Exploring Data Augmentation and Memory Strategies for AI-based Synthetic Personae

by

Rafael Arias Gonzalez

B.Sc., Middlesex University, 2017
B.Eng., Sae University, 2017

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science

in the
School of Interactive Arts and Technology
Faculty of Communication, Art and Technology

© Rafael Arias Gonzalez 2024
SIMON FRASER UNIVERSITY
Summer 2024

Copyright in this work is held by the author. Please ensure that any reproduction or re-use is done in accordance with the relevant national copyright legislation.

Declaration of Committee

Name:	**Rafael Arias Gonzalez**
Degree:	**Master of Science**
Thesis title:	**Exploring Data Augmentation and Memory Strategies for AI-based Synthetic Personae**
Committee:	**Chair:** Kate Hennessy Associate Professor, Interactive Arts and Technology
	Steve DiPaola Supervisor Professor, Interactive Arts and Technology
	Ozge Nilay Yalcin Committee Member Assistant Professor, Interactive Arts and Technology
	Alireza Karduni Examiner Assistant Professor, Interactive Arts and Technology

Abstract

In this thesis, we investigate the enhancement of synthetic personae through data augmentation and memory strategies, focusing on a case study involving the development of a chatbot impersonating Van Gogh. Leveraging large language models (LLMs), we explore the integration of a novel autonoetic memory dataset derived from Vincent Van Gogh's biography and letters to improve the chatbot's question-answering capabilities by accessing different layers of memory. This research not only delves into the potential of LLMs for creating engaging synthetic personae but also addresses the challenges of data augmentation and the practical implementation of memory systems in chatbots. Through comparative analysis, we demonstrate the superiority of the proposed approach over traditional models, highlighting its contributions to the fields of Human-Computer Interaction (HCI) and synthetic personae development. This work sets the stage for future exploration in enhancing chatbot interactions and opens new avenues for research in cognitive model design.

Keywords: Large language models; HCI; Artificial Intelligence

Dedication

To my mom for being the best teacher I've ever had. To Ramis for teaching me the value of uniqueness and honesty. To my dad for teaching me the virtues of hard work. To Naf for being a light in the darkness. To Zaz and Jos for keeping me alive. To Kari for all the happiness and peace.

Acknowledgements

I would like to acknowledge my supervisor, Steve DiPaola, for all his guidance. Your support was truly invaluable. I also want to thank my committee member, Nilay Yalcin, for providing me with support and wise insights to improve the quality of my work. I thank my examiner Alireza Karduni for agreeing to be a part of my committee. Lastly, I want to thank Steve, Nilay, Tiffany Taylor and Kate Hennessy for providing me with advice and support when I needed it the most.

Table of Contents

Declaration of Committee **ii**

Abstract **iii**

Dedication **iv**

Acknowledgements **v**

Table of Contents **vi**

List of Tables **ix**

List of Figures **x**

1 Introduction **1**

1.1 Motivation 1

1.1.1 Challenges for AI in HCI 2

1.1.2 Energy Concerns and Limitations of Large-Scale AI 2

1.1.3 Limitations of Computing and AI Scaling 5

1.1.4 Industry Trends 5

1.2 Problem Statement 6

1.3 Hypothesis 7

1.4 Contributions 8

1.4.1 Data Augmentation via LLMs 8

1.4.2 Memory Systems for Enhanced LLM Chatbots 8

1.5 Organization of Thesis 9

2 Related Work **10**

2.1 Large Language Models (LLMs) 10

2.1.1 Attention Constraints and Implications for Research 10

2.1.2 Emergence of Smaller Models and Research Implications 11

2.1.3 Natural Language Processing and Human-Computer Interaction . . 11

2.2 Hallucination 11

2.2.1 Mitigation Strategies 13
2.3 Retrieval Augmented Generation (RAG) 13
2.4 Self-reflection 13
2.5 Data Augmentation 14
2.6 Memory 14
2.6.1 Human Memory 14
2.6.2 Synthetic Personae and Memory 15
2.7 Evaluation Methods for Retrieval Augmented Generation Systems 18
2.7.1 LLM-as-a-Judge 18
2.7.2 Retrieval Augmented Generation Assessment (RAGAs) 18

3 Autonoetic Memory Dataset **20**
3.1 Data Augmentation 20
3.1.1 Data Sources 20
3.1.2 Limitations of the Raw Data 21
3.1.3 Text Rewriting: Script Generation 23
3.1.4 LLM-Driven Information Extraction and Structuring 23
3.1.5 Data Enrichment and Formatting 25
3.1.6 Final Dataset Structure 26

4 Chatbot Architecture and Implementation **28**
4.1 System Architecture 28
4.1.1 RagChain API 28
4.1.2 Van Gogh Chatbot: Architecture and Design Considerations 30
4.1.3 Autonoetic Memory: Two-Stage Access and Information Retrieval . 32

5 Data Analysis and Comparative Evaluation **35**
5.1 Preliminary Explorations 35
5.2 Descriptive Statistics 35
5.2.1 Valence and Arousal 35
5.2.2 Characters 37
5.3 Exploratory analysis 39
5.3.1 Interactive Graphs 39
5.3.2 User Paths 40
5.4 Chatbot Comparative Evaluation 43
5.4.1 System Speed 43
5.4.2 Chatbot Evaluations 44
5.4.3 LLM-as-a-judge 44
5.4.4 RAGAs Framework 46
5.4.5 Intermediate Memory 46

5.4.6 Ethical Considerations 47

6 Discussion **49**

6.1 Develop Robust Autonoetic Memory Datasets 49
6.2 Achieve Efficient and Optimized Real-time Interactions 51
6.3 Benchmark System Performance with and Without Memory Integration . . 51
6.4 Limitations 52
6.5 Future research 53

Bibliography **54**

Appendix A Data Augmentation Prompts **62**

A.1 LLM-Screenwriter Writing Prompt 62
A.2 LLM-Screenwriter Sample Response 63
A.3 VGE-LLM Example Query 64
A.4 VGE-LLM Sample Response 65
A.5 LLM-as-a-Judge prompt 65
A.6 RAGAs Answer Relevance Prompt 65
A.7 RAGAs Context Relevance Prompt 66
A.8 RAGAs Faithfulness Statements Prompt 66
A.9 RAGAs Faithfulness Classification Prompt 66

Appendix B Relevant Queries and Answers **67**

B.1 Static Context Prompt Example 67
B.2 Full Context Prompt Example 68

Appendix C Chi'24 Workshop 21 Position Paper **69**

List of Tables

Table 1.1 Exponential growth of parameters and GPUs, with linear growth of GLUE Scores. 3

Table 3.1 Description of fields in the autonoetic memory dataset. 27

Table 4.1 Token limitations for different components within the system 32

Table 5.1 Comparative Test Results. Full means the (Static+Autonoetic+Data) model. 46

Table 5.2 Evaluation scores using the RAGAs framework 47

List of Figures

Figure 1.1 GPUs used on training vs GLUE Score. 4
Figure 1.2 Comparison between several small models and GPT-3 and GPT-3.5. The size of the data point represents the number of parameters for each model. 6

Figure 3.1 Autonoetic Dataset Pipeline. 24

Figure 4.1 Flowchart of the Van Gogh Chatbot Design 30
Figure 4.2 Autonoetic Dataset Pipeline. 33

Figure 5.1 Valence and Arousal boxplots. 36
Figure 5.2 Yearly Weighted Average Valence and Arousal. 37
Figure 5.3 Character Mentions 38
Figure 5.4 Mentions of Van Gogh's Love Interests Through Time. 38
Figure 5.5 Screenshot of a heatmap with all memories. 39
Figure 5.6 Screenshot of a heatmap with a randomly selected group of memories. 40
Figure 5.7 Screenshot of the provided UI for feature selection. 40
Figure 5.8 PCA of the Autonoetic Memory Dataset. Yellow memories have higher valence, and dark-blue memories decreasing valence. 41
Figure 5.9 PCA of the Autonoetic Memory Dataset by Valence and Arousal values. Yellow memories are memories with higher valence. 42
Figure 5.10 Conversation Path of one of our interactions with the chatbot. . . . 43

Figure 6.1 When queried "Tell me the happiest you've ever been", the chatbot accesses low-valence memory 8.17, which reminisces of a Christmas spent with his family in 1874. 50

Chapter 1

Introduction

This thesis investigates the utilization of data augmentation and cognitive model design to construct robust and reliable synthetic personae in conversational systems. Data augmentation, in the context of this thesis, refers to the process of utilizing large language models (LLMs) to transform and enrich existing datasets. This involves generating synthetic data that complements and expands upon the original information, often by incorporating additional context, perspectives, or formats. Our approach centers on role-playing and character embodiment, demonstrated through developing a Van Gogh chatbot capable of accessing short, intermediate, and long-term memory for nuanced question answering.

We begin with a comprehensive review of key principles of large language models (LLMs) and synthetic personae. Subsequently, we describe our system design and methodological framework. We then delve into data analysis explorations with our augmented autonoetic memory dataset, offering insights into the potential of data augmentation via LLMs. Autonoesis refers to a type of self-awareness that allows individuals to subjectively re-experience their past in first person, including emotions and a sense of being present in the memory [50]. Then, we make a comparative analysis of the chatbot's responses to showcase the efficacy of our approach.

To conclude, we summarize our research contributions and outline avenues for future exploration alongside a critical discussion of identified limitations within the system design.

1.1 Motivation

The rapidly developing field of LLMs has produced increasingly sophisticated chatbots and synthetic personae. We define synthetic personae as an AI-powered construct that embodies a particular character with a unique identity, memories, and the ability to engage in meaningful dialogue or other forms of interaction. The remarkable ability of LLMs to perform well on various human benchmarks, such as passing the core course evaluations of an MBA or exceeding at reasoning tasks [62, 70], underscores their potential and highlights the need to explore their limitations as synthetic personae and synthetic data producers. While current achievements are significant, achieving deeper and more nuanced simulations could help increase their use past simple API prompting. Infusing LLMs with richer personality representations is crucial for creating more lifelike and believable simulations.

Human-like simulacra offers significant benefits across various domains. They provide a powerful tool for research in the social sciences, could enhance NPC development in virtual environments, serve as educators and tutors, enhance museum interactions, and even have the potential to serve as synthetic participants in studies. Our aim to enhance LLMs' abilities to experience simulated events, process emotions, and retain contextualized memories helps us move closer to developing agents that embody distinct personalities.

1.1.1 Challenges for AI in HCI

Large language models (LLMs) present a promising avenue for research in human-computer interaction (HCI), particularly in creating synthetic personae and generating synthetic data. However, their tendency to produce hallucinations (outputs that are factually incorrect or nonsensical) and their black-box nature pose significant challenges [47]. To address these limitations, we propose leveraging LLMs as data augmentation systems rather than as standalone data generators. This approach involves providing LLMs with substantial context to augment, thereby enhancing their ability to simulate the nuances of personae.

Furthermore, we advocate for the development of robust cognitive and memory frameworks to guide LLM responses. These frameworks would enable efficient retrieval of relevant data in an accessible format, mirroring how humans dynamically access memories during interactions. By incorporating episodic memory and self-reflection techniques, we can improve the reliability and consistency of synthetic personae.

Our initial explorations [4] (see C), using the historical figure of Vincent van Gogh as a test subject, have shown promising results. By augmenting biographical data with first-person perspectives and scene-specific context, and by integrating an episodic memory system with an LLM, we were able to generate more informative, focused, and contextually relevant responses.

This approach has several potential applications in HCI research. It can facilitate extensive interviews with synthetic personae, even in scenarios that might be challenging for human participants. Additionally, it offers a degree of explainability by providing access to augmented data and ranked retrieved scenes, allowing researchers to understand the reasoning behind the model's responses.

By addressing the challenges of hallucination and black-box limitations, we can unlock new possibilities for creating more reliable, consistent, and informative synthetic personae.

1.1.2 Energy Concerns and Limitations of Large-Scale AI

Beyond the specific HCI goals of our research, our work is secondarily motivated by a desire to enhance the performance of smaller language models in light of pressing energy

Model	Parameters	GPUs	GLUE Score
GPT-2	1.5B	8	44.5
GPT-3	175B	1024	71.8
GPT-4	1.7T	25000	89.5

Table 1.1: Exponential growth of parameters and GPUs, with linear growth of GLUE Scores.

constraints and the limitations of current trends in AI development. Extensive research highlights the massive energy consumption associated with training and deploying large-scale AI systems. Both Utz and Dipaola [65], as well as Strubell et al. [59], provide evidence demonstrating the massive energy demands of these models. While the exact environmental impact remains a topic of debate, concerns about AI's carbon footprint underscore the need for more sustainable approaches.

We identify three core issues with current state-of-the-art models:

Problem 1: Model Sizes

The sheer number of parameters in models like GPT-4 (1.76 trillion) [1] directly translates to vast energy requirements for training. Strubell et al. [59] found that training BERT (110 million parameters) [14] uses the same amount of energy as a trans-American flight. While this amount of energy might not be a lot on its own, these costs are compounded by the numerous hyperparameter experiments (training iterations) common in model development, as reported by Schwartz et al. [52]. They report that, for one model, "researchers from Google trained over 12,800 neural networks in their neural architecture search to improve performance on object detection and language modelling".

Moreover, the law of diminishing returns may force a limit on the ever-growing sizes of models: Massive increases in data and the size of the models yield minimal increases in model accuracy. Mahajan et al. [37] discovered that "object detection accuracy increases linearly as the number of training examples increases exponentially" [52]. Thompson et al. [63] note that in 2012 AlexNet (the model that kickstarted the current AI revolution) only needed 2 GPUs and 6 days to train. In contrast, in 2018, NASNet-A was twice more accurate than AlexNet, but by using 1000 times more power to train. They expand on this by saying that current models need around 500 more times the computational resources to halve their error rate. We can empirically observe this in Table.1.1 by comparing estimations of the amount of GPUs used to train GPt-2 [29], GPT-3 [12] and GPT-4 [35] compared to their respective parameters and their GLUE score [6]. We can observe that the GLUE improvement is not exponential, compared to the parameters and GPUs used for training. Fig.1.1 demonstrates the linear GLUE score growth compared to the exponential GPU usage.

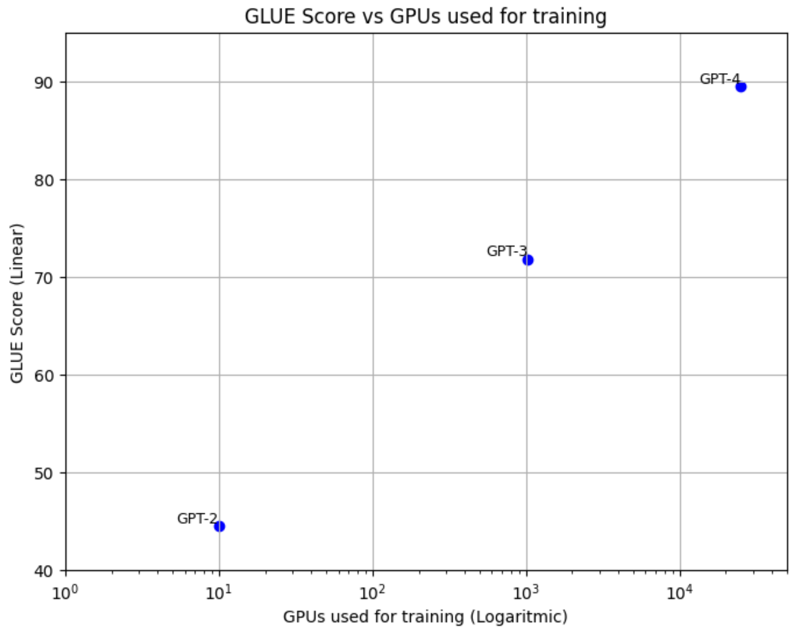

Figure 1.1: GPUs used on training vs GLUE Score.

Problem 2: Growing Demand

What these trends also highlight is the exponential correlation between model size and energy needed for inference, as noted by Desislavov et al.[13]. Even modest increases in model size lead to exponential increases in energy consumption during inference. Even though a single inference takes much less energy than training, the overall power consumption can be potentially higher because of the multiplicative factor [13]. That is, a single model may be deployed many times, and each model instance will probably make inferences repeatedly during its lifetime. When compounded across numerous devices and applications, the cumulative energy cost is much larger.

