Delegating Educational Tasks to LLMs: A Content Analysis of Evaluation Approaches

Badmavasan Kirouchenassamy, Chloé Conrad, Maeva Somny, Léo Nebel

▶ **To cite this version:**

Badmavasan Kirouchenassamy, Chloé Conrad, Maeva Somny, Léo Nebel. Delegating Educational Tasks to LLMs: A Content Analysis of Evaluation Approaches. 27th International Conference on AI in Education, Jun 2026, Séoul, South Korea. ⟨hal-05625072⟩

HAL Id: hal-05625072

https://hal.science/hal-05625072v1

Submitted on 18 May 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC BY-NC-ND 4.0 - Attribution - Non-commercial use - No Derivative Works - International License

Delegating Educational Tasks to LLMs: A Content Analysis of Evaluation Approaches

Badmavasan Kirouchenassamy[1 [0009−0003−6502−154X]], Chloé Conrad[2 [0009−0003−5348−4459]], Maëva Somny[2 [0009−0006−1717−1107]], and Léo Nebel[1,3 [0000−0001−5859−265X]]

[1] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
{badmavasan.kirouchenassamy,leo.nebel}@lip6.fr
[2] Université Claude Bernard Lyon 1, CNRS, École Centrale de Lyon, INSA Lyon, Université Lumière Lyon 2, LIRIS, UMR5205, 69622 Villeurbanne, France
{chloe.conrad,maeva.somny}@univ-lyon1.fr
[3] EvidenceB, Paris, France

Abstract. LLMs have gone from theoretical computer science research to widespread use by the general public, with the emergence of increasingly powerful and popular LLMs models. These effective models have been incorporated into AIED research, but they raise several questions, since in this field, every prediction and every decision can have a significant impact on the learner. In this paper, we propose a study of LLMs' use in education, focusing on the evolution of the presence of these systems, their evaluation, and the tasks assigned to them. To this end, we conducted a meta-analysis, leading to the construction of a codebook based on the articles published from 2023 to 2025 editions of the *AIED* conference, ensuring agreement between annotators on a sample.

Keywords: Large Scale Language Models · Education · Content Analysis

1 Introduction

The landscape of educational technology is undergoing a "generative turn". Large Language Models (LLMs), with their transformer architectures and emergent reasoning capabilities, have moved beyond sophisticated auto-complete to serve as multifaceted instruments within the pedagogical ecosystem, facilitating functions such as learning, supporting content creation, dialogue-based learner assistance, assessment, and feedback [5]. However, this rapid integration has outpaced our theoretical understanding of its impact. Initially confined to low-stakes support functions—such as text summarization or grammar correction—LLMs are increasingly being delegated "High-Level Cognitive Tasks" (HLCTs). In education, these include instructional design, learner misconception modeling, and even acting as pseudo-tutors in collaborative learning environments [7]. This shift makes rigorous socio-technical evaluation all the more important: LLM outputs are probabilistic, sensitive to prompt and context variations, and prone to

generating fluent but nonfactual content–commonly known as "hallucinations" – which poses real risks to educational validity and trust. This article presents early findings from a broader, ongoing review of generative AI in education. This broader review addresses educational uses of LLMs, approaches for evaluating such uses, ethical and governance considerations, and prompting strategies for different educational uses.

We restrict the present synthesis to papers published in the *AIED* (Artificial Intelligence in Education) Conference over the most recent three-year window. This scope is motivated (*i*) temporally because widely accessible LLM systems entered public use at the end of 2022 and the first sustained wave of educational research followed in 2023 and (*ii*) venue-wise because *AIED* is a long-standing, field-defining conference at the intersection of AI and the learning sciences, and its proceedings provide a concentrated view of methodological norms and emergent application patterns in AI-in-education research. This scope might be expanded in future works. Given the identified need for rigorous evaluation methods for LLM systems in education, we narrow the focus to a foundational **research question** that structures the first phase of synthesis: ***How are different tasks assigned to LLMs in educational settings evaluated?***

