Alier-Forment, M., Pereira-Valera, J., Casañ Guerrero, M. J., & García-Peñalvo, F. J. (2025). Enhancing Learning Assistant Quality Through Automated Feedback Analysis and Systematic Testing in the LAMB Framework. In B. K. Smith & M. Borge (Eds.), Learning and Collaboration Technologies. 12th International Conference, LCT 2025 Held as Part of the 27th HCI International Conference, HCII 2025 Gothenburg, Sweden, June 22–27, 2025 Proceedings, Part II (pp. 3–12). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-031-93567-1_1

Enhancing Learning Assistant Quality Through Automated Feedback Analysis and Systematic Testing in the LAMB Framework

Marc Alier-Forment[1(✉)], Juanan Pereira-Valera[2], Maria Jose Casañ-Guerrero[1], and Francisco Jose Garcia-Penalvo[3]

[1] Universitat Politècnica de Catalunya, Barcelona, Spain
{marc.alier,ma.jose.casan}@upc.edu
[2] Universidad Pais Vasco EHUS, Donosti, Spain
[3] Universidad De Salamanca, Salamanca, Spain

Abstract. The Learning Assistant Manager and Builder (LAMB) is an open-source software framework that lets educators build and deploy AI learning assistants within institutional Learning Management Systems (LMS) without coding expertise. It addresses critical challenges in educational AI by providing privacy-focused integration, controlled knowledge bases, and seamless deployment through standard protocols. This paper presents major enhancements that enable systematic quality assurance and continuous improvement of these learning assistants.

The new LAMB includes mechanisms for structured feedback on real-world assistant behavior, transforming it into a test suite with curated prompts and expected correct or incorrect responses. When changes are made—such as prompt engineering, retrieval-augmented generation optimization, or knowledge base expansions—this suite enables automated validation of their impact.

A key innovation is using frontier large language models (LLMs) to evaluate responses automatically, generating detailed reports that reveal improvement areas and confirm performance gains. This systematic feedback-driven testing fosters continuous refinement while preserving quality standards.

Validation studies show measurable boosts in reliability and consistency. In various educational contexts, the framework identifies edge cases, maintains consistency across iterations, and provides actionable insights. Automated testing is especially beneficial for assistants with extensive knowledge bases and complex interaction patterns.

This work advances educational AI by providing a robust methodology for quality assurance and ongoing improvement of learning assistants. Its structured feedback and automated evaluations ensure alignment with educational goals while refining assistants over time. The enhanced LAMB framework offers a scalable and reliable solution for educators aiming to integrate AI-driven support into their LMS environments.

Keywords: Learning assistants · artificial intelligence in education · automated testing · quality assurance · continuous improvement · retrieval-augmented generation · prompt engineering · LLM Evals

1 Introduction

1.1 Learning Assistants Based on LLMs and RAG

The emergence of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) has transformed the educational landscape, creating new opportunities to enhance personalized learning experiences [1]. The widespread adoption of tools like ChatGPT has changed how we approach educational technology, with educators recognizing AI's potential to automate academic tasks, enhance student learning experiences, and provide personalized feedback [2].

Learning assistants powered by LLMs form a new category of educational software that combines advanced Natural Language Processing abilities, few-shot learning, chain-of-thought reasoning, and common-sense reasoning capabilities [3]. These assistants can understand complex queries, provide detailed explanations, and adapt to different learning contexts, making them valuable in educational settings.

However, LLM-based learning assistants face significant challenges. Response quality varies based on the training data's depth and breadth in specific domains [4], and AI "hallucinations" create risks in academic contexts where accuracy is critical [5]. There are also concerns about over-reliance on this technology affecting students' creativity and critical thinking skills [6].

To address these limitations, Retrieval-Augmented Generation (RAG) has become a key technology. RAG combines an LLM's ability to generate responses with information pulled from external databases or documents, improving the accuracy and relevance of responses [7]. This approach ensures that learning assistants can provide responses grounded in authoritative sources while maintaining the natural language processing and reasoning capabilities of LLMs [8].

