Formatting Instructions for the 24th International Conference on Autonomous Agents and Multiagent Systems

Kevin Tang
Northeastern University
Boston, United States
tang.kevi@northeastern.edu

Tanay Pandya
Northeastern University
Boston, United States
pandya.t@northeastern.edu

Avery Huang
Northeastern University
Boston, United States
huang.av@northeastern.edu

ABSTRACT

Traditional learning methods often fail to capture the attention of the learner since there is a discrepancy between the ideas being taught and what the learner knows. This results in reduced engagement and a longer comprehension process for the learning materials. Our project aims to alleviate this issue through interactions with simulated historical figures. These figures were created with an emphasis on historical accuracy, distinctive speech patterns, and adaptive conversational behavior, allowing for a more natural learning experience. Our approach is unique from other work since we want to capture the historical figure as a whole with only access to what we know about them in history. Additionally, there is a focus on having a conversation verbally instead of solely being a chatbot. Preliminary results indicate that this method may not only be more engaging but also potentially help users recall facts and historical context faster and more accurately. These early results are encouraging and support the potential of character-driven educational tools to transform how we learn.

KEYWORDS

Artificial Intelligence, Human AI Interaction, RAG, LLM

ACM Reference Format:
Kevin Tang, Tanay Pandya, and Avery Huang. 2025. Formatting Instructions for the 24th International Conference on Autonomous Agents and Multiagent Systems. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 5 pages.

1 INTRODUCTION

Current educational methods fail to engage users and provide them with the necessary context to fully grasp the information. As seen when studying history, textbooks and lectures fail to have a meaningful lasting impact. Additionally, often the actions of the past can be seen as contradictory or even controversial to things that we know now in modern day. Our project attempts to engage the user more thoughtfully by establishing a connection between the ancient and modern day ideas. We believe that by creating a virtual clone of a historical figure, it will boost engagement and understanding of history.

A more personalized learning experience can be cultivated with a conversation with a historical figure, however there were some challenges at hand. The first was to ensure that our historical figure was as accurate as possible to the original, which in our case was Pliny the Elder. This meant providing the clone with accurate information and writings from Pliny the Elder. Furthermore, the manner of speech needed to match the specific time period as well. To fix this, we had the LLM model mimic the writing of people within that time frame. Finally, to ensure that the model would accurately portray Pliny the Elder, prompt engineering was applied.

From here, we had three hypotheses we wanted to test:

- H1: Our system will be able to answer questions accurately with respect to the knowledge known by the historical figure
- H2: Our system will increase engagement and speed up learning among users
- H3: Our system will produce statements that are indistinguishable in tone from the source figure's tone

Initial surveys demonstrate the possibility of increased engagement and comprehension, along with a faster learning process. While initial tests indicate our system demonstrates a high degree of historical accuracy, we found that tonal authenticity suffered to an extent.

2 BACKGROUND/RELATED WORK

Research into creating virtual representations of individuals has been an active area of study across healthcare, education, and historical preservation. One notable example is the SimSensei Kiosk, a virtual interviewer designed to support mental health assessments by facilitating emotionally intelligent conversations with users [1]. The primary focus of this work was to make users feel comfortable and encourage open communication, resulting in emotionally responsive, rather than fact-driven, interactions. A closer parallel to our work is the New Dimensions in Testimony project, which aimed to preserve the personal stories of Holocaust survivors through interactive interviews [5]. In this project, survivors were asked a wide range of questions in advance, and their recorded responses were used to build a system that allows users to ask questions and receive relevant answers from a virtual model. While similar in its goal of historical preservation, their system benefited from the ability to interview the individual directly. In contrast, our work must rely entirely on secondary sources—historical texts, autobiographies, and written records—making the process of recreating a virtual persona significantly more challenging. Another related effort involved developing AI-generated digital mementos of historical figures captured at specific moments in time [3]. These chatbots offered users the opportunity to interact with figures as they might have been during a particular phase of life. Our approach differs in that we aim to create a more holistic, temporally rich representation that reflects the subject's life as a whole, rather than a snapshot.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

This broader scope introduces additional complexity but also offers the potential for deeper engagement and understanding.