2 Related Works

Recent syntheses have studied LLM's appearance in education. Position and survey-style papers outline opportunities and risks across student-facing support (explanations, dialogue tutoring, practice generation, writing support) and teacher-facing work (lesson/assessment authoring, feedback drafting), emphasizing the socio-technical implications of deploying probabilistic text generators in instructional settings [5, 4]. Scoping and systematic reviews further catalogue applications and recurring challenges (e.g., transparency/replicability, privacy, bias), often grouping findings by use-case families such as feedback, grading, content generation, and teaching support [11, 9, 3]. These syntheses also focus on specific application domains evaluation, especially assessment and tutoring. Reviews and empirical studies report mixed evidence on reliability/validity when LLMs grade or score open-ended work, and highlight sensitivity to prompts, rubrics, and rater disagreement [2, 6, 8]. In tutoring, recent work proposes pedagogy-grounded evaluation taxonomies and benchmarks, reflecting the lack of shared standards for assessing "instructional quality" beyond surface plausibility. Human-centered evaluation frameworks in adjacent fields stress multidisciplinary protocols and validity threats in human judgment of generative systems [10, 1].

But a key gap still remains: most reviews organize studies by application area rather than by what is actually delegated to the LLM. This is important because different delegated tasks require distinct evaluation designs and outcome measures. When studies with fundamentally different delegation choices are grouped, results are harder to compare and may not reflect genuine learning effects. A task-level lens makes the unit of analysis explicit, aligns evaluation

methods with the delegated function, and supports more educationally valid interpretations. Addressing this gap, our content analysis (restricted to the last three years of *AIED*) builds a task-level delegation taxonomy and synthesizes, by task, the evaluation approaches used

3 Methodology

3.1 Annotation Process

Because the application of LLM to education is both recent and rapidly diversifying, existing reviews organize the space from multiple angles (applications, risks, stakeholders, technical approaches) but do not yet provide a shared, ready-to-use taxonomy that directly operationalizes the constructs needed for our research question on the evaluation of the different tasks provided to LLMs. We have therefore developed a purpose-built coding scheme using a hybrid content-analysis strategy: categories were inductively derived and iteratively refined from the corpus, while individual labels and decision rules were anchored in established frameworks, such as canonical prompting and adaptation methods (see 3.2).

To this end, the authors formed a team of **four annotators** and followed an iterative codebook-development procedure common in systematic content analysis. First, we jointly read a small subset of papers to determine which variables were necessary to address the research questions and draft an initial codebook specifying constructs, admissible values, and explicit decision rules. We then conducted a broader pilot annotation on additional papers, during which we applied the draft codebook and logged ambiguities, missing categories, and recurrent edge cases. These observations informed successive refinements (merging/splitting categories, tightening definitions, and adding decision rules) until the coding scheme stabilized—i.e., newly sampled papers rarely prompted new categories and the rules could be applied consistently. To quantify reliability, we ran a two-round agreement assessment: in Round 1, we coded 24 of 86 papers in the 2025 subset, and we computed Cohen's κ for single-choice variables and Krippendorff's α for multi-label variables (alongside percent agreement, following reporting recommendations). We then held a structured calibration meeting to resolve disagreements, clarify definitions, and revise decision rules, to read all papers again with these adjusted guidelines, and repeated the procedure with the same codebook on a second sample of 24 papers to verify improved consistency. Second-round κ and α scores (and percent agreement) are reported for each variable in the next subsection. After this calibration-and-verification cycle, we applied the finalized codebook to annotate the full set of included papers (all papers from 2024 and 2023).

3.2 Coding Scheme

This section documents the operational definitions used in our annotation protocol. We grouped our annotations in three main categories: (A) Technical Architecture and Implementation, (B) Task, and (C) Evaluation (specifically to

answer RQ). Table 1 sum up all the categories, variables, and labels. In line with our research question, we confined our meta-analysis to variables that were both pertinent and supported by strong annotator agreement. Variables deemed non-essential for the questions at hand were omitted, as were those with unresolved annotator disagreement—an issue we interpret as reflecting insufficient definitional clarity. Some variables that were kept still yielded low κ or α scores while having an acceptable percentage of agreement. They represent potentially imbalanced categories that should be clarified. Moreover, our results for these few might be mitigated by this limitation (Category A: reproducibility, Category C: LLM Evaluation, surveys).

