Electronics and Information Technologies, 2026, 33, 5–16
http://publications.lnu.edu.ua/collections/index.php/electronics/index

UDC: 004.89

USING LARGE LANGUAGE MODELS FOR TEXT ANALYSIS IN THE EVALUATION OF UNIVERSITY EDUCATIONAL PROGRAMS

***Mykola Stasiuk*[1], *Vitaliy Kukharskyy*[2], *Bohdan Pavlyshenko*[1],**
[1]*System Design Department, Ivan Franko National University of Lviv, 50, Drahomanova St., Lviv, 79005, Ukraine*
[2]*Applied Mathematics Department, Ivan Franko National University of Lviv, 1, Universytetska St., Lviv, 79000, Ukraine*

Stasiuk, M. I., Kukharskyy, V. M., Pavlyshenko, B. M. (2026). Using Large Language Models for Text Analysis in the Evaluation of University Educational Programs, *Electronics and Information Technologies*, 33, 5–16. https://doi.org/10.30970/eli.33.1

ABSTRACT

Background. Large language models (LLMs) are increasingly used in educational analytics, particularly for processing large volumes of accreditation-related documents. However, it remains unclear how reliably such models can assess the quality of self-evaluation reports for educational programs, and which textual characteristics influence how models form their assessments.

Materials and Methods. In the study, ten self-evaluation reports of educational programs were analyzed: five identified by the expert assessment as the strongest within the higher education institution over the last three years, and five as the weakest over the same period. GPT-5 and Gemini-2.5 models independently evaluated each document using the official ten Ukrainian National Agency for Higher Education Quality Assurance (NAQA) criteria and eight textual metrics reflecting structural, semantic, argumentative, and factual properties of the text. All evaluation grades were generated directly by the models on a unified scale from 1 to 10. To analyze the relationships between NAQA and textual criteria, Pearson's and Spearman's correlation coefficients were used.

Results and Discussion. LLMs demonstrated limited alignment with the NAQA criteria, yielding weak correlations. In contrast, textual criteria, primarily factual density, argumentativeness, semantic coherence, and lexical diversity, consistently differentiated between stronger and weaker reports. GPT-5 exhibited lower variability and reduced sensitivity to stylistic noise, while Gemini-2.5 reacted more strongly to structural and stylistic deficiencies. Correlation matrices showed that textual criteria better capture the latent quality characteristics of documents than the direct application of NAQA criteria.

Conclusion. The results show that LLMs currently do not accurately reproduce expert evaluations based on the formal NAQA criteria but effectively analyze the structural and content-related characteristics of reports using textual metrics. These metrics complement the NAQA criteria by accelerating expert workflows and enhancing document monitoring. Future research will focus on expanding the dataset, standardizing prompts, and comparing a broader range of models.

Keywords: large language models, educational programs, quality assessment.

INTRODUCTION

Large Language Models (LLMs) are gradually being integrated into educational analytics, from automated reviews to the creation of self-assessments and accreditation reports [1, 2]. At the same time, the question arises not only of the accuracy of such

assessments but also of the quality of the assessment process itself. This includes how effectively the model explains its own decisions, refers to sources, and demonstrates reasoned motivation. Studies indicate that models can differ significantly in their ability to build logical chains of reasoning and engage in reasoned thinking, which affects the stability and interpretability of their evaluations [3].

Educational program self-assessment texts are the central element of the accreditation process and reflect the aspects of the educational program's functioning required by the regulatory framework of the National Agency for Higher Education Quality Assurance (NAQA). In such documents, the higher education institution must demonstrate the completeness and systematic nature of student preparation, justify the rationale for the curriculum structure, and describe the faculty, the learning environment, resources, and quality assurance mechanisms.

The analysis of large volumes of texts is a significant burden for expert groups, especially amid the growing number of programs. Meanwhile, modern LLMs demonstrate the ability to work with long texts, summarize information, search for patterns, and measure structural and stylistic characteristics. Previous research, particularly approaches such as G-Eval, shows that LLMs can serve as tools for evaluating and analyzing text quality [4]. In this study, textual criteria are considered an additional tool for the analysis of the document's structural and semantic properties, which are not always explicitly represented in the formal criteria. This opens the possibility of using LLMs to analyze self-assessment materials.

Despite growing interest in the use of LLMs in educational analytics processes, little is currently known about their ability to work with educational program self-assessment texts. In the context of Ukrainian-language research, the first attempts to create specialized benchmarks for evaluating LLM capabilities are already emerging [5]. However, it has scarcely been explored whether models can consistently evaluate such texts, differentiate their quality characteristics, and how their assessments correlate with the official NAQA criteria. In this research, the capabilities of GPT-5 and Gemini-2.5 were evaluated by comparing their results on real self-assessment texts.

