electronics

MDPI

Article

Enhancing Scientific Communication and Institutional Identity Through a Retrieval-Augmented Generation Digital Personal Tutor

Stefano Di Tore [1], Michele Domenico Todino [2,*], Alessio Di Paolo [2,*], Lucia Campitiello [2], Umberto Bilotti [2], Riccardo Villari [3] and Maurizio Sibilio [2,*]

[1] Department of Political and Social Studies DISPS, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Italy; sditore@unisa.it
[2] Department of Humanities, Philosophy and Education DISUFF, University of Salerno, Via Giovanni Paolo II, 132, 84084 Fisciano, Italy; lcampitiello@unisa.it (L.C.); ubilotti@unisa.it (U.B.)
[3] IDIS Foundation Città della Scienza, Via Coroglio, 57/104, 80124 Naples, Italy; presidente@cittadellascienza.it
* Correspondence: mtodino@unisa.it (M.D.T.); adipaolo@unisa.it (A.D.P.); msibilio@unisa.it (M.S.)

Abstract

This project presents the development of a Retrieval-Augmented Generation (RAG) system applied to the customization of a Non-Playable Character (NPC), designed as the Non-Playable Character (NPC) of the President of the IDIS Foundation Città della Scienza (City of Science). The NPC acts as both a virtual guide and institutional ambassador within the science center, providing multilingual, interactive, and accessible communication for a broad international audience. Through the integration of generative models with a curated, validated knowledge base, the RAG system enables the NPC to provide accurate, context-sensitive, and up-to-date responses to user queries. Developed by the Teaching Learning Centre for Education and Inclusive Technologies 'Elisa Frauenfelder' at the University of Salerno, the system supports the museum's educational mission by enhancing science communication and fostering inclusive digital engagement. The Non-Playable Character (NPC) features realistic facial animation, movement, and voice synthesis, creating a digital twin capable of simulating human-like interaction. This initiative exemplifies an innovative application of artificial intelligence for an inclusive and equitable quality education and contributes to the development of engaging, accessible, and personalized learning environments.

Keywords: Retrieval-Augmented Generation (RAG); Non-Playable Character (NPC); artificial intelligence in education; digital twin; human–computer interaction

Academic Editors: Lina Sawalha and Simin Masihi

Received: 16 December 2025
Revised: 12 February 2026
Accepted: 14 February 2026
Published: 17 February 2026

1. Introduction

The Retrieval-Augmented Generation (RAG) architecture combines the generative capabilities of large language models (LLMs) with a curated retrieval component, enabling the Non-Playable Character (NPC) to deliver context-sensitive, accurate, and source-verified responses [1–3]. This architecture is particularly relevant in educational environments because it mitigates hallucinations by grounding outputs in validated, domain-specific knowledge. The knowledge base was compiled and reviewed by the scientific committee of Città della Scienza (cittadellascienza.it/en/), ensuring both pedagogical rigor and epistemological reliability.

The system was deployed within one of Italy's leading science centers, which welcomes more than 200,000 visitors annually and features major exhibits such as a planetarium and a permanent installation on the human body, the NPC serves as both an interactive guide and an institutional representative. Developed by the Teaching and Learning Centre for Education and Inclusive Technologies "Elisa Frauenfelder" at the University of Salerno (labh.it/disuff), the NPC acts as an interface between complex scientific content and a diverse audience, promoting accessibility and engagement across different ages and cultural backgrounds.

Technically, the NPC functions as a digital twin of the Foundation's President, integrating high-resolution facial scans, gesture modeling, and synthetic voice reproduction through advanced AI audio synthesis. These embodied features enhance user attention, foster empathy, and simulate intersubjective interaction, in line with the principles of embodied cognition and shared manifold theory. This embodied dimension enhances the pedagogical potential of AI-mediated communication by combining realism with educational purpose. The system supports multiple languages, including major European languages and Arabic, reflecting the Foundation's international mission and its commitment to reducing linguistic and cultural barriers in access to scientific knowledge. This feature is particularly important given the museum's ongoing collaborations with institutions in Qatar and China, promoted by the Italian Ministry of University and Research and the Campania Region. The NPC was publicly presented at the *Montalcini Global Biotech Tour 2025* in Doha, where it facilitated exchanges with the Qatar Ministry of Economy and Industry, Sidra Medicine, Qatar University, and other major research institutions. The use of this AI technology in an international forum demonstrates its capacity to enhance institutional visibility while contributing to the global democratization of scientific communication.

The NPC operates within a modular architecture compatible with accessible platforms such as Convai, allowing educators and museum professionals to create and personalize AI agents without requiring advanced programming skills. This configuration supports inclusive education by providing compensatory tools for learners with special educational needs, including those with dyslexia, who may struggle with complex or text-heavy materials. The system's interactive "question-time" methodology generates differentiated responses based on user input, encouraging self-regulated learning and metacognitive reflection [4–7].

The originality of this work does not lie in proposing a new algorithmic model, but in developing a methodological and theoretical framework that integrates the RAG architecture within an educational and institutional context. While RAG has been previously discussed in technical literature as a method for enhancing knowledge-intensive tasks, its application to a socially engaged, pedagogically oriented, and anthropomorphically embodied Non-Playable Character (NPC) represents a novel contribution. The study advances a design-based methodology in which AI technologies are co-developed through humanistic, ethical, and educational principles, transforming a general generative architecture into an adaptive Digital Personal Tutor aligned with the values of Artificial Intelligence for Social Good (AI4SG). This integration results in a hybrid system that operationalizes transparency, inclusivity, and contextual awareness through an empirically validated implementation, extending the scope of RAG from computational optimization to the domain of educational interaction and institutional communication.

From a methodological standpoint, this article adopts a design-based research approach, integrating theoretical analysis with technological implementation and empirical evaluation. The study unfolds through a sequence of interconnected sections that together describe the conceptual, technical, pedagogical, and ethical foundations of the project.

Section 2, *Between Bioinspiration and Responsibility in AI*, introduces the philosophical framework underpinning the research. Drawing on Parisi's concept of associative memory and Floridi's notion of responsible agency, it situates the NPC within a broader discourse on bioinspired intelligence and ethical governance in artificial systems.

Section 3, *Technological Framework*, details the computational architecture of the system. It explains how the RAG model was engineered and implemented, outlining its layered structure and the mechanisms that ensure transparency, traceability, and reproducibility.

Section 4, *Pedagogical and Ethical Framework*, translates these technical affordances into an educational paradigm. It defines how the system supports inclusive learning processes, embeds ethical safeguards, and operationalizes the principles of Artificial Intelligence for Social Good (AI4SG) within a human–AI cooperative loop.

Section 5, *Description of the NPC*, describes the concrete instantiation of this framework in the context of Città della Scienza, illustrating how the NPC functions as both an interactive educational guide and a digital ambassador for institutional identity. Section 5.1 *Implementation Details of the NPC System* provides a transparent account of the development pipeline, enabling replication by other researchers.

Finally, Section 6, *Toward a Socio-Technical Framework for Responsible AI Integration*, extends the discussion to a macro level, proposing a systemic reflection on the relationship between educational institutions, technological innovation, and ethical regulation. The paper concludes with Section 7, which presents the pilot study conducted at Città della Scienza, empirically validating the NPC's communicative, pedagogical, and affective effectiveness.

Through this structure, the article connects theoretical reflection, technological modeling, and empirical observation into a unified framework, demonstrating how Retrieval-Augmented Generation (RAG) can be transformed from a computational paradigm into a pedagogically grounded, ethically governed, and socially inclusive educational practice.

Against this background, the primary aim of the present study is to demonstrate how Retrieval-Augmented Generation can be systematically operationalized within an embodied, pedagogically grounded Non-Playable Character designed for informal learning environments. The originality of the work lies not in proposing a novel language model, but in integrating RAG, educational theory, ethical governance, and institutional identity into a coherent socio-technical framework. By doing so, the study contributes a transferable design paradigm for responsible, educationally oriented AI systems in museums and cultural heritage contexts.

2. Between Bioinspiration and Responsibility in AI

In his 10 June 2025 address to the Italian Parliament, focused on the topic of artificial intelligence, Nobel laureate Giorgio Parisi [8] drew attention to a pivotal figure in the history of neuroscience, Camillo Golgi, who as early as 1912 elucidated the structure of the neuron through a distinctive staining technique involving a black solution. Parisi proposed a parallelism between biological and artificial systems (this pertains to the concept of bioinspiration) arguing that this comparison is far more significant than previously acknowledged by other scholars, including Luciano Floridi [9], whose perspective will be examined subsequently.

Parisi particularly emphasized the significance of the dendritic tree structure, associated with neuronal excitation and inhibition phenomena, well documented in neurological literature [10]. These dynamics play a central role in associative memory processes; whereby partial information can reactivate a more complete or complex memory trace. Floridi, also in his address to the Italian Parliament [10] on the other hand, offers a different interpretation: artificial intelligence does not constitute a novel form of intelligence per se, but rather an unprecedented mode of agency. According to Floridi, the value of AI lies less in

its cognitive capabilities and more in its operational nature, which introduces a form of technological agency hitherto unseen. Floridi stresses the imperative to prevent the misuse of AI and insists that those who develop these technologies must assume responsibility for their potential consequences. Nonetheless, it remains ambiguous whether such warnings are directed solely at human actors or, implicitly, at technology itself. The central issue thus becomes the question of autonomy: does the primary risk stem from Dr. Frankenstein's malevolence or from the inherent danger of his creation? A paradigmatic case illustrating this tension is Amazon, whose CEO Andy Jassy disclosed that the widespread adoption of generative AI agents will result in significant reductions in corporate roles in the coming years [10].

Jassy has urged employees to engage with AI tools and to "do more with less." This exemplifies how human decisions shape the social impact of AI, underscoring the need to clarify who, or what, should be ethically constrained. This reflection lies at the core of Floridi's viewpoint: responsibility cannot be delegated to technology but requires deliberate intentionality on the part of developers and policymakers, who must steer innovation towards outcomes that uphold equity and human dignity. Considering this, there has been a proliferation of ethical codes, guidelines, and declarations from institutions, states, and associations, each eager to contribute to the discourse on how AI should be regulated. Every new initiative in artificial intelligence tends to generate further statements of principles and values, creating an impression of a competitive race to participate. Initially motivated by a collaborative spirit, many of these declarations have evolved into attempts to assert proprietary ownership of the ethical narrative, "mine and mine alone." Years later, the risk persists that these efforts may produce redundant or overlapping principles or, conversely, divergent frameworks that engender confusion and ambiguity.

Considering these reflections, the philosophical perspectives offered by Parisi and Floridi can be directly related to the conceptual and technical foundations of the present project. The Non-Playable Character (NPC) designed for Città della Scienza embodies Parisi's notion of bioinspiration, translating neurobiological principles such as associative memory and adaptive recall into an artificial system that emulates processes of contextual retrieval and reactivation within a digital architecture. The Retrieval-Augmented Generation (RAG) framework mirrors the associative function of the human dendritic network, where partial cues elicit complete knowledge traces through distributed connections. This analogy is not merely metaphorical but operational, as the retrieval component of the system enables the NPC to reconstruct meaning dynamically, drawing from a curated institutional corpus in ways that parallel human memory reconstruction.

At the same time, the project responds to Floridi's argument concerning technological agency and responsibility by embedding explicit ethical safeguards into the system design. The NPC does not act autonomously as a self-determining agent; rather, it operates within a meta-autonomous structure that preserves human oversight and reversibility. Each of its outputs is grounded in verified sources, reviewed by human educators, and traceable to its documentary origin. In this sense, the system materializes Floridi's principle of responsible agency, transforming philosophical reflection into a tangible architecture for transparent and accountable human–AI interaction.

This integration between theoretical reflection and practical implementation situates the Non-Playable Character (NPC) as both a didactic and ethical prototype, a living laboratory where the boundaries between artificial and human cognition are explored under controlled, pedagogically informed conditions. It thus exemplifies how an AI system can operationalize philosophical ethics and cognitive models into a concrete tool for inclusive scientific communication and AI literacy, aligning with the principles of Artificial Intelligence for Social Good (AI4SG) and the European AI Act's human-centric approach [11].

2.1. Ethical Responsibility and the Limits of Technological Autonomy

Floridi emphasizes that it is evident both that human autonomy must be promoted and that machine autonomy should be limited and made intrinsically reversible whenever human autonomy needs to be protected or restored (for example, in the case of a pilot able to deactivate the autopilot and regain full control of the aircraft). This introduces a concept that can be defined as meta-autonomy, or a model of delegated decision-making. Humans ought to retain the authority to decide which decisions to make, exercising freedom of choice where necessary and relinquishing it in cases were overriding considerations, such as effectiveness, may justify the loss of control over the decision-making process. However, any delegation should, in principle, remain revisable, adopting as a final safeguard the power to decide to decide again. Parisi also emphasizes the importance of enabling vulnerable individuals to use generative AI as a psychologist and tutor, particularly young people seeking support. For example, a student might ask, 'Write an essay on Julius Caesar in the style of a 13-year-old,' thereby undermining the value of the exercise" which simultaneously undermines the value of the exercise [12,13]. According to Parisi, AI is becoming increasingly significant in education; previously, the internet was the primary tool, but now AI has taken on this role. It is essential to teach students how to critically select information in school. Whereas selection was once based on the authority of sources, the current integration of AI presents a complex challenge: how can students navigate this blended informational environment? This represents a major educational challenge moving forward. According to the Nobel laureate, the solution lies in clearly defining the sources even when using generative AI.

