A Microservice-Based Conversational AI Architecture for Virtual Human Presentations in Cultural Heritage

Fulvio Coppola[1], Antonio Origlia[1,*] and Vincenzo Norman Vitale[1]

[1]University of Naples Federico II

Abstract

This work presents a work in progress for a modular, containerised architecture for a Conversational AI system that drives a virtual human designed for cultural heritage presentation. The pipeline integrates Automatic Speech Recognition (ASR), a large language model for question answering, an embedding model for semantic retrieval over a knowledge graph, and a Text-to-Speech (TTS) module. Dialogue orchestration is handled by a LangGraph-based agent managing multi-step reasoning and tool coordination across the pipeline as a stateful graph of actions. Inter-service communication is strongly decoupled through Apache Kafka topics, allowing producers and consumers to operate asynchronously along the dialogue loop. This design is intentionally modular: by treating each AI component, including the LangGraph agent, as a relocatable microservice, the architecture enables systematic experimentation with different service placement strategies across intelligent network infrastructures.

Keywords

Virtual Humans, Intelligent Networks, Conversational AI

1. Introduction

Presenting cultural heritage is a topic with a long-standing history that touches multiple topics related to Human-Computer-Interaction and museological studies. The use of Virtual Humans, in particular, has collected a lot of interest: [1] presents a multi-modal LLM-powered virtual tour guide with voice narration and avatar-based guidance in a simulated VR museum while [2] investigate five design elements (character, styling, dynamic, sound, scene) and their effect on user satisfaction with virtual humans as cultural mediators. A full review of past approaches to the problem can be found in [3].

Given the rise of Large Language Models (LLMs), Conversational AI is the target of a significant amount of interest from both the industrial and the academic communities. The extremely convincing outputs these systems can provide while interacting using natural language has brought many users, especially during the first months since the publication of ChatGPT by OpenAI, to attribute intelligence to them. This happens regardless of these agents not having a humanoid form: while it is true that anthropomorphic appearance leads people to consider artificial agents more intelligent [4] and believable [5], contextual factors play a crucial role in anthropomorphisation, beyond the agent's looks [6]. More than that, it has been shown that "[...] people tend to predict and explain robot behaviour with reference to mental states without reflecting on the reality of those states" [7, p.1100]. Since the typically meaningful social interactions happen between humans using language, the instinctive approach to understanding the behaviour of an artificial agent adopting language as a communication mean is to attribute mental states to it, regardless of these being actually present.

In this view, the need for deeper representations of cognitive processes and explicit knowledge representation has opened the way to new approaches trying to reconcile model-based and neural-based approaches by creating hybrid models for AI. Such approaches are characterised by heterogeneous technologies exchanging information to collaboratively solve problems. These architectures can be naturally designed to leverage on microservices, which has the interesting consequence of making them

Workshop on Advanced Visual Interfaces for Cultural Heritage (AVICH 2026), June 08, 2026, co-located with the 18th International Conference on Advanced Visual Interfaces, Venice, Italy.

*Corresponding author.

fulvio.coppola@studenti.unina.it (F. Coppola); antonio.origlia@unina.it (A. Origlia); vincenzonorman.vitale@unina.it (V. N. Vitale)

an interesting use case to conduct research on modern intelligent network infrastructure. Interactive experiences with Virtual Humans, in fact, have strict requirements in terms of perceived latency from the users: a case explicitly targeted by [8]. Positioning microservices powering a Virtual Human for cultural heritage presentations in a complex, heterogenous network has also been explored by [9].

In this paper, we present ongoing work on the design of such an architecture, using a basic setup that can be deployed on intelligent networks to take advantage of edge computing capabilities. In this approach, we stress decoupling and extensibility of the design by investing on an agentic orchestrator, which can be easily expanded by modifying the structure of the control graph, and on a message broker supporting both event-based and streamed data: Apache Kafka.

2. Conversational AI for Cultural Heritage

The use of chatbots and conversational agents sparked significant discussion given their implications in ethical consequences of human–machine conversation in the heritage industry [10] and about their applications to digital mannequins impersonating historical characters [11]. The specific case of Virtual Humans as *gatekeepers*, moreover, poses interesting questions concerning the empowerment of users through preparatory experiences [12] and the impact of the use of linguistically motivated disfluent synthetic voices on memorization capabilities [13].

