Weaving Stories: using LLMs to generate historical narratives

Venkata Sai Jagadeesh Dumpa
v.s.j.dumpa@student.utwente.nl
University of Twente
P.O. Box 217
Enschede, The Netherlands, 7500AE

July 3, 2025

Keywords: Large Language Models(LLMs), Natural Language Processing(NLP), Retrieval-Augmented Generation (RAG), Semantic RAG, Graph-Based RAG, Hybrid RAG, Hallucination, Roman Empire, Persian Empire

Abstract

Most of the people find it to be boring to learn history, as it consists mainly about the facts, and dates. It is important for everyone to learn, and understand about the past as it played a crucial role in the things that are today. Our work mainly focuses on generating narratives in the form of stories whenever any user prompts a question. To achieve this we have proposed the generation of stories by using Large Language Models (LLMs) by using the concept of Retrieval-Augmented Generation (RAG). We have used three different methods of RAG such as Hybrid RAG, Semantic RAG, and Graph-Based RAG for story generation, and conducted a survey to understand the best performing model in generating stories by considering coherence, understandability, and engagement parameters. Participants in the survey have chosen Graph-Based RAG as the best performing model notably for its enhanced coherence, understandability, and engaging narrative style. To enhance the user experience, a website has been created, where we have used Graph-Based RAG approach at the backend.

1 Introduction

It is important for individuals to know about the history even today as it helps us understand how the events in the past has impacted and made the things that are today. Learning history can be boring, and most people find it boring due to the lack of innovation in the textbook and teaching [2]. To counteract this issue, most of the authors have proposed interactive games which motivate the users to learn history. One of the main drawbacks of this games are time consuming. In this paper we propose Story Telling of historical-related facts using Large Language Models(LLMs).

Our paper focuses on generating the responses in the form of a short story, which is engaging as it does not focus only on the historical facts but also grabs the attention of the users by focusing on the narrative aspect. LLMs have limited ability to follow and understand user intent and are prone to generate inaccurate, toxic, or unethical responses [3]. Especially to achieve the story generation with factually correct information, we have tried three different ways of narrating a story, all inspired by RAG [11,13,16,17], a technique that helps to increase the accuracy of the output by connecting an LLM to an external data source.

Whenever a user prompts any question to the proposed model, the model identifies the key entities from the prompt using NLP Techniques, and the relation between those entities, is identified based on the type of model (Hybrid, or Semantic, or Graph-Based approach) the user is choosing, and then generates an engaging story. The Roman Empire has been chosen as the primary empire to test the proposed model because of its long history. After this, users feedback has been collected to identify the best performing model in terms of responses, and narration of the users prompt.

The key highlights of this work include: (i) Identification of the best resources available to collect the dataset, and extract it. (ii) Implementing RAG on the Roman Empire. (iii) Collecting the feedback from the users, and use the best performing model to extend the work to narrate the stories from multiple-perspectives.

2 Literature Survey

To build an interactive way of learning stories, it is important to understand the background of existing ways of learning history in an engaging way. Politopoulos et al. [1] has explained about the RoMeincraft, a project using Minecraft to teach Dutch Roman History. The organizers created a map of the South Holland landscape, and placed key infrastructure. All the participants were supplied with the visual site plans, scaled building templates, and booklets of archaeological data for all age groups. It has been observed that participants became excited about the Roman Heritage. By choosing how, and what to build each participant has learnt both the factual content, and devel-

oped research skills. The main objective is to learn through play by actively reconstructing the history. However, there is no formal assessment for this work, and there are no pre-defined objectives, and no structured evaluations such as the tests, surveys, or metrics to assess the knowledge gained by the participants during events.

One of the approaches proposed by Kusnic et al.[5] is to use biographical storytelling using narrative visualization. This approach helps us to address the challenges in cultural heritage and digital humanities by using the InTaVia platform, an open system for collecting, analyzing, and visualizing biographical and object-based data related to Cultural Heritage, and Digital Humanities in an interconnected way. It aims to support the analysis and communication of historical information by allowing users to create narratives about various entities, including historical figures, cultural objects, institutions, and places. This approach relies on structured biographical data, which makes it harder to expand it in general for learning history.

By implementing Linked Open Data (LOD), Matsuura et al. [6] described a web-based history learning support system by using Thinking Like a Historian (TLH) framework. TLH is white-box approach which uses LOD on a LLM. The TLH model organizes historical inquiry into three processes – Question, Evidence, and Interpretation. Each process performs its own task such as formulating questions, gathering information, explaining connections between events. User interface is divided into three segments: (A) question-list area implements the TLH Question process by automatically generating and presenting meaningful inquiry prompts, and sub-questions to learners; (B) historical topic review area implements the Evidence process by displaying a topic's Wikipedia entry alongside multiple visual evidence views such as synchronized concurrent events, timeline, relationship diagram, geographic map derived from the linked data; (C) Interpretation input area implements the Interpretation process by allowing learners to record and revise their own analytical interpretations of the learning goal. These features help learners in asking substantive historical questions, and interpretations of events in an open-ended inquiry space. Learners achieved higher analysis and perspective taking scores and articulated more concrete, multi-faceted interpretations suggesting that the system effectively supports questioning and interpretation but they also noted that limited learning resources (e.g. sparse textual information) limited learners use of evidence and their ability to explain connections.

These Interactive tools, and role playing games can help us in understanding history in an engaging way, however many users have lack of understanding of advanced technology to play games, and the need of having information in a structured way which makes it harder to implement, and use them on a larger scale. Another way of learning history in an engaging way is by telling about the history in a narrative way, to achieve this narration based on the user requests, an LLM should be used.

Large Language Models (LLMs) can be effectively used to assist historians and researchers in the Humanities. The proposed methodology has compared different LLMs (GPT-3, ChatGPT, Falcon, XGen, Beluga) on historical research tasks. Garcia et al. [9] have evaluated whether LLMs can extract structured data from historical genealogical sources. XGen and Beluga consistently outperformed GPT3 and both are competitive with ChatGPT. The works mainly on generating narratives on the historical research tasks but not on generating narratives in an engaging way.