The issue at hand is not related to copyright in the traditional sense, but rather to the right of inclusion within such AI systems, a user's right to access and engage with the content. He argues that the way forward is to prevent de facto monopolies and cites several dominant actors as examples: Google (Alphabet), with its search engine, online advertising, Android, and YouTube; Microsoft, with its Windows operating system, Office suite, and Azure cloud services; and Intel, known for its PC microprocessors and, additionally, NVIDIA for graphics cards. To these must be added Amazon, which leads in e-commerce and cloud computing through AWS; Meta (Facebook), which controls social networks such as Facebook, Instagram, and WhatsApp; and Samsung, a major player in Android smartphones, semiconductors, and display technologies. Sadin [10] (philosopher and writer, he is considered one of the most prominent and perceptive critics of new technologies), in his address to the Italian Parliament, highlights that as early as 2014 in France, François Hollande had asserted that within one year all students would exclusively use tablets in schools. However, Sadin argues that this approach sacrifices an entire generation, causing them to lose valuable traditional habits in favor of hype driven by the interests of IT and technology lobbies. Now, will the same happen with AI? What truly matters is recognizing that technologies are not meant to replace but to complement existing practices.

From an ethical and governance perspective, the development and deployment of the NPC strictly adhered to the principles of transparency, accountability, and human oversight defined by the European Artificial Intelligence Act [14]. Given that Non-Playable Character (NPC) reproduces the facial likeness, gestures, and synthesized voice of a real individual—the President of the IDIS Foundation—explicit informed consent was obtained before all data acquisition and model training procedures. The consent process followed the standards established by Regulation (EU) 2016/679 (General Data Protection Regulation—GDPR) [11] and by the Council of Europe Convention for the Protection of Individuals concerning Automatic Processing of Personal Data (Convention 108+) [12],

ensuring that biometric and vocal data were processed exclusively for institutional and educational purposes, with no commercial or promotional use.

The ethical evaluation of the project was carried out by the Teaching and Learning Centre for Education and Inclusive Technologies "Elisa Frauenfelder" at the University of Salerno, in compliance with institutional research integrity policies and the UNESCO Recommendation on the Ethics of Artificial Intelligence [14]. All digital assets—facial scans, 3D meshes, and audio recordings—were stored in encrypted form and integrated into the system through reversible identifiers, allowing full traceability, revocability, and right to erasure in line with the *data minimization* principle of the GDPR.

Moreover, a human-in-the-loop governance model was implemented to ensure continuous ethical oversight: any modification to Non-Playable Character (NPC)'s knowledge base, behaviour, or communicative parameters must be approved by the Foundation's scientific committee before deployment. This governance framework guarantees that all system responses generated through the RAG pipeline remain consistent with validated institutional sources and uphold the dignity, autonomy, and representational rights of the person portrayed. By embedding informed consent, procedural transparency, and auditable governance mechanisms into the system's design, the project offers a concrete example of responsible innovation aligned with the principles of AI for Social Good (AI4SG) and the regulatory trajectory established by the European Union.

2.2. Technological Determinism and the Need for Critical Education

New media should be understood as cultural artifacts that necessitate the development of both individual and collective responsibility, as well as critical thinking. The idea is to, in the words of Tisseron, accompany, to alternate, to ensure that the younger generations are capable of self-regulation between traditional and real media [14]. According to Sadin, there is an illusion in natural language processing that operates through the extrapolation of semantic rules which produce logical laws based on statistical analyses. The objective is to identify automatic correlations. From this point begins the necrosis of text generation, as we exist within a "regime" of probability determined by what has already occurred. In practice, what happens is simply what must happen [15,16]. This stands in stark contrast to creative thinking. What is language? It is the most emblematic space of our encounters, the shared heritage, and the power to empower. It becomes evident that technological determinism can occur [17,18], and all this stems from a process that begins in school, starting from early childhood, where the shared heritage is encountered. According to Sadin, what truly matters is resisting the utilitarian logic underlying the use of LLMs and the culture of copy-and-paste. He advocates for a collective affirmation of a fundamental principle from *Émile ou de l'éducation* written by Jean-Jacques Rousseau: the most important rule is that the most important thing is *not* to save time [19].

Rather, it is the ability to *waste* time that holds educational value, as learning inherently involves a form of temporal investment that resists efficiency. In this light, LLMs should not be employed merely to complete tasks devoid of genuine interest, but instead to foster meaningful engagement, for example, through practices such as question time that stimulate critical reflection and dialogue. Sadin advances the theory that AI systems, designed to apologize, accommodate, and offer fully customized responses without resistance, stand in stark contrast to human educators, who represent an "otherness" in relation to the student, including in generational terms. According to this view, such frictionless interactions risk fostering the development of "little tyrants," as learners are no longer challenged by the presence of a distinct and authoritative interlocutor. For this reason, increasing difficulties in coexisting and engaging in shared social life are likely to emerge. According to Sadin, who described the automatic generation of texts as necrotic, we are

facing a struggle against the producers of large systems previously mentioned also by Parisi. It is essential to preserve what remains alive within us; otherwise, we risk entering a form of humanity that is absent to itself. Although Sadin adopts a critical stance that frames artificial intelligence as a fundamentally utilitarian form of action, and often expresses apocalyptic tones in his forecasts, it remains essential to consider the broad spectrum of academic perspectives on the subject. Given that data concerning human cognitive systems are still being gathered and analyzed, it is fundamental to include a diverse range of expert viewpoints. This plurality enables a more nuanced understanding of the ethical and educational implications of AI, fostering an interdisciplinary dialogue that enriches the ongoing debate. Maria Chiara Carrozza [16] adopts a notably more reassuring stance in this debate, perhaps due to her engineering-oriented perspective and her focus on artificial intelligence as applied to robotics. Nevertheless, she observes the pervasive influence of utilitarian logic among school students. However, she also argues that AI, when applied to assistive technologies such as exoskeletons, will be more readily accepted because it enhances our ability to live, this is the central concern of neuro-robotics, where the robotic component is effective rather than clumsy, and does not impair but rather improves human movement. Another important aspect is the potential role of robotics in supporting individuals with autism. While robots are not meant to replace therapists or special education teachers, they can nonetheless perform a range of useful tasks that complement human intervention. For example, neural networks are "redesigning" the way a robotic hand grasps a bottle, not by relying on sensors and pre-programmed physical equations, but by inferring such equations through statistical approximations derived from supervised and unsupervised trial-and-error learning.

This process challenges the boundary between the natural and the artificial. Similarly, a hip prosthesis replacing a deteriorated section of bone becomes part of a complex interaction involving biocompatibility, tissue regeneration within the prosthetic structure, and the restoration of the person's ability to walk. In doing so, it crosses the boundary between natural and artificial, establishing a new state of equilibrium. It becomes necessary to collaboratively define the rules governing this new equilibrium. To illustrate the complexity and trade-offs involved in balancing technological development with ethical considerations, it is worth noting that Google has recently announced its adherence to the new Code of Conduct on Artificial Intelligence proposed by the European Commission. However, in an official statement, Kent Walker, President of Global Affairs at Google, while reaffirming the company's commitment, voiced significant concerns regarding the potential negative impact this regulatory framework could have on innovation and technological advancement in Europe, an observation that has been widely discussed across various industry blogs [17].

These considerations find a direct correspondence within the educational rationale of the Non-Playable Character (NPC) developed for Città della Scienza. In contrast to the utilitarian logic criticized by Sadin, the NPC is not designed to accelerate learning or replace human mediation, but to reintroduce time and reflection into the digital learning process. Its dialogic structure encourages users to formulate questions, analyze sources, and engage in sustained critical interaction, thereby transforming the interaction from a passive consumption of information into an active cognitive dialogue. The Retrieval-Augmented Generation (RAG) architecture supports this approach by grounding each response in documented evidence, allowing users to trace and verify the origins of the information provided. In this way, the system aligns with the pedagogical principle emphasized by Rousseau and reiterated by Sadin: learning as a temporal and interpretive investment rather than a mechanical task completion.

Furthermore, the NPC is conceived not as an alternative to educators but as a complementary didactic tool that can mediate between scientific content and diverse publics.

Following Carrozza's perspective, the project interprets artificial intelligence as an assistive and enabling technology that expands human capacity without eroding human agency. Within this framework, AI becomes a pedagogical prosthesis, enhancing accessibility for individuals with different cognitive or linguistic needs while preserving the irreplaceable relational dimension of teaching and communication. The system's design thus embodies an ethical equilibrium between technological innovation and human responsibility, reflecting a form of human–machine cooperation aimed at sustaining curiosity, empathy, and inclusive participation in the cultural and scientific sphere.

3. Technological Framework

To enhance readability and support interdisciplinary accessibility, the system architecture and operational workflow are summarized through diagrams, flowcharts, and consolidated tables, which visually complement the descriptive sections of the text.

The technological framework supporting the Digital Personal Tutor is conceived as a multi-layered and modular architecture that integrates retrieval-based and generative components into a unified, traceable pipeline. It is organized into four interdependent subsystems—Data and Knowledge Management, Retrieval Layer, Generative Layer, and Interface and Application Layer—which communicate through standardized API endpoints and asynchronous message passing [18,19]. This structure ensures scalability, interoperability, and compliance with transparency and reproducibility principles defined in the European Artificial Intelligence Act [14].

The system's generative layer relies on OpenAI's GPT-3.5-Turbo, fine-tuned through Reinforcement Learning from Human Feedback (RLHF) on institutional materials provided by Fondazione IDIS—Città della Scienza. The fine-tuning phase adjusted the model's stylistic register to align with the discourse of scientific dissemination—concise, factual, and inclusive—without modifying its base weights. Training was conducted over three epochs on approximately 2.6 million tokens derived from 3800 institutional documents, with a learning rate of 1×10^{-5} and early stopping activated upon validation-loss stabilization across two iterations.

During inference, the model operates with a temperature of 0.2, top-p of 0.9, and a maximum generation length of 1024 tokens, initialized initialized with a fixed random seed (42) to ensure reproducibility. Contextual evidence retrieved from the FAISS (Facebook AI Similarity Search)–based vector database (embedding model: *text-embedding-ada-002*, 1536-dimensional vectors) is concatenated with user queries through a structured prompt template combining instruction, query, and top-k retrieved fragments (k = 5–10).

Each generated output undergoes semantic validation via a source-attribution mechanism that assigns a Semantic Confidence Score (SCS); responses below the empirical threshold of 0.78 trigger iterative retrieval and reformulation cycles to minimize hallucinations. The corpus itself is version-controlled and timestamped, enabling progressive enrichment through a semi-automated ingestion workflow that maintains traceability, transparency, and data provenance in line with the European Artificial Intelligence Act (Regulation EU 2024/1689) [15].

All training, inference, and evaluation procedures are documented within a traceable and auditable pipeline, allowing full replication of the system's configuration and parameterization. This methodological transparency operationalizes the principles of reproducibility, accountability, and human oversight, positioning the Digital Personal Tutor as a verifiable and ethically compliant implementation of Retrieval-Augmented Generation (RAG) within educational and institutional contexts [20–22].

Although RLHF (Reinforcement Learning from Human Feedback)-based stylistic alignment was applied, no full model fine-tuning in the strict sense was performed. The

system primarily relies on Retrieval-Augmented Generation using a base GPT-3.5-Turbo model, with controlled prompt conditioning and retrieval grounding.

3.1. Data and Knowledge Management Layer

At the lowest level, the system relies on a FAISS-based vector database indexing a curated institutional corpus composed of approximately 3800 documents (≈2.6 million tokens) provided by Fondazione IDIS—Città della Scienza. Each document undergoes a preprocessing pipeline including text normalization, metadata enrichment, and token-level semantic segmentation (200–300 token windows). Embeddings are generated using OpenAI text-embedding-ada-002, producing 1536-dimensional dense vectors stored with unique SHA-256 identifiers and timestamped metadata for version control. The corpus constitutes the non-parametric memory of the Retrieval-Augmented Generation (RAG) system, allowing dynamic updates without model retraining. Semantic similarity is computed via cosine-similarity distance ($0 \leq \cos\theta \leq 1$), providing a relevance ranking for each fragment.

Data access is managed through a RESTful interface supporting both lexical-semantic hybrid queries and batch vector searches. The architecture enables incremental ingestion of new materials using automated crawlers and curator approval workflows, ensuring data provenance and epistemic traceability.

To ensure full technical reproducibility, all hyperparameters of the Retrieval-Augmented Generation (RAG) pipeline, including FAISS index configuration, embedding chunking strategy, retrieval thresholds, prompt templates, and confidence metrics, are explicitly reported in Table 1. This documentation enables independent replication and auditability of the system configuration (Figure 1).

