Intelligent Human Systems Integration (IHSI), Vol. 200, 2026, 697–705
https://doi.org/10.54941/ahfe1007128

A Unified Multimodal Pipeline for Luxembourgish Language Learning: Curriculum-Grounded Retrieval and LAM-Driven Interaction

Hedi Tebourbi[1], Sana Nouzri[1], Piotr Kluczynski[2], Yazan Mualla[3], and Abdeljalil Abbas-Turki[3]

[1]University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg

[2]Bialystok University of Technology, Wiejska 45A, 15-351, Bialystok, Poland

[3]Université de Technologie de Belfort Montbéliard, UTBM, CIAD UR 7533, F-90010, Belfort, France

ABSTRACT

We introduce a unified multimodal system that transforms the official INL Luxembourgish textbook into an interactive, AI-driven tutoring environment. The work combines two complementary lines of research: (1) a pipeline for digitizing, structuring, and retrieving textbook exercises; and (2) a Large Action Model (LAM)–based interaction layer that enables a large language model (LLM) tutor to surface relevant visuals dynamically during conversation. The data pipeline begins with a semi-automated exercise extraction stage, where textbook pages are processed through a computer-vision and GPT-4.1 Vision–assisted workflow to isolate individual exercises, followed by targeted manual correction when needed. Extracted images are then passed through an OCR and cleanup stage, segmented into coherent units, enriched with metadata (chapter, theme, exercise type), and embedded into a vector store. This produces a structured, searchable, curriculum-grounded knowledge base. On top of this representation, a LAM-based Image Retrieval Tool, implemented using LangGraph orchestration and OpenAI function calling, allows the tutor agent to issue tool calls that retrieve pedagogically aligned images from a vector store of textbook illustrations. The backend fuses meta-parameters and multimodal embeddings to identify the most relevant image, which is rendered directly in the learner interface alongside the tutor's text. This design moves beyond static or manually curated visuals towards dynamic, curriculum-aligned multimodal interaction. Evaluation indicates that the retrieval mechanism consistently returns accurate, relevant images, and that most latency stems from LLM generation rather than retrieval or rendering. Pilot users report that multimodal support improves clarity, engagement, and perceived trust. Therefore, the combined pipeline offers a practical blueprint for curriculum-grounded, multimodal language learning systems and highlights future work on interactive exercises, large-scale evaluation, and further optimization of LAM-driven human–agent interaction.

Keywords: Luxembourgish language learning, Multimodal tutoring, Large action models, Function calling, Retrieval-augmented generation, Human–agent interaction, Human-centred design

Received November 19, 2025; Revised December 2, 2025; Accepted December 17, 2025; Available online February 1, 2026

INTRODUCTION

Modern language-learning platforms increasingly rely on large language models (LLMs) to adapt content and feedback to learners. However, aligning visual material, such as exercise figures and task sheets, with dynamically generated tutoring remains underexplored. Visual context improves comprehension and engagement, yet most LLM-driven tutors still present images through static links or manual curation (Mayer, 2009).

Recent work on multi-agent systems for Luxembourgish learning (Tebourbi et al., 2025a) and on Large Action Models (LAMs) (Zhang et al., 2024) shows that tool-capable agents can strengthen Human–Agent Interaction (HAI) by retrieving multimodal content during dialogue rather than producing text-only responses. Surveys further highlight the increasing role of LLMs in adaptive tutoring and educational support (Wang et al., 2024).

This work integrates a LAM with function calling into an existing Luxembourgish learning application to automatically retrieve and display exercise images from the official INL textbook (INL, 2021). The goals are to: (i) surface relevant images in sync with the instructional flow, (ii) preserve conversational continuity, and (iii) maintain a reliable mapping between OCR-extracted exercises and their associated visuals.

The system follows a three-tier architecture: a FastAPI backend with LangGraph-based multi-agent orchestration (LangChain Inc., 2024a); a Chroma vector store embedding OCR text and metadata (Chroma, 2024); and a React front end that renders images on demand. The main contributions are:

- A semi-automated data pipeline combining computer-vision-based exercise cropping with multi-stage OCR, semantic chunking, and metadata enrichment to produce a structured, curriculum-grounded vector store.
- A function-calling integration enabling dynamic image retrieval aligned with textbook content, using a dual-state design that separates user-facing messages from lesson content with image metadata.
- An implementation supporting multi-image lessons with robust fallbacks, and systematic evaluation demonstrating retrieval accuracy, latency characteristics, and positive user perception.

