

<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper presented at *2025 17th International Conference on Education Technology and Computers, ICETC 2025, Barcelona, Sept 18-21, 2025*.

Citation for the original published paper:

Yavariabdi, A., Paudel, B., Carleton, T., Andrade de Almeida, C D. (2025)
Generative AI in Assessment and Feedback Generation in Higher Education: A
Systematic Review
In: *Proceedings of the 2025 17th International Conference on Education Technology
and Computers, ICETC 2025* (pp. 361-371). Institute of Electrical and Electronics
Engineers (IEEE)
<https://doi.org/10.1109/ICETC66579.2025.11387416>

N.B. When citing this work, cite the original published paper.

©2025 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:bth-29449>

Generative AI in Assessment and Feedback Generation in Higher Education: A Systematic Review

Amir Yavariabdi
Department of Computer Science
Blekinge Institute of Technology
Karlskrona, Sweden
amir.yavariabdi@bth.se

Bhuwan Paudel
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
bhuwan.paudel@bth.se

Tamara Carleton
Department of Mechanical Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
tamara.carleton@bth.se

Carlos Diego Andrade de Almeida
Department of Software Engineering
Blekinge Institute of Technology
Karlskrona, Sweden
carlosdiego.andrade.de.almeida@bth.se

Abstract—Assessment and feedback activities in higher education are undergoing significant changes. Many universities and institutes still rely on traditional testing and grading methods, which often fall short in supporting meaningful student learning, especially in large classes. Although educational policies, such as those promoted by the Bologna process, encourage more feedback-oriented and student-centered approaches, these practices can be difficult to implement due to time constraints and limited resources. Generative Artificial Intelligence (GenAI), particularly Large Language Models (LLMs), has shown strong potential in addressing these challenges. This review examines 21 research studies published between 2023 and 2025 that explore the use of GenAI in providing feedback and assessing student work in higher education, with some studies also comparing GenAI’s performance with human instructors. Findings show that LLMs can generate personalized and constructive feedback and/or assist with fair and consistent assessment. However, in most studies, teachers still play a key role, as expert oversight is essential to ensure that grading assessments and assignment feedback are accurate, relevant, and aligned with learning objectives. For GenAI to be used effectively, educators need to understand how to work with these tools, such as learning GenAI prompt design and the basic principles behind LLMs. We recommend that academic institutions provide training for educators in AI literacy, prompt engineering, and the development of teaching strategies that combine the strengths of human judgment with AI support. By effectively integrating LLM tools, major assessment challenges, such as limited time and inconsistent feedback quality, can be addressed while also enhancing student learning and engagement.

Index Terms—Large Language Model (LLMs), Generative Artificial Intelligence (GenAI), Automatic Feedback, Automatic Scoring, Comprehensive Study

I. INTRODUCTION

The rapid growth of AI, particularly LLMs, has begun to transform the educational landscape, especially in the domains of assessment and feedback generation [1]. Assessment

and feedback are different, yet complementary, activities for educators. Assessment is the process of evaluating learning outcomes, often expressed in test scores and grades. Feedback is information given to students to improve their learning, often as short statements or verbal comments. These activities are clearly connected because giving feedback requires some level of assessment and judgment about the quality of student work, and assessment may result in some form of feedback being given or received. Providing high-quality, personalized feedback, as well as fair and consistent assessment, remains a significant challenge, particularly in large-scale university courses and for open-ended assignments. Although educational frameworks such as the Bologna process [2] emphasize the importance of student-centered and feedback-oriented practices, their implementation is often constrained by limited time and instructional resources [3], [4]. In practice, human feedback and assessment can suffer from inconsistencies due to time pressure, cognitive fatigue, and subjective judgment [5]. In addition, unconscious biases related to gender, race, or prior academic performance can influence outcomes, compromising fairness and affecting student confidence and learning.

To tackle these challenges, various AI tools have been explored and developed, including automated writing evaluation [6]–[9], Natural Language Processing (NLP) [10], [11], and most recently, GenAI [12]–[14]. Among GenAI models, LLMs such as OpenAI’s GPT family, Meta’s LLaMA, Claude, Gemini, and Mistral have mainly been used to deliver scalable, personalized, and high-quality feedback and consistent assessments to reduce instructor workload and mitigate bias [1], [15], [16]. Table I presents some well-known LLM models from 2018 to April 2025, highlighting their capabilities. Although some of these models have been examined in educational contexts [14], [17]–[19], many, including multimodal models

TABLE I
TIMELINE OF LARGE LANGUAGE MODELS INTRODUCED FROM 2018 TO
APRIL 2025 WITH KEY CAPABILITIES

| Year | Model | Multimodality | Key Capabilities | Country |
|------|-------------------------|---------------|---------------------|---------|
| 2018 | GPT (OpenAI) | No | Text | USA |
| 2018 | BERT (Google) | No | Text | USA |
| 2019 | GPT2 (OpenAI) | No | Text | USA |
| 2019 | Bart-large-mnli (Meta) | No | Text | USA |
| 2020 | GPT3 (OpenAI) | No | Text | USA |
| 2021 | Claude (Anthropic) | No | Text | USA |
| 2022 | PaLM (Google) | No | Text | USA |
| 2022 | GPT 3.5 (OpenAI) | No | Text | USA |
| 2022 | LaMDA (Google) | No | Text | USA |
| 2022 | Galactica (Meta) | No | Text | USA |
| 2023 | LLaMA (Meta) | No | Text | USA |
| 2023 | GPT 4.0 (OpenAI) | Yes | Text, Visual | USA |
| 2023 | PaLM 2 (Google) | No | Text | USA |
| 2023 | LLaMA 2 (Meta) | No | Text | USA |
| 2023 | Claude 2 (Anthropic) | No | Text | USA |
| 2023 | Mistral (Mistral AI) | No | Text | France |
| 2023 | Grok 1 (xAI) | No | Text | USA |
| 2023 | Gemini 1 (Google) | No | Text | USA |
| 2023 | DeepSeek LLM (DeepSeek) | No | Text | China |
| 2023 | Phi 2 (Microsoft) | No | Text | USA |
| 2024 | Gemini 1.5 (Google) | Yes | Text, Visual, Audio | USA |
| 2024 | Claude 3 (Anthropic) | Yes | Text, Visual | USA |
| 2024 | Phi-3 (Microsoft) | No | Text, Visual | USA |
| 2024 | Granite Code (IBM) | No | Text (Code) | USA |
| 2024 | Qwen2 (Alibaba) | No | Text | China |
| 2024 | Qwen-VL 2 (Alibaba) | Yes | Text, Visual | China |
| 2024 | Llama 3.1 (Meta) | Yes | Text, Visual | USA |
| 2024 | Nemotron 4 (Nvidia) | No | Text | USA |
| 2024 | Pixtral (Mistral AI) | Yes | Text, Visual | France |
| 2024 | GPT 4o (OpenAI) | Yes | Text, Visual, Audio | USA |
| 2024 | GPT O-series (OpenAI) | Yes | Text, Visual | USA |
| 2024 | MM1 (Apple) | Yes | Text, Visual | USA |
| 2024 | DeepSeek V3 (DeepSeek) | No | Text | China |
| 2025 | Qwen2.5 (Alibaba) | No | Text | China |
| 2025 | Qwen-VL 2.5 (Alibaba) | Yes | Text, Visual | China |
| 2025 | Qwen2.5-Math (Alibaba) | No | Text (Math) | China |
| 2025 | Qwen2.5-Code (Alibaba) | No | Text (Code) | China |
| 2025 | DeepSeek R1 (DeepSeek) | No | Text, | China |
| 2025 | Gemini 2.0 (Google) | Yes | Text, Visual, Audio | USA |
| 2025 | Grok 3 (xAI) | Yes | Text, Visual | USA |
| 2025 | GPT 4.5 (OpenAI) | Yes | Text, Visual | USA |
| 2025 | Llama 4 (Meta) | Yes | Text, Visual | USA |

and those specifically designed for tasks such as mathematics or programming, have not yet been thoroughly evaluated. Their effectiveness in different disciplines and assignment formats, which may involve one or more combinations of text, code, equations, figures, or tables, remains underexplored.