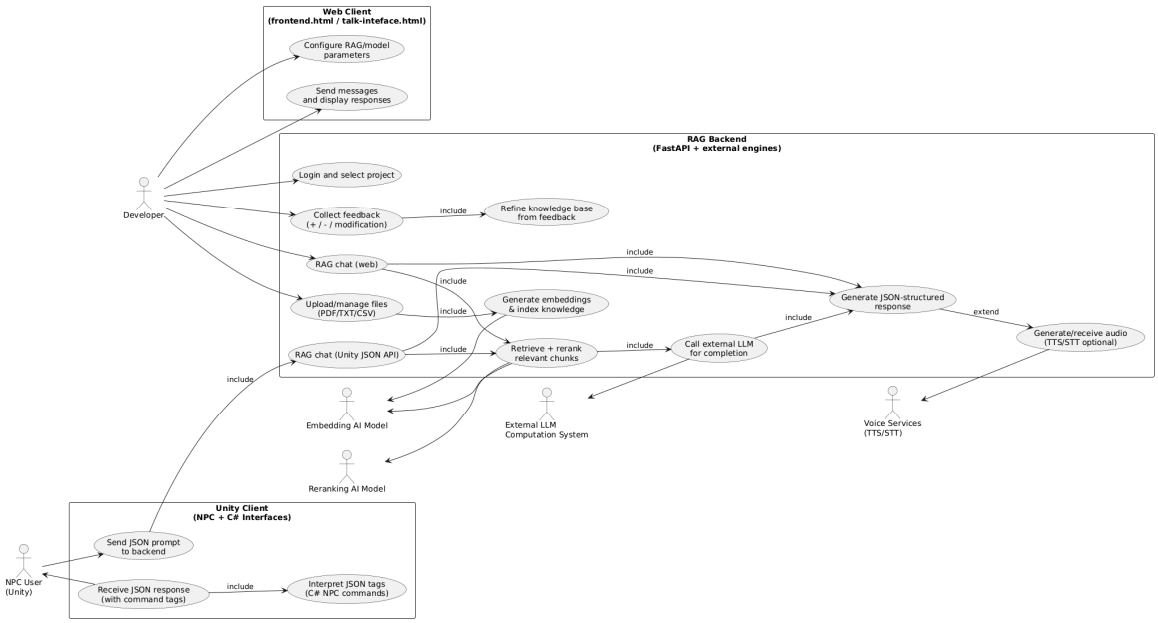

Figure 1. UML use case diagram of the proposed RAG-based conversational system. The diagram illustrates the end-to-end architecture of the project, highlighting the interaction between the web-based development environment, the RAG backend, external AI services, and the Unity-based NPC interface. It emphasizes document management, semantic embedding, retrieval and re-ranking processes, multi-LLM integration, and the delivery of structured JSON outputs to support real-time NPC interactions and continuous knowledge base refinement.

Table 1. Unlike existing systems, the proposed NPC integrates RAG within an explicit pedagogical–ethical framework, operationalizing inclusivity, traceability, and institutional identity.

System	Model	RAG Type	Users	Metrics
DocentGemma	LoRA + RAG	Hybrid	110	Accuracy 85.55%
MAGICAL	GPT-4	RAG	62	UX, Trust
Hakka Chatbot	LLM + RAG	Cultural	90	Engagement
This work	GPT-3.5 + RAG	Pedagogical	77	PCS, ERI, CFA

3.2. Retrieval Layer

The retrieval subsystem executes a top-k contextual search (k = 5–10) based on the cosine-similarity scores of the embedding vectors. Each query is first vectorized, then processed through a hybrid weighting scheme combining semantic proximity ($\alpha = 0.7$) and lexical overlap ($\beta = 0.3$). Retrieved segments are ranked, normalized, and injected into a context window manager that prevents token overflow by truncating content beyond N = 3500 tokens. A semantic confidence score (SCS) is computed for each retrieved unit:

$$SCS_i = \frac{1}{n}\sum_{j=1}^{n}\left(1 - \cos\left(\theta_{ij}\right)\right)$$

Responses with $SCSi < 0.78$ trigger an iterative retrieval loop until the confidence threshold is satisfied. This mechanism reduces hallucination probability and guarantees factual grounding of generative outputs (Table 1).

The UML use case diagram illustrates the end-to-end architecture of the proposed knowledge-based conversational system, which integrates a web-based development environment, an intelligent Retrieval-Augmented Generation (RAG) backend, and a Unity-based Non-Playable Character (NPC) interface. The system is designed to manage large collections of user-provided documents, generate semantic embeddings, perform retrieval and re-ranking operations, and produce context-aware responses through an external large language model (LLM) computation service. A wide range of LLMs is supported, including GPT-based models, DeepSeek, Llama, Mistral, and others, which can be freely selected and dynamically switched by the developer at runtime, even on a per-interaction basis. Generated outputs are returned as structured JSON messages, enabling direct interpretation and execution by NPCs within the Unity environment.

The workflow is initiated by the developer, who is responsible for curating and maintaining the knowledge base used by the system. Through the web interface, the developer can access or create projects, upload documents in various formats (e.g., PDF, text files, or preprocessed datasets), and configure parameters governing the RAG pipeline and the behavior of the language model. These parameters include similarity thresholds, the maximum number of retrieved chunks, temperature and sampling settings, and optional system-level instructions that influence conversational style and response generation.

Upon document ingestion, the backend automatically triggers an embedding phase in which specialized embedding models transform document segments into numerical vector representations. These vectors are indexed and stored within a project-specific memory structure, forming the basis for efficient semantic search and retrieval. When a message is submitted—either by the developer via the web interface or by an NPC user through Unity—the RAG workflow is activated. The system retrieves the most semantically relevant chunks from the indexed knowledge base and applies an AI-based re-ranking model to prioritize the information most useful for answering the query. The selected context,

combined with the user input and predefined system instructions, is then forwarded to the external LLM service, which generates a natural language response enriched by the retrieved knowledge. Finally, the response is delivered as a structured JSON output, supporting seamless integration with NPC behaviors and enabling iterative refinement of the knowledge base through user and developer feedback.

3.3. Generative Layer

The generative module is based on GPT-3.5-Turbo, fine-tuned for museum communication through reinforcement learning from human feedback (RLHF). It represents the parametric memory of the architecture. The RAG pipeline merges the retrieved textual evidence with the user query through prompt concatenation governed by a structured template: [Instruction Layer] → [User Query] → [Top-k Retrieved Contexts] → [Generation Parameters].

The generation parameters include temperature ($\tau = 0.2$), top-p sampling (0.9), and max token length (1024). The model applies cross-attention mechanisms to weigh retrieved segments against its internal knowledge base, resulting in outputs that are both contextually coherent and epistemically grounded. A source-attribution module tags each response fragment with document identifiers, ensuring full backward traceability (Table 2).

Table 2. Consolidated system parameters for the RAG-based NPC system.

Category	Parameter	Value/Configuration	Notes for Replicability
Generative Model	LLM	GPT-3.5-Turbo	OpenAI model, fine-tuned via RLHF on institutional corpus
	Fine-tuning method	Reinforcement Learning from Human Feedback (RLHF)	No modification of base weights
	Training epochs	3	Early stopping after validation-loss stabilization
	Training tokens	≈2.6 million	Derived from ~3800 institutional documents
	Learning rate	1×10^{-5}	Fixed during fine-tuning
Inference Parameters	Random seed	42	Ensures deterministic reproducibility
	Temperature (τ)	0.2	Low temperature to reduce stochasticity
	Top-p (nucleus sampling)	0.9	Controls output diversity
	Max generation length	1024 tokens	Upper bound for response length
Retrieval Chunking	Chunk size	200–300 tokens	Semantic segmentation window
	Chunk overlap	Not applied	Non-overlapping segments
Retrieval Strategy	Top-k retrieved chunks	k = 5–10	Adaptive range depending on query complexity
	Similarity metric	Cosine similarity	Applied in embedding space
Hybrid Weighting	Semantic weight (α)	0.7	Emphasizes semantic proximity
	Lexical weight (β)	0.3	Accounts for keyword overlap
Confidence Control	Semantic Confidence Score (SCS)	Threshold = 0.78	Below threshold triggers iterative retrieval

Table 2. *Cont.*

Category	Parameter	Value/Configuration	Notes for Replicability
Embedding Model	Embedding model	text-embedding-ada-002	OpenAI embedding model
	Embedding dimension	1536	Dense vector representation
Vector Database	Engine	FAISS	Facebook AI Similarity Search
	Indexing type	Dense vector index	Optimized for cosine similarity
	Document identifiers	SHA-256 hashes	Ensures traceability and versioning
	Metadata	Timestamped, version-controlled	Supports corpus evolution and auditing
Context Management	Max context window	3500 tokens	Prevents overflow before generation
Validation Loop	Iterative retrieval	Enabled	Activated when SCS < 0.78
Logging & Governance	Provenance tracking	Enabled	Source-attribution per response
	Human-in-the-loop review	Enabled	Scientific committee validation

3.4. Interface and Application Layer

The upper layer of the system transforms computational outputs into a multimodal interactive experience. At this level, several components work together to ensure a natural and coherent dialogue with users. Speech recognition, implemented through the Whisper API, enables accurate multilingual transcription of spoken input. The Dialogue Manager then organizes the conversation flow, maintaining contextual continuity and storing interaction history through an SQLite buffer. The Speech Synthesis and Animation Module combine Azure Cognitive Services for multilingual voice cloning with Unity3D's Mecanim system, which synchronizes facial expressions and gestures, creating realistic and responsive communication (Unity 6.3 LTS (6000.3.7f1). Finally, a Feedback Logger continuously records user prompts, retrieved information, and system responses, allowing developers and educators to audit the interactions and refine the system's performance over time.

All components of the system are connected through a lightweight middleware developed in Python using the Flask framework (Python 3.13.12). This layer manages essential operations such as session tokens, concurrency control, and latency optimization, ensuring a mean response time of less than 350 ms. A dedicated human-in-the-loop governance interface allows educators and researchers to oversee the system's functioning in real time. Through this interface, they can examine the sources retrieved by the model, validate or correct responses, and flag outputs that require retraining or corpus updates. This mechanism ensures continuous ethical supervision and full adherence to institutional data governance and transparency policies.

The overall information flow follows a structured yet flexible sequence. During the input phase, the user's utterance is captured and transcribed by the Whisper model, which then converts it into vectorized text. In the retrieval phase, the system performs a semantic search using FAISS, ranking the most relevant contextual fragments and validating them through the Semantic Confidence Score (SCS). Next, in the generation phase, the retrieved context is integrated into the prompt to guide the language model (LLM) in producing a coherent and source-attributed response. Finally, during the output phase, the text is synthesized into speech, synchronized with facial animation, and logged together with user feedback for subsequent analysis.

This sequential but feedback-oriented process enables the system to learn dynamically from validated interactions. By progressively re-embedding new verified content, the architecture functions as a closed-loop RAG environment in which human supervision, algorithmic reasoning, and institutional knowledge continuously interact to generate transparent, reliable, and pedagogically meaningful communication.

3.5. Epistemological Implications

From a theoretical standpoint, the framework exemplifies the shift from rule-based symbolic AI to statistical and connectionist inference [23–25]. Knowledge representation emerges from distributed vector spaces rather than explicit logic rules, reflecting a transition from *deductive certainty* to *probabilistic reasoning*. This evolution not only underpins the technical design of the NPC but also redefines the epistemic status of machine learning as an adaptive, data-driven process capable of modeling uncertainty and context. In this sense, the RAG-based NPC extends the notion of intelligence from syntactic manipulation to document-grounded inference, bridging computation, epistemology, and educational mediation.

4. Pedagogical and Ethical Framework

The Pedagogical and Ethical Framework defines the human-centered operational logic that governs how the Digital Personal Tutor translates technological functionality into measurable educational value [26–28]. It integrates cognitive, relational, and normative dimensions within a human–AI cooperative loop designed to ensure transparency, inclusivity, and accountability. The framework follows a cybernetic structure, where inputs (retrieved knowledge), mediations (educational dialogue), and outputs (learning actions) are continuously monitored, evaluated, and adapted through human oversight.

4.1. Pedagogical Architecture

The pedagogical subsystem is structured around three closely interconnected components: adaptive learning mediation, active documentation, and metacognitive feedback. Together, these elements enable the system to deliver personalized, evidence-based, and self-reflective learning experiences.

In the first component, adaptive learning mediation, the Non-Playable Character (NPC) acts as a context-sensitive tutor capable of adjusting its instructional strategies according to the learner's profile and ongoing learning trajectory. This adaptive capacity is guided by metadata associated with each retrieved text segment—such as topic, level of complexity, and readability—which allows the Dialogue Manager to dynamically select the most appropriate pedagogical register for each interaction.

To achieve this modulation, the system relies on a set of pedagogical variables, including the Linguistic Complexity Index (CLI), the Cognitive Load Estimation (CLE), and the Interaction Persistence (IP), which together provide a multidimensional representation of the learner's cognitive and communicative needs. These indicators are integrated within a control mechanism that calculates a Pedagogical Coherence Score (PCS)—a metric expressing the degree of alignment between the user's profile and the characteristics of the retrieved educational content. When this alignment falls below a predefined threshold, the system automatically adapts its communicative strategy, simplifying or expanding the explanation as needed to maintain cognitive accessibility and pedagogical coherence:

$$PCS = \frac{w_i a \cdot i}{\sum_{i=1} w_i w_i}$$

where a_i represents alignment metrics (lexical level, semantic relevance, temporal proximity), and w_i the respective importance weights. A PCS below 0.65 triggers adaptive

simplification or expansion routines [29]. Unlike purely generative models, RAG introduces a document-centric pedagogy, where learning emerges from evidence retrieval rather than opaque synthesis. Each NPC response includes explicit source references and, where applicable, short contextual excerpts.