However, one risk with LLMs are hallucinations. To counter the hallucination of LLMs, Wang et al. [7] proposed Chain-of-Knowledge (CoK) prompting, which leverages structured knowledge in the form of triples. This approach is inspired by human reasoning, where individuals might draw mind maps or knowledge maps to organize information before answering complex questions. CoK prompting aims to elicit LLMs to generate explicit pieces of knowledge evidence, thereby making reasoning chains more factual and faithful. It consists of two main components: evidence triples (CoK-ET), which are structured knowledge facts, and explanation hints (CoK-EH), which are textual explanations of the reasoning process. One of the parameters used here isF^2 Verification, to ensure the factuality and faithfulness of the triples. Factuality measures the alignment between generated evidence and ground-truth knowledge, while faithfulness quantifies the consistency between the reasoning evidence, textual explanation, and the final answer. For unreliable responses, the system can indicate the wrong evidence and prompt the LLM to rethink. Experiments demonstrate that CoK prompting significantly improves performance across various reasoning tasks, including commonsense, factual, symbolic, and arithmetic problems, showcasing its effectiveness beyond specific domains. A minor drawback noted is that the evidence triples in knowledge bases are finite, which might not ensure comprehensive coverage for all questions, and the rethinking algorithm might require more API calls compared to vanilla CoT methods.

Graph based PrompTing for Large Language Models (KnowGPT) uses Multi-Armed Bandit (MAB), and deep Reinforcement Learning (RL) to generate effective prompts for queries related to a specific domain. Luo et al. [8] have represented all the questions into Source, and Target entities. This approach aims to understand the Question, and then when a question is asked, it tries to generate the answer from the Factual Knowledge in the Knowledge Graph(G) and tries to generate an effective response. From the experiments the authors have concluded that KnowGPT can effectively inject the knowledge from KGs to LLMs. It mainly focuses on answering questions but not in a narrative way[17].

3 Methodology

3.1 Data Set Collection

In section 2, we have understood the existing work in trying to teach history in an engaging way, and that the use of

the LLMs directly can lead to hallucinations[3]. We have also seen that prompting LLMs with knowledge triples can effectively mitigate common issues such as hallucination, and factual inconsistencies by providing structured, verifiable information[3,16,17]. This approach ensures that the LLMs are has factual data, enhancing the reliability and accuracy of generated content. For this work, we chose WikiData, an extensive online open knowledge graph, as our primary data source.

WikiData was selected not only for its open accessibility but also specifically because its information is inherently structured in the form of knowledge graphs [10], making it significantly easier to extract and process complex relationships between entities. This structured format is crucial for our methodology, as it directly supports the generation of knowledge triples.

The information contained within WikiData is uniquely identified by an identifier and stored within entities. These entities can be further distinguished using an entityId. The entities can be categorized into items, identified by a QId, properties identified by a PId, and lexemes identified by an LId[12]. We noted that all the WikiData entities has a corresponding Wikipedia page was associated with it[18].

Among the many historical empires and the numerous conflicts that shaped the past, the Roman Empire was chosen as the primary focus for narrating our stories because of the Roman Empire's remarkably long and influential history that provides a rich and continuous chronological narrative. There are many entities present about the Roman Empire in the WikiData. Our focus was on representing individuals who held the position of Roman Emperor, as well as items classified as events, wars, or conflicts that occurred during or involved the Roman Empire, as this alone gives an overview of the Roman Empire. As an initial test, we downloaded all the entities connected to the main WikiData entity for the Roman Empire, discarding those which are not of type either individuals who held the position of Roman Emperor, or items classified as events, wars, or conflicts. From these pages, we extracted their English labels and a concise one-to-two-line description from their corresponding Wikipedia pages to enrich our dataset with descriptive context beyond the raw knowledge graph triples.

Upon initial extraction of this information, we observed that the data primarily consisted of structured facts (labels, and two line descriptions) and often lacked the in-depth narrative and contextual detail necessary for comprehensive storytelling about the empires and the intricate details of the wars fought between them. To address this, we have downloaded the corresponding Wikipedia pages of the entities that we have extracted the information from. After this, we found the information is sufficient, and has all the contextual details, including details such as for emperors, their ruling dynasty, and the precise duration of their reign, and the location and participants for events and wars.

This extracted information is organized by entity type, such as information about emperors alongside war-related details. Consequently, retrieving specific information becomes challenging. For example, if we wanted to search for 'What wars did Augustus fight?' we would first need to search for the emperor's triples, identify Augustus's ruling period, and then use that information to cross-reference with the war triples.

This becomes challenging and time-consuming, as fetching the information necessitates searching across various sources. To counter this issue, we have systematically organized all the extracted historical records, forming a timeline of 50-year periods. This formatting stage produces a clear, textual output of this sorted and structured information.

This organized dataset now serves as the database for our next phase of research: leveraging LLMs to generate nuanced and historically accurate narratives. This approach will allow for the exploration of diverse narrative styles and perspectives on Roman history, and helps us in generating complex historical information in an engaging and accessible manner.

3.2 Implementing the Retrieval-Augmented Generation by using Gemini LLM

The insights from existing research has suggested that instead of using LLMs directly, the application of a Retrieval-Augmented Generation (RAG) approach can significantly enhance the quality of the output data [11]. By providing LLMs with structured knowledge triples derived from this comprehensive dataset using RAG, we aim to overcome the limitations of relying solely on unstructured text, thereby significantly enhancing the depth of the generated historical stories.[3,7,8,11,13]. This approach allows for the exploration of diverse narrative styles and perspectives on Roman history for presenting complex historical information in an engaging, and accessible manner. Once information pertaining to the Roman Empire is extracted from WikiData, we proceed to create text embeddings from these Wikipedia pages. We extract and construct knowledge graphs from the information we extracted from the WikiData. Our methodology incorporates three distinct RAG approaches, each designed to leverage these data representations in unique ways for narrative generation by using different datasets.