The integration of learning assistants in educational settings must consider several critical factors, including privacy concerns [9–12], data protection laws, and adherence to institutional policies. These tools must also complement rather than replace human interactions, which remain vital for students' emotional and social growth [13].

The success of learning assistants in education depends on their ability to balance technological capabilities with teaching needs, ensuring they enhance rather than diminish the learning experience. This includes addressing concerns about assessment integrity [14], fostering critical thinking, and providing meaningful personalized support that aligns with educational objectives [15].

Learning assistants go beyond simple automation of tasks. When implemented with RAG technology, they can serve as effective tools for personalized learning, immediate feedback, and student engagement, while maintaining the connection to authoritative educational content and institutional standards.

1.2 The Learning Assistants Builder and Manager (LAMB) Project

Learning Assistants Manager and Builder (LAMB) is an innovative open-source software framework designed to create AI-powered learning assistants for integration into Learning Management Systems (LMS) [16]. LAMB addresses critical gaps in existing

educational AI solutions by providing a framework specifically tailored for education sector requirements.

The project's main goals include enabling educators to create AI-powered learning assistants without coding skills, ensuring seamless integration with institutional systems, and maintaining high standards of privacy and ethical AI usage [16]. LAMB emphasizes the importance of using authoritative sources and proper citations in its responses, addressing common concerns about AI accuracy and reliability in educational contexts [17].

Learning assistants created with LAMB become part of an institution's educational technology ecosystem through established educational standards. A key feature of LAMB is its implementation of the IMS Learning Tools Interoperability (LTI) standard [16], which allows learning assistants to integrate naturally into any modern Learning Management System like Moodle, Canvas, or Blackboard.

When teachers want to add a learning assistant to their course, they can do so through their familiar LMS interface. The process preserves the institution's authentication systems and privacy policies [12], while allowing the learning assistant to understand its educational context - who the users are, what roles they have, and what course they're working with [16].

The technical foundation of LAMB combines several modern AI technologies. At its core, the system leverages Large Language Models for natural language understanding and generation [18], but extends their capabilities through Retrieval-Augmented Generation. This RAG implementation ensures that learning assistants ground their responses in authoritative course materials rather than relying solely on the LLM's training data [19]. The system organizes educational content using embeddings databases, which enable semantic search and contextually relevant information retrieval [20].

LAMB's deployment options consider the varied needs and resources of educational institutions. The system can run on standard servers without specialized hardware requirements, and institutions can choose between commercial LLM providers like OpenAI or Anthropic, or opt for open-source models like Llama or Mistral [16]. This flexibility helps balance performance needs with budget constraints.

The framework has proven its effectiveness in real educational settings. Beyond the initial Macroeconomics Study Coach implementation, LAMB has supported various learning scenarios, from helping students analyze business cases to providing teaching assistants for faculty training [16]. These implementations have shown that learning assistants can meaningfully enhance teaching while maintaining high standards of privacy and ethical AI usage [22, 23].

The technical architecture of LAMB prioritizes security and scalability through containerization and modular design [16]. This approach ensures that as educational technology evolves, institutions can adapt and extend their learning assistants to meet new pedagogical needs and technological opportunities [24].

The framework has been validated through several real-world implementations. The "Macroeconomics Study Coach" served as an initial case study, effectively integrating

lecture transcriptions and course materials to support student inquiries [16]. Another successful implementation involved creating an expert-based learning assistant for analyzing technology-driven business projects, which showed improved student performance in understanding and applying complex methodologies [16].

Cost analysis of LAMB implementations has shown the system to be financially viable for educational institutions. The framework's design allows for flexibility in choosing between commercial and open-source LLMs, enabling institutions to balance costs with performance needs [21]. The system can run on standard servers without requiring specialized hardware, making it accessible to a wide range of educational institutions [16]. Specially with the publication of new frontier level models distilled to 70B sizes like Llama 3.3 7b, or DeepSeek-R1 70B that can be integrated with LAMB and run on.

