**International Journal of Academic and Applied Research (IJAAR)**
**ISSN: 2643-9603**
**Vol. 9 Issue 10 October - 2025, Pages: 249-263**

# Development and Research of a Method for the Combined Use of Large Language Models for Text Generation

**Anna Suprun, Iryna Tvoroshenko, Volodymyr Gorokhovatskyi, Olena Yakovleva**

Department of Informatics
Kharkiv National University of Radio Electronics
Kharkiv, Ukraine
e-mail: anna.suprun@nure.ua; iryna.tvoroshenko@nure.ua; volodymyr.horokhovatskyi@nure.ua; olena.yakovleva@nure.ua

***Abstract:*** *This paper presents a comparative analysis of three advanced large language models (GPT-4o, Claude 4 Opus, and Gemini 2.5 Pro) applied to creative text analysis and generation tasks. The study examines each model's performance across multiple narrative scenarios of varying complexity, assessing both analytical accuracy and literary expressiveness using integrated linguistic, logical, and stylistic metrics. Experimental evaluation showed that Claude 4 Opus achieved the highest analytical consistency with minimal hallucination rate and strong logical reasoning, while Gemini 2.5 Pro excelled in generation quality, demonstrating superior stylistic coherence, emotional depth, and grammatical precision. GPT-4o, in turn, maintained high contextual completeness but revealed a tendency toward interpretive creativity and higher variance in factual precision. Building on these findings, a new method for combined utilization of LLMs was developed and tested. In this approach, Claude 4 Opus serves as the analytic module, performing structured narrative decomposition and contextual synthesis, while Gemini 2.5 Pro acts as the generative module, transforming the processed analytical output into artistically refined text. Experimental validation demonstrated that the proposed method achieved an average generation quality index that surpassed each model's individual results in coherence, emotional integrity, and stylistic harmony. These outcomes confirm the effectiveness of inter-model collaboration for enhancing both analytical precision and creative depth in LLM-based literary text generation, offering a promising direction for future hybrid human–AI creative systems.*

**Keywords—artificial intelligence; Claude 4 Opus; contextual coherence; Gemini 2.5 Pro; generation quality assessment; GPT-4o; inter-model collaboration; large language models; narrative analysis; prompt engineering; text analysis; text generation**

## 1. INTRODUCTION

The rapid advancement of large language models (LLMs) has moved artificial intelligence beyond purely academic research into widespread application across various fields. A domain of significant interest is their potential within creative endeavors, specifically for the analysis and generation of literary texts. Traditional methods of literary analysis are highly resource-intensive, and the creation of new artistic work has historically been the sole prerogative of the human author. LLMs, with their ability to process massive datasets, understand context, and produce coherent language, offer novel tools for exploring the deeper structures of literature and experimentally generating plots, dialogue, and character descriptions [1].

In the context of creative writing, a critical question remains: to what extent are LLMs capable of genuinely understanding complex literary concepts? This includes elements like character psychology, internal conflicts, and world logic, and how realistically they can embody these concepts in their generated output. A fundamental challenge is the models' tendency to create superficial or template-driven narratives. Although proficient at replicating style and vocabulary, even the most advanced LLMs often struggle to fully capture the emotional depth, psychological nuance, and multilayered thematic coherence essential for high-quality artistic prose.

LLMs are a class of neural networks developed for natural language processing and generation, distinguished by their massive scale – ranging from billions to trillions of parameters – and their reliance on extensive training data. Key determinants of an LLM's effectiveness include its parameter count, the volume and diversity of its training data, and the context window length, which defines the amount of prior text the model can consider during generation. These models function not merely as language processors but as universal AI systems capable of content generation, context analysis, and task adaptation, making their systematic evaluation highly pertinent. They represent an intellectual instrument for inspiration, taking on routine or technical tasks (like generating various scene options), while the human writer retains control over the artistic vision [2].

The research focuses on a comparative analysis of three leading LLMs: GPT-4o (OpenAI), Claude 4 Opus (Anthropic), and Gemini 2.5 Pro (Google DeepMind) – for their efficacy in creative text analysis and generation based on user-provided context. The selection of these models is dictated by their cutting-edge performance and complex context processing. GPT-4o, a multimodal model, is noted for its high accuracy in character psychology and style adherence.

Claude 4 Opus, optimized for multi-step reasoning and structured analysis, provides a benchmark to contrast logical integrity with artistic fluency. Gemini 2.5 Pro, combining deep reasoning with multimodal integration, is effective for generating complex, interwoven plot scenarios [3, 4].

The object of the research is large language models in the tasks of analysis and generation of literary texts.

The objective of the research is the comparison of the capabilities of GPT-4o, Claude 4 Opus, and Gemini 2.5 Pro models, and the development of a methodology that integrates analytical and generative strengths of different LLMs to achieve higher-quality, contextually coherent, and artistically expressive text generation.

## 2. Related Works

To ensure a comprehensive understanding of the research problem, an analysis of scientific sources regarding the use of LLMs for creating and analyzing artistic texts, as well as evaluating their narrative and creative abilities, was conducted.

The study in [5] compared the ability of commercial (GPT-4, Claude 1.2, Gemini 1.5) and open-source LLMs in English-language creative writing (an epic combat description), assessing originality, style, and fluency. It established that some commercial models can equal or surpass human writers, while open models significantly lag. This source provides a valuable evaluation methodology utilizing human expert assessment for tasks that cannot be replicated from training data. In [6], the authors explored the process of human-LLM co-creation during text preparation. Three iterative stages (ideation, elaboration, and realization) were identified, where the dominant role is retained by the human, though initiative shifts between the partners. The work provides a methodology for assessing user-LLM interaction within the creative process.

The research [7] investigated the ability of LLMs to exhibit divergent thinking and semantic diversity (using the Divergent Association Task). It was found that modern models can exceed the average human level but do not reach the capabilities of the most creative authors. This confirms the potential of LLMs in creativity but highlights limitations regarding originality. The paper [8] evaluated the creative ability of LLMs in short fiction writing (stories based on five keywords) using both automated and human metrics. LLMs were found to generate stylistically complex texts but fell short of humans in novelty and diversity. Crucially, the perception of creative quality for machine-generated texts differed based on the evaluator's level of expertise.

The study [9] demonstrated the use of ChatGPT as a creative writing support tool via a multi-vocal prompting technique (where the model simultaneously acts as author and critic). The main conclusion is that the sophistication and complexity of the text directly depend on the structure and quality of the prompt, providing a basis for developing effective interaction strategies.

The work in [10] compared the creative writing of models (BART-large, GPT-3.5, GPT-4o) with human authors. BART-large, overall, surpassed amateur writers with an average skill level in fiction writing, while GPT-4o demonstrated high coherence but was more predictable.

In [11], the summarization of short stories by GPT-4, Claude 2.1, and LLaMA 2 70B was assessed. All models made errors (over 50%), particularly in interpreting subtext and conveying details. The involvement of writers as expert reviewers underscores the limitations of LLMs in comprehending complex narrative structures.

The research [12] proposed a new methodology for evaluating LLM creativity based on modified Torrance Tests of Creative Thinking. Testing models (GPT-4, Claude 2, Gemini-1.5 Pro) against criteria like flexibility, originality, fluency, and elaboration showed that role-playing scenarios and correctly constructed prompts significantly enhance creativity, and combining multiple models helps compensate for individual limitations.

Article [13] compared Gemini and ChatGPT, focusing on architectural features and performance across various (non-creative) domains. It highlighted Gemini's innovative training approaches and ChatGPT's stability in long-term dialogues, providing context on how architecture influences context coherence in creative text. The paper [14] investigated the linguistic abilities of Gemini Pro versus GPT-3.5 Turbo across 10 tasks (logic, knowledge, translation). Gemini Pro slightly lagged in overall accuracy but was stronger in working with non-English languages and processing long, complex logical chains. This justifies its use for multilingual fiction and plot logic. Source [15] explored the factuality of long texts generated by advanced LLMs (GPT-4, Gemini 1.5 Pro, Claude 3 Opus). It was found that information accuracy decreases in later sentences, and the models' ability for self-correction is limited. This suggests the need for strategies to control the logic and factuality of long narrative sequences.

The study in [16] analyzed GPT-4's ability to emulate the style of H. P. Lovecraft. Respondents were unable to reliably distinguish between generated and original texts, confirming the effectiveness of prompt engineering in controlling text style and content.

The literature review confirms that contemporary LLMs are effective support tools for writers, capable of generating plot ideas and dialogues. However, they possess significant limitations: insufficient style individualization, a tendency toward predictable plots, and a restricted ability to create deep internal conflicts and multi-layered narrative structures, especially in long-form texts. These findings underscore the relevance of further research to enhance the quality of artistic generation and develop methods for human-AI co-creation and model combination.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 9 Issue 10 October - 2025, Pages: 249-263

## 3. Characteristics and Application of Selected LLMs in Creative Writing

GPT-4o from OpenAI (where "o" stands for "omni") represents the philosophy of SOTA (State-of-the-Art) and universality, aiming for peak performance across all domains, including complex creative generation. Its architectural advantage lies in native multimodality, allowing for the deep integration of various inputs (text, audio, image). GPT-4o's main value in creative work is its stylistic flexibility and prose quality. Trained on a massive, high-quality data, it can emulate a wide range of literary styles and generate emotionally rich text, making it a good candidate for generating the final artistic prose in research scenarios [3, 17]. A potential weakness is that its universality and tuning for "correct" responses may dull the uniqueness or edge of a character if explicit instructions regarding contradictory motivations are not provided.

