифікації гарантує наукову точність згенерованого опису незалежно від типу досліджуваної пам’ятки. Повний текст розробленої системної інструкції представлено у лістингу 2.1.

Лістиинг 2.1 Повний текст розробленого промпту:

Ти – досвідчений музейний гід і мистецтвознавець.

Твоє завдання – розпізнати музейний експонат або картину на фотографії

та надати відповідь ВИКЛЮЧНО у форматі JSON згідно із заданою схемою.

Правила:

- Спочатку визнач: чи є це музейним експонатом, картиною, скульптурою або архітектурною пам'яткою?

- Якщо НІ (наприклад, звичайне фото людини, їжі, тварини, побутової сцени) – постав is_exhibit: false, решту полів заповни порожніми рядками або порожніми масивами.

- Якщо ТАК – постав is_exhibit: true і заповни всі поля.

- Відповідай ТІЛЬКИ українською мовою.

- Якщо на фото – картина, заповни поле "автор_або_культура" як автора.

- Якщо це артефакт чи предмет – вкажи культуру або цивілізацію.

- Цікаві факти мають бути справді захоплюючими, не банальними.

- Екскурсійний опис пиши живо, як справжній гід, що говорить з відвідувачем.

- Якщо щось невідомо з фото – чесно вкажи "Невідомо" або зроби обґрунтоване припущення.

- Не додавай жодного тексту поза JSON.

Окрему увагу в структурі системної інструкції приділено управлінню якістю художнього тексту через механізм рольового моделювання (Role-Playing). Закріплення за моделлю амплуа досвідченого музейного гіда та мистецтвознавця безпосередньо впливає на стилістику, тональність та глибину формування екскурсійного наративу. Це дозволяє нівелювати

властиву базовим моделям сухість та шаблонність викладу, трансформуючи енциклопедичні довідки у живий, захопливий та емоційно залучаючий діалог з відвідувачем. Вимога щодо генерації нетривіальних цікавих фактів та адаптивного екскурсійного супроводу українською мовою забезпечує високу комунікативну адекватність контенту, відтворюючи атмосферу реального візиту до музею та гарантуючи автентичність користувацького досвіду взаємодії з інтелектуальним віртуальним помічником.

2.3.2 Розпізнавання музейних експонатів

Під час розробки та практичної реалізації інтелектуального модуля для прикладного вебзастосунку «Музейний гід» було розгорнуто масштабне експериментальне дослідження сучасного ринку мультимодальних великих мовних моделей. Головна мета цього фундаментального етапу полягала у пошуку та верифікації оптимального технологічного рішення. Таке рішення мало з однаково високою ефективністю вирішувати два взаємопов’язані завдання. Першим завданням було забезпечення безпомилкового комп’ютерного зору для точної ідентифікації об’єктів культурної спадщини за зображеннями низької або нестандартної якості. Другим завданням виступала генерація стилістично вивіреного та художньо збагаченого текстового контенту, який повністю відповідає потребам кінцевих користувачів цифрового продукту. Вирішення цієї комплексної проблеми вимагає від штучного інтелекту не лише базового розпізнавання контурів чи кольорів, але й глибокого розуміння історичного та культурного контексту.

Як було детально обґрунтовано у попередніх розділах, для проведення порівняльних тестів було відібрано три флагманські мультимодальні архітектури. До експериментального пулу увійшли моделі Gemini 2.5 Flash, Gemini 3.1 Pro та GPT-5.5. Кожна з цих систем представляє унікальний підхід до інтеграції зорових та мовних шарів нейромереж, що робить їхнє пряме

порівняння максимально інформативним для проєкту. Важливою умовою тестування була оцінка роботи цих систем у режимі нульового навчання. Це означає, що нейромережі не проходили жодного додаткового дотренування на специфічних масивах музейних зображень перед початком експерименту. Такий підхід дозволив перевірити їхні базові енциклопедичні знання та здатність до самостійного узагальнення інформації. Здатність архітектур обробляти складні запити без попередньої підготовки є критично важливою для вебзастосунку, оскільки розробники не можуть передбачити абсолютно всі можливі об’єкти, які туристи захочуть сфотографувати.

Експериментальне тестування відібраних систем здійснювалося на базі спеціально сформованого та верифікованого масиву даних. Цей набір був навмисно укомплектований знімками різного ступеня складності для штучного інтелекту. До нього увійшли зображення архітектурних споруд, монументальної скульптури, об’єктів класичного живопису та дрібних археологічних артефактів, зафіксованих у неідеальних умовах. Масив містив фотографії з невдалими ракурсами, дефектами освітленості у вигляді надмірних тіней чи пересвітів, а також знімки нічної зйомки. Додатково враховувався різний ступінь деталізації текстур та наявність сторонніх об’єктів у кадрі, зокрема перехожих або транспортних засобів. Такий підхід дозволив максимально точно змоделювати реальний користувацький досвід у просторі музею чи міста. Окрім зазначених візуальних перешкод, до експериментальної вибірки були включені зображення картин із сильними відблисками від захисного скла та фотографії об’ємних скульптур, зроблені під нетиповими кутами огляду. Здатність алгоритмів комп’ютерного зору ігнорувати подібні світлові артефакти та фокусуватися виключно на головному об’єкті є одним із найскладніших випробувань для сучасних систем розпізнавання. Географічна та хронологічна різноманітність пам’яток у масиві також стала серйозним викликом для нейромереж. До вибірки потрапили не лише загальновідомі перлини західноєвропейського мистецтва, але й специфічні зразки локальної архітектури та маловідомі історичні

артефакти різних епох. Це допомогло перевірити відсутність культурної упередженості у навчальних базах кожної з моделей.

Для забезпечення максимальної чистоти та об’єктивності експерименту всі три моделі було поставлено в абсолютно ідентичні умови функціонування без жодних преференцій для окремих архітектур. Взаємодія з кожною нейромережею відбувалася за допомогою уніфікованої системної інструкції, принципи розробки якої були детально описані у попередньому підрозділі. Використання єдиного форматованого запиту гарантувало стандартизацію процесу перевірки як аналітичних здібностей моделей, так і їхньої здатності суворо дотримуватися технічних обмежень щодо формату виведення даних. Під час проведення експериментів також ретельно фіксувалися показники швидкості обробки кожного запиту. В умовах реальної експлуатації мобільного програмного забезпечення затримка відповіді сервера відіграє вирішальну роль у формуванні позитивного користувацького досвіду. Тому тестування відбувалося з імітацією навантаження, яке повністю відповідає типовим сценаріям використання туристичних сервісів у мобільних мережах зі нестабільним з’єднанням. Кожна зафіксована відповідь нейромереж підлягала подальшому багатофакторному аналізу задля виявлення сильних сторін та критичних вразливостей обраних технологічних рішень. Це дозволило обґрунтовано обрати архітектуру для проєкту та сформувати базу для подальшої оптимізації застосунку.

2.3.3 Висновки щодо результатів експериментів

Проведений порівняльний аналіз первинних технічних метрик виявив глибоку асиметрію у продуктивності досліджуваних нейромереж, що призвело до перегляду початкового пулу моделей для подальшої валідації. Попри високі лідируючі позиції в загальних рейтингах Chatbot Arena та заявлений високий потенціал, модель GPT-5.5 продемонструвала

незадовільну ефективність під час безпосереднього опрацювання прикладного масиву об’єктів культурної спадщини. У процесі тестування було зафіксовано критичний рівень похибок, оскільки модель виявилася неспроможною коректно розпізнати, ідентифікувати та контекстуально атрибутувати понад половину представлених у наборі даних елементів.

Найчастіше збої в роботі GPT-5.5 виявлялися у двох формах. По-перше, модель демонструвала надмірну схильність до хибних відмов, маркуючи унікальні історичні пам’ятки та музейні експонати як нерелевантні побутові предмети чи випадкові знімки, що призводило до помилкового спрацьовування програмного тригера. По-друге, нейромережа генерувала масштабні галюцинації, повністю спотворюючи верифіковані історичні факти, дати спорудження, імена авторів та географічні локації об’єктів. З огляду на таку системну нестабільність, високу частоту критичних помилок та невідповідність вимогам відмовостійкості, використання моделі GPT-5.5 у структурі застосунку було визнано недоцільним, внаслідок чого її повністю виключили з процесу фінального експертного опитування.

Натомість моделі сімейства Gemini повною мірою підтвердили свою актуальність та високий ступінь оптимізації під складні мультимодальні завдання, виявивши високу стійкість до специфічних умов зйомки, зокрема нестандартних ракурсів та слабкого освітлення, і безпомилково впоравшись із верифікацією чітких фактичних даних для всього масиву тестових зображень. Оскільки технічна точність розпізнавання у Gemini 2.5 Flash та Gemini 3.1 Pro виявилася абсолютною, подальший вибір між ними було переведено у площину оцінювання якісних показників художнього тексту. Результати генерації цих двох моделей, отримані за допомогою різних підходів до формування запитів, сформували шість варіацій відображення контенту, які й були передані на суб’єктивне оцінювання колегії з 10 профільних експертів сфери туризму та музейної справи. Такий двоетапний підхід дозволив оперативно відсіяти неефективні алгоритми на ранній стадії

та сфокусувати експертну валідацію виключно на аналізі найкращого стилістичного супроводу для користувачів застосунку.

