Beyond Anachronism: A RAG-Based Framework for Historically Consistent Conversational Agents

Fabio Clarizia[1,*], Pasquale Esposito[1], Annarita Giunto[1] and Rocco Loffredo[2]

[1]Hermes Lab, Department of Humanities, University of Salerno, via Giovanni Paolo II 132, Fisciano, Italy
[2]Department of Electronic Engineering, University of Salerno, via Giovanni Paolo II 132, Fisciano, Italy

Abstract

Conversational agents based on Large Language Models (LLMs) offer new opportunities for educational applications and cultural heritage dissemination, but have limitations in terms of accuracy, temporal coherence, and the risk of anachronism. To address these challenges, the proposed system adopts an architecture based on Retrieval-Augmented Generation (RAG), which integrates a generative language model with a historical documentary corpus, enabling the generation of responses grounded in verifiable sources. The system architecture combines a vector database implemented with ChromaDB, an embedding model and a backend developed with FastAPI. Several language models were tested within the same framework to evaluate their ability to maintain narrative coherence, historical accuracy and hallucination mitigation. Results indicate that medium-sized models offer the best trade-off between generative quality and faithfulness to the sources. A user study with 52 participants shows a high perceived levels of usability, engagement and educational value. The findings suggest that historical conversational agents based on RAG can represent promising tools for teaching and cultural heritage dissemination, supporting interactive and personalized learning experiences.

Keywords

Historical Conversational Agent, Retrieval-Augmented Generation, Large Language Models, Educational Technology

1. Introduction

Conversational agents have become a central component of modern digital systems, with applications spanning education, cultural heritage, and knowledge access [1, 2, 3]. The recent adoption of LLMs has significantly enhanced their fluency and interaction capabilities, enabling more natural and context-aware dialogues [4, 5]. However, these advantages come with well-known limitations, particularly in terms of factual reliability, controllability, and consistency in knowledge-intensive domains [6, 7].

These issues become especially critical in the case of conversational agents designed to embody historical figures [8, 9]. Unlike standard information retrieval systems, these agents adopt a first-person perspective grounded in a specific historical identity [10]. This requires not only linguistic coherence, but also strict adherence to factual accuracy, temporal consistency, and cultural and ideological context. In practice, LLM-based systems often struggle with anachronisms and the generation of plausible but historically incorrect statements, due to their training on heterogeneous and temporally mixed corpora [6, 11].

To address these challenges, this work investigates the integration of RAG with persona-based constraints for the development of historically grounded conversational agents. While RAG improves factual grounding by incorporating external knowledge sources, and persona modelling enforces narrative consistency, their combined use in historical educational settings remains underexplored, particularly with respect to temporal coherence and evaluation strategies.

This paper addresses this gap by proposing a unified framework for the design and evaluation of historical conversational agents. The main contributions are:

WOA 2026 – 27th Workshop "From Objects to Agents", June 15–17, 2026, Salerno, Italy
*Corresponding author.
fclarizia@unisa.it (F. Clarizia); pasesposito@unisa.it (P. Esposito); a.giunto1@studenti.unisa.it (A. Giunto); rloffredo@unisa.it (R. Loffredo)
0000-0003-2660-6759 (F. Clarizia); 0009-0006-3464-5861 (P. Esposito); 0009-0009-6665-0229 (A. Giunto); 0009-0007-0799-1168 (R. Loffredo)
© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-10-30

CEUR-WS.org/Vol-4307/paper16.pdf

- a RAG-based architecture combined with persona-driven generation to ensure factual grounding and narrative consistency;
- a mechanism for enforcing temporal constraints to reduce anachronistic outputs in historical dialogue systems;
- an evaluation framework combining automatic metrics and human assessment for historical conversational quality;
- the definition of the Historical Consistency Score (HCS) as a composite metric for measuring factual, temporal, and persona alignment.

The proposed framework is validated through an experimental study on a conversational agent embodying Carlo Pisacane. Results show that the integration of retrieval mechanisms and persona constraints improves factual reliability, reduces hallucinations, and enhances narrative coherence, supporting its applicability in educational and cultural heritage contexts.

2. Related Work

Conversational agents have been widely studied in recent years across multiple domains, including education, cultural heritage, and knowledge-intensive applications. In educational settings, they have been shown to support learning processes by enabling interactive tutoring, adaptive feedback, and personalized dialogue strategies [12, 13, 3]. In parallel, within the cultural heritage domain, natural language technologies have been used to improve access to historical knowledge by transforming unstructured content into conversational interfaces [14].

A relevant research direction concerns Embodied Conversational Agents (ECAs), which aim to simulate human-like interaction by combining verbal and non-verbal communication. These systems have been shown to improve user engagement and interaction quality by providing more natural and socially rich exchanges [9]. Building on this line of work, persona-based conversational agents extend the concept by enabling systems to adopt consistent identities and perspectives during interaction [10]. This aspect is particularly relevant for narrative-driven applications, where coherence of identity plays a central role.

With the advent of LLMs, conversational systems have significantly improved in terms of fluency, adaptability, and contextual understanding. However, these models also introduce well-documented limitations, including hallucinations and the generation of factually incorrect yet plausible outputs [6, 7, 11]. Such issues are particularly critical in knowledge-intensive domains, where factual correctness is essential.

To mitigate these limitations, RAG has been proposed as an effective approach to improve factual grounding by incorporating external knowledge sources into the generation process. RAG-based systems have demonstrated improved reliability in question answering and dialogue generation tasks by reducing dependence on parametric memory.

