EduDesignEval: A Structured Evaluation Framework for Educational Program Designer Using Large Language Models

Ayanbek Serikov [1,*], Andrii Biloshchytskyi [2,3], Aidos Mukhatayev [4,*], and Balgyn Zheldybayeva [5]

[1] Department of Computer Engineering, Astana IT University, Astana, Kazakhstan
[2] Department of Computational and Data Science, Astana IT University, Astana, Kazakhstan
[3] Department of Information Technologies, Kyiv National University of Construction and Architecture, Kyiv, Ukraine
[4] Department of Social Disciplines, Astana IT University, Astana, Kazakhstan
[5] Department of Physics and Computer Science, Shakarim State University, Semey, Kazakhstan
Email: ayanbek.as@gmail.com (A.S.); a.b@astanait.edu.kz (A.B.); aidos.mukhatayev@astanait.edu.kz (A.M.); balgun@mail.ru (B.Z.)
*Corresponding author
Manuscript received January 16, 2026; revised March 4, 2026; accepted March 23, 2026; published September 15, 2026

***Abstract*—The rapid advancement of Large Language Models (LLMs) has created unprecedented opportunities for transforming educational program design. However, the lack of standardized evaluation frameworks raises concerns regarding the content quality and pedagogical effectiveness. This study proposes EduDesignEval, a comprehensive evaluation framework designed to assess LLM-generated educational content across multiple dimensions of quality. The framework defines three metrics: groundedness for measuring factual accuracy and hallucination mitigation, coherence for evaluating logical flow and thematic integrity, and fluency for assessing linguistic naturalness and readability. These metrics are synthesized into a unified score that provides an adaptable, weighted assessment based on specific educational requirements. We first validated the EduDesignEval framework through metric definitions and demonstrated its utility by employing ten publicly available LLM models across two critical tasks: text generation for learning goals and course descriptions, and multi-label classification for professional standards and competency frameworks, using a dataset derived from over 8,200 educational programs across 150 Kazakhstani institutions. Our fine-tuned model has demonstrated promising zero-shot classification performance with high groundedness while maintaining balanced coherence and fluency, outperforming larger models utilizing one-shot prompting. Preliminary results show that big models excel in groundedness, while smaller models show limitations in classification due to redundancy and low coherence. EduDesignEval bridges gaps in literature by providing multi-level quality control, fostering human-Artificial Intelligence (AI) collaboration, and promoting equitable, adaptive education. Ultimately, this framework ensures reliable LLM integration, enhancing program design efficiency and pedagogical integrity for global educational systems.**

***Keywords*—large language models, educational program design, evaluation framework, curriculum development, ai in education, text generation, multi-label classification**

I. Introduction

The growth of Large Language Models (LLMs) and generative Artificial Intelligence (AI) is fundamentally changing the way educational programs are designed. This rapid transformation is already having a significant impact on educational practice and may serve as a basis for discussing some systemic shifts in the principles of building the educational process. However, it is fair to say that some of the most essential characteristics of the teacher's and students' activities are not affected by such changes. The teacher is still passing on knowledge, and the students are learning it. A certain verifier of cultural heritage (a teacher) transmits it, and the absorbing receiver (students) perceives it with varying degrees. Moreover, this level is again determined from the position of the teacher as an evaluator of the quality of education.

In many educational systems, teachers are no longer solely responsible for assessment, as these functions are often delegated to independent agencies that conduct monitoring, diagnostics, and state final certification. So, these roles could be filled with AI as well. LLMs such as GPT-4, Claude, and Llama are becoming key tools in the transformation of educational processes [1]. The introduction of LLM models into the educational process should lead to a reduction in the academic and extracurricular burden on teachers and expand the opportunities for self-education of all participants in the educational process.

Traditional education is increasingly facing problems such as (1) the collaborative approach of integrating insights, concepts, and methodologies from multiple academic disciplines [2], (2) the need for structured educational programs [3], and (3) the lack of adaptive personalization [4]. Recent advances in AI within educational environments have created new opportunities to overcome these barriers.

Interdisciplinarity has become essential, as modern educational systems increasingly need to adapt knowledge from different domains. For example, a sustainable development program can combine ecology, sociology, and engineering to provide solutions with holistic approaches. Such methods form critical thinking and professional competencies to solve real-world problems from several perspectives. The willingness to look broadly is a necessity. Science is going through another round of development. There were generalists as philosophers and mathematicians, then specialization went on, and now there is a new round when specialists with universal competencies are in demand. We can see it in bioinformatics, cybernetics, robotics, and others. Researchers highlight difficulties matching methodologies, terminologies, and organizational barriers between academic disciplines [2]. Also, it is laborious, costly, and time-consuming. Yet interdisciplinarity is a key to innovative education, especially in the digital world.

The next important necessity in educational programs is structure, which ensures a clear, logical flow and coherence in a limited time frame. It builds up transition from basic to complex, leading to deeper knowledge acquisition, increased retention, and motivation. For example, chemistry class should start with fundamentals and theory, then gradually introduce complex elements so that students can adapt on the

journey. But too much structure could be harmful as well in the form of blocking flexibility and adaptation to various learning styles. Modern research notes the need to balance standardization and modularity via creating "structured flexibility" [3]. While this goal is conceptually appealing, its implementation takes strategic planning and prowess, which is hard to achieve in tight schedules.

The third is to make personalized learning a lever for equitable education. Personalized systems are being developed considering the diversity of students, with their backgrounds and experiences. It is assumed that properly designed AI will reduce bias and help each student, regardless of their capabilities, receive the support they need. For example, adaptive programs for children with special educational needs will automatically adjust to their pace and way of perception, giving them additional time or alternative tasks. It can significantly increase access to education. In addition, it will help compensate for the shortage of teachers and educational materials by offering each student basic personal learning support. Of course, it depends a lot on investments and policies. Studies back this AI improved engagement by 40% and test scores by 25% in some cases [5]. Another research showed 20% in equity and 30% in educator improvements via explainable AI [6]. Although the exact numbers vary, the trend is clear.

The lack of a unified framework for assessing the quality of LLM-generated content leads to risks of introducing ineffective or harmful materials into educational programs. The existing approaches (prompt engineering, few-shot learning) are fragmented and do not provide comprehensive control. Therefore, there is a need for a system (1) to standardize evaluation criteria of educational content of LLM, (2) to provide multi-level quality control, (3) to systematize the process of the program designer and LLM interaction, and (4) to improve the effectiveness of educational program development. This study proposes a structured assessment framework—a specialized benchmark for evaluating LLM content in educational tasks, allowing an objective comparison of models and generation methods.

II. Literature Review

The integration of AI introduces a different structural dynamic in education. We already see some changes that have the most significant impact on the transforming positions of the teacher and student. To begin with, the use of AI in education can be described as a new, effective tool in the work of a teacher from a traditional perspective. The discussion extends beyond general generative AI models, including also specially configured ones, when the model is trained on the texts of a certain historical character, or the model has a preset position from which it interacts with the user. This approach enables outcomes that were previously unattainable prior to the emergence of AI technologies. These include prompt reconciliation of links and citations, overcoming the "fear of the blank slate" when it is necessary to begin creating a narrative but initial ideas do not come to mind, and others [7]. Similarly, you can systematize various new services for students and scientists.

The concept of LLMs as curators and auditors is examined in several publications [4, 8]. For example, Padovano and Cardamone [9] proposed an AI-powered framework using semantic analysis to identify gaps for dynamically updating the WING model for Industry 5.0. Similarly, Brown *et al*. [10] developed the CIRCLE methodology, where LLMs support mapping to NCAE-C Cybersecurity Defense Education, reducing time spent by over 50%. Yet they are susceptible to hallucinations and constrained by their training data.