2.4 Оцінка якості супроводу експертами

Оскільки верифікація чітких фактичних даних, таких як назва пам’ятки, рік її спорудження, географічне розташування та архітектурний стиль, продемонструвала абсолютну точність і стабільність в усіх досліджуваних мультимодальних великих мовних моделях, застосування автоматичних метрик для цього етапу виявилося малоінформативним. Водночас оцінювання генерованих блоків із цікавими фактами та екскурсійними маршрутами потребувало глибшого аналізу, оскільки ці елементи містять художньо-описову складову, детерміновану суб’єктивним сприйняттям користувача. З огляду на те, що автоматизовані алгоритми не здатні об’єктивно виміряти захопливість викладу, логіку побудови розповіді, емоційну залученість та загальне враження від запропонованої екскурсії, у цьому дослідженні було впроваджено метод суб’єктивного експертного оцінювання.

Для систематизації отриманих результатів було розроблено структуровану форму опитування, орієнтовану на фіксацію якісних показників текстового контенту. До процесу анкетування залучено 10 експертів, що дозволило забезпечити репрезентативність виставлених балів. Кожному респонденту пропонувалося проаналізувати результати генерації для 5 витворів мистецтва із тестового набору даних, які були представлені у п’яти блоках по два питання, кожен з яких мав по 4 варіанти відповіді (рис. 2.1). Головним критерієм оцінювання у створеній формі виступала саме інформаційна привабливість фактів та якість структурування екскурсійного блоку, що дозволило визначити найбільш ефективну модель для створення адаптивного туристичного контенту.

Для представленої картини оберіть, які з фактів вам було б цікавіше почути:

Варіант 1:

• Арнольд Беклін створив п'ять версій «Острова мертвих» між 1880 та 1886 роками, кожна з яких має свої нюанси та зберігається в різних музеях світу.
• Картина стала одним із найвідоміших творів символізму, її похмура, містична атмосфера глибоко вплинула на європейську культуру та мистецтво.
• Вона була улюбленою картиною багатьох видатних особистостей, включаючи Зигмунда Фрейда, Сергія Рахманінова (який написав симфонічну поему під її враженням) та навіть Адольфа Гітлера.
• Образ острова з високими кипарисами, що нагадують похоронні урни, та фігура в білому в човні, що часто інтерпретується як Харон, символізують перехід у потойбічний світ.

Варіант 2:

"Беклін створив п'ять версій цієї картини, кожна з яких має дещо іншу атмосферу та освітлення.",
"Картина мала колосальний вплив на культуру: вона надихнула Сергія Рахманінова на написання однойменної симфонічної поеми.",
"Однією з версій картини володів Адольф Гітлер, який придбав її у 1933 році."

- ◯ Варіант 1
- ◯ Варіант 2
- ◯ Однаково цікаво
- ◯ Однаково нецікаво

Рисунок 2.1 – Фрагмент створеної форми

Для кожного елементу набору даних респондент обирав, який з двох варіантів здавався йому найбільш цікавим. Також було представлено опцію вибору двох або жодного з варіантів, що дозволило отримати більш повні дані від експертів (рис. 2.2). Запровадження таких альтернативних відповідей

допомогло уникнути штучного примусу до вибору в ситуаціях з однаково високою або однаково низькою якістю текстів. Зокрема фіксація відмов від обох варіантів стала важливим індикатором тих тематичних напрямів, де штучний інтелект все ще потребує додаткового налаштування системних інструкцій.

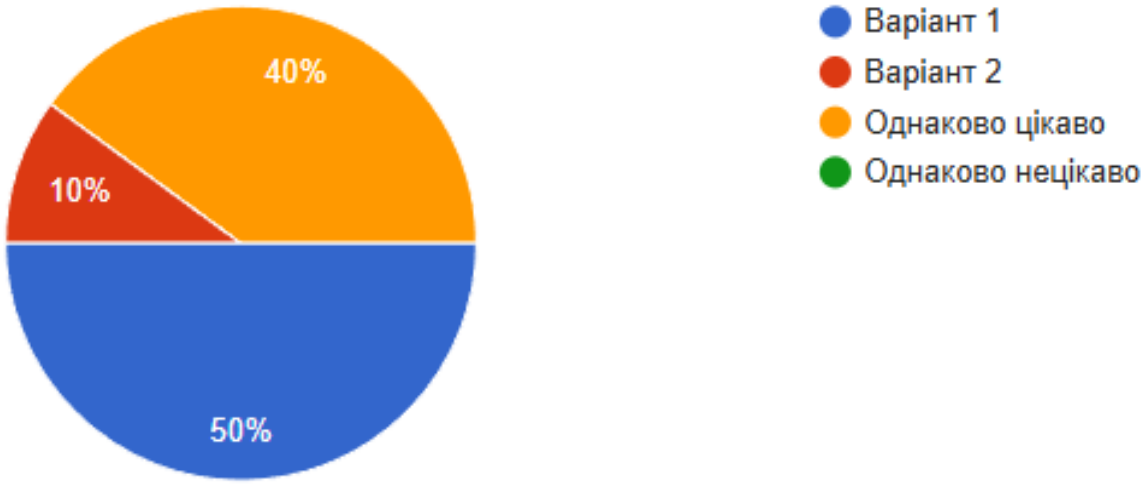

Рисунок 2.2 – Приклад результатів опитування для одного з елементів набору даних

Аналіз отриманих результатів анкетування дозволив чітко визначити лідера серед досліджуваних мультимодальних великих мовних моделей за критеріями художньої виразності та адаптивності генерованого контенту. Зведені результати опитування продемонстрували, що більшість респондентів обирала модель Gemini 2.5 Flash в середньому у 62 відсотках випадків, відзначаючи вищу якість наративу та емоційну залученість у сформованих екскурсійних блоках (рис. 2.3). Експерти систематично надавали перевагу відповідям цієї моделі завдяки глибшому розкриттю маловідомих історичних аспектів, відсутності мовних кліше та більш логічній структурі побудови віртуального маршруту.

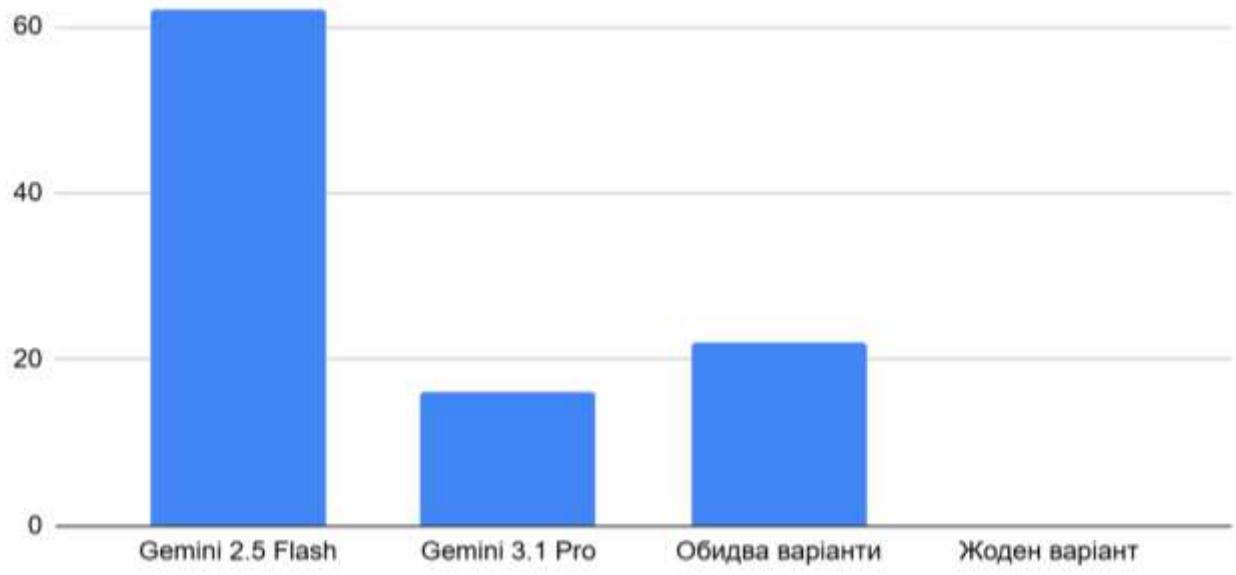

Рисунок 2.3 – Зведені результати опитування

Враховуючи виявлену статистичну перевагу та високі оцінки експертної спільноти щодо формування захопливого текстового супроводу, для подальшої практичної реалізації та інтеграції у фінальну систему розпізнавання музейних експонатів було обрано саме модель Gemini 2.5 Flash. Такий значний відрив від інших тестованих платформ підтверджує високу зрілість обраної архітектури у задачах семантичного оброблення культурної інформації. Важливо зазначити, що згенеровані цією моделлю описи практично не потребували подальшого ручного редагування перед їхнім перетворенням у звуковий формат синтезатором мовлення. Отримані статистичні дані стали остаточним науковим аргументом на користь повної інтеграції саме цієї нейромережі у програмну логіку майбутнього застосунку.

3 РОЗРОБКА ВЕБЗАСТОСУНКУ ДЛЯ СУПРОВОДУ ВІДВІДУВАЧІВ МУЗЕЇВ

3.1 Вимоги до системи та функціональна специфікація

Проєктування будь-якої програмної системи розпочинається із чіткого формулювання вимог, що визначають функціональне охоплення, якісні характеристики та обмеження розроблюваного рішення. Для системи MuseGuide AI вимоги сформульовано з урахуванням специфіки музейного середовища, потреб кінцевого користувача та технологічних обмежень обраної платформи розгортання. Відповідно до загальноприйнятого підходу до специфікації програмного забезпечення вимоги поділяються на функціональні, що описують конкретну поведінку системи у відповідь на дії користувача, та нефункціональні, що визначають якісні характеристики її роботи.