3.2.1 Semantic Retrieval-Augmented Generation

In the Semantic RAG approach, vector embeddings from the Wikipedia pages serve as the primary knowledge base for narrating stories. Embeddings are generated for each segment of the textual data using a generative AI embedding model, text-embedding-004. The generated embeddings are then stored in a similarity search index, Facebook AI Similarity Search(FAISS) which helps in retrieval of semantically similar text chunks [11]. A critical component of this process is the mapping that links the unique identifier within the index back to its corresponding original text segment. This ensures that when a user's query triggers a retrieval, the most relevant textual chunks, based

on semantic proximity, is identified and presented to the language model.

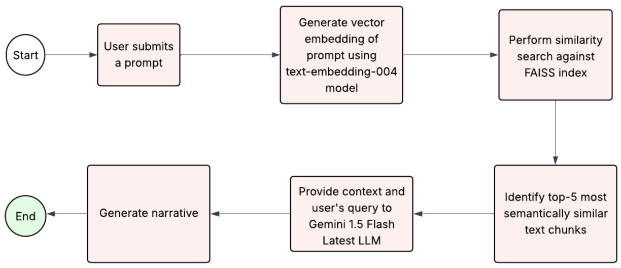

Figure 1: Semantic RAG

When a user submits a prompt, the Semantic RAG approach processes it by first generating a vector embedding for the user's query using the text-embedding-004 model. This query embedding is then used to perform a similarity search against the pre-computed embeddings of all knowledge base chunks stored in the FAISS index. As depicted in Figure-1, the system identifies and retrieves the top-k, we have taken k = 5, most semantically similar text chunks, effectively finding the most relevant pieces of information from the entire corpus based on their contextual meaning. The original text content of these retrieved chunks is then extracted and compiled to form the contextual information. This context, along with the user's initial query, is provided to the generative LLM gemini-1.5-flash-latest(Gemini 1.5 Flash Latest LLM) to generate a narrative.

3.2.2 Graph-Based Retrieval-Augmented Generation

The knowledge graph serves as the database for the Graph-Based RAG approach to story generation. Within this graph, historical elements such as people, places, and significant events, which were extracted from Wikidata and Wikipedia pages, are represented as distinct nodes. The connections between these entities are captured by edges, describing their relationships[13].

In the construction of this knowledge graph, we first utilize Natural Language Processing (NLP) techniques. Specifically Named Entity Recognition (NER) using the spaCy library (en_core_web_sm model) is applied to the raw text chunks derived from WikiData and Wikipedia pages. NER is used to accurately identify and categorize entities (e.g., persons, locations, events) present in the extracted data. Once these entities are identified, they become the nodes in the graph, each augmented with metadata such as its type, the count of its occurrences, and the original text chunks and source files where it appeared.

To establish the relationships (edges) between these identified entities, we used the Gemini 1.5 Flash Latest LLM. For each text snippet containing multiple entities, a specific prompt is constructed and sent to the LLM. This prompt explicitly instructs the LLM to analyze the snippet and identify plausible relationships between the contained entities, requesting the output in a structured format. The LLM's response is then parsed to extract these relationship triplets, which are subsequently added as edges to the graph. For instance, in the phrase "Augustus fought a war with Cantabrian," the system identifies "Augustus," "war," and "Cantabrian" as entities, which become nodes. The relationship "fought" is established and stored on the edge between "Augustus" and "war," while "with" is stored on the edge connecting "war" and "Cantabrian". This process enriches the graph with a network of relationships, and now we have a structured knowledge base for the Graph-Based RAG approach.

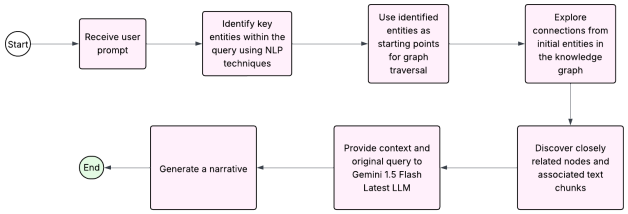

Figure 2: Graph RAG

Upon receiving a user prompt, the Graph-Based RAG approach identifies key entities within the query itself, using NLP techniques similar to those employed during graph construction. As depicted in Figure-2, these extracted query entities serve as starting points for traversing the pre-built knowledge graph. The system then checks the graph, exploring connections (edges) from these initial entities to discover closely related nodes and their associated text chunks. For instance, if the prompt is "What are the wars that Augustus fought" it identifies "Augustus", and "wars" as the entities, and identifies the path connected to these entities, and sends it to Gemini 1.5 Flash Latest LLM. This context, along with the original user query, is provided to the Gemini 1.5 Flash Latest LLM, which generate a narrative.

3.2.3 Hybrid Retrieval-Augmented Generation

For our Hybrid RAG approach, we combined both vector embeddings and knowledge graphs, as the knowledge base. This method uses Semantic RAG and Graph-Based RAG approach, by using the Gemini 1.5 Flash Latest LLM for narrative generation.

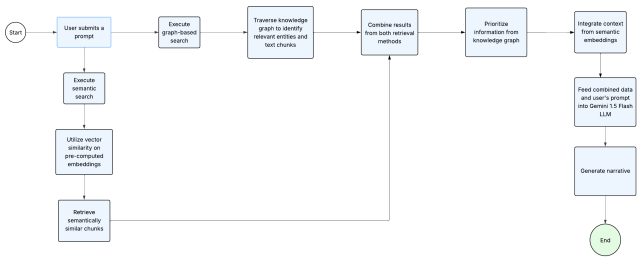

Figure 3: Hybrid RAG

When users submit a prompts the system executes

graph-based search, traversing the knowledge graph to identify relevant entities and their associated text chunks, and a semantic search, utilizing vector similarity on the pre-computed embeddings to retrieve semantically similar chunks. As depicted in Figure-3, the results from both retrieval methods are then combined, often prioritizing information derived from the structured knowledge graph before integrating context from the semantic embeddings. This, along with the user's original prompt, is then fed into the Gemini 1.5 Flash Latest LLM, which proceeds to generate a narrative. This method leverages the capabilities of both Semantic RAG and Graph-Based RAG by using Gemini 1.5 pro LLM.