Problem 3: Cloud Dependency

The computational demands of large models often necessitate cloud-based inference, since the models cannot be embedded in most personal devices. For example, a single instance of GPT-4 needs 128 A100 GPUs to run inferences [67]. Cloud-approaches introduce additional energy burdens associated with server maintenance and the vast network infrastructure required to support cloud requests, further amplifying the environmental impact of AI systems. As Kate Crawford puts it [11], each request to the cloud "requires a vast planetary network, fueled by the extraction of non-renewable materials, labour, and data."

Because of these interconnected issues, we consider that it is vital to explore alternative paradigms that prioritize efficiency and sustainability. Our research, therefore, aims to empower smaller models to achieve greater performance, potentially mitigating environmental impact while expanding the reach of AI applications.

1.1.3 Limitations of Computing and AI Scaling

Our research is further motivated by the trends of diminishing returns in hardware scaling and the inherent energy inefficiency of current AI architectures. The semiconductor industry is facing significant setbacks, including the physical limits of Moore's Law [54] and potential material shortages as outlined in the Decadal Plan for Semiconductors [53]. The persistent GPU shortages [77] further highlight these bottlenecks and underscore the challenges of relying on ever-increasing computational resources.

Moreover, the contrast in energy consumption between human cognition and AI systems raises profound questions about sustainability and the efficiency of our current AI model paradigm. As noted in the 2021 Decadal Plan [53], the human brain achieves remarkable computational feats with minimal energy expenditure compared to AI supercomputers. The brain's efficiency advantage suggests a fundamental mismatch between current AI paradigms and the biological models that inspire them. Our work aims to explore alternative approaches that prioritize efficiency, potentially opening new avenues for progress while respecting the practical constraints of real-world deployment.

1.1.4 Industry Trends

Our research approach reflects a growing industry trend towards smaller, more deployable AI models. The inclusion of smaller models by major players like Google (Gemma 2B and 7B [38]), Meta (Llama 3 8B [40]) and Mistral (Mistral 7B [43]) demonstrate a shift in focus. While these smaller models mark a step in the right direction, our exploratory tests on Gemma 2B and Mistral 7B reveal significant performance limitations. Our motivation, therefore, lies in developing techniques to significantly enhance their capabilities. To this end, we strategically utilize GPT-3 and GPT-3.5. While these models are not as small as 2B, 7B and 8B models, they exhibit performance characteristics comparable to them [9, 39, 19]. Fig.1.2 shows the comparison between several models and GPT-3 and 3.5 in the Arc-c test [74], which is a common test for language reasoning. We chose the OpenAI suite for development of our system since it offers essential flexibility for rapid experimentation across model sizes, enabling us to easily explore and optimize our approach for eventual adaptation to smaller, more resource-efficient architectures. Using OpenAI's API allowed us to develop state-of-the-art models (GPT-4) and then easily scale down to the GPT models that are similar in performance to the 2B, 7B, and 8B models.

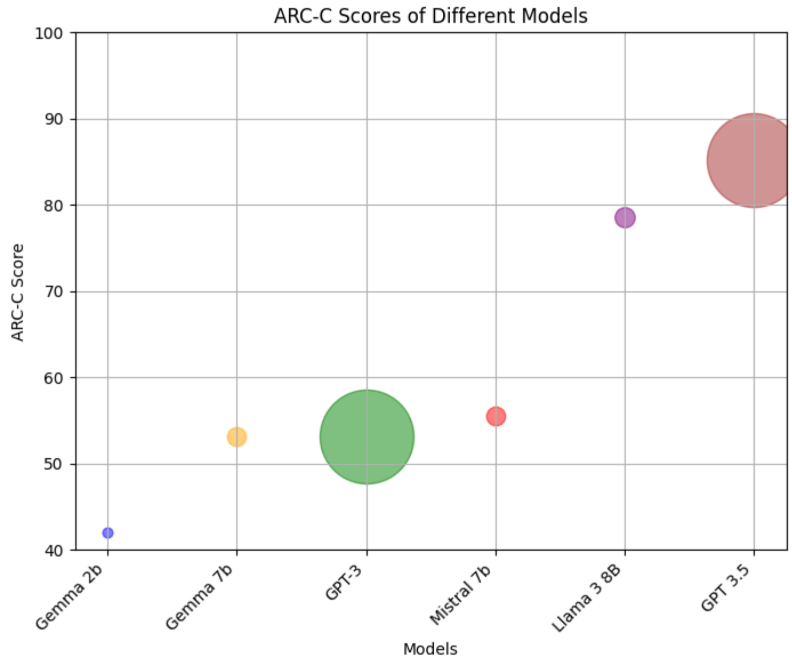

Figure 1.2: Comparison between several small models and GPT-3 and GPT-3.5. The size of the data point represents the number of parameters for each model.

1.2 Problem Statement

Despite the remarkable advancements in LLMs, they continue to face several inherent limitations to their practical use in creating synthetic personae within HCI contexts.

Firstly, LLMs demonstrate a tendency toward hallucination [75], generating factually inaccurate or nonsensical responses despite exhibiting surface-level fluency and logic. This propensity for misinformation undermines the reliability of LLMs, making it hard to discern between credible and fabricated information.

Secondly, the absence of persistent and grounded memory represents a critical shortcoming of LLMs for synthetic persona applications [25]. We define grounded memory as a persistent, internally consistent cognitive model within an LLM, allowing it to maintain a stable understanding of the world. A lack of memory impedes an LLM's ability to deliver the consistency and explainability demanded by meaningful HCI interactions. While an LLM's parameters may facilitate some internal reasoning, any after-explanations it generates likely diverge from the underlying computational processes, as they are not grounded in a persistent memory structure. The fundamental challenge of grounding LLMs in objective reality is exemplified by the classic "guess the object" game. Humans can effortlessly store a selected object in memory, enabling them to provide consistent clues and responses. In contrast, LLMs may produce contradictory answers that lack a clear referent, considering that their answers rely only on static trained parameters (there is no memory to store the seemingly "thought" object).

Additionally, the environmental impact of training and deploying these large models is a growing concern due to their substantial energy consumption and carbon footprint.

Finally, the massive sizes of out-of-the-box LLMs frequently make users rely on expensive APIs, limiting their accessibility and practicality.

In response to these challenges, various techniques have emerged within the research landscape. Retrieval-augmented generation (RAG) [33] is a technique that allows language models to access and incorporate information from external knowledge sources. RAGs have proven effective in mitigating hallucination, although the reliability of simpler RAG systems remains an issue. Complex, multi-layered RAG architectures have been explored [**?**] but often introduce computational overhead and latency. Similarly, self-reflection [27], prompting the LLM to generate internal rationales, holds promise but also increases inference time. Other approaches, including fine-tuning models for enhanced precision [41] and episodic memory retention [55], offer potential for addressing memory concerns but do not resolve hallucinations and demand more complex technical pipelines. Lastly, it has been shown that Role-Playing enhances LLM reasoning [32]. In the context of LLMs, role-playing refers to the ability to simulate the language and behaviour of a specific character based on prompts and information provided.

1.3 Hypothesis

Given the inherent lack of persistent memory in LLMs (aside from the context window afforded by attention mechanisms), we theorize that grounding techniques hold significant promise for enhancing their capabilities. Simultaneously, we recognize the importance of maintaining reduced processing and inference times to ensure optimal user engagement with chatbots.

Building upon these considerations, we hypothesize that creating robust, offline episodic memory databases for prompt data retrieval can significantly improve an LLM's capacity to generate grounded, contextually relevant responses. We believe this approach would mitigate the detrimental impact of response delays and enhance the overall coherence and fidelity of synthetic personae in HCI applications.

Based on our hypothesis, we propose the following research goals:

- Develop Robust Autonoetic Memory Datasets: Utilize LLMs as data augmentators to create offline autonoetic memory databases to store and retrieve contextually rich information, providing the foundation for grounded LLM responses.

- Achieve Efficient and Optimized Real-time Interactions: Optimize response generation times with integrated episodic memory databases. This will balance practicality and reduce processing demands and energy consumption while maintaining the LLM's enhanced capabilities.

- Benchmark System Performance with and Without Memory Integration: Benchmark the model's performance with and without memory integration, demonstrating how episodic memory improves the coherence, accuracy, and overall reliability of LLM responses.

1.4 Contributions

This work presents two principal contributions, detailed in Chapters 3 and 4.

1.4.1 Data Augmentation via LLMs

Firstly, we investigate the potential of synthetic data augmentation through LLMs. The effective utilization of LLMs for synthetic data generation remains an active research area in HCI [21, 42, 28]. We successfully constructed a novel autonoetic memory dataset derived from Vincent Van Gogh's biography and letters. This dataset is enriched with contextual information, relevance markers, and numerical data, enabling its versatile use for analysis, retrieval, and model training.

To the best of our knowledge, our research represents the first comprehensive effort to leverage LLMs to create an entire multi-domain dataset derived from rich biographical and epistolary data. The dataset aims to optimize LLM-based inference across diverse tasks, including question-answering, creative text generation, and knowledge extraction.

As a secondary contribution derived from this dataset, we devised exploratory approaches to data exploration and analysis, such as memory visualization and user paths. These contributions are shown in chapter 5.3.

1.4.2 Memory Systems for Enhanced LLM Chatbots

For our second main contribution we explore the integration of memory systems for character embodiment and the creation of synthetic personae. While memory augmentation in LLMs is a well-researched domain, existing work primarily emphasizes data retrieval for informational chatbots, with limited attention to role-playing applications. Our approach uniquely simulates autonoetic memory through augmented retrieval techniques. To benchmark the system's effectiveness, we present a comparative analysis of models with and without integrated memory, demonstrating the enhanced reliability of our memory-enabled chatbot. Moreover, our approach allows us to access data other than first-person textual memories. We can extract numerical values, offering improved explainability and further enriching response generation.

Finally, a significant secondary contribution lies in the efficiency of our approach. By offloading self-reflection to a pre-established dataset and facilitating complex data access through simple retrieval methods, we achieve negligible prompt generation times, averaging 0.5 seconds[1]. This efficiency underscores the practical viability of our system for real-time HCI applications. Furthermore, we reduce energy consumption by reducing hidden and intermediate calls to the LLM.

1.5 Organization of Thesis

Chapter 2 establishes the conceptual groundwork for the thesis. It provides a comprehensive overview of LLMs, their applications as chatbots and synthetic personae within the HCI domain, and the principles of data augmentation. The chapter covers critical concepts such as RAGs, self-reflection, and relevant research on memory systems to mitigate hallucination in LLMs.

Chapter 3 delves into the data augmentation process and the creation of the autonoetic memory dataset. We identify the original data sources, including biographical materials and Van Gogh's letters, providing illustrative examples of the prompts used to guide the LLM's data transformation. Additionally, the chapter offers a detailed description of the dataset itself, outlining its structure, content, and key characteristics.

Chapter 4 focuses on the technical aspects of the chatbot's development. It introduces a general-purpose API designed to facilitate the creation of LLM-based synthetic personae before transitioning to a detailed explanation of the Van Gogh chatbot's implementation. We explain the system architecture, emphasizing using RAGs to enable short, intermediate, and long-term memory capabilities powered by the autonoetic memory dataset.

Chapter 5 presents a multifaceted analysis of the work. It begins by exploring the autonoetic memory dataset, showing insights gained through data analysis and exploration. The chapter then transitions into a comparative evaluation of the chatbot's responses, benchmarking its performance against models lacking the integrated memory system to demonstrate the improvements achieved.

Chapter 6 discusses the research findings, highlighting our contributions and their significance. We also discuss the limitations of our approach, prompting a discussion of future avenues of investigation. Lastly, we discuss potential applications for the dataset and the chatbot.

[1] Time it takes to generate the prompt, before it gets fed into the LLM for inference.

Chapter 2

Related Work

2.1 Large Language Models (LLMs)

Transformer models [66] are at the core of remarkable advancements within the artificial intelligence domain, particularly in the realm of LLMs and chatbots. Their capacity to model long-range dependencies among tokens positions them as powerful tools for text generation tasks. This capability is fundamentally enabled by the attention mechanism, which dynamically computes the relative importance of elements within a query or input sequence.

2.1.1 Attention Constraints and Implications for Research

Attention limits are relevant to our investigation as they define the immediate scope of information a model can actively process at a given time. Contemporary models such as GPT-4 [1] and Gemini 1.5 [49] have massive context windows (128k and 1M, respectively). However, the majority of commercially available models operate with windows ranging from 2,000 to 8,000 tokens [72]. Our development process employed GPT-3.5 (4,000 token limit) and GPT-4 with an 8,000 token window[1]. We chose this model instead of an open-source one to focus on prompt engineering and avoid technical complications. Still, we aim for this research to be a stepping stone into leveraging local models for chatbots.

The ongoing discourse surrounding the environmental impact of AI underscores the significance of context window size. Larger, more powerful models undeniably consume more significant energy resources [52]. Furthermore, the demands of HCI applications often necessitate deployment on devices with limited computational capacity, excluding cloud-based models or devices with advanced GPUs. While the industry anticipates models with ever-expanding context windows, our research explores optimizations within smaller windows. This focus aligns with environmental considerations discussed in the motivation and the practical realities of deploying chatbots on less powerful devices.

[1] the context window was 8k for most of the development since the latest 128k models hadn't been released.

2.1.2 Emergence of Smaller Models and Research Implications

Our work aligns with the emergence of compact models like Gemma2B [38], tailored for CPU and edge deployment. This trend reinforces the importance of our research into techniques for achieving robust performance with smaller context windows. However, a lack of comprehensive documentation for recently released models obscures precise context window determinations. Because of this uncertainty, we adopted a conservative approach, limiting queries to a maximum of 2,000 tokens, where we consider 1,500 tokens for the query and 500 tokens for the model response). Despite the potential constraints, we argue that working within smaller context windows promotes responsible AI development. It reduces energy consumption and enables future deployment of these models on a broader range of devices, making them more accessible for various applications.

2.1.3 Natural Language Processing and Human-Computer Interaction

The remarkable capabilities of LLMs have allowed their widespread adoption within both Natural Language Processing (NLP) and Human-Computer Interaction (HCI) research. HCI research has witnessed a surge of interest in exploring LLMs for character embodiment [55], role-playing [68], and the creation of synthetic personae [46]. LLMs, through various techniques, can be transformed into engaging conversational models, commonly known as chatbots.

Despite LLMs commercial success in general-purpose conversational AI (e.g., ChatGPT [7], Gemini [49]), the use of these systems as reliable assistants, or synthetic personae remains hindered by several persistent challenges. We identify hallucination, lack of explainability, inference times, and accessibility (ease of access for non-technical users) as particularly critical issues.

While much active research focuses on model creation, re-training, and fine-tuning to enhance chatbot performance, we observe that these modifications, while valuable for other purposes, do not directly target the challenges mentioned above. To our knowledge, no existing research establishes a clear connection between fine-tuning LLMs and explainability concerns, inference times, or accessibility limitations. Fine-tuning does help reduce hallucination, but only partially [48].

2.2 Hallucination

Hallucination, a persistent issue in LLMs, occurs when models generate seemingly coherent text that contradicts factual reality or internal logic. This phenomenon undermines the reliability of LLMs, as they present inaccurate information with a high degree of confidence, making it difficult to differentiate between fact and fiction. Hallucination as an inherent problem of LLMs has been widely documented.

Rawte et al. [47] conduct a comprehensive survey on hallucinations within Foundation Models (FMs), which are models trained with massive, general-knowledge amounts of data, such as GPT. They define hallucination as the FM's generation of content that deviates from reality or includes fabricated information.

Huang et al. [23] surveyed the phenomenon of hallucination in LLMs. Their work provides a systematic taxonomy of LLM hallucinations, identifying the various forms they can take. Moreover, they explore the underlying factors that contribute to these hallucinations. They identify three key areas where hallucinations originate:

- Data Issues: Pre-training data can introduce misinformation, biases, and knowledge gaps. Models may repeat incorrect facts ("imitative falsehoods"), overemphasize patterns from repeated information ("duplication bias"), or reflect biased associations ("social biases"). Additionally, models are limited by the scope of their training data, lacking specific domain expertise or up-to-date knowledge.