Unlike prior systems relying on static or pre-curated visuals, this approach enables dynamic, curriculum-grounded multimodal tutoring, improving pedagogical alignment and interaction quality.

This work builds on a broader research programme on AI-powered Luxembourgish language learning. Earlier work introduced the core multi-agent architecture and personalization pipeline (Tebourbi et al., 2025a), followed by a BPMN-based workflow formalization that added process transparency, explainability, and RAG-grounded validation (Tebourbi et al., 2025b). Parallel research explored LAM-driven multimodal interaction to augment the tutor agent with image-centric tool-use capabilities (Kluczyński et al., 2025). The present paper unifies these lines by introducing a full OCR-to-retrieval multimodal pipeline that grounds image delivery in authoritative curriculum materials and integrates it seamlessly into the existing multi-agent tutoring architecture.

RELATED WORK

AI tutors and adaptive learning. Intelligent tutoring systems have long leveraged AI to personalise exercises and feedback (Xu and Wang, 2006; Kim, 2020; Xiao et al., 2022). Recent work integrates LLMs to generate adaptive exercises, explanations, and assessments (Wang et al., 2024), though most systems focus on text-only interaction with limited exploration of real-time visual alignment. For Luxembourgish, our prior work established a foundational multi-agent architecture for personalized tutoring (Tebourbi et al., 2025a; Nouzri et al., 2025), formalised workflows using BPMN integrated with RAG (Tebourbi et al., 2025b), and demonstrated LAM-driven multimodal augmentation (Kluczyński et al., 2025). The present paper unifies these contributions by introducing a complete OCR-to-retrieval multimodal pipeline grounded in authoritative curriculum materials.

Retrieval-augmented generation. RAG has become a cornerstone for improving factuality in LLM applications (Lewis et al., 2020), with vector stores like Chroma enabling semantic search across embedded text and metadata (Chroma, 2024). This work extends the paradigm from text to image retrieval, indexing cropped exercise images alongside OCR text and metadata for multimodal alignment.

Function calling and multimodal learning. Tool use and function calling allow LLMs to interface with external systems for structured retrieval (Eleti et al., 2023), underpinning advances such as operator-style and computer-using agents (OpenAI, 2025a; 2025b; 2025c). Research on multimedia learning suggests that pairing text with images reduces cognitive load and supports retention (Mayer, 2009). While educational chatbots and commercial systems like Duolingo integrate multimodal support (Mageira et al., 2022; Duolingo Inc., 2023), most rely on pre-curated visuals. In contrast, this system operationalises dynamic multimodal integration, automatically surfacing textbook images through retrieval rather than manual design.

SYSTEM DESIGN AND METHODOLOGY

The system is organised as a three-stage pipeline that transforms static textbook exercises into interactive, dynamically retrievable resources.

Stage 1: Exercise cropping. Page scans from the INL textbook are processed to isolate individual exercises using a semi-automated approach combining computer-vision-based text region detection with classification to identify exercise headers. Due to layout complexity, some manual correction is required. The output is a collection of exercise-level images aligned with the textbook structure.

Stage 2: OCR and vector store preparation. Optical character recognition extracts exercise text, which is cleaned, segmented into coherent chunks, and enriched with metadata (chapter, theme, exercise type) before embedding into a vector store. This provides the semantic foundation for accurate retrieval of both textual and visual content.

Stage 3: Function calling integration. Within the LangGraph multi-agent framework, agents query the vector store and receive text with associated image metadata. Through LAM-style function calling, this metadata is passed to the frontend, which dynamically renders exercise images in sync with conversational tutoring.

This staged architecture maintains robustness and extensibility, errors in earlier stages can be corrected without disrupting later ones, and new features can be layered on top in future iterations.

