

Building Bridges to Student Growth: An LLM-Powered Feedback Generation System for Holistic Development

Hi-kuen Yu(✉), Jacky Keung, and Yicheng Sun

Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong SAR, China

hikuenyu2-c@my.cityu.edu.hk

Abstract. High-quality feedback is crucial for student learning, yet scaling it across both academic and extracurricular achievements remains difficult. This study introduces an LLM-powered system that generates professional feedback from structured extracurricular records. Using a dataset of 100 secondary students with 1,067 entries, five open-source models (DeepSeek-R1, Gemma3, GPT-OSS, Llama-3.1, Qwen3) produced 500 outputs evaluated across faithfulness, coverage, fluency, coherence, and relevance. Results show that GPT (4.48) and Gemma (4.45) achieved the highest overall means, with medium-sized models (20B–27B) offering the best trade-off between accuracy and inference stability. Strengths included consistently fluent and coherent prose (>4.6), while coverage remained the weakest dimension (3.48–4.25). These findings position LLMs as drafting aids rather than replacements. They can reduce teacher workload and accelerate feedback cycles, provide human oversight to ensure factual alignment, fairness, and contextual appropriateness.

Keywords: large language models · generative AI in education · student development · feedback generation

1 Introduction

High-quality feedback is widely recognized as a critical driver of student learning, particularly when it is timely, specific, and actionable (Hattie and Timperley, 2007). Brookhart (2017) provides a complementary discussion that producing individualized narrative comments at scale is labor-intensive, especially when schools wish to include not only academic performance but also extracurricular participation, leadership, and awards. The workload frequently leads to generic, low-utility remarks that offer limited guidance for improvement.

Recent advances in large language models (LLMs) make it feasible to draft short, professional feedback for subsequent human review, potentially streamlining the drafting stage (Song et al., 2024). However, Van der Lee et al. (2021) noted that education facing Natural Language Generation (NLG) must be assessed by human raters and governed rigorously, as automatic metrics (e.g., BLEU) correlate imperfectly with human

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2026
S. K. S. Cheung et al. (Eds.): ICTE 2025, CCIS 2752, pp. 108–120, 2026.
https://doi.org/10.1007/978-981-95-4499-8_10

judgment and are not suitable for scientific hypothesis testing. By focusing on explicit checkpoints and human oversight, this exploratory work seeks to demonstrate not only the efficiency gains offered by Generative AI (GenAI) but also its potential to enhance the quality and personalization of feedback in educational settings (Chen et al., 2024). Furthermore, governance guidelines from Holmes and Miao (2023) highlight that there are risks of hallucination, opacity, and misalignment with local pedagogical norms if systems are deployed without adequate safeguards. For the sustainable adoption of GenAI in education, robust institutional data practices are pivotal, as sufficient and high-quality data are fundamental for building accurate and effective AI models (Chui et al., 2024).

This paper addresses these gaps by studying an LLM-powered system that generates brief, professional feedback from structured extracurricular activity and award records. Our system takes normalized JSON records as input and uses a prompt framework with role instructions, genre guidelines, and few-shot samples to produce feedback that is suitable for teachers to review. Five open-source models were evaluated that executed locally via Ollama – DeepSeek-R1, Gemma3, GPT-OSS, Llama-3.1, and Qwen3 under identical prompting conditions. The empirical study samples of 100 students from a secondary-school dataset (1,067 activity entries across six categories). For each student, each model produces one comment (500 outputs total). Three calibrated professional raters assess every output on five dimensions adapted from NLG evaluation literature: faithfulness to evidence, coverage of salient information, fluency, coherence & clarity, and relevance & non-redundancy, with a five-point Likert scale (Reiter and Belz, 2009; Howcroft et al., 2020).

