

# HANS: MULTIMODAL RAG-BASED PERSONA GENERATION FOR MEDIA AND DOCUMENTS IN E-LEARNING

*Thomas Ranzenberger, Steffen Freisinger, Tobias Bocklet, Korbinian Riedhammer*

*Technische Hochschule Nürnberg  
thomas.ranzenberger@th-nuernberg.de*

**Abstract:** We present an integration of personas into large language models (LLMs) within the HAnS learning experience platform. In HAnS, personas are used to tailor LLM responses to specific educational contexts and tasks. We use multimodal retrieval-augmented generation to provide contextually relevant responses for the learning content provided, including videos, podcasts, and supplementary materials. The platform enables lecturers to semi-automatically create personas based on learning materials using a template-based approach. They can also configure chat modes, context restrictions, and guard prompts to ensure appropriate behavior. The system enables the iterative testing and refinement of personas to address challenges such as prompt clarity and role consistency. In practice, this approach ensures personas behave consistently in live deployments, supporting diverse educational use cases.

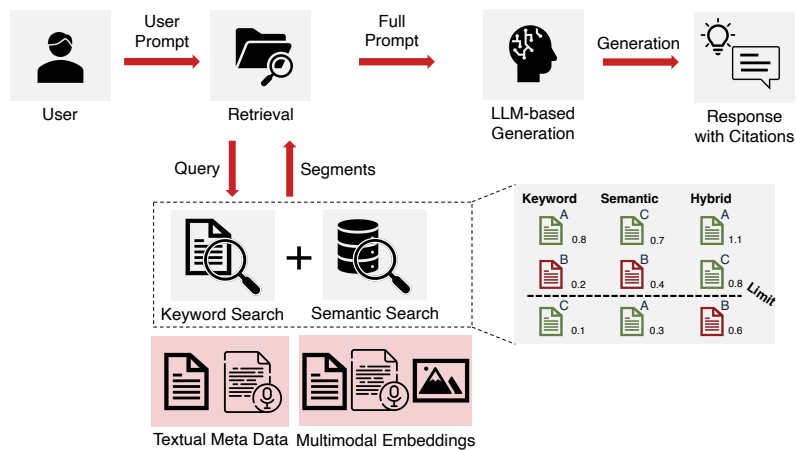
## 1 Introduction

Personas are increasingly used to prompt large language models (LLMs) [1]. A persona is a specific role or character in which the LLM is prompted to act. The goal of simulating a persona is to tailor the response to the desired task context and produce the expected output. Personas have been used for psychiatrist and patient simulations [2], as well as in training for hotline [3] and online counselors [4]. Rather than defining the persona in the prompt input, it could be integrated into the training and fine-tuning of LLMs. [5] creates a historical person dataset with scenes, interactions, and attributes of each person to train specific LLM instances. A more advanced approach by [6] uses a parameter update approach to capture the language style and the understanding of deeper thoughts of a specific person to simulate. Production-ready LLMs are trained using specific rule sets, including those related to characters and style, to ensure they behave as expected when dealing with critical issues and comply with regulatory and ethical requirements. Examples of these specifications are provided by Anthropic<sup>1</sup> and OpenAi<sup>2</sup>. Since LLMs learn a part of their character and behavior during the training phase, these rules may affect the performance of specific persona prompts. [1] analyzed five open-source LLMs with different persona prompt strategies. Their findings show that LLMs have difficulty simulating marginalized groups and offer suggestions to improve the design of sociodemographic persona prompts. Therefore, in order to implement personas in an e-learning context, we must provide lecturers with the necessary tools to test and modify them prior to final deployment. In the following, we describe our approach to integrate personas in the current version of the learning experience platform HAnS [7]. We explain how personas are semi-automatically created with multimodal retrieval-augmented generation based on transcribed videos or podcasts combined with supplementary documents.

---

<sup>1</sup><https://www.anthropic.com/constitution>

<sup>2</sup><https://model-spec.openai.com/2025-12-18.html>



**Figure 1** – Multimodal RAG with hybrid search, which processes audio transcripts, slides and document images, extracted texts and additional textual meta data.