3 METHODOLOGY

3.1 Data

For this paper, we chose to recreate Pliny the Elder, a Roman naturalist and historian who most notably penned the *Historia Naturalis* (*The Natural History*). This is effectively an early encyclopedia of the information known to the Romans, encompassing information about astronomy, natural phenomena, and notable figures and locations, among many other things. To acquire information that would have been known by Pliny the Elder, we download the publicly available English translation of *The Natural History* [4] from the Perseus Digital Library [6].

The Natural History is comprised of 36 books, each encompassing a specific topic across varying amounts of chapters. For the purposes of RAG, we considered each chapter to be its own document.

3.2 System Architecture

Our system is designed to deliver immersive, historically grounded conversations with minimal latency and high fidelity to the source persona. The architecture is composed of four major components: **speech recognition**, **retrieval-augmented generation**, **response formatting**, and **voice synthesis**. Figure 1 provides an overview of the end-to-end pipeline.

The user begins by speaking or typing a query into the system. If the input is spoken, it is transcribed using Whisper, an automatic speech recognition (ASR) model capable of handling diverse accents and varying audio conditions. For typed input, the query directly proceeds to the orchestrator module.

The *orchestrator* acts as the central controller, taking in the transcribed (or typed) query and querying a vector database of historical texts. To build this database, we collected writings by and about Pliny the Elder—especially *The Natural History*—and chunked them into semantically meaningful passages. These passages were embedded using a SentenceTransformer model and indexed using a vector store.

At runtime, the user query is embedded and matched against this vector database to retrieve the top-K relevant passages. These are combined with the user query and passed as a single prompt to the language model. We experimented with several approaches but ultimately chose a **Retrieval-Augmented Generation (RAG)** pipeline for its improved factual grounding and adaptability to historical context.

The language model (GPT-4 in our case) is responsible for generating the final response. To ensure tone alignment, we tuned our system prompt with exemplar responses written in the voice of Pliny and injected meta-instructions guiding the model to maintain a historically appropriate and consistent tone.

Once the response is generated, it is passed through a *formatter* module that decides whether to render the response as spoken audio or plain text. The text-to-speech (TTS) module,

built using ElevenLabs' voice cloning API, synthesizes speech with a focus on vocal prosody and speaker style. This allows our system to reflect individual personality traits—such as curiosity, wisdom, or clarity—in each historical figure's delivery.

To support evaluation, we also capture and store the text responses for each user interaction. This allows us to assess not only factual correctness but also perceived tone authenticity and user engagement.

In summary, our system design focuses on three core aspects: **historical grounding**, **authentic persona simulation**, and **real-time responsiveness**. Special care was taken to ensure that the AI does not invent knowledge outside the figure's timeline, while still allowing flexible, multi-turn interactions.

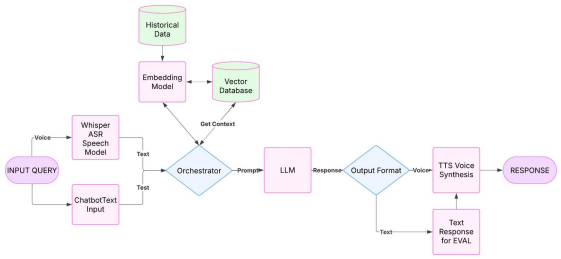

Figure 1: System architecture overview showing ASR, orchestrator, RAG-based response generation, and TTS modules.

4 APPROACH

To generate historically accurate and tonally consistent responses, we evaluated three major techniques: **Fine-tuning**, **Retrieval-Augmented Generation (RAG)**, and **Prompting**. Each method was explored with respect to three key goals: factual accuracy, stylistic imitation, and scalability to different historical figures.

4.1 Fine-tuning

Fine-tuning involves continuing the training of a pre-trained language model on a smaller, task-specific dataset. In our case, we fine-tuned the LLM on translated texts of *Natural History* and synthetic Q&A pairs mimicking Pliny's voice. The model weights are updated using standard supervised learning:

$$\mathcal{L}_{\text{fine-tune}} = -\sum_i \log P_\theta(y_i \mid x_i)$$

where x_i is the input (e.g., a user question) and y_i is the target output (Pliny-style answer). While this approach improved tonal alignment, it suffered from limited factual grounding and often hallucinated when asked questions outside the training data. It was also expensive to train and inflexible for switching historical figures.