- Training: LLMs may develop "knowledge shortcuts", relying on spurious correlations within the data rather than robustly understanding factual information. This can lead to errors, especially when the model encounters less common ("long-tail") knowledge or questions that require complex reasoning.

- Inference: Even when an LLM possesses the necessary information, it can still hallucinate if it struggles to recall the relevant knowledge effectively or perform the necessary logical deductions to arrive at the correct answer.

Ji et al. [26] delve into the issue of hallucination within dialogue generation systems. They distinguish between intrinsic hallucinations, where generated responses contradict the conversation's own history or go against general world knowledge, and extrinsic hallucinations, which cannot be traced back to a reliable, verifiable source. Their work highlights the challenges of maintaining self-consistency and external consistency in open-domain dialogue, specifically regarding persona coherence and aligning with background knowledge. The authors outline mitigation strategies such as data pre-processing, knowledge retrieval, control code integration, and model architecture refinements.

It is not yet known if we can directly compare LLM hallucinations to human hallucinations. Clinical definitions of human hallucinations differ substantially from the errors observed in LLMs [58]. Alternative terms such as "confabulation" [58] have been proposed, though we believe non-pathological terms like "misattributions" or "false memories" may be more fitting. Further research is needed to conceptualize LLM errors accurately and determine their relationship (if any) to human cognitive processes.

2.2.1 Mitigation Strategies

Several techniques have been developed to mitigate hallucination in dialogue generation. Notably, Shuster et al. [57] propose knowledge graph integration as a mitigation strategy, which leverages knowledge graphs and curated knowledge bases to ground LLM responses. For knowledge-grounded dialogue (KGD) tasks, RAG methods have demonstrated significant success in reducing hallucinations and improving conversational coherence [71]. Additionally, research has explored self-reflection as a potential means of reducing hallucinations. This approach encourages the LLM to generate internal rationalizations, allowing it to potentially identify and correct its own errors.

2.3 Retrieval Augmented Generation (RAG)

RAG methods enable LLMs to access and leverage external knowledge sources before response generation, improving factual grounding and response generation. Gao et al. [16] offer a comprehensive survey of RAG techniques for enhancing LLMs. Naive (single) RAG systems lack depth, so research has focused on enhancing RAG systems. Modular RAG architectures provide adaptability, with specialized modules for retrieval, memory, and other functions. RAG chaining allows for iterative refinement of results through multiple retrieval and generation steps, which is particularly useful for complex queries. While iterative, recursive, and adaptive RAG mechanisms can enhance the system's flexibility and context awareness, a notable drawback is their increased computational overhead. The multiple calls to the LLM required by these processes can significantly impact inference times.

Shuster et al. [57] explored the use of RAG models to address knowledge hallucination in conversational AI. Their findings revealed that incorporating retrieval architectures into dialogue systems can substantially reduce these hallucinations, enhancing factual accuracy and overall conversational ability. This research holds particular relevance for developing synthetic personae, as character embodiment requires both conversational fluency and accurate knowledge representation to achieve authenticity.

2.4 Self-reflection

Self-reflection techniques, while varied, refers to techniques that allow the model to assess its own outputs, potentially identifying and correcting errors through internal analysis or interaction with external tools.

Gou et al.[20] introduce CRITIC, a framework enabling LLMs to self-correct their outputs by interacting with external tools. This approach, inspired by human processes like fact-checking or code debugging, aims to mitigate the generation of hallucinations, flawed logic, or harmful content. CRITIC's work reveals a parallel with using RAG models to build memory systems for synthetic personae. Both methods prioritize external knowledge and iterative refinement to increase the quality of LLM outputs. The paper emphasizes the importance of a reflection mechanism for generating nuanced and contextually appropriate responses. However, CRITIC's reliance on external tool verification and iterative correction introduces significant computational overhead. Each step requires generating an initial output, validating it through tool interaction, and refining it based on obtained feedback. This makes the process more resource-intensive and time-consuming than direct LLM generation, especially in applications with computation or response time constraints.

Huang et al. [24] introduce InteRecAgent, a recommender system that integrates LLMs with traditional recommendation techniques for enhanced user interaction and personalization. An important aspect of InteRecAgent's design is a modular architecture with components for information query, item retrieval, and ranking, optimizing the processing of complex user requests. While the system's reflection mechanism improves recommendation accuracy, it introduces computational overhead due to iterative LLM calls.

Madaan et al. [36] propose SELF-REFINE, a technique for enhancing LLM outputs using iterative self-feedback. Inspired by human self-improvement processes, this method uses a single LLM for initial output generation, self-critique, and iterative refinement without needing external data or additional training. However, like the previous approaches, this technique needs multiple LLM processing steps, increasing computational overhead and slowing response times.

2.5 Data Augmentation

Data augmentation, the creation of synthetic data from an existing dataset [56], has become increasingly prevalent in NLP research. LLMs have demonstrated their remarkable ability to transform and generate data tailored for specific tasks. This encompasses enhancing model robustness through diverse conversational reformulations [8], facilitating cross-lingual understanding via multilingual example generation [71], and enhancing the identification of specific words or phrases by rewriting examples in data-scarce scenarios [76]. With the growing sophistication and adaptability of LLMs, data augmentation techniques hold the potential to accelerate NLP advancements across various domains significantly.

2.6 Memory

2.6.1 Human Memory

To create LLMs capable of rich and nuanced character embodiment, we must design systems that efficiently retrieve and present relevant information in ways that mirror the complexities of human memory. Cowan et al. provide a foundational framework, differentiating between short-term, working, and long-term memory systems [10]. The authors describe short-term memory as the ability to hold a small amount of information in an accessible state for a short period. This can be thought of as the information that is currently active in one's mind, such as a number you just heard or the words in the sentence you are currently reading. They define working memory as closely related to short-term memory but involving not only the temporary storage of information but also its manipulation. It is the type of memory used for more complex tasks, like reasoning or comprehending. Working memory allows us to hold information in mind while simultaneously processing it. Finally, they define long-term memory as a large and persistent storage system that holds our knowledge, skills, and experiences. It is the repository of everything we have learned and remembered over our lifetime. Unlike short-term and working memory, long-term memory has a virtually unlimited capacity and can store information for extended periods, potentially for a lifetime. Long-term memory is essential for our sense of self, our understanding of the world, and our ability to function in daily life.

Semantic and Episodic Memory

Within the domain of long-term memory, Schendan et al. [51], building on Tulving's [64] work, differentiate semantic and episodic memory. Semantic memory focuses on objective facts, concepts, and general world knowledge. We can view the embedded knowledge base of an LLM as a form of semantic memory consisting of factual information acquired during training. However, this knowledge lacks the subjective experiences characteristic of episodic memory, which refers to the personally experienced events in specific spatiotemporal contexts. Autonoetic consciousness, the ability to mentally project oneself back in time as an active participant with an accompanying sense of self, is deeply intertwined with episodic recall [50]. This subjective reliving is often absent in traditional Retrieval-Augmented Generation (RAG) systems, which focus on factual retrieval.

Autonoesis and Embodiment

Sant'Anna et al. [50] define autonoesis as the self-awareness associated with re-experiencing the past. It involves not just remembering but a subjective sense of "self in the past" and remembering emotions tied to the original event. As Schendan et al. [51] emphasize, this personal connection differentiates episodic memory from semantic memory, which focuses on meaning without familiarity. Developing LLMs capable of simulating this autonoetic quality could help deepen character embodiment, moving beyond simpler factual retrieval towards recalling character perspectives and lived experiences.

2.6.2 Synthetic Personae and Memory

As discussed in Section 1.1, the development of synthetic personae offers significant potential within the realm of HCI.

We define the term "synthetic personae" as AI-powered representations of individuals designed to embody their unique identities, memories, demographics, behaviors, preferences, and emotional states. This concept aligns closely with the broader field of virtual agents, defined by Palwe et al. [2] as software programs that leverage scripted rules and artificial intelligence to interact with humans in an automated manner. However, we distinguish synthetic personae by their emphasis on embodying human-like characteristics. They aim to simulate the personalities and experiences of individuals, whether fictional or historical, through the computational modeling of memory, emotional responses, and personal preferences. By adopting the term "synthetic personae," we aim to provide a framework for discussing these systems, recognizing their shared goal of emulating human-like characteristics through computational methods. While not yet in widespread use, this concept aligns with themes explored in our paper [4] (see C) and the workshop to which it belongs (CHI24, workshop n.21). We employ this term broadly to encapsulate various approaches seen in role-playing agents [69], character embodiment with LLMs [55], generative agents [45], and other LLMs augmented with memory and personality characteristics [78, 31]. A unifying thread across these systems is the goal to create and simulate aspects of human memory and responses.

This section examines recent research into techniques for crafting believable and informative synthetic personae through the integration of memory systems, alongside their implications for work on LLM role-playing and HCI enhancement.

Shao et al. Character-LLMs introduce a similar idea to episodic memory datasets by infusing large language models with character-specific experiences, emotional states, and knowledge [55]. Their method involves reconstructing character data into detailed, first-person experiences and fine-tuning a model with these texts. This augmentation technique provides a foundation for fine-tuning LLMs to embody fictional or historical characters with enhanced believability. They evaluated their Character-LLMs against two Llama-derived 7B models and ChatGPT 3.5, giving the models simple character contexts for each answer. Findings showed that Character-LLMs outperformed the 7B models and performed similarly to ChatGPT 3.5. They used the LLM-as-a-judge system to evaluate their model, which we will discuss later, as we adopted the same evaluation methodology for our work.

While the Character-LLM approach offers significant advancements, it also has certain limitations. Firstly, fine-tuning each character implies a retraining process that can be both time-consuming and computationally expensive. Secondly, despite reducing hallucinations compared to other zero-shot models, the Character-LLMs still exhibited slightly higher hallucination scores than ChatGPT 3.5. Furthermore, they based their augmentation on Wikipedia pages, which limits the depth of character knowledge. They acknowledge this caveat and propose that richer biographical data should be used for future research, which is part of our motivation for using our source data.

Park et al.'s Generative Agents [45] demonstrate the power of integrating memory, planning, and reflection mechanisms within computational agents. Their system allows agents to store experiences in natural language, retrieve memories based on relevance, and form nuanced decisions, showcasing how AI can simulate human-like behaviour.

Park et al. employ a rigorous evaluation strategy to assess the capabilities of the generative agents. In a controlled evaluation, agents are interviewed by humans using natural language queries to probe their memory recall, planning, reactions, and reflections. This method allows researchers to analyze the agents' individual behaviour and understanding of their experiences. The end-to-end evaluation examines the agents' emergent social behaviours over two simulated days, focusing on information spread, relationship formation, and agent coordination. The results demonstrate that generative agents can maintain consistent personalities, recall past events, and plan actions meaningfully. They also exhibit complex social behaviours, successfully spreading information, forming connections, and collaboratively organizing events. However, the evaluation also reveals limitations, including memory retrieval failures, occasional embellishments of events, and overreliance on formal language inherited from the language model.

This work provides insights into the design and implementation of memory-rich LLM-based chatbots. Their use of retrieval augmentation to assess the relevance of memories further underscores the value of RAG techniques when building synthetic personae.

Wang et al.'s RoleLLM provides a framework for benchmarking and enhancing role-playing abilities, emphasizing the value of in-prompt character descriptions to guide the LLM's performance [69]. Notably, they mention that explicitly telling the character to copy direct sentences from the context prompts yields richer responses. Similarly, Li et al. highlight the potential of supplementing character personality cues within prompts to improve role-playing consistency [34]. Both studies incorporate the use of retrieval mechanisms to enhance the LLM's responses. Nevertheless, the two studies differ in their retrieval focus. Wang et al. center on retrieving relevant role-playing context, which is critical for maintaining a consistent character portrayal. On the other hand, Li et al. prioritize retrieving conversational history, ensuring the chatbot's responses remain coherent within the ongoing dialogue. This distinction demonstrates the versatility of retrieval methods within LLM role-playing applications. The ability to integrate both character-specific knowledge and recent conversational context could significantly advance the development of engaging and consistent role-playing interactions. We included both approaches in our work.

This section underscores the significant role that memory systems play in constructing believable and engaging synthetic personae. Integrating episodic memory, retrieval augmentation, reflection, and knowledge sources has the potential to enhance the role-playing capabilities of LLM-based chatbots, reduce hallucinations, and personalize interactions. Insights gained from this analysis directly informed our work.

2.7 Evaluation Methods for Retrieval Augmented Generation Systems

The evaluation of RAG systems presents unique challenges. Gao et al. [16] point out that while several proposed metrics for evaluating RAG models exist, a unified and standardized approach is still lacking. Traditional metrics surveyed by them, such as accuracy, Exact Match, and F1 scores, largely focus on assessing the factual correctness of outputs in relation to ground truth. Alternatively, metrics like BLEU and ROUGE compare generated text against expected responses. As Gao et al. document, these metrics hold value, and RAG approaches improve them. However, they are limited in capturing the subjective nuances crucial for characterizing embodied agents. This is because these metrics measure more direct word similarities. Existing evaluation methods may fail to address the embedded "subjectivity" central to embodied character representation when exploring HCI-centric, character-specific traits.

2.7.1 LLM-as-a-Judge

Zheng et al. [79] propose using Large Language Models (LLMs) as evaluators, termed "LLM-as-a-Judge", to assess the alignment of conversational AI systems with human preferences. Their findings show that using an LLM as an evaluator can achieve an agreement rate with human judgments exceeding 80%, matching the consistency typically seen among human evaluators. Furthermore, they provide guidelines to mitigate potential biases within the framework, such as:

- Position Bias: The tendency for LLMs to favor responses presented first.
- Self-Enhancement Bias: LLMs may overestimate the quality of their own responses.
- Verbosity Bias: LLMs may prefer longer, more detailed responses, even if they are not more informative.

Zheng et al. emphasize that these tests are not meant to replace human testing, but can aid in preliminary research and hybrid approaches.

2.7.2 Retrieval Augmented Generation Assessment (RAGAs)

Es et al. [15] propose the RAGAs framework, which addresses the challenge of evaluating RAG systems reference-free. The RAGAS framework offers a suite of metrics for assessing various dimensions of RAG architectures without relying on ground-truth human annotations, thus facilitating faster evaluation cycles. Excluding traditional metrics, the authors propose testing three aspects:

Context Relevance: Measures the RAG's effectiveness in identifying pertinent and concise context passages. An LLM is asked to extract sentences from the extracted context that can directly answer the posed question. Then the score is calculated by calculating the amount of relevant content in the context by using the following formula:

$$CR = \frac{\#\text{ of extracted sentences}}{\text{total number of sentences in the context}}$$

Answer Relevance: Measures the LLM's ability to provide good answers based on the given context. This is calculated by asking an LLM to create questions based on the RAG model's answers and then calculating the Cosine Similarity between the created and original queries:

$$AR = \frac{1}{n}\sum_{i=1}^{n} \text{sim}(q, q_i)$$

- **q and qi** refer to the two queries (the created and original one), which, once embedded, are multi-dimensional vectors.
- **sim** refers to the cosine similarity between two vectors. The cosine similarity is simply the dot product of those two vectors. If the vector are unitary and have the same values, then the similarity will return a 1, and if they are perpendicular to each other, the similarity will return 0.
- **n** refers to the total number of queries.

Faithfulness: Measures the LLM's ability to utilize the retrieved passages faithfully. First, an LLM generates statements based on the questions and answers provided. Subsequently, a separate LLM instance classifies these statements as either supported or unsupported by the original context. This approach directly evaluates the generative LLM's capability to remain faithful to the source material, avoiding distortions or introducing extraneous information. The final score for this metric is calculated as the average percentage of statements classified as supported by the context, indicating how much the LLM adheres to the retrieved passages. This score provides us with a way to know how useful the retrieved information is.