Combining multiple technologies to deliver an engaging experience about cultural heritage goes beyond what LLMs can do through their purely associative capabilities. Nuances, cross-references and mutual influences make the landscape of cultural heritage difficult to manage without explicit knowledge representations. Symbolic reasoning and utility-based decision making are also becoming more and more investigated to deliver the idea of Hybrid reasoning. These structurally different technologies need to be orchestrated so that every module performs the tasks it is most adequate for and passes information to the other relevant modules. The current orientation leans towards Agentic AI, where AI Agents are typically considered to be software artifacts that are able to design their own workflows and utilizing available tools. As in these approaches Large Language Models are still central, however, this approach still suffers from an implicit attempt to introduce symbolic reasoning in a deep learning architecture.

In the design we propose here, LLM-centric orchestration routing is limited at tool-usage concerning context sensing: colleting information that is external to the linguistic context but still influences communication like time of day, available gold-data and current position constitute information that can be recovered through a decision process that simply associates the needed typology of information to the problem at hand. On another level, this collection of data needs to be passed to modules with different capabilities, aimed at combining and manipulating relevant data to manage dialogues in a goal-directed way. The need for different levels of decision making and communication, in a rapidly evolving field, motivates the approach we present here to the design of the proposed distributed architecture. Specifically, in cultural heritage settings, visitors engage in short, situated interactions where responses must be grounded in curated content. The knowledge graph serves as the authoritative content repository, while co-located interaction with the Virtual Human imposes latency constraints that motivate the architectural separation we propose, distinguishing it from monolithic, cloud-centric, chat-oriented deployments.

3. A distributed architecture for Explainable Conversational AI

One of the main bottlenecks in the development of such architectures is to be found in the need to keep the coupling between services low and reduce the latency in data processing to a minimum to ensure that they, with reference to the ISO/IEC 25010 standard, satisfy the requirements that state-of-the-art quality dialogue systems must necessarily possess, first of all *Flexibility* (with particular attention to *Scalability*), *Performance Effectiveness* (and in particular *Temporal Behaviour*, *Resource Utilization*) and *Maintenance* (and the sub-property of *Modularity*).

In Figure 1, we detail the underlying Kubernetes architecture that provides the advanced functionalities necessary to deploy the proposed framework. Access to LLM models is achieved via a mounting mechanism relying on Persistent Volumes and related Claims. This approach ensures microservices remain completely stateless—capable of being created, moved across the intelligent network, or destroyed on demand—perfectly aligning with the ephemeral nature of Kubernetes. To access the Data Lake, processes simply raise claims to the repository hosting the required models.

Figure 2 highlights the central role the Kafka ecosystem plays in reducing the degree of coupling between the microservices constituting the presented architecture. While the introduction of specific Kubernetes components, such as Services, allows microservices to communicate through predefined endpoints, automatically managed by Kubernetes, this risks introducing a wide variety of interfaces (REST, proprietary interfaces, etc.), making the system less flexible. Using a message-broking abstraction layer reduces the degree of strict dependency between microservices, as well as bringing with it the advantage of making communication intrinsically asynchronous. Using a message broker that supports streaming, like Apache Kafka, leaves ample space to future developments. Figure 2 also shows the role each micro-service adopts in the subscription to a specific Kafka topic, whether as a *Consumer* or a *Producer*. The service graph in Figure 3 illustrates the resulting system architecture, instantiated over the Kubernetes cluster (Fig. 1) and communicating through the Kafka mechanism described above. The main modules underlying the proposed architecture are independent of one another and, by connecting to Kafka topics, have no knowledge of which module will consume their output. This helps decoupling, while centralizing the representation of data communication formats in a single manager, i.e., Kafka. From the service graph shown in Figure 3, the ASR model records the user's message, converts it into a string format and publishes the result on the topic *user_query*. This is where the Neo4j module (here representing the knowledge graph) is involved, recording the user's new query in the database and then returning the entire conversation history to the Dialog Manager. Similarly, when exiting the Dialog Manager, the AI model's responses are recorded on the appropriate topic *llm_response* to allow the virtual assistant to deliver what the LLM generated and, furthermore, to record in the knowledge graph the output of the LLM invocation. With these considerations in mind, reading the architecture as a service graph in Figure 3 becomes fairly straightforward. As shown in Figure 3, the tool-calling functionality in the agent acting as the Dialogue Manager exposes a Retrieval-Augmented Generation (RAG) tool that queries the knowledge graph via the embedding service and directly invokes the LLM servers for text generation as needed. This choice leverages the standard Model Context Protocol (MCP) to achieve higher speed during LLM tool use. Nevertheless, this connection still relies on an intermediate communication layer, represented by the Kubernetes LoadBalancer. In this context, the explainability emerges as the combination of two complementary design choices. First, grounding LLM responses on the curated knowledge stored in the Neo4j graph, combined with the persistent dialogue history, enables backtracking of erroneous or hallucinated outputs to their evidential support, which is a prerequisite for building robust agents. Second, the separation of responsibilities across stateless, specialized microservices makes it possible to localize failures along the dialogue loop, both from a logical standpoint (which module produced an inconsistent output) and from a Quality of Experience (QoE) standpoint (which module introduced a perceivable degradation). Errors thus become traceable and bottlenecks identifiable, exposing aspects of the interaction that would remain opaque in monolithic conversational deployments.