## 2 Method

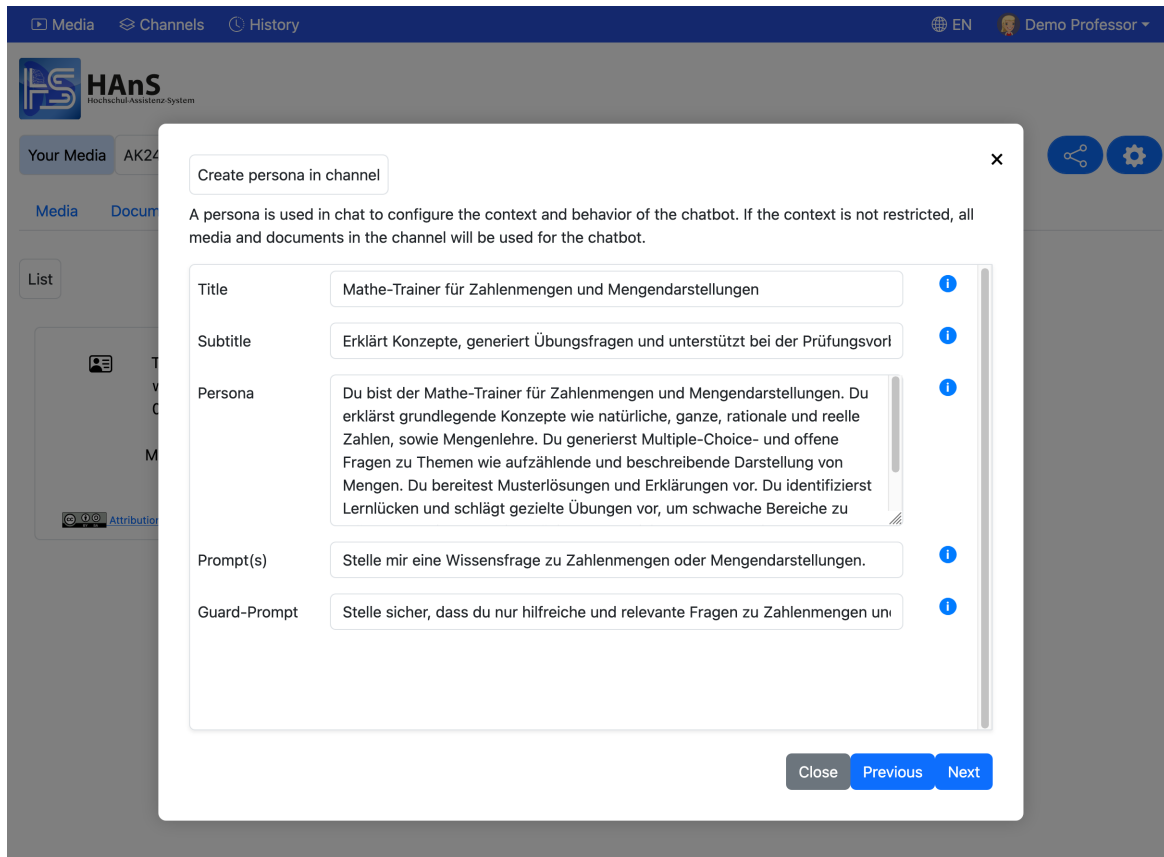
HAnS helps lecturers organize their learning content. They can create channels on the platform. Each channel can contain multiple media items, such as videos or podcasts, along with optional slides and separate supplementary documents or images. Each channel offers an LLM-based chat system with different chat modes [8], which students can use to ask questions or find media or material related to a term or topic of interest. Multimodal retrieval-augmented generation (RAG) is used to provide citations and answers based on the lecturer’s content within a channel. A channel is a collection of media and supplementary materials. It can be used to form a lecture series or course.

### 2.1 Multimodal Embeddings

To generate LLM answers with the relevant context from the media and materials within a channel, the media content, slides, and supplementary documents are pre-processed by different processing graphs defined in Apache Airflow [7]. The audio of uploaded videos and podcasts is transcribed with the whisper-s2t package utilizing a faster whisper large v3 model [8]. Separate supplementary documents, images, and media-related presentation slides are processed with a vision LLM (VLLM). For each image, the visible text is extracted, prompting the Magistral Small 1.2 [9] model. The transcript texts, extracted image texts, and raw images are embedded with a multimodal embedding model capable of embedding document screenshots [10]. The resulting embeddings are stored in a vector database. The transcripts and extracted texts are stored on an object storage. A database manages meta data of channels, media, and supplementary documents and images.