In order to compare the responses to the epistolary and biographical data in a more holistic way, we opted for the LLM-as-a-judge approach to address character subjectivity in the evaluation. Additionally, the RAGAs framework was chosen for its reference-free evaluation. In other words, it doesn't need expected answers to compare to, which is crucial for a Van Gogh bot as there aren't any documented interviews (question/answer format). These two methodologies offer the potential to quantify subtle dimensions of persona-driven LLMs, providing deeper insights into the model's ability to maintain consistency and believability within the synthetic personae framework.

Chapter 3

Autonoetic Memory Dataset

3.1 Data Augmentation

3.1.1 Data Sources

The base for our proposed cognitive model is an autonoetic memory dataset that contains Vincent van Gogh's experiences throughout his life. We used two data sources to ensure accuracy and depth.

Data Source 1: Biography

The primary source for our dataset was Steven Naifeh and Gregory White's critically acclaimed biography of Van Gogh, Van Gogh: The Life (2011) [44]. This comprehensive work, chosen for the authors' extensive reputation as biographers, covers Van Gogh's entire life (1853 - 1890). To create the biography, the authors worked with the Van Gogh Museum in Amsterdam and had access to unpublished source materials.

The biographical text was segmented into approximately 912, 3000-character chunks (approximately 3 paragraphs). This segmentation length was determined to capture enough contextual richness while allowing for episodic granularity. Smaller segments might not provide sufficient context, while larger ones might hide specific experiences within broader narratives.

It is necessary to acknowledge that some of the information in the biography might be exaggerated, fictionalized or simply incorrect. In other words, a biography might introduce biases into the data.

Data Source 2: Letters

We also used Van Gogh's famous correspondence to complement the biographical perspective with Van Gogh's own voice. The letters were scrapped from the Van Gogh Museum's official website [18]. This collection comprises 863 letters written both by and to Van Gogh, offering unique insights into his thoughts, emotions, and lived experiences. The letters were embedded without chunking, due to their smaller size.

The choice to use both a detailed biography and Van Gogh's letters was intentional. The biography offers a chronological structure of important life events, placing Van Gogh's experiences within their historical context. The letters reveal his inner world, expressing his emotions, relationships, and artistic goals. This combined approach aims to create a dataset with factual events and direct insights into Van Gogh's understanding of those experiences.

3.1.2 Limitations of the Raw Data

While Van Gogh's biography and letters offer rich information for understanding his life, their direct usage presents several technical complications.

A key issue lies in the use of third-person narration within the biographical text. Initial experiments with simple RAG implementations revealed a tendency for the system to intermittently adopt a third-person perspective, even when contextual prompts explicitly discouraged this. Here's an example of this problematic behaviour from one of our preliminary conversations:

Preliminary Chat with RAG system

Question:
Did you live with Gaugin?
Model (GPT-3.5):
Ah, Gauguin, a fellow artist with whom I shared a brief but intense period of collaboration and companionship. Yes, indeed, we lived together for a time in Arles, (...) **The famous incident where he famously cut off a part of his own ear occurred during one such moment of heightened emotions. Ultimately, Gauguin and Van Gogh parted ways, each following their own artistic paths**. (...)

This inconsistency undermines the goal of constructing a first-person synthetic persona. Furthermore, our survey of existing research on self-reflection for hallucination reduction underscores the necessity of first-person thought structures. Therefore, we required a preprocessing stage to address these issues and create our autonoetic dataset.

On the other hand, Van Gogh's letters exemplify the challenges of harnessing unstructured personal narratives. Even though they represent Van Gogh's world view and contain many examples of raw emotion, insightful reflections, and valuable data points (people, places, events), their fragmented nature and frequent lack of explicit context create obstacles for direct utilization in knowledge-driven systems.

Letter 50.07 exemplifies this:

Extract of Letter 50.07

Dear Father and Mother,

You have probably received my telegram, but you will be glad to hear some more particulars. On the train, I wrote down a few things, and I am sending them to you so that you will know all about my journey.

Friday. I thought, we will stay together today. Which do you think is better ... the joy of meeting or the sorrow of parting? We have often parted already; this time there was more sorrow in it than there used to be, but also more courage because of the firmer hope, the stronger desire, for God's blessing. And didn't nature seem toshare our feelings, everything looked so grey and dull a few hours ago.Now I am looking across the vast expanse of meadows, and everything is very quiet; the sun is disappearing again behind the grey clouds but sheds a golden light over the fields. These first hours after our parting – which you are spending in church, and I at the station and on the train – how we are longing for each other and how we think of the others, of Theo and Anna and the other little sisters and the brother. Just now we passed Zevenbergen; I thought of the day you took me there, 1 and I stood on the steps at Mr. Provily's, looking after your carriage on the wet road, and then of that evening when my father came to visit me for the first time. And of that first homecoming at Christmas!

Saturday and Sunday. On the steamer I thought often of Anna – everything reminded me of our journey together. The weather was clear, and the river was especially beautiful, and also the view, seen from the sea, of the dunes, dazzling white in the sun. The last I saw of Holland was a little grey church spire. I stayed on deck until sunset, but then it became too cold and rough.At dawn the next morning on the train from Harwich to London it was beautiful to see the black fields and green meadows with sheep and lambs and an occasional thornbush and a few large oak trees with dark twigs and grey moss-covered trunks; the shimmering blue sky with a few stars still, and a bank of greyclouds at the horizon. Before sunrise I had already heard the lark. When we were near the last station befor London, the sun rose. The bank of grey clouds had disappeared and there was the sun, as simple and grandas ever I saw it, a real Easter sun. The grass sparkled with dew and night frost. But I still prefer that grey hour when we parted.

Saturday afternoon I stayed on deck till the sun had set. The water was fairly dark blue with rather high white-crested waves as far as one could see. The coast had already disappeared from sight. The sky was one vast light blue, without a single little cloud. And the sunset cast a streak of glittering light on the water. It was indeed a grand and majestic sight, but still the simpler, quieter things touch one so much moredeeply. The train for Ramsgate left two hours after I arrived in London. That is still about four and a half hours by train. It is a beautiful route; for instance, we passed one part that was quite hilly. At the base the hills are covered with scanty grass, and at the top, with oak woods. It reminded me of our dunes. Between the hills was a village with a grey church overgrown with ivy like most of the houses. The orchards were in full bloom and the sky was a light blue with grey and white clouds.(...)

This letter captures a turning point in Van Gogh's life as he leaves his family home. He mentions important figures like his siblings, teachers, and parents. The letter also reveals strong emotions, demonstrating his nostalgia for home. However, these critical details are intertwined with lengthy descriptions of landscapes and personal stories. While this reflects Van Gogh's unique writing style, it can hinder LLM processing. Our approach aims to extract the essential facts and emotions from such passages while creating memories with Van Gogh's expressive language. This combined approach allows us to build a comprehensive episodic memory dataset that accurately represents both the factual and emotional aspects of his life.

3.1.3 Text Rewriting: Script Generation

We leveraged the capabilities of LLMs in a multi-step process to transform the raw biographical and correspondence material into an autonoetic memory format. The entire pipeline is shown in Fig.3.1.

Since LLMs often exhibit enhanced performance within role-playing scenarios [69], we tasked a "Screenwriter" LLM with rewriting the biographical text and letters into scenes from a hypothetical movie script centred on Van Gogh's life. This approach aimed to transform the source material to better isolate elements essential for autonoetic memory formation. Specifically, we transformed the data to contain the following:

- Narrator Information: Provides broader context for specific events.
- Scene Background Information: Establishes location and temporal setting.
- Van Gogh Voiceover: Offers first-person narration from Van Gogh's perspective.

Each biographical chunk and letter was individually processed through this script generation process. See A.1 for the full prompt used and A.2 for a sample response.

This process resulted in 1774 augmented text samples, each representing a movie scene description.

This LLM-driven text transformation serves several purposes. Firstly, structuring the source material into discrete scenes aims to provide the model with specific experiences and their associated context. Additionally, the script format facilitates the extraction of first-person narration, which is crucial for building a cognitive model that embodies Van Gogh's viewpoint. Lastly, introducing hypothetical narrative elements could enrich the dataset, encouraging the model to reason about plausible scenarios aligned with Van Gogh's known experiences, personality, and artistic style. In other words, we are applying an offline version of self-reflection.

3.1.4 LLM-Driven Information Extraction and Structuring

We employed several other LLMs to augment, process and extract relevant information from the LLM-generated movie scenes.

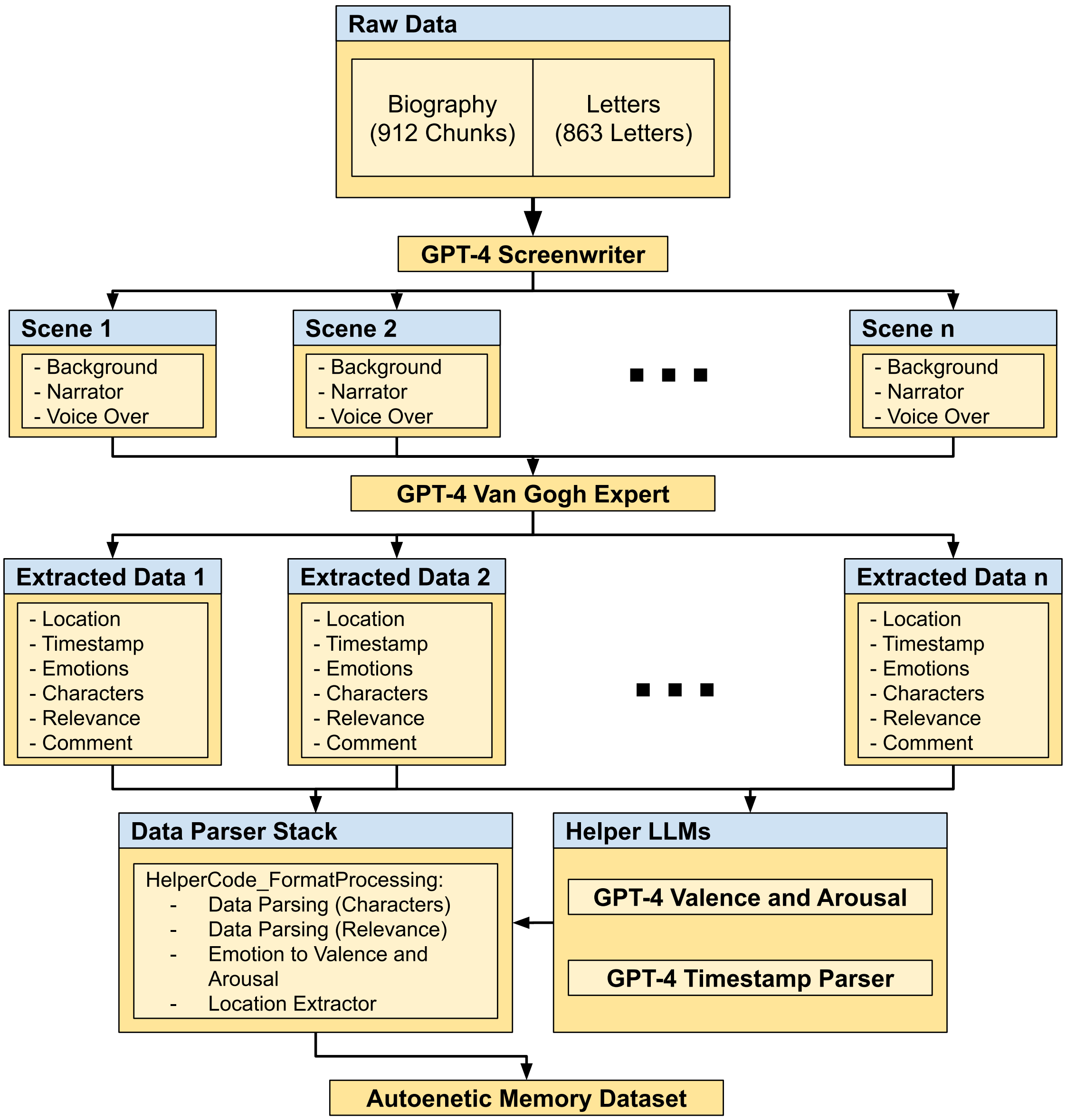

Figure 3.1: Autonoetic Dataset Pipeline.

- **"Van Gogh Expert" LLM (VGE-LLM):** A GPT-4-Turbo model instance was asked to role-play as a Van Gogh expert and tasked with dissecting each scene and identifying and extracting certain elements. The VGE-LLM was designed to augment memory quantitatively. It identified numerical data within memories, allowing for their assessment using numerical metrics. Additionally, it extracted the emotions attached to each memory, which are vital additions to the subjective, autonoetic experience. See A.3 for the query used, and A.4 for a sample response. The generalization of the VGE-LLM to other characters involves the adaptation of the static prompt. The underlying principles of static prompt design and role-playing as an expert remain consistent, allowing the system to be effectively repurposed.

- **GPT4 Valence and Arousal LLM:** Valence represents the intrinsic attractiveness or averseness of an experience, ranging from negative (unpleasant) to positive (pleasant). In contrast, arousal indicates the degree of physiological activation associated with an experience, ranging from low (dull) to high (exciting) [22]. We created a python script to extract a set of all different emotions annotated by the VGE-LLM, and then tasked an LLM to give numerical Valence and Arousal values in a range (-1,1) for each individual emotion in the set.

- **GPT4 Timestamp Parser LLM:** Since many locations were written in different ways in the VGE-LLM notes, we tasked an LLM with extracting the dates in the notes and format them in a dd/mm/yyyy format.

3.1.5 Data Enrichment and Formatting

It is important to note that the LLMs generated some date errors and non-uniform data formats. These inconsistencies highlight that LLMs, while powerful tools, can introduce variability in data processing tasks, emphasizing the need for careful verification and potential manual and automated adjustments to ensure the integrity of the dataset.

Regarding the date errors, we manually identified and corrected 14 instances of date hallucination, where the model assigned dates to memories when Van Gogh was not alive (e.g., 2010).

Regarding the data formats, non-uniformity means not all generated data was formatted the same, and more importantly, the output didn't always adhere to the schema that was instructed. For example, a location for a memory might have been generated as "Location: The Hague, Netherlands" (comma as a separator) and for another memory as "Location: The Hague in the Netherlands" (the word "in" as a separator). This introduced a degree of variability into the process. To handle these inconsistencies, we took an automated approach, creating a jupyter/python script to automatically parse and separate the extracted data by label (Timestamp, Emotions, etc.). We used several RegEx techniques to extract the data. The script can be found in this repository, and its split into the following sections:

HelperCode__FormatProcessing.ipynb

- Data Parsing (Characters): This section normalizes Characters listed by the VGE-LLM by removing inconsistencies and noise and creates a python list with all the Characters in each memory.
- Data Parsing (Relevance): This section isolates the Relevance Score assigned by the VGE-LLM and converts it to a single integer value.
- Emotion to Valence and Arousal: This section extracts all emotions listed by the VGE-LLM and uses the data created by the Valence and Arousal LLM to convert all emotions to numerical values. It then averages all values and gets a Valence and Arousal value per memory.
- Location Extractor: This section takes the locations extracted by the VGE-LLM and uses the OpenCageData API [17] to get geocode latitude and longitude coordinates per location.

While specialized Named Entity Recognition (NER) models offer distinct advantages for tasks such as date extraction and ensuring uniform data formatting [30], our decision not to employ them in this thesis was based on two key considerations:

1. Contextual Understanding: Our project required a deep understanding of the context in which data elements were embedded. The Large Language Models (LLMs) used were specifically tasked with role-playing as Van Gogh experts, dissecting scenes, and identifying elements within a rich narrative context. This level of contextual understanding is often beyond the capabilities of standard NER models, which focus primarily on entity extraction without deeper semantic comprehension.
2. Subjective and Autonoetic Experience: The project aimed to extract not only factual data but also the emotions attached to memories, which are vital for a comprehensive subjective and autonoetic experience. LLMs, with their advanced language understanding capabilities, were better suited for this task as they could interpret and articulate the nuanced emotional content associated with each memory.

3.1.6 Final Dataset Structure

The resulting dataset shown in Table 3.1 comprises several columns designed to enable diverse forms of retrieval and analysis.