The purpose of this study is to investigate how modern LLMs interpret and apply the official NAQA criteria when analyzing educational program self-assessment texts, as well as the extent to which LLM assessments are consistent and justified. The models GPT-5 and Gemini-2.5 were used to assess their ability to work with long normative texts, distinguish between strong and weak reports, and identify the key document properties that influence the evaluation. Additionally, it was investigated whether textual criteria can complement the official evaluation system and serve as a tool for preliminary analytical auditing of such documents. A separate focus was placed on analyzing which elements of the documents' content influence the LLM assessments and correlate with program quality.

MATERIALS AND METHODS

In this work, educational programs are understood as higher education degree programs evaluated within formal accreditation frameworks. The study focuses on the analysis of self-evaluation texts and does not address the evaluation of software or programming code.

Ten self-assessment texts of educational programs that are publicly available on the official website of Ivan Franko National University of Lviv were used in the experiments. The sample included five programs that received higher expert evaluations based on accreditation results, and five programs that received the lowest evaluations. All documents were structured in accordance with NAQA requirements, ensuring their comparability. Each text was processed separately by two LLMs: GPT-5 [6] and Gemini-2.5 [7].

Two sets of criteria were applied for evaluating the self-assessment texts. The first contained 10 official NAQA criteria [8], each with a score ranging from 1 to 10. The criteria

cover educational program design, its structure, staffing, learning environment, internal quality assurance system, and other components, as shown in **Table 1**. Since Criterion 10 applies only to third-level educational programs (PhD level), its scores were not included in the subsequent analysis.

Table 1. **NAQA criteria.**

#	Criterion	Short Description
1	Educational Program Design	Corresponds to the standard of higher education and professional standards. Clear goal, aligned with the HEI's mission. Consideration of labor market needs and trends, and the regional context.
2	Structure and Content of the Educational Program	ECTS credit volume complies with legislation. Logical structure, relevance to the subject area. Possibility of an individual educational trajectory. Practical training, social skills. Consideration of the UN Sustainable Development Goals.
3	Access to the Educational Program and Recognition of Learning Outcomes	Transparent and non-discriminatory admission rules. Procedures for recognizing learning outcomes acquired in other institutions, as well as results of non-formal and informal education.
4	Learning and Teaching within the Educational Program	Student-centered approach, academic freedom. Availability of syllabi. A combination of learning with research. Updating content based on scientific achievements. Internationalization.
5	Monitoring Measures, Evaluation of Higher Education Students, and Academic Integrity	Clear and published evaluation criteria. Clear rules for conducting exams, preventing conflicts of interest, and appeal procedures. Effective policy and culture of academic integrity.
6	Human Resources	Faculty qualifications meet requirements. Transparent competitive selection. Involvement of employers and practitioners in teaching. Promotion of faculty professional development.
7	Educational Environment and Material Resources	Sufficiency of material and technical base, libraries, and software. Safe and inclusive environment. Support for physical and mental health. Anti-corruption policies and anti-discrimination measures.
8	Internal Quality Assurance of the Educational Program	Procedures for program monitoring and review. Involvement of students and employers as partners. Tracking graduate careers. Response to surveys and feedback from previous accreditations.
9	Transparency and Publicity	Clear rules for all participants. Public discussion of program drafts before approval. Publication of complete information on the website.
10	Learning through Research	Preparation for solving complex problems. Alignment of research with the supervisor's field. Presence of specialized academic councils. Resource provision for research. Integration into the international community. Integrity of supervisors.

The formation of the set of textual metrics was based on modern approaches to the qualitative properties of text analysis used in the field of Natural Language Generation (NLG) [4] and in the evaluation of language models. These properties include structural organization, logical coherence, argumentative completeness, semantic connectivity, factual correctness, and lexical variability.

In this study, all metrics were computed directly by the LLMs, which were instructed to assign each metric a value on a scale from 1 to 10 based on the document's content, without any subsequent manual adjustments. The textual criteria used are described in **Table 2**.

Table 2. **Textual Criteria for the Assessment of Self-Evaluation Reports.**

Metric	Abbreviation	Essence of the Metric
Structural Consistency	S	Integrity and logical coherence of the document's structure; presence of all key sections.
Criteria Coverage	C	The proportion of NAQA criteria for which substantive explanations are provided.
Argumentative Saturation	A	The quantity and quality of arguments, examples, and evidence-based statements in the text.
Lexical Diversity	L	The degree of vocabulary variety, including the proportion of unique lexemes, the presence of terminology, synonyms, and different language registers.
Evaluation Variability	V	Diversity of types of evaluative judgments in the text (positive, negative, neutral), as well as the balance between them and their sources of origin.
Semantic Connectivity	SC	Logical and linguistic coherence between adjacent sentences in the document.
Factual Saturation	F	The number of references to regulatory documents or standards.
Model Motivation Clarity	M	The presence of an explicit explanation for the logic the model followed when forming the evaluation.