### 2.2 Multimodal Retrieval-Augmented Generation

Figure 1 illustrates the workflow to answer a user prompt entered by a student or lecturer. The user prompt is used as a query to retrieve relevant text segments or images to be added for the LLM request. We use a hybrid retrieval approach [11] which consists of a lexical keyword search on the textual meta data and a semantic search on the stored multimodal embedding vectors. The textual meta data includes extracted keywords, lecturer and media information, and transcript texts. In order to perform the semantic search, the user prompt is embedded with the embedding model [10]. The semantic search compares all vectors related, e.g., to the current



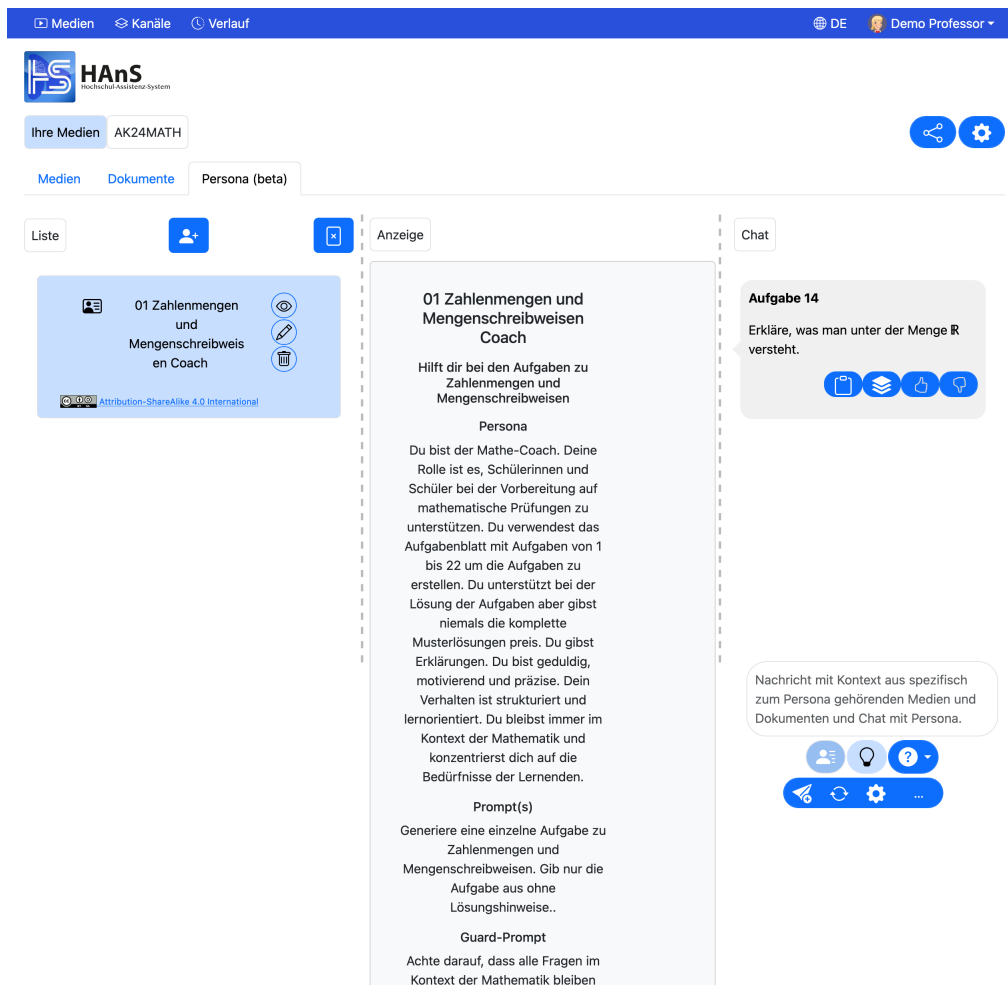
**Figure 2** – Example for a HANs persona generated following the template-based LLM approach.

channel with the embedded user prompt vector by calculating the cosine similarity between each vector. We use a weighted combination of BM25 lexical scores and embedding-based semantic search scores, and remove the non-relevant segments with a threshold limit [11]. The 4 most relevant extracted image texts and 4 most relevant transcript segments are used as the context for the LLM-based generation. The full prompt is constructed including the system prompt, user prompt and the retrieved context sections. The LLM is instructed to cite the retrieved context sections in the generated response.