Category / Variable	**Definition & Rules**	**Labels (Values)**	**Reliability (κ/α & %)**
Category A: Technical Architecture & Implementation			
Model Used	Name of the model used.	*Free text*	
Adaptation Level	Level of adaptation to the task.	**Zero-Shot, Few-Shot, Fine-Tuning...**	$\alpha = 0.84$ - 83.3%
Deployment Mode	How is the model accessed and executed?	**Local, API, Unspecified**	$\kappa = 0.75$ - 91.7%
Reproducibility	Was the prompt given or the method detailed?	**Prompt, Method, Unspecified**	$\kappa = 0.58$ - 74.8%
Category B: Task			
Task Category	What is the category of the educational task that was delegated to LLM according to our taxonomy	**Evaluation, Generation, Chatbot, Other**	$\alpha = 0.72$ - 84.7%
Task Type	What educational task was delegated to LLM ?	**Scoring, Feedback...**	$\alpha = 0.64$ - 72.2%
Category C: Evaluation			
Evaluation Focus	What was the LLM evaluated on?	**Performance, Usefulness, Usability**	$\alpha = 0.84$ - 91.7%
Human evaluation	Were there any human evaluation? And if so, were they experts?	**Yes/No**	$\kappa = 0.75$ - 91.7% $\kappa = 0.81$ - 91.7%
LLM evaluation	Were there any LLM-as-judge? And if so, were they validated by experts?	**Yes/No**	$\kappa = 0.59$ - 87.5% $\kappa = 0.00$ - 91.7%
A/B Testing	Were there any A/B Testing protocol?	**Yes/No**	$\kappa = 1.0$ - 100%
Statistical tests	Were there any statistical tests ?	**Yes/No**	$\kappa = 0.83$ - 91.7%
Surveys	Were there any standardized survey used and if so, which one?	**Yes/No** *+ Free text*	$\kappa = 0.57$ - 83.3%
Datasets	Were there any external dataset used and if so, which one?	**Yes/No** *+ Free text*	$\kappa = 0.80$ - 91.7%
Metrics	What metrics were used?	*Free text*	

Table 1: Annotation Scheme Summary: Features and Reliability

Category A: Technical Architecture & Implementation captures implementation-level descriptors that can be coded directly from the methods: **(A1)** the model family used (verbatim model IDs and parameter scales were first recorded, then collapsed to families such as GPT/Llama/Mistral for analysis), **(A2)** the adaptation level describing how an off the shelf LLM was personalized for the use case (from prompt-based in-context learning and structured prompting to parameter-efficient and full fine-tuning) and, **(A3)** the deployment mode at inference time (API/hosted, local/self-hosted, or unspecified)

Category B: Task identifies the educational task(s) assigned to the LLM (coded as multi-label when applicable) in order to situate each study's evaluation within its functional intent. We also define a task-level delegation taxonomy: **Generation** (educational content generation, feedback...), **Evaluation** (assessment, educational content difficulty...), **Chatbots**, and **Other**. This category could also be coded as multi-label when necessary.

Category C: Evaluation records how studies evaluate LLM-based systems and is the primary basis for our **research question**: we coded an **Evaluation Focus** as **Performance** (output quality against a reference/criterion), **Usefulness** (perceived or measured educational value/outcomes), or **Usability** (interaction quality in terms of effectiveness/efficiency/satisfaction), and additionally captured structured metadata about the evaluation design (e.g., instruments, raters, agreement metrics, statistical tests).

Overall, the methodology combines data-driven codebook construction with iterative annotator calibration and inter-annotator agreement checks, yielding a coding scheme that is both grounded in the corpus and sufficiently operationalized for reproducible application.

4 Results

4.1 Tendencies and Global Observations

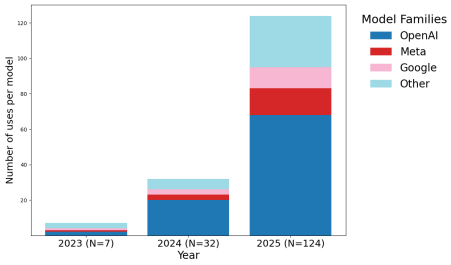

Fig. 1: Model supplier evolution across years

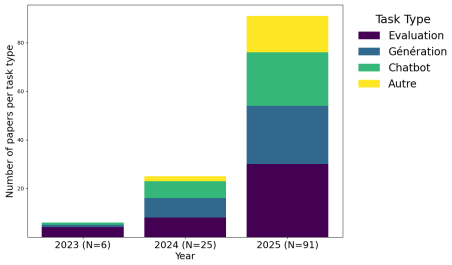

Fig. 2: Task type distribution evolution across years

A global examination of the results confirms a steady rise in the use of LLMs either as standalone systems or as components within systems over the last