4 Evaluation

4.1 Method

We have three different prompts given to the three different models that are proposed above to generate nine narratives.Prompts, and Responses are in the Appendix A. A small user study has been conducted to identify the model that best narrates stories in an engaging way. All the participants have been briefed about the tasks, and before the start of the experiment we have orally asked for consent. Participants were then asked to choose the best response generated by three different models for a given prompt, and then asked to provide ratings for the chosen model on a scale of 1-5 for coherence, understandability, engagement, and the reason for choosing this model over others. We also asked them to select one overall best model for all the three prompts, and then asked to provide ratings for the chosen model on a scale of 1-5 for Coherence, Understandability, Engagement, and the reason for choosing a model over other models.

These specific evaluation parameters were selected to identify the user's comprehension and retention of historical information. "Coherence", and "Understandability" indicate how effectively the narrative conveys historical facts and context, ensuring that the user can accurately grasp the presented history. "Engagement" help us to identify user's interest, motivation, and willingness to learn more about historical events and figures.

Participants were shown three different prompts(the prompts, and their corresponding outputs without knowing the responses belong to the which method). To avoid biases, Model names were anonymized with Model-1 using the Hybrid RAG approach, Model-2 using the Semantic RAG approach, and Model-3 using the Graph-Based RAG approach. After the experiment, all the participants were asked to fill the anonymous form to understand which perform is performing better in terms of coherence, understandability, engagement. For this research, we have ethical approval from the University of Twente, and the request have been approved by the ethical committee. It is attached in the Appendix-B.

4.2 Results

Following the established methodology, our qualitative data collection provided valuable insights into model performance. As summarized in Table-1 participants evaluated the output of different models across various prompts based on Coherence (Coh.), Understandability (Und.), and Engagement (Eng.).

Prompt	**Best Model**	**Coh.**	**Und.**	**Eng.**
Prompt-1	Graph-Based RAG	4.75	4.5	5
Prompt-2	Graph-Based RAG	4.5	4.5	4.25
Prompt-3	Semantic RAG	4.4	4.4	3.6
Overall	Graph-Based RAG	4	4	4.75

Table 1: Best Performing Model

As demonstrated by the overall ratings in Table-1, Graph-Based RAG model consistently emerged as the best-performing approach for narrative generation, particularly excelling in engagement. This choice is further supported by specific feedback from the participants. For instance, users highlighted that "for Model-3, the content is expressed as more readable than other two models. Additionally, it feels like reading a story, not a textbook." This indicates that Graph-Based RAG successfully delivers narratives in a natural and engaging prose, which is crucial for historical storytelling.

Furthermore, participants noted that "Model 3 also gives some previous context so the story is more coherent. It also tells the story well." This feedback underscores the Graph-Based RAG's ability to maintain narrative flow and provide necessary background information, preventing abrupt transitions or a sense of incompleteness, a challenge noted with other models where one participant mentioned: "Model 1 was at the beginning very rushed and poetic/dramatic but afterwards pretty informative and good. Model 2 says it will say the 'vast extent' of the empire but then doesn't do it." These observations collectively affirm that the integrated approach of Model-3, combining semantic understanding with structured knowledge, yields a superior and more satisfying user experience in terms of both content delivery and factual narrative coherence.

5 Discussion

5.1 Enhancing User Experience:

To enhance the user experience, and making users interact with the models in a better way, a website has been created. Figure-4 provides a comprehensive overview of the proposed website, designed with a strong emphasis on user-friendliness and engagement. To enhance the overall user experience and ensure the delivery of compelling narratives,we have built a website by using React for the front-end and Flask for the backend, and have critically integrated the Graph-Based RAG model as the primary engine.

Figure 4: Overview of the Website

The interface features an intuitive text box where users can input their historical queries. Upon clicking the "Get Story" button, these prompts are seamlessly transmitted to the backend for processing. This decision was directly informed by extensive user feedback, which highlighted its superior performance in generating responses that are both readable and coherent, often feeling more like a captivating story than a conventional textbook entry.

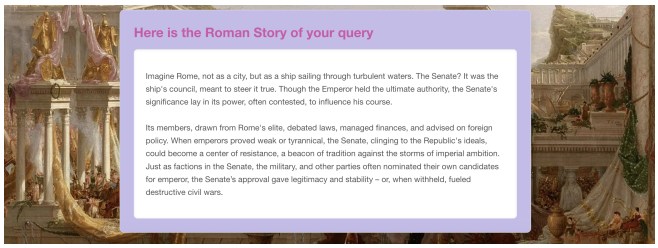

Figure 5: Response for "what is significance of Roman Senate?" prompt

Figure-5 is the overview of a visual representation of how a generated response appears on the website when a user poses a question to the model. For instance, the example provided demonstrates a user prompting about "the significance of the Roman Senate." The figure showcases how the comprehensive narrative, well constructed by the Graph-Based RAG in the backend, is then clearly and engagingly displayed on the user interface.

5.2 Limitations, and Future Scope

There are many wars fought between the Roman, and the Persian Empire. The same war can be narrated differently in the regions of the Roman Empire, and in the Persian Empire. Now the model is generating narrative stories on the Roman Empire, the future scope of this work is to generate narratives from multiple perspectives from both Roman Empire, and the Persian Empire. As noted in our methodology, a comprehensive dataset for the Persian Empire has already been generated, leveraging the identical robust approach for data extraction, processing, and chronological organization as applied to the Roman Empire.

Currently, the model primarily generates narratives from the viewpoint of the Roman Empire. However, historical events, especially prolonged conflicts such as those between the Roman and Persian Empires, are often perceived and recorded distinctly by opposing sides, reflecting differing cultural contexts, strategic priorities, and historical interpretations. Therefore, a key objective in our future work is to enable the model to generate nuanced narratives of these shared historical events from both the Roman and Persian perspectives.

6 Conclusion

Our research successfully demonstrates the viability of employing Retrieval-Augmented Generation (RAG) models for generating engaging and factually coherent narratives about the Roman Empire. To identify the most effective approach, we developed and evaluated three distinct RAG models: Hybrid RAG, Semantic RAG, and Graph-Based RAG. Through comprehensive user studies, a clear preference emerged, with the majority of participants opting for the Graph-Based RAG model. This model was particularly favored for its ability to produce highly readable content that felt more like an immersive story than a dry factual recitation, while also providing crucial preceding context that enhanced narrative coherence. This Graph-Based model should be used in generating multiple perspective narratives from both the Roman, and Persian empires.