One of LAMB's key achievements has been demonstrating that learning assistants can enhance teaching methodologies while maintaining high standards of privacy and ethical AI usage. The framework's focus on institutional policy compliance and use of authoritative sources has helped address common concerns about AI in education [22, 23]. Additionally, LAMB's integration capabilities with existing LMS platforms through standard protocols have made it a practical solution for educational institutions looking to implement AI-powered learning tools [24].

2 Automated Feedback Analysis and Systematic Testing in the LAMB Framework

2.1 LLM Evaluations

Large Language Model evaluation frameworks (LLM Evals) aim to systematically assess and validate the performance, reliability, and behavior of LLM-based systems.

The primary goals of LLM evaluations are to:

1. Verify response accuracy and factual correctness
2. Ensure consistent behavior across multiple interactions
3. Validate adherence to specified guidelines and constraints
4. Measure improvements across system iterations
5. Identify potential failure modes and edge cases

Traditional metrics like BLEU and ROUGE, while useful for general natural language generation tasks, often prove insufficient for evaluating complex interactive systems like learning assistants [25]. These metrics primarily focus on lexical similarity but may miss important aspects like factual accuracy, contextual appropriateness, and instructional effectiveness.

The evaluation of LLM-based systems presents unique challenges due to their probabilistic nature and context-dependent behavior [26]. A key consideration is that responses can vary significantly even for identical prompts, making deterministic testing approaches inadequate. Additionally, the evaluation must consider not just the correctness of individual responses, but also the consistency of behavior across multiple interactions [27].

LLM Evals typically involve creating test suites that combine reference prompts with expected responses and evaluation criteria. Recent advances in evaluation approaches

have introduced the concept of using frontier LLMs themselves as evaluators [28]. This approach leverages the sophisticated understanding capabilities of advanced models to assess responses along multiple dimensions, providing more nuanced evaluation than traditional metrics alone.

The systematic collection and analysis of test results enables both quantitative and qualitative assessment of model performance. This data-driven approach helps identify patterns in model behavior, highlight areas needing improvement, and validate that changes to the system result in measurable improvements [29].

For educational applications, evaluation frameworks must additionally consider pedagogical effectiveness, alignment with learning objectives, and appropriateness for the target audience [30]. This requires specialized evaluation criteria that go beyond general language model metrics to assess educational value and impact.

The development of robust evaluation frameworks is particularly critical for systems that will be deployed in educational institutions, where reliability and adherence to educational standards must be consistently maintained [31]. Systematic evaluation helps ensure that AI-powered educational tools remain aligned with pedagogical goals while becoming increasingly refined through iterative improvement. ¡

2.2 Evaluation Data Gathering Framework

The LAMB system incorporates a comprehensive evaluation data gathering framework that leverages Open WebUI's feedback interface to collect and analyze model performance data. This framework enables systematic collection of user interactions, feedback, and performance metrics to support continuous improvement of learning assistants.

The evaluation framework consists of several key components:

1. Feedback Collection Interface: The system utilizes Open WebUI's native feedback mechanism, which allows users to provide both quantitative and qualitative feedback on model responses. Users can rate responses positively or negatively and provide detailed comments explaining their assessment. This feedback is particularly valuable as it captures the immediate usefulness and accuracy of responses from the users' perspective.
2. Dataset Creation and Management: The framework includes a robust dataset management system that enables:
    - Creation of evaluation datasets from collected interactions.
    - Storage of prompt-response pairs along with associated feedback.
    - Organization of datasets by specific topics or evaluation criteria.
    - Export capabilities for external analysis.
    - Ability to edit, duplicate, and refine dataset entries.
3. Model Evaluation Pipeline: The evaluation pipeline allows systematic testing of models against curated datasets. Key features include:
    - Ability to run complete datasets against specific model configurations.
    - Support for comparing responses across different model versions.
    - Integration with established metrics including BLEU and ROUGE scores.
    - Calculation of both aggregate and per-response performance metrics.