The proposed decomposition is not without trade-offs: introducing a message broker and multiple containerized services increases system complexity, complicates end-to-end debugging, and adds a non-zero latency contribution from inter-service serialization and Kafka brokering. However, moving from a monolithic deployment to specialized microservices enables selective replication and placement of only those components handling bandwidth-intensive payloads, such as the ASR service ingesting raw audio streams. Once audio is transcribed near the edge, the downstream pipeline exchanges compact textual messages, whose size is several orders of magnitude smaller than the corresponding waveform. The same modularity that introduces brokering overhead therefore enables targeted latency reduction that would be unattainable in a co-located monolithic design.

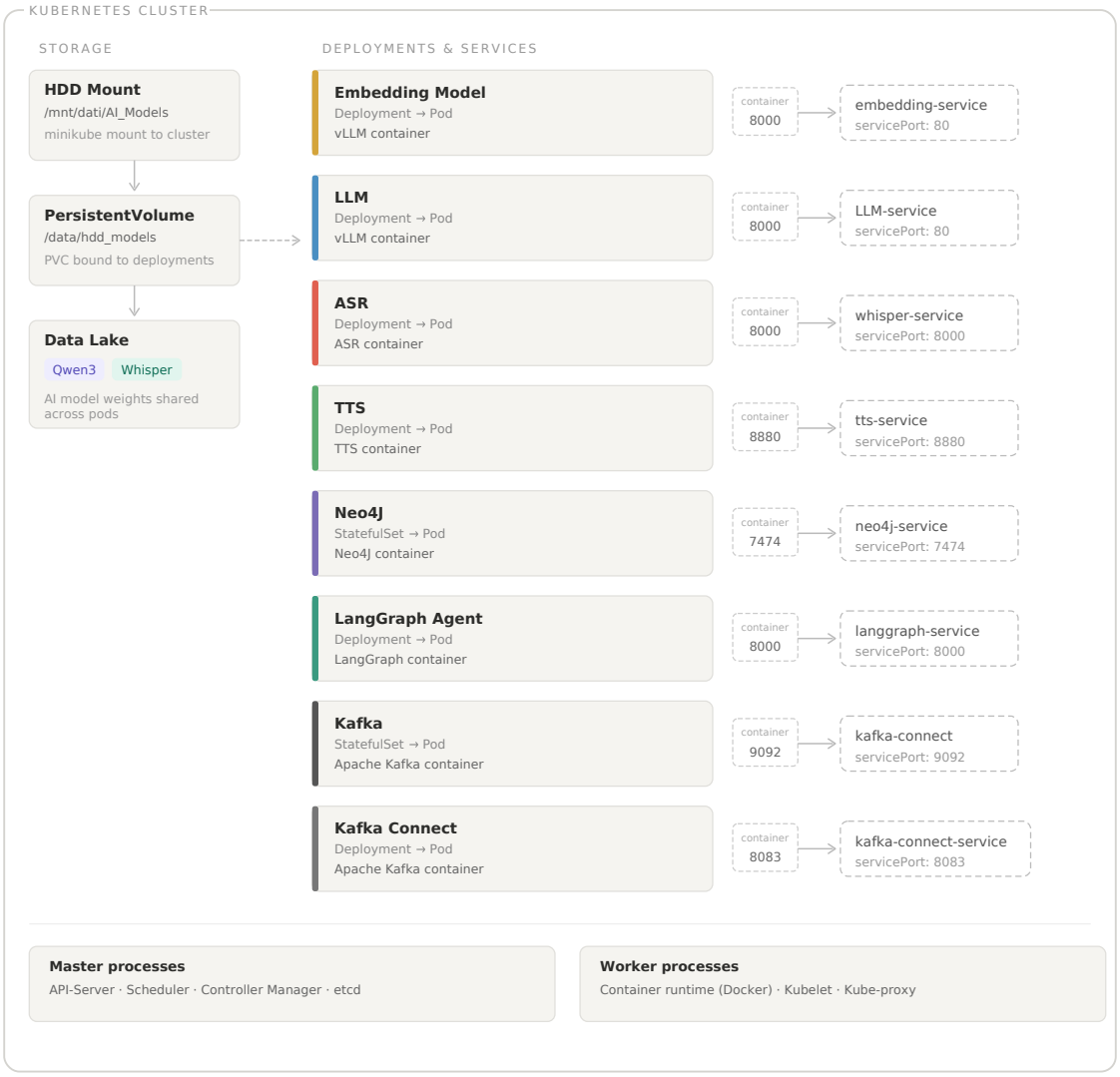

Figure 1: Involved modules in the Kubernetes architecture

4. Conclusions

We discussed an ongoing work on a distributed architecture implementing a hybrid approach to Conversational AI that can be deployed in cultural heritage settings to support, at the same time, two lines of research. The first one strictly concerns the development of Conversational AI for the specific case of cultural contents, using Virtual Humans as interactive *beacons* helping people explore and make sense of large, unfamiliar, information sets. The second one aims at exploring the role of intelligent networks in managing the positioning of multiple microservices on machines with different capabilities. The proposed design, founded on strong decoupling, supports both extensibility of functionalities and services through the use of an agentic AI orchestrator and a Kafka message broker.

5. Declaration on Generative AI

Generative AI is used in this paper to implement the verbalisation capabilities of the conversational agent. Also, figures have been generated with the assistance of Anthropic Claude.