IMPLEMENTATION

Stage 1: Exercise cropping. A Differentiable Binarization (DB-50) model with ResNet backbone detects text regions on each page (Liao et al., 2019; He et al., 2016). GPT-4.1 Vision classifies bounding boxes as exercise headers or non-exercise regions, and OpenCV (OpenCV, 2024) crops images accordingly. Classification errors on complex layouts require manual correction, making the system semi-automated but significantly faster than fully manual segmentation.

Stage 2: OCR and vector store preparation. GPT-4 Vision extracts Markdown-formatted exercise text preserving Luxembourgish diacritics. A post-processing pipeline standardises headings and restructures dialogues, followed by GPT-based semantic splitting into coherent chunks. A second prompting step assigns structured metadata (Kapitel, Thema, category, responsible agent). Embeddings are generated with BAAI bge-large-en (BAAI, 2023) and stored in ChromaDB with timestamp-based versioning, following the RAG paradigm (Lewis et al., 2020).

Stage 3: Function calling integration. A tool named getChunks queries Chroma with semantic similarity, returning both text and image metadata (image path, chapter, theme, source). If the LLM returns insufficient chunks despite multiple available images, a fallback algorithm searches for exercise indicators to split content further. A second tool, getLearningContent, produces lessons enriched with inline metadata. The backend exposes REST endpoints for retrieving lesson buffers and serving image files.

A dual-channel state maintains clean conversation history (user-facing text) and an internal lesson buffer (text plus image metadata). The React UI queries whether lessons contain metadata, requests associated images if present, and renders them responsively with loading and error states. The backend validates paths, restricts file types, and sanitises inputs to ensure security and robustness.

EVALUATION

Metrics. The system was evaluated on: (1) cropping accuracy, (2) OCR and vectorisation quality, and (3) retrieval correctness, latency, and user perception.

Cropping Tool Reliability. The semi-automated process reduced manual workload but produced errors on pages with multi-column exercises, embedded tables, and mixed dialogues, requiring manual correction for a non-trivial subset of pages.

OCR and Vector Store Preparation. The pipeline performed consistently—text extraction preserved Luxembourgish diacritics, semantic chunking produced coherent segments, and metadata enrichment supported reliable downstream retrieval.

Function Calling Integration and Latency. Measurements collected via LangSmith during a representative tutoring session show that LLM generation dominates latency, while retrieval operations complete in under three seconds.

Table 1: Latency breakdown from LangGraph conversation trace.

Component	Duration(s)	Notes
Communicator (first response)	4.21	Initial LLM reply
Orchestrator + retrieval	97.79	Includes getChunks and metadata resolution
VectorStoreRetriever	2.66	ChromaDB semantic query
Tutor (longest generation)	108.83	Main tutor agent LLM generation
Router_tutor	77.16	Routing and agent coordination overhead
Session finalisation	3.65	Final state update
Total (end-to-end)	**≈ 386.09**	Full multi-turn session duration

These results confirm that optimising LLM prompt efficiency, caching, or selecting a smaller model would yield far greater benefits than tuning retrieval or preprocessing modules.

Accuracy and User Perception. Manual inspection confirmed consistent retrieval of correct exercise images when metadata was available. Pilot users reported improved clarity and engagement, particularly for reading and listening tasks. Figures 1 and 2 illustrate the interface before and after retrieval.

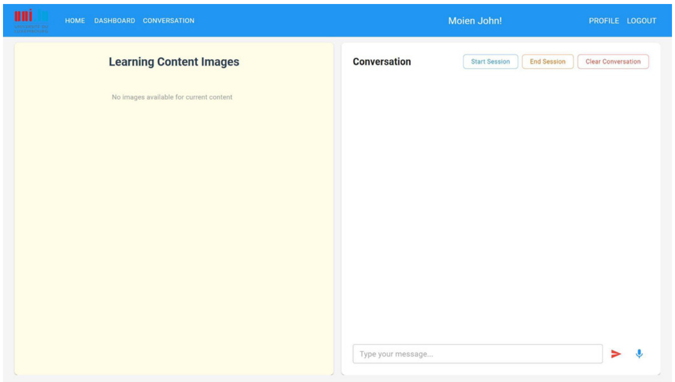
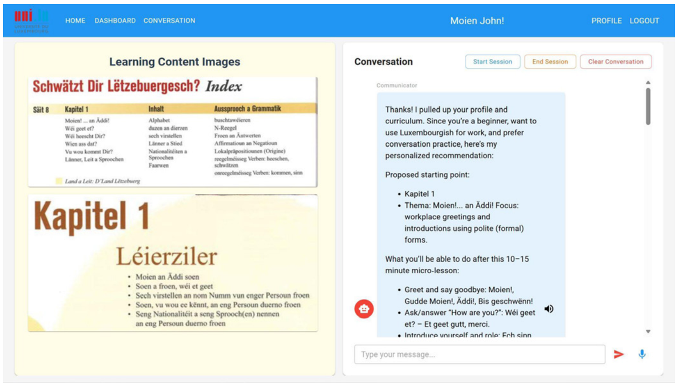