This paper aims to provide a comprehensive and up-to-date review of the application of LLMs in assessment and feedback activities within higher education, focusing on studies published between 2023 and 2025. Beyond simply summarizing the current literature, the paper critically evaluates existing approaches, identifies gaps and challenges, assesses the practical effectiveness and limitations of these applications, and outlines directions for future research and development in this rapidly evolving field. In contrast to earlier reviews (e.g., [20]) that focus on discrete tasks, such as automated scoring or feedback generation, our study emphasizes the integration of both feedback and scoring. It highlights the intrinsic link between assessment and feedback, where effective grading improves feedback quality and meaningful feedback enhances grading. This paper categorizes empirical studies into three areas: (1) LLMs for feedback generation, (2) LLMs

for assessment, and (3) studies combining both tasks. It also focuses on the comparative performance of human evaluators versus GenAI systems in grading assessment and feedback generation, emphasizing LLMs’ ability to improve efficiency, personalization, and consistency. Furthermore, the paper addresses AI’s pedagogical, ethical, and institutional challenges in education, including concerns about bias, privacy, and the need for human oversight. Lastly, we highlight the importance of educators’ training and AI literacy for the responsible integration of GenAI systems into higher education.

To guide our systematic review and analysis, this review paper addresses the following Research Questions (RQs):

RQ₁: How are LLMs being utilized to support and automate assessment practices in higher education?

RQ₂: How do LLMs support feedback delivery in higher education?

RQ₃: What are the emerging practices and challenges in combining feedback and assessment with LLMs?

RQ₄: How do LLM-generated feedback and assessments compare to those provided by human instructors in terms of accuracy, fairness, and learner reception?

These questions aim to provide a multi-dimensional understanding of how LLMs are currently shaping assessment and feedback, highlighting both the technological potential and the pedagogical responsibilities involved in their implementation.

Paper structure: Section 2 describes the research methodology, including the research protocol, keyword selection, inclusion and exclusion criteria, and the paper selection process. Section 3 presents the findings of the literature review in four themes aligned with the research questions: (1) assessment (RQ1), (2) feedback (RQ2), (3) integration of assessment and feedback (RQ3), and (4) comparison with human educators (RQ4). Each subsection reviews current practices, key advances, and main challenges in the application of LLMs in higher education. Section 4 discusses current limitations, unresolved challenges, and implementation considerations, offering practical recommendations for institutions and researchers. Section 5 concludes the review by highlighting the key insights and identifying gaps in the literature to advance the effective use of LLMs in educational assessment and feedback.

II. BACKGROUND

To ensure conceptual clarity and avoid potential terminology confusion, this section provides brief definitions of the key terms and methodologies used throughout this paper. Concepts such as prompting, retrieval augmentation, and fine-tuning are defined to facilitate a comprehensive understanding of the techniques discussed and to support the reader in following the subsequent analyses and discussions.

In the rapidly advancing domain of LLMs, the quality and reliability of model-generated responses critically rely on the design and structure of input prompts, a practice

TABLE II
DISTRIBUTION OF SELECTED PAPERS BY CATEGORY AND YEAR (2023–2025), COVERING FEEDBACK, ASSESSMENT, INTEGRATION, AND HUMAN VS. LLM COMPARISONS.

| Category | Year | Subject Areas | LLMs | # Papers |
|--------------|------|--|---|----------|
| Assessment | 2023 | Data Science, information systems | GPT 3.5 | 1 |
| Assessment | 2024 | Operating system, Multidisciplinary, Language Education | GPT 4.0, Transformer Model, GPT 4.0 | 3 |
| Assessment | 2025 | Pedagogy (for mathematics) | GPT 4o | 1 |
| Feedback | 2023 | Physics | GPT 3.5 | 1 |
| Feedback | 2024 | English Literacy, Database Systems, Java, Object Oriented | GPT 3.5, GPT 4.0, PaLM 2, Bart-large-mnli | 4 |
| Integration | 2024 | MSc courses, Communication Networks, Essays | GPT 4.0, LLaMa2 (7 & 13B), Mistral-7B | 3 |
| Integration | 2025 | Statistics, Bioinformatics, Software Engineering & Computer Science courses | GPT 4.0, GPT 4o, LLaMA 3.1, Nemotron | 3 |
| Human vs LLM | 2023 | Hyothetical Case | — | 1 |
| Human vs LLM | 2024 | Reflective Essays, Learning Design and Leadership, Multidisciplinary, Logistics and Transport Management | GPT 3.5, GPT-4 | 4 |

known as `prompting`. Prompting involves providing instructions, questions, or context to guide the LLM output. Various prompting strategies have been designed and developed to enhance model performance in different scenarios. `Zero-shot` prompting, often referred to as basic prompting, is when the model generates a response without any prior examples. In contrast, `few-shot` prompting provides the model with a number of examples to learn from before generating a response. `Chain of thought` prompting breaks down complex problems into intermediate reasoning steps, encouraging the model to think step-by-step, while `multi-step` prompting involves a series of interconnected prompts to refine or extend answers iteratively. `Batched` prompting allows multiple prompts to be processed simultaneously, improving efficiency. `Self-reflective` prompting enables the model to evaluate and revise its own outputs.

An increasingly popular enhancement to prompting is `Retrieval-Augmented Generation (RAG)`. Unlike standard prompting, RAG augments the model’s knowledge by retrieving relevant information from external sources, such as proprietary knowledge bases or curated documents. This retrieved content is then dynamically integrated into the prompt fed to the LLM, generating responses using up-to-date or domain-specific information and thereby reducing hallucinations. By combining retrieval and generation, RAG is considered a powerful framework for improving consistency and handling queries that exceed an LLM’s static training data.

In addition to prompting, `fine-tuning` is another method for adapting LLMs to specific tasks by further training them on a specialized dataset. While `direct fine-tuning` can be highly effective, it is often computationally intensive and requires substantial resources, which are frequently unavailable to many higher education institutions. To address these computational challenges, alternative fine-tuning techniques such as `Low-Rank Adaptation (LoRA)` and `Quantized Low-Rank Adaptation (QLoRA)` have been proposed. These techniques significantly reduce the required computational resources by updating only a small selected portion of the model parameters, allowing fine-tuning to be more efficient and accessible for specialized applications. Among these, QLoRA is generally faster and more memory-efficient due to its use of quantization techniques, which compress

model weights to lower precision formats without significant loss in performance.