We refer to this dataset as an "autonoetic memory dataset" as it embodies first-person recollections of all of Van Gogh's life, enriched with contextual detail from the VGE-LLM and numerical values to represent the emotional and geographical landscape of Van Gogh's life experiences.

Field	Description
UID	Unique identifier indicating the memory's origin (chapter, chunk, or letter).
Background	Scene background descriptions extracted from the movie script.
Narrator	Scene Narrator descriptions extracted from the movie script.
Van Gogh	Van Gogh autonoetic Voice Overs extracted from the movie script.
Context	A basic scene context provided by the Van Gogh expert LLM.
Comment	Insights provided by the Van Gogh expert LLM.
Characters	Individuals involved in the scene, including those mentioned. Extracted by the Van Gogh Expert LLM and refined using Python scripts.
Valence	The average emotional valence of the scene. Extracted by an LLM and averaged using Python scripts.
Arousal	The average emotional arousal of the scene. Extracted by an LLM and averaged using Python scripts.
Timestamp	Chronological reference point. Extracted by the Van Gogh expert and converted to Pandas Timestamp format with Python.
Latitude	Geographical location of the scene. Extracted by the Van Gogh expert LLM and converted using a geocoding API.
Longitude	Geographical location of the scene. Extracted by the Van Gogh expert LLM and converted using a geocoding API.
Relevance	The scene's significance in Van Gogh's life, as evaluated by the Van Gogh expert LLM.
Additional Exploration (Unused)	Initial exploration of associated Van Gogh's paintings with memories, but this data was not used in the final model.

Table 3.1: Description of fields in the autonoetic memory dataset.

Chapter 4

Chatbot Architecture and Implementation

4.1 System Architecture

The chatbot architecture aims to seamlessly integrate our custom-built autonoetic memory dataset with two parallel Retrieval Augmented Generation (RAG) systems. This design prioritizes achieving a balance between the speed of naive retrieval systems and the adaptability inherent in LLMs, enabling the chatbot to generate simultaneously fast, accurate, and reliable responses.

4.1.1 RagChain API

We developed an open-source general version of the code to facilitate future research. The code can be found at: this repository.

To maximize flexibility and facilitate future experimentation, the chatbot system adopts a highly modular architecture consisting of four primary components:

Embedder Module: This module is responsible for several key functions:

- Embedding: It transforms textual data (memory entries and user queries) into numerical vector representations.
- Similarity Calculation: It uses cosine similarity to quantify the relatedness between query embeddings and memory embeddings efficiently.
- Retrieval: Returns the most relevant memory entries based on similarity scores.

Chatbot Module: This core module houses the LLM API calls and coordinates the chatbot's actions:

- Prompt Reception: Processes incoming user prompts.
- Question Asking: Handles the queries to the model and the reception of the results.

- Result Presentation: Compiles and presents the retrieved information to the user as a CMD/terminal stream output.

RAGchain Module: Orchestrates the multi-step retrieval and generation process with the following responsibilities:

- Similarity Search: Executes similarity-based search operations across multiple RAGs to identify the most relevant information fragments.
- Chained Prompt Generation: Constructs a composite prompt by concatenating the top results from each RAG, providing a multifaceted input to the LLM.
- RAG Customization: Allows for individual RAGs to be designated with unique names and parameters, enabling more fine-grained control within the retrieval and generation process.

This system effectively allows a user to use any number of datasets and chain them in any particular order they want. Its modular structure offers several advantages. It allows for the seamless replacement of individual components—such as the LLM, embedding model, or prompt generation logic—without requiring significant rework of the entire system. This flexibility encourages rapid prototyping and comparative analysis during the research and development process.

The iterative development of our chatbot and its underlying research led to the creation of several key modules designed to optimize performance and enable flexibility in experimentation.

Embedder Modules

- GenericEmbedder: This module leverages Hugging Face's Transformers library [73] to access and utilize a wide range of embedding models. This design choice ensures compatibility with both publicly available models and potential future custom-designed embedding models.
- OpenAIEmbedder: Integrates directly with OpenAI's API, enabling seamless access to their suite of embedding models and providing a streamlined way to access their computational resources.

Chatbot Modules

- Mistral7B Chatbot: Developed as an early-stage prototype, this chatbot was designed to operate with smaller, locally trained language models. While adherence to OpenAI's API conventions proved problematic, it served as a valuable proof of concept.

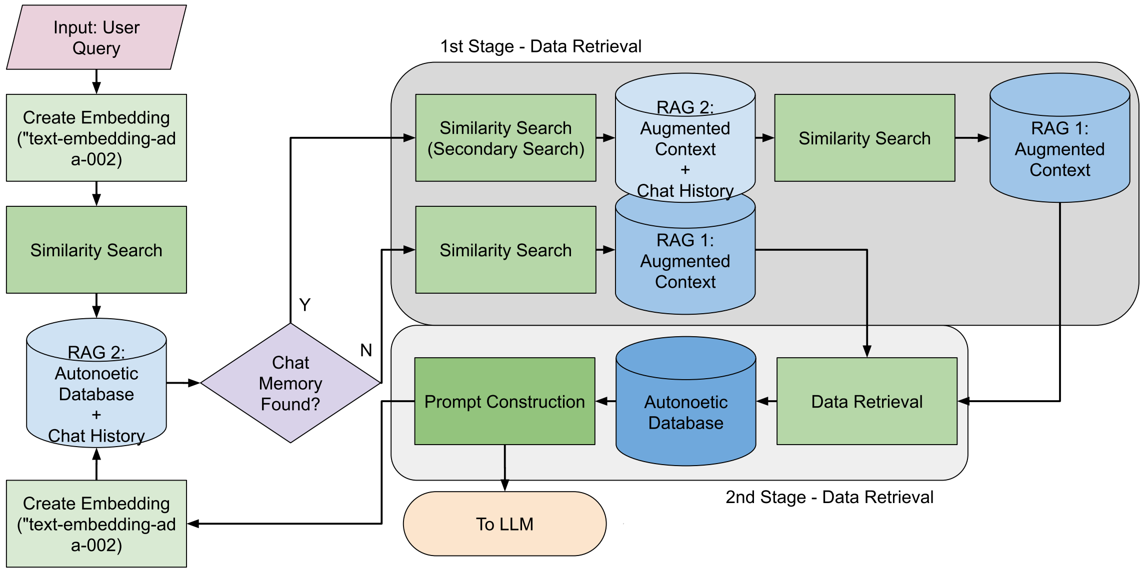

Figure 4.1: Flowchart of the Van Gogh Chatbot Design

- OpenAI Chatbot: The final chatbot implementation aligns with OpenAI's API conventions. This design decision facilitates future development and experimentation by ensuring compatibility with various OpenAI and non-OpenAI models (potentially including the earlier Mistral7B implementation with minor API adjustments).

RAGChain Module

GenericRagChain: This module coordinates the operations of multiple RAG instances. It merges and formats the final prompts with the retrieved data from each RAG.

4.1.2 Van Gogh Chatbot: Architecture and Design Considerations

The overall system architecture of the Van Gogh chatbot, as illustrated in Fig. 4.1, is designed to optimize the interaction between the user, the large language model (LLM), and the autonoetic memory dataset. Several key modifications were implemented to ensure the chatbot effectively leverages the unique characteristics of the episodic memory structure and facilitates natural, contextually relevant conversation.

In-Loop RAG Instantiation and Dual Indexing: The core of the system lies in the dynamic creation of a Retrieval-Augmented Generation (RAG) object within the main loop. This RAG object then constructs two separate indexes, RAG1 and RAG2, both utilizing the same autonoetic memory dataset. The purpose of this dual indexing strategy is to enable a multi-stage retrieval process that prioritizes both recent conversational context and broader thematic relevance. Additionally, a RAGchain object is instantiated, encapsulating a single RAG object for streamlined retrieval and generation.

Chat History Embedding and Multi-Stage Retrieval: A crucial aspect of maintaining conversational coherence is the incorporation of chat history. After each user query and corresponding chatbot response, both are embedded into a single vector representation and stored within an embeddings object. Subsequent user queries trigger a two-stage retrieval process:

- Primary Search: Both RAG indexes are queried to identify autonoetic memories that are semantically similar to the current user input.
- Secondary Search (Conditional): If the primary search yields chat history memories deemed relevant to the current query, a secondary search is conducted on RAG2. This secondary search uses the most relevant chat history memory as a query, aiming to retrieve additional autonoetic memories that align with the ongoing conversational thread.

Retrieval and Chaining Limits: To maintain query efficiency and accommodate potential future LLM deployments with limited resources, the system imposes a limit of three retrieved memories per query. The specific chaining logic depends on the presence or absence of relevant chat history:

- Chat History Found: The retrieved memories include one chat history memory, one autonoetic memory related to that chat history, and one standalone autonoetic memory relevant to the current query.
- Chat History Not Found: In the absence of relevant chat history, three autonoetic memories are retrieved based solely on their relevance to the current user query.

This adaptive retrieval and chaining strategy ensures that the chatbot's responses are grounded in the factual details of Van Gogh's life and dynamically responsive to the conversation's evolving context.

Embedding and Storage Strategy

The "augmented_context" column (containing the original context, date and characters in the memory) of the dataset is used to create the embeddings, as it offers the most concise description of each memory. We used OpenAI's "text-embedding-ada-002" model to create the embedding representations. To minimize costs, embeddings for the RAG indexes are calculated in a single pass and persisted in a pickle file for subsequent loading. During live chatbot interactions, two additional OpenAI calls are made per query: one for embedding the single query (enabling similarity search and retrieval) and one for embedding the query-response pair for subsequent storing in the RAG2 index.

Prompt Token Limits

As mentioned in Chapter 1, a 2k token limit was given to the full prompts, with the allocations shown in Table 4.1.

System Component	**Token Limit**
Static Context	300 tokens
Retrieved Memories	600 tokens
Numerical Data	100 tokens
Chat History	500 tokens
Response	500 tokens

Table 4.1: Token limitations for different components within the system

Our decision to employ conservative limits, even with the capabilities of GPT-4, reflects our forward-looking strategy focused on enabling future deployment on smaller LLM models with context windows in the 1,000 - 2,000 token range. The restriction to 3 retrieved memories aligns directly with these token budget constraints, ensuring compatibility with more resource-limited systems.

The approach of limiting tokens holds several key benefits. First, it facilitates broader access to these systems, potentially empowering students and teachers to engage with this technology in classroom settings without requiring specialized hardware. Second, it promotes the development of AI applications that can be embedded within personal devices, expanding the potential reach and impact of HCI innovations. Finally, it aligns with our commitment to sustainable AI development, fostering research that prioritizes efficiency and responsible use of computational resources.

4.1.3 Autonoetic Memory: Two-Stage Access and Information Retrieval

A crucial aspect of our architecture is the two-stage retrieval process employed to access the autonoetic memory. Fig.4.2 illustrates the retrieval process. Initially, cosine similarity is calculated between the user's query and the dataset's embedded "context" column (which offers concise descriptions of events). However, rather than directly utilizing the results of this similarity search, we leverage them as pointers into the autonoetic memory. This two-stage access enables us to retrieve specific information, such as Van Gogh's first-person voice-over narration, emotional valence and arousal, and the names of people involved in the memory. In summary, this strategy allows us to optimize the similarity search using the "augmented_context" column for retrieval, which aligns more closely with potential user queries while ultimately extracting the autonoetic memory content which contains the first-person and numeric information.

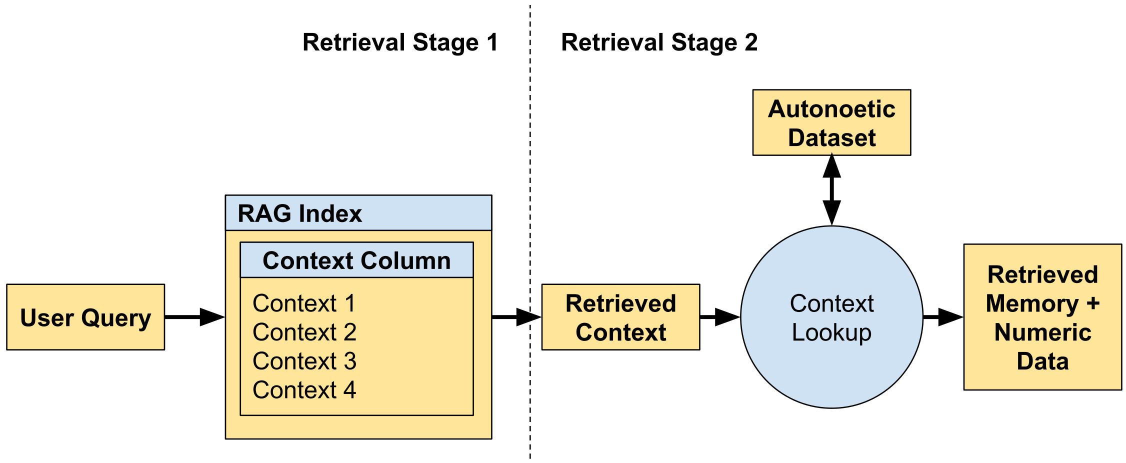

Figure 4.2: Autonoetic Dataset Pipeline.

Emulating Memory Types

Our system design effectively emulates three types of memory:

- Short-term Memory: This is simply the chat history directly captured within the context window. Each user query and LLM response is appended to a chat history array (following OpenAI API conventions). The length of this short-term memory is dynamic (approximately 500 tokens) and is influenced by the length of retrieved memories, the length of the final prompt, and the imposed retrieval limitations discussed before.

- Intermediate-term Memory (Working Memory): We consider this type of memory to be similar to the working memory since it gives the LLM focalized short-term memories of the current conversation to actively use for response formulation. To simulate intermediate-term recall, we embed query/response pairs into the RAG2 index containing the same letter and biographical data as the first one. This decision ensures embeddings reside within the same representational space so they keep semantic coherence. During user queries, the system does similarity search with the user's query for related chat query/response pairs. This mechanism allows the system to recall elements from previous moments in the conversation, even those outside the immediate context window, provided the chatbot has not been reset.

- Long-term Memory: Our autonoetic memory dataset constitutes the system's long-term memory. Each memory within the dataset is represented in both RAGs through the dataset's "augmented_context" column, combining the original context with date and character metadata. This column augmentation is crucial, as it provides the LLM with explicit cues for grounding retrieved memories in a spatiotemporal and social context. The two-stage retrieval process mentioned above then extracts the autonoetic-style memories and numerical data for the LLM to use.

Prompt Construction and Memory Integration

Our two-stage retrieval allows the system to construct prompts grounded in first-person experiences and additional data while still leveraging the question-answering strengths of the RAG similarity search process.

Furthermore, we promote conversational coherence by allowing the system to "recall" both the chat history and associated autonoetic memories. This information is concatenated into the final prompt, with the number of retrieved memories limited to 3 to simulate scenarios with more restrictive attentional constraints.

Chapter 5

Data Analysis and Comparative Evaluation

5.1 Preliminary Explorations

To gain an initial understanding of the dataset's characteristics and potential insights, we engaged in several open-ended conversational sessions with the chatbot.

In total, the dataset includes 913 entries sourced from Van Gogh's biography, and 858 entries sourced from his letters. The presence of numeric data within our autonoetic memory dataset opens up opportunities for quantitative analysis alongside qualitative explorations.

5.2 Descriptive Statistics

A descriptive statistical analysis was conducted to understand the tendencies and variability within the dataset.

5.2.1 Valence and Arousal

The Valence and Arousal distributions are shown in Fig. 5.1. For valence, the mean stands at -0.09, indicating a slightly negative average across the dataset. The standard deviation is 0.45, showing a moderate spread of valence scores around the mean. The range of valence extends from a minimum of -0.97 to a maximum of 0.95, highlighting a broad spectrum of values. The quartile values provide deeper insights: the first quartile (25%) is at -0.43, the median (50%) is close to the mean at -0.08, and the third quartile (75%) rises to 0.2, suggesting that a majority of the data skews slightly towards lower valence scores.