4. Advanced LLM-Based Response Analysis: The system employs a sophisticated approach to response evaluation by leveraging large language models as semantic comparison tools. This goes beyond traditional metrics by enabling:

   - Semantic Understanding: The evaluation LLM analyzes responses for semantic equivalence rather than just lexical similarity. This means it can recognize when different phrasings convey the same meaning, even if they use entirely different words.
   - Contextual Evaluation: The LLM considers the broader context of both the prompt and the response, understanding nuances and implicit information that might be missed by traditional metrics.
   - Multi-dimensional Analysis: The evaluation LLM assesses responses across multiple dimensions:

     – Factual accuracy
     – Contextual relevance
     – Logical coherence
     – Completeness of information
     – Appropriateness of tone and style

   - Nuanced Feedback Generation: The LLM can provide detailed explanations of why responses do or don't meet expectations, highlighting specific strengths and weaknesses.

   The framework supports iterative refinement through a structured workflow:

1. Collection of user interactions and feedback through the Open WebUI interface.
2. Curation of evaluation datasets from collected data.
3. Systematic evaluation of model performance using established metrics.
4. Deep semantic analysis using the evaluation LLM.
5. Integration of findings into model refinement process.

   The LLM-based evaluation process works as follows:

1. Initial Comparison: For each prompt-response pair, the evaluation LLM receives:

   – The original prompt
   – The expected response from the dataset
   – The actual response generated by the model being tested
   – Any context or special requirements

2. Analysis Phase: The LLM performs a multi-stage analysis:

   – First, it assesses whether the generated response captures the core meaning of the expected response
   – Then, it evaluates additional aspects like accuracy, completeness, and appropriateness

   – Finally, it considers any specific requirements or constraints from the learning context

1. Scoring and Feedback: The LLM provides:
   – A numerical score for different aspects of the response
   – Detailed explanations of its assessment
   – Specific suggestions for improvement
   – Identification of any critical issues or missing elements

This sophisticated evaluation approach allows LAMB to assess responses in a way that better mirrors human judgment, understanding that there can be multiple valid ways to express the same information or answer a question. The LLM-based evaluation is particularly valuable for:

- Complex Responses: Where simple metric-based comparison would miss nuanced differences or similarities
- Educational Contexts: Where the way information is presented is as important as the information itself
- Domain-Specific Evaluation: Where responses need to be assessed against field-specific criteria or standards

Results from this evaluation framework have proven particularly valuable in assessing how well learning assistants maintain context in extended conversations and how accurately they incorporate information from their knowledge bases. The combination of traditional metrics, user feedback, and LLM-based semantic analysis provides a comprehensive picture of assistant performance that would be impossible to achieve with any single approach.

The evaluation data gathering framework has been instrumental in validating improvements to LAMB's learning assistants, particularly in areas such as response accuracy, contextual relevance, and adherence to institutional guidelines. This systematic approach to evaluation ensures that developments in the system are driven by empirical evidence rather than subjective assessment alone.

3 Conclusions and Future Work

This paper has presented LAMB (Learning Assistant Manager and Builder), an innovative open-source framework for creating and evaluating AI-powered learning assistants, along with a comprehensive evaluation system for assessing their performance. The research makes several significant contributions to the field of educational technology and AI-powered learning systems.

First, LAMB demonstrates that it is possible to create a framework that enables educators to develop and deploy AI-powered learning assistants without requiring programming expertise. The system's integration with standard LMS platforms through LTI protocols shows that AI assistants can be seamlessly incorporated into existing

educational technology ecosystems while maintaining institutional privacy and security requirements.

