УДК 655.3:658.8

КОМПЛЕКСНЕ ОЦІНЮВАННЯ ЯКОСТІ МУЛЬТИМЕДІЙНОГО КОНТЕНТУ, СТВОРЕНОГО ГЕНЕРАТИВНИМИ МОДЕЛЯМИ ШТУЧНОГО ІНТЕЛЕКТУ

[1]Олександр СУПРУН, [2] Ірина ЛУЩЕНКО

[1]доцент, кафедра МСТ, ХНУРЕ [2]студентка, кафедра МСТ, ХНУРЕ

Abstract. *This paper investigates the quality of multimedia content generated by modern artificial intelligence models. The study compares leading generative systems used for creating textual, graphical, and multimedia materials based on quality, creativity, response speed, and content relevance. A generalized methodology for evaluating generated content using integral quality indicators is proposed. Comparative analysis of ChatGPT, Gemini, Claude, and Copilot demonstrates differences in content generation performance and output characteristics. The results indicate that modern generative models provide high-quality multimedia materials and can significantly support content production processes in publishing, media, education, and digital communications. The proposed evaluation approach enables objective comparison of AI systems and identification of the most effective solutions for multimedia content creation.*

Keywords: generative artificial intelligence, multimedia content, large language models, content quality assessment, multimodal systems, digital media, content generation, AI technologies.

Поширення генеративного штучного інтелекту змінює процеси створення мультимедійного контенту для цифрових видань, освітніх ресурсів, реклами, соціальних мереж і інтерактивних середовищ [1]. Сучасні моделі здатні генерувати текст, зображення, аудіо, відео та комбіновані матеріали за текстовими або мультимодальними запитами. Водночас швидкість генерації не гарантує належної якості: результат може містити фактичні помилки, візуальні артефакти, порушення композиції, неузгодженість між текстом і зображенням, дефекти мовлення або невідповідність завданню [2]. Тому актуальним є розроблення комплексного підходу до порівняння результатів різних генеративних моделей. Порівняння моделей доцільно проводити на однаковому наборі завдань, використовуючи зіставні промпти, однакові вимоги до формату та наперед визначені правила оцінювання. Для кожного результату варто фіксувати назву й версію моделі, дату генерації, параметри, кількість спроб і час отримання відповіді.

До змістових критеріїв належать відповідність запиту, фактична коректність, повнота, логічна цілісність і відсутність суперечностей [3]. Візуальна якість охоплює композицію, читабельність, деталізацію, природність зображення та сталість стилю. До технічних характеристик належать роздільна здатність, тривалість генерації, формат, розмір файлу та відтворюваність результату. Для практичного застосування пропонується оцінювати кожний критерій за шкалою від 1 до 5, де 1 означає неприйнятний результат, а 5 — високий рівень якості. Щоб зменшити суб’єктивність, кожний матеріал бажано оцінювати щонайменше двома експертами за спільною інструкцією. Якщо застосовуються автоматичні метрики, їх слід добирати відповідно до модальності та завдання: текстові показники не замінюють перевірки фактів, а метрики зображень не завжди відображають естетичну чи комунікаційну цінність [4, 5].

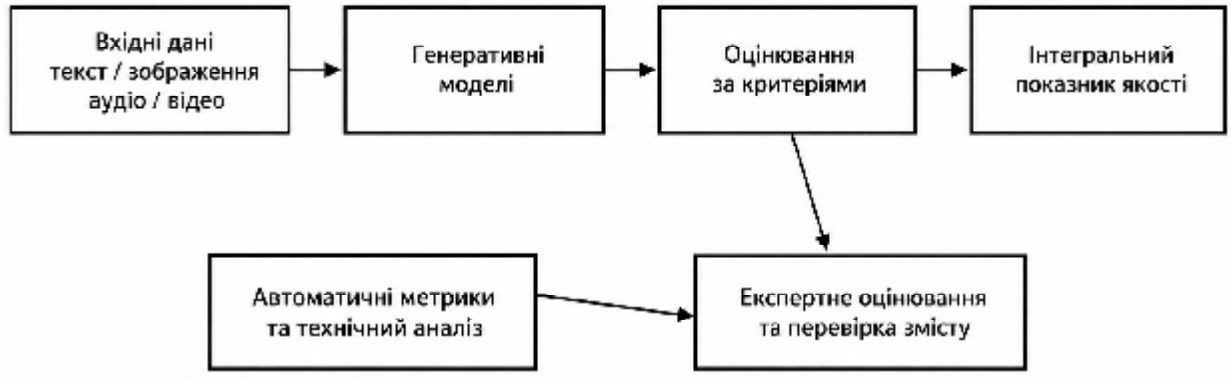


Рис. 1. Структурна модель комплексного оцінювання якості

Табл. 1

Критерії комплексного оцінювання мультимедійного контенту

Група критеріїв	Що оцінюється	Способи оцінювання
Змістова якість	Відповідність запиту, точність, повнота	Експертний чекліст; фактологічна перевірка
Візуально-звукова якість	Композиція, природність, розбірливість, синхронізація	Експертна шкала; спеціалізовані метрики
Технічна якість	Роздільна здатність, артефакти, стабільність формату	Автоматичний аналіз файлу; контроль відтворення
Комунікаційна ефективність	Зрозумілість, доступність, відповідність аудиторії	Користувацьке тестування; оцінка доступності
Ресурсна ефективність	Час, обсяг обчислень, вартість генерації	Вимірювання часу та вартості одного результату

Для зіставлення результатів можна використати зважену суму нормованих критеріїв. Ваги встановлюють залежно від призначення контенту: наприклад, для навчального відео більшої ваги набувають точність і зрозумілість, а для рекламного візуалу композиція та відповідність бренду. Усі показники перед об’єднанням потрібно привести до спільної шкали.

$$Q = \Sigma_{i=1}^{n} w_i q_i, \quad \Sigma_{i=1}^{n} w_i = 1 \qquad (1)$$

де Q — інтегральна оцінка; q_i — нормоване значення і-го критерію в межах від 0 до 1; w_i — ваговий коефіцієнт критерію; n — кількість критеріїв.