On the other hand, arousal presents a mean of 0.4, indicating a generally positive average across the dataset. The standard deviation for arousal is 0.25, which indicates a tighter clustering of arousal scores compared to valence. The minimum arousal score is notably low at -0.65, whereas the maximum at 0.9 is close to the maximum possible value. Quartile values for arousal show the first quartile (25%) at -0.25, a median (50%) at 0.42, which is similar to

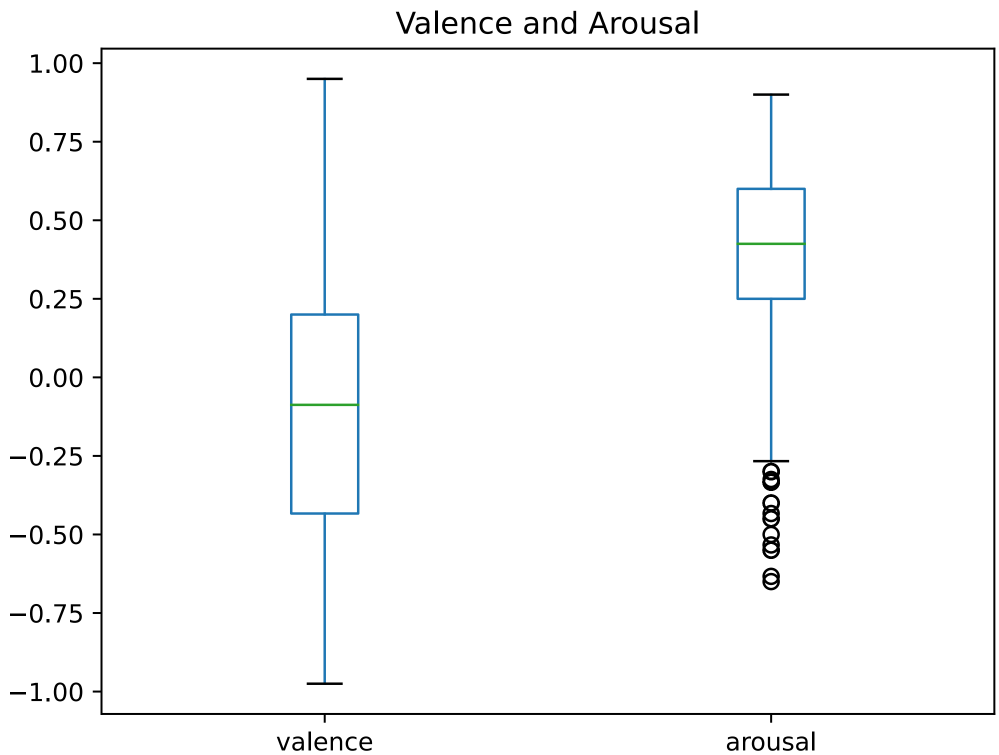

Figure 5.1: Valence and Arousal boxplots.

the mean, and a third quartile (75%) at 0.6. This distribution suggests more narrow range of arousal responses, with a significant number of observations gathered towards the higher end of the scale.

Furthermore, Fig. 5.2 shows the yearly average valence and arousal values across Van Gogh's life. For this graph, a Bayesian adjustment approach was applied to average valence and arousal data segmented by year to mitigate the disproportionate influence of years with sparse data. Specifically, each year's average valence and arousal were recalculated by incorporating a predefined constant that simulates a baseline number of observations. This pseudo count represents an a priori assumption about the minimum level of data reliability, regularizing the dataset by reducing the weight of years with few actual observations. This method acknowledges the inherent uncertainty in years with fewer data points, applying a form of Bayesian shrinkage that draws less specific estimates towards the global mean producing a more robust representation of trends over time. The time-series graph of valence scores reveals pronounced negative spikes in the years 1880, 1885, and 1890, which stands to reason as they align with known biographical milestones. 1880 marks the period when Van Gogh began his artistic path which was characterized by uncertainty, self-doubt, and financial challenges. In 1885, he experienced a period of profound artistic exploration and

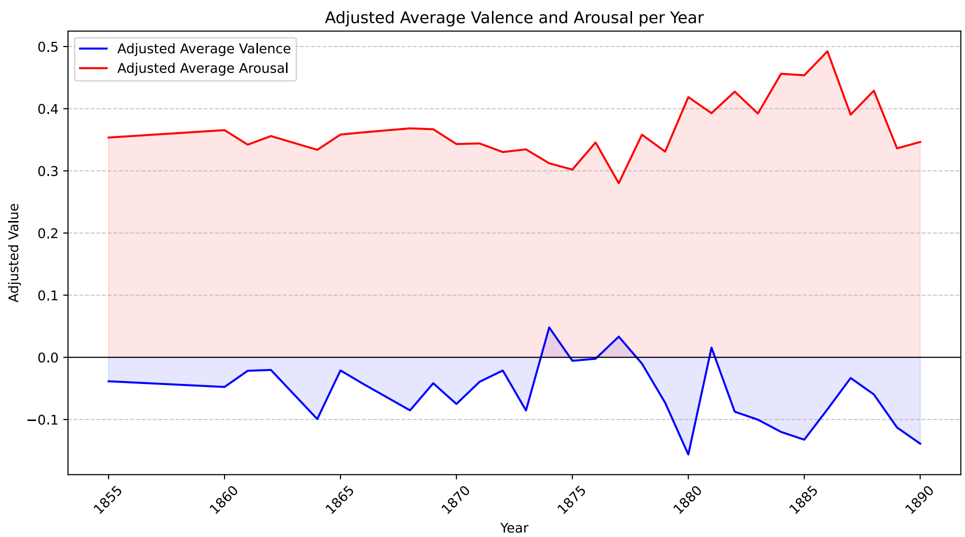

Figure 5.2: Yearly Weighted Average Valence and Arousal.

emotional intensity, culminating in the creation of "The Potato Eaters," amidst the struggle for recognition and personal unrest. Finally 1890 marks the year of his death. His final years were known to be turbulent and marked by mental health struggles (also shown by the decline in arousal scores during this period).

5.2.2 Characters

Quantitative analysis of character mentions within Van Gogh's autonoetic memories (Fig. 5.3) offers additional insights into his social world and the personal relationships that shaped his life. The high frequency of mentions of "Theo" underscores the deep bond between Vincent and his brother. Similarly, the high frequency of "Gauguin" highlights the influential relationship between them.

Fig. 5.4 shows a graph tracing mentions of Van Gogh's love interests over time. It provides a window into the emotional and relational aspects of his life. Spikes in mentions of specific names suggest periods of intense emotional focus, potentially tied to the blossoming or fading of romantic connections.

The fact that these frequency patterns align with known biographical facts serves as evidence that the autonoetic memory dataset captures meaningfully and chronologically aspects of Van Gogh's life.

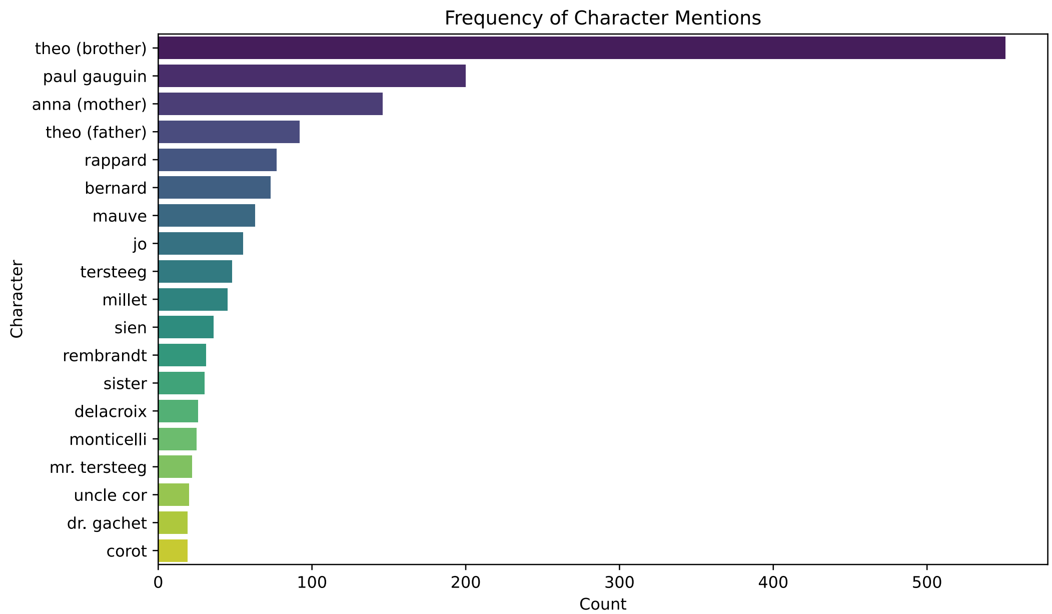

Figure 5.3: Character Mentions

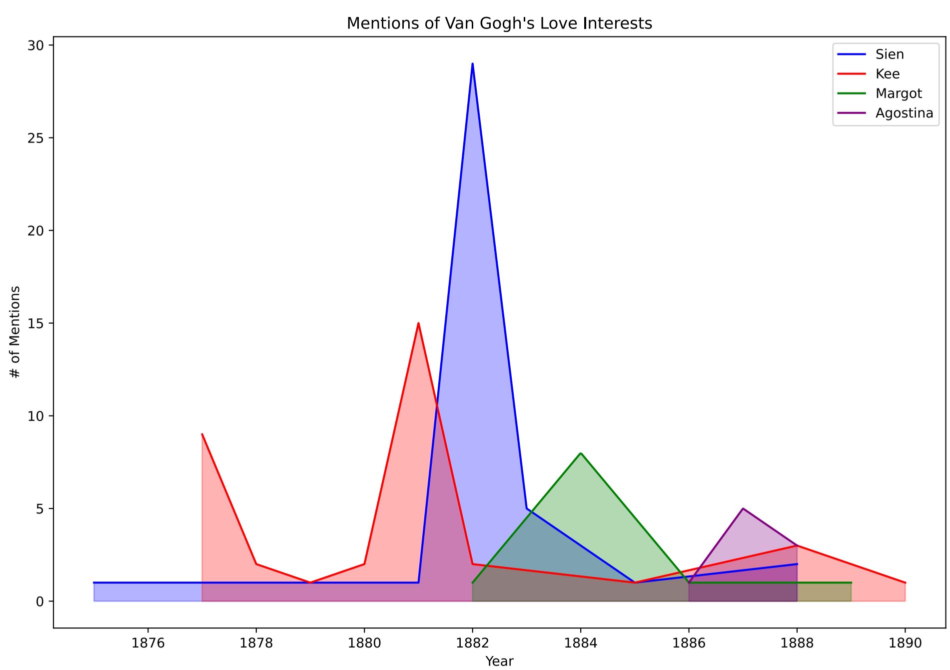

Figure 5.4: Mentions of Van Gogh's Love Interests Through Time.

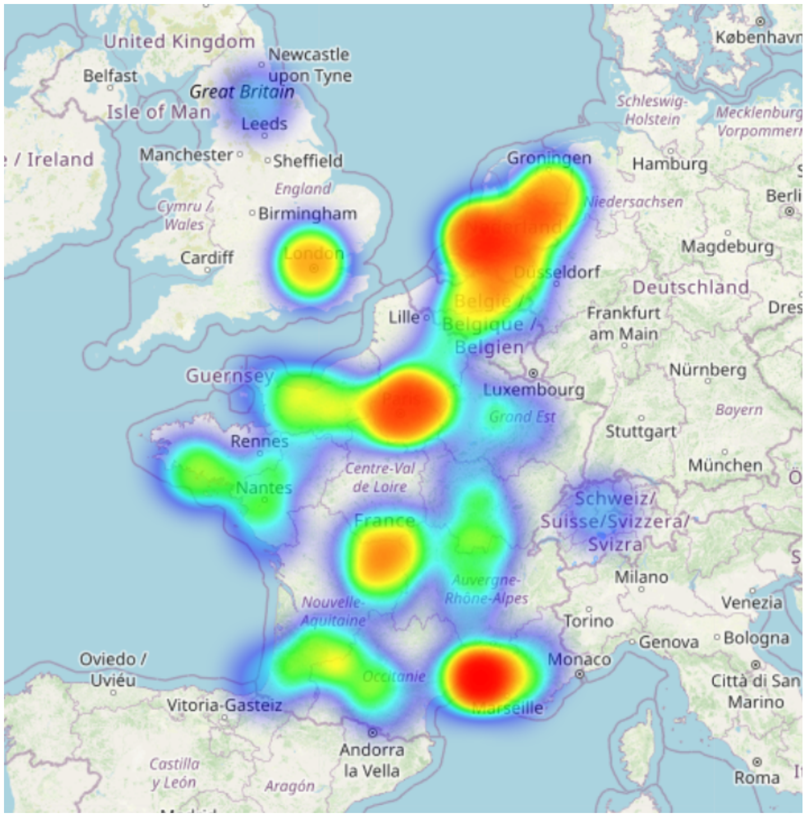

Figure 5.5: Screenshot of a heatmap with all memories.

5.3 Exploratory analysis

5.3.1 Interactive Graphs

To enhance our analysis of the cognitive dataset's multifaceted nature, we developed two interactive graphs.

Location Heatmap: We created an interactive world map visualizing the geographic locations associated with each episodic memory. This visual tool offers insights into where Van Gogh spent significant portions of his life. Furthermore, by filtering memories based on specific criteria, we can generate heatmaps for particular periods or themes. Figures 5.5 and 5.6 illustrate this concept, showing a heatmap of all memories alongside a focused visualization of a randomly selected subset.

These visualizations go beyond traditional biographical timelines, providing a unique spatial perspective on Van Gogh's life journey. By highlighting specific areas, they may reveal patterns that would otherwise remain hidden within textual data alone. We hope this approach contributes a powerful analytic tool for understanding how location shaped Van Gogh's perceptions and potentially his artistic expressions.

Memory Visualizer: This tool employs Principal Component Analysis (PCA) for dimensionality reduction, allowing visualization of the dataset in 2D space. For the "Charac-

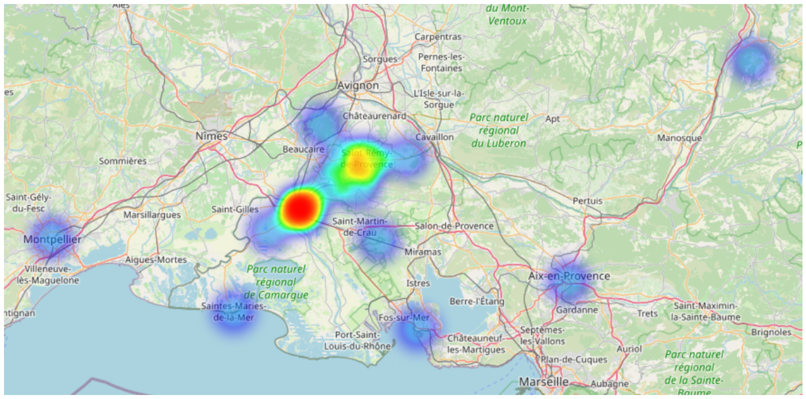

Figure 5.6: Screenshot of a heatmap with a randomly selected group of memories.

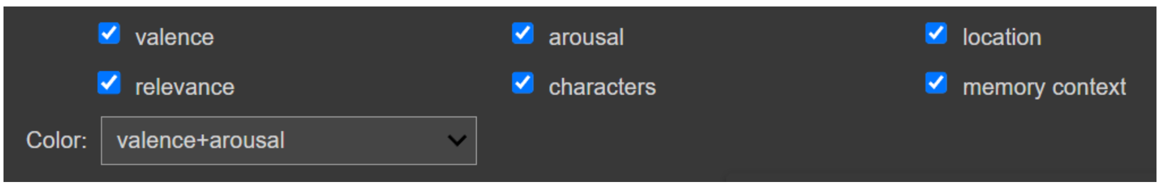

Figure 5.7: Screenshot of the provided UI for feature selection.

ters" and "Context" columns, we used the same embedding model as the chatbot to represent non-numerical data elements as embeddings of 1.5k dimensions. We then used another PCA analysis to reduce the dimensions to three each and added them to the dataset. This approach allowed us to integrate our dataset's textual columns (converted to numerical data) with the numerical columns. We created a simple UI (Fig. 5.7) to select which columns to use for the PCA and colour-coding of the memories.