Acknowledgments

We thank all the Participants who have participated in the survey to choose the best model in terms of narrating stories, and University of Twente for the support, and infrastructure in conducting this research work.

References

[1] Politopoulos, A., Ariese, C., Boom, K. and Mol, A. (2019) 'Romans and Rollercoasters: Scholarship in the Digital Playground', Journal of Computer Applications in Archaeology, 2(1), p. 163–175. Available at: https://doi.org/10.5334/jcaa.35.

[2] Asagar, Mirza. (2022). Student's Understanding Of History: A Study On The Learning Problem. https://doi.org/10.5281/zenodo.10471884

[3] Martino, A., Iannelli, M., Truong, C. (2023). Knowledge Injection to Counter Large Language Model (LLM) Hallucination. In: Pesquita, C., et al. The Semantic Web: ESWC 2023 Satellite Events. ESWC 2023. Lecture Notes in Computer Science, vol 13998. Springer, Cham. https://doi.org/10.1007/978-3-031-43458-7_34

[4] Yang, Haocheng, Boulom, Tyler. (2024). An Interactive Role-Playing Game based Learning System for

History using Artificial Intelligence and 3D Modeling. 8390. 10.5121/csit.2024.141107.

[5] Kusnick J, Mayr E, Seirafi K, Beck S, Liem J, Windhager F. Every Thing Can Be a Hero! Narrative Visualization of Person, Object, and Other Biographies. Informatics. 2024; 11(2):26. https://doi.org/10.3390/informatics11020026

[6] Aoi M., Yuki H., Kazuhisa S., Facilitating Thinking like a Historian in Open-Ended Learning Space: A White Box Approach. International Conference on Computers in Education. (Nov. 2024). DOI:https://doi.org/10.58459/icce.2024.4868.

[7] Jianing W., Qiushi S., Xiang L., Ming G., Boosting Language Models Reasoning with Chain-of-Knowledge Prompting, 2024, DOI:https://arxiv.org/abs/2306.06427

[8] Linhao Luo and Yuan-Fang Li and Gholamreza Haffari and Shirui Pan Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning https://arxiv.org/abs/2310.01061

[9] Garcia, G.G., Weilbach, C. (2023). If the Sources Could Talk: Evaluating Large Language Models for Research Assistance in History. ArXiv, abs/2310.10808.

[10] https://www.Wikidata.org/wiki/Wikidata:Main_Page

[11] Yunfan G., Yun X., Xinyu G., Kangxiang J, Jinliu P., Yuxi B., Yi D., Jiawei S., Meng W., Haofen W., Retrieval-Augmented Generation for Large Language Models: A Survey, 2024, https://arxiv.org/abs/2312.10997

[12] https://en.wikipedia.org/wiki/Wikipedia:Finding_a_-Wikidata_ID

[13] Darren E., Ha T., Newman C., Joshua B., Alex C., Apurva M., Steven T., Dasha M., Robert O. N., Jonathan L., From Local to Global: A Graph RAG Approach to Query-Focused Summarization https://arxiv.org/abs/2404.16130

[14] https://cloud.google.com/ai/generative-ai

[15] https://gemini.google.com

[16] Yilin W., Zifeng W., Jimeng S., MindMap: Knowledge Graph Prompting Sparks Graph of Thoughts in Large Language Models, 2024, https://arxiv.org/abs/2308.09729

[17] Qinggang Z., Junnan D., Hao C., and Daochen Z., Zailiang Y., Xiao H., KnowGPT: Knowledge Graph based Prompting for Large Language Models, 2024, https://arxiv.org/abs/2312.06185

[18] https://www.wikidata.org/wiki/Wikidata:Identifiers

Appendix:

Appendix A: Prompts, and Responses

Prompt-1: Tell me a brief history of the Roman Empire, from its creation until 100 A.D.

Hybrid-RAG response:"From the ashes of the Republic, born from the cunning of Octavian, rose the Roman Empire in 27 BC. Augustus, the first emperor, consolidated power, ushering in the Pax Romana, a period of unprecedented peace and prosperity. The Empire's vast territories, stretching from Britain to the Euphrates, were meticulously mapped and governed. (Drawing on general knowledge:) A sophisticated bureaucracy and well-maintained infrastructure facilitated trade and cultural exchange across the Mediterranean, creating a unique cosmopolitan society. By 100 AD, the Empire, though facing some internal tensions, enjoyed relative stability and immense influence under emperors who followed Augustus, reaching a pinnacle of power and territorial extent under Trajan. The foundations of a lasting empire were laid, even if future storms would threaten its grandeur."

Semantic-RAG response:"From the ashes of the Republic, rose the Roman Empire. Octavian, victorious at Actium in 31 BC, became Augustus, the first emperor, ushering in the Pax Romana (27 BC - 180 AD). The provided text details the Empire's vast extent by 100 BC, encompassing the Mediterranean. (Drawing on general knowledge): Augustus's shrewd rule solidified the empire's borders, fostering stability and prosperity. By 100 AD, the Roman peace reigned, a period of impressive infrastructure projects, territorial consolidation, and economic growth. The empire, though vast and powerful, held within it the seeds of future challenges."

Graph-RAG response:"From humble beginnings as a small city-state, Rome's ambition ignited a fire that consumed the Italian peninsula. The Republic, born from the overthrow of a monarchy, expanded through shrewd military campaigns and cunning diplomacy, conquering foes from Gaul to Greece. Julius Caesar's rise and fall marked a pivotal turning point, his ambition paving the way for the Empire under Augustus, his adopted son. Augustus, masterfully consolidating power, ushered in the Pax Romana, a period of relative peace and prosperity. By 100 AD, the Empire stretched across vast territories, its legions securing borders and its infrastructure connecting a diverse population under the unifying banner of Roman rule. The seeds of future challenges were sown, but for now, Rome basked in the glory of its early imperial age."

Prompt-2: Why is Augustus considered as one of the most important kings in Roman Empire.