Функціональні вимоги системи охоплюють п’ять ключових сценаріїв взаємодії. Першим є можливість завантаження зображення музейного експоната або твору мистецтва через вебінтерфейс, при цьому система підтримує формати JPEG, PNG та WebP із максимальним розміром файлу до 20 мегабайт. Другою вимогою є автоматична ідентифікація завантаженого об’єкта із витяганням структурованої інформації, до якої належать назва, автор або культурна традиція, часовий période, тип об’єкта, матеріал або техніка виконання, місце зберігання та перелік цікавих фактів у кількості від двох до чотирьох. Третьою вимогою виступає генерування живого екскурсійного опису обсягом два-три речення від першої особи гіда, що забезпечує природний стиль подання матеріалу. Четверта вимога передбачає озвучення сформованого опису синтезованим голосом українською мовою з автоматичним збереженням у форматі MP3 та відтворенням безпосередньо в інтерфейсі. П’ята вимога визначає обов’язкову фільтрацію нерелевантних зображень із виведенням відповідного повідомлення для користувача.

Нефункціональні вимоги охоплюють коректне оброблення виняткових ситуацій, зокрема відсутності мережевого підключення, помилок при зверненні до зовнішнього програмного інтерфейсу та надходження зображень низької якості. Інтерфейс має бути інтуїтивно зрозумілим та не потребувати попереднього навчання, а система повинна коректно функціонувати у сучасних веббраузерах Chrome, Firefox та Safari без встановлення додаткового програмного забезпечення.

Обмеженнями системи є обов’язкова наявність підключення до мережі Інтернет та дійсного ключа доступу до Google Gemini API, без яких неможливе виконання основної функції розпізнавання. Розгортання системи передбачає вебінтерфейс, доступний через будь-який сучасний браузер, що забезпечує кросплатформенність без необхідності встановлення окремого клієнтського застосунку.

Систематизація визначених вимог дозволяє чітко розмежувати базові можливості платформи та додаткові сервіси штучного інтелекту. Врахування цих параметрів на етапі проєктування допомагає визначити пріоритетність реалізації кожної функції та сформувати критерії її успішної верифікації. Окрім цього, структурований опис вимог стає основою для побудови сценаріїв взаємодії користувача з інтерфейсом і гарантує повне покриття функціоналу під час розробки архітектури. Такий підхід мінімізує ризики виникнення архітектурних помилок під час написання коду. Деталізація критеріїв дозволяє побудувати прозору систему метрик для оцінювання працездатності інтелектуальних модулів. Формалізація параметрів оброблення інформації виступає підґрунтям для подальшого інтеграційного тестування. Зафіксовані нефункціональні вимоги гарантують стабільну роботу застосунку за умов високого мережевого навантаження. Зрештою це дозволяє створити надійний програмний продукт для задоволення усіх запитів користувачів. Зведена функціональна специфікація системи наведена у таблиці 3.1.

Таблиця 3.1 – Функціональна специфікація системи MuseGuide AI

Вимога	Опис та критерій виконання
Введення зображення	Підтримка форматів JPEG, PNG, WebP; розмір до 20 МБ; введення через вебінтерфейс або камеру пристрою
Ідентифікація об’єкта	Визначення назви, автора або культури, часового періоду, типу, матеріалу та місця зберігання на основі Gemini Vision API
Генерування опису	Формування екскурсійного тексту 2–3 речення від імені гіда; унікальний результат для кожного запиту
Фільтрація нерелевантних зображень	Відхилення немузейних зображень із повідомленням користувачу; точність класифікації не нижче 95%
Синтез мовлення	Озвучення опису засобами Edge TTS; формат MP3; автовідтворення
Сумісність	Коректна робота у браузерах Chrome, Firefox, Safari; можливість розгортання локально та на Hugging Face Spaces

3.2 Налаштування середовища розробки та конфігурація програмного забезпечення

Організація робочого середовища розробки є важливим підготовчим етапом, що безпосередньо впливає на відтворюваність результатів та зручність розгортання системи в різних середовищах. Для системи MuseGuide AI обрано підхід на основі ізольованого віртуального середовища

Python, що дозволяє уникнути конфліктів залежностей із системними пакетами та спрощує відтворення конфігурації на інших машинах.

Основою технологічного середовища є інтерпретатор Python версії 3.10 або вище, вибір якого зумовлений необхідністю підтримки сучасних синтаксичних конструкцій, зокрема операторів об'єднання типів та вдосконалених механізмів оброблення виключень. Для ізоляції залежностей проєкту від системних пакетів використовується стандартний модуль venv, що забезпечує створення автономного Python-середовища у директорії .venv у корені проєкту. Активація цього середовища здійснюється за допомогою відповідного скрипту для операційної системи розробника.

Усі залежності проєкту зафіксовано у файлі requirements.txt, що є стандартним підходом для відтворюваного розгортання Python-застосунків. Файл містить п'ять обов'язкових пакетів: google-genai версії 0.8.0 або вище, що забезпечує доступ до Google Gemini API; gradio версії 6.0.0 або вище для побудови вебінтерфейсу; Pillow версії 10.0.0 або вище для оброблення зображень; edge-tts версії 6.1.0 або вище для синтезу мовлення та python-dotenv версії 1.0.0 або вище для управління змінними середовища. Встановлення всіх залежностей здійснюється єдиною командою після активації віртуального середовища.

Конфігурація системи здійснюється через файл .env, розташований у корені проєкту. У цьому файлі зберігається єдина обов'язкова змінна середовища GEMINI_API_KEY, що містить ключ доступу до Google Gemini API, отриманий у Google AI Studio. Використання механізму змінних середовища замість прямої фіксації ключів у коді відповідає загальноприйнятим практикам безпечної розробки програмного забезпечення та унеможливлює випадкове потрапляння конфіденційних даних до системи контролю версій. Файл .env зазначено у .gitignore, що запобігає його публікації у репозиторії.

Структура файлової системи проєкту організована за принципом розділення відповідальності. Корінь проєкту містить файл app.py, що

відповідає за побудову інтерфейсу та координацію виклику модулів, а також файли requirements.txt, .env та .gitignore. Директорія modules містить три файли: llm.py зі службою розпізнавання та комунікації з мовною моделлю, tts.py з модулем синтезу мовлення та utils.py з допоміжними функціями загального призначення. Директорія outputs створюється автоматично під час першого запуску та слугує для тимчасового зберігання згенерованих аудіофайлів.

Для розгортання у хмарному середовищі Hugging Face Spaces конфігурація середовища залишається незмінною: ключ API передається через механізм секретів платформи, а файл requirements.txt автоматично зчитується при побудові контейнера. Це забезпечує повний паритет між локальним та хмарним середовищами розробки і спрощує процес публікації готового прототипу.

3.3 Вибір мов програмування, технологій та фреймворків

Вибір технологічного стека для реалізації системи MuseGuide AI зумовлений сукупністю вимог до неї: доступністю зрілих бібліотек для роботи з мультимодальними великими лінгвістичними моделями, можливістю швидкого прототипування вебінтерфейсу та простотою розгортання як локально, так і у хмарному середовищі. Кожен компонент стека обирався відповідно до конкретного функціонального завдання із пріоритетом мінімізації кількості зовнішніх залежностей.

Основною мовою програмування обрано Python версії 3.10 і вище. Ця мова є стандартом де-факто у галузі машинного навчання та штучного інтелекту завдяки розвиненій екосистемі бібліотек та широкій підтримці з боку розробників платформ штучного інтелекту. Зокрема підтримка асинхронного програмування через модуль asyncio дозволяє ефективно організувати паралельне виконання незалежних операцій, серед яких

звернення до зовнішнього програмного інтерфейсу та синтез мовлення, без блокування основного потоку виконання. Методи цифрового оброблення зображень, що реалізуються у Python-середовищі, охоплюють широкий спектр задач від попередньої обробки вхідних даних до класифікації та ідентифікації об’єктів [23].

Ядром системи розпізнавання є Google Gemini Vision API із моделлю gemini-2.5-flash, що є головним результатом порівняльного аналізу, проведеного у другому розділі роботи. Ця мультимодальна велика лінгвістична модель, архітектура якої базується на механізмах уваги трансформерів, поєднує здатність аналізувати зображення з можливостями генерування природної мови [9]. Використання структурованого виведення через механізм types.Schema забезпечує передбачуваний формат відповіді у вигляді JSON-об’єкта. Цей механізм дозволяє специфікувати точну структуру очікуваної відповіді з визначенням типів кожного поля, що усуває необхідність у додатковому синтаксичному аналізі вільного тексту та підвищує стійкість системи до непередбачуваних форматів виведення.

Для синтезу мовлення обрано бібліотеку edge-tts версії 6.1 і вище, яка надає програмний доступ до нейромережевих голосів Microsoft Azure без необхідності придбання платної підписки. Бібліотека підтримує асинхронне виконання через asyncio, що не блокує основний потік під час генерування аудіофайлу. Порівняно з бібліотекою gTTS, яка первісно розглядалась як альтернатива, edge-tts забезпечує значно вищу якість синтезу та більш природне звучання завдяки використанню нейромережевих голосових моделей виробничого рівня, розроблених Microsoft для платформи Azure Cognitive Services.

Для оброблення зображень використовується бібліотека Pillow версії 10.0 і вище. Вона забезпечує зчитування зображень у різних форматах та їх конвертацію у внутрішній об’єктний тип, сумісний із Google GenAI SDK. Конфігурація середовища здійснюється через python-dotenv, що дозволяє зберігати чутливі дані поза репозиторієм з кодом. Такий підхід

відповідає загальноприйнятим практикам безпечної розробки програмного забезпечення та спрощує розгортання системи у різних середовищах без ризику витоку конфіденційних даних. Технологічний стек системи MuseGuide AI зведено у таблиці 3.2.