The metrics for structural consistency and argumentative saturation rely on the principles of logical organization of text analysis described in modern NLG evaluation systems [4]. Completeness of coverage reflects the model's ability to encompass all structural components of the document correctly and is based on coverage-based assessment approaches in NLG. The factual saturation metric is based on practices for verifying the reliability of statements in language model responses and related research on factuality [9]. The semantic connectivity metric applies approaches to sentence similarity in embedding space analysis, a typical approach for evaluating the coherence of generated texts [10].

The variability metric reflects the diversity of evaluative judgments present in the document's text. It takes into account the balance between positive, negative, and neutral statements, as well as different sources of evaluation, such as claims about strengths, acknowledgment of shortcomings, and neutral descriptions of processes. The lexical diversity metric assesses the degree of vocabulary variation, including the proportion of unique lexemes, the use of terminology, synonyms, and different linguistic registers. It reflects the richness of the language and the stylistic saturation of the text and operates based on the Type-Token Ratio [11]. Clarity of motivation assesses the presence of explicit explanations regarding the logic of evaluation formation and is related to approaches for evaluating the transparency of reasoning processes in LLMs.

Thus, the set of metrics combines generally accepted approaches for studying academic texts with modern methods for evaluating the quality of LLM responses, and is used in this study as an additional analytical tool.

The research focused on three aspects:

- Relationship between NAQA criteria scores and textual metrics. It was checked whether self-assessment text metrics are associated with high or low scores on the NAQA criteria.
- Inter-model consistency. The correlation between the two models' scores was analyzed separately for the NAQA criteria and the textual metrics. This allowed for determining how stable and reproducible the decisions of different LLMs are when working with identical documents.
- Internal structure of each model's evaluations. It was investigated which textual properties most strongly influence the NAQA scores within each model, and whether the models form similar patterns.

The Pearson correlation coefficient [12] was used for quantitative analysis, enabling assessment of the linear relationships between the criterion values on a unified scale. Additionally, Spearman's rank correlation coefficient [13] was employed to analyze relationships based on score rankings. Together, these correlation measures were used to examine whether the textual metrics and NAQA scores generated by the models are sufficient to distinguish between program self-assessment documents previously identified by experts as stronger or weaker.

Since the study sample comprises only 10 documents, the obtained Pearson correlation coefficients should be considered indicative. With so few observations, a single atypical document can significantly influence the values of the correlation indicators. Therefore, the interpretation of correlations is presented as identified tendencies rather than statistically confirmed patterns.

To simplify the analysis and comparison, the results were structured using conditional notation. The five programs that experts previously identified as higher quality are denoted as B1-B5 (Best). The five programs that received the lowest expert evaluations are denoted as W1-W5 (Weakest). All graphical materials, tables, and correlation analysis results use these notations for data unification.

RESULTS AND DISCUSSION

The results of the models' work, presented in **Fig. 1**, demonstrate varying behavior between the NAQA and textual criteria. Both LLMs are generally prone to assigning high scores based on the NAQA criteria. However, significantly more variation is observed in the textual criteria, especially for metrics such as lexical diversity, semantic connectivity, and evaluation variability. The NAQA scores are less sensitive to the text's actual properties than the textual criteria, which more accurately capture the quality of the document's content.

Notably, the Gemini-2.5 model shows a significantly wider spread of textual criteria scores, whereas the GPT-5 model assigns scores less frequently at 5-6 and below. This points to different internal text analysis strategies: GPT-5 focuses on structural integrity and general style, while Gemini-2.5 concentrates more on argumentation, facts, and logical connectivity.

The correlation analysis showed that the relationship between the NAQA criteria and the textual characteristics is weak or unstable in most programs. Even in cases where the text contains a significant number of facts, examples, and clear arguments, the model may assign very high scores for structural criteria but react weakly to the quality of the content. In program W5, GPT-5 assigns 10 points for Criterion 1, "Educational Program Design" but the textual criteria indicate poor semantic connectivity (SC = 6). This highlights a disconnect between the formal fulfillment of the structure and the low quality of the content.