Second, the evaluation framework developed for LAMB provides a robust methodology for assessing learning assistant performance. By combining traditional metrics like BLEU and ROUGE with user feedback collection and structured dataset management, the system enables both quantitative and qualitative assessment of assistant responses. This dual approach has proven particularly valuable in educational contexts, where both technical accuracy and pedagogical effectiveness must be considered.

The implementation of Retrieval-Augmented Generation (RAG) within LAMB has demonstrated significant improvements in response accuracy and relevance. Our evaluation results show that RAG-enhanced models consistently outperform base models in maintaining factual accuracy and providing responses grounded in authoritative course materials. This is particularly evident in the comparative analysis of models with and without Knowledge integration, where metrics showed substantial improvements in response quality.

The framework's modular architecture and use of Docker containerization ensure that institutions can deploy and scale the system according to their needs. The flexibility to use either commercial or open-source LLMs provides institutions with cost-effective options while maintaining performance standards. This adaptability, combined with the system's minimal hardware requirements, makes LAMB an accessible solution for a wide range of educational institutions.

Real-world implementations, including the Macroeconomics Study Coach and expert-based learning assistants, have validated the system's effectiveness in diverse educational contexts. User feedback and performance metrics from these implementations demonstrate that LAMB-created assistants can effectively support student learning while maintaining high standards of accuracy and reliability.

Future work should focus on several key areas:

- Expanding the evaluation framework to include additional metrics specific to educational effectiveness.
- Developing more sophisticated feedback collection mechanisms to capture nuanced aspects of learning assistant performance.
- Implementing advanced analytics to identify patterns in user interactions and assistant responses.
- Exploring the integration of emerging LLM technologies as they become available.
- Extending the framework's capabilities to support multiple languages and diverse educational contexts.

The open-source nature of LAMB encourages collaborative development and continuous improvement, ensuring that the framework can evolve alongside advances in AI technology and educational practices. As AI continues to transform education, frameworks like LAMB will play an increasingly important role in ensuring that these technologies are deployed effectively, ethically, and in service of improved learning outcomes.

In conclusion, LAMB represents a significant step forward in making AI-powered learning assistants accessible and effective for educational institutions. The combination of robust development tools, seamless integration capabilities, and comprehensive evaluation frameworks provides a solid foundation for the future of AI in education.

Funding and Acknowledgements. The authors give thanks and acknowledge the grad student Joel Corredor for his contribution to the lamb project in the evals implementation. This research is partially funded by the Ministry of Science and Innovation through the AvisSA project (reference PID2020-118345RB-I00), by the Department of Research and Universities of the Catalan Government through the 2021 SGR 01412 grant for research groups, and by the University of the Basque Country/Euskal Herriko Unibertsitatea under contract GIU21/037 as part of the "Call for Grants for Research Groups at the University of the Basque Country/Euskal Herriko Unibertsitatea (2021).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. OpenAI.: GPT-4 Technical Report. arXiv:2303.08774v4 (2023)
2. Crawford, J., Cowling, M., Allen, K.A.: Leadership is needed for ethical ChatGPT: character, assessment, and learning using artificial intelligence (AI). J. Univ. Teach. Learn. Pract. **20**(3), 1–12 (2023)
3. Levesque, H., Davis, E., Morgenstern, L.: The winograd schema challenge. In: 13th International Conference on Principles of Knowledge Representation and Reasoning, pp. 552–561. AAAI Press, USA (2012)
4. Nazir, A., Wang, Z.: A comprehensive survey of ChatGPT: advancements, applications, prospects, and challenges. Meta-Radiology **1**(2), 100022 (2023)
5. Thorp, H.H.: ChatGPT is fun, but not an author. Science **379**(6630), 313 (2023)
6. Dwivedi, Y.K., et al.: So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int. J. Inf. Manag. **71**, 102642 (2023)
7. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Larochelle, H., et al. (eds.) Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474. Curran Associates Inc. (2020)
8. Gao, Y., et al.: Retrieval-augmented generation for large language models: a survey. arXiv:2312.10997v5 (2024)
9. Llorens-Largo, F., García-Peñalvo, F.J.: La inteligencia artificial en el gobierno universitario. Universidad (2023)
10. Lim, W.M., et al.: Generative AI and the future of education: ragnarök or reformation? A paradoxical perspective from management educators. Int. J. Manag. Educ. **21**(2), 100790 (2023)
11. Wang, T., et al.: Security and privacy on generative data in AIGC: a survey. arXiv:2309.09435v2 (2023)
12. Alier, M., Casañ, M.J., Amo, D., Severance, C., Fonseca, D.: Privacy and e-learning: a pending task. Sustainability **13**(16), 9206 (2021)