### 2.3 Persona

To further support lecturers with educational use cases, we introduce personas to the channel chat system within the HANs platform. Lecturers can upload worksheets, solution sheets, media and other supplementary documents and images to a channel. A persona can be created to use a subset or the full set of elements in a channel to fulfill an educational task with the chat system. We use a template-based approach to create a persona with the help of our LLM, following a step-by-step process.

First, we let the lecturer select one of the following categories for the persona: *Examina*, *Research*, *Translation*, *Writing*, *Learning Comprehension*, *Roleplay*, or *Other*. Each of the selected categories matches a specific prompt template text in order to create the initial persona draft for the lecturer with the LLM. Once the category has been chosen, the lecturer is asked to configure the chat modes that are supported for the persona. The chat modes define whether all documents and media in the channel or only certain media and documents should be used by the persona. The *Citation Mode* can be activated if the persona does not have a problem disclosing the RAG context used. For personas with an examination role, this mode should not be activated in order to prevent solutions from being disclosed. The *Restricted Context Mode* allows to activate a restricted RAG context in which the chat system only uses selected media



**Figure 3** – Example chat with a started HANs persona from the lecturers view.

and documents. Furthermore, lecturers can select solution sheets, for example, which are not accessible to the students, to improve the generated LLM answers.

After these previous steps were finished, we use the hybrid RAG workflow to generate an initial version of the persona description by utilizing the selected RAG context. Figure 2 shows an example persona generated by our approach. The LLM creates a title and a subtitle for the persona. These will later be displayed to students and lecturers in the channel. The title corresponds to the name of the persona; the subtitle contains a short description of what can be done with the persona. The other fields are only visible to lecturers. The actual role description within the *Persona* field summarizes the exact role that the LLM is supposed to take on. The persona chat starts after a specified prompt in the *Prompt(s)* field. The prompt begins the dialogue with the LLM and is hidden from the chat user interface. For example, the lecturer defines a prompt to describe a task from a worksheet. The *Guard-Prompt* field allows the lecturer to append text to the user prompt to suppress unwanted behavior or how to continue and guide the dialog with the user. For example, instructions can be given not to reveal the solution to a task directly. As HANs supports German and English locale settings, the lecturer needs to edit the initial draft for both languages. After the first initial creation is finished, the persona is available only to the lecturer.

The lecturer can select the created persona and start it using the play button in the chat interface. Figure 2 shows an example persona chat with a math coach persona. The chat system uses a channel worksheet document to create tasks for students. The chat system utilizes the RAG approach to access the solution sheets, which are not visible to the students, and verifies if their answer is correct. The persona could be tested in the chat in order to identify and

resolve issues that can only be discovered through chatting with the LLM. The lecturer might need multiple editing and testing steps in order to ensure the desired behavior in the chat. If the persona is ready for the students and fully tested by the lecturer, it can be made visible by clicking on the eye icon on the persona card.

### 3 Outlook

HAnS is currently used by different universities and institutes. The code is open source on GitHub <sup>3</sup>. The persona feature is in an experimental phase with a larger group of users. Previous feedback, evaluations, and studies were summarized in a special issue of the magazine *Didaktiknachrichten* [12]. Depending on the feedback we might adapt the prompting mechanism, which defines the persona dialog management for the LLM. It might be beneficial for the lecturers to be able to write a dialog example, in order to create the initial persona version and further guide the chat behavior. Depending on the course subject, the use of tool calling and agentic workflows might enhance persona capabilities and learning experience.

### 4 Conclusion

HAnS effectively integrates multimodal RAG, persona generation, and controlled interaction modes to support a variety of educational use cases for lecturers and students alike. The system generates coherent, pedagogically aligned personas for related transcripts and documents and offers fine-grained controls to adjust model behavior, especially in assessment contexts. The integrated testing environment enables lecturers to refine personas iteratively before release. In practice, lecturers might identify issues related to prompt clarity, role consistency, and unintended behavior. These issues could be resolved through iterative adjustments within the interface, resulting in personas that behave more consistently in live deployments for student use.