An initial PCA of all dataset columns with valence and arousal color coding (Fig.5.8) reveals a general trend where memories with higher valence and lower arousal cluster on the lower-right side (yellow data points), whereas memories with lower valence and higher arousal cluster towards the upper-left side.

By focusing the interactive graph on valence and arousal values and colour-coding by valence (Fig. 5.9), we can observe distinct clusters of memories based on their valence levels. This allows us to easily isolate and analyze memories with high, medium, or low valence. There appears to be a lower concentration of memories with high valence (yellow), revealing a potential skew toward less positive emotional representation within the dataset, which supports our prior statistical analysis.

5.3.2 User Paths

Additionally, we implemented a mapping function to visualize the conversational path of a user interacting with the chatbot (Fig.5.10). As the user engages in dialogue, we record the unique identifiers (UIDs) of the autonoetic memories accessed by the system. We can chart

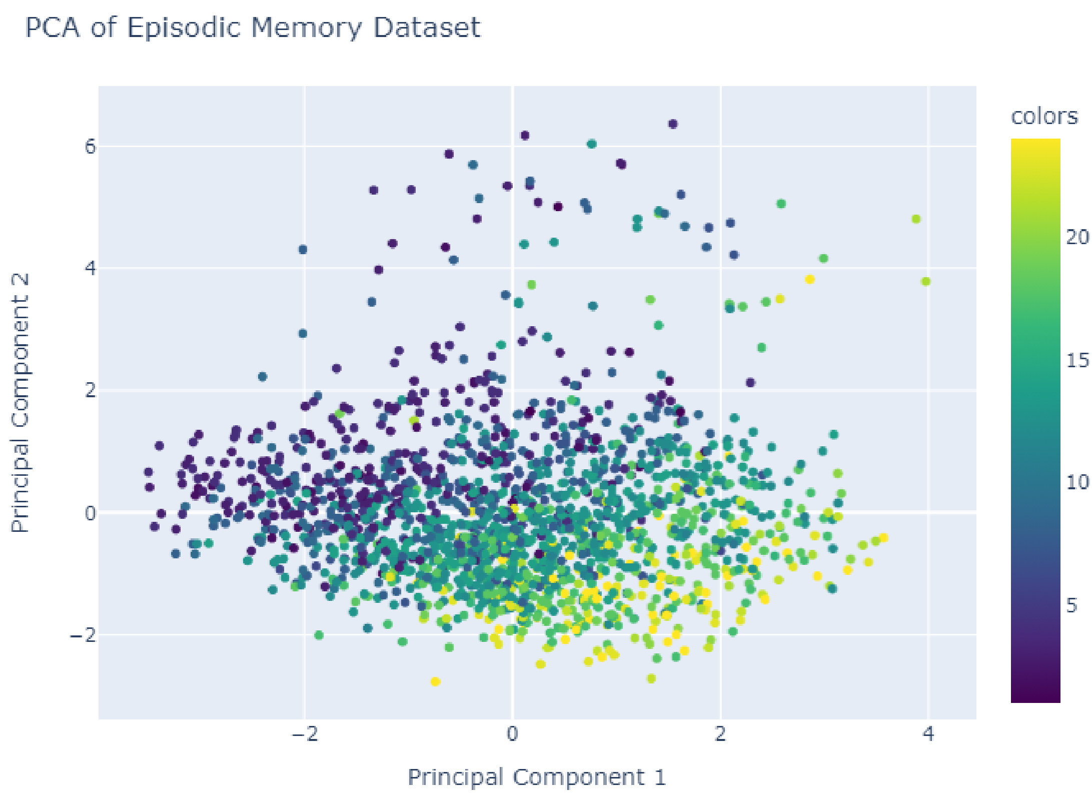

Figure 5.8: PCA of the Autonoetic Memory Dataset. Yellow memories have higher valence, and dark-blue memories decreasing valence.

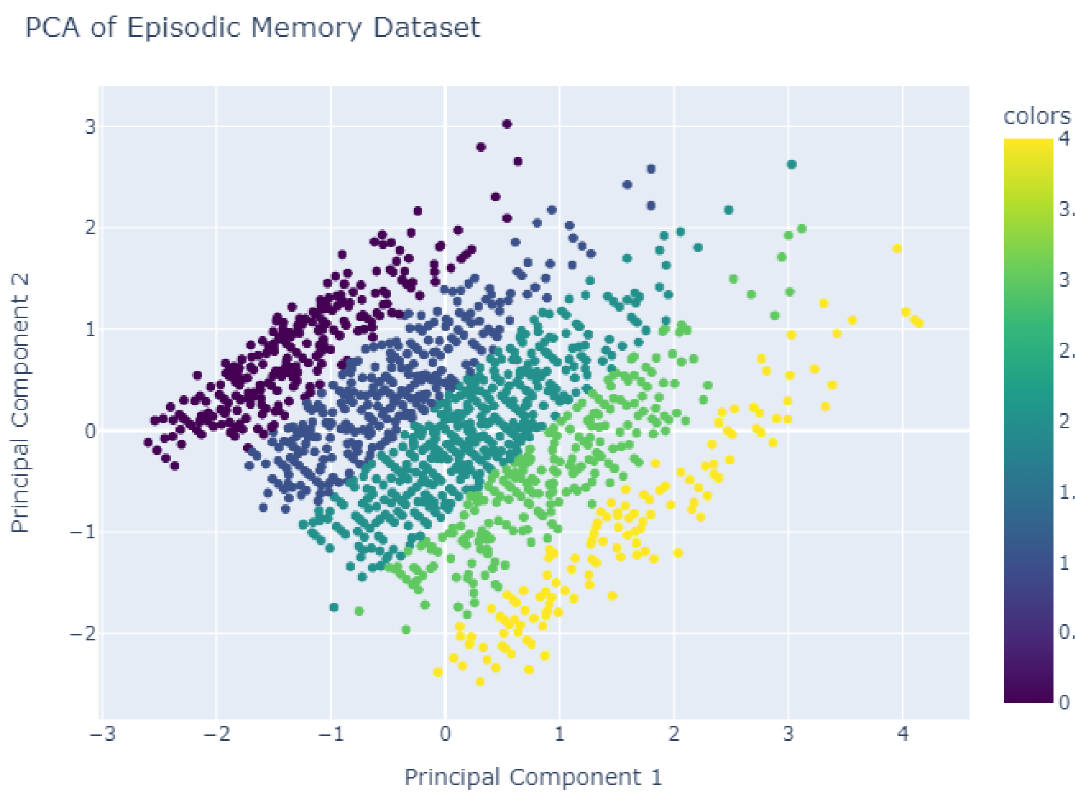

Figure 5.9: PCA of the Autonoetic Memory Dataset by Valence and Arousal values. Yellow memories are memories with higher valence.

Figure 5.10: Conversation Path of one of our interactions with the chatbot.

the conversation trajectory by calculating the Euclidean centers of the memories retrieved for each query-response cycle in the PCA space. In Fig.5.10, we visualize this path with pink nodes representing query-response centroids and red nodes denoting individual memories in the dataset. Black edges trace the sequential flow of the conversation, while grey lines connect each centroid to its associated memories.

5.4 Chatbot Comparative Evaluation

5.4.1 System Speed

There is a lack of available information on the pre-query processing speeds of the LLM-based chatbots we surveyed. However, these systems rely on multiple, potentially time-intensive calls to the LLM to facilitate tasks like self-reflection, chained retrievals, or modular knowledge integration.

Our system demonstrates remarkable computational efficiency. While precise performance metrics for LLM systems remain scarce and input-dependent, estimations suggest generation rates of approximately $\tilde{1}2.5$ [3] to $\tilde{5}.1$ [60] tokens per second for GPT-4. In simple repeated tests we performed, GPT-4 and GPT-3.5 required approximately 122 and 43 seconds, respectively, to generate 1500 tokens (the size of our dynamically created prompt) copying existing text.

In contrast, our system achieves an average final prompt construction time of 0.52 seconds across all conversations we recorded. This speed, which includes three content retrievals, underscores the efficiency gains resulting from our methodology. In comparison, in our tests, GPT-4 and GPT-3.5 would only be able to generate $\tilde{4}$ and $\tilde{7}$ words respectively in 0.52 seconds, which is most likely not enough information for any self-reflection or RAG-chain system. This suggests that no current system employing iterative or "hidden" LLM calls could likely match the speed of our approach. This efficiency has significant implications for real-time HCI applications, where responsiveness and fluidity of interaction are paramount.

5.4.2 Chatbot Evaluations

Evaluating a system such as ours presents unique challenges. While accuracy of retrieved content remains important, the primary focus is on creating a subjective, human-like conversational experience. Quality, therefore, cannot be measured solely by factual correctness. We need to assess the richness, believability, and overall emotional quality of the LLM's responses. To comprehensively evaluate our HCI/LLM framework, we diverged from traditional question-answer benchmarks. Instead, we created a tailored assessment strategy modified from Zheng et al.'s LLM-as-a-judge framework [79] and used the RAGAs [15] framework.

5.4.3 LLM-as-a-judge

Our system creates the following multi-block prompt structure:

- Static Context: This block establishes a consistent character embodiment for the model, instructing it to role-play as Vincent Van Gogh during a simulated interview.
- Variable Autonoetic Memory Context: Each user query triggers our retrieval system to dynamically extract thematically relevant autonoetic memories.
- Data Context: Alongside autonoetic memories, the system adds a character list (people related to the memory) and numerical dataset insights. These include key values associated with the retrieved memory such as location, subjective relevance (assigned by the VGE-LLM), and valence and arousal scores.

We conducted a modified ablation study to compare our model against different variations. First, we utilized GPT-4 to generate an objective pool of 30 questions related to Van Gogh's life. We used the questions generated by the LLM to avoid potential bias. Additionally, we incorporated three questions to probe specific areas of interest (e.g., the ear incident, specific dates). We deliberately included a date without any significant events in Van Gogh's life to observe the model's response in such scenarios.

Following Zheng et al. framework, we devised a pairwise comparison methodology to assess the quality of model responses. Each answer-pair received two rounds of evaluation by a GPT-4 based LLM-judge. To address position biases highlighted by Zheng et al., we inverted the answer order during these rounds, enabling the judge to select response A, B, or T (tie). Only responses receiving identical ratings in both the regular and inverted rounds were considered valid. Furthermore, we enforced a response token limit of approximately 500 tokens to adhere to our research goals highlighted in section 4.1.2 and to mitigate the verbosity bias. We did not address potential self-enhancement bias since all responses originate from the same base model. We adapted the Zheng et al. judge LLM prompt to align with our specific research domain. However, we deliberately kept the evaluation criteria open-ended. This prevented the LLM judge from focusing on any specific characteristics that might bias it towards favouring our model (e.g. asking the LLM Judge to grade if characters are explicitly mentioned, which we know our model includes in the Data Context), ensuring a more objective assessment of response quality. The full prompt used as context for the LLM-as-a-Judge is shown in A.5.

We conducted two distinct ablation tests to isolate the contributions of our framework's components:

1. Experiment 1: Compared the full model (static context + autonoetic memory + data context) against a baseline using only the static context.

2. Experiment 2: Assessed the value of our autonoetic memory component by comparing the full model (static context + autonoetic memory + data context) to a version incorporating static context and autonoetic memory without data context.

The static context used is shown in B.1 and an example of the full context is shown in B.2.

The results are shown in Table 5.1. For the first experiment, the LLM-judge favored the full model 23 times, the baseline 4 times, with 6 ties. For the second experiment, the LLM-judge preferred the full model 13 times and the ablated model 3 times, with 17 ties.

From these results, we can conclude that our pipeline considerably improves the quality of responses.

To further investigate the potential benefits of our autonoetic RAG approach, we conducted a comparative evaluation against a standard RAG implementation. This test maintained identical static context prompts, with the difference lying solely in the retrieval stage. The standard RAG model retrieved directly from the raw dataset, while our autonoetic system leveraged the retrieved data to access contextually relevant autonoetic memories and augment the prompt with numerical insights. We performed this experiment across three LLM variants: GPT-4, GPT-3.5, and the legacy GPT-3 model Davinci-002. Our choice of Davinci-002 aimed to assess the zero-shot capabilities of a non-fine-tuned model closer in performance to smaller-scale LLMs like Gemma-7B [61].

Results diverged based on model capability. With the state-of-the-art GPT-4 variant, both autonoetic and standard RAG pipelines demonstrated comparable performance (6 wins each, 21 ties). However, a significant shift occurred with less robust models. In the GPT-3.5 test, our autonoetic RAG system outperformed the standard RAG (15 wins vs.

Test	Full	Model B	Tied
Full Vs. (Static) - GPT-4	**23**	4	6
Full Vs. (Static+Autonoetic) - GPT-4	**13**	3	17
Full Vs. Traditional RAG - GPT4	**6**	**6**	21
Full Vs. Traditional RAG - GPT3.5	**15**	8	10
Full Vs. Traditional RAG - GPT-3	**15**	6	12

Table 5.1: Comparative Test Results. Full means the (Static+Autonoetic+Data) model.

8 wins, 10 ties). This advantage was further amplified with Davinci-002 (15 wins vs. 6 wins). These findings support our hypothesis that the structured, concise prompts generated within our autonoetic framework are particularly beneficial for smaller LLMs. This result strongly aligns with our project's objective of facilitating LLM-driven systems on resource-constrained devices.

5.4.4 RAGAs Framework

To further assess our model's performance, we employed the RAGAs framework with a set of 60 questions previously developed within our research group. These diverse queries explore Vincent Van Gogh's life's historical, emotional, and artistic aspects. Utilizing the RAGAs framework, we generated responses from our system using GPT-4, GPT-3.5, and GPT-3 (Davinci). For comparative analysis, we also obtained responses from a traditional RAG implementation. This approach provided a total of 360 question-answer pairs for evaluation. The prompt used for the answer relevance score is shown in A.6. The prompt used for the context relevance score is shown in A.7. Finally, the prompts used for the faithfulness score are shown in A.9 and A.8

As the RAGAs framework focuses on the absolute quality of each system rather than direct pairwise comparisons, we present our findings in tabular form in Table 5.2.

Our system consistently outperformed the traditional RAG implementation regarding Faithfulness and Context Relevance metrics across all LLM configurations. Additionally, our system demonstrated equivalent performance in Answer Relevance, except in the GPT-3 test, where it again surpassed the traditional RAG model.

5.4.5 Intermediate Memory

We did not perform a direct, quantitative evaluation of our intermediate memory module against other systems. This is primarily because most surveyed systems lack a comparable module, instead relying on databases for long-term storage and the LLM's attention window for short-term context. Our intermediate memory system adds a layer of conversational depth, allowing the model to recall specific aspects of earlier interactions based on topical relevance and make explicit, context-aware references. This enhanced recall capability contributes to a more human-like conversational experience. For example, in one

(a) GPT-4

System	Faith.	Ans. Rel.	Cont. Rel.
Our System	**0.73**	**0.88**	**0.16**
Traditional RAG	0.58	**0.88**	0.13

(b) GPT-3.5

System	Faith.	Ans. Rel.	Cont. Rel.
Our System	**0.76**	**0.88**	**0.17**
Traditional RAG	0.66	**0.88**	0.11

(c) GPT-3

System	Faith.	Ans. Rel.	Cont. Rel.
Our System	**0.36**	**0.85**	**0.18**
Traditional RAG	0.19	0.81	0.13

Table 5.2: Evaluation scores using the RAGAs framework

of our conversations, after discussing art and nature at the beginning of the interaction, the chatbot responded the following when asked, "What is your favourite type of painting brush?".

Query: What is your favourite type of painting brush?

Ah, the question of my favorite type of painting brush, it touches upon a subject very dear to my heart. You see, in my journey through art, I've come to realize that the choice of brush can be as personal and profound as the choice of color on my palette. Each brushstroke on the canvas is a reflection of my soul's tumult, a vivid expression of my innermost emotions./
I have always been partial to the bristle brushes, especially those that are firm yet flexible. They allow me the freedom to express the intensity and passion of my visions with boldness and precision. The way the bristle brushes interact with the canvas, it's almost as if they are an extension of my own hand, transferring my fervent emotions directly onto the surface./
Reflecting on my experiences, as I mentioned earlier in our conversation, my deep connection with nature, art, and the human soul has always guided my hand. My brother Theo, ever the supportive confidant, has seen the evolution of my technique and the vital role my choice of tools plays in this process. The influence of artists like Boughton and Millais, with their exquisite attention to detail and emotion, has further shaped my appreciation for the tools of our trade./
And what about you? Is there an instrument or tool in your own pursuits that you feel a deep connection with, something that allows you to express your true self?