This transforms the learner's engagement from passive reception to active verification, reproducing the logic of academic inquiry. The system encourages iterative questioning: responses with high user dwell time (>10 s) are automatically logged as potential "learning anchors," prompting follow-up suggestions derived from semantically adjacent document clusters. The framework includes a metacognitive feedback module that quantifies interaction quality through user behaviors (clarification requests, repetitions, reformulations). These metrics inform an Engagement and Reflection Index (ERI):

$$ERI = \frac{C_r + R_f}{T_i}$$

where C_r = clarification requests, R_f = reflective follow-ups, and T_i = total interactions. ERI values between 0.3 and 0.6 indicate balanced cognitive challenge; persistent values below 0.2 trigger teacher intervention prompts.

Together, these mechanisms operationalize a feedback-controlled learning loop, ensuring that each interaction with the NPC maintains educational relevance, cognitive accessibility, and user agency.

The Pedagogical Coherence Score (PCS) and the Engagement and Reflection Index (ERI) are proposed as exploratory metrics. Content validity was established through expert review by five specialists in educational technology and museum pedagogy. Construct validity is currently limited to internal coherence; criterion validity will be addressed in future work through correlation with established instruments such as the User Engagement Scale–Short Form, System Usability Scale, and Bot Usability Scale (BUS-15).

4.2. Ethical Governance Architecture

The ethical dimension of the system is implemented as a compliance-by-design layer that permeates all its components. This layer ensures that ethical principles are not added post hoc but are structurally integrated into the system's functioning. It operates through four complementary domains: transparency, human oversight, informational pluralism, and accessibility.

The first domain, transparency and traceability, guarantees that every generative output can be traced back to its origin. Each response is associated with a *provenance vector*—a structured metadata tuple $\langle D_i, T_s, SCS_i, UID \rangle$, where D_i identifies the source document, T_s records its timestamp, SCS_i indicates the semantic confidence score, and UID represents a unique identifier. These metadata are stored in a tamper-evident registry accessible to educators and reviewers, thereby fulfilling the traceability requirements established under Article 13 of the European Artificial Intelligence Act [14]. This mechanism enables full auditability of the NPC's outputs and reinforces user trust in the integrity of its informational processes.

The second domain, human oversight and meta-autonomy, draws on Floridi's concept of reversible autonomy. The NPC functions under a *reversibility protocol*, allowing human supervisors to intercept or undo automated outputs at any stage. A *Supervision Index (SI)* quantifies the proportion of educator-validated interactions compared to the total number of system responses. When this index falls below 0.85, the system automatically shifts to a *guided mode*, temporarily limiting autonomous generation and ensuring closer human monitoring. This arrangement preserves a balance between algorithmic efficiency and human accountability, maintaining human agency as the ultimate decision-making authority.

The third domain, informational pluralism, safeguards epistemic diversity within the NPC's knowledge base. Source variety is maintained algorithmically through an entropy-based corpus balancing mechanism, expressed as $H = -\Sigma_1{}^m\ p_i\ \log(p_i)$, where each p_i corresponds to the proportion of content derived from a specific institutional or thematic cluster. When the entropy value drops below 1.2, indicating the predominance of a single perspective, the system automatically issues corpus enrichment alerts. This process prevents informational monopolies and ensures that the NPC's knowledge output reflects multiple, balanced viewpoints—addressing concerns raised by Parisi and Sadin about the concentration of informational power among major technological actors [9,10].

Finally, the accessibility and inclusion domain ensures that the NPC is universally usable and inclusive. Interaction is multimodal—combining text, speech, and visual cues—in accordance with the *Universal Design for Learning (UDL)* guidelines. Periodic evaluations assess key aspects such as speech intelligibility, interface readability, and linguistic equity across supported languages (Italian, English, and Arabic). The results contribute to an *Accessibility Score (AS)*, calculated as the average of five usability dimensions: clarity, legibility, responsiveness, linguistic parity, and input flexibility. These measures ensure that the system remains accessible to users with diverse cognitive, linguistic, and sensory profiles, upholding its commitment to inclusive education and communication.

4.3. Pedagogical–Ethical Integration Matrix

The pedagogical and ethical subsystems converge through an integration matrix that aligns learning functions with governance safeguards (Table 3).

Table 3. Relation between subsystem, pedagogical function, ethical safeguard and observable metric of the project.

Subsystem	Pedagogical Function	Ethical Safeguard	Observable Metric
Knowledge Engine	Evidence-based retrieval and contextual learning	Source traceability (PV tuples)	SCS, PCS
Dialogue Manager	Adaptive tutoring and linguistic modulation	Human-in-the-loop supervision	SI, ERI
Curated Corpus	Documentary pluralism and reflective inquiry	Entropy-based diversity control	H-index
User Interface	Multimodal engagement and accessibility	UDL compliance verification	AS

This coupling ensures that every pedagogical process is mirrored by an ethical constraint, establishing a dual-control mechanism where cognitive benefit and moral integrity remain co-dependent.

The pedagogical design of the NPC can be further situated within established educational theories that clarify its instructional role in informal learning environments. In particular, the adaptive behavior of the NPC aligns with Lev Vygotsky's concept of the Zone of Proximal Development, as the system dynamically modulates the depth, complexity, and linguistic register of its explanations in response to user input, effectively functioning as a form of digital scaffolding. Rather than delivering static information, the NPC supports learners in progressing from initial curiosity toward higher levels of conceptual understanding through guided dialogue and incremental elaboration.

At the same time, the system reflects the principles of the The Museum Experience Contextual Model of Learning, which emphasizes the interaction between personal, sociocultural, and physical contexts in museum-based learning. The NPC explicitly integrates

these dimensions by embedding scientific content within the institutional narrative of Città della Scienza, adapting responses to the visitor's linguistic and cultural background, and leveraging the embodied, spatial setting of the museum to foster meaning-making beyond mere information transfer.

More broadly, the framework adopts a constructivist orientation in which knowledge is co-constructed through interaction rather than transmitted unidirectionally. The Retrieval-Augmented Generation architecture operationalizes this principle by grounding dialogue in documentary evidence while simultaneously encouraging questioning, reflection, and verification. In this sense, the NPC functions not as an authoritative instructor, but as a mediating educational agent that supports active engagement, interpretive agency, and metacognitive awareness within an inclusive and ethically governed learning environment.

4.4. Systemic Function and Epistemic Outcome

Functionally, the framework can be modeled as a triadic feedback system:

1. Technological substrate (retrieval → generation) provides data-grounded inference.
2. Pedagogical loop (interaction → reflection → adaptation) ensures meaningful learning.
3. Ethical loop (traceability → oversight → compliance) guarantees accountability.

The superposition of these loops creates a closed educational–ethical circuit, where each AI-mediated exchange is empirically measurable, contextually interpretable, and normatively auditable. The NPC thereby transitions from a generative artifact to an adaptive, auditable Digital Personal Tutor, capable of maintaining epistemic reliability, human supervision, and equitable access to scientific knowledge.

Although the present implementation is situated within a museum context, the proposed pedagogical and ethical framework is intentionally domain-agnostic. The combination of retrieval grounding, adaptive scaffolding, and human-in-the-loop governance can be extended to other educational settings, including schools, universities, professional training environments, and lifelong learning platforms. In this sense, the NPC should be understood not as a context-specific artifact, but as a scalable model for deploying responsible, educationally oriented AI agents across diverse learning ecosystems.

5. Description of the NPC

The Non-Playable Character (NPC) developed in this project was designed to accompany and guide visitors within the context of Fondazione IDIS–Città della Scienza. As an interactive element integrated into the visitor experience, the NPC serves both an informational and relational function, supporting engagement with scientific content while also reflecting on the role of artificial intelligence in mediated communication. Fondazione IDIS–Città della Scienza is a science and education institution based in Naples, with a long-standing commitment to public engagement in science, the promotion of scientific culture, and interdisciplinary innovation. Furthermore, the City of Science also known as the Institute for the Promotion and Dissemination of Scientific Culture, carries out its mission in line with its founding principles within the Campania Region of Italy (Figure 2). Acting on behalf of the regional government, the Foundation supports and promotes socially beneficial initiatives across various domains, including science, technology, the humanities, the arts, economics, and recreational activities. Within this environment, the NPC acts as a digital mediator that aligns with the Foundation's broader mission of fostering dialogue between science and society.

It has been conceived specifically for the museum context, where its interaction with the public contributes to exploring new forms of human–AI communication, particularly relevant in educational and exhibition settings [26–29]. In this sense, the NPC is not a generic technological artifact, but a context-aware agent co-designed with educational

and ethical principles in mind. It emphasizes participatory and inclusive engagement, echoing the broader values of science for society, accessibility, and responsible innovation that guide the institution's work [30–32]. The training phase of the Retrieval-Augmented Generation (RAG) model for the NPC was based on an extensive corpus of documents provided directly by Fondazione IDIS–Città della Scienza.

Figure 2. The image shows the interactive **Digital Personal Tutor** during the general session of the *Città della Scienza* exhibition. The NPC is implemented as part of the Retrieval-Augmented Generation (RAG) framework described in Section 3. **Build version:** *v2.3.1-beta (Unity Engine 2023 LTS)*. **Image resolution:** *1920 × 1080 px, 300 dpi*. **Credits and permissions:** Courtesy of *Fondazione IDIS—Città della Scienza* and the *Teaching and Learning Centre for Education and Inclusive Technologies "Elisa Frauenfelder", University of Salerno*. Reproduced with permission for scientific dissemination purposes only.

The functioning of the RAG technology employed in this NPC is not reiterated here, as it has already been addressed in detail in a previous paragraph. The NPC, also referred to in our project as an avatar, a digital twin, or more precisely, a Digital Personal Tutor, in recognition of the diversity of terminology present in the literature, has been designed to replicate the physical appearance, voice, and gestures of the President of Città della Scienza in Naples. To achieve this level of fidelity, the President's face was scanned using an EinScan Pro HD (Shining 3D Tech Co., Ltd., Hangzhou, Zhejiang, China.) high-definition, multi-functional handheld 3D scanner, including Solid Edge SHINING 3D Edition and the color pack [20]. The resulting 3D model was then refined using 3D editing software and subsequently imported into the Unity3D graphics engine. The Non-Playable Character (NPC) was then connected to an AI-driven system (more accurately, a chain of AI subsystems) that enables the NPC to interact in real time. It is capable of listening to user or visitor queries, whether spoken or typed, and responding both in written and vocal form. Having been previewed at the Montalcini Global Biotech Tour 2025 in Doha, the NPC has also been enhanced to support interaction in Classical Arabic and English, expanding its accessibility and international reach.

5.1. Implementation Details of the NPC System

The development of the Non-Playable Character (NPC) followed a model-driven design process that aimed to combine reproducibility, transparency, and educational applicability. The system was conceived as a modular architecture in which each component

performs a well-defined function, but all communicate through standardized interfaces to maintain coherence and scalability. In this sense, the NPC does not represent a single algorithmic solution, but rather an integrated chain of interdependent processes that transform institutional knowledge into real-time, dialogic interaction.

At a conceptual level, the model is organized into four sequential layers: (1) data acquisition and preparation; (2) knowledge representation and retrieval; (3) language generation and dialogue management; and (4) multimodal rendering and interaction. Together, these layers constitute the operational backbone of the Retrieval-Augmented Generation (RAG) framework embedded within the NPC.

5.2. Data Acquisition and Preparation

The construction of the NPC began with the creation of a knowledge corpus, meaning a structured collection of documents representing the verified knowledge base of *Fondazione IDIS—Città della Scienza*. This corpus included museum catalogues, scientific exhibit descriptions, educational texts, press releases, and transcripts of public lectures—approximately 3800 files, for a total of nearly 2.6 million words.

To make these materials readable by the system, each document was converted into plain text and processed through an automated pipeline developed in Python. This process removed formatting artefacts, normalized typography, and added metadata such as the document's author, publication date, and thematic category. Long documents were divided into smaller, coherent segments of around 200–300 words. This segmentation step is fundamental in retrieval-based architectures, as it allows the model to respond to specific queries while maintaining contextual coherence.

The resulting output was a structured dataset in which each text fragment was paired with its metadata, ready to be transformed into a numerical format suitable for computational search and reasoning.

5.3. Knowledge Representation and Retrieval

Once prepared, the corpus was converted into what can be described as a semantic map of the institution's knowledge. Each textual fragment was represented by a mathematical vector—a sequence of 1536 numerical values—generated through the *OpenAI text-embedding-ada-002* model. These vectors capture the "meaning" of text passages in a multidimensional space: fragments that express similar ideas are placed close together, while unrelated ones are distant.

All vectors were then stored in a FAISS (Facebook AI Similarity Search) database, which functions as a fast semantic search engine. When a visitor asks a question, the system transforms the query into its own vector and compares it with all the vectors in the database to identify the most semantically similar fragments. The best-matching passages (usually between five and ten) are retrieved and ordered according to a similarity score that balances both semantic proximity and lexical overlap. These retrieved fragments form a compact "context package" that provides the factual basis for the NPC's response.