III. RESEARCH METHODOLOGY

A structured and systematic approach was adopted to ensure the findings’ reliability, transparency, and reproducibility in conducting this comprehensive review. The study followed a protocol inspired by several best practices in systematic literature reviews (e.g., [20], [21]), focusing on identifying, selecting, analyzing, and synthesizing peer-reviewed publications that explore using LLMs in assessment and feedback within higher education between January 2023 and March 2025. We collected articles from various major academic databases such as IEEE Xplore, ACM Digital Library, ScienceDirect, SpringerLink, Wiley Online Library, and ArXiv, which were selected for their high relevance and comprehensive coverage in the fields of educational technology, AI, and computer science. A set of targeted search strings combining keywords such as “LLM,” “large language model,” “AI feedback,” “automated grading,” “automated scoring,” “AI assessment,” “generative AI,” “education,” “higher education,” and “human versus LLM performance” were applied to titles, abstracts, and full texts to identify the most relevant papers.

After removing duplicates and screening for relevance, a total of **21** empirical studies were selected based on inclusion criteria that required studies to: (1) *focus on higher education*; (2) *report empirical findings*; and (3) *specifically address the use of LLMs in assessment or feedback generation*. Exclusion criteria filtered out review papers, theoretical papers without implementation, studies not in English, works focused solely on K–12 subjects or corporate training contexts, and traditional NPL methods not involving LLMs. Two researchers reviewed each study independently to extract relevant data on the purpose, methodology, LLM models used, evaluation methods, human-AI comparison, outcomes, and implications. Discrepancies in study selection and data extraction were resolved through discussion and consensus. The findings were then synthesized through thematic analysis and categorized into four categories: LLMs used for feedback generation, for assessment, for integrated tasks involving both, and human versus LLMs.

Additionally, we ensured the inclusion of multidisciplinary applications of LLMs across various subject areas in higher

education for each category to provide a comprehensive understanding of the influence, impact, and effectiveness of GenAI tools across diverse academic disciplines. An overview of the distribution of studies by year, category, and subject area is presented in Table II. This research methodology allowed for a balanced representation of LLM capabilities while highlighting emerging challenges, limitations, and areas for further research in aligning LLM-based tools with pedagogical goals in higher education.

IV. LARGE LANGUAGE MODELS IN FEEDBACK AND ASSESSMENT PRACTICES IN HIGHER EDUCATION

This section presents our findings across four key themes: feedback, assessment, integration of feedback and assessment, and comparisons between human and LLM performance. The studies reviewed for this literature analysis, along with their core attributes such as focus area, subject domain, models used, and methodological strategies, are summarized in Table III.

A. Large Language Models for Assessment

Assessment is one of the core components of higher education, helping to evaluate student learning, guide academic progress, and ensure the achievement of learning outcomes. Effective assessment practices not only measure performance but also shape instructional strategies and motivate learners through clear benchmarks and outcomes. In recent years, researchers have explored the use of LLMs in various aspects of assessment, including rubric-based evaluation and automated grading [40], [41]. In this part, we review key studies and applications that explore the role of LLMs in assessment, addressing RQ₁ of this study.

Rubric-based assessment is a structured approach in higher education that uses predefined criteria and performance levels to evaluate student work. This method enhances grading consistency and fairness by clearly outlining expectations and reducing subjectivity. Traditionally, instructors design and apply rubrics manually. Recent advances in LLMs, however, have enabled more dynamic and scalable implementations. For instance, Xie et al. [15] investigated a multi-agent grading system with GPT 4.0 that employed few-shot, self-reflection, and batched prompting to dynamically refine rubrics based on student responses in an operating systems course. This approach demonstrated high grading accuracy and consistency, with LLM-generated rubrics performing similarly to expert-designed ones. In a contrasting approach, Li et al. [25] proposed a human-in-the-loop grading framework using GPT 4o with zero-shot prompting to assess open-ended student responses from a pedagogical knowledge of mathematics course. This framework allows instructors to refine rubric criteria based on LLM interactions. While Xie et al. [15] focused on automated rubric refinement, Li et al. [25] prioritized human oversight to ensure highly controllable grading standards, highlighting a trade-off between full automation and human control. Li et al. [25] also found that Reinforcement Learning (RL)-based Q&A retrieval improved grading accuracy,

suggesting the potential of combining LLMs with RL and human expertise. However, due to the complexity of rubric-based assessment, most existing works have focused primarily on automating the grading process itself rather than fully automating rubric generation and refinement.

Existing methods to automate the assessment process using static rubrics or simplified scoring systems have primarily focused on fine-tuning LLM models or prompt engineering strategies. In this manner, Gobrecht et al. [23] proposed a direct fine-tuning approach using a transformer model to grade 16 different courses with open-ended responses. The training framework used the question, a reference answer, the maximum points, and the student's response as inputs. This approach demonstrated grading accuracy comparable to human evaluators while outperforming instructors in both consistency and speed. However, the authors observed significant variation in model performance across courses, with the highest accuracy achieved in technical subjects like machine learning and artificial intelligence and lower accuracy in diversity management and public law. This highlights the necessity for course-specific model adaptations. While fine-tuning strategies can enhance accuracy, they require deep AI knowledge, computational resources, and diverse datasets to avoid bias, which is often impractical for many instructors. Hence, LLMs with prompt-based strategies have become a popular alternative. These methods generate assessments through prompts, avoiding the need for retraining or large datasets. While they may not reach the accuracy of fine-tuned models, prompt-based approaches are more accessible, flexible, and easier to implement, making them a feasible choice for instructors with limited resources.

Building on prior work that emphasizes fine-tuning, most studies have explored prompting-based approaches as a more accessible alternative. In this context, Schneider et al. [22] evaluated GPT 3.5 with zero-shot prompting for short textual responses in two university-level exams, including data science (English, 21 participants, 16 questions) and information systems (German, 34 participants). The model was prompted with the question and the student's answer, then asked to classify responses on a scale from extremely good to bad, followed by justification. The study found that GPT 3.5 assessments often lacked specificity, offered generic feedback, and showed a bias toward middle categories regardless of actual answer quality. Furthermore, minor alterations in student responses significantly shifted the model's grading, highlighting the limitations of basic zero-shot prompting. Similarly, Pack et al. [24] investigated GPT 4.0 with zero-shot prompting in language education using static rubrics. The study addressed ethical considerations by obtaining informed consent and ensuring voluntary participation, transparency in data use, and student autonomy. The study revealed that GPT 4.0 struggled to consistently follow the evaluation criteria, frequently produced incorrect formats, and was susceptible to prompt hacking, where students manipulated the LLM to achieve higher scores. Hence, the study underscored the challenge of prompt hacking, a concern often raised in safety research, which also affected

TABLE III

OVERVIEW OF SELECTED STUDIES ON LLMs IN HIGHER EDUCATION. INTEGRATION REFERS TO THE COMBINATION OF ASSESSMENT AND FEEDBACK.