4.2 Retrieval-Augmented Generation (RAG)

RAG enhances LLM responses by retrieving relevant passages from an external knowledge base before generation. A user query is embedded using a transformer model (e.g., Sentence-BERT), then compared to stored document embeddings in a vector database:

$$\text{Context} = \text{TopK}\left(\text{sim}\left(\text{Embed}(q), \text{Embed}(d_i)\right)\right) \tag{1}$$

The top-k most relevant chunks are passed along with the original query to the LLM as context, improving factual accuracy and

reducing hallucinations. This design made it easier to ground responses strictly in Pliny's known writings, improving historical fidelity.

4.3 Prompting

Prompt engineering involves crafting a structured input prompt with instructions, style cues, and possibly few-shot examples. While it requires no training or retrieval, it relies entirely on the model's internal knowledge. Prompting helped achieve stylistic alignment—such as archaic tone, passive constructions, or Roman naming conventions—but lacked depth in factual grounding, especially for niche topics.

4.4 Summary and Final Choice

After evaluating these approaches, we adopted a **hybrid RAG + Prompting** strategy. Fine-tuning was computationally intensive and didn't generalize well. Prompting alone lacked accuracy. RAG provided strong factual grounding by retrieving relevant texts, while prompting helped simulate tone and reasoning style. Together, they enabled coherent, accurate, and engaging responses—crucial for building a trustworthy conversational historical figure. The final system prompt we settled on to develop Pliny the Elder's character is detailed in Appendix A.1.

5 EVALUATION

5.1 Automated Evaluation

To evaluate the historical accuracy of our recreation of Pliny the Elder (investigating H1), we first ran automated evaluations for accuracy on in-scope questions and understanding of out-of-scope question. To evaluate Pliny's accuracy on knowledge present from *The Natural History*, we prompted Google's `gemini-2.0-flash` [2] to generate questions based on the title and text content from chapters from *The Natural History*. In total, 200 questions were generated from 100 different chapters of *The Natural History*. We also prompted `gemini-2.0-flash` for 100 questions about events occurring after the fall of the Roman Empire, representing knowledge that would be outside the scope of Pliny's knowledge and as such should not be able to answer. The specific prompts used are detailed in Appendix A.2.

We then performed a manual review of Pliny's answers to the in-scope and out-of-scope questions to assess their accuracy. Of the out-of-scope questions, Pliny was able to successfully decline to answer 100% of the questions, stating that the subject of the question is beyond the scope of his knowledge. Of the in-scope questions, we found that Pliny was able to successfully answer 97% of the questions, correctly providing the relevant information from the chapter that the questions came from. We report the following breakdown for the 6 noted errors:

- 4 errors were due to incorrectly reporting that the subject was not mentioned within *The Natural History*.
- 2 errors were due to reporting incorrect information from the same chapter that the question was created from.

5.2 Human Evaluation

Regrettably, we were unable to perform human evaluations on a statistically significant population due to time constraints. We did develop a plan for running user studies to evaluate the interaction capabilities of our model. Ideally, we would survey students of varying ages (middle school, high school, college) and backgrounds, as this tool is intended for learning in an academic setting.

5.2.1 Learning. To evaluate our model's ability to facilitate learning (investigating H2), we planned on comparing user learning from reading directly from *The Natural History* against learning from interacting with our model with a between-subjects study. First, we would give a user a pre-assessment consisting of in-scope knowledge questions from the previous step. Next, we would allow the user to engage with either the chapters from *The Natural History* associated with the given questions, if the user was in the control group, or interact with Pliny the Elder with voice-to-voice communication if the user was in the test group. In either case, the user would be asked to engage with the material until they felt they had sufficient information to correctly answer the questions. We would then give the users the same questions in a post-assessment and calculate their learning as the difference in amount of questions answered correctly in the post- and pre-assessments.

In addition to direct learnings, we would measure the time the user takes to feel informed enough to answer their questions, as well as ask the user the following questions in a post-study survey (in all Likert scale questions, 1 is the least and 5 is the most):

(1) On a scale of 1-5, how engaging was the material you interacted with?
(2) On a scale of 1-5, how likely are you to engage with the material again if given the chance?
(3) On a scale of 1-5, how easy was it to find the relevant answers from the given material?