Our contributions are threefold:

- Task formulation and system design. Extracurricular-to-feedback generation is formulated as a constrained NLG task over structured, school-curated records and presents a reproducible pipeline that includes data retrieval, normalization to JSON, prompt construction with rule-guided constraints, and model-agnostic inference, all of which teachers can run locally.
- Comparative evaluation across open-source LLMs. Five widely used open-source models are compared under identical prompts and hardware, reporting dimension-wise outcomes and variability, thereby offering practical guidance on model selection for resource-constrained school deployments.
- Error analysis and governance implications. Recurring failure modes are documented, including coverage gaps on longer records, factual infidelity via fabricated achievements, translation issues for Chinese award names, and redundancies. Prompt rules, formatting checks, and human oversight are discussed as strategies to mitigate risks in school settings without requiring external retrieval components.

By centering extracurricular evidence and adopting a rubric-based human evaluation, this paper advances an underexplored but educationally salient use case for LLMs. The findings inform model choice, prompt design, and oversight practices for schools seeking to augment teacher workflows, while maintaining control and accountability.

2 Literature Review

2.1 Foundations of Effective Feedback

Effective feedback clarifies performance standards, closes the gap between current and desired performance, and supports motivation and self-regulation when delivered in a timely, specific, and actionable manner (Hattie and Timperley, 2007). Brookhart (2017) offers a related perspective. In large classes, teachers struggle to provide individualized narrative comments, and the workload pressures often degrade the specificity and usefulness (Brookhart, 2017). Agostini and Picasso (2024) also report similar observations. This motivates tooling that can draft personalized summaries to help accelerate the effectiveness of human review rather than replacing the teachers' judgment.

2.2 Automating Narrative Feedback with LLMs

LLMs have been explored as assistants for drafting educational feedback and assessment artifacts, with studies reporting feasibility in producing short comments that humans can refine (Song et al., 2024). Zhou et al. (2018) report comparable findings. Much work has focused on academic performance; comparatively fewer studies consider structured non-academic evidence (e.g., activity logs, roles, awards) that schools also value to it. Our study extends this line by conditioning generation on structured extracurricular records and comparing multiple open-source models.

2.3 Human Evaluation in Education Facing NLG

Human evaluation remains essential in education. Automatic metrics such as BLEU provide limited validity beyond machine translation, show weak correlations with human judgments, and should not substitute for hypothesis testing (van der Lee et al., 2021). Callison Burch et al. (2006) discuss the limits of automatic metrics. Accordingly, a rubric-based human assessment was adopted, targeting five dimensions: faithfulness to evidence, coverage of salient information, fluency, coherence and clarity, and relevance and non-redundancy. Each dimension was rated on a five-point Likert scale to balance granularity with rater workload.

2.4 Guardrails, Factuality and Governance

Hallucination is a persistent risk in abstractive summarization and in LLM outputs (Maynez et al., 2020). Sector guidance recommends transparency, data protection, and educator control (Holmes and Miao, 2023). Technical mitigations include retrieval augmented generation (RAG) to ground outputs in external sources (Lewis et al., 2020) and prompting strategies that explicitly align generation with structured evidence.

2.5 Summary and Research Gap

The literature establishes both the importance of individualized feedback and the necessity of human evaluation and governance for education facing NLG. What remains underexplored is how well different LLMs transform extracurricular records into concise, evidence-aligned feedback, and which quality dimensions constitute the primary failure modes at scale. This study therefore examines:

- RQ1: How does the proposed LLM-powered feedback generation system perform in generating personalized student feedback?
- RQ2: What are the strengths and weaknesses of different LLMs in generating student feedback?

3 Methodology

This study proposed an LLM-powered feedback generation system designed to automatically generate personalized feedback based on students' extracurricular activities and award records. While examination scores provide a quantifiable measure of academic achievement, extracurricular engagement and recognition represent complementary aspects of student development that are less amenable, compare to numeric evaluation. The proposed system addresses this gap by using LLMs to generate narrative feedback that captures students' contributions, achievements, and growth potential beyond academics.

3.1 System Overview

The system consists of three core modules: Data Management, Feedback Generation and Evaluation. The Data Management Module manages the structured collection, organization and storage of student activity and feedback records. The Feedback Generation Module, orchestrated by the Feedback Agent, retrieves activity data, reformats it into JSON, prepares a prompt, and generates feedback through multiple LLMs. Finally, the Evaluation Module ensures the quality and reliability of generated feedback through systematic human review. Figure 1 shows the overall system architecture.