Claude 4 Opus from Anthropic is built on a philosophy of safety and deep contextual reasoning. For creative text, it is better suited as a logical analyst capable of deeply understanding and reproducing the internal rules of a fictional world and character motivations. Its technical advantage is an extended context window (around 200K tokens), which is critical for maintaining narrative integrity across large inputs [18, 19]. The model exhibits exceptional Chain-of-Thought reasoning, enabling it to plan the internal logic of an analysis (e.g., predicting a character's reaction by sequentially assessing their motivation and external trigger). A potential drawback is that its safety philosophy might lead to "cautious" generation, smoothing over artistically justified, but controversial, character actions.

Gemini 2.5 Pro from Google DeepMind combines high reasoning efficiency with multimodality and an unprecedented context window of up to 1 million tokens. This advantage allows the model to analyze and generate text within the context of an entire novel, crucial for artistic coherence and integrating all user-provided data (descriptions, rules, history) [20–22]. Gemini is a powerful scene planner and world logic analysis tool due to its multi-step reasoning. The main risk is the need for empirical verification of its generation quality and stylistic sophistication in the target language (Ukrainian) compared to established competitors [4, 23].

The selected models share several fundamental common features: they are all built upon the Transformer architecture (multi-head attention), are multimodal, have undergone additional tuning via Reinforcement Learning (RLHF/Constitutional AI), and are universal generative models capable of creative writing [24–26]. However, unique differences in their architectural optimizations and focus on Reasoning provide the basis for the empirical study of their complementary properties and potential synergy in complex creative tasks.

## 4. Developing Methods for Literary Text Analysis and Generation Using Large Language Models

For an objective comparative study of the selected LLMs in the context of creative text analysis and generation, a standardized set of input data is essential. This dataset must simulate real-world conditions where a user (writer) provides fragmented and unstructured descriptions of the world, characters, and plot. The quality of LLM output directly depends on their ability to perform effective knowledge retrieval and correctly integrate these key elements into the final generated scene [27, 28]. The input dataset is composed scenarios, each containing two main elements: unstructured contexts and a structured user query (prompt).

### 4.1 Formalization of Input Contexts

The input contexts are formalized as a set of unstructured descriptions $D$, containing all necessary information about the fictional world and its elements. The unstructured nature of these descriptions is chosen to empirically test the effectiveness of the Retrieval-Augmented Generation and Reasoning mechanisms within the tested LLMs [29, 30]. The set of unstructured descriptions $D$ is defined as the union of subsets describing key aspects of the narrative, as in

$$D = \{D_{char}, D_{loc}, D_{rule}, D_{story}\}, \quad (1)$$

where $D_{char}$ is the set of character descriptions, including motivations and psychological characteristics:

$$D_{char} = \{d_1^{char}, \dots, d_{n_c}^{char}\}; \quad (2)$$

$D_{loc}$ is the set of location descriptions, covering physical and atmospheric features:

$$D_{loc} = \{d_1^{loc}, \dots, d_{n_l}^{loc}\}; \quad (3)$$

$D_{rule}$ is the set of worldbuilding rules (e.g., laws of physics, magic systems, social rules) [31]:

$$D_{rule} = \{d_1^{rule}, \dots, d_{n_r}^{rule}\}; \quad (4)$$

$D_{story}$ is a generalized summary of the plot events that occurred prior to the scene being generated, ensuring logical sequence and narrative coherence.

### 4.2 Structure of the User Query (Prompt)

The user query $q$ contains the target instruction for scene generation and controlling parameters. Its components are designed to test the model's ability to execute complex, multi-faceted requirements. The query $q$ has the following structural components:

$$q = \{g, s, C_{must}, C_{avoid}, P\}, \quad (5)$$

where $g$ is the generation goal, defining the scene that must occur; $s$ is the desired tone and genre; $C_{must}$ are the mandatory conditions and/or events that must take place in the scene; $C_{avoid}$ are the prohibitions and/or restrictions for the scene; $P$ are the generation parameters.

### 4.3 Obtaining Analytical and Creative Outputs from LLMs

A method for obtaining analysis and generation results from the LLMs is developed to subsequently identify the best model for each task. The methodology is based on the Zero-shot Learning approach, utilizing a general input prompt containing context $D$ and user query $q$. This minimizes the influence of additional fine-tuning, allowing an objective assessment of the models' intrinsic capabilities based solely on their generalized knowledge [32]. Initially, the input prompt is formed by combining a specialized system prompt, the context $D$, and the user query $q$ into a single, identical large prompt for all three models.

The system prompt for analysis $Sp_A$ is designed to enforce a structured, analytical output:

"You are an analytical model whose task is to extract and structure all key facts from any provided text. Adhere to the output rules. Document structure: each main category starts with a heading ('Entity: Name'). Facts within each category must be atomic key-value pairs ('Key: Value'). Emotion, character, and psychological aspects must be recorded as separate facts. Following the factual extraction, a separate section for Possible Actions must be added. For each character, list possible actions (consistent with their character, backstory, and world rules) within the scene requested by the user ('Possible Actions: Character Name'). Do not add any free text, summaries, or interpretations; output only facts."

The system prompt for generation $Sp_G$ is designed to elicit high-quality creative output: "You are a literary author. Your goal is to create high-quality artistic scenes based on the user's request. Adhere to the following requirements: the text must be cohesive and logical; use an artistic, evocative style, not technical description; the tone must match the user's request; character dialogues must be vivid and reveal their personalities; do not exceed the limits of the user's request; and always conclude the scene with a logical endpoint. Write an artistic scene that meets the user's request."

Testing is performed for each model in the set:

$$\{LMM_1, LLM_2, LLM_3\}, \quad (6)$$

where $LMM_1$ is GPT-4o, $LLM_2$ is Claude 4 Opus, and $LLM_3$ is Gemini 2.5 Pro.

The model receives the corresponding formatted prompt and provides either the analysis or the generated text.

Analysis output ($O_A$) is defined by the formula below

$$O_A = LLM(Sp_A, D, q). \quad (7)$$

Generation Output ($O_G$) is defined by the formula below

$$O_G = LLM(Sp_G, D, q). \quad (8)$$

The results are stored for subsequent quality evaluation.

### 4.4 Context Analysis Quality Assessment Method

The methodology for assessing context analysis quality is aimed at objectively measuring the ability of LLMs to effectively extract, structure, and logically harmonize key knowledge from the provided prompt.

The resulting assessment will indicate whether there is room for improvement in the existing method of obtaining text analysis from the models. If so, it will serve to identify the best Analyst Model capable of ensuring the psychological authenticity of characters and the logic of the world in the subsequently proposed method for combining the work of models. The assessment is carried out using automated tools by comparing the model's structured output list $O_A$ with a reference analysis – a list created by an expert containing ideally extracted facts and logical connections. The quality of analysis is measured as a comprehensive indicator that combines six scientifically recognized metrics covering the completeness, reliability, and relevance of the extracted information.

The following key metrics will be utilized:

1. Context Coverage Score. This metric is an indicator of the effectiveness of the model's RAG mechanism and its ability to utilize a long context window. It is calculated by the formula

$$Coverage = \frac{Facts_{extracted} \cap Facts_{ref}}{Facts_{ref}}, \quad (9)$$

where $Facts_{extracted}$ is the set of extracted facts; and $Facts_{ref}$ is the set of reference facts.

The metric evaluates how completely the model has extracted all key facts and rules from the input context $D$. A high score minimizes the risk of the LLM "forgetting" critically important details that should influence the plot [33].

2. Hallucination Rate. This metric determines the proportion of information in the output that has no direct confirmation in the input context $D$. This is a key indicator of the reliability of the model's analytical apparatus. It is calculated as

$$Hallucination\ Rate = \frac{neutral + contradiction}{total}, \quad (10)$$

where $neutral$ is the number of neutral (unconfirmed) statements; $contradiction$ is the number of contradictory statements; and $total$ is the total number of statements.

Models with a low Hallucination Rate demonstrate higher accuracy and reliability, which is critical for preventing plot errors (Lore Breaking) [34].

3. Entity Linking Accuracy. The metric evaluates the correctness of identifying and classifying unique entities (characters, locations, artifacts, rules).

Precision, Recall, and the $F1$-score are measured at the entity level.

Their calculation uses metrics such as True Positives ($TP$) the number of entities correctly identified by the model (found by the model and matching the reference); False Positives ($FP$)

– the number of entities incorrectly identified by the model (found by the model but absent in the reference, perhaps because the model mistakenly labeled a word as an entity or assigned an incorrect type); and False Negatives ($FN$) – the number of entities missed by the model but present in the reference.