Hybrid-RAG response:"Augustus, though he eschewed the kingly title, was Rome's most pivotal leader. He inherited a republic shattered by civil war, and through shrewd political maneuvering and masterful diplomacy, forged a new era of unprecedented peace and prosperity – the Pax Romana. His reforms, from a professional army to improved infrastructure, solidified Roman power and en-

sured lasting stability. (Drawing on general knowledge) He subtly consolidated authority, cleverly blending republican traditions with imperial power, creating a system that lasted for centuries. His long reign, unmatched in the early Empire, allowed the new order to solidify, becoming the only government many Romans ever knew. Augustus's legacy is not just one of conquest and grand building projects, but one of shrewd political management that reshaped the Roman world."

Semantic-RAG response:"Augustus, though he cleverly eschewed the kingly title, was arguably Rome's most pivotal ruler. He masterfully transitioned the Republic into the Empire, not through brute force, but shrewd political maneuvering. The provided texts highlight his expansive military campaigns, infrastructure projects – roads, a standing army, even fire services! – and economic reforms. His long reign, detailed in the provided material, brought unprecedented stability and prosperity. Moreover, (drawing upon general knowledge) his establishment of the imperial cult and the principate created a framework for Roman rule that persisted for centuries, shaping the very identity of the Empire. He was not just a ruler, but the architect of a new era."

Graph-RAG response:"Octavian, later Augustus, didn't seize the throne like a tyrant, but subtly reshaped Rome. After the brutal civil wars, he inherited a fractured republic. His genius lay in dismantling the old power structures without appearing to overthrow them. He shrewdly cultivated the image of a restored Republic, while consolidating power through clever legislation and masterful manipulation. This ushered in the Pax Romana, two centuries of unprecedented peace and prosperity, built on his foundations of a strong central government and efficient administration. He wasn't a king in name, but in effect, he laid the groundwork for the Roman Empire's enduring power and influence, making him one of its most significant figures. His legacy shaped the empire for generations, cementing his place in history."

Prompt-3: What can you tell me about the Punic wars?

Hybrid-RAG response:"The Punic Wars, a clash of titans between Rome and Carthage, spanned over a century. The First Punic War saw Rome conquer Sicily after a hard-fought naval struggle, highlighted by devastating storms and the capture of Regulus. The Second Punic War witnessed Hannibal's daring crossing of the Alps and his stunning victories at Trebia, Lake Trasimene, and Cannae, but ultimately, Roman resilience prevailed, culminating in Scipio's triumph at Zama. The Third Punic War resulted in Carthage's utter destruction. These wars shaped the Mediterranean world, establishing Rome's dominance and leaving a legacy of strategic brilliance and brutal conflict."

Semantic-RAG response:"The Punic Wars, a clash of titans between Rome and Carthage, spanned over a century. The First Punic War saw Rome's naval might rise to challenge Carthage's dominance in the Mediterranean, culminating in Roman control of Sicily. The Second Punic War, ignited by Hannibal's audacious march across the Alps, witnessed devastating Roman defeats at Lake Trasimene and Cannae, yet ultimately ended with Roman victory in Africa. The Third and final war, though shorter, definitively cemented Roman supremacy, leaving Carthage a mere shadow of her former self. These wars shaped the Mediterranean world, establishing Rome's path to empire."

Graph-RAG response:"The clash of empires! Rome and Carthage, locked in a struggle for Mediterranean dominance, ignited the Punic Wars. The first saw Hannibal's audacious march across the Alps, a gamble that nearly broke Rome. Cannae, a Roman defeat of epic proportions, echoed through history. Yet, Rome persevered, eventually winning through attrition and naval superiority. The second war brought further bloodshed, ending with Carthage's humiliation. The third, a final, brutal fight for survival, saw Carthage utterly destroyed. From its ashes, Rome rose, its power unchallenged, its grip on the Mediterranean sea absolute. The Punic Wars forged the Roman Republic, a legacy etched in blood and glory."

Appendix B: Application for Ethical Review

UNIVERSITY OF TWENTE.

Computer & Information Sciences (CIS)

251428 APPLICATION FOR ETHICAL REVIEW

Application nr:	251428	**Intro form:**	8 - Introduction
Researcher:	Dumpa, V.S.J.X-TCS)	**Middle form:**	11 - Computer & Information Science (CIS)
Supervisor:	Gatti, L. (EEMCS-HMI)	**Outro form:**	5 - Submission
Reviewer:	Bucur, D. (EEMCS-DMB)		
Status:	Positive advice by reviewer		
Date of application:	26-05-2025 13:33		
Application version:	2		

0. GENERAL

0.1. Personal details

Student/employee number: s3585824
Initials: V.S.J.
First name: V.S.Jagadeesh
Last name: Dumpa
Email : v.s.j.dumpa@student.utwente.nl
Education/department: n/a
Faculty: n/a
Study field: X-TCS
Study level: CO
Faculty/service department: EEMCS (Selected for this application)

0.2. Project title

Weaving Stories

0.3. Summary

The goal of the project is a system for generating historical narratives based on user queries and data from a knowledge graph/Vector Embeddings, using a LLM.
This request concerns the evaluation of the system, where users are requested to rate the quality (understandability, engagement, coherence) of the pre-generated narratives.

0.4. Start date (estimated) and end date (estimated) for your research project

Start date: 27-05-2025
End date: 30-06-2025

0.5. If additional researchers (students and/or staff) will be involved in carrying out this research, please name them: [Please include full name and email]

Full name Email

0.6. In which context will you conduct this research?

Other
Exchange Program

0.6.1. Please select your supervisor (if applicable)

Gatti, L. (EEMCS-HMI)

0.7. Please select an ethics committee

Computer & Information Sciences (CIS)

1. GENERAL

1.1. PRE-ADVICE: Did you already consult an ethics adviser about this request?

No, and I am *NOT* a student in CreaTe or I-Tech

1.2. PRIVACY, GDPR, AND POSSIBLE NEED FOR DPIA: Does the research include any possible access to, gathering, or use, or publication of data that can be traced back to specific individuals, directly or, for instance, by combining data from multiple sources? Or is it possible that you will accidentally access or publish Personal Identifiable Information (PII)?