Figure 1: Baseline interface before retrieval (empty state).



Learning Content Images

Schwätzt Dir Lëtzebuergesch? *Index*

Säit 8	Kapitel 1	Inhalt	Aussprooch a Grammatik
	Moien! ... an Äddi!	Alphabet	buschtawéieren
	Wéi geet et?	duzen an dierzen	N-Reegel
	Wéi heescht Dir?	sech virstellen	Froen an Äntwerten
	Wien ass dat?	Länner a Stied	Affirmatioun an Negatioun
	Vu wou kommt Dir?	Nationalitéiten a Sproochen	Lokalpräpositiounen (Origine)
	Länner, Leit a Sproochen	Faarwen	reegelméisseg Verben: heeschen, schwätzen
			onreegelméisseg Verben: kommen, sinn

Land a Leit: D'Land Lëtzebuerg

Kapitel 1

Léierziler

- Moien an Äddi soen
- Soen a froen, wéi et geet
- Sech virstellen an nom Numm vun enger Persoun froen
- Soen, vu wou ee kënnt, an eng Persoun duerno froen
- Seng Nationalitéit a seng Sprooch(en) nennen an eng Persoun duerno froen

Conversation

Communicator

Thanks! I pulled up your profile and curriculum. Since you're a beginner, want to use Luxembourgish for work, and prefer conversation practice, here's my personalized recommendation:

Proposed starting point:

- Kapitel 1
- Thema: Moien!... an Äddi! Focus: workplace greetings and introductions using polite (formal) forms.

What you'll be able to do after this 10–15 minute micro-lesson:

- Greet and say goodbye: Moien!, Gudde Moien!, Äddi!, Bis geschwënn!
- Ask/answer "How are you?": Wéi geet et? – Et geet gutt, merci.
- Introduce yourself and role: Ech sinn

Figure 2: Dynamically populated interface with retrieved textbook images displayed during tutoring.

Across all metrics, three conclusions emerge:

- The cropping tool accelerates preprocessing but remains a key bottleneck due to layout complexity.
- The OCR–RAG pipeline is robust, accurate, and scalable.
- Function-calling integration retrieves the correct images reliably, with latency dominated by LLM computation rather than system architecture.

DISCUSSION AND LESSONS LEARNED

The staged integration of dynamic image retrieval into the Luxembourgish language learning application yields several insights.

Pipeline robustness and preprocessing challenges. The semi-automated cropping pipeline accelerates dataset preparation but highlights the limits of generic text detection and classification for complex educational materials. The variety of layouts in the INL textbook makes bounding-box detection error-prone, and early preprocessing quality directly impacts downstream retrieval accuracy. In contrast, the OCR–RAG pipeline demonstrates the value of stable middle layers, reusing an existing architecture meant that extending it to images and metadata required minimal additional engineering. Robust, reusable components at this level provide leverage for iterative innovation.

Performance characteristics. System performance analysis confirms that LLM generation dominates latency. LangSmith traces show that retrieval and image serving are comparatively cheap. Optimisation efforts should therefore focus on prompt efficiency, caching, model selection, and orchestration overhead rather than low-level retrieval mechanics.

User experience and pedagogical implications. User interface design contributes to perceived authenticity and trust. The contrast between an empty baseline interface and the dynamically populated version demonstrates to learners that content is retrieved at run-time and grounded in a specific textbook, transparency that is important in educational settings where trust in AI systems is still emerging. Pilot feedback further suggests that contextual images reduce cognitive load, particularly for beginner learners who benefit from immediate visual anchoring. By grounding LLM-driven tutoring in authoritative visual material, the system also mitigates hallucinations, a known limitation of large language models (Huang et al., 2023; Niu et al., 2024; Rzepka et al., 2023).