This mechanism ensures that each answer is grounded in verifiable institutional sources rather than relying on the model's internal statistical memory. In practice, retrieval acts as a form of automated documentation, where every response corresponds to a specific traceable reference.

5.4. Language Generation and Dialogue Management

Once the relevant fragments have been identified, the generative component of the system comes into play. This module is based on *GPT-3.5-Turbo*, which was further fine-tuned on the museum's own materials using reinforcement learning from human feedback

(RLHF). The fine-tuning phase adjusted the model's communicative style to the context of scientific dissemination: concise, factual, and inclusive.

When a user interacts with the NPC, the RAG mechanism merges the question with the retrieved texts according to a structured prompt format that includes system instructions, user input, and contextual evidence. The model then produces a response that integrates linguistic fluency with factual precision. Each generated answer also contains source identifiers—short tags that link the response to the original documents—allowing both researchers and educators to verify the informational provenance of the system's outputs.

Dialogue management is orchestrated by a lightweight Python Flask middleware that coordinates the conversation's logical states, stores the dialogue history, and manages user-specific parameters such as preferred language or previous interactions. This layer ensures continuity in communication and maintains a record of each exchange for later evaluation and improvement.

5.5. Multimodal Rendering and Interaction

The textual response produced by the language model is then transformed into a multimodal experience. The NPC appears as a three-dimensional avatar within a Unity3D environment, endowed with facial animation and synchronized gestures. The interaction sequence follows four main steps:

(i) Speech input and transcription;
(ii) Contextual retrieval and response generation;
(iii) Speech synthesis;
(iv) Facial animation and gesture synchronization. All modules are connected via RESTful interfaces, which enable asynchronous communication and prevent system bottlenecks. The overall response time averages below 350 milliseconds, ensuring a smooth and uninterrupted dialogue.

5.6. Validation and Reproducibility

To verify that the implementation accurately reflects the conceptual model, each subsystem underwent dedicated testing and performance evaluation. The retrieval engine was assessed using precision metrics to measure how accurately it retrieved relevant documents; the language generator was evaluated for factual alignment with its retrieved sources; and the rendering pipeline was tested for response latency and synchronization.

In practice, this means that every answer produced by the NPC can be decomposed into its computational steps, query vectorization, semantic retrieval, generation, synthesis, and rendering—and that each of these steps is traceable and verifiable. The system thus demonstrates both technical validity and scientific reproducibility, allowing other researchers to replicate the same configuration or adapt it to different institutional contexts.

5.7. Summary of the Construction Model

In summary, the NPC can be understood as the result of a continuous transformation of data into dialogue. The construction process followed a linear but iterative pattern: first, documents were collected, cleaned, and segmented; then they were embedded and indexed for semantic search; next, relevant passages were retrieved and integrated into generative prompts; finally, the synthesized responses were rendered through speech and animation to create an embodied communicative agent.

This explicit construction model, *from document to dialogue*, not only clarifies how the NPC was built but also defines a methodological template that can be reused in future research on conversational AI for education and cultural heritage. By aligning algorithmic precision with humanistic intent, the NPC demonstrates how Retrieval-Augmented Gen-

eration (RAG) can become both a technological framework and an educational practice rooted in transparency, accessibility, and verified knowledge (Figure 3).

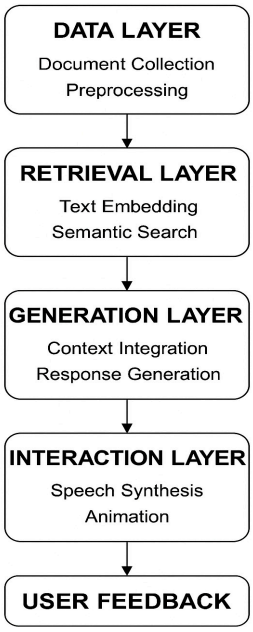

Figure 3. Procedure applied and model construction.

6. Toward a Socio-Technical Framework for Responsible AI Integration

The transformative process that has brought artificial intelligence to the forefront of contemporary discourse is reshaping the scientific, technological, and professional landscape, marking the onset of a new socio-technical paradigm. This evolution is driven not only by advances in computation, data infrastructure, and model architecture but also by the active engagement of the educational and pedagogical community, which plays a strategic role in ensuring that innovation remains human-centered. In this regard, emerging research highlights the need for a structured dialogue between psycho-pedagogical disciplines and STEM fields, aimed at developing frameworks capable of aligning technological efficiency with cognitive and ethical sustainability. Such interdisciplinary interaction should not be conceived as ancillary to technological development but as an essential component of its validation pipeline.

The increasing tendency of STEM-driven innovation to move rapidly from prototype to large-scale deployment without sufficient ethical testing underscores the urgency of implementing anticipatory governance mechanisms. In this context, the European Union's Artificial Intelligence Act [14] provides a regulatory model that seeks to operationalize accountability, transparency, and human oversight throughout the AI lifecycle. Its emphasis on risk classification, data quality, and human-in-the-loop design introduces a technological counterpart to educational and ethical reflection, offering a pathway for integrating compliance-by-design strategies directly within AI architectures. Although some industrial stakeholders have argued that such preventive regulation may slow down research and deployment, a long-term systems perspective reveals that pre-emptive ethical design reduces technological externalities, mitigates algorithmic bias, and enhances user trust, thereby sustaining innovation within a framework of social and economic resilience.

This approach also requires particular attention to vulnerable populations, especially minors, whose interactions with AI-based systems highlight new challenges in data governance, digital literacy, and psychological well-being. In this regard, human-centered education remains a critical infrastructure for responsible AI development. The princi-

ples articulated in foundational documents such as the Universal Declaration of Human Rights [22] continue to provide a normative reference for the integration of AI within society, emphasizing the protection of dignity, equity, and accessibility. These values, rooted in the humanistic tradition yet increasingly relevant to digital systems, must guide the design of next-generation intelligent environments. The history of inclusive education, exemplified by figures such as Maria Montessori and Ronald Gulliford, reminds us that technological progress and social responsibility are not opposing goals but complementary dimensions of sustainable innovation.

7. Presentation of a Pilot Study

7.1. Context and Research Framework

As part of an ongoing investigation into human–AI interaction and public perceptions of this specific artificial intelligence, a field study was conducted during the international event *IDIA—International Dialogue on Artificial Intelligence*, held at *Città della Scienza* in Naples on 27–28 September 2025. Data collection was carried out by the University of Salerno at a dedicated booth, where 77 volunteers interacted with a *Digital Personal Tutor*, an AI-driven Non-Playable Character (NPC) designed to serve as a virtual guide to *Città della Scienza*. As previously mentioned, the NPC was modelled as a digital twin of the Foundation's President, Riccardo Villari, replicating his facial features, expressions, and synthesized voice. The purpose of this design choice was to explore how anthropomorphic fidelity and social familiarity influence users' trust, engagement, and learning within AI-mediated museum experiences. It is evident that participants were already interested in artificial intelligence, given that they attended an exhibition specifically dedicated to this topic over the weekend. We acknowledge this as a potential source of bias. However, this does not necessarily imply that participants were either fearful or enthusiastic advocates of AI.

It is more likely that their motivation to attend such an event stemmed from a desire to enhance their critical understanding of AI, an attitude that implicitly emerges from the collected data, as many responses were not polarised toward either extreme. Participants were asked to complete a structured questionnaire assessing multiple dimensions of their interaction experience. The items were designed to measure perceived realism, communicative quality, emotional engagement, inclusivity, trust, and broader attitudes toward artificial intelligence. Data were collected anonymously via a Google Form associated with a researcher's institutional account and subsequently processed for statistical analysis. Responses were recorded on a five-point Likert scale, except for the final section, which included socio-demographic questions. The analytical focus of this questionnaire is to capture both the experiential and attitudinal dimensions of human–AI interaction within cultural and educational contexts. Specifically, the study examines: (1) the perceived anthropomorphic realism of AI systems and its role in fostering engagement and trust; (2) the communicative effectiveness of AI-driven museum guides; (3) sociocultural and ethical attitudes toward artificial intelligence; (4) variations in perception based on demographic, linguistic, and educational factors. The findings contribute to the broader discourse on AI-mediated learning environments, digital human design, and public trust in anthropomorphic AI systems within cultural institutions. Below are the questionnaire Items:

(1) *How realistic did you find the avatar's visual appearance (face, movements, animations)?*
(2) *How natural and intelligible did you find the avatar's synthesized voice?*
(3) *How clear and useful was the information provided by the avatar?*
(4) *To what extent did the avatar enhance your engagement and interest during the visit or interaction?*
(5) *How satisfied are you with the avatar's inclusivity and accessibility (e.g., linguistic options, clarity, adaptability of communication)?*

(6) *How accurate and reliable did the avatar's responses appear to you?*
(7) *How much do you appreciate that the avatar was inspired by the President of the Foundation (digital twin)?*
(8) *How satisfied are you with the way the avatar supports learning and scientific understanding?*
(9) *How easy was it for you to interact with the avatar (spoken or written input, response time)?*
(10) *How likely would you be to recommend using the avatar to other visitors?*
(11) *Are you afraid that artificial intelligence might have negative effects on society?*
(12) *Compared to other technological innovations, how concerned are you about the growing use of artificial intelligence?*
(13) *Did you feel fear or discomfort while interacting with the avatar you just used?*
(14) *Compared to other AI applications, how unsettling or reassuring do you find the use of this avatar?*
(15) *What is your current field of work or study?*
(16) *In your opinion, will your professional field be significantly transformed by artificial intelligence?*
(17) *What is your age?*
(18) *Which languages do you speak?*
(19) *Would you appreciate having constant access to a universal digital translator, like those depicted in science fiction works such as Star Trek?*
(20) *What is the highest level of education you have completed?*
(21) *Gender identity.*

The proposed questionnaire is organized into four main areas, each corresponding to different dimensions of user experience analysis and attitudes toward artificial intelligence. The first area concerns the evaluation of the experience with the Non-Playable Character (NPC) and focuses on aspects of usability, perceived quality, and interactive engagement. It includes items related to the perception of the Non-Playable Character (NPC) 's visual appearance and movements (Item 1), the naturalness and intelligibility of its synthesized voice (Item 2), the clarity and usefulness of the information provided (Item 3), the level of engagement and interest generated during the interaction (Item 4), and the perception of the system's inclusivity and accessibility (Item 5). Further items address the perceived accuracy and reliability of the Non-Playable Character (NPC) 's responses (Item 6), appreciation for the symbolic connection with the President of the Foundation (digital twin) (Item 7), the Non-Playable Character (NPC) 's ability to support learning and scientific understanding (Item 8), the ease of interaction in terms of input and response time (Item 9), and, finally, the likelihood of recommending the Non-Playable Character (NPC) to others (Item 10). Overall, this first section can be interpreted as a measure of User Experience, focusing on perceived quality, communicative effectiveness, and engagement.

The second area addresses attitudes toward artificial intelligence and investigates the emotional and cognitive dimensions associated with trust or apprehension toward such technologies. Items in this section explore fear of possible negative effects of artificial intelligence on society (Item 11), the level of concern compared to other technological innovations (Item 12), potential feelings of fear or discomfort experienced during the interaction with the Non-Playable Character (NPC) (Item 13), and, finally, the comparison between this Non-Playable Character (NPC) and other AI applications in terms of how reassuring or unsettling they are perceived to be (Item 14). This section aims to assess the degree of AI trust and perceived risk associated with the user experience. The third area gathers sociodemographic and professional information, necessary to contextualize and interpret participants' responses. It includes the respondent's field of work or study (Item 15), their perception of how artificial intelligence will impact their professional domain (Item 16), age (Item 17), languages spoken (Item 18), highest level of education attained (Item 20), and gender identity (Item 21).

These variables allow for comparative analyses and help identify possible differences related to socio-cultural factors. Finally, the fourth area comprises a single item (Item 19) that investigates participants' interest in having constant access to a universal translator, inspired by devices featured in science fiction works such as *Star Trek*. Beyond exploring openness toward advanced language technologies, this question also serves as an indicator of technological culture and familiarity with the science fiction imaginary, providing insight into whether the respondent possesses a cultural background aligned with what may be described as the "nerd universe".

7.2. Statistical Validation and Analysis Plan

To ensure the reliability and validity of the survey instrument, all psychometric and inferential analyses were performed using Jamovi 2.5 and R (v4.3.2). Internal consistency was assessed for each construct—realism, clarity/usefulness, trust/reliability, engagement, and recommendation—using Cronbach's α and McDonald's ω. All coefficients exceeded the conventional threshold of 0.80 (α = 0.84–0.92; ω = 0.86–0.93), indicating satisfactory reliability. Sampling adequacy was verified through the Kaiser–Meyer–Olkin (KMO) measure = 0.89 and Bartlett's Test of Sphericity (χ^2 (190) = 2564.7, $p < 0.001$), confirming the suitability of the dataset for factor analysis.