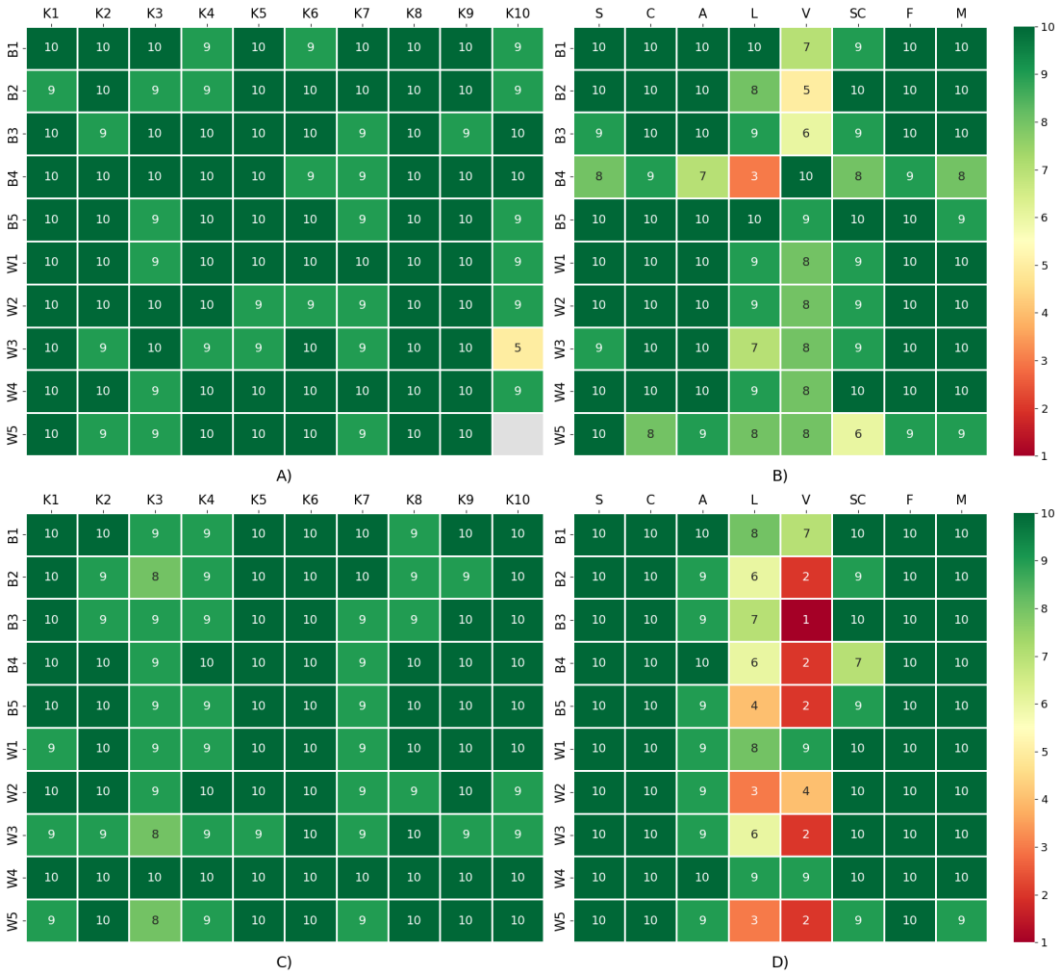

Fig. 1. Self-evaluation report evaluation scores obtained from GPT-5 and Gemini-2.5 models: (A) NAQA criteria evaluation by GPT-5; (B) Textual criteria evaluation by GPT-5; (C) NAQA criteria evaluation by Gemini-2.5; (D) Textual criteria evaluation by Gemini-2.5.

Furthermore, it is one of the key reasons for the weak correlation between the groups of criteria in the "weaker" programs: the text appears to be correctly constructed but lacks objective evidence or substantiation.

For program self-assessment texts that experts rated as stronger or B-group, a generally higher correspondence between the NAQA criteria and the textual criteria is observed compared to the texts in the W-group. **Fig. 1** shows that most documents in this group exhibit high values for argumentativeness, factual saturation, and structural consistency. High scores often accompany these indicators on the NAQA criteria.

Despite the general trend, individual strong programs demonstrate localized shortcomings in specific criteria:

- In B1, both models lower the scores for the Lexical Diversity criterion, despite high scores on other indicators, likely due to the homogeneity of terminology.
- In B2, the Variability of Evaluations criterion is lower, indicating the document's self-critical nature, even though the NAQA criteria scores are high.
- In B3 and B4, specific textual criteria have lower values than others, indicating weaker internal logical density or an insufficient number of facts.
- Some NAQA criteria receive lower scores even in strong programs.

Thus, strong programs demonstrate a higher correlation between text quality and model scores. This suggests that LLMs can notice and reflect the quality of strong programs, but even in strong documents, specific textual characteristics may be weak, and the models capture this.