FUTURE WORK

Several directions can extend the system:

- **Improved preprocessing.** Enhance the cropping pipeline for higher accuracy across different textbooks, reducing manual correction. Extend OCR to output JSON-based metadata, enabling structured representation of fill-in-the-blank, table, or matching tasks.
- **Interactive exercises.** Move beyond static rendering by enabling learners to complete exercises directly in the interface, including fill-in-the-blank, multiple-choice, drag-and-drop, and voice-based pronunciation tasks.
- **Performance optimization.** Reduce LLM generation times through smaller models, prompt compression, response caching, and pre-computed retrieval. Integrate event-driven APIs or WebSockets for real-time updates without page reloads.
- **Scalability and evaluation.** Transition to a more modular backend and extend beyond Luxembourgish to other languages. Conduct controlled studies measuring learning outcomes and long-term retention, not only usability and perception.

CONCLUSION

This paper presented a three-stage pipeline for multimodal educational content delivery in a Luxembourgish language learning application: a semi-automated cropping tool to extract exercises from the INL textbook, an OCR–RAG pipeline that transforms cropped images into structured, searchable chunks with rich metadata, and a function-calling integration that enables tutoring agents to dynamically retrieve and display relevant images during live conversations. Together, these stages bridge the gap between static textbook exercises and interactive digital learning.

Evaluation shows that the cropping tool accelerates preprocessing but suffers from detection errors in complex layouts, requiring manual correction. In contrast, the OCR–RAG pipeline is robust and accurate, seamlessly extending a text-only vector store to incorporate exercise images. The function-calling stage successfully links these resources to the tutoring interface, providing dynamic, contextually aligned visuals that improve learner engagement and comprehension. Latency analysis confirms that performance bottlenecks originate primarily from LLM generation rather than retrieval or rendering.

The system demonstrates the feasibility and educational benefits of combining computer vision, OCR, RAG, and LAM-style function calling within a unified pipeline. Grounding AI-driven tutoring in authoritative textbook material increases both trust and learning effectiveness and offers a replicable blueprint for future multimodal, curriculum-grounded AI tutors.

This work completes the progression initiated in our previous contributions (Tebourbi et al., 2025a; Tebourbi et al., 2025b; Kluczyński et al., 2025) by adding a fully unified, end-to-end multimodal retrieval pipeline grounded in authoritative curriculum materials.

ACKNOWLEDGMENT

The authors thank everyone who contributed to the design, development, and evaluation of this system.

REFERENCES

BAAI (2023) *BGE Embedding Models: Technical Report*. Available at: https://huggingface.co/BAAI/bge-large-en-v1.5 (Accessed: 1 July 2025).

Chroma Inc. (2024) *Chroma: Open-Source Embeddings Database for AI Applications*. Available at: https://www.trychroma.com (Accessed: 1 July 2025).

Duolingo Inc. (2023) *Duolingo Max Uses GPT-4 For New Learning Features*. Available at: https://blog.duolingo.com/duolingo-max.

Eleti, A., Harris, J. and Kilpatrick, L. (2023) Function Calling and Other API Updates. OpenAI. Available at: https://openai.com/index/function-calling-and-other-api-updates/.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778.

Huang, L. et al. (2023) 'A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions', *arXiv preprint*, arXiv:2311.05232.

INL – Institut national des langues (2021) *Schwätzt Dir Lëtzebuergesch? – Niveau A1*. Luxembourg: INL. Available at: https://sdl.inll.lu/book-a1-2017-2018-2020-2021/.

Kim, W-H. and Kim, J-H. (2020) 'Individualized AI tutor based on developmental learning networks', *IEEE Access*, 8, pp. 27927–27937.

Kluczyński, P., Mualla, Y., Nouzri, S., Tebourbi, H., Picard, A., Gechter, F. and Abbas-Turki, A. (2025) 'Augmenting an LLM-based tutor agent with a Large Action Model for multimodal interaction', Proceedings of the 2nd International Workshop on Causality, Agents and Large Models (CALM 2025), University of Luxembourg, Luxembourg, December 2025.