three years in our research domain, increasing from 7 articles in 2023 (8.98% of articles) to 21 in 2024 (27.63%) and 86 in 2025 (34.40%). These figures exclude papers that discuss LLMs without actually deploying them, which also grow over time: from 1 (1.28%) in 2023 to 13 (17.11%) in 2024 and 53 (21.2%) in 2025. These results show that LLMs are a subject that is growing enormously in importance, accounting for half of all articles published in 2025. Regarding the specific models employed, Figure 1 presents their evolution across years. A total of 128 different models were used. While only 1 article used more than one model in 2023, this number rose to 9 in 2024 (42.9% of LLM-operating papers) and 29 in 2025 (33.7% of LLM-operating papers), enabling broader comparative evaluations across systems. We observe the strong predominance of OpenAI models (specifically coming from GPT-4o in the 2025 graph) in the last two years, along with the emergence of an increasingly diverse range of alternatives, captured in the expanding "Other" category. Regarding models' openness, $70,3\%$ of the 78 models used in our scope were proprietary, $21,9\%$ the models were open weights, and 7.8% were open source. 72.6% were online models only usable through APIs or interfaces, and 27.4% were downloadable models. Figure 2 shows how the tasks delegated to LLM evolved throughout the year according to our taxonomy. Although there are slightly more papers with an evaluation task, the proportions of the 3 task types remain balanced throughout the year.

4.2 Evaluation Methods

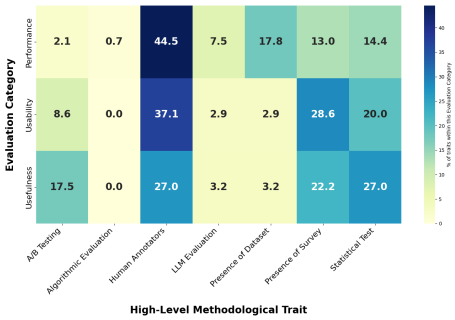

Fig. 3: Methodologies of evaluation used depending on evaluation category

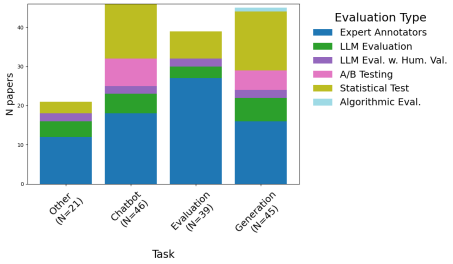

Fig. 4: Methodologies of evaluation per task delegating to the LLM system

Of the 115 papers that used LLM as or within a system, 19 evaluated it on multiple dimensions. Most assessed performance ($N = 86$, 74.8%), while fewer examined usefulness ($N = 26$, 22.6%) or usability ($N = 17$, 14.8%). Only three papers did not evaluate either the LLM or its host system.

When diving into details, 67.8% of articles evaluate their models through human annotation, among which 84.6% are experts' annotations (the others might

be from the learners or users of the system). 31.3% of articles performed statistical tests, 30% used surveys, 22.6% used existing datasets. LLM was used as a judge in 14 (12.2%) articles, and these evaluations were validated by humans in 6 cases. A/B testing protocol was used in 10.4% of studies. Figure 3 illustrates the prevailing methods used to evaluate the system, broken down by the performed evaluation types. However, this visualization has a limitation: several papers include multiple evaluation types, sometimes applying a method to only one aspect. For example, a study might assess both performance and usefulness, while A/B testing is used only for usefulness, yet it still appears on the performance line. Nearly half of the studies (44.5%) used human annotation to evaluate performance. When it was not the case, they often relied on existing datasets (17.8%). Usability shows an important survey uses (28.6%), usefulness of statistical tests (27.0%), and A/B Testing (17.5%).

Figure 4 shows how different tasks are evaluated. We can see that systems for Chatbot or Generation tasks are generally similarly assessed. For Evaluation tasks, there is a clear prevalence of expert labeling and an absence of A/B testing. This tends to indicate that, for this type of system (mostly classification systems), authors aim to develop systems focusing on performance metrics but do not consider the possible pedagogical impact these systems could have.