| Study | Category | Year | Subject Areas | LLMs | Strategy |
|--------------------------------|--------------|------|--|-----------------------------------|------------------------|
| Schneider et al. [22] | Assessment | 2023 | Data Science and Information System | GPT 3.5 | Prompting |
| Xie et al. [15] | Assessment | 2024 | Operating System | GPT 4.0 | Prompting, Rubrics |
| Gobrecht et al. [23] | Assessment | 2024 | Multidisciplinary | Transformer Model | Fine-tuning |
| Pack et al. [24] | Assessment | 2024 | Language Education | GPT 4.0 | Prompting, Rubrics |
| Li et al. [25] | Assessment | 2025 | Pedagogy (Mathematical Courses) | GPT 4o | Prompting, Rubrics |
| Wan et al. [26] | Feedback | 2023 | Physics | GPT 3.5 Turbo | Prompting |
| Meyer et al. [27] | Feedback | 2024 | English Literacy | GPT 3.5 Turbo | Prompting |
| Riazi et al. [28] | Feedback | 2024 | Database Systems | GPT 4.0 | Prompting |
| Estévez et al. [29] | Feedback | 2024 | Programming Language | GPT 3.5, PaLM 2 | Prompting |
| Jia et al. [30] | Feedback | 2024 | Object-Oriented Course | BART-large-mnli, GPT 3.5, GPT 4.0 | Fine-tuning, Prompting |
| Jauhainen et al. [11] | Integration | 2024 | MSc Level courses | GPT 4.0 | Prompting, Rubric, RAG |
| Katuka et al. [16] | Integration | 2024 | Communication Networks | LLaMA 2 (7B and 13B) | Fine-tuning, Rubric |
| Stahl et al. [31] | Integration | 2024 | Secondary level Essay | Mistral-7B, LLaMA 2 | Prompting, Rubric |
| Yeung et al. [32] | Integration | 2025 | Statistics | GPT 4.0 | Prompting, Rubric |
| Poličar et al. [33] | Integration | 2025 | Bioinformatics | GPT 4o, LLaMA 3.1, Nemotron | Prompting, Rubric |
| Diyab et al. [34] | Integration | 2025 | SE and CS course | GPT 4.0 | Prompting |
| Kumar [35] (hypothetical case) | Human Vs LLM | 2023 | — | — | — |
| Awid [36] | Human Vs LLM | 2024 | Reflective Essays (Engineering Course) | GPT Models | Prompting, Rubric |
| Saini et al. [37] | Human Vs LLM | 2024 | Learning Design and Leadership | GPT 3.5 | — |
| Nazaratsky et al. [38] | Human Vs LLM | 2024 | Multidisciplinary | GPT 4.0 | Prompting |
| Flodén [39] | Human Vs LLM | 2024 | Logistics and Transport Management | GPT 3.5 | Prompting, Rubric |

academic assessments. These studies underscore the challenges associated with basic prompting strategies, particularly for complex assessment tasks, which often lead to lower reliability and consistency compared to human evaluations.

Collectively, these studies highlight the potential of LLMs to enhance grading practices in higher education, particularly in terms of scalability, speed, and consistency. However, they also reveal several challenges, including issues with model transparency, temporal stability, and ethical concerns such as trust and bias. Even though publicly available LLMs have shown potential in certain cases, custom-trained models, especially those using dynamic rubric refinement or fine-tuning, provide more robust solutions for course assessments. However, developing these models can be complex and resource-intensive, as they must be designed for each specific course, making them less accessible for many instructors and institutions. Given these challenges, a hybrid approach, combining LLMs with human oversight and prompt-based strategies, emerges as a practical solution to ensure both reliability and fairness in automated grading.

B. Large Language Models for Feedback

Feedback is another key element in higher education, helping learners in understanding their progress, correcting misconceptions, and improving performance. Feedback promotes self-regulated learning by encouraging students to reflect and adapt their learning strategies. For educators, feedback offers insights into student understanding and informs instructional adjustments. Recent studies show that LLMs can support feedback generation across various disciplines and course types in higher education via prompting to guide responses and fine-tuning to improve task-specific accuracy. Here, we summarize relevant literature and discuss how LLMs have been used to enhance feedback, addressing RQ₂ of this study.

Prompting has been a commonly used strategy to generate feedback with LLMs, and the results vary based on the subject

area and framework used. For example, Wan et al. [26] utilized GPT 3.5 Turbo with few-shot learning to provide personalized feedback on student-written responses in a Physics course. The study reported a high level of student satisfaction (70%), with feedback frequently rated as comparable and indistinguishable from the feedback provided by human instructors. However, this approach required significant initial labor for prompt design to generate high-quality examples, and its effectiveness was highly dependent on prompt patterns and contents. In a separate study, Meyer et al. [27] also employed GPT 3.5 Turbo, but with a zero-shot prompting strategy to generate structured feedback on argumentative English essays. While their approach significantly improved revision quality and student motivation in the short term, it showed no transfer effects to subsequent tasks and received only moderate ratings for the absolute usefulness of feedback. This contrasts with Wan et al. [26] findings on indistinguishable feedback, suggesting that the effectiveness and transferability of LLM-generated feedback can vary considerably based on prompting strategy and subject domain.

More advanced prompting techniques have also been explored. For instance, Riazi et al. [28] introduced an LLM-based system using GPT-4.0 with a multi-step prompting strategy to support conceptual design learning in database systems courses. Their approach involved converting diagrams to JSON format and employing targeted prompts to generate detailed feedback and even create Frequently Asked Questions (FAQs). This method achieved high usefulness ratings (84%) and demonstrated strong precision in identifying structural errors. However, the system still faced challenges in detecting more complex constructs. In contrast to the positive outcomes reported by Wan et al. [26], Meyer et al. [27], and Riazi et al. [28], Estévez et al. [29] found that ChatGPT 3.5 and PaLM 2 struggled significantly in providing feedback for complex programming code in a Java programming course, particularly for issues like race conditions and deadlocks. Despite employ-

ing zero-shot, few-shot, and contextual prompting strategies, both models showed low precision, frequently producing false positives, and failed to consistently match expert feedback, highlighting the current limitations of LLMs in highly specialized and complex problems.

Many state-of-the-art feedback generation strategies using LLMs fail to address three important factors: data privacy, hallucination, and evaluation of model fine-tuning. To overcome these limitations, Jia et al. [30] investigated the trade-offs between a data-driven fine-tuning approach and a few-shot prompting strategy for generating feedback on student project reports in an object-oriented course. This study fine-tuned BART-large-mnli [42] and ChatGPT 3.5, and used few-shot learning with ChatGPT 4.0. Their findings indicated that while few-shot learning slightly reduced hallucination compared to fine-tuning (23.5% vs. 27.1% of sentences with hallucinations), neither approach fully mitigated this challenge. Fine-tuning tended to produce more intrinsic hallucinations (16%), whereas few-shot prompting resulted in more extrinsic ones (13.7%), primarily due to limited contextual grounding. The findings suggest that while few-shot learning with ChatGPT 4.0 slightly reduced hallucination, neither approach fully mitigated it, underscoring the need for further refinement in LLM-generated feedback mechanisms.

Collectively, the reviewed studies suggest that LLMs have strong potential to generate high-quality, timely, and personalized feedback in higher education, while also positively influencing student motivation. Their application across diverse subject areas highlights their versatility and ability to enhance learning outcomes and reduce instructor workload. However, most existing studies are short-term pilots involving limited feedback instances, offering insufficient evidence on how LLM-generated feedback supports knowledge transfer or influences long-term student performance across multiple assignments. Moreover, the effectiveness of feedback depends heavily on prompt design, model capabilities, and task complexity. Challenges, including ensuring feedback specificity, avoiding hallucinated content, and supporting sustained learning, remain active areas of investigation. As a result, integrating LLM-generated feedback into educational settings may currently benefit most from hybrid approaches that follow the trade-off between scalability and pedagogical soundness by combining LLMs with human oversight.