LangChain Inc. (2024a) *LangGraph: Building Stateful Multi-Agent Applications*. Available at: https://langchain.com/langgraph (Accessed: 1 July 2025).

LangChain Inc. (2024b) *LangSmith: Interactive Tooling for Explainable LLM Workflows*. Available at: https://langchain.com/langsmith (Accessed: 1 July 2025).

Lewis, P. et al. (2020) 'Retrieval-augmented generation for knowledge-intensive NLP tasks', in *Advances in Neural Information Processing Systems (NeurIPS)*, 33, pp. 9459–9474.

Liao, M. et al. (2019) 'Scene text detection via differentiable binarization', in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15091–15100.

Mageira, K., Stamati, T., Boursinos, V. and Tsiatsos, T. (2022) 'Educational AI chatbots for content and language integrated learning', *Applied Sciences*, 12(7), 3239.

Mayer, R.E. (2009) *Multimedia Learning*. 2nd ed. New York: Cambridge University Press. doi: 10.1017/CBO9780511811678.

Niu, M. et al. (2024) 'Mitigating hallucinations in large language models via self-refinement-enhanced knowledge retrieval', *arXiv preprint*, arXiv:2408.07061.

Nouzri, S. et al. (2025) 'Beyond chatbots: Enhancing Luxembourgish language learning through multi-agent systems and large language models', in *PRIMA 2024: Principles and Practice of Multi-Agent Systems*. LNCS 15395, pp. 385–401. Springer. doi: 10.1007/978-3-031-77367-9_29.

OpenAI (2024) *GPT-4.1 Model Documentation*. Available at: https://platform.openai.com/docs/models/gpt-4.1 (Accessed: 1 July 2025).

OpenAI (2025a) *Introducing Operator*. Available at: https://openai.com/index/introducing-operator.

OpenAI (2025b) *Computer-Using Agent: Introducing a Universal Interface for AI to Interact with the Digital World*. Available at: https://openai.com/index/computer-using-agent.

OpenAI (2025c) *Introducing ChatGPT Agent: Bridging Research and Action*. Available at: https://openai.com/index/introducing-chatgpt-agent.

OpenCV (2024) *Open Source Computer Vision Library*. Available at: https://opencv.org (Accessed: 28 August 2025).

Rzepka, R., Araki, K. and Kojima, K. (2023) 'Addressing hallucinations in educational AI: A critical analysis', *International Journal of Artificial Intelligence in Education*, 33, pp. 245–263.

Tebourbi, H., Nouzri, S., Mualla, Y., El Fatimi, M., Najjar, A., Abbas-Turki, A. and Dridi, M. (2025b) 'BPMN-based design of multi-agent systems: Personalized language learning workflow automation with RAG-enhanced knowledge access', Applied Sciences, 16(9), 809. Available at: https://www.mdpi.com/2078-2489/16/9/809 (Accessed: 1 July 2025).

Tebourbi, H., Nouzri, S., Mualla, Y. and Najjar, A. (2025a) 'Personalized language learning: A multi-agent system leveraging LLMs for teaching Luxembourgish', in *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)*, pp. 3032–3034. Detroit, USA: IFAAMAS.

Wang, S., Liu, Y. and Chen, H. (2024) 'RAG applications in educational AI: reducing hallucinations and improving accuracy', *Journal of AI Education*, 11, pp. 156–171.

Wang, S. et al. (2024) 'Large language models for education: A survey and outlook', *arXiv preprint*, arXiv:2403.18105.

Xiao, J. and Bai, Q. (2022) 'iTutor: Promoting AI-guided knowledge interaction in online learning', in *Proceedings of the 2022 International Symposium on Educational Technology (ISET)*, pp. 253–257. IEEE.

Xu, D. and Wang, H. (2006) 'Intelligent agent supported personalisation for virtual learning environments', *Decision Support Systems*, 42(2), pp. 825–843.

Zhang, J. et al. (2024) 'xLAM: A family of Large Action Models to empower AI agent systems', in *Proceedings of the 2025 NAACL-HLT*, pp. 11583–11597.