Precision shows reliably the model identified entities. It is calculated by the formula below

$$Precision = \frac{TP}{TP+FP}. \quad (11)$$

Recall shows how completely the model identified all entities from the reference. It is calculated by the formula below

$$Recall = \frac{TP}{TP+FN}. \quad (12)$$

$F1$-score is an averaged measure, which is the main metric for the quality of entity extraction. It is calculated by the formula below

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \quad (13)$$

Additionally, experts evaluate the logicality of the possible actions identified by the models for each character in the context of the query scene $q$. Experts will provide their assessment using the Likert Scale – a psychometric scale used to measure attitudes or opinions, where respondents rate their degree of agreement or disagreement with a statement by choosing one of a fixed number of points (from 1 – "very poor" to 5 – "exceptionally good"). The average value of the expert ratings, the Actions Logic Score, will be included in the integral assessment.

The Integral Context Analysis Quality Index $Q_A$ for each model is calculated as a weighted complex index based on six metrics, allowing the identification of the best analyst model. It is defined as

$$Q_A = \omega_1 \cdot Coverage + \omega_2 \cdot Actions\ Logic + \\ + \omega_3 \cdot (1 - Hallucination\ Rate) + \\ + \omega_4 \cdot Consistency\ Score + \omega_5 \cdot F1, \quad (14)$$

where $\omega_n$ is a weighting coefficient, determined by expert judgment, reflecting the importance of each quality for identifying the best model.

### 4.5 Creative Generation Quality Assessment Method

The goal of assessing creative generation quality is to empirically determine the capacity of the selected LLMs to write complex, high-quality, and engaging literary works. Furthermore, this assessment seeks to identify the best model for this task – the Generator Model – that demonstrates the highest quality of prose, stylistic sophistication, and emotional depth, to be used in the subsequent LLM combination method. Since artistic quality is inherently subjective, the methodology's main focus is shifted to human evaluation [35].

Experts evaluate each generated scene based on five main criteria using the Likert Scale.

The following criteria will be assessed:

- Prose Quality ($P_1$) – the complexity and imagery of the language, syntactic variety, and the absence of tautologies and stylistic errors.
- Emotional Depth ($P_2$) – the text's ability to evoke an emotional response, the depth of character feelings, and adherence to the tone specified in the user query $q$.
- Psychological Plausibility ($P_3$) – the plausibility of actions and dialogues within the scene context. This assesses how logically aligned the characters' actions are with their motivations and the world's rules.
- Tone Adherence ($P_4$) – the extent to which the generated text conforms to the genre and stylistic requirements $s$ specified in the user query $q$.
- Technical Literacy ($P_5$) – the absence of spelling, punctuation, grammatical, and syntactic errors. This also includes factual consistency within the text itself (e.g., whether a character's eye color remains consistent) [36].

For each model $LLM_i$, the final Generation Quality Score ($Q_G$) is calculated. This is based on the average ratings $\overline{P_j}$ for each criterion $P_j$, where $j = 1 \dots 5$. The overall score $Q_G$ is calculated as the weighted arithmetic mean of the five criteria by the formula below

$$Q_G(LLM_i) = \sum_{j=1}^{5} \omega_j \cdot \overline{P}_j, \quad (15)$$

where $\omega_j$ is the expert-determined weighting coefficient.

### 4.6 Development and Justification of a Combined LLM Utilization Method

If it is found that no single model outperforms the others in both the analysis and generation categories, a method for the combined use of LLMs is proposed. This method is developed to overcome the limitations of individual models in performing complex, multi-faceted tasks that require both deep logical analysis and high-quality artistic generation. The method is based on an architecture that utilizes an Analyst Model and a Generator Model, which are determined by the preceding tests and evaluation methods.

This method ensures the transformation of unstructured input data ($D$, $q$) into the final artistic text $T$ through a sequence of intermediate, structured forms, applying iterative validation. The process is divided into three key stages: analysis, generation, and validation/iteration [37]:

Stage 1. Analysis: the Analyst Model ($LMM_A$) performs a logic-semantic analysis of every world element and constructs a reliable analytical scene model. For each character $c_i$, the model conducts a logic-semantic analysis of their description $d_i^{char}$ within the context of all available information. It is defined as

$$Descr_i = LLM_A\left(d_i^{char}, q, D_{loc}, D_{rule}, D_{story}\right), \quad (16)$$

where $LLM_A$ is the Analyst Model; and $Descr_i$ is a structured interpretation that aggregates critical elements to create an authentic character profile:

$$Descr_i = \{M_i, A, C_i, R_i, L_i, B_i\}, \quad (17)$$

where $M_i$ is the set of character motivations; $A$ is the description of appearance; $C_i$ is the set of character traits; $R_i$ is the role/status in the world; $L_i$ represents lexical and stylistic markers (typical vocabulary, phraseology, syntax); and $B_i$ is the character's backstory.

$LLM_A$ uses the structured description $Descr_i$ and the constraints from prompt $q$ to predict a logically consistent set of permissible actions for each character in the scene. Set of permissible actions is defined as

$$Act_i = LLM_A\left(Descr_i, q, D_{loc}, D_{rule}, D_{story}\right), \quad (18)$$

where $Act_i$ incorporates the mandatory character actions $c_{must}$ and prohibitions $c_{avoid}$ from prompt $q$. The analysis results for all characters are aggregated into a single structured scene model defined as

$$Scene = \left\{Char_1, \ldots Char_n, q, D_{loc}, D_{rule}, D_{story}\right\}, \quad (19)$$

where $Char_i$ is the complete character description.

The Analyst Model forms the final prompt ($Prompt$) for the Generator Model, using the $Scene$ model as the primary structured context. $Prompt$ is defined as

$$Prompt = LLM_A(Scene, I), \quad (20)$$

where $I$ represents additional technical instructions for generation control.

Stage 2. Creative Text Generation: the Generator Model ($LLM_G$) receives and considers all parameters of the $Scene$ and creates the text $T$ that defined as

$$T = LLM_G(Prompt). \quad (21)$$

Stage 3. Validation and Iterative Correction (Quality Control): this stage ensures that the final text $T$ meets high requirements for consistency and artistic quality by employing a validation and iteration loop. Validation is conducted using the same metrics as those employed in the creative generation quality assessment. If the quality score does not reach the predefined threshold, the text is returned for regeneration, where the model uses the previous failed output $T$ as additional context for correction.

The diagram for the combined LLM usage method is depicted in Fig. 1.

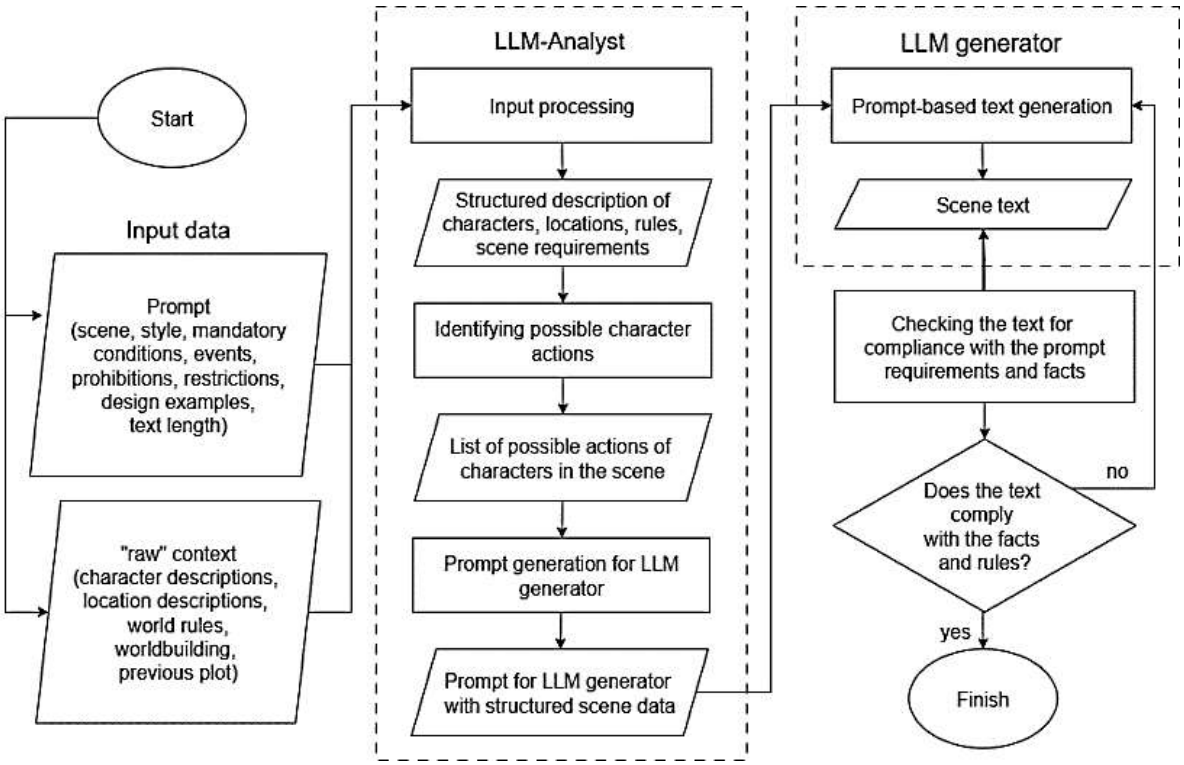


**Fig. 1.** *Combined LLM Usage Method Diagram*

The proposed method is based on the fact that LLM effectiveness is differentiated by task type: one model may excel in deep logical analysis, while another may excel in artistic sophistication.