Research shows that LLMs can generate code fragments and provide debugging assistance that could rival human capabilities in clarity, although they require human verification [11, 12]. For example, it achieved a 95.83% correctness rate in code generation and 75% on the Aizu Online Judge platform. It demonstrates how LLMs have been used to provide context-specific outputs, allowing for a degree of personalization that was previously unachievable on a scale. It underscores the potential of human-AI collaboration in educational program design. Despite the numerous generated data, the pedagogical alignment is a crucial factor, with some models giving linguistic fluency precedence over instructional soundness.

Language models don't ignore optimization and satisfaction. Empirical studies demonstrate that automating routine tasks reduces the workload of developers, speeds up revisions, and increases overall job satisfaction, achieving 37% time reductions and 0.4 SD quality improvements [13]. Research conducted by a group at Stanford University demonstrates that language models can be used not only as a text generation tool but also as a means of approximately replicating expert assessments [14]. The authors demonstrate that such models can predict learning outcomes and reproduce robust effects described in educational psychology, replicating the "Expertise Reversal Effect and Variability Effect".

A separate development area is LLM-based multi-agent systems. This approach allows you to consistently perform the tasks of generating, clarifying, structuring, and editing the material in the various stages of content creation, which contributes to the adaptation to the age characteristics of students. The use of such systems also facilitates the inclusion of morally oriented elements and visual support while complying with ethical and pedagogical standards, thereby enhancing students' perception of the material [15]. Research in the field of technology adoption among students of pedagogical training areas shows that ease of use, reduction of perceived risks, and accumulated user experience have a significant impact on the willingness to apply LLM in educational activities [16]. This contributes to the formation of a consistently positive attitude towards the use of language models to support written work and increases satisfaction with digital educational practices.

At the same time, the development of AI-oriented educational solutions requires rethinking approaches to assessing the quality of created content. Several authors note that standard Natural Language Processing (NLP) metrics such as BLEU and ROUGE are primarily focused on formal text matching and do not reflect the pedagogical correctness or meaningful reliability of educational materials [17, 18]. In response to these limitations, Huang *et al*. [19] proposed a multifactorial system for evaluating digital educational resources based on Delphi methods and analytical hierarchy. Within the framework of this system, meaningful, expressive,

user-defined, and technical aspects of quality are highlighted, among which authenticity, accuracy, legitimacy, and relevance are recognized as key. This approach creates the basis for a more objective and reproducible assessment of educational content. Yet it relies on iterative expert consensus and therefore is neither automated nor scalable across LLMs. Complementing these ideas, Khosravi *et al*. [20] emphasizes the need to combine automated checks with the participation of experts to assess the pedagogical applicability of materials, and Fuchs [21] draws attention to the importance of analyzing educational consistency, that is, the correspondence of the generated tasks to the stated learning goals. Their findings support the potential of AI applications in enhancing educational quality when properly evaluated and implemented. EduDesignEval, on the other hand, addresses these collective shortcomings by integrating three dimensions. It addresses a particularly underexplored gap. It is the simultaneous evaluation of LLMs across both generative and classificatory educational tasks.

Building on these considerations, later observational research on stakeholder perceptions in Kazakhstan in the field of higher instruction quality assurance highlights deep-rooted, interconnected problems such as irrelevant content, aging staff, and a lack of needed infrastructure. To address low satisfaction with the educational program's relevance and practical skills, data underscores the need for reforms to integrate AI [22]. Alternatively, basic models for interconnected data and instructive frameworks recommend unified systems that utilize ontological databases and microservices to combine different information sources, advertising real-time analytics and feedback to upgrade quality control and accommodate worldwide benchmarks like the Bologna Process. These models may be adjusted to evaluate AI-generated guideline materials [23].

Fewer studies focus on educational program evaluation. As noted in Ref. [24], "a growing trend among universities to promote systems of curricula and academic course evaluation entails more responsibility for faculties and departments." There is a noticeable gap in the research on curricula evaluation concerning their use and validation. Authors examine the content aspect of validity in a rubric-based assessment system for course syllabi.

Despite the obvious advantages, the use of machine learning models in education comes with several limitations. A major limitation concerns the reproduction and amplification of distortions present in the training data. Although language models have no intentions and do not pursue their own goals, they are not subjects of responsibility, which excludes the possibility of direct ethical obligations imposed on them. An additional factor is the different quality of language processing: currently, English remains the most reliable language for interacting with most models due to its dominance in teaching materials. When evaluating academic performance, LLMs demonstrate high internal consistency; however, they tend to assign lower grades than human experts and require clearly defined reliability criteria. In addition, excessive reliance on automated tools can negatively affect the development of critical thinking and reflective skills of students, which emphasizes the need for a balanced and methodically verified implementation of such technologies.

Several studies highlight the importance of bias, fairness, and ethical considerations. Bulut *et al*. [25] and Lee *et al*. [26] explored the various structures where AI is used in educational dimensions, from development to adaptation. They identified that bias could occur at any stage of the lifecycle, which leads to biases in automation and the dataset used for fine-tuning. There is a need for regulated metrics and ethical standards. The development of rigorous assessment methodologies remains fundamental to ensuring the reliability of LLMs in educational applications.

There are still fundamental gaps in the literature. On the one hand, while quality evaluation tools exist for specific instructional tasks, there is no complete benchmark framework that could compare LLMs across several quality metrics simultaneously. Standardized metrics for assessing educational suitability and cognitive skills remain underdeveloped. Furthermore, research on workflows between program designers and LLMs, including prompt engineering, quality assurance, and iterative refinement, remains limited. The dual-task gap is not merely technical. Practitioners require a framework that can work with produced content and classified content. The current study aims to develop standardized evaluation criteria for educational content generated by LLM within the context of educational program design.

It addresses the following Research Questions (RQs):

RQ1: To what extent can LLMs generate educationally valid content meeting factuality, coherence, and fluency metrics?

RQ2: How do LLMs differ in terms of text generation and multi-label classifications for educational content?

III. Materials and Methods

This section presents the evaluation framework developed to measure the quality of LLMs in educational program design. Fig. 1 outlines the steps of the proposed framework.

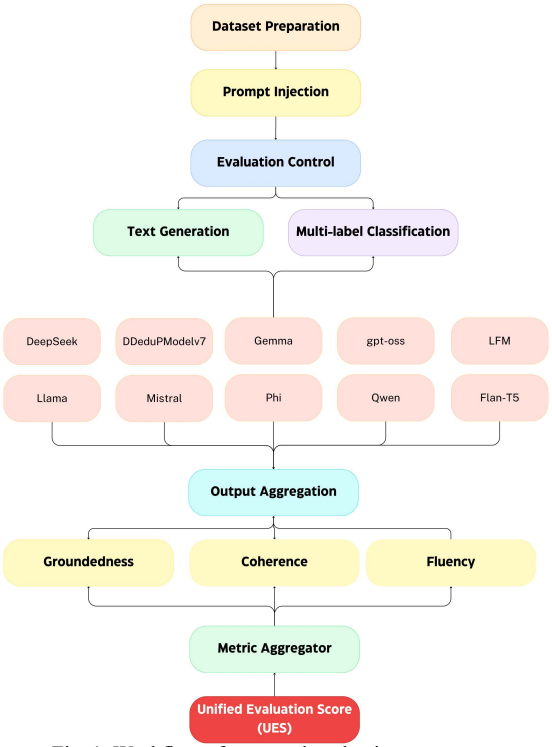

Fig. 1. Workflow of proposed evaluation system.

Several task categories were selected to represent LLM use

cases, such as learning goals, course descriptions, the Atlas of new professions and competencies in Kazakhstan, professional standards, etc. For multi-label classification, 1475 and 202 clean prompts were designed for Atlas professions and professional standards, respectively. While for text generation, every 200 clean prompts were done for learning goals and course descriptions. These prompts are available at the Fralet/DDEduPDatasetv2 repository, which contains data from 8200+ educational programs across 150+ Kazakhstani educational institutions. The dataset was created through manual and automatic scraping of publicly available curricula on the respective university's website.