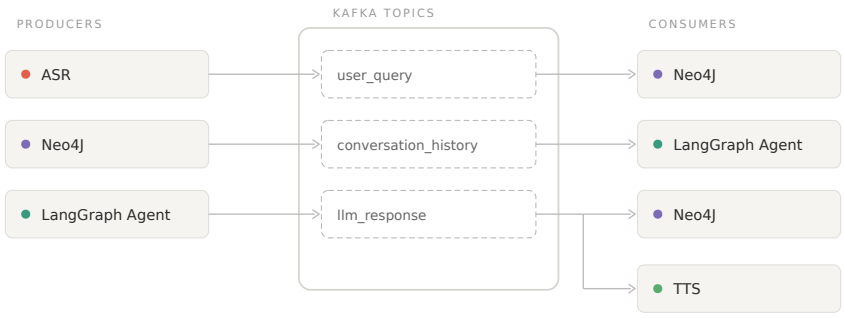

Figure 2: Kafka topics detail with relevant Producers and Consumers

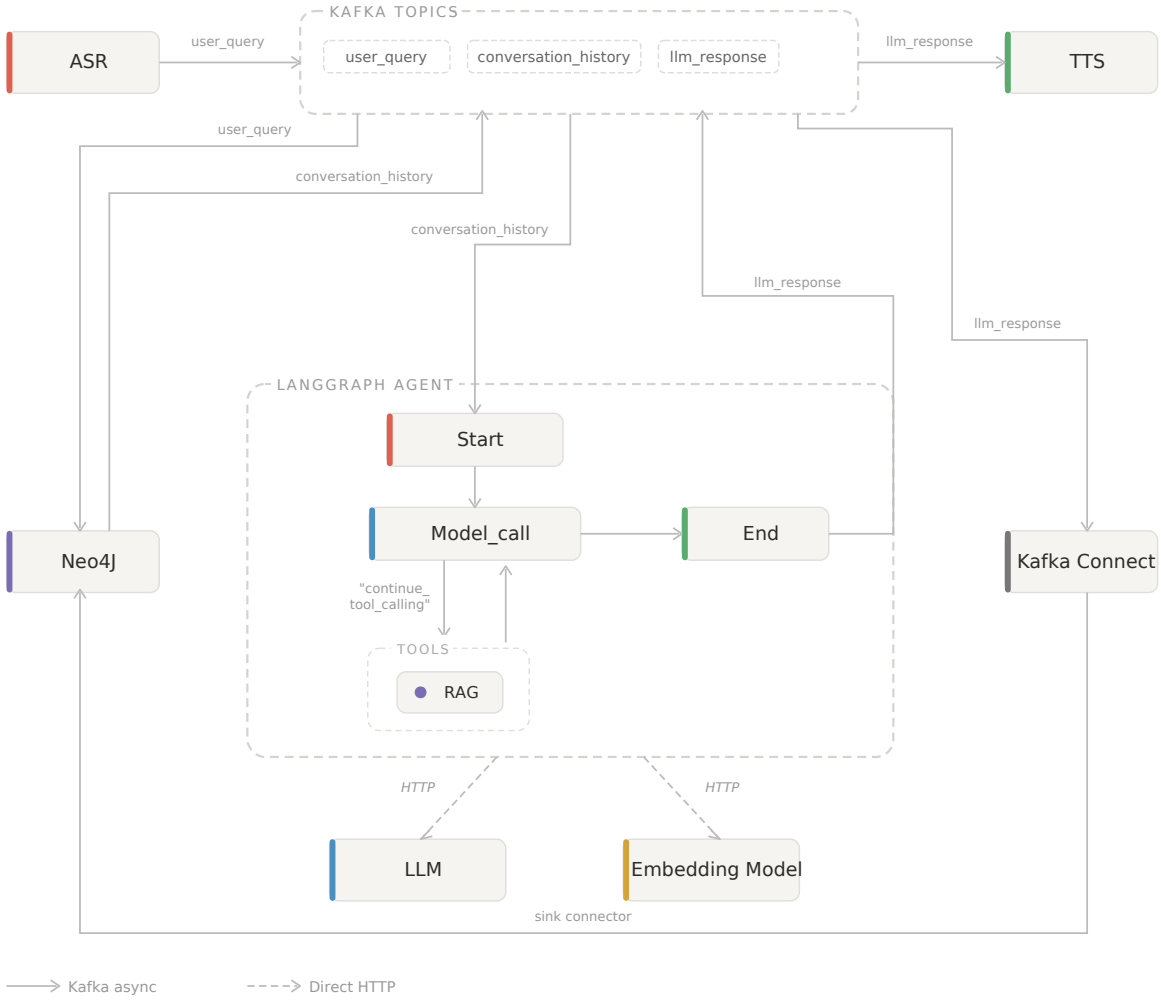

Figure 3: The proposed architecture presented in the form of a service graph.

References

[1] Z. Wang, L.-P. Yuan, L. Wang, B. Jiang, W. Zeng, Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models, in: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, 2024, pp. 1–20.

[2] J. Li, The impact of hyperrealistic virtual humans on cultural heritage dissemination, npj Heritage

Science 13 (2025) 515.
[3] O. M. Machidon, M. Duguleana, M. Carrozzino, Virtual humans in cultural heritage ict applications: A review, Journal of Cultural Heritage 33 (2018) 249–260.
[4] K. S. Haring, D. Silvera-Tawil, T. Takahashi, K. Watanabe, M. Velonaki, How people perceive different robot types: A direct comparison of an android, humanoid, and non-biomimetic robot, in: 2016 8th international conference on knowledge and smart technology (kst), IEEE, 2016, pp. 265–270.
[5] D. Zanatto, M. Patacchiola, A. Cangelosi, J. Goslin, Generalisation of anthropomorphic stereotype, International Journal of Social Robotics 12 (2020) 163–172.
[6] M. Dubois-Sage, B. Jacquet, F. Jamet, J. Baratgin, We do not anthropomorphize a robot based only on its cover: context matters too!, Applied Sciences 13 (2023) 8743.