C. Integration of Assessment and Feedback

Feedback and assessment are often studied as distinct components in educational research. However, they are interdependent processes in real-world academic practice. In higher education, instructors often assess student work by either grading first and justifying the score with feedback or by providing feedback first, then combining the student's answer, the feedback, and the rubric to assign a grade. However, most existing work on LLMs in higher education treats feedback and assessment as isolated tasks, optimizing one while neglecting their pedagogical interconnection. This separation can weaken the effectiveness of AI-based educational tools.

Consequently, their overall ability to support teaching and learning can be diminished. In this part, we summarize relevant literature and explore how LLM-based frameworks have been applied to enhance feedback and assessment together. We also investigate which of the two is typically prioritized in existing work, addressing RQ₃ of this study.

Recent studies have begun to explore frameworks that integrate feedback and assessment using LLMs, acknowledging that these processes influence each other. However, many of these studies primarily focus on generating feedback or assigning grades without explicitly analyzing the prioritization between the two, or how one affects the other. To address this gap, prompting strategies and fine-tuning methods have been employed to enable LLMs to perform both tasks in a more interconnected and pedagogically meaningful way. For example, Jauhiainen et al. [1], investigated ChatGPT 4.0's ability to assess and provide feedback on open-ended English responses. Their framework used one-shot and ten-shot grading guided by a chain of thought prompting and Retrieval-Augmented Generation (RAG), showing high grading consistency (68.7% identical grades, 96.1% within one grade point). However, teachers generally gave higher scores. While the model generated individualized feedback approved by teachers, it struggled with evaluating higher-level thinking skills. In a similar manner, Yeung et al. [32] applied a GPT 4.0-based framework for automated feedback and assessment in a statistics course, where a zero-shot prompt simultaneously assigned scores and generated personalized feedback. This system demonstrated strong correlations with human classification and improved student motivation. While both studies highlight the potential for integrated assessment and feedback, Jauhiainen et al. [1] noted challenges in assessing higher-level cognitive skills, a significant limitation not explicitly reported by Yeung et al. [32].

Expanding on integrated approaches, Poličar et al. [33] conducted a more extensive study evaluating six different LLMs, including GPT 4o, LLaMA 3.1, and Nemotron, for automated grading and feedback in a bioinformatics course. Their few-shot prompting strategy integrated various inputs such as question, correct answer, student submission, grading rubric, and 10 grading examples into a single prompt for both tasks. Most models achieved a strong grading accuracy (85–90%) comparable to human grading, and student satisfaction with LLM feedback was generally high, although Nemotron received more negative ratings. A key limitation noted was the stochastic nature of LLM outputs, as identical prompts could yield different results across runs, raising concerns about reliability. Furthermore, while measures were taken to prevent prompt hacking, the study did not observe such attempts, leaving this vulnerability unexamined. Another limitation was the integration of scoring and feedback generation into a single prompt, which excluded analysis of how these components might influence each other. Furthermore, unlike Jauhiainen et al. [1] and Yeung et al. [32], which focused on specific subject areas, Poličar's broader model comparison provides a more comprehensive view of integrated performance across

different LLM architectures. Complementing these, Diyab et al. [34] presented a comprehensive framework using ChatGPT 4.0 with both zero and few-shot prompts to automate four core assessment tasks in software engineering and computer science education: grading, weakness identification, feedback provision, and question generation. Their few-shot prompt significantly improved grading accuracy (from 16% to 80% exact match), demonstrating the benefit of providing examples. This finding aligns with the general effectiveness of few-shot learning observed in other studies for improving LLM performance. However, Diyab et al. [34] also highlighted common limitations such as potential risks of bias, hallucinations, and the crucial need for transparency and human oversight, resonating with concerns raised across the broader LLM literature.

While prompting strategies are widely adopted due to their accessibility, recent research has also explored fine-tuning to improve consistency and relevance in integrated assessment and feedback tasks. Katuka et al. [16] used fine-tuning techniques, specifically LoRA and QLoRA, to adapt quantized LLaMA 2 (7B and 13B) models for automated scoring and feedback generation. The study utilized a combination of the Short Answer and Feedback (SAF) dataset [43] and proprietary datasets on communication networks topics to train and evaluate models. Their experiments showed remarkable efficiency and strong performance (less than 3% grade prediction error), outperforming traditional baselines. In particular, the quality of the feedback improved significantly when the model was provided with predicted scores as input, suggesting a beneficial interdependence between scoring and feedback generation when fine-tuning is applied. This contrasts with prompting-centric studies, which often treat these as distinct or simply co-generated outputs, highlighting fine-tuning potential for deeper task integration and improved output quality.

To the best of our knowledge, no existing work in higher education has explicitly studied the prioritization between feedback and assessment, specifically, whether providing feedback before scoring or vice versa yields better outcomes. However, Stahl et al. [31] conducted a relevant study at the secondary school level. They investigated how different prompting strategies affect LLMs in joint essay scoring and feedback generation, specifically exploring task instruction types (scoring before or after feedback). Their findings indicated that strategies generating feedback prior to scoring were generally more effective and that feedback was most beneficial when produced independently of scoring. This suggests that the sequencing of these tasks can impact outcomes, a significant aspect that most other integrated frameworks have not yet thoroughly explored.

In summary, the integration of feedback and assessment in higher education through LLMs remains an evolving area of research. While numerous studies have employed prompting strategies, including single prompts that generate both feedback and scores, and multiple prompts handling each task separately, some have also explored fine-tuning approaches. There is a notable absence of comprehensive evaluations ad-

ressing the sequencing of feedback and scoring. Specifically, questions such as whether providing feedback before scoring yields better outcomes than the reverse remain unexplored. Similarly, whether combining both tasks in a single prompt is more effective than separating them is also not addressed in most of the existing studies.

D. Enhancing or Replacing Human Feedback and Assessment in Higher Education

One of the most debated questions in the integration of LLMs in higher education is whether these models can enhance or even replace human input in the processes of scoring and feedback. This part addresses RQ₄, which explores how LLMs compare with human educators in performing assessment and feedback tasks in higher education. Specifically, we examine existing studies that assess the accuracy, quality, consistency, and pedagogical usefulness of AI-generated scores and feedback relative to those provided by instructors and peers. This includes analyses of how students perceive the fairness, trustworthiness, and usefulness of LLM outputs, although discussion of how these perceptions vary across learner demographics and assignment types remains limited. Understanding these comparisons is essential for determining the appropriate role of LLMs in academic environments, whether as supportive assistants, co-evaluators, or autonomous evaluators, and for identifying the limitations that must be addressed before broader adoption.

Much discussion of AI tools in higher education has focused on academic integrity, especially plagiarism concerns, but the conversation is evolving. A growing number of studies are examining AI's potential to support grading and provide personalized feedback in a timely and scalable way. One line of research has compared human versus AI evaluation in grading and feedback, exploring student perceptions and aspects of efficiency, consistency, and reliability. Although exams using multiple-choice questions have relied on automated scoring for the past few decades, an emerging line of research has investigated perceptions of automated scoring of essay-based exams using AI. A growing number of research studies show that AI scoring can be mistaken for human scoring. Also, current AI tools - namely ChatGPT (models 1.0 and 3.5) in these studies- are proving adept at assessing student answers and generating plausible and comprehensible feedback. Awidi [36] compared the grading reliability of human expert tutors against ChatGPT scripts when reviewing short essays produced in multiple first-year engineering courses. This study suggests that it is premature to rely solely on ChatGPT for grading reflective essays; however, the study did not specify which model of ChatGPT was used. One study insight was that ChatGPT demonstrated more consistency and rigor in scoring student writing compared to human expert tutors, but the ChatGPT scripts required consistent iteration of improved prompts. However, ChatGPT grading did not provide personalized feedback to nearly a third of the cases, and much of the feedback was repetitive and generic. Moreover, two independent reviewers both agreed that the

expert tutors consistently awarded higher scores to students, while ChatGPT's scoring tended to be more strict following the assessment criteria.