In documents B1-B4, elevated values for Argumentative and Factual Saturations are associated with the presence of specific details directly relevant to the educational program. In B1, this consists of a list of specific software packages. In B2, it links to the national context. In B3, it is the program's adaptation to changing learning conditions. In B4, it refers to the employer and real cases of cooperation.

The Variability of Evaluations showed interesting patterns. In documents where the program openly acknowledges shortcomings, Evaluation Variability decreases: the text becomes more stylistically uniform, restrained, specific, and lacks excessive self-praise. Despite this, such documents received high scores for the "Internal Quality Assurance" criterion, indicating that honesty and self-reflection do not lower the model's evaluation but rather align with the signs of quality programs. In contrast, documents where shortcomings are hidden behind general phrases demonstrate higher Evaluation Variability values - that is, a greater diversity of evaluative statements dominated by generalized claims not substantiated by specifics. Such texts often also have lower Factual Saturation values, indicating a deficit of concrete information.

Self-assessment documents for the W-group programs demonstrate a characteristic pattern: structural consistency is always high, but factual saturation is low. For example, document W2 directly acknowledges the lack of resources for several program components, and W3 notes an insufficient number of courses on Moodle, which is reflected in lower scores for the "Educational Environment" criterion. This is a typical situation in which the text has a formal structure but lacks specifics, as reflected by low Factual and Argumentative Saturation metrics.

Although both models were evaluated on identical input texts, the results suggest that they differ in their sensitivity to stylistic noise present in the self-evaluation reports. In this study, stylistic noise refers to superficial stylistic features inherent to the documents, such as template-based phrasing, repetitive generic statements, lexical redundancy, and declarative claims without supporting evidence, that do not add substantive content. While these features are part of the input texts, the models appear to weigh them differently, leading to varying levels of score variability and sensitivity to document quality.

Summarizing the results, both models demonstrate sensitivity to the content characteristics of self-assessment documents, but in different ways. In stronger programs, moderate or high consistency is observed between the NAQA criteria and the textual metrics, whereas weaker programs are characterized by a disconnect between formal structure and substantive content. Textual metrics, Argumentative and Factual Saturations, and Semantic Connectivity proved to be more informative indicators of the actual quality of the documents than the scores based on the NAQA criteria, which both models are prone to assigning mainly in the high range. Furthermore, the GPT-5 model demonstrated more stable and conservative behavior, while the Gemini-2.5 model showed greater variability and sensitivity to specific textual deficiencies. The results indicate that LLMs can identify substantial qualitative differences between programs, but their assessments based on formal criteria require careful interpretation and supplementation with textual characteristics of the documents.

The correlation matrices presented in **Fig. 2 (A–D)** illustrate the relationships between the NAQA criteria and the textual metrics using Pearson’s and Spearman’s correlation coefficients for Gemini-2.5 and GPT-5. Panels **(A)** and **(C)** show Pearson correlation matrices, while panels **(B)** and **(D)** present Spearman rank correlation matrices. In all cases, correlations were computed across the complete set of 10 self-assessment documents, enabling the analysis of general associations between formal accreditation criteria and textual characteristics irrespective of the expert-based grouping of reports.