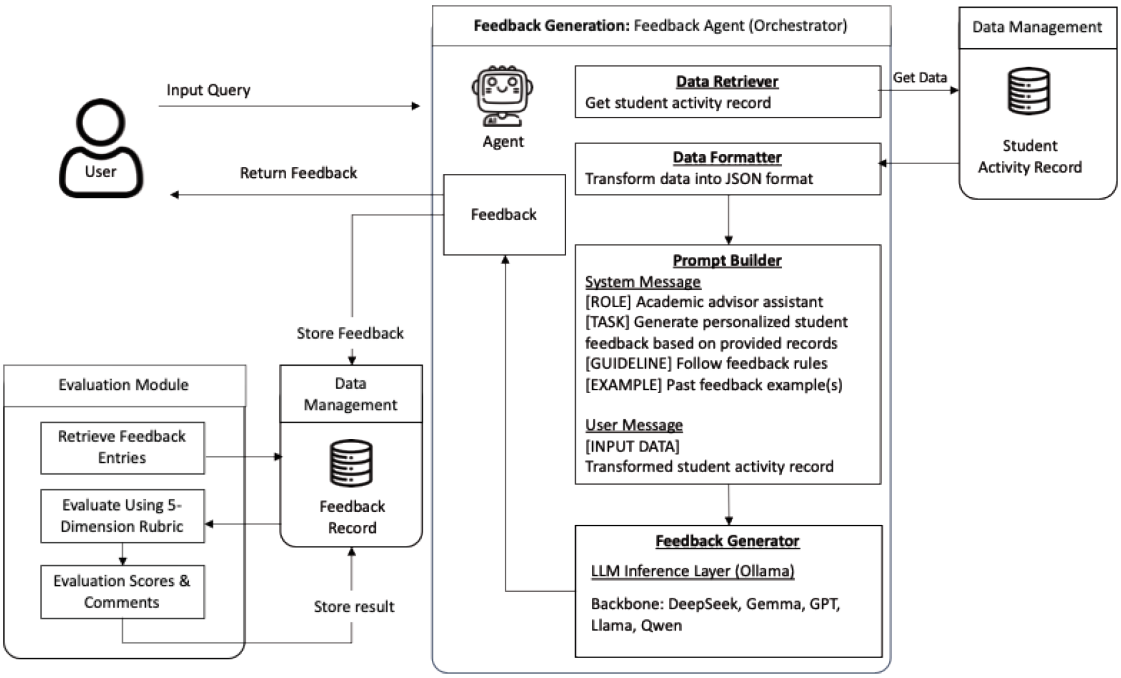


Fig. 1. System architecture of the LLM-Powered Feedback Generation System.

3.2 Feedback Generation

The feedback generation process is orchestrated by a Feedback Agent, which transforms raw extracurricular activities and award records into structured prompts and produces narrative feedback. The process includes four key components:

Data Retriever: Queries the student activity database to obtain relevant records.
Data Formatter: Normalizes and translates retrieved data into a JSON-compatible format.
Prompt Builder: Assembles the system and user messages, embedding role instructions, explicit writing guidelines and few-shot examples.
Feedback Generator: Executes the prompt using multiple LLMs via Ollama inference. Outputs are validated against formatting before being stored in the feedback repository.

3.2.1 Configuration

All models were executed locally via Ollama on a workstation running Windows Server 2025, equipped with an Intel® Core™ Ultra 9–285 K (3.70 GHz), 192 GB RAM, and an NVIDIA GeForce RTX 5090 GPU (32 GB VRAM).

3.2.2 Model Selection

Five open-source LLM variants were selected for Feedback Generator, summarized in Table 1. All models were executed locally through the Ollama inference framework. The selection was guided by their status as the most popular models on the Ollama platform, as indicated by official download statistics (Ollama, 2025).

- DeepSeek-R1 (32 B, 128 K): An open-source reasoning-oriented LLM.
- Gemma3 (27 B, 128 K): Google's lightweight multimodal model family, supporting multilingual inputs.