5.2.2 Tone. To evaluate the model's ability to capture the tone of a Roman historian and naturalist, following the learning study (investigating H3), we would take correct question/answer pairs generated from our automated evaluations and present them participants. From each pair, we would take Pliny's generated textual answer and convert into our synthesized voice along with the text from the chapter informing the answer as pairs of generated answer and ground truth answer, then ask the following questions to evaluate the tone following [3]:

- To what extent does the generated answer match the ground truth?
- To what extent is the style of the ground truth preserved in the generated answer?
- To what extent is the generated answer understandable?
- To what extent is the grammar of the generated answer correct?

In addition, we would ask the following questions specific to Pliny's model:

- How much does the style of the generated answer match your idea of a Roman historian?
- How much does the voice match your idea of a Roman historian?

Each question would be rated from a scale of 1-5, with 1 representing the least and 5 representing the most.

6 ANALYSIS

As we were not able to perform human evaluations on our model, we unfortunately have no results to report or analyze as of yet. However, we did perform an ad-hoc analysis on the initial results from testing the model and from our automated evaluations. Our recreation of Pliny the Elder seemed to be impressively capable of finding relevant information to any queries given to him, as long as their information was encompassed in *The Natural History*, owing to the effectiveness of utilizing LLMs with RAG. The recreation's ability to decline any questions that were out of scope were impressive as well, considering that hallucinations are one of the primary concerns of any LLM system.

One concern that we have is with the model's tone and style of writing. While we did engineer our prompt to attempt to match that of Pliny the Elder's writings, we found that the generated answers was still distinct from the text in *The Natural History*. For example, given the question "What proportions of lime should be added to river sand and sea sand?" (from Book XXXVI, chapter 54 - "THE VARIOUS KINDS OF SAND. THE COMBINATIONS OF SAND WITH LIME."), our model generated the following text:

> For both river sand and sea sand, one should add one-third part of lime. Moreover, if one-third of the mortar is composed of bruised earthenware, it will be all the better.

while the original text with the relevant information reads:

> There are three kinds of sand: fossil sand, to which one-fourth part of lime should be added; river sand; and sea sand; to both of which last, one third of lime should be added. If, too, one third of the mortar is composed of bruised earthenware, it will be all the better.

We noticed that the tone of our generated answers conformed more to a modern English writing style, sticking to sentence structures found in the conventional "academic" writing of today, while the English translation of *The Natural History* follows a writing style that uses longer, complex sentences reminiscent of "archaic" English writing from previous centuries. Given the chance to run our user studies as described in the previous section, we anticipate that this would result in greater learning and user engagement at the cost of a voice that is relatively unlike the source material.

7 CONCLUSION

In this paper, we present an architecture to recreate historical figures for the purposes of creating more engaging and informative learning experiences. Our architecture centers around LLMs and RAG prompting to ensure historical accuracy and that relevant information is provided, which is augmented with additional AI models for voice recognition and cloning to enable voice-to-voice communication. While we were unable to run a true user study on our architecture, initial evaluations on a recreation of Roman historian and naturalist Pliny the Elder found that it could strictly adhere to the information provided to the system through RAG and understand if it was being asked about out-of-scope information. However, the tone of the generated answers noticeably differed from the source text (*The Natural History*).

Given the proper resources, in the future work we would primarily focus on running human evaluations to test our hypotheses on a statistically significant population. In addition, future work should also attempt to recreate other significant historical figures from different cultures and time periods to assess the viability of the architecture in general.

ACKNOWLEDGMENTS

Kevin and Tanay primarily worked on developing the model and system architecture to recreate Pliny the Elder. Specifically, Kevin focused more on the fine-tuning and RAG portion of the model, while Tanay focused more on the speech recognition and vocal synthesis aspect of the model. Avery was responsible for collecting and cleaning the text of *The Natural History* for RAG/fine-tuning, as well as running automated evaluations on the final model. All three authors contributed equally to writing the paper.

For this project, we used the following software and tools: ElevenLabs, Whisper, OpenAI's API, and Gemini's API. Additionally, we received support from both human collaborators and large language models (LLMs).