Despite these advances, the integration of retrieval mechanisms, persona-based modelling, and temporal consistency constraints within a unified framework remains relatively underexplored, particularly in the context of historical conversational agents and educational applications.

3. System Methodology

The proposed methodology defines a unified framework for the design of historical conversational agents, combining retrieval-based grounding, persona modelling, and controlled language generation. The system is structured as a coherent pipeline in which each component contributes to ensuring factual accuracy, temporal consistency, and narrative fidelity. To generalize the approach beyond a single historical figure, we define a design framework composed of four core principles. This framework provides methodological guidance for constructing historically grounded conversational agents, ensuring accuracy, temporal coherence, and persona consistency.

3.1. Design Framework

The first principle is documentary grounding, which ensures that the agent relies exclusively on verified historical sources. This provides a solid epistemological foundation and reduces the risk of generating anachronistic or unsupported statements. The second principle is temporal constraint modelling, which introduces explicit chronological restrictions to prevent references to events or knowledge outside the historical period of the represented figure. The third principle is persona identity modelling, which ensures that the system maintains a stable first-person perspective consistent with the linguistic and stylistic characteristics of the historical subject. The fourth principle is retrieval-grounded reasoning, where responses are generated by integrating retrieved documentary evidence into the generation process, ensuring that outputs remain both factually grounded and narratively coherent.

3.2. System Architecture

The proposed framework is implemented through a multi-layered system designed to operationalize its core principles in a computational pipeline. The system is organized into four main layers: document level, semantic representation level, generative level, and interaction level. Each layer corresponds to a specific functional component of the architecture and contributes to ensuring retrieval grounding, temporal coherence, and persona-consistent generation. The document level consists of a curated collection of historical sources related to Carlo Pisacane, including academic articles, historical essays, original letters, and scholarly publications. These sources are selected to ensure balanced coverage of both biographical events and ideological aspects of the historical figure. Rather than being treated as passive input, these documents actively constrain and guide the generation process, improving epistemic reliability.

The semantic representation level is based on embedding models and a vector database that enable efficient retrieval of contextually relevant information. Texts are segmented into chunks of 500 tokens with an overlap of 80 tokens to preserve semantic continuity between adjacent segments. Embeddings are generated using BGE-M3, while ChromaDB is used as the vector database to support semantic search, persistence, and modular retrieval.

The retrieval process uses a top-k value of 4, selecting the most relevant document segments for each query. This configuration balances contextual completeness and noise reduction, ensuring that the model receives sufficient but controlled contextual information. The overall dataset consists of 936 chunks derived from eight historical documents, ensuring a compact yet semantically rich knowledge base.

The generative level is based on an open-source large language model deployed through the Ollama framework. In this work, LlaMA 3.2 3B is selected due to its balance between generative capability, contextual control, and computational efficiency [5]. The model is configured with a limited maximum token length and a low temperature setting to reduce variability and improve response stability. It is constrained to respond in the first person, adopting the persona of Carlo Pisacane, while restricting outputs to historically grounded and verifiable information. Overall, the main configuration parameters of the proposed system, including document collection, chunking strategy, and retrieval settings, are summarized in Table 1. For clarity, Table 2 reports the acronyms and symbols used throughout the paper.

Table 1
RAG Configuration and Computational Parameters

Parameter	Value
Documents	8
Total chunks	936
Chunk size (tokens)	500
Chunk overlap (tokens)	80
Embedding model	BGE-M3
Embedding dimension	1024
Vector database	ChromaDB
Top-k retrieval	4
Generative model	LlaMA 3.2 3B
Inference framework	Ollama

Table 2
Acronyms and symbols used in the proposed framework.

Acronyms	
RAG	Retrieval-Augmented Generation
LLM	Large Language Model
HCS	Historical Consistency Score
ECA	Embodied Conversational Agent
NLP	Natural Language Processing
Symbols	
F	Factual correctness score
T	Temporal coherence score
P	Persona consistency score
HCS	Composite metric: $(F+T+P)/3$
k	Top-k retrieval parameter

The interaction level is responsible for managing communication with the user through a FastAPI backend, exposing a conversational endpoint that orchestrates retrieval and generation in real time. Figure 1 illustrates the overall architecture of the proposed system and summarizes the interaction between the main components of the methodology. The implementation of the proposed system follows