Для порівняння експертних оцінок доцільно також обчислювати середнє значення за кількома оцінювачами та контролювати розбіжності. Середня експертна оцінка та вибіркове стандартне відхилення:

$$\bar{q} = (1 / m) \Sigma_{j=1}^{m} q_j, \quad s = \sqrt{[\Sigma_{j=1}^{m}(q_j - \bar{q})^2 / (m - 1)]} \qquad (2)$$

де m — кількість експертів, q_j — оцінка окремого експерта, $\bar{q}$ — середня оцінка, s — показник розкиду оцінок. Низьке значення s свідчить про більшу узгодженість суджень, однак саме по собі не гарантує об’єктивності оцінювання.

На етапі експерименту результати різних моделей потрібно порівнювати окремо за кожним критерієм, а не лише за одним підсумковим числом. Це дає змогу виявити сильні й слабкі сторони систем: одна модель може створювати композиційно виразні зображення, інша точніше дотримуватися змісту запиту, а третя забезпечувати кращу технічну стабільність. На рис. 2 наведено умовний приклад такого порівняння.

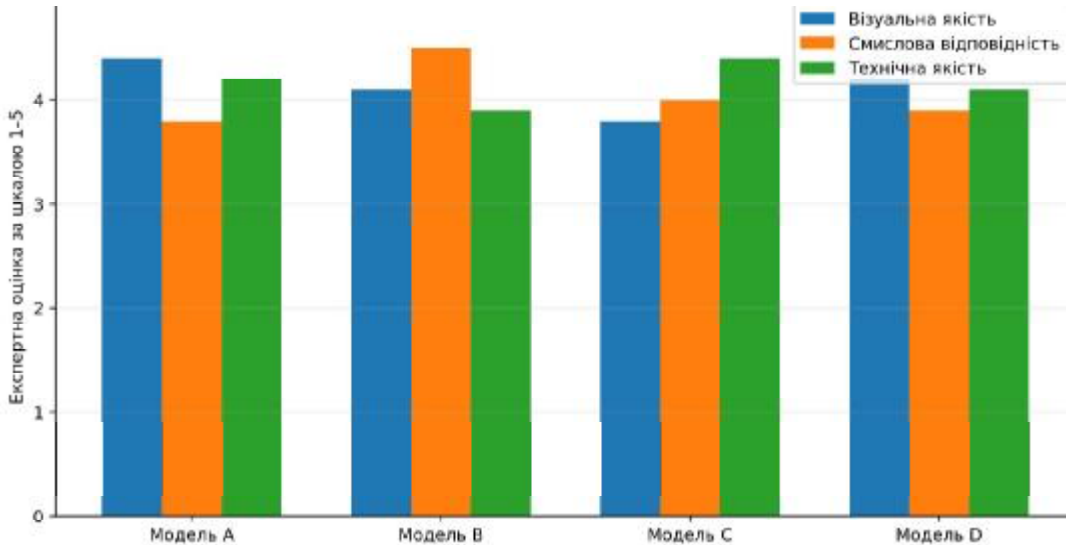


Рис. 2. Умовне порівняння якості результатів чотирьох моделей

Для забезпечення надійності експерименту необхідно використовувати однакові завдання та повторні генерації, оскільки результат може змінюватися навіть за незмінного запиту. Доцільно повідомляти не лише середній бал, а й розкид оцінок, кількість невдалих генерацій, частку результатів, що відповідають вимогам, та витрати часу. Важливо також фіксувати версії моделей і дату тестування: оновлення сервісів можуть змінювати якість результатів. Якість мультимедійного контенту, створеного генеративними моделями, є багатовимірною характеристикою, яку неможливо коректно описати однією універсальною метрикою. Запропонований підхід передбачає оцінювання змістової, візуально-звукової, технічної, комунікаційної та ресурсної складових із подальшим розрахунком інтегрального показника. Поєднання автоматизованих вимірювань, експертних оцінок і повторних генерацій дає змогу зробити порівняння прозорішим і відтворюваним.

Список використаних джерел

1. ISO/IEC 25010:2023. Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Product quality model. Geneva: ISO, 2023.

2. Zhang R. et al. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018. P. 586–595. DOI: 10.1109/CVPR.2018.00068.

3. Radford A. et al. Learning Transferable Visual Models From Natural Language Supervision. Proceedings of ICML. 2021. P. 8748–8763.

4. Oppenlaender J. The Creativity of Text-to-Image Generation. Proceedings of the 25th International Academic Mindtrek Conference. 2022. P. 192–202. DOI: 10.1145/3569219.3569350.

5. Weidinger L. et al. Taxonomy of Risks posed by Language Models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency. 2022. P. 214–229. DOI: 10.1145/3531146.3533088.