[7] S. Thellman, T. Ziemke, The intentional stance toward robots: Conceptual and methodologicalconsiderations, in: Proceedings of the Annual Meeting of the Cognitive Science Society, volume 41, 2019.
[8] C. Centofanti, W. Tiberti, A. Marotta, F. Graziosi, D. Cassioli, Taming latency at the edge: A user-aware service placement approach, Computer Networks 247 (2024) 110444.
[9] M. Di Bratto, A. Origlia, J. Llorca, A. Detti, A. Mauro, M. Grazioso, V. N. Vitale, V. Russo, A. Mancini, A. Perrino, et al., Automatic positioning of ai microservices on nextg networks to support interactive holograms (2025).
[10] K. Brown, Artificial exchange: Chatbots and the ethics of museum experience design, in: Transformative museum experiences, Springer, 2025, pp. 247–262.
[11] M. P. Engstrøm, A. S. Løvlie, Using a large language model as design material for an interactive museum installation, in: Companion Publication of the 2025 ACM Designing Interactive Systems Conference, 2025, pp. 386–390.
[12] A. Origlia, M. L. Chiacchio, M. Grazioso, F. Cutugno, Increasing visitors attention with introductory portal technology to complex cultural sites, International Journal of Human-Computer Studies 180 (2023) 103135.
[13] L. Schettino, A. Origlia, F. Cutugno, Though this be hesitant, yet there is method in't: Effects of disfluency patterns in neural speech synthesis for cultural heritage presentations, Computer Speech & Language 85 (2024) 101585.