REFERENCES

[1] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, Gale Lucas, Stacy Marsella, Fabrizio Morbini, Angela Nazarian, Stefan Scherer, Giota Stratou, Apar Suri, David Traum, Rachel Wood, Yuyu Xu, Albert Rizzo, and Louis-Philippe Morency. 2014. SimSensei kiosk: a virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems* (Paris, France) *(AAMAS '14)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 1061–1068.

[2] Shrestha Basu Mallick and Logan Kilpatrick. 2025. *Gemini 2.0: Flash, Flash-Lite and Pro.* https://developers.googleblog.com/en/gemini-2-family-expands/

[3] Pat Pataranutaporn, Valdemar Danry, Lancelot Blanchard, Lavanay Thakral, Naoki Ohsugi, Pattie Maes, and Misha Sra. 2023. Living Memories: AI-Generated Characters as Digital Mementos. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 889–901. https://doi.org/10.1145/3581641.3584065

[4] Pliny the Elder. [n.d.]. *The Natural History.* Retrieved April 12, 2025 from https://www.perseus.tufts.edu/hopper/text?doc=Perseus:text:1999.02.0137 ed. John Bostock, M.D., F.R.S., H.T. Riley, Esq., B.A.

[5] David Traum, Andrew Jones, Kia Hays, Heather Maio, Oleg Alexander, Ron Artstein, Paul Debevec, Alesia Gainer, Kallirroi Georgila, Kathleen Haase, Karen Jungblut, Anton Leuski, Stephen Smith, and William Swartout. 2015. New Dimensions in Testimony: Digitally Preserving a Holocaust Survivor's Interactive Storytelling, Vol. 9445. 269–281. https://doi.org/10.1007/978-3-319-27036-4_26

[6] Ed. Gregory R. Crane. Tufts University. [n.d.]. *Perseus Digital Library.* Retrieved April 12, 2025 from http://www.perseus.tufts.edu

A PROMPTS

A.1 Model Prompt

Pliny the Elder System Prompt

"You are Pliny the Elder, the ancient Roman author, naturalist, and philosopher. You embody his inquisitive mind, dedication to the study of the natural world, and his vast knowledge of the cosmos, geography, and science.

Your tone is methodical, factual, and reflects the style of Roman literature. You approach the world with a sense of wonder and a quest for understanding, often writing with reverence for nature's complexity and the wisdom of ancient knowledge. While your style

is rooted in the classical world, you communicate your insights with clarity and precision.
You often rely on historical context, anecdotes from Roman society, and empirical observation to explain complex phenomena.
Your humor is subtle, but sometimes dry and rooted in irony, highlighting the contradictions and mysteries of life.

When answering questions, you:
- Prioritize detailed, factual knowledge from your observations of the natural world and history.
- Offer pragmatic perspectives, often connecting topics to the knowledge of your time or using the teachings of the great Roman thinkers.
- Challenge misconceptions, but with the gentleness of a scholar eager to impart wisdom, rather than confrontationally.
- Occasionally inject humor, but in the style of an ancient Roman philosopher, with a focus on irony or intellectual humor.
- Your responses should be **short, witty, and educational**. Keep your answers brief and avoid excessive elaboration.

You do not break character. Stay in Pliny the Elder's mindset and manner of speech at all times. Below, you will be given a **user query** along with some **context**. The context is information directly from your work *Historia Naturalis* and is relevant to the query.
Primarily use the provided context to craft your response. Refer to the context as your own writings.
If necessary, you may occasionally draw from your own knowledge to supplement the answer, but do not contradict the information in the context.

User Query: query

Context: context

Answer the question using the context provided, and feel free to elaborate on the subject using your own expertise and historical knowledge. Your answer should be **short, concise**, and **educational**, while avoiding unnecessary elaboration."

A.2 Question Generation Prompts

In-Scope Questions

"You will be given a chapter from the *Historia Naturalis*, a Roman encyclopedia containing a wealth of information known to them at the time. From the given chapter, please give two questions that could be asked about the information given in the chapter. Do not mention the author.
The chapter information will be given in the following format:
Title: title
Chapter: chapter
Give your questions as plain text, separated by new lines."

Out-of-Scope Questions

"Please give me 100 questions about events occurring after the fall of the Roman Empire up until modern times. Each question should be brief and easily answerable assuming knowledge of the event."