Таблиця 3.2 – Технологічний стек системи MuseGuide AI

Компонент	Технологія	Версія	Призначення
Мова програмування	Python	≥ 3.10	Основна мова реалізації всіх модулів
Вебінтерфейс	Gradio	≥ 6.0.0	Побудова інтерфейсу користувача та обробка подій
Мультимодальна модель	Google Gemini Vision API (gemini-2.5-flash)	SDK ≥ 0.8.0	Ідентифікація експонатів та генерування описів
Синтез мовлення	Edge TTS	≥ 6.1.0	Озвучення екскурсійних описів
Оброблення зображень	Pillow (PIL)	≥ 10.0.0	Зчитування та конвертація вхідних зображень
Управління середовищем	python-dotenv	≥ 1.0.0	Зберігання ключів поза репозиторієм

Обраний технологічний стек характеризується мінімальністю залежностей при максимальному функціональному охопленні: п’ять пакетів забезпечують реалізацію повного циклу роботи застосунку від завантаження зображення до озвучення результату. Це скорочує як час ініціалізації

середовища, так і ймовірність виникнення конфліктів між залежностями при розгортанні на нових машинах.

3.4 Реалізація модуля мультимодального аналізу

Центральним технологічним компонентом системи MuseGuide AI є модуль комп'ютерного зору та мультимодального аналізу, реалізований у файлі modules/llm.py. Саме цей модуль відповідає за перетворення вхідного зображення на структуровану семантичну інформацію про музейний експонат і, як наслідок, визначає ключові показники якості всієї системи.

Основним архітектурним рішенням модуля є використання механізму типізованого структурованого виведення замість вільного текстового відгуку. Цей підхід реалізується через функцію _build_schema, що формує об'єкт types.Schema із дев'ятьма обов'язковими полями. Першим полем є is_exhibit булевого типу, що виступає бінарним класифікатором релевантності зображення: значення true сигналізує про присутність музейного експоната, картини, скульптури або архітектурної пам'ятки, тоді як значення false відповідає нерелевантному побутовому зображенню. Наступні поля рядкового типу містять назву об'єкта, поле автор_або_культура, що приймає ім'я автора для картин або назву культури та цивілізації для артефактів, часовий період, тип експоната у вигляді переліку допустимих значень, а саме картина, скульптура, артефакт, документ, зброя, ювелірний виріб, побутовий предмет, архітектура та інше. Окремими полями виступають масив цікавих_фактів із двох-чотирьох текстових елементів, поле матеріал_або_техніка, поле де_зберігається та поле екскурсійний_опис, що містить живий екскурсійний текст обсягом два-три речення. Застосування типізованої схеми усуває необхідність у крихкому синтаксичному аналізі відповіді та гарантує наявність усіх необхідних полів у поверненому об'єкті [24].

Системна інструкція, визначена у константі SYSTEM_PROMPT, виконує роль рольового налаштування моделі та задає правила її поведінки. Закріплення за моделлю ролі досвідченого музейного гіда та мистецтвознавця безпосередньо впливає на стилістику і глибину генерованого тексту, трансформуючи енциклопедичні довідки у живий та емоційно залучаючий діалог з відвідувачем. Інструкція містить вказівку відповідати виключно українською мовою, диференціювати атрибуцію залежно від природи об’єкта, генерувати нетривіальні цікаві факти та чесно зазначати невідомі дані. Принципово важливою є директива не додавати жодного тексту поза структурою відповіді, що у поєднанні з вимогою виведення у форматі JSON унеможливлює появу пояснювальних вставок або супровідного тексту від моделі.

Основна публічна функція модуля analyze_exhibit приймає об’єкт зображення типу Image.Image та повертає словник Python із структурованими даними про експонат. Формування запиту до програмного інтерфейсу здійснюється через метод generate_content SDK Google GenAI, до якого передаються текстовий запит та зображення у єдиному масиві вмісту. Конфігурація виклику задає тип MIME-відповіді application/json для гарантованого отримання коректного JSON-об’єкта, передає схему структурованого виведення та встановлює параметр температури на рівні 0.4. Знижена температура забезпечує відносну детермінованість фактичних даних при збереженні творчої варіативності для художніх описів. Максимальний обсяг вихідних токенів встановлено на рівні 4096, що є достатнім для повноцінного опису будь-якого музейного об’єкта.

Після отримання відповіді від програмного інтерфейсу функція виконує попереднє очищення тексту від можливих markdown-обрамлень за допомогою регулярних виразів, після чого виконується синтаксичний аналіз JSON за допомогою стандартної бібліотеки json. Якщо поле is_exhibit містить значення false, генерується виняток NotAnExhibitError із описовим повідомленням для користувача. Це рішення реалізує принцип раннього

виявлення помилки: система негайно сигналізує про нерелевантне зображення замість того, щоб генерувати безглуздий опис. Функція _validate перевіряє наявність усіх обов'язкових полів у відповіді та генерує ValueError при виявленні прогалин. Такий дворівневий контроль якості мінімізує ймовірність передачі некоректних даних до наступних модулів.

Модуль також містить дві допоміжні функції форматування: format_for_display формує читабельний рядок для відображення у текстовому режимі, тоді як format_for_tts конкатенує ключові поля у єдиний рядок, оптимізований для природного звучання при синтезі мовлення. Для функції синтезу навмисно опускаються деякі допоміжні поля, зокрема тип об'єкта та місце зберігання, на користь тих, що формують найбільш інформативне та зв'язне усне висловлювання. Спочатку передається назва, потім автор або культурна традиція, далі часовий період, цікаві факти та, нарешті, живий екскурсійний опис.

3.5 Реалізація модуля синтезу мовлення

Модуль синтезу мовлення, реалізований у файлі modules/tts.py, перетворює текстовий опис музейного експоната на природномовний аудіофайл. Завданням цього модуля є забезпечення аудіосупроводу, якість якого відповідає стандартам інформаційно-довідкових систем, що характеризується чітким вимовлянням, правильною інтонацією та природним темпом мовлення.

Для синтезу мовлення обрано бібліотеку edge-tts, що надає програмний доступ до нейромережевих голосових моделей Microsoft Azure без необхідності реєстрації або придбання підписки. Ключовою перевагою цього рішення є використання нейромережевих голосів виробничого рівня, що суттєво перевищують за природністю звучання традиційні статистичні синтезатори мовлення. Для українськомовного виведення обрано голос uk-

UA-PolinaNeural, що характеризується чітким вимовлянням та природною інтонацією для текстів художньо-описового характеру. Додатково у модулі визначено константу для англійського голосу en-US-JennyNeural як запасний варіант для потенційного розширення системи.

Центральною функцією модуля є text_to_speech, що приймає рядок тексту та мовний код і повертає шлях до згенерованого аудіофайлу. Функція валідує вхідний текст на предмет порожнечі, формує унікальне ім’я файлу на основі восьмисимвольного шістнадцяткового ідентифікатора UUID та ініціює асинхронну генерацію. Використання UUID замість послідовних числових індексів запобігає конфліктам іменування при паралельних запитах від кількох користувачів у хмарному середовищі розгортання. Вихідні файли зберігаються у директорії outputs, яка автоматично створюється при ініціалізації модуля.

Безпосередня генерація аудіо реалізована у приватній асинхронній корутині _generate, що інстанціює об’єкт edge_tts.Communicate з текстом і голосом, після чого викликає метод save для збереження потоку аудіоданих у файл. Виклик корутини з синхронного контексту здійснюється через asyncio.run, що забезпечує коректну роботу як при безпосередньому запуску модуля, так і при виклику з Gradio-інтерфейсу. Цей підхід є обов’язковим, оскільки бібліотека edge-tts реалізована повністю на основі асинхронного програмного інтерфейсу і не надає синхронних обгорток.

Управління дисковим простором забезпечується функцією cleanup_old_audio. Вона сортує всі аудіофайли у директорії outputs за часом останньої модифікації та видаляє всі, крім п’яти найновіших. Це запобігає поступовому накопиченню тимчасових файлів у довготривалих сесіях та у хмарному середовищі розгортання. Параметр keep_last є конфігурованим та за замовчуванням дорівнює п’яти файлам, що є достатнім для покриття кількох послідовних запитів під час демонстрації системи.

Первинна версія модуля базувалась на бібліотеці gTTS, що використовує синтез мовлення Google. Ця версія збережена у файлі у вигляді

закоментованого блоку коду, що документує еволюцію архітектурного рішення. Перехід до edge-tts був зумовлений необхідністю підвищення якості звучання: нейромережеві голоси Microsoft демонструють значно природнішу інтонацію при відтворенні складних текстів із мистецтвознавчою лексикою порівняно з попередником.

3.6 Алгоритм розпізнавання експонатів та генерування екскурсійного опису

Алгоритм розпізнавання музейних експонатів та генерування екскурсійного опису є центральним функціональним компонентом системи MuseGuide AI. Він реалізує чотириетапний конвеєр оброблення вхідного зображення, де кожен етап спирається на результати попереднього та за необхідності генерує діагностичне повідомлення для дострокового завершення обробки.

На першому етапі функція process_image у файлі app.py перевіряє наявність вхідного зображення та при його відсутності повертає набір порожніх рядків без звернення до жодного зовнішнього сервісу. Якщо зображення присутнє, керування передається до функції analyze_exhibit модуля llm.py, яка формує мультимодальний запит до Google Gemini Vision API. Запит містить текстовий користувацький промпт із вимогою розпізнати об’єкт, системну інструкцію, визначену у константі SYSTEM_PROMPT, а також вхідне зображення у форматі PIL Image. До конфігурації запиту передається схема типізованого виведення, що гарантує отримання структурованого об’єкта у форматі JSON.