Comparing human and AI, Saini et al. [37] investigate the experience of 295 professional students who received feedback on project drafts from both peers and an AI-enabled ChatGpt-3.5 feature on an online course platform. They found that peer feedback consistently received higher ratings than AI reviews. Although a significant percentage (41%) of students preferred feedback from both peer and AI reviews, over a third (35%) exclusively preferred peer reviews. In addition, the study found different patterns between education level and that group's perceptions of the AI feedback quality, usefulness, and actionability. For example, individuals pursuing a doctorate degree rated AI feedback the highest, and those enrolled in a master's degree rated it much lower. This difference in perception suggests that AI feedback should be tailored to specific demographic groups. Notably, several differences became apparent on the perceived tradeoffs between peer reviews and AI reviews (pg.36): Peer feedback was characterized by delays, contrasting with the immediacy of AI feedback. Peer reviews typically entail a single attempt per student, contrasting with the ability of AI reviews to allow for multiple iterations. Peers often seem to offer feedback in various modes, such as text, audio, PDF, or video, whereas AI feedback primarily consists of one mode (text). Peer reviews are perceived as time-intensive, while AI reviews are perceived as time-efficient.

Other studies further discuss the comparison between human and AI feedback. Based on a study of 457 students, Nazaretsky et al. [38] found that students' ability to differentiate between AI and human feedback depends on the task at hand. Before students identify whether a human or AI is giving them feedback, they do not perceive a difference in quality or in friendliness. Slightly more than half (59%) of students correctly guessed which feedback came from a human or AI source. When disclosing the feedback source, students consistently prefer human feedback over AI-generated feedback. Moreover, they either lower the score of the AI or increase the score of the human, which the researchers infer that the students do not yet trust AI as a source. An implication is that students who do not trust AI as a valid feedback source are less likely to improve their learning. Similarly, Flodén [39] compared 463 responses from master-level exams in Sweden that were graded by ChatGPT 3.5 versus human teachers. ChatGPT tended to score exams slightly higher than human teachers but tended to avoid very high or very low scores on individual questions. During the interviews, teachers were surprised by how well ChatGPT's grading matched their own scores. Cases are useful in exploring deeper issues as a preventative thought exercise. Kumar [35] presents a hypothetical case of a junior university professor using a third-party AI tool to grade student papers. The exact AI tool is not specified because the discussion focused more on the potential career implications for faculty, especially adjunct instructors with increasing workloads and those seeking tenure. One point

raised in the case is how students would feel if they knew their assignments were graded by AI. How should faculty consider the use of AI tools when the institution lacks a clear policy or access to suitable tools, and when can instructors use tools?

V. DISCUSSION

The application of LLMs in feedback and assessment in higher education has progressed in three interconnected areas: assessment automation, feedback generation, and their integration. In assessment, both static and dynamic rubric-based systems have been developed, along with approaches that automate grading without predefined rubrics. LLMs have been used to evaluate student responses with consistency and efficiency. Among all, techniques such as few-shot, chain-of-thought, and self-reflective prompting help to align model outputs with rubric criteria and human judgment. In feedback, few-shot prompting has shown strong potential in guiding LLMs to generate individualized, context-aware, and pedagogically relevant responses. Currently, integrated approaches link assessment and feedback into unified workflows that support continuous learning. These systems are capable of automatically grading responses, generating personalized feedback for students, and producing feedback for educators to identify concept thresholds or areas where learners commonly struggle. Additionally, they can generate individualized follow-up questions based on a student's performance and the feedback provided, allowing for adaptive and targeted learning interventions. The existing studies indicate that while LLMs can approach or match human accuracy in narrow domains, challenges related to reliability, context sensitivity, and fairness persist.

To address these challenges, it is important to examine the current limitations that restrict the effectiveness of LLMs in educational settings. One of the most critical issues is LLM reliability, particularly in feedback tasks. Models often hallucinate or misinterpret the context, especially when zero-shot prompting is used. While zero-shot prompting is easy to implement, it tends to provide unstable outputs and lacks adaptability across diverse student responses. Zero-shot methods are slightly more effective for scoring tasks, where structured rubrics guide the responses, but they fall short when more detailed feedback is required. Few-shot prompting offers more stable and accurate outputs, but crafting examples that are general enough to cover a wide range of cases remains challenging. If the examples used for training are too narrow or biased, the model's responses may become skewed, raising significant concerns about fairness and consistency in the generated output. Another reliability-related vulnerability is prompt hacking, in which students exploit the structure of prompts to deceive or influence the model, thus compromising the validity and integrity of LLM-based assessments. These issues highlight the need for safeguards and robust prompt design in high-stakes applications. For a comprehensive analysis of some of these limitations, including the prevalence of hallucinations, the influence of question framing, and the

impact of system instructions, see [44], [45] for a detailed examination.

To address the limitations of prompting, some researchers have employed fine-tuning approaches. Fine-tuning LLMs for specific educational tasks can improve performance and reduce the need for complex prompts. However, this approach demands specialized expertise, access to large and high-quality datasets, and significant computational resources. Without these, fine-tuned models risk overfitting, reduced generalizability, and potential bias. These challenges currently make fine-tuning a less favored approach in practice.

While model customization can enhance the performance of feedback and assessment frameworks, the choice of models used in educational research also plays a vital role. Most current work focuses on general-purpose models like GPT 3.5, GPT 4.0, GPT 4o, or LLaMA. These models have been widely used as benchmark in education tasks, but there are some domain-specific LLMs, such as those designed for coding, mathematics, and other disciplines (see Table I), that remain largely unexplored in the context of educational feedback and assessment. Evaluating these task-specific models in educational contexts could yield better performance and more relevant outputs. Moreover, existing evidence suggests that different models may perform better depending on the assignment type or question, highlighting the need to evaluate multiple LLMs rather than relying on a single one. Additionally, many educational materials, such as essays and reports, often include multimodal elements like tables, figures, or diagrams. However, there is a lack of research assessing the capabilities of multimodal LLMs in processing and providing feedback on these richer content formats. These gaps present important directions for future research to explore the comparative effectiveness of both domain-specific and multimodal LLMs in diverse educational scenarios.

Beyond the models themselves, the availability and quality of datasets limit the research progress in this domain. A significant obstacle to the broader applicability and standardization of LLMs in educational assessments is the absence of robust and universally applicable datasets. While several datasets such as AGIEval [46], AlpacaFarm [47], BBQ [48], BoolQ [49], GSM8K [50], MMLU [51], WinoGrande [52], ASAP [53], SAF [43], Mohler [54], and others, have been developed to benchmark feedback and scoring tasks, they remain underutilized in large-scale studies or lack alignment with institutional standards for assignments, essays, and exams. Their limited scope, often confined to specific domains or educational systems, hinders the creation of LLM solutions that can be applied universally. This limitation undermines the reproducibility, comparability, and scalability of research efforts in the field.