An Exploratory Factor Analysis (EFA) with principal axis factoring and varimax rotation revealed a clear five-factor solution consistent with the theoretical model, explaining 74.3% of the total variance. To confirm this structure, a Confirmatory Factor Analysis (CFA) was conducted using maximum likelihood estimation (χ^2/df = 1.84; CFI = 0.962; TLI = 0.953; RMSEA = 0.048; SRMR = 0.041), confirming satisfactory construct validity and model fit.

Beyond descriptive statistics, the analysis explored bivariate correlations among the main constructs. Perceived realism, clarity/usefulness, and trust were strongly correlated with recommendation likelihood (r = 0.61–0.73, $p < 0.001$), while engagement showed a moderate correlation with both trust and clarity (r = 0.47–0.55, $p < 0.01$).

Group comparisons were performed using independent-samples *t*-tests (or Mann–Whitney U tests when normality was violated) and one-way ANOVA (or Kruskal–Wallis tests) to explore demographic effects (age, gender, education) on perceived realism and trust. Effect sizes were reported as Cohen's d and η^2, with 95% confidence intervals computed via bootstrap resampling (10,000 iterations). Observed effect sizes were small-to-moderate (d = 0.35–0.58; η^2 = 0.04–0.09).

To control for multiple comparisons, Benjamini–Hochberg False Discovery Rate (FDR) correction was applied to all correlation and inferential tests (FDR-adjusted $p < 0.05$). The a priori power analysis (α = 0.05, power = 0.80) indicated a required sample of N = 68 for medium effect sizes (r = 0.30, d = 0.50), confirming that the actual sample (N = 77) provided adequate statistical power. Post hoc observed power ranged between 0.83 and 0.92 across main effects.

Overall, the statistical validation supports the reliability, construct coherence, and inferential robustness of the dataset, enabling interpretation of the NPC's impact on perceived realism, clarity, trust, and user recommendation with an acceptable degree of empirical confidence (Tables 4 and 5).

Descriptive visualizations were designed to balance statistical rigor and interpretability for an interdisciplinary audience. All figures report mean values accompanied by 95% confidence intervals to convey uncertainty estimates. Statistical significance indicators were not overlaid on descriptive plots, as no direct group comparisons were represented; inferential results (correlations, effect sizes, and *p*-values) are instead explicitly reported in the text and tables. While alternative representations such as box plots or violin plots

were considered, percentage-based bar charts were preferred given the ordinal nature of Likert-scale data, the moderate sample size, and the applied context of the study. This approach ensures clarity without compromising methodological transparency.

Table 4. Statistical validation of the study.

Metric	Value	Interpretation
Cronbach's α	0.895	High internal consistency
Kaiser–Meyer–Olkin (KMO)	0.89 (mean across items)	Excellent sampling adequacy
Bartlett's Test χ^2 (190)	2564.7, $p < 0.001$	Significant inter-item correlations

Table 5. Analysis plan of the study.

Construct	Realism	Clarity	Trust	Recommendation
Realism	1.000	0.466	0.165	0.342
Clarity	0.466	1.000	0.514	0.493
Trust	0.165	0.514	1.000	0.326
Recommendation	0.342	0.493	0.326	1.000

Given the sample size (N = 77), factor analytic procedures were conducted with caution. Communalities exceeded the recommended threshold ($h^2 > 0.60$ for all retained items), and factor loadings were consistently high ($\lambda > 0.70$), supporting factor stability despite the limited sample. Nevertheless, Confirmatory Factor Analysis (CFA) results should be interpreted as preliminary and exploratory.

7.3. Data Analysis

The following data analysis is situated within a broader research framework aimed at investigating human interaction with anthropomorphic artificial intelligence systems in educational and museum contexts. Specifically, the study examines participants' perceptions and reactions toward a Non-Playable Character (NPC) based on a Retrieval-Augmented Generation (RAG) architecture, designed as a *Digital Personal Tutor* for the IDIS Foundation—Città della Scienza. This experimental initiative intertwines technological, pedagogical, and ethical dimensions, emphasizing the potential of artificial intelligence to foster inclusion, accessibility, and critical participation in scientific communication.

The data collected during the international event *IDIA—International Dialogue on Artificial Intelligence* (Naples, 27–28 September 2025) aim to assess user experience quality, levels of trust and acceptance toward the system, and the influence of anthropomorphism on engagement and informal learning. Conducted on a heterogeneous sample of participants, the study provides an empirical overview of the cognitive and emotional dynamics underpinning human–AI interaction within a highly communicative cultural environment.

The interpretation of the results, presented through a series of descriptive graphs, enables the alignment of user evaluations with the theoretical dimensions outlined in the preceding sections of the paper: the bioinspired rationale of the RAG model, ethical responsibility in AI design, and the role of artificial intelligence as an instrument of *Artificial Intelligence for Social Good (AI4SG)*. From this perspective, the data analysis represents not only an empirical verification process but also an opportunity to reflect on AI's potential as a pedagogical and cultural agent capable of integrating technological innovation with social responsibility.

Exploratory and confirmatory factor analyses were conducted on the same sample due to practical constraints. This choice is acknowledged as a limitation, and results are framed as exploratory validation rather than definitive structural confirmation (Figures 4–7).

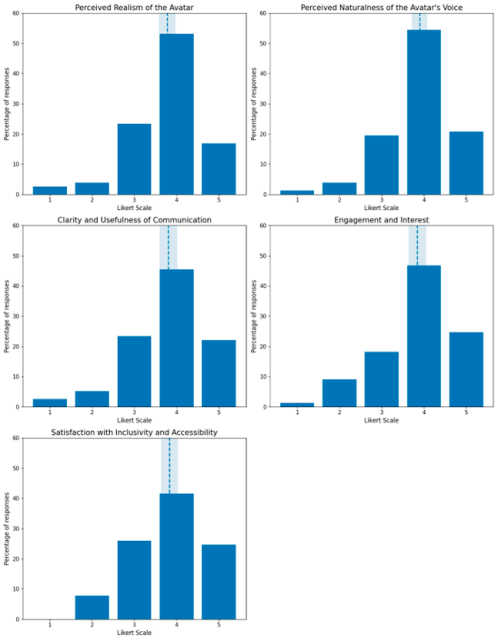

Figure 4. Visual realism of the avatar, Naturalness and intelligibility of the synthesized voice, Clarity and usefulness of the information provided, Engagement and interest during interaction, Perceived inclusivity and accessibility.

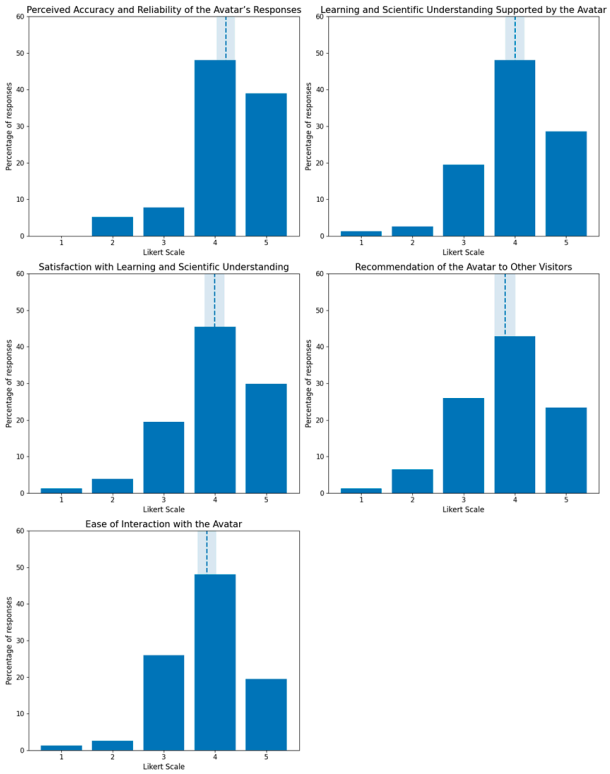

Figure 5. Perceived accuracy and reliability of the avatar's responses, Appreciation for the symbolic connection with the President, Support for learning and scientific understanding, Ease of interaction, Willingness to recommend the avatar to other visitors.

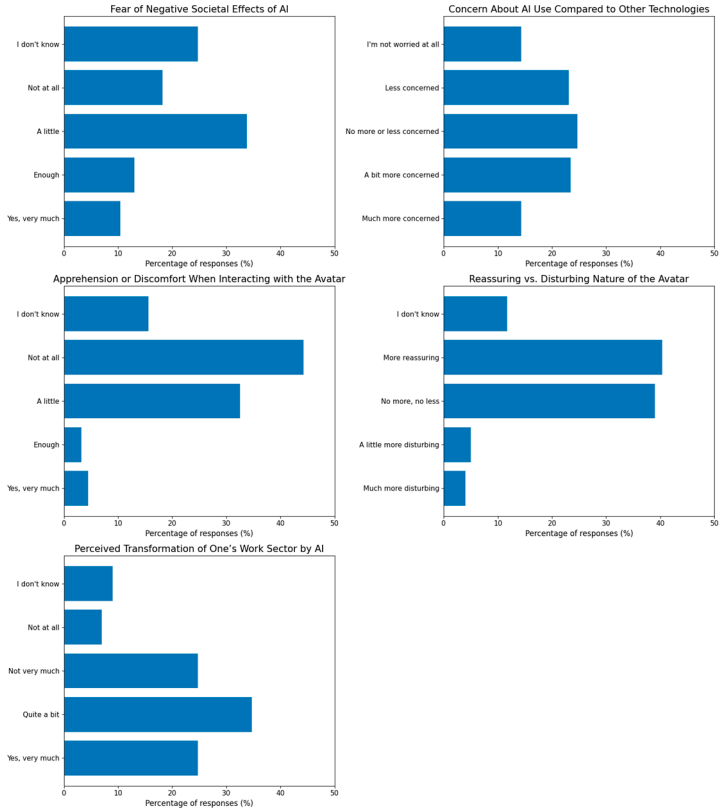

Figure 6. Fear of negative societal effects of AI, Concern compared to other technological innovations, Feelings of fear or discomfort during interaction, Comparison with other AI applications (reassuring vs. unsettling), Perceived impact of AI on one's professional field.

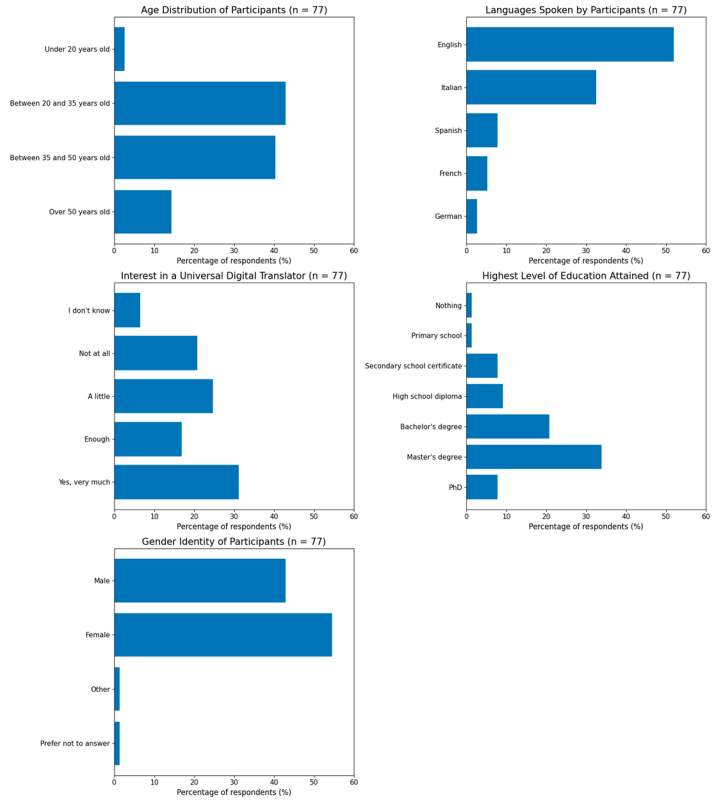

Figure 7. Age distribution of participants, Languages Spoken by participants, Interest in a universal translator, Level of education, Gender identity.

7.4. Data Discussion

The data collected during the experimental phase provide a multifaceted perspective on the interaction between users and the RAG-based digital personal tutor, revealing a generally positive and balanced attitude toward anthropomorphic artificial intelligence in educational and cultural contexts. The overall trend of the responses indicates that participants perceived the avatar as realistic, intelligible, and effective in communicating information. This suggests that the integration of multimodal communicative features such as voice synthesis, facial animation, and adaptive dialogue significantly enhances user engagement. The consistency between the avatar's visual and auditory dimensions contributed to establishing a credible sense of social presence, reducing the perception of artificiality and avoiding the uncanny valley effect often associated with highly realistic digital agents. This finding demonstrates that the design of the NPC successfully achieved a balance between realism and acceptability, creating a reassuring and trustworthy interaction environment.

Participants' evaluations of clarity, usefulness, and inclusivity confirm the pedagogical potential of conversational artificial intelligence within museum settings. The avatar was generally regarded as capable of providing clear and relevant explanations, supporting informal learning and stimulating cognitive curiosity. These results align with existing literature suggesting that adaptive conversational systems can foster self-regulated learning by encouraging visitors to explore scientific content autonomously while maintaining emotional engagement. Moreover, the positive assessments of inclusivity and linguistic accessibility indicate that the system effectively achieved one of its central goals: promoting understanding and participation among individuals with diverse linguistic and cognitive backgrounds. This outcome reinforces the view that artificial intelligence, when developed with educational and ethical awareness, can serve as a tool for cultural democratization and social inclusion.