To validate the EduDesignEval framework's practical utility, ten openly available LLM models (Table 1) were chosen and quantized to 4-bit. Our fine-tuned model was created by fine-tuning llama-3-8b-bnb-4bit followed by LoRA adaptation on DDEduPDatasetv2 (train). All experiments were conducted on Google Colab Pro using an NVIDIA A100 GPU. All ten models are open-source and publicly available via the HuggingFace hub. It means results are fully reproducible without licensing fees, and institutions can deploy them locally.

Text generation tasks used zero-shot prompting for our fine-tuned model DDeduPModelv7, as shown in the template above. Models were executed locally with max_new_ tokens = 3000, temperature = 0.7, and top_p = 0.9. The rest of the LLM models received one high-quality example in the prompt (one-shot prompting) using the same generation parameters. The zero-shot prompt is presented in Fig. 2.

```
def build_prompt(input_text: str, instruction: str) -> str:
    prompt = f"""
You are a Design Developer of Educational Programs (DDeduP). Below is an
instruction that describes a task, paired with an input that provides
further context. Write a response that appropriately completes the request.

### Instruction:
{instruction}

### Input:
{input_text}

### Response:
"""
    return prompt.strip()
```

Fig. 2. Employed prompt for models in evaluation system.

In the case of the multi-label classification task, for our model we used pure zero-shot prompting again. For the other models we implemented Retrieval-Augmented Generation (RAG). A knowledge base was constructed from (1) the Atlas of Emerging Jobs of Kazakhstan (463 entries) and (2) Professional Standards (3462 standards). Models used the same inference hyperparameters but with max_new_tokens = 1000. Texts were embedded using sentence-transformers/paraphrase-multilingual-mpnet-base-v2.

Table 1. Selected models for educational program design

Model	Parameters	Context Window	Paradigm Used (MLC / TG)
LiquidAI/LFM2-1.2B	1.2B	32,768 tokens	RAG / one-shot
unsloth/llama-3-8b-bnb-4bit	8B	8,192 tokens	RAG / one-shot
deepseek-ai/DeepSeek-R1-Distill-Llama-8B	8B	131,072 tokens	RAG / one-shot
microsoft/phi-4	14B	16,384 tokens	RAG / one-shot
unsloth/mistral-7b	7B	32,768 tokens	RAG / one-shot
google/gemma-7b	7B	8,192 tokens	RAG / one-shot
Qwen/Qwen3-8B	8B	32,768 tokens	RAG / one-shot
openai/gpt-oss-20b	20B	131,072 tokens	RAG / one-shot
google/flan-t5-large	780M	512 tokens	RAG / one-shot
Fralet/DDeduPModelv7	8B	8,192 tokens	RAG / zero-shot

Source: developed by the authors.

To measure quality, we introduce new metrics such as Information Groundedness Score (IGS), Generation Coherence Coefficient (GCC), and Text Fluency Metric (TFM). They compute factuality, logical consistency, and linguistic quality. This triad is not arbitrary, as it was chosen with the "content-structure-form" model.

From an instructional design perspective, Merrill's First Principles of Instruction establish that effective educational content must be organized around accurate problem-centered knowledge (content), presented in a progressive and integrated sequence (structure), and delivered in a form accessible to the learner (form) [27]. Mayer's Cognitive Theory of Multimedia Learning similarly argues that instructional material fails when it is factually unreliable, structurally incoherent, or linguistically inaccessible [28]. These three failure modes map directly onto our three dimensions.

The NLP has converged in the same dimensions. Fabbri et al. introduced SummEval, which is the most widely adopted benchmark for summarization evaluation. It identifies consistency (faithfulness), coherence, fluency, and relevance for assessing generated text quality [29]. It is a direct parallel to the EduDesignEval triad. Liu *et al*. [30] further validated this structure in G-Eval, demonstrating that GPT-4-based evaluation of coherence, consistency, and fluency aligns more closely with human judgment than single-metric alternatives. EduDesignEval targets the minimum necessary conditions for educational content to be deployable without causing harm.

This IGS metric measures how much the generated text is based solely on the provided context and does not contain unconfirmed statements or hallucinations. It evaluates whether generated content has a direct basis in the sources.

$$G_{\mathrm{TG}} = \frac{\text{GPT-4}_{\mathrm{Rating}}(context, expected\_facts, output)}{10} \quad (1)$$

where

- $\text{GPT-4}_{\mathrm{Rating}}$ —the automatic G-Eval score (1–10) generated by GPT-4.
- context—supporting information and instruction.
- expected_facts—list of factual elements the output should contain.
- output—generated texts.

Higher values (≥0.7) are considered acceptable for industrial use in educational design, indicating stronger

performance and weaker hallucination. Values between 0.5 and 0.7 are acceptable, and those less than 0.5 are prone to hallucinations. IGS is crucial because incorrect or hallucinated information can distort curricula. It is a major concern in AI-generated content, and IGS directly combats hallucinations. Using GPT-4 evaluation ensures consistent, high-level factual validation across all models.

While in multi-label classification, it checks whether labels are covered in label arrays. It is a percentage of the overall number of labels that are covered in the source dataset. It was chosen so that the models don't overproduce unseen or irrelevant labels and always provide reliable knowledge transmission. The higher the value (≥0.7), the better the alignment, while lower values (<0.4) suggest poorer coverage and potential overfitting.

$$G_{\text{MLC}} = \frac{|L_{\text{pred}}|}{|L_{\text{overall}}|} \quad (2)$$

where

- L_{pred}—overall number of predicted labels that appear in the source dataset.
- L_{overall}—overall number of predicted labels.

In text generation, GCC evaluates the logical flow and thematic integrity of generated text. It measures whether sentences connect smoothly and remain aligned with the input prompt.

$$C_{\text{TG}} = \frac{1}{3}(\text{BERT}_{\text{F1}} + \text{ROUGE}_{\text{L}} + \text{Overlap}) \quad (3)$$

where

- BERT_{F1}—average BERTScore F1 between consecutive sentences.
- ROUGE_{L}—ROUGE-L F-measure between the prompt and generated text.
- $\text{Overlap} = \frac{|W_p \cap W_g|}{|W_p \cup W_g|}$—vocabulary overlap ratio between prompt words W_p and generated words W_g.

Higher scores (≥0.4) indicate a highly connected and logically structured text, while lower scores (<0.3) indicate fragmented and off-topic values.

The reason for choosing is that the calculation avoids the weaknesses of each individual one (BERTScore—semantics, ROUGE-L—n-gram intersection, Jaccard—thematic proximity), giving a more stable assessment than any single metric.

While in classification, GCC assesses how logically compatible a predicted set of labels is with each other and with the input data. It measures how well the output corresponds to expected educational standards and professional structures.

$$C_{\text{MLC}} = \frac{1}{2}(\text{BERT}_{\text{match}} + \text{ExactMatch}) \quad (4)$$

- $\text{BERT}_{\text{match}}$—a semantic similarity between predicted and true label sets.
- ExactMatch—a percentage of exactly matched labels with ground truth.

Higher scores (≥0.4) indicate that the predicted labels are structurally and semantically appropriate, while lower scores (<0.3) show the opposite. The reason for choosing exact match is too strict for synonymous formulations in professional standards of Kazakhstan (Russian/Kazakh languages, different formulations of the same competence), so we added a weighted exact match on BERT embeddings, which significantly increased the correlation with expert assessment.

In text generation, TFM measures the correctness of the output format and the absence of excessive duplication or unnecessary labels. It evaluates how naturally and consistently the text flows according to linguistic norms. Higher fluency (≥0.6) makes text more natural, while lower fluency (<0.4) makes it repetitive or mechanical.

$$F_{\text{TG}} = \frac{1}{3}(\text{Perplexity} + \text{Consistency} + \text{Repetition}) \quad (5)$$

where

- $\text{Perplexity} = exp!(-\mathcal{L}/10)$ is the normalized inverse perplexity score, where $\mathcal{L}$ denotes the average token-level cross-entropy loss produced by the XGLM language model.
- $\text{Consistency} = 1 - \frac{\text{Var}(L)}{100}$ —represents the variance-based sentence length consistency.
- $\text{Repetition} = \frac{|unique\_words|}{|total\_words|}$ denotes the repetition penalty.