13. Choi, E.P.H., Lee, J.J., Ho, M.H., Kwok, J.Y.Y., Lok, K.Y.W.: Chatting or cheating? The impacts of ChatGPT and other artificial intelligence language models on nurse education. Nurse Educ. Today **125**, 105796 (2023)
14. Cotton, D.R.E., Cotton, P.A., Shipway, J.R.: Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov. Educ. Teach. Int. **61**(2), 228–239 (2024)
15. García-Peñalvo, F.J.: Generative artificial intelligence and education: an analysis from multiple perspectives. Educ. Knowl. Soc. **25**, 31942 (2024)
16. Alier, M., Pereira, J., García-Peñalvo, F.J., Casañ, M.J., Cabré, J.: LAMB: an open-source software framework to create artificial intelligence assistants deployed and integrated into learning management systems. Comput. Stand. Interfaces **92**, 103940 (2025)
17. Gašević, D., Siemens, G., Sadiq, S.: Empowering learners for the age of artificial intelligence. Comput. Educ. Artif. Intell. **4**, 100130 (2023)
18. Zhao, W.X., et al.: A survey of large language models. arXiv:2303.18223v13 (2023)
19. Lewis, P., et al.: Retrieval-augmented generation for knowledge-intensive NLP tasks. In: Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474 (2020)
20. Sahoo, P., et al.: A systematic survey of prompt engineering in large language models: techniques and applications. arXiv:2402.07927v1 (2024)
21. Yao, Y., et al.: A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly. High-Confidence Comput. **4**(2), 100211 (2024)
22. Triberti, S., Di Fuccio, R., Scuotto, C., Marsico, E., Limone, P.: Better than my professor? How to develop artificial intelligence tools for higher education. Front. Artif. Intell. **7**, 1329605 (2024)
23. Alier, M., García-Peñalvo, F.J., Camba, J.D.: Generative artificial intelligence in education: from deceptive to disruptive. Int. J. Interact. Multimedia Artif. Intell. **8**(5), 5–14 (2024)
24. Su, J., Yang, W.: Unlocking the power of ChatGPT: a framework for applying generative AI in education. ECNU Rev. Educ. **6**(3), 355–366 (2023)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318. ACL Press, Philadelphia (2002)
26. Zheng, L., Gao, J., Ding, N., Liu, Z.: Challenges and opportunities in LLM-based assessment. arXiv:2402.12287v1 (2024)
27. Liu, H., Guo, T., Fang, H., Li, X.: On the evaluation of large language models: a framework and best practices. arXiv:2403.00285v1 (2024)
28. Zheng, C., et al.: A survey on evaluation of large language models. arXiv:2307.03109v5 (2024)
29. Aspillaga, C., Bhargava, R., Goel, A.: Test case generation using large language models. In: Proceedings of the 10th International Conference on Learning Analytics and Knowledge, pp. 1–10. ACM, New York (2024)
30. García-Peñalvo, F.J., Llorens-Largo, F., Vidal, J.: The new reality of education in the face of advances in generative artificial intelligence. RIED: Revista Iberoamericana de Educación a Distancia **27**(1), 9–39 (2024)
31. Wu, J., et al.: Assessment criteria for AI-powered educational tools: a systematic review. Comput. Educ. **185**, 104767 (2024)