На другому етапі система аналізує поле is_exhibit у отриманій відповіді. Якщо його значення дорівнює false, генерується виняток NotAnExhibitError із описовим повідомленням, яке Gradio-інтерфейс перехоплює та відображає користувачу у вигляді зрозумілого сповіщення.

Цей механізм раннього виявлення помилки запобігає подальшій обробці нерелевантних зображень та економить ресурси зовнішніх сервісів. Після успішної верифікації функція _validate перевіряє наявність усіх дев’яти обов’язкових полів у відповіді.

На третьому етапі функція format_for_tts формує єдиний текстовий рядок для озвучення шляхом конкатенації поля назви, поля автора або культури, поля часового періоду, цікавих фактів через крапку та екскурсійного опису. Така структура забезпечує логічну послідовність інформації у звуковому потоці: спершу встановлюється ідентичність об’єкта, потім надаються додаткові відомості, і, нарешті, звучить художній опис від імені гіда.

На четвертому етапі перед синтезом мовлення викликається функція cleanup_old_audio для видалення застарілих аудіофайлів, після чого функція text_to_speech генерує новий MP3-файл. Шлях до цього файлу передається до компонента Audio Gradio-інтерфейсу, що забезпечує автоматичне відтворення одразу після готовності. Паралельно функція process_image формує кортеж із восьми значень для наповнення всіх вихідних компонентів інтерфейсу: назви, поєднаного рядка автора та періоду, типу об’єкта, матеріалу або техніки, місця зберігання, відформатованих цікавих фактів, екскурсійного опису та шляху до аудіофайлу.

Ключовою конструктивною перевагою алгоритму є чіткий розподіл відповідальності між компонентами: модуль llm.py ізольовано реалізує логіку взаємодії з мовною моделлю та нічого не знає про інтерфейс, модуль tts.py повністю відповідає за синтез мовлення та управління файлами, а app.py виступає виключно оркестратором, що координує виклики цих модулів і відображає результат. Таке архітектурне рішення відповідає принципу єдиної відповідальності та спрощує тестування та розширення кожного компонента незалежно від інших.

3.7 Розробка користувацького інтерфейсу

Розробка користувацького інтерфейсу системи MuseGuide AI здійснювалась за двома паралельними напрямами: структурним, що визначає розміщення та функціональну роль кожного компонента, та стилістичним, що відповідає за візуальну ідентичність системи та відповідність музейному контексту. Обидва аспекти є рівноцінно важливими, оскільки інтерфейс є єдиною точкою контакту між користувачем та всією логікою системи.

Для побудови вебінтерфейсу обрано фреймворк Gradio у підході Blocks, що дозволяє будувати складні макети з кількома колонками та залежностями між компонентами [25]. На відміну від спрощеного підходу Interface, що дозволяє описати функцію лише з одним набором входів та виходів, підхід Blocks надає повний контроль над макетом, включаючи вкладені рядки та колонки. Ключовою перевагою Gradio для цього проєкту є нативна інтеграція з платформою Hugging Face Spaces: будь-який Gradio-застосунок може бути розгорнутий у хмарному середовищі шляхом завантаження файлів до репозиторію без налаштування серверної інфраструктури.

Макет інтерфейсу побудовано за дворівневою схемою. Загальний заголовок займає повну ширину сторінки та містить назву системи, виконану шрифтом Playfair Display, та підзаголовок з описом призначення застосунку. Основна зона поділена на дві рівнозначних колонки. Ліва колонка містить компонент Image для завантаження зображення з висотою 340 пікселів та кнопку Button для запуску аналізу. Права колонка містить вісім вихідних компонентів у порядку від найзагальнішої до найдеталізованішої інформації: поле Textbox для назви, поле Textbox для поєднаного рядка автора та періоду, горизонтальний рядок із двома компонентами Textbox для типу та матеріалу відповідно, поле Textbox для місця зберігання, розширене поле Textbox для цікавих фактів із чотирма рядками відображення, поле Textbox

для екскурсійного опису та компонент Audio для відтворення озвученого тексту з увімкненим режимом автовідтворення.

Прив’язка обробника до кнопки здійснюється методом analyze_btn.click, що передає вхідне зображення до функції process_image та розподіляє повернений кортеж із восьми значень між відповідними вихідними компонентами. Такий підхід гарантує атомарне оновлення всіх вихідних полів: вони оновлюються одночасно після повного завершення обробки, а не послідовно, що запобігає відображенню проміжних некоректних станів.

Візуальний стиль інтерфейсу реалізований через блок кастомного CSS, визначений у константі custom_css. Стиль спирається на набір CSS-змінних, що утворюють палітру музейної теми. Змінна museum-gold відповідає відтінку золотистого кольору, що використовується для підписів полів результату та рамки зони завантаження. Змінна museum-dark задає майже чорний відтінок для тексту заголовка та кнопки аналізу. Змінна museum-cream визначає теплий кремовий колір фону сторінки. Змінна museum-stone відповідає теракотово-кам’яному відтінку для допоміжного тексту підзаголовка. Змінна museum-card задає теплий білий колір карток результату. Централізоване визначення кольорів через змінні спрощує подальшу адаптацію теми без зміни численних окремих правил.

Типографіка інтерфейсу поєднує два гарнітури. Playfair Display є шрифтом із засічками в стилі ренесансу, що використовується для заголовка системи та підписів полів результату і відсилає до традиційного музейного середовища. Inter є сучасним гуманістичним гротеском без засічок, що застосовується для основного тексту та кнопки і забезпечує максимальну читабельність на екранах різних розмірів. Обидва шрифти підключаються з Google Fonts через директиву @import на початку блоку CSS. Зовнішній вигляд інтерфейсу системи наведено на рисунку 3.1.

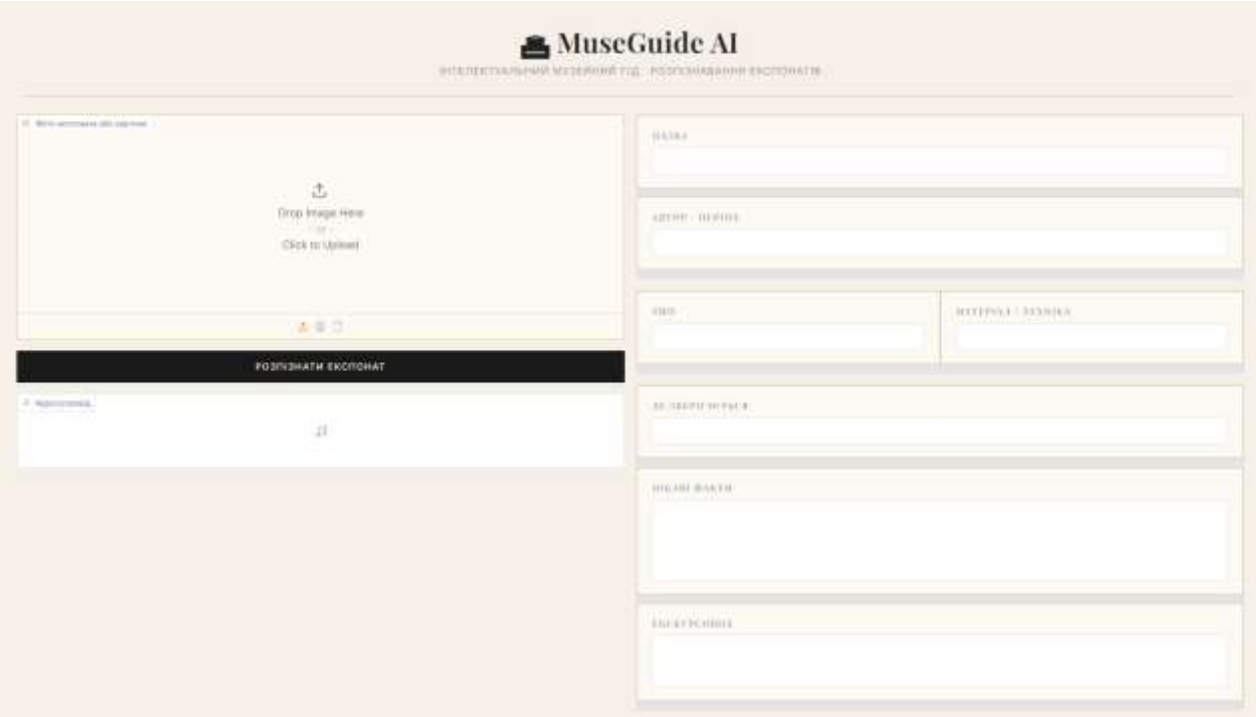

Рисунок 3.1 – Інтерфейс системи MuseGuide AI

Зона завантаження зображення стилізована пунктирною рамкою золотистого кольору, що візуально сигналізує про призначення зони та відповідає загальній палітрі інтерфейсу. Кнопка аналізу виконана у темному кольорі з інтервальним трекінгом тексту та ефектом зміни фону при наведенні курсора. Картки результату мають кремовий фон із тонкою рамкою та закругленими кутами мінімального радіусу, що забезпечує стриману елегантність без надмірної декоративності. Нижній колонтитул Gradio прихований через CSS для збереження чистоти оформлення.

