A Modular LLM-Enhanced Agent-Based System for the Generation and Evaluation of Journalistic Interview Questions

Roberto Jiménez de la Torre, Carlos Á. Iglesias
Intelligent Systems Group (GSI)
Universidad Politécnica de Madrid, Madrid, Spain
roberto.jimenez.delatorre@alumnos.upm.es, carlosangel.iglesias@upm.es

Introduction

This work comes from a collaborative initiative with the Prodis Foundation, an institution dedicated to promoting the social and professional inclusion of people with intellectual disabilities. As part of its educational programs, the Foundation envisioned an innovative interview show co-hosted by one of its trained communicators and a conversational Artificial Intelligence (AI) system. The objective was to create an inclusive and supportive environment that enables individuals with cognitive disabilities to actively participate in public dialogue, aided by AI. Rather than focusing solely on task automation, this project explores how AI can assume communicative and pedagogical roles within structured interviews. The developed system assists the human interviewer by retrieving relevant biographical information about a guest, synthesizing that content, and generating contextually appropriate, coherent questions. In doing so, it adapts to the goals, tone, and topics defined for each interview, facilitating fluid and engaging conversations.

This work addresses a broader challenge: integrating Large Language Models (LLMs) into socially meaningful and accessible applications. Although LLMs have already demonstrated their potential in domains such as automated journalism or education, their deployment in inclusive communication settings has not yet been explored. This work contributes to this gap by addressing not only the technical design of such systems but also their ethical, pedagogical, and usability dimensions. The work thus combines strong social motivation with a research-driven approach. The purpose of this study is to demonstrate that LLM-based systems can be effectively adapted to support communicative inclusion, especially when designed with empathy, transparency, and accessibility in mind [5].

Research Question

This work examines the role of LLMs in facilitating inclusive and guided interview experiences, particularly for individuals with intellectual disabilities who act as interviewers. At its core, the system addresses how AI can be leveraged not only to generate coherent and relevant questions but also to support the structure, tone, and pedagogical value of journalistic interviews in socially sensitive real-world contexts.

The central research question guiding this work is : “How can LLM-based conversational systems be designed and orchestrated to support inclusive, structured, and evaluable interviews in social and educational settings?”.

To address this overarching question, the study investigates the following research questions.

- **RQ1**: Are current LLMs suitable for generating biographies for journalistic interviews and which LLMs are more suitable for this task?
- **RQ2**: What strategies are most effective for generating diverse and context-aware interview questions using LLMs?
- **RQ3**: How reliably can LLMs evaluate user responses across multiple qualitative dimensions?
- **RQ4**: How do different LLMs compare in performance, narrative quality, and evaluative consistency within this framework?
- **RQ5**: What design principles can ensure that the system remains accessible, ethically grounded, and adaptable to the needs of users with cognitive disabilities?

These questions situate the project at the intersection of technological development, inclusive communication, and the evaluation of the Social Sciences and Humanities (SSH) using artificial intelligence. The goal is not only to build a functioning prototype but to contribute meaningful insights into how LLMs can enhance human agency and participation in structured conversational tasks.

Methodology

1. System design

The system developed in this work adopts a modular architecture that separates user interaction, reasoning processes, data storage, and communication with external services. This design not only facilitates scalability and maintainability but also enables transparent orchestration and targeted evaluation of each module. The backend is built in Python using the FastAPI framework[1], which handles user session management, asynchronous requests, and inter-agent coordination in real time.

Users interact through a single-page responsive application (SPA) developed with React[2], which integrates an animated 3D avatar rendered using Three.js[3]. The avatar is capable of real-time lip-syncing via viseme alignment from Rhubarb Lip Sync[4]. It delivers questions generated by the system using high-quality neural text-to-speech (TTS) from ElevenLabs[5]. The system also supports voice interaction: user responses are captured and transcribed using OpenAI's Whisper[6], enabling a fluid and accessible conversational experience. The backend logic is structured around a series of reasoning agents responsible for specific tasks such as context retrieval, question generation, evaluation, and dialogue management. These agents interact with LLMs and rely on a semantic memory layer built with vector embeddings stored in Pinecone[7]. External APIs, such as SerpApi[8] and Tavily[9], are used to collect structured and up-to-date information about public figures, while LangChain[10] orchestrates the prompt flows, memory retrieval, and inter-agent communication in a transparent and extensible manner.