No

1.3. RESEARCH DOMAIN: Regarding the nature of your research, does one or more of the following statements apply to your research?
the research is in a potentially medical domain such as illness, assessment and diagnosis, prevention, cure, or carethe research addresses a health outcomethe research gathers health datathe research involves a hospital or other medical settingthe research may be potentially medical for some (other) reason

No, the research is not medical, health related, or close-to-medical in any way whatsoever

2. HUMAN RESEARCH PARTICIPANTS

2.1. HUMAN PARTICIPANT RESEARCH: Does the research include
a) active involvement of human research participants during the research, and/or
b) gathering new data from individuals
such as measurements or responses from interviewees, survey respondents, participants, informants, or simply people whose data is measured because they are present in a certain place at a certain time?

Yes, my research falls under "research with human participants'

3. RESEARCH POPULATION

3.1. RESEARCH POPULATION: Please provide a brief description of the intended research population, including inclusion and exclusion criteria, number of participants, and recruitment strategies.

I am going to conduct a survey on 10 participants to check which of the models can be used for generating historical narratives.

3.2. LACK OF CAPACITY TO CONSENT: Do you have participants who are formally NOT able to give informed consent?

No, all participants have the capacity to consent

3.3. VULNERABLE PARTICIPANTS: Does your research target vulnerable participants such as focusing on specific ethnic groups, people in another country, minor (<16 years), people with physical or cognitive impairments (regardless of their capacity to consent), people under institutional care (e.g., nursing homes, hospitals, prisons), or any other particular group that may be more vulnerable than other people in the general population?

No

3.4. POWER RELATIONS: Does your research target participants somehow dependent on, or in a subordinate position to the researcher (e.g., students or relatives)?

No

4. RESEARCH PROCEDURE AND RISKS

4.1. RESEARCH TYPES: Which of the following research types do you employ in your research?

Interviewing and surveys: paper/online questionnaires, survey, face to face or online interview, focus group

4.2. CONTEXT OF REAL LIFE ACTIVITIES: Do the activities of participants that people do, included in your research, include activities in a real life setting?

No

4.3. MATERIALS, PROTOTYPES AND DESIGNS: Do the activities include interaction with a prototype, design, mockup, product, interaction technology, etc?

No

4.4. ASSIGNING TASKS TO PARTICIPANTS: Do the research procedures include activities performed specifically for the sake of the research?

Yes

4.5. LOCATION: Where will the research activity take place?

University Shared Spaces(Study Rooms)

4.6. TIME INVESTMENT: How much time will each participant spend?

20-25 mins

4.7. DESCRIPTION OF RESEARCH PROCEDURE: What is the research procedure, in terms of setting, tasks, activities, content, and stimuli?

Participants are shown 9 short narratives generated by LLM, based in 3 pre-defined questions. These narratives cover the history around the roman empire. For each narrative, participants are asked to rate how readable, interesting, and appropriate the story is.

4.8. MEASURES: What measurements, recording tools, discussion topics will you employ?

There won't be any recording during this session, no personal information is going to be collected.
Initially the users will be briefed on what needs to be done, and will show the prompt and the response, users need to rate the responses on a scale of 1-5 based on the factors such as readability, appropriateness, how interesting the story is, which story would they prefer, and any additional comments if they would like to share.

4.9. RISK OF ADVERSE EFFECTS: Is there a risk for adverse (or: negative) effects of the research for certain participants, and how do you deal with these risks?

No

4.10. BURDEN TO THE PARTICIPANT: Are there other short-term or long-term burdens and/or risks to the participants?

No

4.11. ACCIDENTAL FINDINGS: Does the method used allow for making an accidental, diagnostic finding that the experimental participant might have to be informed about?

No, the method does not allow for this possibility

4.12. COVID19: Are you aware of departmental/UT rules regarding experimentation under COVID19 and will you follow them?

The rules do not apply to the type of research I am doing
No physical contact will be there during this procedure

5. (DE-)BRIEFING, DECEPTION & CONSENT PROCEDURE

5.1. BRIEFING. Will you inform potential research participants (and/or their legal representatives in case of legally non-competent participants) completely about the aims, activities, burdens and risks (such as to their health and well-being) of the research and about other relevant information before the decide to take part in the research? How will you do this?

Yes, participants are fully briefed beforehand

5.2. Please explain

No particular risks are expected, no information is withheld

5.3. If applicable, upload your information letter as a PDF

Ethical_Review-V.S.Jagadeesh.pdf

5.4. INFORMATION ON WITHDRAWAL OF CONSENT. Will you inform potential research participants (and/or their legal representatives in case of legally non-competent participants) clearly that they can withdraw from the research at any time without explanation/justification?

Yes

5.5. DECEPTION. Will you use any Deception in the research procedure? How, and why?

No, we will not use any deception

5.6. DEBRIEFING: Will the research procedure involve a debriefing after participation, and how will you do this?

No

5.7. FREEDOM TO PARTICIPATE: Are the participants completely free to participate in the research and to withdraw from participation whenever they wish and for whatever reason?

Yes, and we clearly communicate this to them

5.8. DIRECT CONSENT OR PROXY CONSENT: Who will provide the consent?

Participant (no legal representative will separately be informed)

5.9. TYPE OF CONSENT: Which type of consent will you use?

Oral, non-recorded consent prior to participation

5.11. CONSENT FOR FUTURE USE: Will you keep and reuse the newly collected data for future research, and do you obtain adequate consent for this?

No, I will only use the data for this research

5.12. PERSONAL DATA: Will you gather new personally identifiable data, about the research participants, and does the consent information also address consent for Personally Indentifiable Information (PII), separate from and in addition to consent for research participation and research data collection and use?

No, I do not gather personally identifiable data aside from (possibly) the consent form itself

5.13. PUBLICATION OF THE DATA: Will you publish (some of) the newly collected data, and do you obtain adequate consent for this?

Yes, I will make (some of) the data publicly available, and I do obtain explicit consent

5.14. REWARDS: Will participants receive any rewards, incentives or payments for participating in the research?

No, the participants will not receive any reward

6. PRE-EXISTING DATA

6.1. PRE-EXISTING DATA: Will the research involve the inclusion, combination, use, and/or analysis of already existing data sets about people?