The reason for choosing Perplexity is the performance for anticipating the token sequence and syntactic regularity, the stability of sentence length reflects the rhythm of professional texts, and the uniqueness of words punishes excessive repetition.

TFM in classification evaluates the structural cleanliness and non-redundancy of predicted label sets. It checks whether the model outputs well-formatted, non-repetitive, and valid educational labels.

$$F_{\text{MLC}} = \frac{1}{2}(\text{Structure} + \text{NoRepetition}) \quad (6)$$

where

- Structure—binary indicator assessing whether labels follow the required formatting.
- $\text{NoRepetition} = \frac{|unique_{labels}|}{|total_{labels}|}$—repetition penalty for label predictions.

Lower scores (<0.05) indicate that the predicted label set is clean, non-redundant, and structurally valid, while higher values (≥0.15) are prone to duplications and malformed outputs. Any extra text or duplicates break the pipeline, so we focused on structural purity and the absence of duplication, which is especially important for zero-shot/one-shot modes.

Unified Evaluation Score (UES) measures the overall quality of a model by aggregating IGS, GSS, and TFM into a single scalar. It captures a balanced view of factuality, logical consistency, and linguistic quality across either text generation or multi-label classification tasks. It is not presented as an absolute indicator of pedagogical quality.

$$\text{UES} = \alpha\,\text{IGS} + \beta\,\text{GSS} + \gamma\,\text{TFM} \quad (7)$$

subject to

$$\alpha + \beta + \gamma = 1, \alpha, \beta, \gamma \in [0,1] \quad (8)$$

where

α, β, γ —user-selected weights controlling the relative importance.

A higher UES indicates stronger overall model performance, reflecting better factual grounding, coherence, and fluency. If it is less than 0.3, then it is not suitable for educational use. UES from 0.3 to 0.5 means acceptable but with human review, while a value higher than 0.5 means excellent overall quality.

Combining measurements into a single score disentangles the comparison and determination when designers must select among numerous LLMs and deduction procedures. The weighted form preserves adaptability: analysts and professionals can emphasize factuality (increase α), coherent structure (increase β), or readability (increase γ) depending on application needs.

After UES was aggregated, for each model, the mean, standard deviation, and Standard Error of the Mean (SEM) were chosen to calculate uncertainty for each model. 95% confidence intervals were calculated using the Student's t-distribution, and a comparison was conducted using Welch's T-test. Effect sizes were quantified using Cohen's d to measure the practical magnitude of differences.

The primary contribution of this study is the design and structural validation of formulated metrics and their aggregation using UES. The comparison between LLM performances served as an applied illustration of how the proposed evaluation system behaves across diverse architectures. Additionally, it demonstrated how sensitive the system is to metric trade-offs.

IV. Result

Ten LLMs were evaluated on text generation and multi-label classification tasks, revealing subtle superiority across EduDesignEval measures (Table 2). The models' three main aspects of groundedness, coherence, and fluency were considered. Each of these dimensions consists of several component metrics that collectively offer a thorough understanding of the models' performance in the creation of instructional content.

Table 2. Performance metrics of selected models

Model	UES (TG / RAG)	IGS (TG / RAG)	GCC (TG / RAG)	TFM (TG / RAG)
DeepSeek	0.5240 / 0.6938	0.7885 / 0.7283	0.3595 / 0.4148	0.4243 / 0.0616
DDeduPModelv7	**0.5997** / 0.6918	0.7615 / **0.7837**	0.3769 / 0.4142	0.6608 / 0.1224
Gemma	0.5184 / 0.6120	0.7330 / 0.4546	0.3626 / 0.4151	0.4599 / 0.0335
gpt-oss	0.4557 / 0.5535	0.6460 / 0.5141	0.3548 / 0.3272	0.3663 / 0.1808
LFM	0.5738 / 0.5514	0.7560 / 0.2489	0.3811 / 0.4119	0.5843 / **0.0064**
Llama	0.5918 / 0.5944	0.6815 / 0.3984	0.4052 / 0.4108	0.6887 / 0.0258
Mistral	0.5491 / 0.5784	0.6255 / 0.3519	**0.4162** / 0.4135	0.6057 / 0.0301
Phi	0.5146 / 0.6986	0.7030 / 0.7055	0.2101 / **0.4194**	0.6308 / 0.0288
Qwen	0.5370 / **0.7065**	**0.8675** / 0.7242	0.3311 / 0.4162	0.4125 / 0.0206
Flan-T5	0.5984 / 0.5273	0.6235 / 0.2000	0.3656 / 0.4110	**0.8062** / 0.0289

Source: done via https://huggingface.co/Fralet/DDeduPModelv7 and https://huggingface.co/datasets/Fralet/DDeduPDatasetv2 by the authors.

The text generation results (Fig. 3) show that Qwen leads in groundedness with a score of 0.868, followed by DeepSeek (0.789) and DDeduPModelv7 (0.762). This information demonstrates that the zero-shot methods may lag in a few complex assignments, but this can be moderated through fine-tuning. Models like LFM and Gemma perform more ineffectively, scoring 0.756 and 0.733 points, respectively. Other models, such as Mistral (0.626) and FlanT5 (0.624), show variability in addressing training accuracy. An examination of the variance models uncovers critical contrasts in model reliability. The basis variance of DDeduPModelv7 is in the range of 2–5% (0.263), whereas Mistral is around 4% (0.293) and gpt-oss is at 0.300. These models perform very consistently in a few zones but suddenly fall flat in others. In educational contexts that require disciplinary diversity and high-quality standards, reliability assessment, namely, the extent to which a model performs consistently across a specific domain, may be more imperative than small performance contrasts.

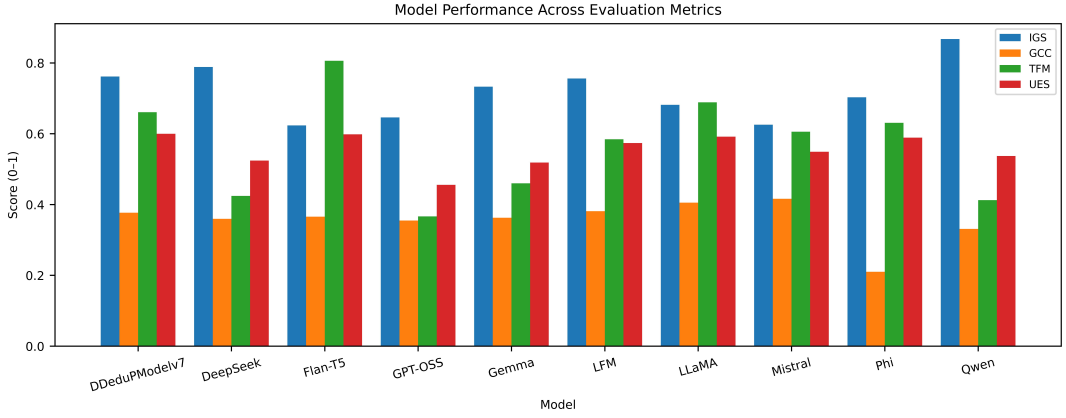

Fig. 3. Model performance across evaluation metrics in text generation.

Component metric analysis of fluency uncovers unpretentious quality markers. FlanT5's word repetition score of 0.874 is the highest among all models, suggesting that the model demonstrates regularly reused expressions inside single responses. It could be a pattern that would produce monotonous, repetitive educational content or model

hallucinations. On the other hand, DDeduPModelv7's word repetition of 0.598 and sentence length variance of 0.401 show more natural linguistic diversity. The perplexity component, where lower normalized values demonstrate superior familiarity, shows DDeduPModelv7 at 0.765 compared to Qwen's 0.833, recommending DDeduPModelv7 produces more normally streaming content even if slightly less factually comprehensive.