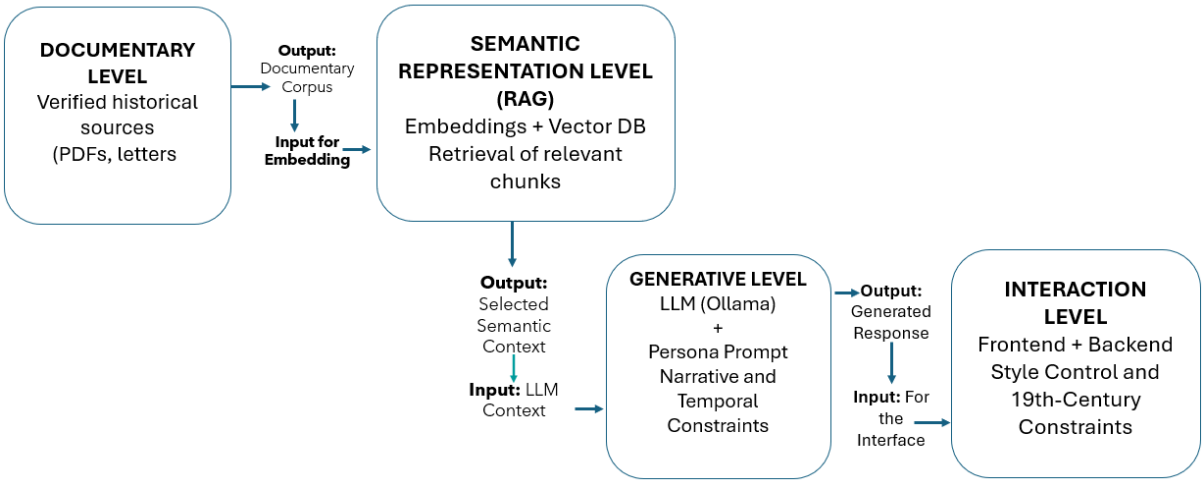


Figure 1: Overview of the historical conversational agent's architecture.

a modular architecture designed to support RAG and ensure controllable, context-aware dialogue generation. In the initial development phase, a lightweight prototype was implemented using an SQLite-based

storage system, where textual information was stored as structured records. Semantic representations were generated using the all-MiniLM-L6-v2 embedding model, selected for its efficiency and suitability for CPU-based execution. Although this configuration enabled rapid prototyping, it exhibited several limitations, including restricted contextual understanding, limited ability to prioritize relevant historical information, and poor scalability, since any update to the dataset required recomputation of all embeddings.

To overcome these limitations, the final system adopts a more robust retrieval infrastructure based on a chunked document representation and vector-based semantic search. Textual sources are segmented using a Recursive Character Text Splitter, producing coherent chunks that preserve local semantic continuity. Embeddings are generated using the BGE-M3 model, while ChromaDB is used as the vector database to enable efficient similarity search, persistence, and modular integration with the rest of the system.

The embedding model BGE-M3 produces dense vector representations with a fixed dimensionality of 1024, ensuring a balance between semantic expressiveness and computational efficiency. The retrieval system is optimized for CPU-based execution and does not require GPU acceleration during inference.

Retrieval latency is dependent on corpus size and indexing state; ChromaDB supports approximate nearest neighbor search, enabling efficient similarity computation even without GPU acceleration.

The system operates under a lightweight inference setup, where the language model is executed locally through the Ollama framework. This design choice ensures portability and reproducibility while maintaining acceptable response times for interactive use cases.

The retrieval process is orchestrated through a pipeline that transforms the user query into a vector representation and retrieves the most semantically relevant document segments from the knowledge base. This mechanism allows the system to ground generation in historical evidence rather than relying solely on parametric knowledge.

To improve system reliability, an explicit validation step is introduced to assess the quality of retrieved documents. When no sufficiently relevant context is found, the system refrains from generating speculative responses and instead signals the absence of reliable information. The input prompt to the language model is constructed by combining three elements: the persona definition of Carlo Pisacane, the retrieved contextual documents, and the user query. This structured prompt design ensures consistency between historical grounding and narrative coherence, while also constraining the model to remain within the defined temporal and ideological boundaries.

The system is deployed through a FastAPI-based backend that exposes a conversational endpoint, providing a lightweight and efficient interface for real-time interaction between the user and the RAG pipeline.

All experiments were conducted on a single machine equipped with an Intel Core i3-8145U processor (8th generation, 2.10 GHz), Intel UHD Graphics 620 integrated GPU, with inference running in CPU-only mode via the Ollama framework. Retrieval latency and end-to-end response time were not formally measured in the current study; characterising these aspects quantitatively represents a planned direction for future work, alongside larger-scale benchmarking of the full pipeline.

3.3. Historical Consistency Score (HCS)

To systematically evaluate the quality of responses generated by the historical conversational agent, we introduce the HCS, a composite evaluation metric designed to measure the fidelity of generated responses to historical figures.

The metric evaluates three dimensions:

- **Factual correctness**: accuracy of information with respect to verified historical sources.
- **Temporal coherence**: adherence to chronological constraints, avoiding anachronistic references.
- **Persona consistency**: alignment with the linguistic style, perspective, and identity of the historical figure.

Each dimension is independently assessed by expert evaluators on a normalized scale from 0 to 1. The final score is computed as the arithmetic mean:

$$HCS = \frac{F + T + P}{3} \tag{1}$$

where F, T, and P denote factual correctness, temporal coherence, and persona consistency respectively.

This metric complements standard evaluation procedures by providing a structured and interpretable measure of historical reliability and narrative consistency in generated responses.

4. Experimental Setup

4.1. Models and Evaluation Conditions

The experimental evaluation was conducted to assess the behavior of different LLMs within the same RAG framework. A comparative analysis was performed using a set of small and medium-sized models, including Phi-3, Gemma 3:4B, Gemma 3:1B, Gemma 2B, Mistral, Qwen 2.5 3B, and LlaMA 3.2 3B [15, 16, 17, 18, 5]. All models were evaluated under identical conditions in order to ensure a fair comparison.

4.2. Benchmark and Question Types

Each model was queried using a standardized set of prompts designed to cover different cognitive and factual dimensions: factual questions (e.g., birth and death dates of Carlo Pisacane), descriptive questions (e.g., role in the Sapri Expedition), motivational questions (e.g., political and ideological motivations behind the revolutionary action), and evaluative questions (e.g., assessment of the expedition outcome and relationship with Giuseppe Garibaldi).