This methodology eliminates the drawbacks of a monolithic approach by dividing responsibilities between the identified Analyst Model and Generator Model. The Analyst creates a verified, structured plan, which minimizes the risk of hallucinations and plot inconsistencies. The Generator then uses this reliable plan as high-quality context, focusing exclusively on stylistic quality and the emotional depth of the text. Such sequential two-phase processing maximizes the strengths of each model, ensuring the necessary logical coherence and quality control of the artistic output [38, 39].

## 5. Modeling the Application Structure for Creative Text Analysis and Generation

### 5.1 Database Architecture Modeling

To effectively conduct experiments with LLMs and evaluate their responses, a database structure is being developed. This database is designed to store input test scenarios, data on the LLMs used, and the quantitative and expert results obtained during the analysis, generation, and hybrid method stages. A relational database model was chosen because it ensures data integrity, provides a clear structure for relationships between objects (test scenarios and results), and simplifies the subsequent aggregation and analysis of experimental metrics. The database consists of five tables, with relationships established using foreign keys. The LLM_Models table stores metadata about each Large Language Model (LLM) involved in the experiment. Each model is uniquely identified by its model_id (Primary Key). It also records the model's common name (model) and its technical identifier (model_name), which is used for API requests.

The Test_Scenarios table holds the complete set of input data necessary for testing. Each record represents a unique experimental scenario identified by scenario_id (Primary Key). It contains the full unstructured context provided to the model (context_d), the user's specific instruction (prompt_q), the expert-created gold standard or reference output (gold_ref), and the type of test being conducted (test_type) which specifies "analysis," "generation," or "hybrid." The Analysis_Results table archives the results from the context analysis phase, which is crucial for selecting the optimal Analyst Model. Each result is uniquely identified by analysis_result_id (Primary Key) and linked to the specific scenario and model via scenario_id and model_id (Foreign Keys). The table stores the raw analysis text generated by the LLM (output_o_an) alongside the quantitative metric scores: coverage_score, hallucination_rate, the entity_f1_score, and the expert-rated actions_logic_score. A single aggregated value, q_an_final, stores the final quality score for the analysis.

The Generation_Results table records the outcomes from the creative text generation phase, used to evaluate models for the Generator Model role. Records are linked to the scenario and model via Foreign Keys. It stores the raw generated artistic text (output_text) and the scores from the five human evaluation criteria: $P_1$ (Prose Quality), $P_2$ (Emotional Depth), $P_3$ (Plausibility), $P_4$ (Tone Adherence), and $P_5$ (Technical Literacy). The table concludes with the final combined generation quality score, q_gen_final.

The Hybrid_Results table captures the results produced by the developed hybrid approach, which combines the specialized Analyst and Generator Models. Records are linked back to the scenario using scenario_id. It stores the intermediate analysis text produced by the Analyst Model (output_an_hybr) and the final text generated by the hybrid method (output_text_hybr). Like the Generation table, it includes the five expert evaluation scores ($P_1$ through $P_5$) and the final combined generation quality score (q_gen_final) for the hybrid output.

The structure of the database tables is illustrated in Fig. 2.

### 5.2 System Architecture

The developed system is based on a modular architecture designed for flexibility, scalability, and progressive development. The core objective of this architecture is to create an environment where data undergoes multi-level transformation: from the initial parsing of the input text to the formation of structured knowledge, and ultimately, the generation of new creative material.

The architecture strictly separates functions into the following structural blocks:

- Database and Scenario Management Block. Responsible for structuring the database, managing queries, and forming the controlled test scenarios for experiments. This ensures reliable organization and correctness verification of the input data.
- Model Interaction Block (API Handler). Manages all API calls to the LLMs and controls the application of System Prompts. It ensures standardized information transfer and correct interpretation of the requests by the models.
- Analysis Output and Quality Assessment Block. Performs the quantitative evaluation of the text analysis quality using specialized metrics (e.g., F1-Score, Hallucination Rate). This block provides essential data for optimizing subsequent text generation.
- Generation Output and Quality Assessment Block. Responsible for creating textual fragments based on the input data and the analytical output. Concurrently, it assesses the quality of the generated text, controlling for accuracy, stylistic coherence, and logical consistency.

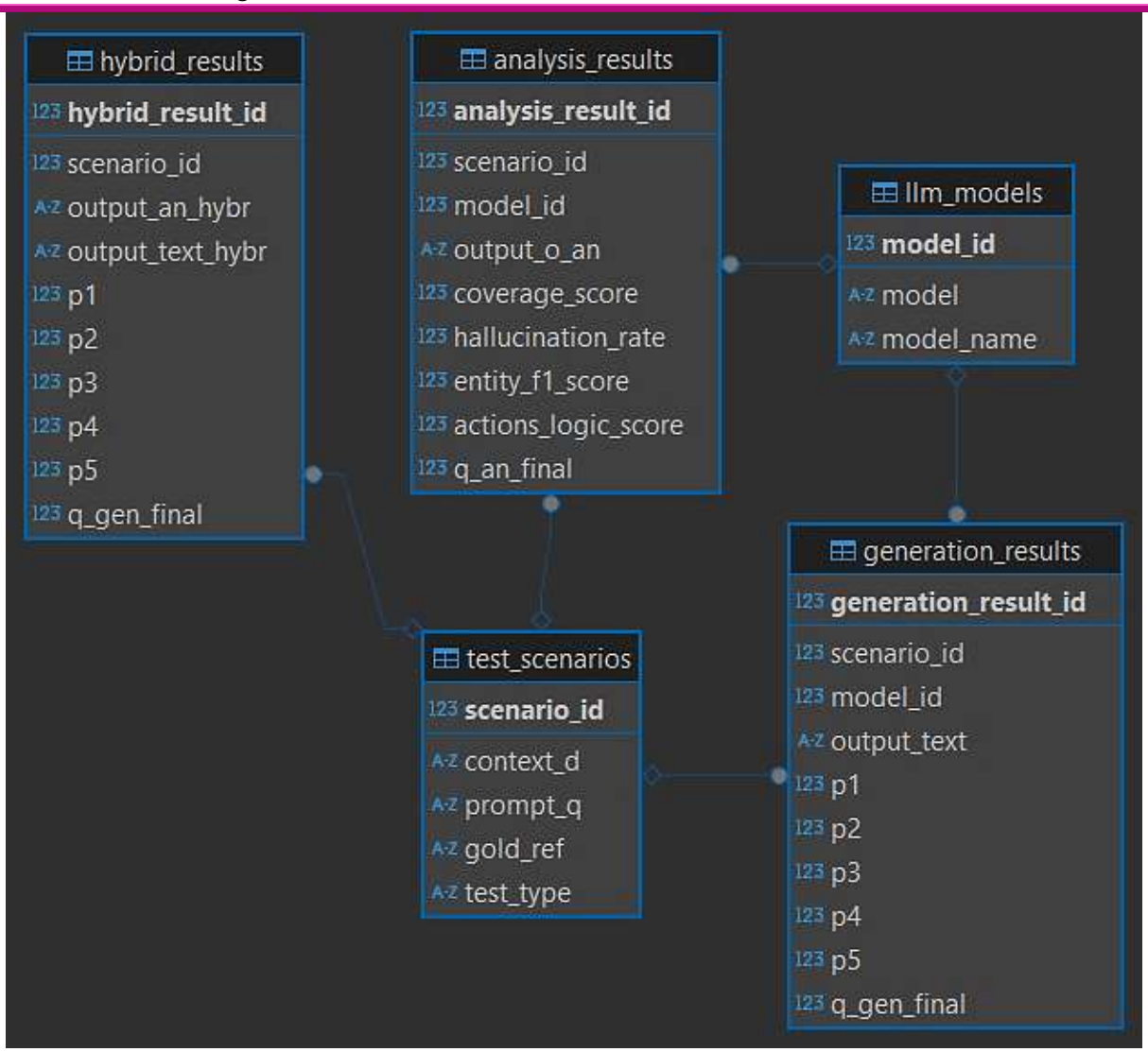


**Fig. 2.** *Database Structure*

- Hybrid Collaboration Testing Block. Integrates the preceding steps into a single experimental cycle: it submits the request for analysis, passes the analytical results to the generative module, and performs the final quality evaluation. This block specifically validates the effectiveness of the combined, inter-model approach to text creation.

- Results Visualization Block: Responsible for outputting the experimental results in the form of graphs and tables.

The interaction of the application blocks is illustrated in Fig. 3.

The technical solutions focus on combining computational efficiency with scalability. The system supports asynchronous task execution, which allows for simultaneous processing of multiple data and request streams, significantly reducing latency and increasing experimental throughput. Furthermore, the modular design ensures straightforward integration of various LLMs and tools, enabling flexible configuration changes and functional expansion without substantial structural modifications.

The research experiments are conducted in the Google Colab environment, chosen for its Python integration, GPU access, and minimal local configuration requirements. The core language of implementation is Python 3.10.