When it comes to text generation, Mistral is the most coherent (0.416), followed by Llama (0.405) and LFM (0.381). DDeduPModelv7 is positioned in the upper-middle tier (mean of 0.377 with 0.825 BERTScore performance). DeepSeek (0.359) and Gemma (0.362), which maintain a reasonable cohesive structure but fall short of expectations in semantic layout when compared to the top models, are examples of moderate performers. Although occasionally less tightly matched with reference text, their BERTScore and ROUGE-L results confirm that outputs are often coherent. Phi, with a mean of 0.210 and a remarkably high standard deviation (0.194), and gpt-oss, with a mean of 0.354 but a greater variance (0.046), exhibit worse coherence. These models sometimes deliver divided or incoherent outputs, particularly in multi-sentence or multi-paragraph responses. FlanT5 sits near the middle (0.365), coherent on average but with a risk of inconsistencies in longer outputs. These failures likely stem from hallucinations or formatting errors that completely detached outputs from source material, highlighting dangers in deploying models without vigorous error handling.

FlanT5, with a score of 0.806, ranks with the highest fluency, receiving low perplexity and varied sentencing. Llama (0.689) and DDeduPModelv7 (0.731) show similar performances. Indeed, models with a somewhat lower level of grounding, such as Flan-T5, can deliver astonishingly natural prose, though at the cost of intermittent factual blunders. So far, LFM shows good fluency (0.584) while maintaining good coherence and groundedness, making it a balanced model. Gemma and DeepSeek offer moderate readability but can be marginally repetitive in style. Lower fluency in gpt-oss (0.366) and Phi (0.631) led to broken sentences or uneven word choice. Interestingly, some models with very high groundedness, such as Qwen, have lower fluency (0.412), indicating a trade-off of prioritizing actual accuracy that can sometimes be attributed to stylistic smoothness. Breaking down the components revealed DDeduPModelv7 normalized perplexity of 0.765 compared to Qwen's 0.833, suggesting reasonably natural language flow and Qwen's awkward phrasing. The sentence length variance of 0.401 indicates appropriate structural diversity, while word repetition of 0.598 shows some tendency toward repeated terminology, likely reflecting the specialized lexicon required (Fig. 4).

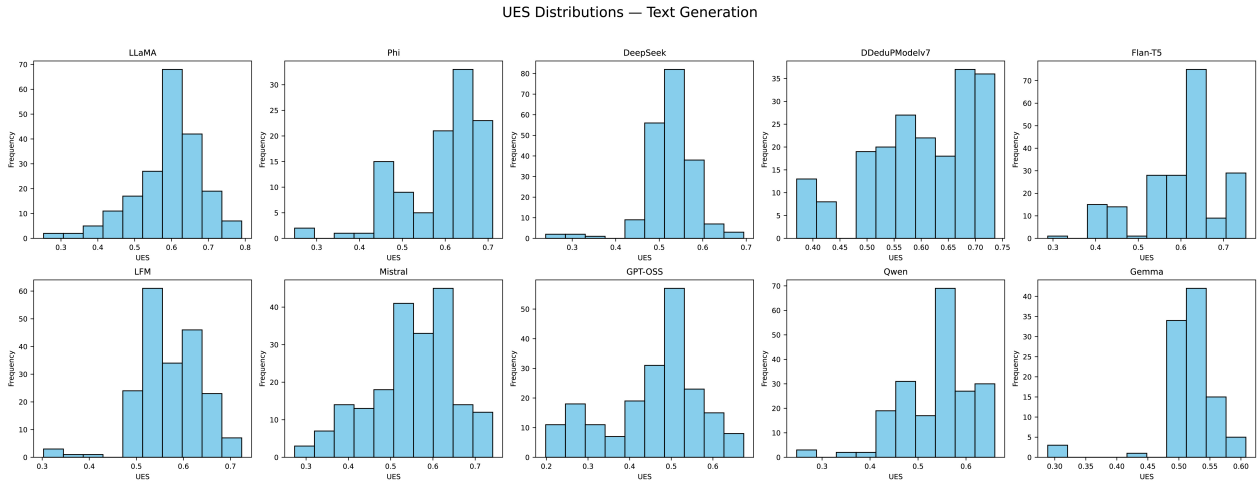

Fig. 4. UES distributions in text generation.

For multi-label classification, including RAG techniques for labeling educational standards and professions, DDeduPModelv7 accomplished the most noteworthy groundedness at 0.784 with zero-shot conditions, outperforming DeepSeek (0.728) and Qwen (0.724), which benefited from one-shot prompting. Phi at 0.705 is next, whereas FlanT5 (0.200) and LFM (0.249) battled catastrophically with label accuracy (Fig. 5).

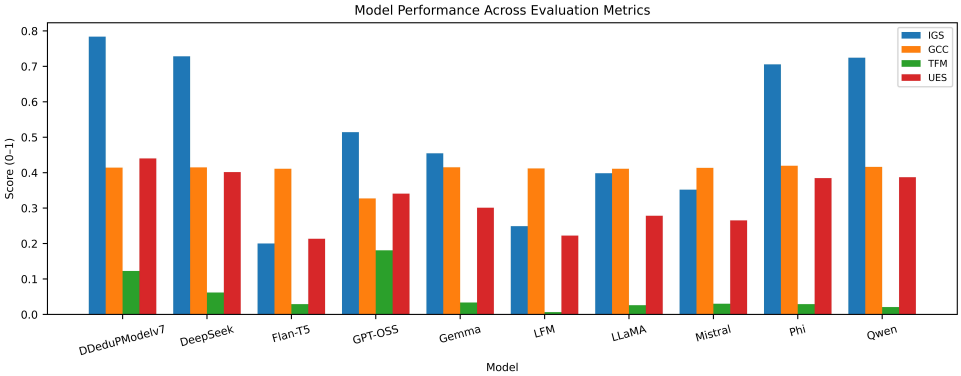

Fig. 5. Model performance across evaluation metrics in multi-label classification.

Phi took the lead with an average coherence of 0.419, followed by Qwen (0.416), DDeduPModelv7 (0.414), and Mistral (0.414). Component metrics, particularly BERTScore, showed that most models had strong semantic understanding, with values ranging between 0.814 and 0.827. However, the exact match showed that the best value the models could give was 0.012, with others having zero as well. This methodology suggests that, although models are effective at capturing semantic meaning, they rarely anticipate sets of names with pinpoint accuracy, suggesting semantically comparable alternatives that can still serve educational purposes.

Fluency here is recognized by average label repetition and the unreadable/empty proportion; lower scores indicate superior execution by minimizing issues. The LFM score is 0.007, followed by Qwen (0.021) and Llama (0.026), whereas DDeduPModelv7 (0.123) shows a higher number of reiterations (0.213) but a low unparsable rate (0.032)—but it is superior to gpt-oss (0.181). FlanT5 has a 10 times worse unparsable/empty rate than competitors. It is imperative to note that DDeduPModelv7 has kept up a low unparsable ratio (0.032), which demonstrates that, despite intermittent repetition, the outputs remained structurally valid. In contrast, LFM achieved the best fluency by minimizing redundancy but sacrificed groundedness, effectively prioritizing format over content (Fig. 6).