- GPT-OSS (20 B, 128 K): An open-source language model developed by OpenAI, designed for reasoning and agentic tasks.
- Llama-3.1 (8 B, 128 K): Meta's open-source model family with multilingual and reasoning capabilities.
- Qwen3 (32 B, 40 K): A dense and mixture-of-experts model with reasoning and multilingual capabilities.

Table 1. Open-source LLMs used in the inference layer, with download counts reported in the Ollama model library (Ollama, 2025)

	Model	Version	Parameter	Context Size	Downloads
1	Deepseek	Deepseek-R1	32 B	128 K	62.3 M
2	Gemma	Gemma3	27 B	128 K	16.8 M
3	GPT	GPT-OSS	20 B	128 K	2.4 M
4	Llama	Llama3.1	8 B	128 K	102.5 M
5	Qwen	Qwen3	32 B	40 K	8.3 M

3.2.3 Prompting and Generation Protocol

The prompting framework consisted of two parts: a system message that established global rules for the LLM and a user message that delivered student-specific data.

System Message.

The system message defined the role, task, and operational constraints for the LLM. Example 1 illustrates the structure:

- Role: Act as an academic advisor assistant.
- Task: Generate personalized narrative feedback based on student extracurricular activities and award records.
- Guidelines: Restrict the feedback to ≤ 100 words, adopt a professional and encouraging tone, translate all non-English content into English, and address aspects such as personality, initiative, peer relationships, and areas for improvement.
- Few-shot Exemplars: To guide style and structure, three authentic past comments previously written by teachers were embedded as examples in the system prompt.

Example 1. Excerpt of system prompt

```
You are an academic advisor assistant.
Your task is to generate personalized feedback for a student based on his/her extracurricular activities and award records.

***GUIDELINES***
- The feedback must be written in English and under 100 words.
- If any terms, award names, or descriptions are in Chinese, translate them to English.
- Feedback may cover: personality, relationship with peers, attitude towards studies, performance or service in class/school, areas of strength, and areas for improvement.
- The feedback must be clear, objective, and encouraging and in professional tone.

***EXAMPLES***
<example>
The student's steady contributions as class librarian reflect …
</example>

<example>
The student consistently engages in reading activities …
</example>

<example>
The student participates in the STEM robotics challenge…
</example>
```

User Message

The user message supplied the transformed student's extracurricular activities and award records in JSON format. Example 2 illustrates a sample of the transformed data.

Example 2. JSON-formatted user message.

```
<activity-record>
[
    {
     "category": "Activity",
     "semester": "First Semester",
     "roleOfParticipation": "Participant",
     "programDescription": "Internal activity relevant to Aesthetic and Physical activities"
    },
    {
     "category": "Competition",
     "roleOfParticipation": "Competitor",
     "programDescription": "External competition relevant to National Education",
     "awardsOrCertifications": "「閱讀之星」金獎"
    },
…
]
<activity-record>
```

3.3 Empirical Setup

To assess the quality of the generated feedback, a human evaluation of model outputs was conducted. A random sample of 100 students was drawn from the dataset. For each student, one feedback comment was generated by each of the five models, yielding 500 outputs in total. The sampled profiles varied in complexity, containing between 1 to 32 activity entries. All outputs were independently reviewed by human evaluators.

3.3.1 Dataset

A set of student activity records was obtained from a partner secondary school in Hong Kong. Data were collected over one academic year (2024 - 25) through the school management platform, which consolidates inputs from program organizers, class teachers, and extracurricular coordinators. Each entry included a unique student identifier, semester, activity category, program description, organizer of the program, role of participation, and award or certification details. Student names were anonymized and replaced with identifiers prior to analysis.

The final dataset contained 1,067 entries from 100 students across Forms 1 to 6. Table 2 shows the sample distribution. The number of activity records per student varied substantially, ranging from 1 to 32 entries. Activities were further grouped into six major categories, as summarized in Table 3.