For working with models, we rely on the PyTorch framework. This is coupled with the Hugging Face Transformers library, which provides a unified interface for tasks like tokenization and semantic similarity calculation. The Sentence-Transformers module is specifically integrated to build multilingual sentence embeddings using the distiluse-base-multilingual-cased-v2 and paraphrase-multilingual-MiniLM-L12-v2 models, enabling robust analysis across Ukrainian and other target languages.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 9 Issue 10 October - 2025, Pages: 249-263

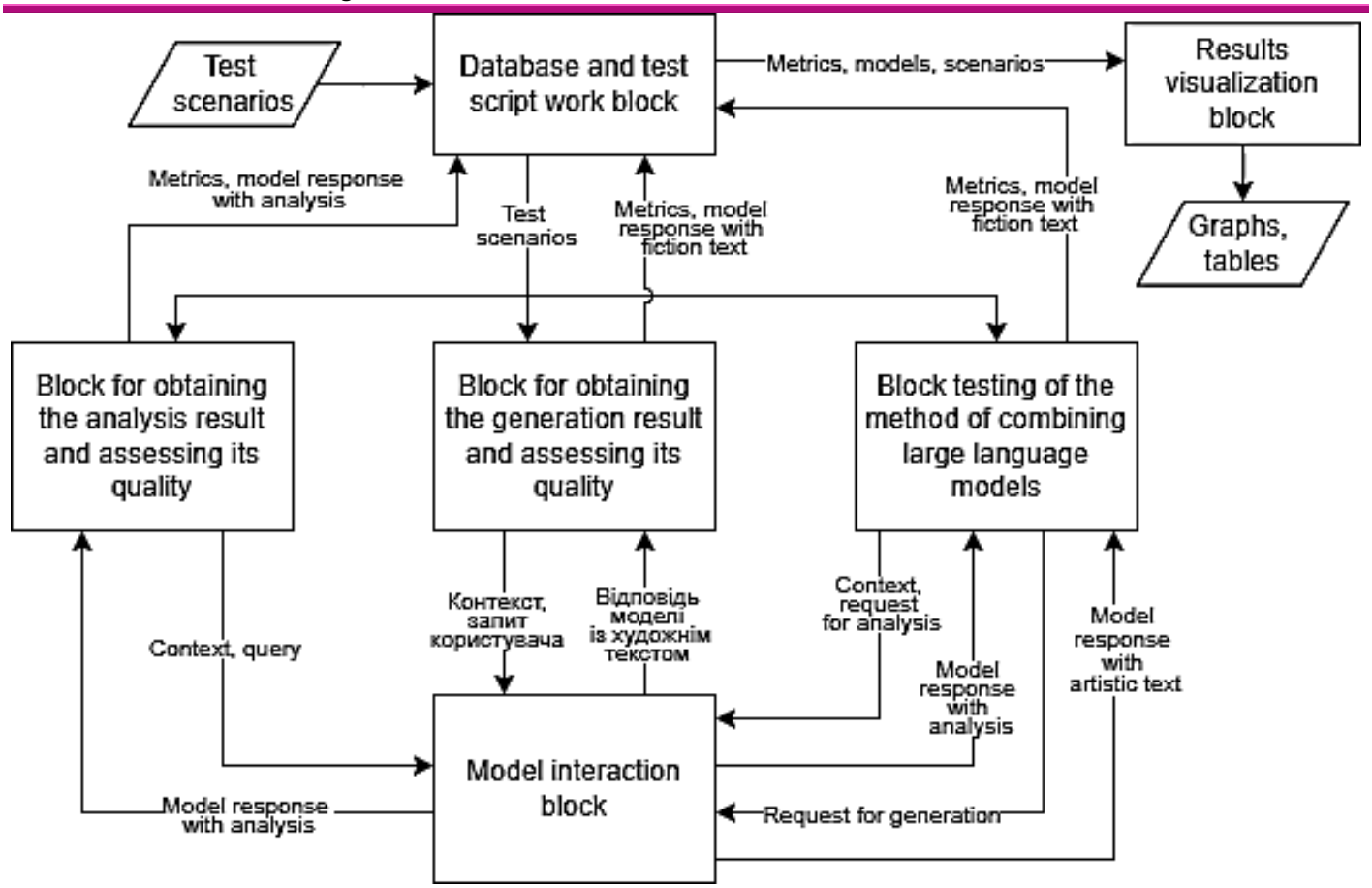


**Fig. 3.** *Interaction of the application blocks*

External LLM access (GPT-4o, Claude 4 Opus, Gemini 2.5 Pro) is facilitated via the AsyncOpenAI client through the OpenRouter platform, which standardizes asynchronous API calls for comparative testing. Results are stored in an embedded SQLite relational database for simple, server-less data management, with the Pandas library used for subsequent statistical analysis. This setup ensures a flexible and efficient environment for evaluating LLM outputs.

## 6. Design and Selection of Test Scenarios and Prompts

### 6.1 Test Scenarios

The scenarios form the set of input data for testing the LLMs. Each scenario was structured in the format of a context, a user query, and a reference output (gold standard).

The scenarios covered various genres (techno-fantasy, military strategy, neo-noir) and presented the models with complex tasks requiring not only the direct use of facts but also the integration of characters' psychological traits into dialogues and internal monologues. Brief descriptions of the scenarios are given below.

Scenario 1. "Aether Smuggler" (Techno-fantasy):

- Context: The world of Aria and the floating city of Starlight Port, where magic (Aether) leads to dangerous "crystallization" of people. Main characters: Liam Skywalker (a cynical Aether smuggler) and Ayla (a cold, logical architect of Aether circuits) who seeks revenge against the Paladins of Light system. Conflict: They possess a rare Stabilization Crystal, which Liam wants to sell (money) and Ayla wants to destroy (safety and revenge for her sister). The Paladins are pursuing Liam.
- Query: A scene of Liam and Ayla meeting in a hideout, where Liam, having just escaped a patrol, tensely discusses the Crystal's fate with the logical Ayla. Emphasis on the contrast of characters and the description of Aether circuits.
- Testing Goal: To check the models' ability to integrate complex magic world rules, reflect psychological contrast and tension in dialogue (cynicism vs. logic), and maintain the user-specified style and genre.

Scenario 2. "The Warlord and Blood Magic" (Military Fantasy):

- Context: A military camp in the Dragon's Shadow mountain pass. Main character: Lord Cael, a young warlord who hates magic but is guided by the strategy of "slow encirclement." He has a scar and wears the Corvus Amulet.

Conflict: Cael is preparing to besiege the "Dead Rock" fortress, held by the Sorcerer Lord Zane.

- Query: A scene where Cael receives news that Zane has fortified the fortress with blood magic. Cael must react cynically to the mention of magic, display his speed of decision, but reference his signature strategy.
- Testing Goal: To assess the models' ability to reflect an internal conflict (hatred of magic versus the necessity of victory), utilize symbolic details (scar, amulet), and accurately reproduce strategic military terms while maintaining a cynical tone.

Scenario 3. "The Elusive and the Pharaoh's Clock" (Neo-Noir Genre):

- Query: A scene where Cael receives news that Zane has fortified the fortress with blood magic. Cael must react cynically to the mention of magic, display his speed of decision, but reference his signature strategy.
- Context: A city in 1978, the "Silent Haven" tavern. Main characters: Victor "The Elusive" Koval, a veteran engineer, a smuggler with his own principles (no violence/drugs) and deep pessimism. He has a younger sister, Lida. Threat: Commissioner Ivan Draganov, who seeks to destroy Victor's reputation. Conflict: Victor must urgently smuggle out the stolen national relic, the "Pharaoh's Clock," but receives news of Lida's accident.
- Query: A scene in the backroom where Victor is disassembling a radio receiver and receives bad news. Emphasis on Victor's cynicism, his mechanical ingenuity, and the sharp internal conflict between business (the Clock) and family duty (Lida).
- Testing Goal: To check the models' ability to accurately integrate mechanical details (radio receiver description), reflect an internal monologue, and create a high-tension scene where family duty outweighs professional risk.

### 6.2 Prompts

Three system prompts were developed to obtain results from the LLMs. They were designed to maximize the effective use of the models' properties: the ability for accurate data extraction (analysis) and the ability for creative synthesis (generation).

System Prompt for Analysis was designed to implement the analysis stage – converting unstructured or semi-structured context text into a strictly formalized dataset. It mandated the use of rigid markers ("##" for entity headings, "* Key: Value" for facts) to ensure the output document always had the same structure, regardless of the input.

Models were instructed to break down complex descriptions (e.g., biography, character) into separate, atomic facts to ensure data extraction completeness.

A crucial part of the analysis task, beyond mere fact extraction, was requiring the model to use those facts (character, motivation, current situation) to propose a list of logical and psychologically justified actions the character could perform in the given scene. This was separated into a special block ("@@ Possible Actions: Character Name") and served as a direct bridge to the generation stage, providing the author or another model with ready-made scenario material.