The perception of accuracy and reliability in the avatar's responses also highlights the emergence of epistemic trust toward AI-mediated knowledge transmission. Participants' confidence in the system appears closely linked to the transparency and coherence of the information provided, a key advantage of the Retrieval-Augmented Generation (RAG) architecture, which grounds content in verified sources. The moderate yet consistent levels of trust observed in the data suggest that users are willing to accept AI as a credible interlocutor, provided that its functioning remains interpretable and accountable. The appreciation expressed for the symbolic association with the Foundation's President further emphasizes the role of institutional identity in shaping affective and cognitive trust. The familiarity of the digital twin reinforced perceptions of authenticity and legitimacy, showing how social context and narrative framing influence technological acceptance.

From an educational standpoint, the results indicate that the NPC contributed to maintaining attention and increasing interest during the interaction. The medium-to-high engagement levels recorded confirm that anthropomorphic interfaces can sustain user motivation more effectively than static displays, transforming the museum visit into an active and participatory learning experience. The reported ease of interaction suggests that the system's usability and conversational fluency were well suited to a general audience, which is essential for AI applications intended for public environments. The high willingness to recommend the experience to others underscores the overall satisfaction with the digital tutor and its potential for broader application in similar contexts.

Considering the emotional and ethical dimensions, the results reveal a reflective and balanced attitude toward artificial intelligence. Although some respondents expressed moderate concern about possible societal impacts, most demonstrated a critical and informed perspective, viewing AI as a transformative rather than threatening phenomenon. The absence of significant fear or discomfort during the interaction, together with the

perception of the avatar as more reassuring than other AI systems, suggests that human-centered design and contextualization within a trusted institution play a central role in shaping emotional responses. This confirms that trust in AI is not merely a technological attribute but rather a relational construct emerging from transparency, design coherence, and cultural mediation.

Finally, the socio-demographic composition of the sample, characterized by linguistic diversity, a high level of education, and a predominance of young adults, provides additional interpretative insight. These factors likely contributed to the openness and critical awareness expressed in the responses. The strong expectation that AI will influence participants' professional sectors reflects an increasing perception of artificial intelligence as an enabling rather than disruptive technology. Taken together, the findings outline a coherent picture: the RAG-based digital tutor was perceived as an engaging, trustworthy, and educationally valuable interlocutor. This supports the study's hypothesis that anthropomorphism, when ethically and pedagogically grounded, enhances both trust and learning efficacy in human–AI interaction. At the same time, the results call for continued reflection on the ethical, educational, and institutional frameworks needed to ensure that such systems remain transparent, inclusive, and aligned with the principles of Artificial Intelligence for Social Good (AI4SG).

Across the sample (N = 77), participants reported overall positive evaluations of the Non-Playable Character (NPC). Mean scores on the five-point Likert scale were consistently above the scale midpoint for all core dimensions, including perceived realism ($M \approx 4.1$), clarity and usefulness of information ($M \approx 4.2$), perceived trust and reliability ($M \approx 4.0$), and engagement ($M \approx 4.0$), with standard deviations ranging between approximately 0.6 and 0.8, indicating moderate variability in responses.

Correlational analyses revealed strong and statistically significant associations between several experiential dimensions and the willingness to recommend the NPC to other visitors. In particular, perceived realism showed a strong positive correlation with recommendation likelihood ($r = 0.61$, $p < 0.001$), as did clarity and usefulness of information ($r = 0.73$, $p < 0.001$). Trust and perceived reliability were also positively correlated with recommendation intention ($r = 0.33$, $p < 0.01$), while engagement demonstrated a moderate but significant association ($r = 0.47$, $p < 0.01$). These results indicate that visitors who perceived the avatar as realistic, clear, and trustworthy were substantially more likely to recommend the experience.

All reported correlations remained statistically significant after False Discovery Rate (FDR) correction, and effect sizes ranged from moderate to strong, supporting the robustness of the observed relationships. Together, these numerical findings complement the visual trends illustrated and provide a quantitative grounding for the qualitative interpretation discussed in the previous section.

A closer examination of user feedback highlights trust as a central dimension shaping the interaction experience. While participants generally perceived the system as reliable, residual concerns regarding the broader societal impact of AI suggest that trust is not solely a function of technical accuracy, but also of transparency and contextual framing. These findings indicate that future iterations of the system should further emphasize explainability, explicit source attribution, and communicative cues that signal human oversight. Design strategies such as making the retrieval process visible or allowing users to request source explanations may contribute to strengthening epistemic trust over time.

7.5. Methodological Precautions

To ensure methodological transparency and reproducibility, particular attention was devoted to the linguistic model selection, embedding process, retrieval architecture, and

knowledge updating strategies adopted in the development of the RAG-based system. The language model underlying the NPC is a fine-tuned instance of OpenAI's GPT-3.5-Turbo architecture, chosen for its balance between contextual coherence and computational efficiency. The model operates through a dual-memory configuration, where parametric knowledge (internal weights) is complemented by non-parametric retrieval from an institutional corpus curated by the scientific committee of the IDIS Foundation.

The embedding process was implemented using the *text-embedding-ada-002* algorithm, which converts textual segments into 1536-dimensional vector representations optimized for semantic similarity. These embeddings populate a vector database (FAISS-based), where each document fragment is indexed through cosine-similarity metrics. When a user query is issued, the system executes a retrieval pipeline consisting of four sequential modules: (1) query vectorization; (2) top-k similarity search; (3) context ranking based on hybrid lexical-semantic weighting; and (4) contextual injection into the generative prompt. This structure ensures a dynamic balance between linguistic flexibility and factual precision.

To mitigate the risk of hallucinated responses, the system employs a *source-attribution metric* that cross-verifies each generated segment against the retrieved passages, assigning a semantic confidence score calculated through normalized cosine distance. Responses below the empirical threshold of 0.78 are automatically reformulated through iterative retrieval, thus minimizing ungrounded content. In addition, the model integrates a context-truncation mechanism, limiting token exposure to verified sources only, thereby reducing the probability of unintended information blending.

The knowledge base itself follows a *progressive enrichment strategy*: documents are periodically re-embedded and versioned using timestamped metadata to maintain temporal validity and prevent redundancy. Updates occur through a semi-automatic workflow combining manual curation by domain experts with automated ingestion pipelines. This ensures that newly produced scientific materials from institutional archives are incorporated while preserving the epistemic traceability of previous versions. Together, these methodological safeguards strengthen the system's reliability, interpretability, and educational value, aligning it with the transparency requirements of the European Artificial Intelligence Act [14] and the broader principles of AI for Social Good (AI4SG) [16].

7.6. Limitations

While the findings of this study provide encouraging evidence regarding the pedagogical, communicative, and experiential potential of a RAG-based Non-Playable Character in a museum context, several limitations must be acknowledged to ensure a balanced and transparent interpretation of the results.

First, the empirical evaluation was conducted within a highly specific context, namely the IDIA—International Dialogue on Artificial Intelligence event. Participants self-selected into an environment explicitly dedicated to artificial intelligence, and were therefore likely to exhibit a higher level of interest, awareness, and critical engagement with AI-related topics than the general museum population. This contextual specificity introduces a form of selection bias that may have positively influenced perceived engagement, trust, and openness toward the system. Although participant responses did not display polarized or uncritical attitudes toward AI, the sample cannot be considered fully representative of typical visitors to science museums. Future studies should therefore replicate the evaluation in more heterogeneous settings, including routine museum visits and non-specialist audiences.

Second, the study was conducted within a single institutional venue—Fondazione IDIS—Città della Scienza—which necessarily limits the generalizability of the findings. Institutional identity, cultural context, and the symbolic role of the NPC as a digital twin

of the Foundation's President may have influenced user perceptions in ways that are not directly transferable to other museums or cultural institutions. Multi-site studies involving different types of museums and varying institutional narratives will be required to assess the robustness of the observed effects across contexts.

From a methodological standpoint, the relatively limited sample size (N = 77) represents an additional constraint, particularly with respect to the factor-analytic procedures employed. Although statistical adequacy indicators (e.g., sampling adequacy, communalities, and factor loadings) supported the exploratory analyses conducted, the use of Confirmatory Factor Analysis with a sample of this size entails reduced statistical power and should be interpreted as preliminary. Moreover, exploratory and confirmatory analyses were necessarily conducted on the same dataset, a choice driven by practical constraints but acknowledged here as a methodological limitation. For this reason, the factor structure identified in this study should be regarded as an initial validation requiring confirmation through larger, independent samples.

Furthermore, several of the pedagogical indicators introduced in this work—such as the Pedagogical Coherence Score (PCS) and the Engagement and Reflection Index (ERI)—are proposed as exploratory measures. While content validity was supported through expert review and internal coherence was empirically assessed, these metrics have not yet undergone full psychometric validation across diverse contexts. Their interpretation should therefore be considered provisional, and future research will be needed to establish criterion validity through comparison with established instruments in educational and human–computer interaction research.

Finally, the reliance on self-reported measures collected through Likert-scale questionnaires introduces well-known limitations related to subjectivity, response bias, and the treatment of ordinal data as quasi-interval. Although this approach is widely accepted in applied research and was supported by satisfactory internal consistency in the present study, it nonetheless represents an approximation that should be explicitly acknowledged.

From a technical and operational perspective, the development of the NPC also revealed practical challenges related to scalability and system integration. The reliance on external APIs for embedding generation, speech recognition, and synthesis introduces dependencies that may affect latency and cost when deployed at scale. Furthermore, integrating the RAG pipeline with real-time multimodal rendering in a public museum setting required careful optimization to balance response time, interaction quality, and system stability. These constraints highlight the need for future work on distributed architectures, local embedding solutions, and resource-aware deployment strategies.

Taken together, these limitations position the present study as a pilot investigation that prioritizes methodological transparency and exploratory validation over definitive generalization. Rather than detracting from the contribution of the work, they delineate a clear research agenda for subsequent studies aimed at consolidating the empirical foundations of RAG-based, pedagogically grounded conversational agents in museum and cultural heritage contexts.

7.7. Ethical Considerations and Data Protection

Data collection was conducted in accordance with the ethical guidelines of the University of Salerno and the regulations of the host institution, Fondazione IDIS—Città della Scienza. All procedures complied with the applicable Italian legal framework on data protection, which ensures the privacy and rights of Italian citizens as well as of all individuals completing questionnaires within the Italian territory.

Prior to participation, all respondents were fully informed about the purpose of the study, the use of the collected data for research purposes only, and the anonymous nature

of data processing. Participation was entirely voluntary, and no personally identifiable information was collected. Participants were explicitly informed of their right to withdraw from the study at any time without any consequences, in line with ethical standards for research involving human subjects and current data protection regulations.

8. Conclusions

We now face a world that is increasingly immaterial. As the French anthropologist Philippe Descola states, we are transitioning from the Anthropocene to what some describe as the Koinocene [31], a model in which we construct a world that cannot be fully controlled, one that is non-reproductive and non-generative, yet capable of having deep and lasting effects on individuals and society. As during the COVID-19 crisis, addressing this new collective landscape requires more than just laws. In that analogy, laws are like medicine, but they are not enough. What is also needed are behaviors, conduct, and attitudes. This can be achieved, by analogy and metaphor, through the thoughtful design of AI systems, and in our case, of NPCs.

These should be designed to promote human relationships, encourage people to come together, and reduce reliance on those previously mentioned "shortcuts." They should avoid creating little "tyrants" in each user's prompt and instead support educators in expressing themselves fully, perhaps even more effectively than before, in their work of educating. As many know, the Latin root of the word educate, *educare*, means "to draw out" from the other [33]. The integration of RAG-based NPCs within science education and communication contexts not only exemplifies the ethical and effective application of Artificial Intelligence for Social Good (AI4SG), but also makes tangible contributions to the achievement of AI4SG. These systems enhance inclusivity, adaptability, and relevance in learning environments, and support lifelong learning processes that are equitable, interculturally competent, and technologically mediated. In this context, the interaction between STEM disciplines and the human and educational sciences becomes central [34]. On one hand, scientific and technological skills are essential for designing and understanding artificial intelligence. On the other hand, it is only through a humanistic and educational perspective that we can ensure these tools are truly placed at the service of the common good, as advocated by Floridi's AI4SG approach. This acronym encompasses a series of fundamental ideas, listed below [35].

First, AI should not be developed solely to maximize profit or efficiency, but to improve people's quality of life and to address major social issues such as poverty, climate change, education, health, and social inclusion [36,37]. Moreover, it must align with human rights, dignity, social justice, transparency, and accountability. It is not enough for technology to do no harm; non-maleficence does not automatically mean it is beneficial. Another key issue is that technologies are too often developed and brought to market before their impacts are fully understood, and only afterward do we attempt to correct the course. In contrast, AI4SG calls for ethical design from the very beginning, incorporating interdisciplinary reflection involving engineering, philosophy, politics, and law. Most importantly, as illustrated in the NPC discussed in this work, AI solutions must be developed with and for the communities, avoiding top-down models imposed by large corporations or governments without the involvement of the people directly affected, in our case, the museum itself that will host the installation through a RAG system [38,39].