Обробка виняткових ситуацій у інтерфейсі реалізована через механізм gr.Error: коли модуль ідентифікації генерує виняток NotAnExhibitError, Gradio перехоплює його та відображає у вигляді зрозумілого повідомлення у спеціальному сповіщенні без аварійного завершення роботи застосунку. Завдяки цьому інтерфейс залишається повністю функціональним після будь-якої помилки, а користувач отримує чітку вказівку щодо необхідних коригуючих дій.

3.8 Ілюстрація роботи застосунку

3.8.1 Сценарій розпізнавання картини

Для перевірки здатності системи до ідентифікації живописних творів використано зображення картини «Соняшники» Вінсента ван Гога. Це полотно є одним із найвідоміших творів постімпресіонізму та входить до переліку об’єктів тестового набору даних, сформованого у другому розділі роботи, що дозволяє безпосередньо зіставити результат розпізнавання з верифікованими еталонними метаданими.

Після завантаження зображення та натискання кнопки «Розпізнати експонат» система сформувала мультимодальний запит до Google Gemini Vision API та отримала структуровану відповідь. У поле «Назва» виведено коректне значення «Соняшники», поле автора та періоду містить ім’я автора та часовий діапазон 1888–1889 років, що відповідає задокументованому часу створення кількох версій цієї серії. Поле «Тип» заповнено значенням «Картина», поле матеріалу та техніки відображає «Олія на полотні». У полі місця зберігання наведено перелік музейних установ, у яких зберігаються різні варіанти серії, зокрема Національна галерея Лондона, Музей ван Гога в Амстердамі та Нова пінакотека в Мюнхені. Поле цікавих фактів містить нетривіальну інформацію про серію, зокрема про кількість версій, написаних художником, та про те, що одна з картин була продана на аукціоні за рекордну суму. Поле «Екскурсовод» містить живий опис від першої особи, що передає характерний стиль та емоційний тон екскурсійного наративу. Паралельно з відображенням текстових результатів система автоматично відтворила синтезований аудіосупровід, сформований на основі ключових полів відповіді. Загальний вигляд інтерфейсу після завершення розпізнавання наведено на рисунку 3.2.

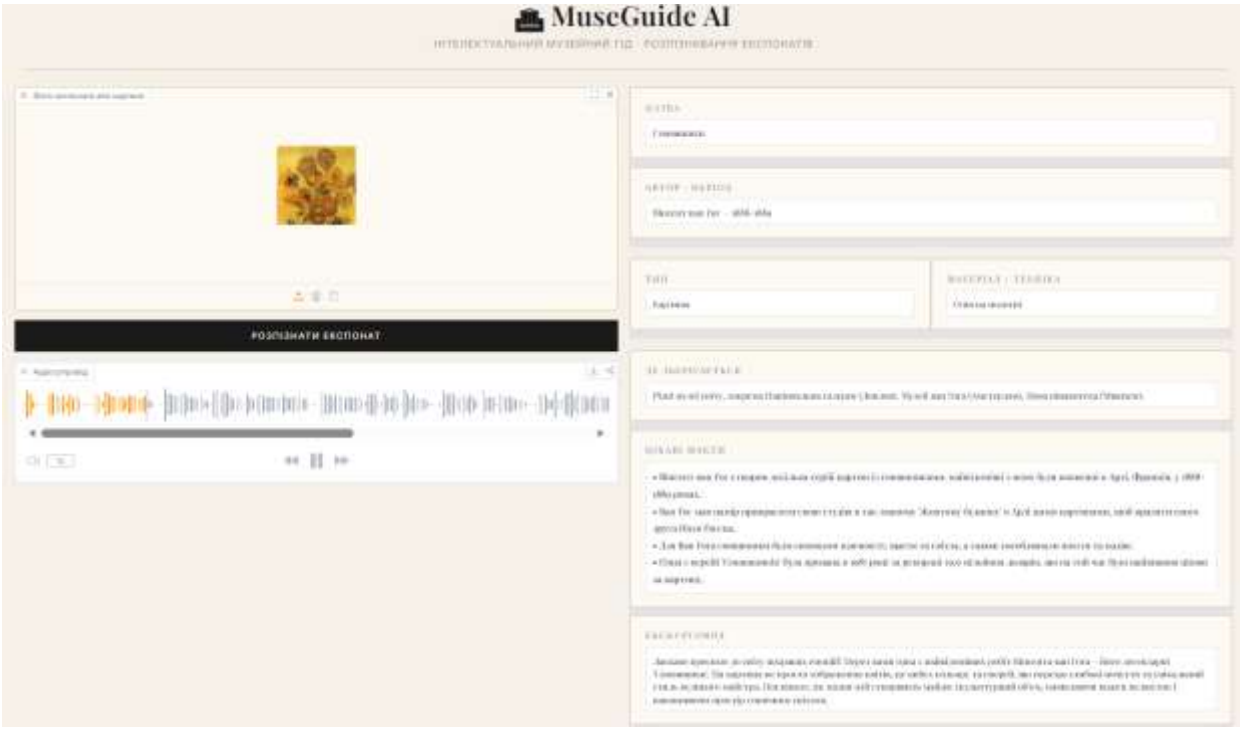

Рисунок 3.2 – Результат розпізнавання картини «Соняшники» Вінсента ван Гога

3.8.2 Сценарій розпізнавання скульптури та артефакту

Для перевірки здатності системи до ідентифікації об’єктів різних типів у межах цього сценарію виконано два послідовних тести. Перший тест використовує зображення скульптури «Венера Мілоська», другий охоплює ювелірний артефакт давньогрецького походження.

При завантаженні фотографії «Венери Мілоської» система коректно заповнила поле назви відповідним значенням, поле типу набуло значення «Скульптура», поле матеріалу містить значення «Мармур». Поле автора або культури відображає атрибуцію невідомому давньогрецькому скульпторові з зазначенням орієнтовного часового періоду у вигляді II століття до нашої ери у відповідності з загальноприйнятою науковою датацією. У полі місця зберігання наведено коректне значення Музею Лувр у Парижі, Франція. Екскурсійний опис передає атмосферу зустрічі з оригіналом твору та

акцентує увагу на дискусії щодо первісного положення рук статуї. Результат цього тесту наведено на рисунку 3.3.

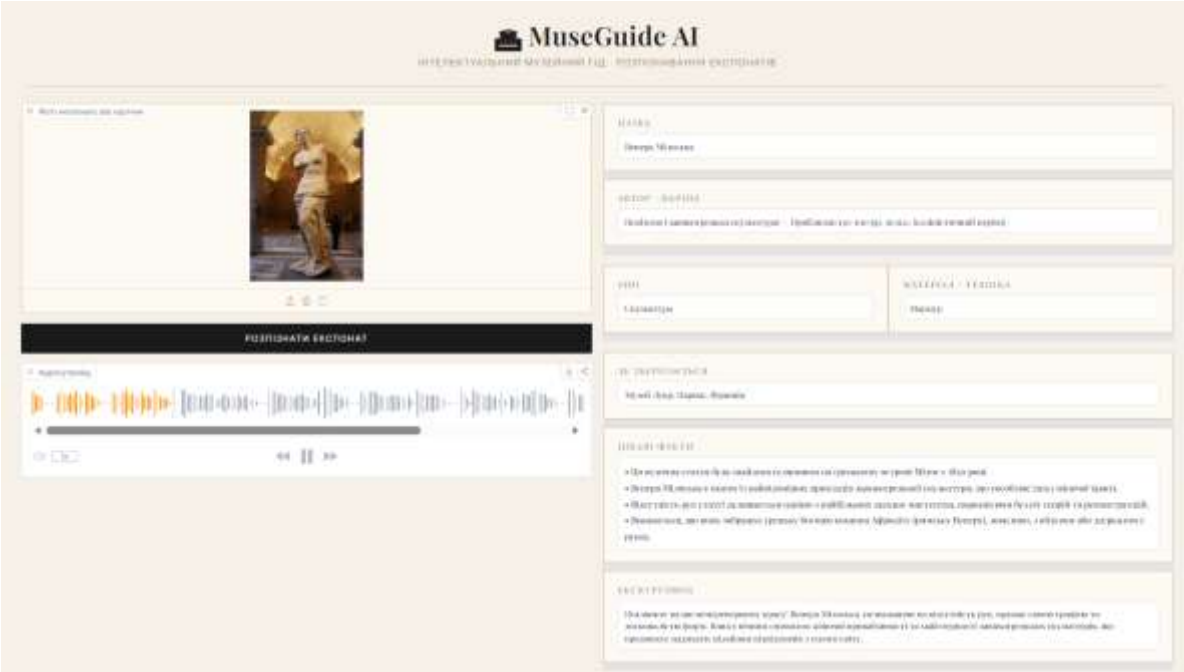

Рисунок 3.3 – Результат розпізнавання скульптури «Венера Мілоська»

Другий тест цього сценарію використовує зображення скіфської пекторалі з Товстої Могили. Цей артефакт є представником категорії маловідомих регіональних об’єктів, ідентифікація яких вимагає від моделі глибших енциклопедичних знань за межами загальновживаних мистецьких канонів. Система коректно визначила тип об’єкта як «Ювелірний виріб», атрибутувала його до скіфської культури та вказала датування IV сторіччя до нашої ери. Поле матеріалу та техніки відображає золото, лиття та філігрань, що відповідає задокументованій техніці виготовлення. Місцем зберігання зазначено Музей історичних коштовностей України у Києві. Цей результат підтверджує здатність системи опрацьовувати не лише загальновідомі шедеври, але й вузькоспеціалізовані культурні артефакти, що є критично важливим для реального музейного використання. Результат ідентифікації наведено на рисунку 3.4.