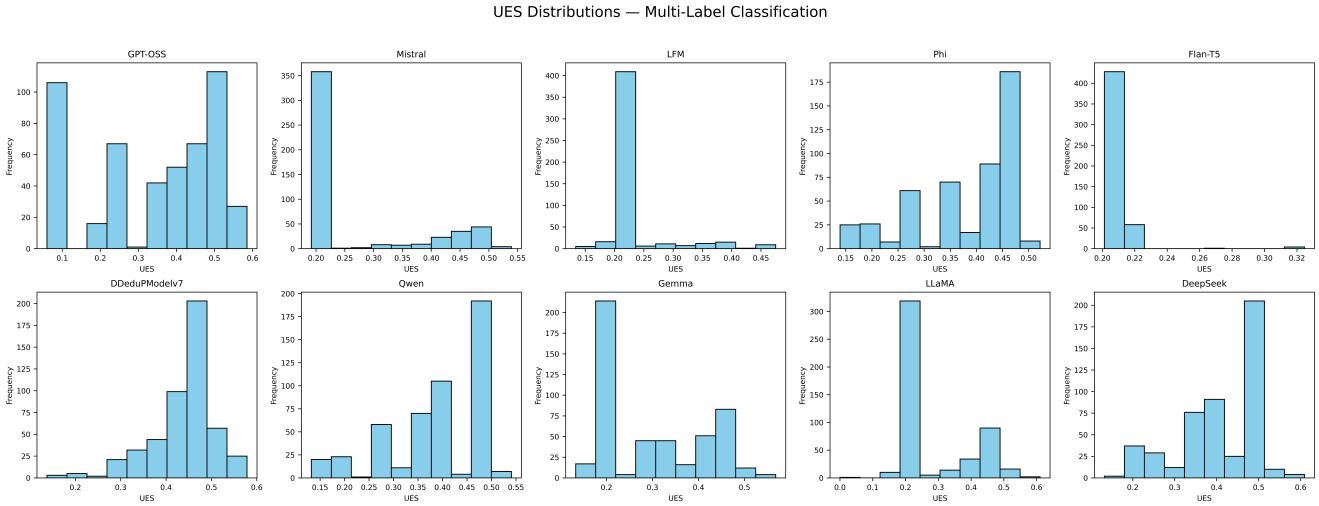

Fig. 6. UES distributions in multi-label classification.

DDeduPModelv7's potential exceeds expectations in classification assignments where comprehensive conceptual scope is pivotal, especially in ranges such as organizing instructive content, making taxonomies, and labeling learning assets. Its zero-shot generation capability makes it perfect for different instructive settings where making high-quality samples for each theme is impractical. Qwen performs best in content generation tasks requiring the most extreme genuine precision and unwavering quality, such as referring materials or evaluating content. It shows that organizations need to invest in effective one-shot solutions. DeepSeek offers adjusted execution in both tasks with its one-shot generation capability, which is reasonable for complex workflows. LFM and Phi ought to be kept at a strategic distance, as they are insufficient for essential instructive text generation. Phi is considered appropriate for specialized applications where great formatting is required rather than content integrity.

Strong text generation (0.868 and 0.789) and classification (0.724 and 0.728) capabilities of the Qwen and DeepSeek models, designating them as generalist models with one-shot advice that can manage a variety of educational procedures. In a comparable manner, DDeduPModelv7 demonstrates cross-task competence (0.784 classification, 0.762 generation), accomplishing this in a novel way without the need for examples.

On the other hand, some models have a limited specialization. The modest capacity to create text (0.624), which suggests applicability solely for generation tasks, contrasts with the catastrophic efficiency of the FlanT5 classification (0.200 groundedness). Limitations associated with a particular task are also indicated by difficulties with categorization in LFM (0.249) and enhanced text generation (0.756). According to these specialization models, educational institutions that need a variety of content-generating functions should either select proven generalists like DeepSeek or Qwen or use many specialized models (DDeduPModelv7).

Table 3. Statistical validation of model performance in multi-label classification

Model	Mean	Std	SEM	95% CI
DDeduPModelv7	0.440	0.071	0.003	[0.434, 0.446]
DeepSeek	0.402	0.090	0.004	[0.394, 0.409]
Qwen	0.387	0.096	0.004	[0.378, 0.395]
Phi	0.385	0.100	0.004	[0.376, 0.393]
GPT-OSS	0.341	0.177	0.008	[0.325, 0.356]
Gemma	0.301	0.113	0.005	[0.291, 0.311]
LLaMA	0.278	0.118	0.005	[0.268, 0.289]
Mistral	0.265	0.107	0.005	[0.256, 0.274]
LFM	0.222	0.056	0.002	[0.217, 0.227]
Flan-T5	0.213	0.010	0.000	[0.212, 0.214]

To address validation in Table 3, in MLC DDeduPModelv7 was seen as the top performer with a mean UES of 0.440, followed by DeepSeek. A Welch's T-test against the weakest performer, Flan-T5, produced $t = 70.19$ and $p = 0$. It confirmed the superiority of DDeduPModelv7 in MLC. Confidence intervals are non-overlapping with the fine-tuned model. It suggested a genuine performance hierarchy. Standard deviation ranged from 0.01 to 0.177. It confirms stable performance across samples except for GPT-OSS with its unreliable behavior. SEM are mostly small values, so averages are stable. Cohen's d is 4.48, dramatically

better for classification.

In Table 4, Welch's T-test showed $t = 0.136$ and $p = 0.892$, while Cohen's $d = 0.01$ in TG tasks. Also, DDeduPModelv7 was the top performer in TG tasks, which is comparable with Flan-T5. Yet results show no statistically significant difference, confirming that differences are not substantively meaningful and revealing that variances were larger. A performance plateau is revealed by overlapping intervals. SEM is small as well.

Table 4. Statistical validation of model performance in text generation

Model	Mean	Std	SEM	95% CI
DDeduPModelv7	0.600	0.099	0.007	[0.586, 0.614]
Flan-T5	0.598	0.093	0.006	[0.585, 0.611]
LLaMA	0.592	0.089	0.006	[0.579, 0.604]
Phi	0.589	0.095	0.009	[0.571, 0.607]
LFM	0.574	0.066	0.005	[0.565, 0.583]
Mistral	0.549	0.097	0.007	[0.536, 0.563]
Qwen	0.537	0.073	0.005	[0.527, 0.547]
DeepSeek	0.524	0.058	0.004	[0.516, 0.532]
Gemma	0.518	0.049	0.005	[0.509, 0.528]
GPT-OSS	0.4557	0.1130	0.0080	[0.4399, 0.4715]

V. Discussion

The evaluation's findings highlight fundamental clashes in LLM design for instructive uses. The negative connection between groundedness and fluency appeared in a few models focusing on an architectural trade-off: models that prioritize readability have a higher chance of hallucinations, while those that are optimized for factuality may compromise phonetic naturalness. Yet models that hallucinate in classification don't mean they hallucinate in text generation. This is a task-structure dependency. This phenomenon has critical repercussions for the creation of guidelines, since both viewpoints are important. A course must be truthful, but using false data is just as much a failure in its educational objective as engaging in misinformation.

DDeduPModelv7 has strong semantic alignment and lexically diverse output content, which is ideal for educational content. While its lower fluency shows its controlled generation, which is not a weakness. For example, Qwen and DeepSeek have high groundedness, but they produce less pedagogical and less fluent text. So high groundedness doesn't directly mean good educational explanation. Groundedness in classification is strict and unforgiving. It measures constraint adherence, penalizing over-prediction and speculative labels. A prime example is FlanT5, which shows excellent fluency but collapses at 0.2 groundedness.

Fluency in classification is range compressed. All models range below 0.1, except gpt-oss and DDeduPModelv7. It captures formatting and repetition behavior, contributing to explanatory power in a structural sense. Fluency is orthogonal to correctness. Fluency in text generation strongly separates models in a linguistic sense. It can be seen because of sentence fragmentation, over-compression, and repetitive phrasing. That's why FlanT5 rises even with moderate grounding, gpt-oss is at the bottom with low fluency and low grounding, and DDeduPModelv7 is strong, but not dominant. Fluency here serves as a proxy for pedagogical usability.

It is comparable to more common LLM evaluations in education, where educational measurements past non-specific criteria are underscored by systems such as ELMES. Comparable to EduDesignEval's groundedness-fluency pressures, these designs adjust with systems like ELMES, which evaluates LLMs in instructive settings utilizing measurements for clarification technique and guided instruction, distinguishing trade-offs between information authority and personalization [31]. Also, EduDesignEval does this by employing classification tasks. This feature is complemented by EducationQ's emphasis on educating capacities through multi-agent exchanges [32]. Both ELMES and EducationQ capture important pedagogical aspects, but neither provides a unified quantitative score to benchmark.