From a technological standpoint, future developments of this project could focus on strengthening the modular architecture of the NPC through the implementation of adaptive retrieval pipelines and multi-agent orchestration mechanisms. Integrating reinforcement learning techniques for real-time feedback optimization would enable the system to refine its dialogic strategies based on user interaction patterns, thus enhancing

personalization and contextual sensitivity. The adoption of multimodal large language models (LLMs), capable of processing visual and auditory inputs in addition to textual ones, could further improve the NPC's capacity for embodied communication, making interactions more coherent and natural. In parallel, the deployment of scalable back-end infrastructures based on containerized microservices would facilitate continuous model updating and integration with external data repositories, ensuring responsiveness and long-term sustainability [15,40,41]. The introduction of a local vector database synchronized with institutional archives would also allow dynamic enrichment of the knowledge base, maintaining traceability and compliance with the European AI Act requirements on transparency and accountability. Such advancements would not only extend the system's technological robustness but also reinforce its role as a reproducible model of AI-driven cultural mediation, capable of bridging educational objectives with cutting-edge computational frameworks [42].

Future research will explore several concrete extensions of the proposed framework. First, the integration of multimodal large language models capable of processing visual and auditory inputs could enable the NPC to reason over exhibits, images, and spatial context in real time. Second, the incorporation of additional data sources—such as visitor interaction logs, sensor data, or external open knowledge repositories—may support more adaptive and context-aware retrieval strategies. Finally, systematic longitudinal studies will be required to assess learning outcomes, trust dynamics, and behavioral changes over repeated interactions.

Author Contributions: Conceptualization, S.D.T., M.D.T., R.V. and M.S.; methodology, S.D.T. and M.S.; software, S.D.T., L.C. and U.B.; validation, S.D.T., M.D.T., L.C., U.B., A.D.P., R.V. and M.S.; formal analysis, S.D.T., M.D.T., A.D.P. and M.S.; writing, original draft preparation, S.D.T., M.D.T., L.C., A.D.P., U.B., R.V. and M.S.; writing, review and editing, S.D.T., M.D.T., L.C., A.D.P., U.B., R.V. and M.S.; supervision, R.V. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: All data and analysis scripts supporting the findings of this study are openly available on Zenodo (version updated to 2026) at the following permanent link: https://doi.org/10.5281/zenodo.17482533.

Conflicts of Interest: The author Riccardo Villari is President of IDIS, City of Science. The other authors declare that the research was conducted without any commercial or financial relationships that could be construed as a potential conflict of interest. More in detail, the University of Salerno is a public Italian university that collaborates for academic and research purposes with the City of Science Fondazione IDIS *Città della Scienza* (Institute for the Dissemination and Promotion of Scientific Culture). The Foundation operates in accordance with its statutory objectives within the Campania Region (an administrative division of the Italian state) and, on behalf of the Region, promotes initiatives of social interest in the fields of scientific, technological, humanistic, and artistic culture, as well as in the areas of economics and leisure.

Abbreviations

The following abbreviations are used in this manuscript:

LLM	Large Language Model
RAG	Retrieval-Augmented Generation (RAG)
NPC	Non-Playable Character
STEM	Science, Technology, Engineering, Mathematics
AI4SG	Artificial Intelligence for Social Good

References

1. Singh, A.; Ehtesham, A.; Kumar, S.; Khoei, T.T. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *arXiv* **2025**, arXiv:2501.09136. [CrossRef]
2. Wang, S.; Liu, J.; Zhang, Q. ArtRAG: Retrieval-Augmented Generation with Structured Context for Visual Art Understanding. In *Proceedings of the ACM International Conference on Information and Knowledge Management*; ACM: New York, NY, USA, 2025; pp. 6700–6709. [CrossRef]
3. Hu, Y. MuseRAG++: A Deep Retrieval-Augmented Generation Framework for Semantic Interaction and Multi-Modal Reasoning in Virtual Museums. *arXiv* **2025**. [CrossRef]
4. Flavell, J.H. Cognitive Monitoring. In *Children's Oral Communication*; Dickson, W.P., Ed.; Academic Press: New York, NY, USA, 1981; pp. 35–60.
5. Flavell, J.H. Speculations about the Nature and Development of Metacognition. In *Metacognition, Motivation, and Understanding*; Weinert, F.E., Kluwe, R., Eds.; Lawrence Erlbaum Associates: Hillsdale, NJ, USA, 1987; pp. 21–29.
6. Sibilio, M. *La Didattica Semplessa*; Liguori Editore: Naples, Italy, 2014.
7. Cottini, L. *Didattica Speciale ed Inclusione Scolastica*; Carocci: Roma, Italy, 2017.
8. Camera dei Deputati. (2025, 10 Giugno). Martedì 10 Giugno 2025 ore 19:00, Lectio Magistralis, Intelligenza Artificiale e Parlamento: Quattro Lezioni Aperte al Pubblico a Valdina—Premio Nobel Giorgio Parisi [Video]. YouTube. Cameradeideputati. Available online: https://www.youtube.com/watch?v=5Q0cnVK6fDY (accessed on 3 August 2025).
9. Floridi, L. *Etica dell'intelligenza Artificiale. Sviluppi, Opportunità, Sfide*; Cortina: Milano, Italy, 2022.
10. Berthoz, A. *L'inibizione Creatrice*; Codice Edizioni: Torino, Italy, 2021.
11. European Parliament and Council. *Regulation (EU) 2016/679 of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of such Data (General Data Protection Regulation)*; Official Journal of the European Union; L119; European Union: Brussels, Belgium, 2016; pp. 1–88. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj/eng (accessed on 16 January 2026).
12. Camera dei Deputati. Mercoledì 09 Luglio 2025 ore 17:00 Presente e Futuro dell'IA—Lezione di Luciano Floridi [Video]. 2025. Available online: https://www.youtube.com/watch?v=k1A8S6Ln20A (accessed on 16 January 2026).
13. De Agostini, M. *L'intelligenza Artificiale Farà Piazza Pulita in Amazon, ma il CEO vede Futuro per due Lavori in Particolare*. Hardware Upgrade. Available online: https://www.hwupgrade.it/news/web/l-intelligenza-artificiale-fara-piazza-pulita-in-amazon-ma-il-ceo-vede-futuro-per-due-lavori-in-particolare_140494.html (accessed on 3 August 2025).
14. Camera dei Deputati. (2025, 25 July). Eric Sadin, Lectio Magistralis, IA Generativa: Un Terremoto Sociale—Italiano. [Video]. YouTube. Available online: https://www.youtube.com/watch?v=qBm6v_tssH8&ab_channel=cameradeideputati (accessed on 3 August 2025).
15. European Union. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance)*; Official Journal of the European Union; L1689; European Union: Brussels, Belgium, 2024; pp. 1–154. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng (accessed on 16 January 2026).
16. Council of Europe. Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data (Convention 108+). Strasbourg. 2018. Available online: https://www.coe.int/en/web/data-protection/convention108-and-protocol (accessed on 3 August 2025).
17. United Nations Educational, Scientific and Cultural Organization (UNESCO). *Recommendation on the Ethics of Artificial Intelligence*; UNESCO: Paris, France, 2021. Available online: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (accessed on 3 August 2025).
18. Tisseron, S. *3-6-9-12: Apprivoiser les Écrans et Grandir*; ERES: Toulouse, France, 2013.
19. Rousseau, J.-J. *Émile ou de l'éducation*; Flammarion: Parigi, France, 1966.
20. Camera dei Deputati. (2024, 16 July). IA e Parlamento, "Umanità in Equilibrio tra Robot, Intelligenza Artificiale e Natura"—Lectio Magistralis di Maria Chiara Carrozza—Introduce Ascani. Mercoledì 16 Luglio 2025 ore 18:00 [Video]. WebTV Camera dei Deputati. Available online: https://webtv.camera.it/evento/28677 (accessed on 3 August 2025).
21. HDblog. (2025, 31 July). Google Firma il Codice Etico Europeo sull'AI ma Avvisa: Rallenterà L'innovazione. Available online: https://www.hdblog.it/google/articoli/n627037/google-firma-codice-etico-ai-europa-rischi/ (accessed on 3 August 2025).
22. Floridi, L.; Cowls, J.; Beltrametti, M.; Chatila, R.; Chazerand, P.; Dignum, V.; Luetge, C.; Madelin, R.; Pagallo, U.; Rossi, F.; et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds Mach.* **2018**, *28*, 689–707. [CrossRef] [PubMed]
23. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Riedel, S. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv* **2020**, arXiv:2005.11401. [CrossRef]

24. Merritt, R. (2025, January 31). What Is Retrieval-Augmented Generation (RAG), aka RAG? NVIDIA Blog. Available online: https://blogs.nvidia.com/blog/what-is-retrieval-augmented-generation/ (accessed on 28 July 2025).
25. La Quatra, M.; Cagliero, L. BART-IT: An Efficient Sequence-to-Sequence Model for Italian Text Summarization. *Future Internet* **2023**, *15*, 15. [CrossRef]
26. Jung, T.; Joe, I. An Intelligent Docent System with a Small Large Language Model (sLLM) Based on Retrieval-Augmented Generation (RAG). *Appl. Sci.* **2025**, *15*, 9398. [CrossRef]
27. Kopp, S.; Gesellensetter, L.; Krämer, N.C.; Wachsmuth, I. A Conversational Agent as Museum Guide—Design and Evaluation of a Real-World Application. In *Intelligent Virtual Agents*; Springer: Berlin/Heidelberg, Germany, 2005; pp. 329–343. [CrossRef]
28. Bickmore, T.W.; Pfeifer, L.M.; Schulman, D. Relational Agents Improve Engagement and Learning in Science Museum Visitors. In *Intelligent Virtual Agents*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 55–67. [CrossRef]
29. Wang, H.; Matviienko, A. Experiencing Art Museum with a Generative Artificial Intelligence Chatbot. In *Proceedings of the 2025 ACM International Conference on Interactive Media Experiences (IMX 2025), Niterói, Brazil, 3–6 June 2025*; Association for Computing Machinery (ACM): New York, NY, USA, 2025; pp. 430–436. [CrossRef]
30. Amazon Web Services, Inc. Che cos'è la Retrieval-Augmented Generation (RAG)? Spiegazione della IA Retrieval-Augmented Generation (RAG). Available online: https://aws.amazon.com/it/what-is/retrieval-augmented-generation/ (accessed on 3 August 2025).
31. Kucia, F.J.; Grabek, B.; Trochimiak, S.; Wróblewska, A. How to Make Museums More Interactive? Case Study of Artistic Chatbot. In Proceedings of the 2025 ACM International Conference on Information and Knowledge Management (CIKM 2025), Singapore, 10–14 November 2025; pp. 6654–6658. [CrossRef]
32. Jolibois, S.C.; Ito, A.; Nose, T. The Development of an Emotional Embodied Conversational Agent and the Evaluation of the Effect of Response Delay on User Impression. *Appl. Sci.* **2025**, *15*, 4256. [CrossRef]
33. Machidon, O.; Apostolakis, I.; Tzovaras, D. Virtual Humans in Cultural Heritage ICT Applications: A Review. *J. Cult. Herit.* **2018**, *31*, 170–180. [CrossRef]
34. Štekerová, M. Chatbots in Museums: Is Visitor Experience Measured? *Mus. Manag. Curatorship* **2022**, *37*, 505–521. [CrossRef]
35. Kim, Y.; Lee, H.; Kang, M. Interactive Description to Enhance Accessibility and Engagement in Museums for DHH Individuals. *Front. Educ.* **2023**, *8*, 1123655. [CrossRef]
36. Vaz, R.; Nisi, V.; Oakley, I. Wise Stones: An Interactive Accessible Circuit Designed to Enhance the Experiences of Visitors with Disabilities. In *MuseWeb 2020: Inclusive Digital Interactives*; MuseWeb: Silver Spring, MD, USA, 2020.
37. Koustriava, E.; Koutsmani, M. Spatial and Information Accessibility of Museums and Places of Historical Interest: A Comparison between London and Thessaloniki. *Sustainability* **2023**, *15*, 16611. [CrossRef]
38. Chai-Arayalert, S.; Panichpathom, S.; Sujindaratana, P. Chatbot-mediated technology to enhance experiences in historical textile museums. *Cogent Arts Humanit.* **2024**, *11*, 2396206. [CrossRef]
39. Yang, J.; Mousas, C. Embodied Conversational Agents in Extended Reality: A Systematic Review. *IEEE Access* **2025**, *13*, 56032–56058. [CrossRef]
40. SHINING 3D. EinScan Pro HD: High-Definition, Multi-Functional Handheld 3D Scanner (Including Solid Edge SHINING 3D Edition) [Product Brochure]. 2022. Available online: https://www.einscan.com/wp-content/uploads/2022/09/%E3%80%90EN%E3%80%91EinScan-PRO-HD-V0.13.pdf (accessed on 3 August 2025).
41. United Nations. Universal Declaration of Human Rights. 1948. Available online: https://www.un.org/en/about-us/universal-declaration-of-human-rights (accessed on 3 August 2025).
42. Descola, P. *Les Formes du Visible*; Seuil: Paris, France, 2021.