Generalized structure of a gold standard and an analysis output is illustrated in Fig. 4.

```
## [Entity Type]: [Entity Name]
* [Attribute 1]: [Value]
* [Attribute 2]: [Value]
* [Attribute 3]: [Value]
...
(Additional attributes depending on entity type)

## [Entity Type]: [Entity Name]
* [Attribute 1]: [Value]
* [Attribute 2]: [Value]
...

@@ Possible Actions: [Entity Name]
- [Action 1 description]
- [Action 2 description]
- [Action 3 description]
...
```

**Fig. 4.** *Generalized structure of a gold standard and an analysis output*

System Prompt for Generation was created for the scenario where the model acts as an independent creative author, receiving the same unstructured context as in the analysis experiment. Its goal was to maximize the activation of the model's creative and stylistic abilities.

Key requirements include:

- Using an "artistic, evocative style, not a technical description," which is critical for transforming dry facts into living prose.
- The requirement for "vivid, plausible dialogues" that reveal character, forcing the model to integrate the context-defined traits into the characters' language and actions.
- The demand to always conclude the scene with a "logical endpoint," countering the common issue of abrupt generation termination and compelling the model to create a sense of completeness and accomplished micro-task.

## 7. Experimental Results Acquisition and Evaluation

Once the analysis responses for the scenarios have been obtained from the models, the evaluation process begins. Automatic metrics are calculated by the custom application, while the score for the Actions Logic metric is provided by human experts.

These experts familiarize themselves with the scenarios, the raw model outputs, specifically focusing on the section detailing the characters' possible actions within the scene.

Fig. 5 shows a screenshot of the application's work in calculating the scores for the analysis results.

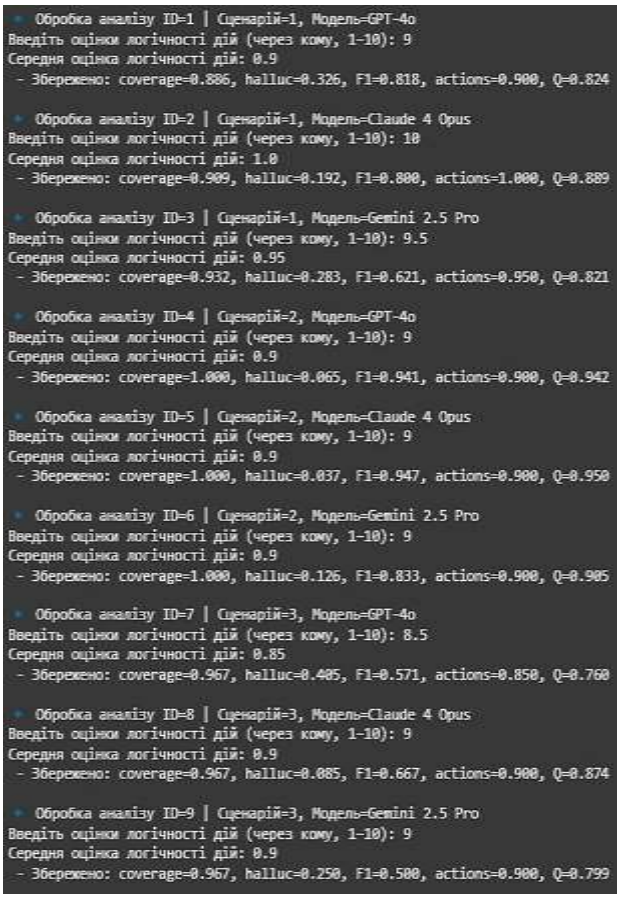


**Fig. 5.** *Calculations of analysis scores for all scenarios*

The generation experiment is conducted similarly: the models produce creative texts. These outputs are then reviewed by human experts who provide their evaluation scores based on the established criteria.

### 7.1 Comparison of Models in the Analysis Task

The results of testing the three models in the analysis task across three scenarios of varying genre and complexity are presented in Table 1.

The Claude 4 Opus model demonstrated the highest overall effectiveness, achieving an average integral score $Q_A$ of ≈ 0.91, which surpasses the other systems. Its strengths lie in an exceptionally low Hallucination Rate (0.1) and high accuracy in Named Entity Recognition ($F1$ = 0.81). It maintains logical integrity even in complex scenarios, formulating conclusions clearly and concisely, without a tendency to invent additional details. Thus, Claude can be characterized as a model with "critical thinking."

**Table 1**: Results of Testing Models in the Analysis Task on All Scenarios

| Model | Scenario | Coverage | Hallucination Rate | $F1$ | Actions Logic | $Q_A$ |
|---|---|---|---|---|---|---|
| GPT-4o | 1 | 0.89 | 0.33 | 0.82 | 0.90 | 0.82 |
| Claude 4 Opus | | 0.91 | 0.19 | 0.80 | 1.00 | 0.89 |
| Gemini 2.5 Pro | | 0.93 | 0.28 | 0.62 | 0.95 | 0.82 |
| GPT-4o | 2 | 1.00 | 0.07 | 0.94 | 0.90 | 0.94 |
| Claude 4 Opus | | 1.00 | 0.04 | 0.95 | 0.90 | 0.95 |
| Gemini 2.5 Pro | | 1.00 | 0.13 | 0.83 | 0.90 | 0.91 |
| GPT-4o | 3 | 0.97 | 0.41 | 0.57 | 0.85 | 0.76 |
| Claude 4 Opus | | 0.97 | 0.09 | 0.67 | 0.90 | 0.87 |
| Gemini 2.5 Pro | | 0.97 | 0.25 | 0.50 | 0.90 | 0.80 |

GPT-4o showed comparatively stable results, maintaining a balance between logic and completeness, but had a higher Hallucination Rate (0.27). This may indicate its tendency toward a more creative interpretation of the text. Despite this, the average level of accuracy and consistency remained high $Q_A$ (≈ 0.84), and its ability to preserve semantic structure was nearly flawless.

Gemini 2.5 Pro demonstrated the highest content coverage (Coverage ≈ 0.97), but had a lower entity accuracy score ($F1$ ≈ 0.65) and a slightly higher propensity for generating inaccurate statements. Despite this, the overall quality remained competitive ($Q_A$ ≈ 0.84). Gemini handles summarization and logical inference well but lags behind competitors in the depth of detail analysis.

All models demonstrated a high level of cognitive consistency in performing the analytical task. However, Claude 4 Opus emerged as the most reliable and consistent system in terms of logical construction of judgments and minimization of factual distortions.

GPT-4o stands out with more flexible and variational reasoning, making it valuable for interpretive and creative analysis types, while Gemini 2.5 Pro shows potential in systematic context coverage but requires improvement in precision of detail.

In the analytical processing of creative texts, Claude 4 Opus demonstrated the highest quality, allowing it to be considered the optimal foundation for the subsequent method of combining models.

### 7.2 Comparison of Models in the Generation Task

The results of testing the three models in the generation task across three scenarios of varying genre and complexity are presented in Table 2.

**Table 2**: Results of Testing Models in the Generation Task on All Scenarios

| Model | Scenario | $P_1$ | $P_2$ | $P_3$ | $P_4$ | $P_5$ | $Q_G$ |
|---|---|---|---|---|---|---|---|
| GPT-4o | 1 | 0.75 | 0.70 | 0.78 | 0.60 | 0.80 | 0.74 |
| Claude 4 Opus | | 0.80 | 0.73 | 0.80 | 0.75 | 0.88 | 0.79 |
| Gemini 2.5 Pro | | 0.93 | 0.88 | 0.90 | 0.90 | 0.95 | 0.91 |
| GPT-4o | 2 | 0.80 | 0.70 | 0,73 | 0,70 | 0.90 | 0.77 |
| Claude 4 Opus | | 0.80 | 0.75 | 0.80 | 0.85 | 0.88 | 0.81 |
| Gemini 2.5 Pro | | 0.88 | 0.90 | 0.88 | 0.85 | 0.95 | 0.89 |
| GPT-4o | 3 | 0.73 | 0.80 | 0.80 | 0.73 | 0.83 | 0.78 |
| Claude 4 Opus | | 0.78 | 0.80 | 0.75 | 0.80 | 0.93 | 0,81 |
| Gemini 2.5 Pro | | 0.93 | 0.90 | 0.90 | 0.90 | 0.98 | 0.92 |

The results in Table 2 allow us to examine generation quality through five aspects: Prose Quality ($P_1$), Emotional Depth ($P_2$), Plausibility ($P_3$), Tone Adherence ($P_4$), and Technical Literacy ($P_5$). The final Generation Quality Indicator ($Q_G$) summarizes these measurements for comparison across specific scenarios.

A clear advantage of Gemini 2.5 Pro is immediately evident across all three scenarios: its $Q_G$ ranges from 0.89 to 0.92, indicating that this model combines language with grammatical accuracy and logical plot coherence.

Prose Quality: Gemini shows the highest scores compared to Claude and GPT-4o. This signifies that Gemini generates richer linguistic constructions, and greater syntactic diversity.

Emotional Depth: Gemini also leads. This suggests the model is superior at conveying the emotional nuances of the scene and eliciting a more pronounced emotional response from the reviewers. Claude shows moderate results, while GPT-4o generally has lower $P_2$ values, indicating a sometimes "drier," less emotionally saturated narrative style.