The overall architecture reflects current trends in multi-agent LLM-based systems [3], designed to maximize modularity and functional clarity. Unlike monolithic chatbots, this system integrates complementary reasoning components with clearly defined roles, supporting extensions such as multilingual capability, affective computing, or even embodied robotics. The emphasis on transparency and real-time multimodal interaction ensures that the system remains adaptable to future research and deployment contexts.

2. Acquisition and processing of information

To generate relevant and personalized interview questions, the system must first gather accurate and up-to-date information about the interviewee. This process begins with the guest's name, which is used to launch a retrieval pipeline through two Web search APIs, SerpApi and Tavily, that access structured and unstructured sources, including news articles, encyclopedic entries, and social media profiles. The retrieved documents are processed using keyword filtering and semantic similarity techniques. Each text is embedded using OpenAI's text embedding-ada-002 model and indexed in Pinecone, allowing for a semantic search tailored to the user's chosen topics. This enables the system to prioritize context segments aligned with interview themes, such as personal life or professional achievements. Then, an LLM is prompted via LangChain to synthesize a biographical summary from the selected content. This summary serves as both a verbal introduction and a knowledge base for generating subsequent questions. To ensure factual consistency, the system employs quality control mechanisms that discard low-confidence or duplicate content and prioritize reputable sources over informal ones.

The approach follows the principles of Retrieval-Augmented Generation (RAG) [1], enhancing the factual basis while preserving flexibility. By combining real-time web retrieval with semantic memory, the system ensures a rich, context-aware foundation to guide conversational flow.

[1] https://fastapi.tiangolo.com/

[2] https://react.dev/

[3] https://threejs.org/

[4] https://github.com/scaredyfish/blender-rhubarb-lipsync

[5] https://elevenlabs.io/

[6] https://platform.openai.com/docs/guides/speech-to-text

[7] https://www.pinecone.io/

[8] https://serpapi.com/

[9] https://www.tavily.com/

[10] https://www.langchain.com/

3. Generation of questions

Once contextual information on the interviewee has been collected and synthesized, the system generates coherent and personalized interview questions aligned with the selected themes. This task is handled by a reasoning agent that prompts a large language model using LangChain's prompt orchestration framework, a key example of prompt engineering in practice. The input includes a structured summary, a list of preferred topics (e.g., personal life, professional milestones), and a stylistic guideline aligned with the tone of the interview.

Prompt engineering techniques, including Chain-of-Thought (CoT) [4] and Outline-based prompting, are employed to guide the model's reasoning process. The Chain-of-Thought method [4] involves step-by-step reasoning to ensure that each question is deeply rooted in the context, whereas the Outline-based method organizes the prompt according to specific interview goals, creating a clear path for generating questions.

The system applies post-processing heuristics to ensure that the questions are clear, diverse, and suitable for the target audience. Throughout development, different prompt strategies were tested, including template chaining and role-based prompting. A three-block prompt structure, guest profile, interview goals, and generation instructions proved to be more effective in improving consistency and minimizing hallucinations. The resulting output adapts in real time to prior answers, ensuring that each question builds naturally on the evolving dialogue.

This dynamic generation capability is central to the project's goal of inclusive and adaptive communication. Moving beyond static question sets, the system leverages the creative and contextual strengths of LLMs through Prompt Engineering [4] to support engaging interviews that are coherent and responsive.

4. Automated evaluation of responses

A distinctive feature of the system is its ability to evaluate, in real-time, the responses given by the interviewee, whether human or simulated, based on multiple qualitative dimensions. This evaluation process is handled by a dedicated reasoning agent, which prompts a large language model to assess each answer using a structured framework grounded in inclusive communication practices. The selected criteria, developed in collaboration with project stakeholders, include informativeness, contextual relevance, motivational tone, stylistic clarity, and potential for engagement.

For each response, the model receives the original question, the guest's response, and the relevant background context, then assigns qualitative labels and optional feedback. This methodology builds upon recent advances in the LLM-as-a-Judge paradigm [2], where models serve as evaluators using interpretable, prompt-driven metrics. To maintain coherence over time, the agent retains conversational memory via LangChain, allowing it to evaluate each reply to previous turns. The results are formatted as JSON objects and visualized through radar charts, providing immediate and accessible feedback. This is especially useful in inclusive settings, where users benefit from precise, structured representations of performance. During the test, participants reported that this visual evaluation helped them reflect more effectively on their contributions. In addition, this evaluation framework was used to compare the performance of several LLMs integrated into the system. By applying uniform metrics in GPT-4, Gemini, and LLaMA3, the system allowed internal benchmarking of fluency, coherence, and adaptability.