To evaluate the impact of retrieval augmentation, an additional control setting was introduced. In this configuration, models were tested both with access to retrieved contextual documents and without any external context. This allowed the assessment of how strongly each model depends on external grounding for generating historically consistent responses.

The analysis focuses on three main evaluation dimensions: (i) factual consistency with respect to historical sources, (ii) degree of hallucination or inclusion of unsupported information, and (iii) ability to maintain the impersonated persona of Carlo Pisacane.

4.3. Baseline Configurations

To further validate the architecture, baseline configurations were introduced to isolate the contribution of each system component:

- **Baseline 1:** LLM without retrieval
- **Baseline 2:** LLM + retrieval without persona
- **Baseline 3:** LLM + persona without retrieval

5. Results

This section presents the experimental results obtained from the evaluation of the proposed historical conversational agent across different language models and configurations.

Table 3
Qualitative Comparison of Language Models

Model	Factual Accuracy	Narrative Coherence	Hallucination Control	Response Completeness	Without Documents
Gemma 3 4B	High	Medium	Low	High	Very Low
Mistral	High	Medium	Low	High	Very Low
Phi-3	Medium	Medium	Medium	Medium	Low
Qwen 2.5 3B	Medium	Medium	Medium	Medium	Low
LlaMA 3.2 3B	High	High	High	High	Low
Gemma 3 1B	Low	Low	High	Low	Very Low
Gemma 2B	Low	Low	High	Low	Very Low

Table 4
Evaluation Metrics for Generated Responses

Response ID	Faithfulness	Answer Relevance	Context Precision	Hallucination (%)
R1	0.95	0.92	0.90	5%
R2	0.90	0.88	0.85	10%
R3	0.92	0.85	0.88	8%
R4	0.97	0.95	0.93	3%

5.1. Qualitative Model Comparison

The experimental analysis was conducted by testing a series of small and medium-sized language models, including Phi-3, Gemma 3:4B, Gemma 3:1B, Gemma 2B, Mistral, Qwen 2.5 3B, and LlaMA 3.2 3B, within the same application framework. All models were queried using an identical set of questions designed to cover different types of requests: factual, descriptive, motivational, and evaluative questions.

The results, summarized in Table 3, show clear differences across model scales in terms of factual accuracy, narrative coherence, hallucination control, and response completeness. A qualitative interpretation of these patterns is provided in Section 6.

5.2. Impact of Retrieval Augmentation

To assess the contribution of retrieval mechanisms, responses generated with access to vectorized documents were compared to responses produced without contextual information [19]. Results confirm that access to retrieved documents substantially improves factual grounding across all model configurations. The detailed behavioral differences observed across model scales are discussed in Section 6.

A qualitative comparison of the evaluated models across key dimensions is reported in Table 3. The comparison highlights differences in factual accuracy, narrative coherence, hallucination control, and response completeness under both retrieval and no-context conditions.

5.3. Automatic Evaluation Metrics

To further evaluate system performance, an automatic assessment was conducted using standard RAG metrics [20], including faithfulness, answer relevance, context precision, and hallucination rate. The results are reported in Table 4.

5.4. Benchmark Evaluation

To ensure a structured evaluation, a benchmark of 40 historical questions was constructed, covering factual, descriptive, interpretative, and evaluative categories. Each response was evaluated in terms of correctness, hallucination rate, and narrative coherence.

A quantitative comparison of model performance across correctness, hallucination rate, and narrative coherence is presented in Table 5. The table also includes a baseline configuration without document retrieval, highlighting the critical role of external knowledge grounding.

The comparison shows that models without document access exhibit extremely low correctness and very high hallucination rates, confirming the fundamental role of retrieval mechanisms. Similarly,

Table 5
Model Performance Comparison

Model	Correctness	Hallucination Rate	Narrative Coherence	Notes
LlaMA 3.2 3B	0.92	4%	0.95	Best balance of accuracy and style
Gemma 3:4B	0.88	10%	0.90	Fluent but slightly hallucinating
Phi-3	0.85	8%	0.87	Moderate accuracy and coherence
Qwen 2.5 3B	0.83	12%	0.85	Some errors and missing context
Mistral	0.87	11%	0.88	Fluent but introduces some anachronisms
Gemma 3:1B	0.78	18%	0.80	Often cautious, incomplete answers
Gemma 2B	0.75	20%	0.78	Limited coverage, low factual accuracy
Without Documents	0.10	95%	0.20	Cannot answer without context

configurations without persona lose narrative coherence, while persona-only configurations tend to produce fluent but historically unsupported responses.

Overall, the full system consistently outperforms all baseline configurations, achieving higher correctness, lower hallucination rates, and stronger narrative coherence. These results demonstrate that effective historical conversational agents require a balanced integration of retrieval mechanisms and persona modelling.

5.5. Historical Consistency Score Results

The HCS was used to evaluate factual correctness, temporal coherence, and persona consistency across different system configurations. The results are reported in Table 6.