Psychological Plausibility: the ability of actions and dialogues to be logically justified also correlates with integral quality. Gemini's high $P_3$ values suggest its texts rarely contain illogical motives or unmotivated character actions, proving superior when character behavior consistency.

Tone Adherence: In terms of stylistic imitation, the adherence to the specified genre and tonal requirements, GPT-4o shows a noticeable weakness in Scenario 1, whereas Claude and Gemini yield significantly higher results. This indicates that GPT-4o is less reliable in following tonal instructions, sometimes deviating from genre constraints. Conversely, Gemini demonstrates stability in reproducing the requested genre and tone.

Technical Literacy: Technical correctness is generally less problematic for all models, but Gemini still exhibits the highest values (0.95–0.98), followed by Claude (0.88–0.93), and GPT-4o (0.8–0.9). This means that spelling, punctuation, and factual detail inconsistencies are rarest in Gemini's texts.

Gemini 2.5 Pro consistently demonstrated the best performance across all artistic quality metrics. Therefore, it is identified as the optimal Generator Model for the proposed combined LLM method.

### 7.3 Evaluation and Superiority of the Combined LLM Method

The evaluation of individual models revealed that no single LLM consistently led in both analysis and generation tasks. Claude 4 Opus excelled in analytical tasks, specifically in structuring plot elements and maintaining logical coherence. Conversely, Gemini 2.5 Pro was significantly more effective in generation scenarios, noted for its stylistic integrity and artistic expression.

Based on these findings, a combined method was tested, integrating the strengths of both:

- Analyst Model (Claude 4 Opus): performs a detailed structural analysis of the input context, creating a coherent, formalized knowledge base of facts, characters, and plot.
- Generator Model (Gemini 2.5 Pro): uses this detailed analysis as a highly structured prompt (a systematized knowledge base) instead of the raw, unstructured user text.

Fig. 6 shows the Combined Method results on all three scenarios.

Table 3 directly compares the generation results of the three individual LLMs against the Combined Method.

This method shows how preliminary analytical processing improves the quality of artistic generation. The method achieved an average integral $Q_g$ score above 0.94 across all three scenarios, significantly surpassing the results of individual models. The highest score was obtained in Scenario 3, which required complex structure and emotional integrity. This confirms that the combined approach successfully optimizes the analytical foundation and enhances the creative aspect, balancing cognitive accuracy with artistic depth by ensuring better coherence and reducing semantic breaks and stylistic inconsistencies.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 9 Issue 10 October - 2025, Pages: 249-263

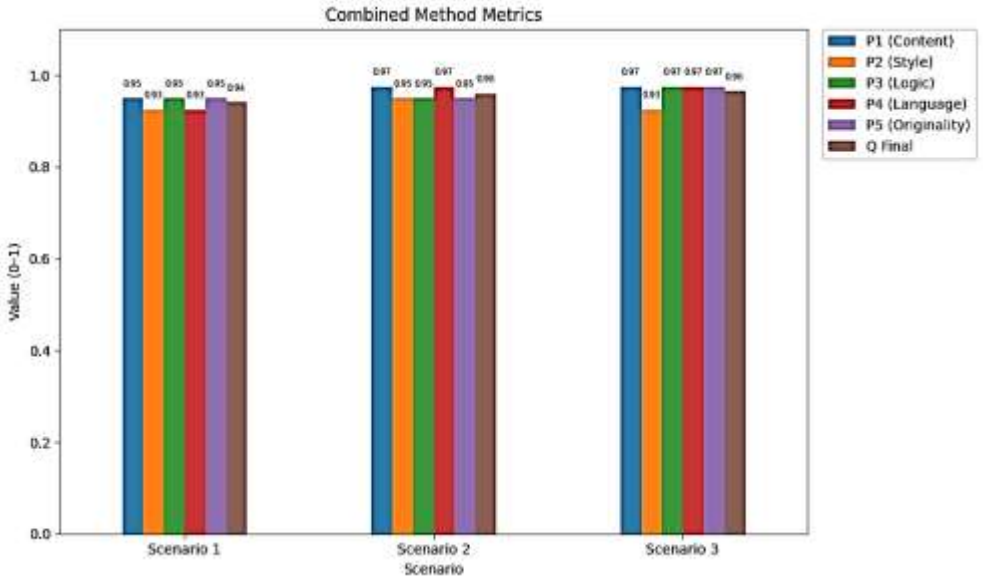


**Fig. 6.** *Combined Method generation*

**Table 3**: Comparison of Generation Results of All Models and Combined Method

| Scenario | Metrics | GPT-4o | Claude 4 Opus | Gemini 2.5 Pro | Combined Method |
|---|---|---|---|---|---|
| 1 | $P_1$ | 0.75 | 0.80 | 0.93 | 0.95 |
| | $P_2$ | 070 | 0.73 | 0.83 | 0.93 |
| | $P_3$ | 0.78 | 0.80 | 0.90 | 0.95 |
| | $P_4$ | 0.60 | 0.75 | 0.90 | 0.93 |
| | $P_5$ | 0.80 | 0.88 | 0.95 | 0.95 |
| | $Q_G$ | 0.74 | 0.79 | 0.91 | 0.94 |
| 2 | $P_1$ | 0.80 | 0.80 | 0.88 | 0.98 |
| | $P_2$ | 0.70 | 0.75 | 0.90 | 0.95 |
| | $P_3$ | 0.73 | 0.80 | 0.88 | 0.95 |
| | $P_4$ | 0.70 | 0.85 | 0.85 | 0.98 |
| | $P_5$ | 0.90 | 0.88 | 0.95 | 0.95 |
| | $Q_G$ | 0.77 | 0.81 | 0.89 | 0.96 |
| 3 | $P_1$ | 0.73 | 0.78 | 0.93 | 0.98 |
| | $P_2$ | 0.80 | 0.80 | 0.90 | 0.93 |
| | $P_3$ | 0.80 | 0.75 | 0.90 | 0.98 |
| | $P_4$ | 0.73 | 0.80 | 0.90 | 0.98 |
| | $P_5$ | 0.83 | 0.93 | 0.98 | 0.98 |
| | $Q_G$ | 0.78 | 0.81 | 0.92 | 0.97 |

The final $Q_g$ scores show that the combined method is superior to all individual models in every scenario. The Combined Method surpassed the individual models across all criteria, only matching the score of the Gemini 2.5 Pro model in the Technical Literacy $P_5$ criterion. This outcome is expected, as Gemini 2.5 Pro – the Generator Model in the approach – is responsible for the output's fundamental spelling and punctuation (which are unaffected by the Analyst Model's structured input).

## 8. Conclusion

This study successfully addressed the challenge of inconsistent output quality in LLMs for complex creative writing tasks, specifically artistic text analysis and generation. The core contribution is the validation of a combined LLM methodology designed to leverage model specialization.

The initial comparative analysis established a clear functional distinction: Claude 4 Opus proved significantly more reliable and accurate in analytical tasks, making it the optimal Analyst Model for structuring complex plot and character data. Conversely, Gemini 2.5 Pro excelled in creative generation, consistently demonstrating superior prose quality, emotional depth, and stylistic adherence, thereby serving as the ideal Generator Model.

The implemented combined approach integrated the verified, structured analytical output from the Analyst Model as a highly systematic prompt for the Generator Model. This strategic division of labor was critical.

The final evaluation confirmed the hypothesis: the combined method generated outputs of a quality that substantially surpassed the performance of any single monolithic model.

The hybrid approach consistently produced texts that were judged superior across all human-evaluated artistic criteria, including psychological plausibility and overall narrative coherence. This demonstrates that by preceding the creative phase with rigorous, formalized analytical processing, the system effectively mitigates common LLM weaknesses, such as factual inconsistency and stylistic drift. In essence, the developed Combined Method successfully resolved the inherent trade-off between cognitive accuracy and creative depth, representing a significant step toward developing controllable, high-fidelity content generation systems for specialized creative domains.

## 9. ACKNOWLEDGMENT

The papers acknowledge the support of the Department of Informatics, Kharkiv National University of Radio Electronics, Ukraine, in numerous help and support to complete this paper.

The research results were obtained as part of the EU Horizon Europe international research project INITIATE (grant No. 101136775).

## 10. REFERENCES

**[1]** Franceschelli, G., & Musolesi, M. (2024). On the creativity of large language models. [Online]. Available: https://arxiv.org/abs/2304.00008.

**[2]** Vaswani, A. et al. (2017). Attention is all you need. [Online]. Available: https://arxiv.org/abs/1706.03762.

**[3]** Achiam, J. et al. (2023). GPT-4 technical report. [Online]. Available: https://arxiv.org/abs/2303.08774.

**[4]** Georgiev, P. et al. (2024). Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. [Online]. Available: https://arxiv.org/abs/2403.05530.

**[5]** Gómez-Rodríguez, C., & Williams, P. (2023). A confederacy of models: A comprehensive evaluation of LLMs on creative writing. [Online]. Available: https://arxiv.org/abs/2310.08433.

**[6]** Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., & Lu, Z. (2024). "It felt like having a second mind": Investigating human-AI co-creativity in prewriting with large language models. [Online]. Available: https://arxiv.org/abs/2307.10811.