Table 2. Sample distribution by form (F1 – F6)

F1	F2	F3	F4	F5	F6	Total
17	23	14	21	16	9	100

Table 3. Activity category distribution

No.	Category	Description	No. of Entries
1	Activity	Covers school-based and externally organized cultural, aesthetic, and physical activities	610
2	Community Service	Covers voluntary participation and service-learning activities	47
3	Competition	Covers internal and external contests and tournaments across domains	297
4	Course / Workshop	Covers short-term training programs, enrichment courses, or skill-building workshops	59
5	Study Tour	Covers structured local or overseas visits, cultural exchanges, and experiential learning opportunities beyond the classroom	52
6	Others	Covers extracurricular activity records that do not fall into the above categories	2

3.3.2 Evaluators

Three university lecturers served as evaluators. They were instructed to judge clarity and structure rather than apply subject-specific teaching criteria. Each evaluator assessed all 500 outputs. Before scoring, they completed a 30-min calibration using a two-page guide and five pilot items, discussing differences to align standards. Items were then shown in random order, with model identities masked to avoid brand effects. After calibration, evaluators scored independently without further discussion.

3.3.3 Rubric

The rubric dimensions were adapted from established evaluation frameworks in NLG (Reiter and Belz, 2009; Howcroft et al., 2020). Table 4 shows the details of the rubric, and each dimension was scored on a five-point Likert scale (1 = very poor, 5 = excellent). All dimensions are equally-weighted.

Table 4. Rubric dimensions

No.	Dimension	Description
1	Faithfulness	To measure the degree to which the content of an output is correct, accurate, or true relative to its input
2	Coverage	To quantify the amount of information conveyed by an output
3	Fluency	Refers to how well a text flows, ensuring it is not a sequence of unconnected parts
4	Coherence and Clarity	To assess the degree to which an output's content and meaning are presented in a well-structured, logical way and can be understood without effort
5	Relevance and Non-redundancy	To determine an output's usefulness for a given task or information need, while also checking that its content is free from redundant elements like repetition

4 Results

This section reports the outcomes of the evaluation across the five selected LLMs. Both quantitative ratings and qualitative evaluator feedback were analyzed.

4.1 Overall Model Performance

Table 5 presents the mean and standard deviation (SD) of scores across five evaluation dimensions. Among the five models, GPT achieved the highest overall mean (4.48 $\pm$ 0.64), followed closely by Gemma (4.45 $\pm$ 0.63). Qwen ranked third (4.24 $\pm$ 0.94), while DeepSeek and Llama obtained lower averages (4.04 $\pm$ 1.15 and 4.03 $\pm$ 0.80, respectively).

These results indicate that medium-sized models (20 B–27 B) provided the best trade-off between reasoning quality and inference stability under the given hardware setup. In contrast, the smaller 8 B Llama and the larger 32 B models showed greater variability, reflecting less consistent performance across records.

Table 5. Mean (±SD) evaluation scores across five models (Likert 1–5)

Model	Faithfulness	Coverage	Fluency	Coherence & Clarity	Relevance & Non-redundancy	Overall Mean
Deepseek	3.91 ± 1.15	3.48 ± 1.18	4.79 ± 0.43	4.60 ± 0.64	3.40 ± 1.13	4.04
Gemma	4.50 ± 0.63	4.20 ± 0.69	4.87 ± 0.35	4.69 ± 0.49	3.97 ± 0.62	4.45
GPT	4.68 ± 0.64	4.25 ± 0.66	4.86 ± 0.36	4.69 ± 0.54	3.93 ± 0.66	4.48
Llama	4.25 ± 0.80	3.69 ± 0.97	4.66 ± 0.55	4.54 ± 0.59	3.00 ± 0.83	4.03
Qwen	4.39 ± 0.94	3.80 ± 0.95	4.79 ± 0.44	4.64 ± 0.63	3.58 ± 0.87	4.24