Table 6
Historical Consistency Score (HCS) evaluation results

Response ID	Factual	Temporal	Persona	HCS
R1	0.90	1.00	0.95	0.95
R2	0.85	0.90	0.90	0.88
R3	0.80	0.85	0.90	0.85
R4	0.95	0.95	0.98	0.96

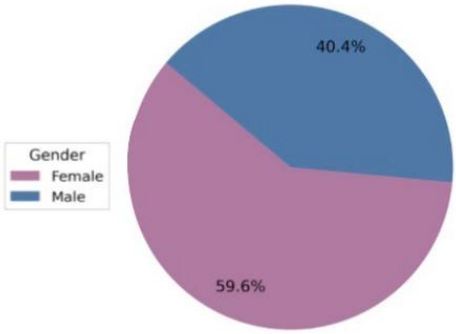
5.6. User Study

To contextualize the user study and support the interpretation of the experimental findings, this section reports a set of descriptive statistics related to the participant sample and their interaction with digital technologies. In particular, Figures 2a–4 illustrate the demographic composition of the participants, including gender distribution, age distribution, educational background, and patterns of technology usage.

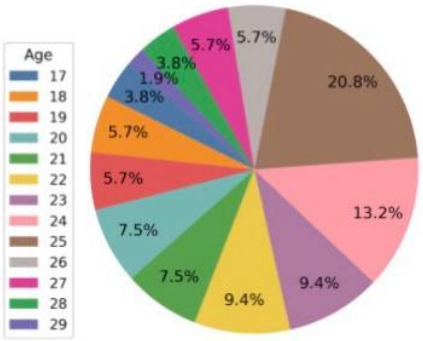
These variables are relevant for interpreting user perceptions of the conversational agent, as prior familiarity with digital tools and educational level may influence both interaction behavior and evaluation outcomes. Additionally, the reported distributions provide transparency regarding the composition of the evaluation sample, ensuring a clearer understanding of the context in which the system was tested.

Overall, communication with the agent was perceived as clear, coherent, and functional, contributing to a positive educational experience. The agent was described as friendly, reassuring, and somewhat human-like, characteristics that supported emotional engagement and user involvement. The interaction successfully stimulated curiosity and interest, confirming the potential of conversational agents as tools for supporting historical learning.

Participants showed a high willingness to interact with the agent (see Figure 5), and the perceived age of the agent did not significantly influence this attitude, suggesting that communication quality plays a more important role than anthropomorphic attributes in shaping user acceptance.

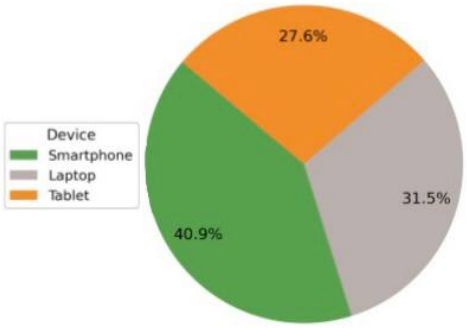


(a) Gender Distribution

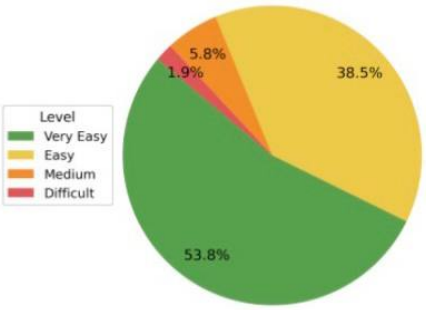


(b) Age Distribution

Figure 2: Distributions overview



(a) Daily Technology Usage (by Device)



(b) Perceived Difficulty with Devices

Figure 3: Technology usage and perceived difficulty

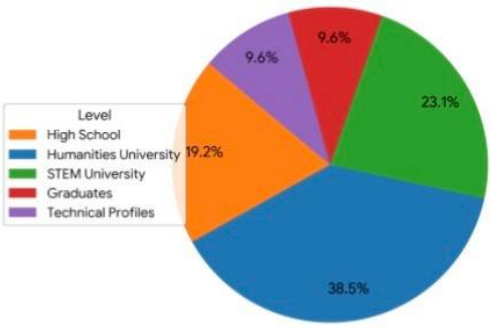


Figure 4: Educational Background.

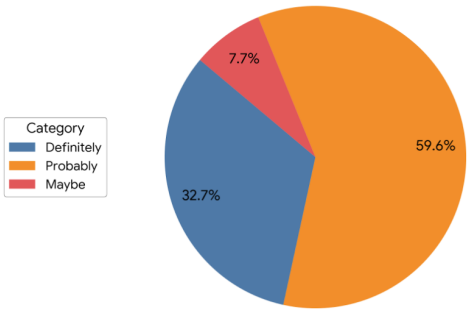


Figure 5: Willingness to Interact.

The conversational agent received positive evaluations across all dimensions, including historical coherence, content accuracy, clarity, engagement, and educational effectiveness (see Figures 6, 7, and 8). Users found the communicative style natural and approachable, with the virtual character fostering curiosity and participation. These results highlight the potential of conversational agents as effective mediators of historical knowledge, supporting comprehension and learning, with potential extensions to other figures, contexts, and multi-modal interactions.