Beyond data-related issues, there are broader educational design concerns, particularly regarding the transferability of LLMs. LLM performance varies across disciplines and assignment types, particularly in non-technical or interpretive fields where context is critical. Additionally, little is known about the longitudinal effectiveness of AI-generated feedback. Most

studies focus on isolated tasks, neglecting whether feedback from one assignment supports improvement in future assignments. This gap limits our understanding of the sustained pedagogical impact of LLM integration in education.

In addition to technical, data-related, and educational design issues, ethical considerations are central to the responsible implementation of LLMs in education. Ethical and legal concerns demand attention. Risks include the generation of biased or inappropriate outputs, data privacy violations, and dependency on non-transparent decision-making systems. To mitigate these concerns, institutions adopt various strategies such as obtaining student consent, anonymizing inputs, or deploying offline LLM instances on university-managed cloud infrastructure to ensure compliance with data regulations like FERPA and GDPR. However, the ethical implications such as AI influencing grades or feedback require continuous oversight and clear governance. In addition, many studies adopt a human-in-the-loop model, wherein instructors review, refine, and validate AI-generated feedback and grades before they are delivered to students. Studies suggest that this human oversight is essential to avoid producing misleading feedback or misgrading students, ensuring that AI tools remain assistive rather than authoritative.

Based on the aforementioned concerns and limitations in the existing works such as reliability issues in prompting strategies, limited availability of high-quality datasets, domain-specific performance variation, ethical risks, and the need for human oversight institutions must take a strategic and responsible approach to implementing LLMs in educational settings. A practical and low-risk starting point is to adopt LLMs as assistive tools for feedback generation or assessment individually as the literature mostly focus on this rather than integration. In this model, educators use carefully crafted prompts to generate feedback (few-shot) or scores (zero-shot), which is then reviewed and refined by educator before being delivered to students. This ensures that LLMs enhance, rather than replace, human judgment and pedagogy. With minimal technical overhead, educators can incorporate LLMs into their existing workflows using prompting and open-source AI models like LLaMA, making the technology more accessible and adaptable for everyday teaching scenarios.

For institutions aiming to maximize the long-term pedagogical benefits of LLMs, a more advanced strategy involves developing fully integrated systems that unify assessment and feedback. These systems can automatically grade student responses, provide personalized feedback, and generate diagnostic insights for instructors by identifying common misconceptions or learning gaps (concept thresholds). In this manner, the system can support adaptive learning by creating individualized follow-up content based on each student's needs. While this integrated approach requires more substantial investments in data infrastructure, model customization, and faculty training, it offers greater potential for scalability, personalized education, and continuous learning. Ultimately, by aligning implementation strategies with the current limitations and various considerations, institutions can take the benefits of

LLMs while maintaining academic integrity, fairness, and educational effectiveness.

VI. CONCLUSION

LLMs have shown transformative potential to enhance feedback and assessment in higher education. However, their deployment must be approached thoughtfully, with attention to limitations in reliability, dataset generalizability, ethical safeguards, and pedagogical design. Addressing current gaps such as the lack of diverse model evaluation, underuse of domain-specific LLMs, challenges in fine-tuning, and the need for human-in-the-loop systems will be essential. Through well-designed frameworks, robust validation, and close collaboration between educators and AI developers, LLMs can support a more personalized, scalable, and equitable future of education.

REFERENCES

- [1] A. G. G. Jussi S. Jauhiainen, "Generative ai in education: Chatgpt-4 in evaluating students' written responses," *Innovations in Education and Teaching International*, vol. 0, no. 0, pp. 1–18, 2024.
- [2] D. Agostini and F. Picasso, "Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework," *Intelligenza Artificiale*, vol. 18, no. 1, pp. 121–138, 2024.
- [3] W. Dai, J. Lin, F. Jin, T. Li, Y.-S. Tsai, D. Gasevic, and G. Chen, "Can large language models provide feedback to students? a case study on chatgpt," in *IEEE International Conference on Advanced Learning Technologies*, 2023, pp. 323–325.
- [4] M. G. Hahn, S. M. B. Navarro, L. De La Fuente Valentín, and D. Burgos, "A systematic review of the effects of automatic scoring and automatic feedback in educational settings," *IEEE Access*, vol. 9, pp. 108 190–108 198, 2021.
- [5] D. A. Wiley, "Learning objects in public and higher education," in *Innovations in Instructional Technology*. Routledge, 2006, pp. 1–9.
- [6] D. Ramesh and S. K. Sanampudi, "An automated essay scoring systems: A systematic literature review," *Artificial Intelligence Review*, vol. 55, pp. 2495–2527, 2022.
- [7] S. A. Crossley, P. Baffour, Y. Tian, A. Picou, M. Benner, and U. Boser, "The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (persuade) corpus 1.0," *Assessing Writing*, vol. 54, 2022.
- [8] T. T. N. Ngo, H. H.-J. Chen, and K. K.-W. Lai, "The effectiveness of automated writing evaluation in efl/esl writing: A three-level meta-analysis," *Interactive Learning Environments*, pp. 1–18, 2022.
- [9] A. Nunes, C. Cordeiro, T. Limpo, and S. L. Castro, "Effectiveness of automated writing evaluation systems in school settings: A systematic review of studies from 2000 to 2020," *Journal of Computer Assisted Learning*, vol. 38, no. 2, pp. 599–620, 2022.
- [10] E. Bauer, M. Greisel, I. Kuznetsov, M. Berndt, I. Kollar, M. Dresel, M. R. Fischer, and F. Fischer, "Using natural language processing to support peer-feedback in the age of artificial intelligence: A cross-disciplinary framework and a research agenda," *British Journal of Educational Technology*, vol. 54, pp. 1222–1245, 2023.
- [11] S. Yang, O. Nachum, Y. Du, J. Wei, P. Abbeel, and D. Schuurmans, "Foundation models for decision making: Problems, methods, and opportunities," *ArXiv*, vol. abs/2303.04129, 2023.
- [12] Y. Chen, R. Wang, H. Jiang, S. Shi, and R. Xu, "Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study," *ArXiv*, vol. abs/2304.00723, 2023.
- [13] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, and X. Xie, "A survey on evaluation of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, 2024.
- [14] Z. Hou, A. Ciuba, and X. Li, "Improve llm-based automatic essay scoring with linguistic features," *ArXiv*, vol. abs/2502.09497, 2025.
- [15] W. Xie, J. Niu, C. J. Xue, and N. Guan, "Grade like a human: Rethinking automated assessment with large language models," 2024. [Online]. Available: <https://arxiv.org/abs/2405.19694>
- [16] G. A. Katuka, A. Gain, and Y.-Y. Yu, "Investigating Automatic Scoring and Feedback using Large Language Models," May 2024, arXiv:2405.00602. [Online]. Available: <http://arxiv.org/abs/2405.00602>
- [17] K. Laak and J. Aru, "Ai and personalized learning: bridging the gap with modern educational goals," *ArXiv*, vol. abs/2404.02798, 2024.
- [18] C. Cao, "Leveraging large language model and story-based gamification in intelligent tutoring system to scaffold introductory programming courses: A design-based research study," 2023. [Online]. Available: <https://arxiv.org/abs/2302.12834>
- [19] S. Al Faraby, A. Romadhony, and Adiwijaya, "Analysis of llms for educational question classification and generation," *Computers and Education: Artificial Intelligence*, vol. 7, 2024.
- [20] D. Agostini and F. Picasso, "Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework," *Intelligenza Artificiale*, vol. 18, no. 1, pp. 121–138, 2024.
- [21] B. Dong, J. Bai, T. Xu, and Y. Zhou, "Large language models in education: A systematic review," in *2024 6th International Conference on Computer Science and Technologies in Education (CSTE)*, 2024, pp. 131–134.
- [22] J. Schneider, B. Schenk, C. Niklaus, and M. Vlachos, "Towards llm-based autograding for short textual answers," *ArXiv*, vol. abs/2309.11508, 2023.
- [23] A. Gobrecht, F. Tuma, M. Möller, T. Zöller, M. Zakhvatkin, A. Wuttig, H. Sommerfeldt, and S. Schütt, "Beyond human subjectivity and error: a novel ai grading system," *arXiv preprint arXiv:2405.04323*, 2024.
- [24] A. Pack, A. Barrett, and J. Escalante, "Large language models and automated essay scoring of english language learner writing: Insights into validity and reliability," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100234, 2024.
- [25] H. Li, Y. Chu, K. Yang, Y. Copur-Gencurk, and J. Tang, "Llm-based automated grading with human-in-the-loop," 2025. [Online]. Available: <https://arxiv.org/abs/2504.05239>
- [26] T. Wan and Z. Chen, "Exploring generative ai assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning," *Physical Review Physics Education Research*, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:265129035>
- [27] J. Meyer, T. Jansen, R. Schiller, L. W. Liebenow, M. Steinbach, A. Horbach, and J. Fleckenstein, "Using llms to bring evidence-based feedback into the classroom: Ai-generated feedback increases secondary students' text revision, motivation, and positive emotions," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100199, 2024.
- [28] S. Riazzi and P. Rooshenas, "Llm-driven feedback for enhancing conceptual design learning in database systems courses," in *Technical Symposium on Computer Science Education*, 2024. [Online]. Available: <https://api.semanticscholar.org/CorpusID:274992071>
- [29] I. Estévez-Ayres, P. Callejo, M. A. Hombrados-Herrera, C. Alario-Hoyos, and C. D. Kloos, "Evaluation of llm tools for feedback generation in a course on concurrent programming," *International Journal of Artificial Intelligence in Education*, 2024.
- [30] Q. Jia, J. Cui, R. Xi, C. Liu, P. Rashid, R. Li, and E. Gehringer, "On assessing the faithfulness of llm-generated feedback on student assignments," in *International Conference on Educational Data Mining*. Atlanta, Georgia, USA: International Educational Data Mining Society, July 2024, pp. 491–499.
- [31] M. Stahl, L. Biermann, A. Nehring, and H. Wachsmuth, "Exploring llm prompting strategies for joint essay scoring and feedback generation," *arXiv preprint arXiv:2404.15845*, 2024.
- [32] C. Yeung, J. Yu, K. C. Cheung, T. W. Wong, C. M. Chan, K. C. Wong, and K. Fujii, "A zero-shot llm framework for automatic assignment grading in higher education," 2025. [Online]. Available: <https://arxiv.org/abs/2501.14305>
- [33] P. G. Poličar, M. Špendl, T. Curk, and B. Zupan, "Automated assignment grading with large language models: Insights from a bioinformatics course," 2025. [Online]. Available: <https://arxiv.org/abs/2501.14499>
- [34] A. Diyab, R. M. Frost, B. D. Fedoruk, and A. Diyab, "Engineered prompts in chatgpt for educational assessment in software engineering and computer science," *Education Sciences*, vol. 15, no. 2, p. 156, 2025.
- [35] R. Kumar, "Faculty members' use of artificial intelligence to grade student papers: a case of implications," *Int J Educ Integr*, vol. 19, no. 1, p. 9, 2023.
- [36] I. T. Awidi, "Comparing expert tutor evaluation of reflective essays with marking by generative artificial intelligence (ai) tool," *Computers and Education: Artificial Intelligence*, vol. 6, p. 100226, 2024.