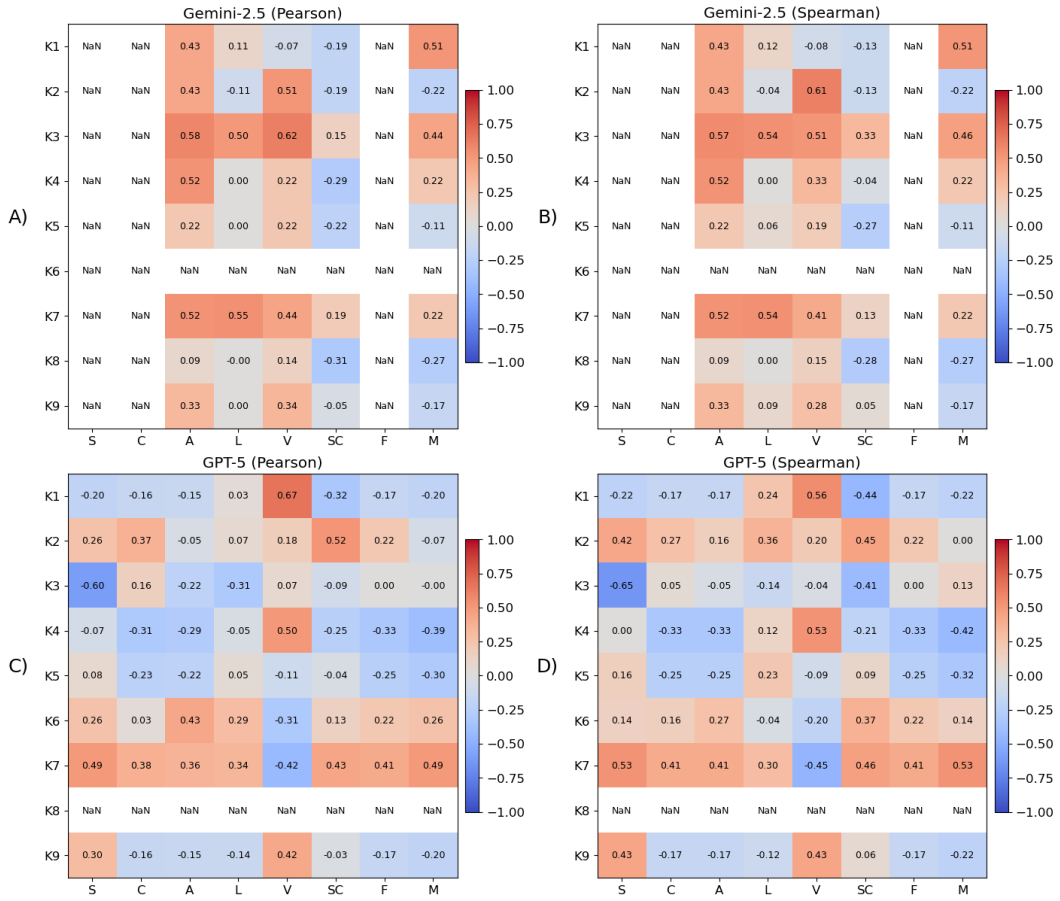

Fig. 2. Correlation between NAQA criteria and textual metrics computed using Pearson's and Spearman's coefficients. Panels show: (A) Gemini-2.5 (Pearson), (B) Gemini-2.5 (Spearman), (C) GPT-5 (Pearson), and (D) GPT-5 (Spearman).

Given the limited sample size, the obtained Pearson and Spearman correlation coefficients should be interpreted as indicative tendencies rather than statistically confirmed relationships. Pearson's correlation reflects linear associations on the unified numerical scale produced by the models, whereas Spearman's rank correlation captures monotonic relationships based on the relative ordering of scores. The use of both coefficients provides complementary perspectives on the same data and allows for assessing the stability of the observed patterns under different correlation assumptions.

The overall similarity between Pearson and Spearman correlation matrices indicates that the identified associations are not driven by individual extreme values or rank-specific effects, but represent stable tendencies present in the analyzed sample. At the same time, the correlation analysis reveals only isolated local relationships between specific textual metrics and NAQA criteria, with no consistent or universal dependence across criteria. This finding supports the conclusion that textual quality characteristics alone do not constitute a reliable proxy for formal accreditation assessment, and that LLM-based evaluations reflect document properties that are only partially aligned with the official NAQA criteria.

The heatmaps show that GPT-5 primarily assigns scores of 9-10 and rarely assigns low scores. This may be a consequence of RLHF (Reinforcement Learning from Human Feedback) [14], which can make the model overly polite and "helpful." In auditing tasks, this may not be very appropriate, and the model should be specifically prompted to criticize.

In the present study, expert evaluations are treated as correct and authoritative. These evaluations were produced by formal expert committees and served as the basis for accreditation decisions. Therefore, they constitute the only available, institutionally validated reference point for distinguishing stronger from weaker educational programs. In this sense, the expert-based classification of programs reflects real decision-making outcomes rather than abstract or hypothetical judgments.

At the same time, it is important to note that the expert assessments used in this study did not include explicit cumulative numerical scores. The information accessible to the authors consisted of the categorical distinction between programs evaluated as stronger and weaker, without a unified quantitative scale that could be directly used for correlation analysis. As a result, a direct comparison between cumulative expert scores and aggregated model-based evaluations was not feasible within the scope of the present dataset.

Given these constraints, the analysis focused on criterion-level scores generated by the language models and on their ability to differentiate between programs previously identified by experts as stronger or weaker. This design allows for investigating whether LLM-based evaluations, expressed either through formal accreditation criteria or through textual quality metrics, are consistent with expert judgments at the group level, even in the absence of explicit numerical expert ratings.

From this perspective, the observed differences in alignment between expert judgments and model-generated scores should not be interpreted as questioning the validity of expert evaluations. Instead, they highlight the extent to which different sets of criteria capture aspects of the documents that are implicitly or explicitly reflected in expert decisions. The results thus provide insight into how expert judgments may be reflected in textual properties of self-evaluation reports, rather than serving as a direct validation or refutation of either the expert assessments or the accreditation criteria themselves.