These findings position Qwen and DeepSeek as factual specialists with high groundedness and moderate-low coherence and fluency. Llama, Mistral, and DDeduPModelv7 are balanced in each of the three metrics and could be used in general-purpose applications. Phi is excellent in classification, while FlanT5 is creative with exceptional fluency. Performance swings suggest architectural biases cannot be overcome by prompting alone. Models like Phi and FlanT5 have fundamental structural preferences that dominate their behavior.

This trade-off is observed to be negated by the DDeduPModelv7 performance. The model appears to show that fine-tuning on domain-specific instructive information may adjust contradicting advancement points by getting competitive text generation groundedness while protecting conventional fluency and overwhelming classification tests without one-shot prompting. Every metric is not extreme, but still at the high middle tier with no pathological weaknesses. The same could be said in classification with strong groundedness, except for slight fluency due to label repetition.

But the pattern revealed high factuality but less fluent output. This observation echoed the factual strength of AI-assisted competency mapping documented by Padovano and Cardamone [9]. High fluency means readability, but speculative labeling in classification. This is comparable with more common patterns in LLM, where studies illustrating unfaltering enhancements in content classification precision illustrate that refined, smaller models often beat greater zero-shot generative ones in particular assignments. Also, by expelling content from each prompt, the zero-shot operation brings down idleness and token costs, which are important factors on a scale. These come about to suggest that fine-tuning moves forward significance in guideline settings by lessening generalist limits. Within these constraints, DDeduPModelv7 and the base LLaMA-3-8B architecture offer the most favorable cost-to-performance ratio. Their 8-billion parameter footprint requires approximately 6–8GB of VRAM at inference time, making deployment feasible on a single mid-range GPU or a modest cloud instance.

In text generation, UES showed the highest positive correlation with fluency, while in classification, it was with groundedness. In generation, IGS had a weak negative correlation with TFM. It mirrored the results that a model optimized for factual adherence produces less natural text from Molenaar [4]. Coherence and fluency show moderate positive relationships. In RAG tasks, fluency showed a very weak correlation since lower values are better here. Coherence is relatively stable.

The 20-billion parameter gpt-oss model approaches the

memory limits of this configuration and would exceed the capacity of most GPUs. In practice, this means that for institutions operating without dedicated GPU servers or cloud computing budgets, the effective choice set is constrained to models at or below approximately 8–14 billion. This hardware ceiling, rather than licensing or data access, is the primary adoption barrier for most models in real educational deployments.

Model selection for educational deployment should not be based on peak benchmark performance alone. Institutions with constrained hardware should prioritize fine-tuned compact open-source models. Institutions with cloud infrastructure and higher throughput requirements may justify the computational premium of larger open-source generalists like Qwen or DeepSeek.

In practice, consider this use case. A curriculum designer is tasked with updating the learning goals for a new interdisciplinary program in Data Science. The designer begins by authoring a structured prompt specifying the program's level, disciplinary scope, and target competencies. In the case of Kazakhstan, it is drawn from the national professional standards database. After submitting to the AI model, it generates five candidate outputs and computes IGS, GCC, TFM, and UES scores for each output. The designer reviews the flagged outputs, refines the prompt to more explicitly reference the relevant professional standards, and submits a revised query. Final prompt, outputs, and scores are logged for institutional quality assurance and forwarded for peer review.

This workflow illustrates three key principles of EduDesignEval's framework. First, the framework functions as a diagnostic filter rather than a gatekeeper. Scores do not automatically accept or reject outputs but direct the designer's attention to specific quality metrics for revision. Second, the iterative loop preserves human authorship and judgment at every decision point. AI generates candidates, and the framework measures them, but the designer controls final selection. Third, the scoring system provides auditable quality evidence that can satisfy institutional accreditation requirements.

Therefore, the proposed evaluation framework, particularly its unified scoring metric, may be interpreted in two distinct ways. Text generation unified score means "How well does the model balance factual correctness, semantic alignment, and readable explanation?" whereas classification means "How well does the model avoid label hallucination while remaining semantically aligned?" So, comparisons between each metric of text generation and classification are not needed. This is subtle but critical. It avoided hallucination, respected document boundaries, revealed semantic values, and didn't overoptimize fluency. It is too conservative by nature, which is ideal for education.

Nevertheless, limitations incorporate conceivable biases, hallucinations, and the need for human observation, as shown in less coherent models yielding divided results. For example: reliance on GPT-4 as a judge for IGS, dataset bias from Kazakhstani-centric data, and domain constraints. The dataset was constructed by scraping publicly available digital educational programs and introduces a selection bias toward more digitally established institutions. Researchers should be aware of whether their target institutions share comparable levels of curricular documentation maturity when applying EduDesignEval. Also, it reflects the multilingual reality of Kazakhstani higher institutions. It may be strengthening the framework's relevance for Central Asian educational systems. It limits the direct transferability of benchmark thresholds. When applied where educational priorities differ (inquiry-based learning in the US or exam-oriented approaches in East Asia), EduDesignEval could not perform as well as in the Kazakhstani context. These constraints do not invalidate the framework but define its scope.

In keeping with concerns for responsible AI in education, integrating LLMs necessitates ethical issues, including promoting AI literacy and guaranteeing fair access. Despite being careful, this assessment framework distorts the complex issue of guidelines for material quality. Although component measurements capture critical viewpoints, they are incapable of precisely delineating educational viability, social relevance, openness, or engagement potential. Future studies should cover expert reviews for pedagogical usefulness, A/B testing of AI-generated and human-generated educational program validation, and the incorporation of EU and US institutions to test cross-cultural transfer. Do materials produced by various approaches genuinely improve students' learning outcomes? Do educators favor using the results of models?

VI. Conclusion

In conclusion, the EduDesignEval framework marks a pivotal step toward developing standardized evaluation criteria for LLM-generated educational content. Our proposed metrics, including IGS for factuality, GCC for logical integrity, TFM for linguistic naturalness, and the aggregated UES, have shown initial promise that transcends fragmented approaches like prompt engineering or isolated NLP metrics (e.g., BLEU or ROUGE). By addressing RQ1, we demonstrated that LLMs can produce educationally valid content. These tools not only standardize quality control but also facilitate objective comparisons across models and methods, addressing gaps in existing literature where pedagogical alignment and cognitive suitability were often overlooked.

The practical implications are substantial, yet initial results are preliminary. In relation to RQ2, models differ across tasks based on specific requirements: Qwen for reference materials demanding maximum accuracy, DDeduPModelv7 for diverse classification tasks requiring comprehensive conceptual coverage, and DeepSeek for balanced workflows. EduDesignEval establishes a foundational methodology for standardizing LLM assessment in education. As AI integration accelerates, systematic evaluation frameworks become essential infrastructure. Not only for technical validation, but also for ensuring educational quality, equity, and responsible AI deployment.

While EduDesignEval underscores automated metrics, it is designed to support human-AI collaboration rather than replace human expertise. It operates through iterative feedback loops. AI handles initial tasks, produces output, and calculates scores. Designers, on the other hand, design prompts, then interpret calculated scores and validate. This work contributes to the broader imperative of developing AI-enhanced education that serves learners effectively while

maintaining pedagogical integrity and ethical standards. Future extensions would incorporate human-centered validation with educators and curriculum designers to assess whether metrics reflect improvements in teaching effectiveness.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Conceptualization, A.S. and A.B.; methodology, A.S. and A.B.; software, A.S.; validation, A.S., A.M. and B.Z.; formal analysis, A.M. and B.Z.; investigation, A.M. and B.Z.; resources, A.M. and B.Z.; data curation, A.S.; writing—original draft preparation, A.S.; writing—review and editing, A.M. and B.Z.; visualization, A.S.; supervision, A.B.; project administration, A.M. All authors have read and agreed to the published version of the manuscript.