**[7]** Bellemare-Pepin, A. et al. (2024). Divergent creativity in humans and large language models. [Online]. Available: https://arxiv.org/abs/2405.13012.

**[8]** Ismayilzada, M., Stevenson, C., & van der Plas, L. (2024). Evaluating creative short story generation in humans and large language models. [Online]. Available: https://arxiv.org/abs/2411.02316.

**[9]** Shanahan, M., & Clarke, C. (2023). Evaluating large language model creativity from a literary perspective. [Online]. Available: https://arxiv.org/abs/2312.03746.

**[10]** Marco, G., Rello, L., & Gonzalo, J. (2024). Small language models can outperform humans in short creative writing: A study comparing SLMs with humans and LLMs. [Online]. Available: https://arxiv.org/abs/2409.11547.

**[11]** Subbiah, M., Zhang, S., Chilton, L. B., & McKeown, K. (2024). Reading subtext: Evaluating large language models on short story summarization with writers. [Online]. Available: https://arxiv.org/abs/2403.01061.

**[12]** Zhao, Y., Zhang, R., Li, W., & Li, L. (2025). Assessing and understanding creativity in large language models, Mach. Intell. Res., vol. 22, no 3, pp. 417-436.

**[13]** Rane, N., Choudhary, S., & Rane, J. (2024). Gemini versus ChatGPT: Applications, performance, architecture, capabilities, and implementation, J. Appl. Artif. Intell., vol. 5, no 1, pp. 69-93.

**[14]** Akter, S. N. et al. (2023). An in-depth look at gemini's language abilities. [Online]. Available: https://arxiv.org/abs/2312.11444.

**[15]** Tu, L., Meng, R., Joty, S., Zhou, Y., & Yavuz, S. (2024). Investigating factuality in long-form text generation: The roles of self-known and self-unknown. [Online]. Available: https://arxiv.org/abs/2411.15993.

**[16]** Garrido-Merchán, E. C., Arroyo-Barrigüete, J. L., & Gozalo-Brizuela, R. (2023). Simulating H. P. Lovecraft horror literature with the ChatGPT large language model. [Online]. Available: https://arxiv.org/abs/2305.03429.

**[17]** Tvoroshenko, I. S. (2021). Decision-making technologies in information systems: Study guide. Kharkiv, Ukraine: KNURE. Творошенко, І. С. (2021). Технології прийняття рішень в інформаційних системах: Навчальний посібник. Харків, Україна: ХНУРЕ.

**[18]** Tvoroshenko, I., Gorokhovatskyi, V., Kobylin, O., & Tvoroshenko, A. (2023). Application of deep learning methods for recognizing and classifying culinary dishes in images, Int. J. Acad. Appl. Res., vol. 7, no. 9, pp. 57-70.

**[19]** Kobylin, O. A., & Tvoroshenko, I. S. (2021). Digital image processing methods: Study guide. Kharkiv, Ukraine: KNURE. Кобилін, О. А., & Творошенко, І.С. (2021). Методи цифрової обробки зображень: Навчальний посібник. Харків, Україна: ХНУРЕ.

**[20]** Daradkeh, Y. I., Gorokhovatskyi, V., Tvoroshenko, I., & Zeghid, M. (2022). Tools for fast metric data search in structural methods for image classification, IEEE Access, vol. 10, pp. 124738-124746.

**[21]** Pomazan, V., Tvoroshenko, I., & Gorokhovatskyi, V. (2023). Development of an application for recognizing emotions using convolutional neural networks, Int. J. Acad. Inf. Syst. Res., vol. 7, no. 7, pp. 25-36.

**[22]** Daradkeh Y.I., Gorokhovatskyi V., Tvoroshenko I., & Zeghid M. (2024). Improving the effectiveness of image classification structural methods by compressing the description according to the information content

criterion, Comput. Mater. Contin., vol. 80, no. 2, pp. 3085-3106.

**[23]** Gorokhovatskyi, V., & Tvoroshenko, I. (2023). Identification of visual objects by the search request. In Int. Sci. Symp. «INTELLIGENT SOLUTIONS-S». Comput. Intellig. (results, problems and perspectives). Decision making theory: proceedings of the international symposium, September 28, 2023, Kyiv-Uzhorod, Ukraine, pp. 25-27.

**[24]** Gorokhovatskyi, V., Tvoroshenko, I., Yakovleva, O., & Hudáková, M. (2025). Image description compression in classification structural methods, IEEE Access, vol. 13, pp. 43631-43641.

**[25]** Yakovleva O., Matúšová S., Tvoroshenko I., & Isaiev Y. (2024). Visitor counting based on video stream analysis from surveillance cameras to solve various business problems, vol. XX, no 1, pp. 67-87. [Online]. Available: https://www.vsemba.sk/portals/0/Subory/vedecky%20casopis%2001%20-%202024%20-%20web%20-%20281024.pdf

**[26]** Gorokhovatskyi, V., Tvoroshenko, I., Yakovleva, O., Hudáková, M., & Gorokhovatskyi, O. (2024). Application a committee of Kohonen neural networks to training of image classifier based on description of descriptors set, IEEE Access, vol. 12, pp. 73376-73385.

**[27]** Gorokhovatskyi, V., & Tvoroshenko, I. (2022). Analysis of multidimensional data described in the form of a set of components: Monograph. Kharkiv, Ukraine: KNURE. Гороховатський, В. О., & Творошенко, І. С. (2022). Аналіз багатовимірних даних за описом у формі множини компонент: Монографія. Харків, Україна: ХНУРЕ.

**[28]** Daradkeh Y. I., Gorokhovatskyi V., Tvoroshenko I., & Zeghid M. (2022). Cluster representation of the structural description of images for effective classification, Comput. Mater. Contin., vol. 73, no. 3, pp. 6069-6084.

**[29]** Gorokhovatskyi, V., Tvoroshenko, I. & Yakovleva, O. (2024). Transforming image descriptions as a set of descriptors to construct classification features, Indones. J. Elec. Eng. Comput. Sci., vol. 33, no. 1, pp. 113-125.

**[30]** Gorokhovatskyi, V., Chmutov, Y., Tvoroshenko, I., & Kobylin, O. (2025). Reducing computational costs by compressing the structural description in image classification methods, Adv. Inf. Syst., vol. 9, no. 1, pp. 5-12.

**[31]** Gorokhovatskyi, V., Peredrii, O., Tvoroshenko, I., & Markov, T. (2023). Distance matrix for a set of structural description components as a tool for image classifier creating. Adv. Inf. Syst., vol. 7, no. 1, pp. 5-13. Гороховатський, В. О., Передрій, О. О., Творошенко, І. С., & Марков, Т. Є. (2023). Матриця відстаней для множини компонентів структурного опису як інструмент для створення класифікатора зображень. Сучасні інформаційні системи, 7(1), С. 5-13.

**[32]** Gorokhovatskyi, V., & Tvoroshenko, I. (2024). An effective method for transforming an image description into a compact vector for classification. [Online]. Available: https://openarchive.nure.ua/server/api/core/bitstreams/13b84919-5169-4089-9b68-881e15812deb/content.

**[33]** Gorokhovatskyi, V., Tvoroshenko, I., Kobylin, O., & Vlasenko, N. (2023). Search for visual objects by request in the form of a cluster representation for the structural image description, Adv. Electr. Electron. Eng., vol. 21, no. 1, pp. 19-27.

**[34]** Daradkeh, Y. I., Gorokhovatskyi, V., Tvoroshenko, I., Gadetska, S., & Al-Dhaifallah, M. (2023). Statistical data analysis models for determining the relevance of structural image descriptions, IEEE Access, vol. 11, pp. 126938-126949.

**[35]** Tvoroshenko, I., Pomazan, V., Gorokhovatskyi, V., & Kobylin, O. (2023). Application of video data classification models using convolutional neural networks, Int. J. Acad. Appl. Res., vol. 7, no. 11, pp. 134-145.

**[36]** Daradkeh, Y. I., Gorokhovatskyi, V., Tvoroshenko, I., & Al-Dhaifallah, M. (2022). Classification of images based on a system of hierarchical features, Comput. Mater. Contin., vol. 72, no. 1, pp. 1785-1797.

**[37]** Pomazan, V., Tvoroshenko, I., & Gorokhovatskyi, V. (2023). Handwritten character recognition models based on convolutional neural networks, Int. J. Acad. Eng. Res., vol. 7, no. 9, pp. 64-72.

**[38]** Gorokhovatskyi, V., Tvoroshenko, I., & Chmutov, Y. (2022). Application of systems of orthogonal functions for formation of sign space in image classification methods. Adv. Inf. Syst., vol. 6, no. 3, pp. 5-12. Гороховатський, В. О., Творошенко, І. С., & Чмутов, Ю. В. (2022). Застосування систем ортогональних функцій для формування простору ознак у методах класифікації зображень. Сучасні інформаційні системи, 6(3), С. 5-12.

**[39]** Suprun, A. (2025). Architectural features of modern large language models, Proc. IV Int. Sci. Practic. Conf. «Technologies, theories and developments: modern scientific teaching», Valencia, Spain: International Science Group, pp. 23-26.