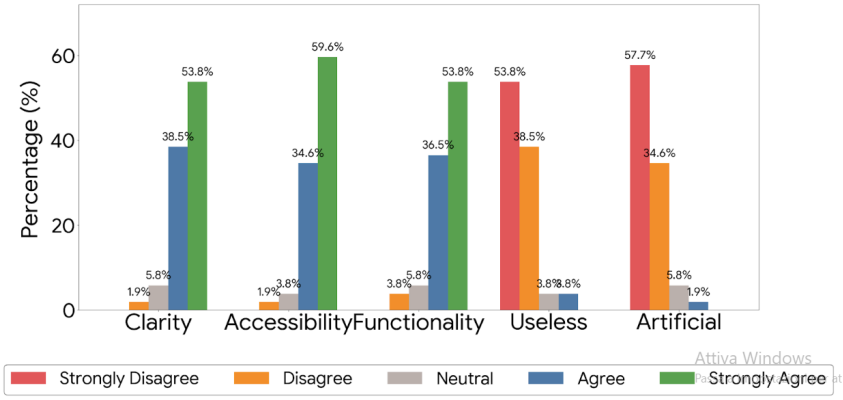


Figure 6: Pragmatic Qualities.

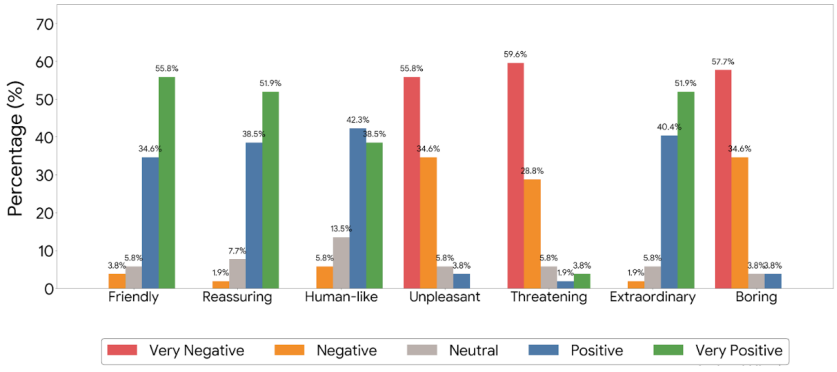


Figure 7: Hedonic Qualities and Feeling.

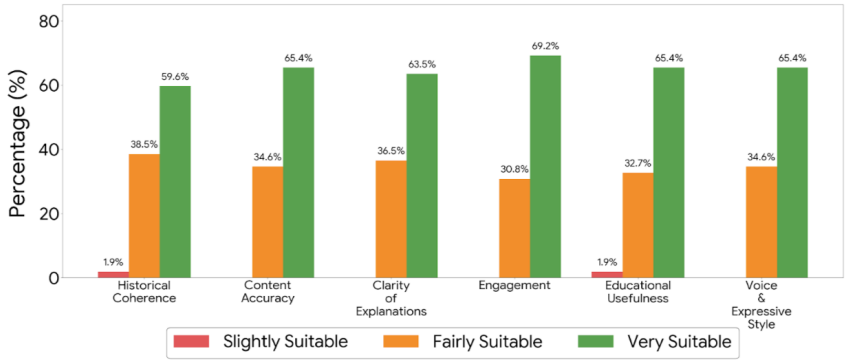


Figure 8: Overall Assessment.

6. Discussion

The results of the experimental evaluation provide several insights into the design of retrieval-augmented historical conversational agents. Building on the quantitative data reported in Tables 3–6, the following analysis interprets the observed patterns across model scales, retrieval configurations, and persona constraints.

First, the comparison across models confirms that model size alone is not a sufficient indicator of performance in RAG-based systems. Larger models, such as Gemma 3:4B and Mistral, produce fluent and well-structured responses, but frequently introduce information not supported by the retrieved documents. This behavior reflects a tendency to prioritize linguistic plausibility over factual grounding, increasing the risk of hallucinations, as highlighted in prior work [6, 7]. Conversely, smaller models such as Gemma 3:1B and Gemma 2B demonstrate a more conservative strategy, often avoiding responses when contextual information is uncertain. While this reduces hallucination rates, it negatively impacts interaction quality, leading to incomplete answers and limiting the system's usability in conversational settings [16].

Intermediate models, including Phi-3 and Qwen 2.5 3B, exhibit a trade-off between these two extremes. Although they are capable of generating contextually relevant responses, their performance remains unstable, particularly in maintaining historical coherence and persona consistency [15, 18].

These findings confirm the importance of combining retrieval mechanisms and persona modelling for achieving both factual reliability and narrative consistency.

From a broader perspective, the results confirm that the effectiveness of RAG-based systems is determined by the interaction between retrieval quality, prompt design, and model behavior, rather than by model size alone. This has important implications for the design of educational and historical conversational agents, where factual accuracy and interpretative coherence are critical. The combination of qualitative and quantitative evaluation provides a comprehensive view of system performance and supports reproducibility across RAG-based architectures.

6.1. Comparison with Existing Evaluation Metrics

Traditional evaluation approaches for natural language generation, such as ROUGE, BLEU, and BERTScore, primarily focus on lexical overlap or general semantic similarity between generated and reference texts. More recent approaches for factual consistency, including entailment-based hallucination detection models, extend this perspective by assessing whether generated statements are supported by external evidence.

However, these methods are not specifically designed for historical conversational settings, where additional dimensions such as temporal coherence and persona consistency play a crucial role. In contrast, the proposed Historical Consistency Score (HCS) explicitly integrates three complementary dimensions: factual correctness, temporal coherence, and persona consistency.

This design allows HCS to jointly capture:

(i) factual alignment with historical sources, (ii) chronological validity with respect to the time period of the simulated entity, (iii) stylistic and identity consistency in persona-driven dialogue.

Compared to standard metrics, HCS provides a task-specific evaluation framework tailored to historical conversational agents, where both narrative identity and temporal grounding are first-class constraints rather than secondary properties.