REFERENCES

[1] T. Brown, B. Mann, N. Ryder *et al*., "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1877–1901, 2020.

[2] J. Oudenampsen, M. Van De Pol, N. Blijlevens, and E. Das, "Interdisciplinary education affects student learning: A focus group study," *BMC Medical Education*, vol. 23, no. 1, 169, 2023. doi: 10.1186/s12909-023-04103-9

[3] J. Cohen, A. Boguslav, J. Wyckoff, V. Katz, K. Sadowski, and E. A. Wiseman, "Core requirements, structured flexibility, and local judgment: Balancing adherence and adaptation in the design and implementation of district-wide professional development," *Educational Evaluation and Policy Analysis*, vol. 47, no. 1, pp. 263–291, 2023. doi: 10.3102/01623737231210285

[4] I. Molenaar, "Towards hybrid human-AI learning technologies," *European Journal of Education*, vol. 57, pp. 632–645, 2022. doi: 10.1111/ejed.12527

[5] J. Li, Y. Yan, and X. Zeng, "Exploring artificial intelligence in inclusive education: A systematic review of empirical studies," *Applied Sciences*, vol. 15, no. 23, 2025. doi: 10.3390/app152312624

[6] S. B. Sukhavasi and S. B. Sukhavasi, "Exploring AI in education: A review of its impact on classrooms, learning management, and pedagogical strategies," presented at 2025 Northeast Section Conference, University of Bridgeport, Bridgeport, CT, Mar. 2025. doi: 10.18260/1-2-1115-55003

[7] J. Leung and Z. Shen, "Prompt engineering for curriculum design," in *Proc. 2024 4th International Conference on Educational Technology (ICET)*, Wuhan, China, 2024, pp. 97–101. doi: 10.1109/ICET62460.2024.10869091

[8] L. McNeill, M. M. Uddin, M. Pei, and L. Regalado, "Generative AI in instructional design: Adoption, benefits, and best practices," *Journal of Applied Instructional Design*, vol. 14, no. 3, pp. 108–135, 2024. doi: 10.59668/2223.22720

[9] A. Padovano and M. Cardamone, "Towards human-AI collaboration in the competency-based curriculum development process: The case of industrial engineering and management education," *Computers and Education: Artificial Intelligence*, vol. 7, 2024. doi: 10.1016/j.caeai.2024.100256

[10] E. L. Brown, D. A. Talbert, and J. Roberts, "Toward the use of LLMs to support curriculum mapping to established frameworks," presented at 2025 ASEE Annual Conference & Exposition, Montreal, Quebec, Canada, Jun. 2025. doi: 10.18260/1-2--57295

[11] M. M. Rahman and Y. Watanobe, "ChatGPT for education and research: Opportunities, threats, and strategies," *Applied Sciences*, vol. 13, no. 9, 2023. doi: 10.3390/app13095783

[12] H. Saiedian, "Leveraging large language models in education: Enhancing learning and teaching," presented at 2023 ASEE Midwest Section Conference, University of Nebraska-Lincoln, Lincoln, Nebraska, Jul. 2024. doi: 10.18260/1-2-119-46353

[13] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187–192, 2023. doi: 10.1126/science.adh2586

[14] J. He-Yueya, N. D. Goodman, and E. Brunskill, "Evaluating and optimizing educational content with large language model judgments," in *Proc. 17th International Conference on Educational Data Mining*, 2024, pp. 68–82. doi: 10.5281/zenodo.12729776

[15] I. A. Solaiman, T. S. Maria, and M. Milanova, "Educational content generation using multi-LLM agents," *International Robotics & Automation Journal*, vol. 10, no. 3, pp. 85–87, 2024. doi: 10.15406/iratj.2024.10.00288

[16] Y. Gong, C. Xu, S. Luo *et al.*, "Modeling teacher education students' adoption of large language models through an extended technology acceptance framework," *Scientific Reports*, vol. 15, 32208, 2025. doi: 10.1038/s41598-025-03298-9

[17] Y. Chen, Y. Li, Y. Ren, Y. Liu, and Y. Ma, "Educational evaluation with MLLMs: Framework, dataset, and comprehensive assessment," *Electronics*, vol. 14, no. 18, 3713, 2025. doi: 10.3390/electronics14183713

[18] G. G. Lee, E. Latif, X. Wu, N. Liu, and X. Zhai, "Applying large language models and chain-of-thought for automatic scoring," *Computers and Education: Artificial Intelligence*, vol. 6, 2024. doi: 10.1016/j.caeai.2024.100213

[19] Q. Huang, C. Lv, L. Lu, and S. Tu, "Evaluating the quality of AI-generated digital educational resources for university teaching and learning," *Systems*, vol. 13, no. 3, 2025. doi: 10.3390/systems13030174

[20] H. Khosravi, S. B. Shum, G. Chen *et al*., "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, 100074, 2022. doi: 10.1016/j.caeai.2022.100074

[21] K. Fuchs, "Exploring the opportunities and challenges of NLP models in higher education: is Chat GPT a blessing or a curse?" *Frontiers in Education*, vol. 8, 2023. doi: 10.3389/feduc.2023.1166682

[22] A. Mukhatayev, S. Omirbayev, K. Kassenov, and Y. Idiyatova, "Quality assurance system of higher education in Kazakhstan through stakeholders' eyes: An empirical study to identify its challenges," *Education Sciences*, vol. 14, no. 12, 2024. doi: 10.3390/educsci14121297

[23] A. Biloshchytskyi, S. Omirbayev, A. Mukhatayev, O. Kuchanskyi, M. Hlebena, Y. Andrashko, N. Mussabayev, and A. Faizullin, "Structural models of forming an integrated information and educational system 'quality management of higher and postgraduate education'," *Frontiers in Education*, vol. 9, 2024. doi: 10.3389/feduc.2024.1291831

[24] J. L. Menéndez-Varela and E. Gregori-Giralt, "The reliability and sources of error of using rubrics-based assessment for student projects," *Assessment & Evaluation in Higher Education*, vol. 43, no. 3, pp. 488–499, 2018. doi: 10.1080/02602938.2017.1360838

[25] O. Bulut, M. Beiting-Parrish, J. M. Casabianca *et al*., "The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges," *Chinese/English Journal of Educational Measurement and Evaluation*, vol. 5, no. 3, 2024. doi: 10.59863/MIQL7785

[26] J. Lee, Y. Hicke, R. Yu, C. Brooks, and R. F. Kizilcec, "The life cycle of large language models in education: A framework for understanding sources of bias," *British Journal of Educational Technology*, vol. 55, pp. 1982–2002, 2024. doi: 10.1111/bjet.13505

[27] M. D. W. Merrill, "First principles of instruction," *Educational Technology Research and Development*, vol. 50, no. 3, pp. 43–59, 2002. doi: 10.1007/BF02505024

[28] R. E. Mayer, *Multimedia Learning*, 2nd ed. Cambridge University Press, 2009. doi: 10.1017/CBO9780511811678

[29] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, and D. Radev, "Summeval: Re-evaluating summarization evaluation," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 391–409, 2021. doi: 10.1162/tacl_a_00373

[30] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-eval: NLG evaluation using gpt-4 with better human alignment," in *Proc. 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023. doi: 10.18653/v1/2023.emnlp-main.153

[31] S. A. Wei, X. Wang, S. Bi *et al*., "ELMES: An automated framework for evaluating large language models in educational scenarios," arXiv, 2025. doi: 10.48550/arXiv.2507.22947

[32] Y. Shi, R. Liang, and Y. Xu, "EducationQ: Evaluating LLMs' teaching capabilities through multi-agent dialogue framework," in *Proc. ACL 2025*, vol. 1, pp. 32799–32828, Jul. 2025. doi: 10.18653/v1/2025.acl-long.1576