4.2 Dimension-Wise Results

- Faithfulness: GPT (4.68 ± 0.64) and Gemma (4.50 ± 0.63) achieved the highest faithfulness scores, while DeepSeek obtained the lowest (3.91 ± 1.15). Qwen showed moderate performance but with greater variability.
- Coverage: Coverage was one of the weakest dimensions across all models, with scores ranging from 3.48 to 4.25. GPT achieved the highest mean (4.25 ± 0.66), while DeepSeek scored the lowest (3.48 ± 1.18).
- Fluency: All models scored consistently high in fluency (> 4.6). Gemma (4.87 ± 0.35) and GPT (4.86 ± 0.36) led this dimension, whereas Llama obtained the lowest score (4.66 ± 0.55), though still within a strong range.
- Coherence & Clarity: Scores for coherence and clarity were uniformly high, ranging between 4.54 and 4.69. GPT and Gemma tied at the top (4.69), while Llama was lowest (4.54 ± 0.59).
- Relevance & Non-Redundancy: This dimension yielded the lowest overall scores. Gemma (3.97 ± 0.62) and GPT (3.93 ± 0.66) achieved the highest ratings, while Llama (3.00 ± 0.83) ranked lowest.

4.3 Qualitative Error Patterns

Evaluator feedback highlighted recurring issues that align with the quantitative results:

- Hallucination: DeepSeek and Qwen occasionally fabricated achievements or inserted irrelevant boilerplate (e.g., generic academic advice), reducing faithfulness despite otherwise coherent prose.
- Translation: Gemma and Llama often left award names untranslated; Llama sometimes retained Chinese terms, which reduced clarity in English.
- Redundancy: Gemma tended to produce longer outputs with repetitive phrasing that added little informational value.
- Coverage: On longer activity records, models sometimes overlooked important participations or major awards. GPT generally handled extended records better, whereas DeepSeek more often omitted key events or defaulted to stock text.

5 Discussion

This section interprets the findings with respect to the two research questions and outlines practical implications and limitations.

5.1 RQ1: How Does the Proposed LLM-Powered Feedback Generation System Perform in Generating Personalized Student Feedback?

Across five models, the system consistently produced fluent and coherent comments, indicating that an LLM-assisted pipeline can efficiently draft short, professional feedback from structured extracurricular activity records. At the same time, coverage was the weakest dimension and faithfulness varied across items. In practice, this means the outputs read well but may omit salient events (e.g., major awards) or include details that do not fully align with the underlying records. Taken together, the results support positioning the system as a drafting aid rather than a stand-alone replacement: it can accelerate first-pass comment generation, provided that schools retain a human review step focused on completeness and factual alignment.

5.2 RQ2: What are the Strengths and Weaknesses of Different LLMs in Generating Student Feedback?

The models exhibited distinct profiles. GPT and Gemma delivered the most balanced performance, combining strong language quality with comparatively better grounding in records. Qwen achieved competitive means but with higher variance, occasionally introducing tangential or fabricated details. DeepSeek underperformed on faithfulness and coverage due to more frequent hallucinations and omissions. Llama, while lightweight, scored lowest on relevance, at times inserting generic academic advice or leaving award names untranslated. A practical takeaway is that medium-sized models ($\approx$20 B–27 B) offered the best trade-off between reliability and compute cost under our setup, whereas both the smaller (8 B) and larger (32 B) variants were less consistent.

5.3 Implications for Deployment and Governance

The error patterns suggest concrete safeguards for school use:

- Evidence anchoring: Require each comment to reference at least N concrete items (e.g., "top-3 achievements") from the record; flag drafts that fall below the threshold.
- Translation normalization: Maintain a controlled glossary for common award and activity names; enforce automatic checks to avoid untranslated or mixed-language terms.
- Role-appropriate scope: Constrain prompts to extracurricular contexts and block academic-advice templates unless explicitly present in the input.
- Human-in-the-loop review: Focus evaluator attention on coverage and faithfulness, since fluency/coherence are already strong and require less effort to verify. These measures can be implemented without external retrieval components and are compatible with on-premise deployment.

6 Conclusion

This study demonstrates the potential of an LLM-powered feedback generation system for producing personalized feedback on students' extracurricular activity records. The findings indicate that the system's value lies in its ability to rapidly produce polished, coherent, and encouraging narrative feedback. Model selection also plays a critical role in balancing reliability and computational feasibility. Medium-sized models (20 B–27 B) emerged as the most effective, outperforming both smaller and larger variants by combining reasoning capability with stability. The system is therefore best positioned not as a replacement for human evaluators, but as a supportive tool to standardize and scale feedback generation in educational settings.