Incorporating real-time evaluation enriches the interaction by transforming the interview into a pedagogical experience. This promotes constructive and adaptive dialogue, and supports transparency, which is particularly critical in contexts where user comprehension and confidence are essential for participation.

5. Comparison between LLMs

To analyze how different language models perform within the system, a comparative evaluation was conducted, focusing on their behavior across the core tasks involved in the interview workflow. Three models were selected for analysis: GPT-4 (OpenAI), Gemini (Google), and LLaMA3 (Meta). Each model was embedded in the same modular architecture, ensuring consistency in prompt structure, memory configuration, retrieval mechanisms, and evaluation pipeline.

The evaluation framework was structured around four main functional components of the system: (1) biographical summary generation, where models synthesized guest profiles from retrieved web content; (2) question generation, assessing their capacity to produce diverse, coherent, and contextually aligned questions; (3) evaluation of individual responses, where each LLM acted as a judge scoring answers along pedagogically relevant criteria; and (4) final interview evaluation, in which the models delivered an overall assessment of the whole interaction sequence. The LLM-as-a-Judge methodology [2] was applied to compare the models' ability to provide evaluative feedback during interviews. In addition to these functional tasks, models were also compared using quantitative indicators that reflect deployment and usability factors. These included average processing speed per interaction turn, memory consumption, pricing structure, and openness of access. These metrics were recorded using standardized input prompts and session logs in controlled test scenarios.

Rather than treating model selection as a matter of ranking, the goal of this comparison was to understand the trade-offs involved in using each model under real-time, structured, and socially sensitive conditions. This evaluation methodology enabled a grounded analysis of how LLMs behave not just in isolation, but when integrated into multi-agent architectures serving inclusive communication tasks.

Preliminary Findings

Preliminary testing of the system confirms its technical feasibility and functional robustness as an autonomous platform for conducting personalized, structured interviews. The complete pipeline, comprising information retrieval, biographical synthesis, question generation, multimodal presentation, and automated evaluation, was successfully executed in controlled interview scenarios. The use case involving the public figure Rafael Nadal served to validate the entire workflow and illustrate the system's ability to maintain topic coherence and stylistic consistency throughout an extended dialogue. Functionally, the system was able to operate with minimal human intervention, generating interview content that aligned with the user's thematic preferences and adapting dynamically to each stage of the interview. The biographical summaries were coherent and faithful to the retrieved content, while the generated questions reflected the appropriate diversity in tone and complexity. The response evaluation mechanism, integrated through radar chart visualizations, provided accessible and structured feedback — a particularly valuable feature for users with cognitive disabilities.

The comparative implementation of three LLMs revealed distinct patterns in how each model handled the core tasks of the system. GPT-4 produced interviews with greater narrative fluidity and contextual alignment; Gemini showed higher discriminative power in evaluating answers, albeit with a slower response time; and LLaMA3 proved efficient for basic generation tasks, making it a lightweight option despite limitations in stylistic expressiveness. In general, these results support the viability of the proposed architecture and its potential for inclusive communication. Beyond its immediate use case, the system serves as a model for how LLMs can be orchestrated transparently and ethically in high-stakes, socially relevant contexts such as education, journalism, and accessibility services.

Impact

This project demonstrates the potential of Large Language Models to enhance inclusive communication by facilitating structured, meaningful dialogue for individuals with intellectual disabilities. By integrating speech synthesis, visual guidance, and automatic evaluation, the system reduces cognitive and linguistic barriers, allowing greater participation in journalistic and educational activities. Its real-time feedback and multimodal interface reflect a user-centered design that prioritizes clarity, accessibility, and engagement.

Beyond its immediate application, the system contributes to the responsible deployment of LLMs in socially relevant contexts. Its modular architecture and internal evaluation framework offer a replicable methodology for designing transparent and adaptable AI systems. By embedding model comparison and reasoning traceability into a real use case, the project advances the practical integration of multi-agent LLM platforms in fields such as education, accessibility, and digital humanities.

References

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[2] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, and Y. Liu. Llms-as-judges: A comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

[3] H. Touvron, T. Lavril, G. Izacard, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[5] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.