6.2. Educational Implication

The experimental results suggest that retrieval-augmented historical conversational agents can provide meaningful support in educational contexts, particularly in the learning of complex historical content.

The observed balance between factual accuracy, narrative coherence, and interaction quality, especially in the case of LlaMA 3.2 3B, indicates that such systems can effectively present historically grounded information while maintaining an engaging conversational style. This combination is particularly relevant in educational settings, where both correctness and engagement play a central role in facilitating learning [21].

From a pedagogical perspective, the integration of persona modelling allows the system to simulate first-person historical narratives, fostering a more immersive learning experience. This narrative dimension can support deeper cognitive processing by encouraging users to engage with historical events not only as abstract facts but as lived experiences, potentially improving both comprehension and retention.

Moreover, the use of retrieval mechanisms ensures that the generated responses remain anchored to verifiable sources, addressing a key limitation of traditional conversational agents. This aspect is crucial in educational applications, where the reliability of information directly impacts learning outcomes [22].

The system also supports a form of self-paced and exploratory learning. Users can interact with the agent according to their own interests, asking follow-up questions and focusing on specific topics. This flexibility aligns with constructivist learning approaches, in which knowledge is actively built through interaction and inquiry.

However, the results also highlight important limitations. The effectiveness of the system depends on the balance between model capability and control mechanisms: overly generative models risk introducing inaccuracies, while overly conservative models may limit interaction. Additionally, the reliance on textual interaction may pose challenges for users with lower digital literacy or limited prior knowledge of the subject.

To further validate the educational impact of the proposed approach, future work should include controlled user studies based on pre- and post-interaction assessments. Such evaluations would allow the measurement of knowledge acquisition and provide quantitative evidence of learning gains.

Finally, the integration of multimodal features, such as audio narration or visual historical materials, represents a promising direction for enhancing engagement and accessibility. Extending the framework to other historical figures or domains could further demonstrate its potential as a scalable tool for digital education and cultural dissemination [3].

6.3. Limitations and Future Works

Despite the promising results, several threats to validity should be considered when interpreting the findings of this study. **Internal validity.** The experimental results depend on specific design choices, including prompt structure, retrieval parameters (e.g., top-k value), and model configuration (temperature and token limits). Variations in these parameters may lead to different behaviors, particularly in terms of hallucination control and narrative coherence. Consequently, the observed performance may not be solely attributable to the model itself, but also to the overall system design. **Construct validity.** The evaluation relies on a combination of qualitative judgments and quantitative metrics, including RAG-based measures and the HCS. While these metrics provide useful indicators of system performance, they may not fully capture all dimensions of conversational quality, such as interpretative

depth or pedagogical effectiveness. Furthermore, some evaluation criteria involve expert judgment, which may introduce a degree of subjectivity [22, 21]. **External validity.** The study focuses on a single historical domain, namely Carlo Pisacane and the Sapri expedition. Although this enables a controlled and in-depth evaluation, it limits the generalizability of the results to other historical figures, time periods, or domains. In addition, the documentary corpus consists of eight curated sources; while reliable, this limited dataset may constrain the diversity of perspectives and the overall coverage of the knowledge base.

Another important limitation concerns the dependence on curated historical material. Historical interpretation is often influenced by historiographical perspectives, cultural context, and source selection. As a consequence, generated responses may implicitly reflect biases present in the underlying corpus. Future systems should therefore investigate mechanisms for handling conflicting sources and representing multiple historical viewpoints in a transparent manner. **Statistical conclusion validity.** The experimental evaluation is based on a relatively small benchmark (40 questions) and a limited number of response samples. Although sufficient for comparative analysis, this scale may not fully capture the variability of model behavior across a broader range of queries. Larger benchmarks could provide more robust and statistically significant results. **User-related validity.** Potential biases may arise from the characteristics of the participants involved in the evaluation. Similar levels of educational background, domain knowledge, or familiarity with conversational systems could influence both usability assessments and perceived engagement, limiting the representativeness of the findings.

Inter-rater reliability was not formally quantified, as the qualitative assessment relied on a single expert evaluator per dimension. While the evaluation protocol was standardized to reduce subjectivity, future studies will employ multiple independent raters and report agreement via Cohen's kappa or Krippendorff's alpha. Similarly, the comparative results are reported descriptively; significance testing over larger benchmarks is left to future work.

Future Research Directions Future work will focus on extending the proposed framework to additional historical figures, educational contexts, and multilingual environments. Although the architecture is designed to be domain-independent, scaling the system to broader historical settings requires larger documentary corpora, more heterogeneous sources, and more sophisticated retrieval strategies.

Another promising direction concerns the integration of multimodal interaction features, such as audio narration, historical images, maps, and archival materials, in order to enhance accessibility and user engagement. Furthermore, future research could explore automated approaches for detecting anachronisms and historically inconsistent generations through temporal reasoning and specialized evaluation modules.

Finally, larger-scale educational studies based on longitudinal evaluation protocols and pre-/post-interaction assessments could provide stronger evidence regarding the pedagogical impact of retrieval-augmented historical conversational agents.

Overall, addressing these limitations by expanding the dataset, increasing the scale of evaluation, diversifying user samples, and refining evaluation methodologies would strengthen the robustness and generalizability of the proposed approach.