Future research should focus on three key areas. First, mechanisms for automatic fact-checking and improved coverage are essential to further reduce the hallucinations and ensure completeness of feedback. This could involve integrating retrieval-augmented generation or domain-specific validation modules to cross-check activity descriptions against official school records. Second, broader evaluation with diverse assessor groups, including teachers and education professionals. Such evaluation will also help assess feedback quality, fairness, cultural appropriateness, and real-world usability in authentic school contexts. Third, adaptive prompting strategies and multimodal data integration should be explored to enhance personalization and contextual relevance; for example, tailoring prompts to student performance history or incorporating behavioral logs, portfolios, and attendance records could yield more nuanced and developmentally aligned feedback. These directions will strengthen the system's technical robustness and support its adoption as a practical innovation in technology-enhanced education.

References

Agostini, D., Picasso, F.: Large language models for sustainable assessment and feedback in higher education: Towards a pedagogical and technological framework. Intelligenza Artificiale **18**(1), 121–138 (2024)

Brookhart, S.M.: How to give effective feedback to your students. Ascd (2017)

Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the role of BLEU in machine translation research. In: 11th Conference of the European Chapter of the Association for Compuational Linguistics, pp. 249–256. Association for Computational Linguistics (2006)

Chen, Z., Cross, S., Rienties, B.: Empowering Assessors in Providing Quality Feedback with GenAI Assistance: A Preliminary Exploration. In: International Conference on Technology in Education, pp. 134–148. Springer Nature Singapore, Singapore (2024)

Chui, K.T., Lee, L.K., Wang, F.L., Cheung, S.K., Wong, L.P.: A review of data augmentation and data generation using artificial intelligence in education. In: International Conference on Technology in Education, pp. 242–253. Springer Nature Singapore, Singapore (2023)

Hattie, J., Timperley, H.: The power of feedback. Rev. Educ. Res. **77**(1), 81–112 (2007)

Holmes, W., Miao, F.: Guidance for generative AI in education and research. Unesco Publishing (2023)

Holmes, W., Tuomi, I.: State of the art and practice in AI in education. Eur. J. Educ. **57**(4), 542–570 (2022)

Howcroft, D.M., et al.: Twenty years of confusion in human evaluation: NLG needs evaluation sheets and standardised definitions. In: 13th International Conference on Natural Language Generation 2020, pp. 169–182. Association for Computational Linguistics (2020)

Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive nlp tasks. Adv. Neural. Inf. Process. Syst. **33**, 9459–9474 (2020)

Maynez, J., Narayan, S., Bohnet, B., McDonald, R.: On faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:2005.00661 (2020)

Ollama: Model library (2025). Retrieved 15 August 2025, from https://ollama.com/search

Reiter, E., Belz, A.: An investigation into the validity of some metrics for automatically evaluating natural language generation systems. Comput. Linguist. **35**(4), 529–558 (2009)

Shute, V.J.: Focus on formative feedback. Rev. Educ. Res. **78**(1), 153–189 (2008)

Song, Y., Zhu, Q., Wang, H., Zheng, Q.: Automated essay scoring and revising based on open-source large language models. IEEE Trans. Learn. Technol. **17**, 1880–1890 (2024)

Van der Lee, C., Gatt, A., Van Miltenburg, E., Krahmer, E.: Human evaluation of automatically generated text: current trends and best practice guidelines. Comput. Speech Lang. **67**, 101151 (2021)

Wan, T., Chen, Z.: Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. Phys. Rev. Phys. Educ. Res. **20**(1), 010152 (2024)

Xie, Y., Huang, L., Wang, Y.: The impact of AWE and peer feedback on Chinese EFL learners' English writing performance. In: International Conference on Technology in Education, pp. 258–270. Springer Singapore, Singapore (2020)

Zhou, Q., et al.: Neural question generation from text: A preliminary study. In: National CCF Conference on Natural Language Processing and Chinese Computing, pp. 662–671. Springer International Publishing, Cham (2017)