- [37] A. K. Saini, B. Cope, M. Kalantzis, and G. C. Zapata, "The future of feedback: Integrating peer and generative ai reviews to support student work," *Pre-print, DOI*, vol. 10, 2024.
- [38] T. Nazaretsky, P. Mejia-Domenzain, V. Swamy, J. Frej, and T. Käser, "Ai or human? evaluating student feedback perceptions in higher education," in *Technology Enhanced Learning for Inclusive and Equitable Quality Education*. Springer Nature Switzerland, 2024, pp. 284–298.
- [39] J. Flodén, "Grading exams using large language models: A comparison between human and ai grading of exams in higher education using chatgpt," *British Educational Research Journal*, vol. 51, no. 1, pp. 201–224, 2025.
- [40] N. S. H. Sally, H.-y. Chan, J. H. K. Wong, L. S. C. Foong, and A. J. Privitera, "A scoping review of the use of generative ai in assessment in higher education," 2024.
- [41] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang *et al.*, "A survey on evaluation of large language models," *ACM transactions on intelligent systems and technology*, vol. 15, no. 3, pp. 1–45, 2024.
- [42] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. rahman Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics*, 2019.
- [43] A. Filighera, S. Parihar, T. Steuer, T. Meuser, and S. Ochs, "Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022.
- [44] A. Algaba, V. Holst, F. Tori, M. Mobini, B. Verbeken, S. Wenmackers, and V. Ginis, "How deep do large language models internalize scientific literature and citation practices?" *arXiv preprint arXiv:2504.0276*, 2025.
- [45] P. L. Jeune and D. Berenstein, "Phare: Analysis of hallucination in leading llms," <https://huggingface.co/blog/davidberenstein1957/phare-analysis-of-hallucination-in-leading-llms>, May 2025, accessed: 16-May-2025.
- [46] W. Zhong, R. Cui, Y. Guo, Y. Liang, S. Lu, Y. Wang, A. S. S. Saied, W. Chen, and N. Duan, "AgiEval: A human-centric benchmark for evaluating foundation models," in *NAACL-HLT*, 2023.
- [47] Y. Dubois, X. Li, R. Taori, T. Zhang, I. Gulrajani, J. Ba, C. Guestrin, P. Liang, and T. Hashimoto, "AlpacaFarm: A simulation framework for methods that learn from human feedback," *ArXiv*, vol. abs/2305.14387, 2023.
- [48] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, "Bbq: A hand-built bias benchmark for question answering," in *Findings*, 2021.
- [49] C. Clark, K. Lee, M.-W. Chang, T. Kwiatkowski, M. Collins, and K. Toutanova, "Boolq: Exploring the surprising difficulty of natural yes/no questions," *ArXiv*, vol. abs/1905.10044, 2019.
- [50] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman, "Training verifiers to solve math word problems," *ArXiv*, vol. abs/2110.14168, 2021.
- [51] P. Liu, "Mmlu dataset," 2023. [Online]. Available: <https://www.kaggle.com/ds/3638509>
- [52] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande," *Communications of the ACM*, vol. 64, pp. 99 – 106, 2019. [Online]. Available: <https://api.semanticscholar.org/CorpusID:198893658>
- [53] B. Hamner, J. Morgan, lynnvande, M. Shermis, and T. V. Ark, "The hewlett foundation: Automated essay scoring," <https://kaggle.com/competitions/asap-aes>, 2012, kaggle.
- [54] M. Mohler, R. Bunesco, and R. Mihalcea, "A larger collection of short student answers and grades for a course in computer science," <https://web.eecs.umich.edu/~mihalcea/downloads.html>, 2011, dataset available online.