It is worth mentioning that the study has certain limitations. First, the sample is small, which does not allow for generalizing about all educational programs. Second, the analysis is limited to two models, which, although modern, may yield different results depending on the version or operating mode. Third, the NAQA scores generated by the models depend on the specific prompt formulation and may change under other conditions. Furthermore, LLMs do not have access to the contextual data about educational programs used in real accreditation expert evaluations, and can only analyze what is presented in the text.

Despite these limitations, the results demonstrate the potential of using LLMs for the preliminary analysis of self-assessment documents. The combination of formal NAQA criteria with textual metrics allows for obtaining a multidimensional view of document quality and for identifying weaknesses that are not always obvious from the text's primary structure. This opens the door to creating semi-automated tools to support expert groups, monitor documents, and enhance transparency in accreditation procedures.

CONCLUSION

This work assessed the ability of LLMs to analyze the quality of educational program self-assessment texts using official NAQA criteria and a set of textual metrics. The results demonstrated that GPT-5 and Gemini-2.5 generally correctly interpret the logic of the NAQA criteria but exhibit different sensitivity to their individual components. Specifically, the models well identified the structural, semantic, and factual properties of the text, but reacted more weakly to those elements of the criteria that do not have a direct textual representation or require the broader context of the educational program's functioning. This indicates that the current form of the NAQA criteria, when interpreted as a text query, does not always allow the models to reproduce the depth of expert evaluation. The identified tendencies were consistent across both Pearson and Spearman correlation analyses, indicating that the observed patterns are robust to the choice of correlation measures.

In contrast, the textual metrics proved effective at distinguishing between strong and weak self-assessment documents. High values for argumentative and factual saturation, as well as semantic connectivity, were associated with programs that received better expert evaluations. In contrast, low scores on these metrics corresponded to texts with less qualitative content. This confirms that LLMs can not only identify the structure of a document but also perform a preliminary substantive analysis in many aspects similar to the logic of expert evaluation. It is important to emphasize that textual metrics do not replace the NAQA criteria and are not an automated assessment. Instead, they complement expert judgment, helping identify potential weaknesses in the document quickly.

A separate finding is that LLMs demonstrate sensitivity to honesty and self-reflection in documents. Programs that openly describe their shortcomings received more balanced evaluations and higher scores on criteria related to internal quality assurance. This indicates that LLMs can consider not only the strengths of the document but also its integrity and realism.

Despite significant differences between the models, GPT-5 and Gemini-2.5 demonstrated similar general tendencies, suggesting the approach's scalability and independence from a specific architecture. At the same time, the work revealed several limitations, including a small sample of documents, models' sensitivity to prompt formulations, and a lack of contextual information about the real operating conditions of the educational programs.

Overall, the results confirm the potential of LLMs as a tool for analytical support in accreditation processes. The combination of formal NAQA criteria with textual metrics creates opportunities for developing semi-automated systems for monitoring and preliminary analysis of the quality of educational programs. Future research can be directed toward expanding the document corpus, comparing a greater number of models, standardizing instructions, and creating specialized tools to support higher education expert groups.

ACKNOWLEDGMENTS AND FUNDING SOURCES

The authors received no financial support for the research, writing, and/or publication of this article.

COMPLIANCE WITH ETHICAL STANDARDS

The authors declare that the research was conducted in the absence of any conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, [M.S.]; methodology, [M.S., V.K.]; validation, [V.K.]; writing – original draft preparation, [M.S.]; writing – review and editing [M.S., V.K., B.P.]; supervision, [B.P.].

All authors have read and agreed to the published version of the manuscript.

ДЖЕРЕЛА

[1] Mazzullo, E., Bulut, O., Wongvorachan, T., & Tan, B. (2023). Learning analytics in the era of large language models. *Analytics*, 2(4), 877–898. Doi: https://doi.org/10.3390/analytics2040046

[2] Aboalela, R. (2024). Harnessing technology to achieve the highest quality in the academic program of university studies. *International Journal of Advanced Computer Science and Applications*, 15(8). https://doi.org/10.14569/IJACSA.2024.0150829