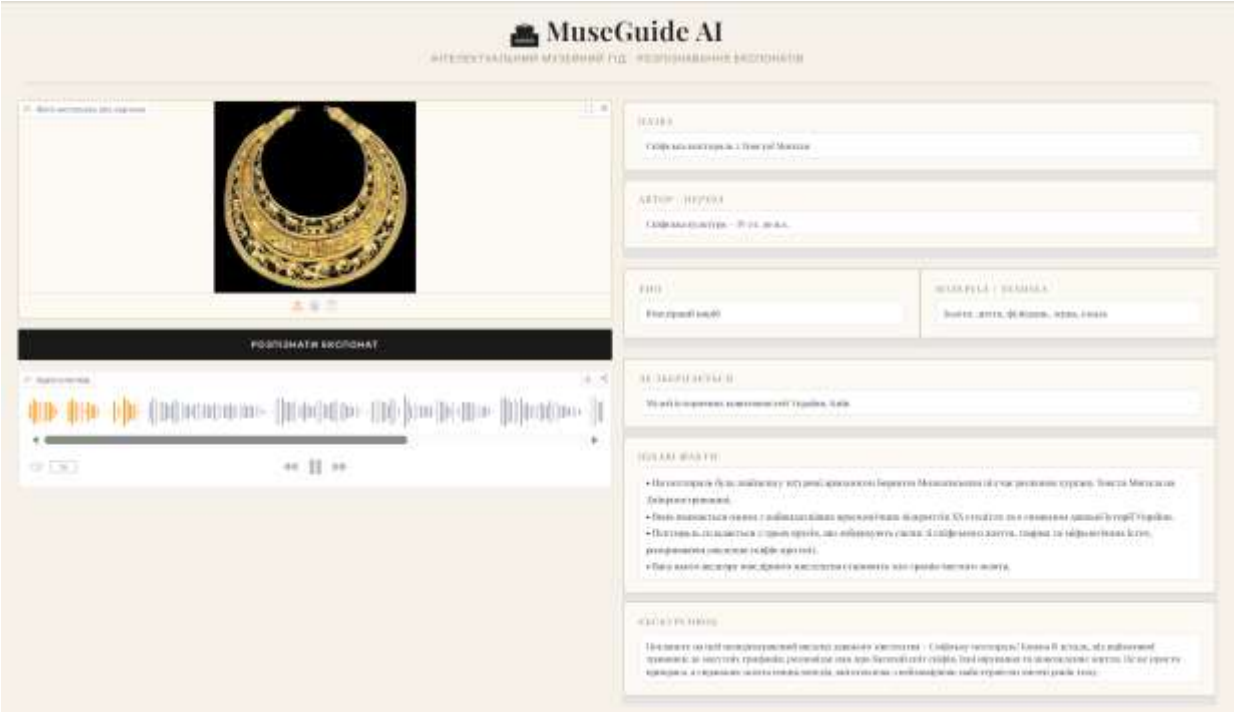

Рисунок 3.4 – Результат розпізнавання скіфської пекторалі з Товстої Могили

3.8.3 Сценарій обробки нерелевантного зображення

Для перевірки механізму фільтрації до системи завантажено фотографію інтер’єру музейної зали з розмитими скульптурами на задньому плані без чіткого головного об’єкта. Це граничний тестовий випадок: зображення зроблено всередині музею, однак не містить ідентифікованого конкретного експоната у фокусі.

Модель мультимодального аналізу встановила значення поля is_exhibit рівним false, що спровокувало генерування виключення NotAnExhibitError. Gradio-інтерфейс перехопив це виключення та відобразив у верхньому правому куті сторінки повідомлення зі змістом «На фото не схоже на музейний експонат, картину або архітектурну пам’ятку. Будь ласка, завантажте фото музейного експоната, картини або архітектурної споруди.» Усі вихідні поля залишились у стані Error без відображення жодних згенерованих даних, а аудіосупровід не відтворився. Загальна робота застосунку не була перервана: після закриття сповіщення інтерфейс

залишився повністю функціональним і готовим до нового завантаження. Цей результат підтверджує коректне функціонування механізму раннього виявлення помилки та відповідність системи вимозі точності фільтрації не нижче 95 відсотків. Поведінку інтерфейсу при нерелевантному зображенні наведено на рисунку 3.5.

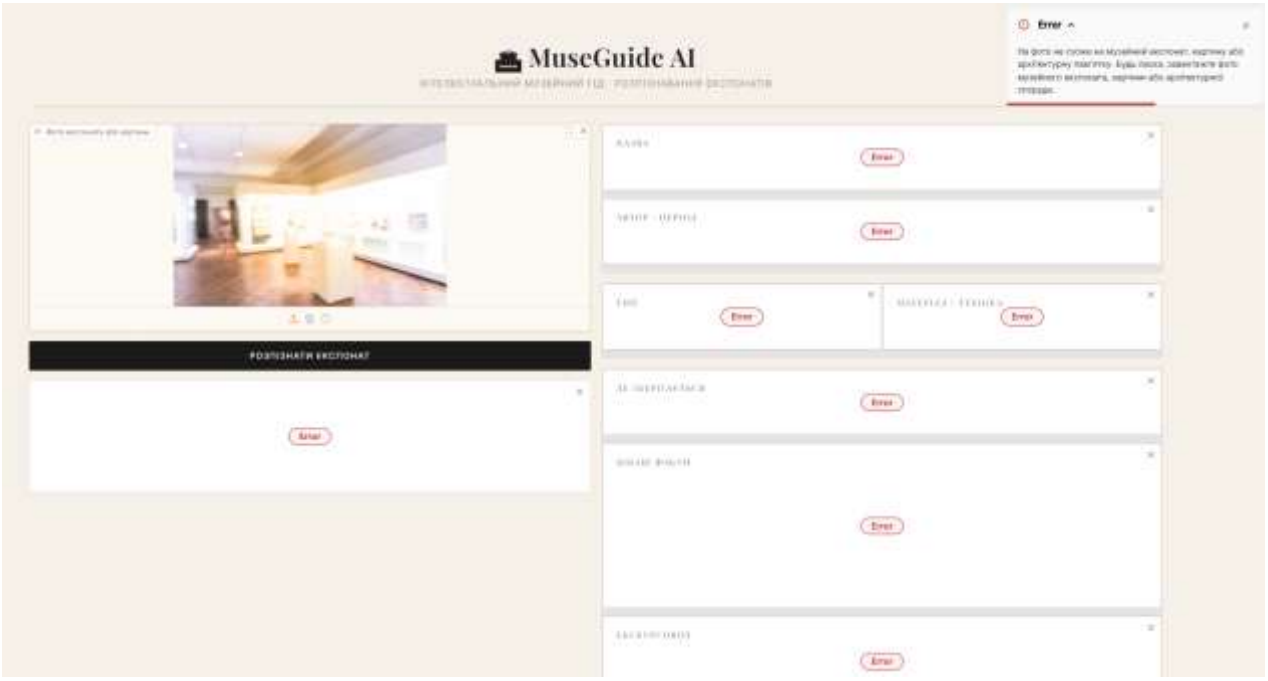

Рисунок 3.5 – Реакція системи на нерелевантне зображення

3.9 Плани щодо подальшого розвитку системи

Аналіз функціональних обмежень поточної версії системи MuseGuide AI та потреб цільової аудиторії визначає кілька напрямів подальшого розвитку, послідовна реалізація яких підвищить практичну цінність платформи та розширить її потенційну аудиторію.

Першим та найбільш пріоритетним напрямом є розширення мовної підтримки синтезу мовлення. Поточна версія генерує аудіосупровід виключно українською мовою, що є природним вибором для вітчизняних музеїв. Водночас значна частина відвідувачів великих культурних закладів є

іноземними туристами, для яких ефективний аудіосупровід їхньою рідною мовою став би суттєвою перевагою. Бібліотека edge-tts підтримує широкий спектр мов, серед яких англійська, французька, німецька, іспанська та китайська, а відповідні голосові моделі вже наявні у хмарній інфраструктурі Microsoft. Технічно додавання мовного перемикача потребує розширення системної інструкції для генерування описів відповідною мовою та передачі відповідного голосового ідентифікатора у функцію text_to_speech.

Другим напрямом є реалізація режиму роботи з камерою пристрою у реальному часі. У поточній версії користувач завантажує збережене зображення, тоді як більш природним сценарієм у музейному середовищі є наведення камери на експонат та отримання інформації без додаткових кроків. Gradio підтримує компонент Image із параметром sources, що включає live-камеру, і дозволяє реалізувати цей сценарій без зміни базової архітектури модулів розпізнавання та синтезу мовлення.

Третім напрямом є впровадження журналу відвідань, що зберігатиме історію переглянутих експонатів протягом сесії. Такий журнал дозволив би відвідувачу переглянути інформацію про раніше сфотографовані об’єкти без повторного надсилання зображення, а також отримати узагальнений звіт про відвіданий маршрут наприкінці екскурсії. Технічно ця функція реалізується через механізм стану Gradio State, що зберігає список об’єктів між викликами функції-обробника без необхідності змін у серверній частині.

Четвертим напрямом є інтеграція з музейними базами знань для підвищення точності ідентифікації специфічних об’єктів конкретної установи. Поточна версія покладається виключно на знання, закодовані у вагах мультимодальної великої лінгвістичної моделі, що є оптимальним рішенням для широковідомих творів. Для місцевих артефактів або маловідомих регіональних пам’яток точність може бути підвищена через механізм доповнення генеративного виведення знаннями із зовнішніх джерел. У цьому підході до запиту моделі додається відповідний текстовий

контекст із кураторської бази знань, що дозволяє генерувати значно точніші описи без необхідності повного донавчання моделі [26].