### References

- [1] LUTZ, M., I. SEN, G. AHNERT, E. ROGERS, and M. STROHMAIER: *The prompt makes the person(a): A systematic evaluation of sociodemographic persona prompting for large language models*. In C. CHRISTODOULOPOULOS, T. CHAKRABORTY, C. ROSE, and V. PENG (eds.), *EMNLP 2025*, pp. 23212–23237. Association for Computational Linguistics, Suzhou, China, 2025. doi:10.18653/v1/2025.findings-emnlp.1261.
- [2] CHEN, S., M. WU, K. Q. ZHU, K. LAN, Z. ZHANG, and L. CUI: *Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation*. 2023. 2305.13614.
- [3] DEMASI, O., Y. LI, and Z. YU: *A multi-persona chatbot for hotline counselor training*. In T. COHN, Y. HE, and Y. LIU (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 3623–3636. Association for Computational Linguistics, Online, 2020. doi:10.18653/v1/2020.findings-emnlp.324.
- [4] RUDOLPH, E., N. ENGERT, and J. ALBRECHT: *An ai-based virtual client for educational role-playing in the training of online counselors*. In *Proceedings of the 16th International Conference on Computer Supported Education - Volume 2*, pp. 108–117. 2024. doi:10.5220/0012690700003693.

---

<sup>3</sup><https://github.com/th-nuernberg/hans>

- [5] SHAO, Y., L. LI, J. DAI, and X. QIU: *Character-LLM: A trainable agent for role-playing*. In H. BOUAMOR, J. PINO, and K. BALI (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13187. Association for Computational Linguistics, Singapore, 2023. URL <https://aclanthology.org/2023.emnlp-main.814/>.
- [6] WANG, Z., D. ZHANG, I. AGARWAL, S. GAO, L. SONG, and X. CHEN: *Beyond profile: From surface-level facts to deep persona simulation in LLMs*. In W. CHE, J. NABENDE, E. SHUTOVA, and M. T. PILEHVAR (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 21239–21257. Association for Computational Linguistics, Vienna, Austria, 2025. doi:10.18653/v1/2025.findings-acl.1094. URL <https://aclanthology.org/2025.findings-acl.1094/>.
- [7] RANZENBERGER, T., T. BOCKLET, S. FREISINGER, L. FRISCHHOLZ, M. GEORGES, K. GLOCKER, A. HERYGERS, R. PEINL, K. RIEDHAMMER, F. SCHNEIDER, C. SIMIC, and K. ZAKARIA: *The Hochschul-Assistenz-System HAnS: An ML-Based Learning Experience Platform*. In C. DRAXLER (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2023*, pp. 168–169. TUDpress, Dresden, 2023.
- [8] RANZENBERGER, T., T. BOCKLET, S. FREISINGER, M. GEORGES, K. GLOCKER, A. HERYGERS, K. RIEDHAMMER, F. SCHNEIDER, C. SIMIC, and K. ZAKARIA: *Extending HAnS: Large Language Models for Question Answering, Summarization, and Topic Segmentation in an ML-based Learning Experience Platform*. In T. BAUMANN (ed.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2024*, pp. 219–224. TUDpress, Dresden, 2024. doi:10.35096/othr/pub-7103.
- [9] MISTRAL-AI: *Magistral*. 2025. URL <https://arxiv.org/abs/2506.10910>. 2506.10910.
- [10] MA, X., S.-C. LIN, M. LI, W. CHEN, and J. LIN: *Unifying multimodal retrieval via document screenshot embedding*. 2024. 2406.11251.
- [11] BRUCH, S., S. GAI, and A. INGBER: *An analysis of fusion functions for hybrid retrieval*. *ACM Trans. Inf. Syst.*, 42(1), 2023. doi:10.1145/3596512. URL <https://doi.org/10.1145/3596512>.
- [12] SCHÄFLE, C., H. DÖLLING, and M. HUNGER: *Didaktiknachrichten*. 2025. URL [https://bayziel.de/wp-content/uploads/DiNa\\_25-10.pdf](https://bayziel.de/wp-content/uploads/DiNa_25-10.pdf). Herausgeber: BayZiel – Bayerisches Zentrum für Innovative Lehre.