[3] Huang, Y., Tang, K., Chen, M., & Wang, B. (2024). A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2404.15777*. https://doi.org/10.48550/arXiv.2404.15777
[4] Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023, December). G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2511–2522). https://doi.org/10.18653/v1/2023.emnlp-main.153
[5] Syromiatnikov, M., Ruvinskaya, V., & Troynina, A. (2025). ZNO-Eval: Benchmarking reasoning capabilities of large language models in Ukrainian. *arXiv preprint* arXiv:2501.06715. https://doi.org/10.48550/arXiv.2501.06715
[6] OpenAI. (2025). GPT-5 system card. https://cdn.openai.com/gpt-5-system-card.pdf
[7] Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., ... & Mehta, S. V. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint* arXiv:2507.06261. https://doi.org/10.48550/arXiv.2507.06261
[8] Ministry of Education and Science of Ukraine. (2024). On approval of the Regulations on the accreditation of educational programs for the training of higher education applicants (in Ukrainian). Order No. 686 on May 15, 2024. https://zakon.rada.gov.ua/laws/show/z1013-24
[9] Muhlgay, D., Ram, O., Magar, I., Levine, Y., Ratner, N., Belinkov, Y., … & Shoham, Y. (2024, March). Generating benchmarks for factuality evaluation of language models. In: *Proceedings of the 18th conference of the european chapter of the association for computational linguistics* (Vol. 1: Long papers) (pp. 49–66). https://doi.org/10.18653/v1/2024.eacl-long.4
[10] Pavlyshenko, B., & Stasiuk, M. (2025). Semantic Similarity Analysis Using Transformer-Based Sentence Embeddings. *Electronics and information technologies*, (30), 43–58. https://doi.org/10.30970/eli.30.4
[11] Templin, M. C. (1957). *Certain Language Skills in Children: Their Development and Interrelationships* (NED-New edition, Vol. 26). University of Minnesota Press. 208 p. http://www.jstor.org/stable/10.5749/j.ctttv2st
[12] Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Routledge. 535 p. https://doi.org/10.4324/9780203774441
[13] Conover, W. J. (1999). *Practical nonparametric statistics*. John Wiley & Sons. 608 p.
[14] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA. *Advances in neural information processing systems*, 30.

ВИКОРИСТАННЯ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ ДЛЯ ТЕКСТОВОГО АНАЛІЗУ В ОЦІНЮВАННІ УНІВЕРСИТЕТСЬКИХ ОСВІТНІХ ПРОГРАМ

Микола Стасюк[1]*, Віталій Кухарський[2], Богдан Павлишенко[1]
[1]Кафедра системного проектування,
Львівський національний університет імені Івана Франка,
вул. Драгоманова 50, 79005 Львів, Україна
[2]Кафедра прикладної математики,
Львівський національний університет імені Івана Франка,
вул. Університетська 1, 79000 Львів, Україна

АНОТАЦІЯ

Вступ. Великі мовні моделі дедалі частіше застосовуються в аналітиці освіти, зокрема цікавим питанням є дослідження, пов'язані з опрацюванням великих обсягів акредитаційних документів. Відкритим, наприклад, є питання щодо того, наскільки надійно великі мовні моделі можуть аналізувати якість текстів самооцінювання освітніх програм і які характеристики тексту впливають на те, як моделі формують свої оцінки.

Матеріали та методи. У дослідженні проаналізовано десять звітів самооцінювання освітніх програм: п'ять, за загальною оцінкою експертів виділені як найкращі у закладі вищої освіти за три останні роки та п'ять, як найслабші за цей же ж період. Моделі GPT-5 і Gemini-2.5 незалежно оцінювали кожен документ з використанням офіційних десяти критеріїв НАЗЯВО та восьми текстових метрик, що відображають структурні, семантичні, аргументаційні й фактологічні властивості тексту. Усі метрики були згенеровані безпосередньо моделями у єдиній шкалі від 1 до 10. Для аналізу взаємозв'язків між критеріями НАЗЯВО та текстовими оцінками використано коефіцієнти кореляції Пірсона та Спірмена.

Результати. Великі мовні моделі продемонстрували обмежену узгодженість з критеріями НАЗЯВО, виявивши слабкі та нестабільні кореляції між ними. Натомість текстові критерії, передусім фактологічна насиченість, аргументованість, семантична зв'язність і лексична різноманітність, стабільно розрізняли сильніші та слабші звіти. GPT-5 демонструвала меншу варіативність і слабшу залежність від стилістичних шумів, тоді як Gemini-2.5 активніше реагувала на структурні та стилістичні недоліки тексту. Кореляційні матриці підтвердили, що текстові критерії краще відображають приховані якісні властивості документів, порівняно з прямим застосуванням критеріїв НАЗЯВО.

Висновки. Результати свідчать, що великі мовні моделі наразі недостатньо точно відтворюють експертне оцінювання за формальними критеріями НАЗЯВО, але ефективно аналізують структурні й змістові характеристики звітів за допомогою текстових метрик. Ці метрики слід розглядати як допоміжний інструмент аналізу, який може прискорити роботу експертів та підвищити якість моніторингу документів. У подальших дослідженнях планується розширення вибірки, стандартизації запитів і порівняння ширшого кола моделей.

Ключові слова: великі мовні моделі, освітні програми, якість оцінювання.

Received / Одержано	Revised / Доопрацьовано	Accepted / Прийнято	Published / Опубліковано
04 December, 2025	06 January, 2026	12 January, 2026	30 March, 2026