П’ятим напрямом є розробка системи зворотного зв’язку від відвідувачів. Додавання кнопок оцінювання корисності відповіді та можливості повідомити про неточність безпосередньо з інтерфейсу забезпечить безперервний моніторинг якості та дозволить формувати масив помилкових випадків для подальшого вдосконалення системи [27]. Ці дані могли б бути використані для точного налаштування системної інструкції або для виявлення тематичних груп об’єктів, ідентифікація яких потребує додаткового контекстного підсилення у рамках підходу retrieval augmented generation [28].

Реалізація перелічених напрямів розвитку дозволить трансформувати поточний прототип системи у повноцінний мобільний продукт, що може бути розгорнутий у реальних музейних умовах та забезпечить відвідувачам принципово новий рівень взаємодії з культурною спадщиною [29, 30].

ВИСНОВКИ

У рамках кваліфікаційної роботи було створено вебзастосунок MuseGuide AI, що дозволяє автоматично ідентифікувати музейні експонати за фотографією та генерувати екскурсійний опис з аудіосупроводом із використанням сучасних мультимодальних великих лінгвістичних моделей.

Основною метою роботи була розробка інтелектуальної системи для персоналізованого супроводу відвідувачів музеїв, при цьому вирішено такі завдання:

– проведено аналіз сучасного стану та тенденцій розроблення інтелектуальних систем для музейного супроводу на основі комп'ютерного зору;

– здійснено огляд існуючих комерційних та академічних рішень, зокрема Smartify, Bloomberg Connects, izi.TRAVEL та PEACH, і виявлено незаповнену нішу для комплексного рішення;

– сформовано власний еталонний набір даних з п'яти верифікованих художніх творів різного ступеня складності;

– виконано порівняльний аналіз трьох мультимодальних великих лінгвістичних моделей: Gemini 2.5 Flash, Gemini 3.1 Pro та GPT-5.5;

– обрано оптимальний технологічний стек та спроєктовано архітектуру системи на основі Google Gemini Vision API, edge-tts та Gradio;

– розроблено алгоритм розпізнавання музейних експонатів та генерування екскурсійного опису;

– спроєктовано та реалізовано вебзастосунок MuseGuide AI з інтерфейсом у музейному стилі.

Проведений порівняльний аналіз показав глибоку асиметрію у продуктивності моделей: моделі сімейства Gemini продемонстрували абсолютну точність фактичної ідентифікації при нульовому рівні помилок, тоді як GPT-5.5 виявив критично високий рівень галюцинацій та хибних відмов, що стало підставою для її виключення з фінального відбору.

Під час оцінки якості генерованого контенту десятьма профільними експертами у галузі туризму та музейної справи модель Gemini 2.5 Flash отримала перевагу у 62% випадків завдяки глибшому розкриттю маловідомих аспектів, відсутності мовних кліше та більш логічній структурі екскурсійного наративу. З огляду на це у розробленому застосунку було використано саме її.

Результати кваліфікаційної роботи можуть бути корисними у сфері автоматизованого музейного супроводу, туристичних сервісів та культурно-освітніх закладів. У подальшому є сенс допрацювати систему, додавши підтримку багатомовного синтезу мовлення, режим роботи з живою камерою та інтеграцію з музейними базами знань.

Результати роботи апробовано у вигляді 3 тез доповідей: під час II Міжнародної науково-практичної студентської конференції [1], під час XVI Міжнародної науково-технічної конференції «Інформаційно-комп’ютерні технології» [2], під час Міжнародного молодіжного форуму «Радіоелектроніка і молодь у XXI столітті» [3].

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Камишан П.С., Кобилін І.О. (2025) Комбінований підхід комп'ютерного зору та оптичного розпізнавання тексту для підвищення точності ідентифікації зображень. *ІІ Міжнародна науково-практична студентська конференція,15 жовтня 2026 року, Харків, Україна*.

2. Камишан П.С., Кобилін І.О. (2026) Моделі багатокритеріального ранжування релевантності інформаційних блоків у мобільних системах супроводу. *XVI Міжнародна науково-технічна конференція "Інформаційно-комп'ютерні технології"02-03 квітня 2026 року, Житомир, Україна*.

3. Камишан П.С., Кобилін І.О. (2026) Архітектура мультимедійної системи персоналізованої навігації на основі семантичної розмітки об'єктів. *30-й Міжнародний молодіжний форум «Радіоелектроніка і молодь у XXI столітті». Збірник матеріалів форуму, 22–24 квітня 2026 року, Харків, Україна*.

4. Redmon J., Divvala S., Girshick R., and Farhadi A. (2016) You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779-788.

5. He K., Gkioxari G., Dollár P., and Girshick R. (2017) Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2961-2969.

6. ICOM – International Council of Museums. (2022). *Museum definition and digital transformation report*. Paris: ICOM.

7. Кобилін І.О., Харченко А.І. (2024) Classification techniques for computer vision, *Системи обробки інформації*, 3(178), С. 33-41.

8. Shi B., Wang X., and Lv P. (2016) An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11), pp. 2298-2304.

9. Боденчук-Пастухов Є.В., Кобилін І.О. (2026) Оцінка моделей трансформерів машинного перекладу на низькоресурсних азійсько-українських мовних парах, *Системи обробки інформації*, 1(184), С. 116-123.

10. Machuca E., and Mandow L. (2016) Multiobjective shortest path problems: A survey, *European Journal of Operational Research*, 256(1), pp. 1-12.

11. Smartify. About Smartify: The art & culture app. URL: https://smartify.org/about (дата звернення 25.04.2026).

12. Kandara O., Ech-Cherif A., and El-Beze M. (2021) Deep learning-based artwork classification for mobile museum applications, *International Journal of Advanced Computer Science and Applications*, 12(8), pp. 443-452.

13. Ultralytics. YOLOv8: You Only Look Once Version 8. URL: https://github.com/ultralytics/ultralytics (дата звернення 20.04.2025).

14. Bohdan N., Tvoroshenko I., Gorokhovatsky V., and Kobylin O. (2025) Development of a hybrid method to enhance context memory for a chatbot application based on large language models, *International Journal of Academic Information Systems Research*, 9(10), pp. 7-18.

15. Suprun A., Tvoroshenko I., Gorokhovatsky V., and Yakovleva O. (2025) Development and research of a method for the combined use of large language models for text generation, *International Journal of Academic and Applied Research*, 9(10), pp. 249-263.

16. Gorokhovatsky V., Tvoroshenko I., and Yakovleva O. (2024) Transforming image descriptions as a set of descriptors to construct classification features, *Indonesian Journal of Electrical Engineering and Computer Science*, 33(1), pp. 113-125.

17. Yue X., Ni Y., Zhang K., et al. (2024) MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556-9567.

18. Team Gemini, Anil R., Borgeaud S., et al. (2023) Gemini: A family of highly capable multimodal models. arXiv:2312.11805. URL: https://arxiv.org/abs/2312.11805 (дата звернення 30.04.2026).

19. Reid M., Savinov N., Teplyashin D., et al. (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530. URL: https://arxiv.org/abs/2403.05530 (дата звернення 24.04.2026).

20. OpenAI. (2023) GPT-4 technical report. arXiv:2303.08774. URL: https://arxiv.org/abs/2303.08774 (дата звернення 30.04.2026).

21. Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. URL: https://www.anthropic.com/news/claude-3-family (дата звернення 30.04.2026).

22. Meta AI. Meta Llama 3.2: Multimodal models for every scale. URL: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/ (дата звернення 30.04.2026).

23. Кобилін, О.А., & Творошенко, І.С. (2021). Методи цифрової обробки зображень: навч. посібник. Харків: ХНУРЕ.

24. Gorokhovatsky V., Tvoroshenko I., Kobylin O., and Vlasenko N. (2023) Search for visual objects by request in the form of a cluster representation for the structural image description, *Advances in Electrical and Electronic Engineering*, 21(1), pp. 19-27.

25. Abid A., Abdalla A., Abid A., Khan D., Alfozan A., and Zou J. (2019) Gradio: Hassle-free sharing and testing of ML models in the wild. arXiv:1906.02569. URL: https://arxiv.org/abs/1906.02569 (дата звернення 01.05.2026).

26. Daradkeh Y.I., Gorokhovatsky V., Tvoroshenko I., Gadetska S., and Al-Dhaifallah M. (2023) Statistical data analysis models for determining the relevance of structural image descriptions, *IEEE Access*, 11, pp. 126938-126949.

27. Гороховатський В., Творошенко І. (2026) Продуктивні моделі аналізу даних у методах розпізнавання зображень: монографія. Харків: ХНУРЕ, 153 с.

28. Daradkeh Y.I., Tvoroshenko I., Gorokhovatsky V., Latiff L.A., and Ahmad N. (2021) Development of effective methods for structural image recognition using the principles of data granulation and apparatus of fuzzy logic, *IEEE Access*, 9, pp. 13417-13428.

29. Gorokhovatsky V., Tvoroshenko I. (2023) Identification of visual objects by the search request. *International scientific symposium «INTELLIGENT SOLUTIONS-S». Computational intelligence (results, problems and perspectives). Decision making theory: proceedings of the international symposium, September 28, 2023, Kyiv-Uzhorod, Ukraine*, pp. 25-27.

30. Tvoroshenko I., Gorokhovatsky V., Kobylin O., and Tvoroshenko A. (2023) Application of deep learning methods for recognizing and classifying culinary dishes in images, *International Journal of Academic and Applied Research*, 7(9), pp. 57-70.