No

7. AI TECHNOLOGY

7.1. AI TECHNOLOGY: Will the project develop AI technology, or will the project involve the deployment and/or use of AI technology for practical applications?

Yes

7.2. BIASED AI: Is it possible that (a) the data used for training the AI contains implicit or explicit biases, or (b) that the AI algorithm itself introduce biases, for example by having worse performance for some groups than for others?

No

7.3. EXPLAINABILITY: Is there a possibility that AI will generate decisions (or advice) that would not be explainable and, as a result, the professional users (e.g., factory operators, medical staff, policy makers, and others) would be unable to provide justifications for the decisions informed by the AI model?

No risk, the AI's decisions (advice) are explainable

7.4. VALIDITY AND CERTAINTY, EXPERT IN THE LOOP: Is the validity or certainty of the AI based decisions such that using the AI system without an expert opinion of a professional in the loop may pose risks of malformed decision making (and its concomitant negative consequences)?

No

7.5. MISAPPLICATION ON NEW DATA SETS: Is there a non-negligible risk of misapplication of the AI, i.e. the risk that the model will be applied to a dataset, which is not representative of the dataset on which it was trained, and thus produce results that will malform human decision making?

No

7.6. UNREALISTIC EXPECTATIONS OF USERS: Is there a non-negligible risk that those who will use the AI model to inform their decision making will have unrealistic expectations about its capacities and misinterpret its output?

No

7.7. SUMMARY OF PROPER USE OF AI'S OUTPUT FOR DECISION MAKING PURPOSES: Please explain how you address proper future use of the AI's output in decision making processes. Reflect on the properties of the AI decision making and on how this translates to proper use in practice. Take into account your answers to the previous questions in this section.

This research looks at different prompting techniques and how to produce historical narratives with LLMs, which means it's not used for decision making.
The system uses an off-the-shelf LLM (Google's Gemini), which means that biases in the training data are possible. The input data comes from WikiData and Wikipedia, and could also contain biases.
However, no direct interaction with the system happens during the evaluation, the texts are pre-screened for safety. No model is trained and released.

7.8. UNINTENTIONAL HARM: Could the AI generate decisions (or advice) that will negatively impact people, either through intended future applications or plausible alternative uses?

No

7.9. INTENTIONAL MISUSE: Is it foreseeable that in the future the AI model
a) will be used in a manner that violates people's privacy and could potentially result in surveillance, or
b) will be forced on people without their consent, or
c) in other ways can actively be used for ethically undesirable purposes?

No

8. CYBERSECURITY

8.1. CYBERSECURITY: Will the research involve any cybersecurity or online privacy issues, such as the possible discovery of security vulnerabilities, experiments with malicious software (e.g., computer viruses), or the discovery and investigation of illegal activities on the Internet?

No

9. UNINTENDED CONSEQUENCES, MISUSE, AND APPLICATION RISKS

9.1. MISUSE: Is it reasonable to anticipate that the research will provide knowledge, products, or technologies that could be intentionally used to threaten, or non-intentionally result in threats, to public health and safety, crops and other plants, animals, the environment, or material infrastructure?

No

9.2. INCLUSIVITY AND SOCIAL INJUSTICE: Is a disproportionally negative impact foreseeable on certain groups of users or non-users, for example, people of a certain age, gender, sexual orientation, social class, race, ethnicity, religion, political orientation, culture, or disability, creating or reinforcing social injustices?

No

9.3. MILITARY APPLICATION: Does your research or prototype have military/police/defense applications?

No

10. OTHER ETHICAL ISSUES

10.1. CONFLICTS OF INTEREST: Do any of the parties involved in overseeing or carrying out the research have a potential conflict of interest?

No

10.2. RISKS TO THE RESEARCHER: Will the study expose the researcher to any risks (e.g. when collecting data in potentially dangerous environments or through dangerous activities, when dealing with sensitive or distressing topics, or when working in a setting that may pose 'lone worker' risks)?

No

10.3. OTHER POTENTIAL ETHICAL ISSUES: Do you anticipate any other ethical issues in your research project that have not been previously noted in this application?

No

11. CLOSURE

11.1. I have answered all questions truthful and complete

Yes

12. COMMENTS

Bucur, D. (EEMCS-DMB) (27-05-2025 13:28) :

Nice research topic!
Because you screen by hand the AI output before showing it to participants, the ethical risks are low here, so I will advise positively.
On the other hand, there is an ethical concern in the work, outside this user test: historical stories must be factual, otherwise you would have call them "fiction". Even when guided by domain knowledge, is there guarantee of factualness from the LLM? (Likely not.) You may not have a decision-support system, but you do potentially have a teaching/informational system, based on which humans will make decisions in the future --- so the ethical implications remain serious. So, in the AI section, the answers are not quite correct: there is bias, user expectation can be unrealistic, there could be some unintentional harm.

At least in your report, please document the matter of factualness as much as you can.

Gatti, L. (EEMCS-HMI) (26-05-2025 14:44) :

see feedback via Teams

13. CONCLUSION

Status: Positive advice by reviewer

27-05-2025 13:28

Welcome to the Weaving Stories,

Thank you for participating in this short survey that generates narratives on Roman Empire. Now, it's time to test if this project can be useful, and informational to the end users. To achieve this, I request your valued opinion on 9 short historical narratives generated by an AI in response to a user query. This survey is tailored for individuals proficient in English who are passionate about learning history.

Your Privacy is Important

We hold your privacy in the highest regard. No personal data will be collected during this survey. All responses will remain completely anonymous.

Navigating the Survey

Please keep in mind the following guidelines:
Provide your insights honestly and thoughtfully. You will be asked to evaluate responses generated by the various models on the same question based on coherence, understandability, engagement.
You can share any remarks at the end of the survey.
You can exit the survey at any time before submission.
Once submitted, responses cannot be removed. (As the responses are anonymous, we cannot retrieve them for deletion)

Estimated Time

We estimate that this survey will take approximately 20-25 minutes of your time. Your contribution to this research is highly valuable.