We can note that 28 studies (24.5%) did not evaluate model performance, focusing instead on the usefulness and/or usability aspects. This shift is particularly evident in studies utilizing proprietary models (89.3% of the 28 studies), where the research objective moves from benchmarking technical capabilities to assessing educational impact. These studies employed human-centric methodologies such as surveys and A/B testing, representing a "qualitative turn" in LLM-assisted education. In a sensitive area like education, the accuracy of systems and their performance in a given task seems to be a first and necessary assessment step. The fact that only 67% of the total corpus utilized either Expert Annotators or existing datasets **suggests a potential "validation gap" where the pedagogical safety of a system is assumed rather than tested**, especially when automated performance metrics are bypassed. A weakness of the LLMs lies in their exploratory nature, especially for proprietary models where results are harder to replicate. To try to answer this issue, we underline that 45.6% of articles detailed their prompting methods while 28.9% even gave an example of their prompt. Lastly, 30 papers (26.3%) did not propose any method or prompt to illustrate their approach, limiting reproducibility of the work.

An interesting methodological shift identified in the recent AIED proceedings is the emergence of the "LLM-as-judge" paradigm, which reflects a growing tension between scalability and pedagogical validity. Our analysis shows that 14 articles (12.3%) entrust the assessment of learner outputs or system performance to an LLM, yet this automation is not always subjected to rigorous oversight. Specifically, only 6 of these 14 integrate expert validation to verify alignment with educational standards. This trend reveals a growing validation gap, where system reliability is assumed rather than empirically proven. Without a robust "human-in-the-loop" framework to audit these automated judges, **we risk em-**

bedding algorithmic biases that remain invisible to both learners and educators.

5 Conclusion

This work on the evaluation of LLM-based systems raises important questions about the trust placed in such technologies, especially in education, where sensitive data is handled. Although LLMs are powerful and effective tools, it is crucial to remember that every decision or output they produce can significantly affect learners, as sustaining their motivation, confidence, and engagement is inherently challenging. Our work also shows how some studies try to keep "Expert-in-the-loop" and persist in rigorous evaluation protocols, evaluating performance, usability, and utilisability. The growing use of LLMs on HLCTs in education should still make us keep these conclusions in mind to ensure pedagogical safety.

In this paper, we present a preliminary overview of our work. Although we are focusing on a subset of our annotations, there is already much to be seen in the AIED publications of the last three years. We still explore other aspects of the corpus to evaluate more precisely the techniques used or how the use of LLM was compared to existing works. This global study aims to engage a discussion within our research community over our practices and the use of these technologies.

Acknowledgments. This work was partly supported by Génération 5 (Cifre PhD scholarship).

References

1. A. Elangovan, L. Liu, L. Xu, S. B. Bodapati, and D. Roth. Considers-the-human evaluation framework: Rethinking human evaluation for generative large language models. In *Proceedings of ACL'2024 (Volume 1)*, pages 1137–1160, 2024.
2. E. Emirtekin. Large language model-powered automated assessment: A systematic review. *Applied Sciences*, 15(10):5683, 2025.
3. S. Guizani, T. Mazhar, T. Shahzad, W. Ahmad, A. Bibi, and H. Hamam. A systematic literature review to implement large language model in higher education: issues and solutions. *Discover Education*, 4(1):1–25, 2025.
4. W. Holmes, F. Miao, et al. *Guidance for generative AI in education and research*. Unesco Publishing, 2023.
5. E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.
6. 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*, 6:100234, 2024.
7. S. Pal Chowdhury, V. Zouhar, and M. Sachan. Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 5–15, 2024.

8. H. Seo, T. Hwang, J. Jung, H. Kang, H. Namgoong, Y. Lee, and S. Jung. Large language models as evaluators in education: Verification of feedback consistency and accuracy. *Applied Sciences (2076-3417)*, 15(2), 2025.
9. Y. Shi, K. Yu, Y. Dong, and F. Chen. Large language models in education: a systematic review of empirical applications, benefits, and challenges. *Computers and Education: Artificial Intelligence*, page 100529, 2025.
10. T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy, H. Osterhoudt, X. Wu, S. Visweswaran, S. Fu, et al. A framework for human evaluation of large language models in healthcare derived from literature review. npj digital medicine, 7 (1), 2024.
11. L. Yan, L. Sha, L. Zhao, Y. Li, R. Martinez-Maldonado, G. Chen, X. Li, Y. Jin, and D. Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *BJET*, 55(1):90–112, 2024.