Regarding scalability and domain transferability, while the proposed architecture is designed to be domain-agnostic, its current evaluation is limited to a single historical figure. Future work will investigate the extension of the framework to multiple historical personalities and heterogeneous corpora, in order to assess robustness across different temporal contexts, narrative styles, and knowledge distributions.

7. Conclusion

This study presented the design and evaluation of a retrieval-augmented conversational agent for the simulation of a historical figure, specifically Carlo Pisacane. The proposed system integrates document-grounded retrieval, persona modelling, and controlled language generation to ensure historically consistent and contextually reliable interactions.

The experimental results demonstrate that the combination of retrieval mechanisms and constrained generation is essential to reduce hallucinations and improve factual accuracy in historical dialogue systems. In particular, models that integrate external knowledge sources consistently outperform those relying solely on parametric memory, confirming the importance of grounding generation in verified documentary evidence [19, 7].

The comparative analysis of different language models further highlights that system performance does not depend exclusively on model size, but on the interaction between retrieval quality, prompt design, and generation control. Among the tested configurations, LlaMA 3.2 3B achieves the most balanced behavior, combining factual reliability, narrative coherence, and computational efficiency.

From a broader perspective, the study shows that retrieval-augmented and persona-based architectures can effectively support the development of historically grounded conversational agents, with potential applications in education and cultural dissemination. However, achieving reliable performance requires careful integration of all system components rather than optimization of a single element.

Future work will focus on extending the approach to multiple historical figures and domains, improving the robustness of evaluation methodologies, and exploring multimodal extensions to enhance engagement and accessibility.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4 in order to: Text Translation, Improve writing style. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] E. C. Acikgoz, C. Qian, H. Wang, V. Dongre, X. Chen, H. Ji, D. Hakkani-Tur, G. Tur, A desideratum for conversational agents: Capabilities, challenges, and future directions, arXiv preprint (2025). URL: https://arxiv.org/abs/2504.16939.

[2] X. H. H.-Y. Shum, D. Li, From eliza to xiaoice: Challenges and opportunities with social chatbots, arXiv preprint (2018). URL: https://arxiv.org/abs/1801.01957.

[3] L. Labadze, M. Grigolia, L. Machaidze, Role of ai chatbots in education: systematic literature review, International journal of Educational Technology in Higher education 20 (2023) 56.

[4] D. Khurana, A. Koli, K. Khatter, S. Singh, Natural language processing: state of the art, current trends and challenges, Multimedia tools and applications 82 (2023) 3713–3744.

[5] A. Grattafiori, A. Dubey, A. J. et al., The llama 3 herd of models, arXiv preprint (2024). URL: https://arxiv.org/abs/2407.21783.

[6] E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the dangers of stochastic parrots: Can language models be too big?, FAccT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 610–623. URL: https://doi.org/10.1145/3442188.3445922. doi:10.1145/3442188.3445922.

[7] Z. J. et al., Survey of hallucination in natural language generation, arXiv preprint (2023). URL: https://arxiv.org/abs/2202.03629.

[8] W. L. Johnson, J. W. Rickel, J. C. Lester, et al., Animated pedagogical agents: Face-to-face interaction in interactive learning environments, International Journal of Artificial intelligence in education 11 (2000) 47–78.

[9] E. André, C. Pelachaud, Interacting with embodied conversational agents, in: Speech technology: Theory and applications, Springer, 2010, pp. 123–149. doi:10.1007/978-0-387-73819-2_8.

[10] J. L. et al., A persona-based neural conversation model, arXiv preprint (2016). URL: https://arxiv.org/abs/1603.06155.

[11] S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, Detecting hallucinations in large language models using semantic entropy, Nature 630 (2024) 625–630.

[12] R. Beale, Dialogic pedagogy for large language models: Aligning conversational ai with proven theories of learning, arXiv preprint arXiv:2506.19484 (2025).
[13] J. Ceha, K. J. Lee, E. Nilsen, J. Goh, E. Law, Can a humorous conversational agent enhance learning experience and outcomes?, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, 2021, pp. 1–14. doi:10.1145/3411764.3445068.
[14] A. Iannella, T. Labruna, L. Santercole, B. Viti, Reclaiming conversation: Introducing a novel approach to using conversational ai at school, Conversational agents as online learning tutors (2021) 119–152.
[15] M. I. Abdin, S. A. Jacobs, A. A. A. et al., Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, Technical Report, Microsoft Research, 2024. URL: https://arxiv.org/abs/2404.14219, technical Report MSR TR 2024-12.
[16] T. M. et al., Gemma: Open models based on gemini research and technology, arXiv preprint (2024). URL: https://arxiv.org/abs/2403.08295.
[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, et al., Mistral 7b, arXiv preprint (2023). URL: https://arxiv.org/abs/2310.06825.
[18] Q. Team, Qwen2.5 technical report, arXiv preprint (2024). URL: https://arxiv.org/abs/2412.15115.
[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.
[20] Y. Gao, Y. Xiong, X. G. et al., Retrieval-augmented generation for large language models: A survey, arXiv preprint (2023). URL: https://arxiv.org/abs/2312.10997.
[21] F. Guo, A. Metallinou, C. Khatri, A. Raju, A. Venkatesh, Topic-based evaluation for conversational bots, arXiv preprint (2018). URL: https://arxiv.org/abs/1801.03622.
[22] H. Ding, J. Simmich, A. Vaezipour, N. Andrews, T. Russell, Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review, Journal of the American Medical Informatics Association (2023). doi:10.1093/jamia/ocad222.