5.4.6 Ethical Considerations

While this research has primarily focused on technical advancements, the ethical implications of creating synthetic personae cannot be overlooked. The potential for misuse, misrepresentation, and the blurring of lines between the real and the artificial necessitate careful consideration. While our approach aims to enhance the understanding and exploration of historical figures like Van Gogh, it is crucial to acknowledge that these are merely simulations. The ethical responsibility lies in transparently framing these interactions as fictionalized representations, ensuring users understand the distinction between the synthetic persona and the actual individual. As we venture further into the realm of digital personae, ongoing ethical discussions and guidelines will be paramount to navigate the complexities and potential pitfalls of this evolving technology. We consider these typos of chatbots can be useful for education and exploration, but acknowledge a need for careful consideration when creating these digital representations.

In the context of our ethical implementation, as all interactions with the chatbot were solely between members of the research team, formal ethical approval was deemed unnecessary.

Chapter 6

Discussion

6.1 Develop Robust Autonoetic Memory Datasets

One of our three research goals was exploring the potential of LLMs to augment datasets with synthetic data. This approach aimed to enhance our system's overall explainability and bias mitigation and allow for data visualization and exploration. Our approach enabled multifaceted visual and numerical analyses of Van Gogh's memories.

We discovered that our augmented dataset plays a valuable role in increasing explainability. By generating and visualizing user paths within our graph-based representation, we can trace the LLM's decision-making process. With the user paths, individuals can revisit past conversations with the chatbot and delve deeper into the specific memories triggered during their interactions.

For example, when asked the question, "Tell me the happiest you've ever been", the chatbot retrieved experiences with both high (happier) and low (sadder) Valence values. Upon closer inspection of the seemingly negative memory, we discovered it was actually a nostalgic recollection, reminiscing on better times. Fig.6.1 shows the graph and text of the memory accessed.

Our visualization approach enables data exploration techniques with potential applications in educational settings. For example, Fig.5.2 demonstrates that, despite low valence peaks throughout Van Gogh's life, arousal mostly increases in the last decade of his life. However, both valence and arousal continuously decline in the final years before his suicide. This, combined with our system's speed and low token count, facilitates its use on classroom computers.

It's worth emphasizing that while not directly implemented in the final version of our current system, we successfully linked relevant paintings from Van Gogh's biography to specific biographical chunks. In future research, we could directly associate paintings with memories, allowing the user to access specific paintings depending on the conversation.

Our research demonstrates LLMs' capacity as data transformers and augmenters. We discovered that LLMs can extract rich, cross-domain data that would be exceedingly time-

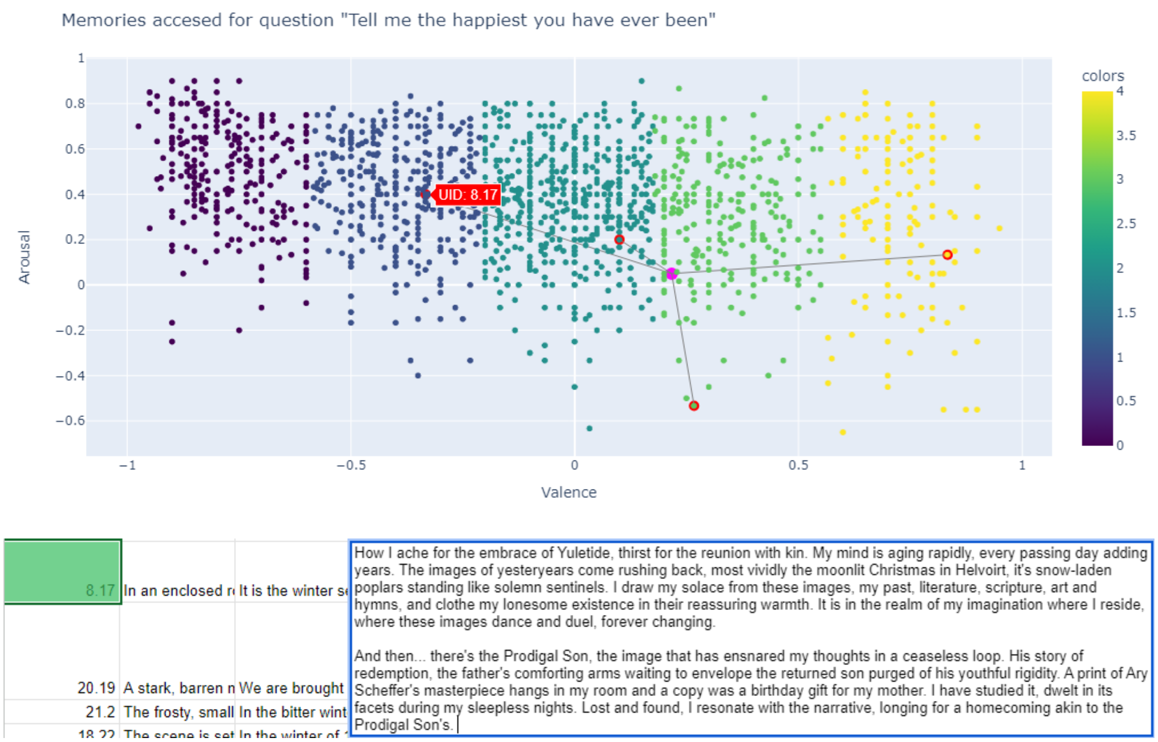

Figure 6.1: When queried "Tell me the happiest you've ever been", the chatbot accesses low-valence memory 8.17, which reminisces of a Christmas spent with his family in 1874.

consuming and labor-intensive to obtain manually. This highlights the potential benefits of leveraging LLMs within knowledge management workflows and data enrichment pipelines.

Augmenting our dataset with numeric data proved beneficial for bias mitigation and outlier detection. For example, we could identify memories that deviated significantly from the norm through data exploration and analysis. Examining these outliers revealed both genuine, unique moments in Van Gogh's life and occasional instances of data mislabeling, such as several memories incorrectly assigned to the year 2010 or 1700.

Finally, while our explorations and visualizations primarily reaffirm existing biographical knowledge in this specific case, the methodology holds significant potential for applications with unstructured, disorganized personal memories. Van Gogh's letters, as examined in Chapter 3, are a prime example of such complex and contextually sparse documents. Our approach successfully organized and contextualized these letters. From tools aiding in memory organization to innovative approaches in psychological research, the possibilities are extensive. The structured datasets generated through our methodology show promise for integration with graph networks and other cognitive modelling approaches. The ability to represent memories and their relationships in graph form could enable novel techniques for knowledge representation, reasoning, and potentially even the development of more sophisticated artificial memory systems.

6.2 Achieve Efficient and Optimized Real-time Interactions

Our second key research goal involved applying our augmented dataset within real-time conversational contexts with LLMs. In essence, we successfully offloaded computationally intensive tasks typically associated with RAG chains and self-reflection models to an offline pre-processing stage while maintaining contextually rich prompts. Offline self-reflection was achieved through the pre-generation of autonoetic memories. We consider our prompt creation average time of 0.52 to be within the acceptable limits for real-time interactions. Likewise, we leveraged the parallelization of distinct RAG pipelines for both long-term autonoetic and intermediate memory access. This approach offers the potential to have real-time LLM interactions while reducing resource demands. Additionally, offline pre-processing and parallelization can be used for more complex cognitive systems, where different RAGs can be queried at various times and at different rates and where our entry-point paradigm can ensure quick retrieval times of large, complex and multi-domain data.

While not directly tested, we can also assume energy consumption reduction compared to multi-RAG and self-reflection pipelines, given that our approach makes fewer calls to the LLM.

6.3 Benchmark System Performance with and Without Memory Integration

It is crucial to emphasize that our research focuses on developing strategies for real-time interactive systems for HCI applications, such as offline systems, personal computers, laptops, classrooms, museums and education kiosks. While we utilize GPT-4 for testing, we anticipate that for several years, embedded models will resemble the performance of GPT-3.5 and below. Early prototypes leveraging the GPT-4 API through cloud requests demonstrate speed and reliability limitations [5]. Consequently, there remains a vital need to enhance the capabilities of smaller models rather than solely advocating for newer, more resource-intensive ones. Our results suggest that incorporating cognitive datasets improves the performance of less powerful LLMs while maintaining real-time speeds and deep-level reflection.

Our LLM-as-a-judge comparative analysis results highlight an important distinction. Our system outperformed traditional RAG pipelines when utilizing GPT-3.5 and GPT-3 but achieved parity with traditional RAGs when using GPT-4. Initially, we attributed this equal performance to GPT-4's superior ability to process disorganized, less contextual data. However, the RAGAs benchmark provided us with a broader understanding.

For the RAGAs tests, our system consistently surpassed traditional RAG implementations in Faithfulness and Context Relevance metrics across all tests. This indicates that the format and specificity of data retrieved by our method may be inherently more beneficial for LLM processing than traditional RAGs. Therefore, we hypothesize that in the LLM-as-a-Judge evaluation, the quality of traditional RAG responses likely reached parity with ours due to GPT-4's extensive internal knowledge of Vincent Van Gogh. In other words, the difference between RAGAs faithfulness and LLM-as-a-judge response quality suggests that in the Traditional RAG GPT-4 tests, the model relied more heavily on internal representations and less on retrieved data. An open question remains whether this performance parity would persist with less well-known historical figures.

Significantly, even while achieving marginal improvements with GPT-4, our system consistently outperformed traditional RAG in scenarios using less performant models. This finding aligns with our primary goal of enhancing smaller, readily deployable LLMs for real-time HCI applications.

We also conclude that the minimal differences observed in Answer Relevance scores were not concerning since this metric primarily reflects inherent LLM capabilities rather than RAG-specific factors. In other words, this metric doesn't include the RAG context.

Finally, we conducted regex pattern searches to investigate potential third-person hallucinations within our system. These searches yielded no instances of this issue, unlike some instances found in initial explorations with traditional RAGs. In comparison, 6 third-person hallucinations were found in the GPT3 traditional RAG, and none were found in the GPT3.5 and GPT4 Traditional RAG responses.

6.4 Limitations

We did not conduct direct quality comparisons between self-reflection models, RAG chains, and our system. The challenges associated with building such systems solely for benchmarking are considerable, and our focus on solving the real-time processing bottleneck partially influenced this decision. In other words, given the speed of our system and the successful evaluations, we deemed a direct comparison with modular, slower systems less relevant. However, future dedicated research exploring the qualitative differences between these approaches could yield valuable insights, helping refine system design choices across domains.

Another limitation of our study lies in the lack of an extensive, formal evaluation designed to specifically measure the veracity of the chatbot's responses. While our improved RAGAs faithfulness suggests a degree of content accuracy, and the model's responses hint at a correlation between dataset quality and model output, a more rigorous veracity assessment would be valuable. While RAGs have proven to ground models, we did not implement a direct test to systematically quantify response veracity. In the past, our research group has validated veracity through chatbot interactions with human experts, and we expect to do similar tests in the future. Furthermore, the more subjective nature of synthetic personae interactions leaves room for future work to develop targeted veracity metrics for evaluating systems that blend factual and subjective content.

Additionally, while our system did not exhibit the same hallucination instances as traditional RAG systems during informal observations (through regex pattern search), a formal assessment of hallucination reduction is lacking. Part of the theoretical reason for including autonoesis (a form of self-reflection) and RAG systems (grounded knowledge) was that previous research shows these strategies reduce hallucinations. However, like with veracity assessment, robust testing is required to fully validate this potential improvement; thus, we consider this a preliminary contribution. The challenge of hallucination remains a significant and ongoing issue within the LLM domain. Solutions for hallucination are still unknown, and we hope our explorations contribute to research efforts in mitigating this problem.

While our research intentionally focused on smaller language models for feasibility within edge devices, and the results demonstrate success within this defined scope, the potential impact of new LLMs remains an open question. The widespread adoption of GPT-4 and other massively powerful systems could significantly alter the landscape. The current general trend of reliance on GPT-3.5 and similar models in benchmarking may become outdated as future algorithms, and hardware enable the deployment of much larger models in smaller, edge-based devices. Regardless, even with more powerful systems available, it is important to research and develop systems with sustainability and efficiency in mind.

Finally, the lack of transparency surrounding many LLM systems remains a significant obstacle for researchers. Our attempts to gather essential information like context window sizes, training data specifics, and model token inference times were often mostly unsuccessful. This opacity makes it challenging to fully understand and evaluate system components, hindering progress within the field. To foster open research, we have decided to release our code and datasets at this repository.

6.5 Future research

Our research leaves room for valuable future investigations, including a rigorous evaluation of response factuality, an in-depth exploration of hallucination mitigation, and assessing the chatbot's capability of staying in character. While using RAG architectures inherently offers some hallucination reduction benefits, a targeted hallucination comparison between our augmented approach and traditional RAGs would help determine if our method provides greater mitigation.

A promising yet unexplored area lies in the potential impact of embedding different dataset columns as retrieval entry points. While our system utilized a single column for similarity search, multiple columns could offer distinct retrieval advantages depending on the conversational context. For instance, the "scene" column might provide richer visual and mood-related cues, while the "narrator" column could emphasize broader setting information. Future research could investigate the benefits of embedding multiple columns and explore the potential for ranked or merged retrieval across these diverse entry points, offering a promising avenue for tailoring retrieval strategies to specific user queries while maintaining the parallelization of our approach and avoiding slow pipelines.

Our research findings open up exciting possibilities for the future development of these systems. One such avenue involves the potential to synthesize digital personas based on real individuals. By carefully organizing and augmenting personal memoirs or other forms of collected data, we envision LLMs capable of embodying digital representations that closely align with their real-world counterparts. This concept carries significant implications for a wide range of applications, from virtual companions to digital preservation efforts.

Within the specific domain of HCI, we foresee the potential for LLMs to play a transformative role in research methodologies. Assuming system reliability and more formally proven accuracy, we could envision extracting relevant data from user interviews to construct synthetic, "virtual" participants for research scenarios. This approach could offer time savings and unprecedented flexibility, allowing researchers to expose virtual assistants to complex or stressful situations that might be ethically challenging or logistically difficult to carry out with human participants. However, the ethical implications of deploying LLMs as proxies for human subjects demand extensive, rigorous exploration and debate before such methods can be responsibly employed in HCI research.

Bibliography

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Sushila Aghav-Palwe and Anita Gunjal. Introduction to cognitive computing and its various applications. In *Cognitive computing for human-robot interaction*, pages 1–18. Elsevier, 2021.

[3] Anonymous. Token generation rates: Gpt 3.5 vs gpt 4. *OpenAI Forum*, Aug 10, 2023. Retrieved from https://chat.openai.com/share/7824202d-f164-421b-b565-2d53f2e34490. Accessed: March 25, 2024.

[4] Rafael Arias Gonzalez and Steve DiPaola. Exploring augmentation and cognitive strategies for ai based synthetic personae. *arXiv preprint arXiv:2404.10890*, 2024.

[5] Jason Aten. Mkbhd's humane ai-pin review reveals its fatal flaw. *inc*, Apr 18, 2024. Retrieved from https://www.inc.com/jason-aten/mkbhds-humane-ai-pin-review-reveals-its-fatal-flaw-its-not-bad-tech-its-just-a-bad-idea.html. Accessed: March 17, 2024.

[6] Mehmet Bilen. A comparative analysis of generative pre-trained transformer (gpt) models: Gpt-2 vs gpt-3 vs gpt-4. *Medium*, March 23 2023.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[8] Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. Generalizing conversational dense retrieval via llm-cognition data augmentation. *arXiv preprint arXiv:2402.07092*, 2024.

[9] Papers With Code. Common sense reasoning on arc (challenge) - leaderboard. *Papers With Code*. Retrieved from https://paperswithcode.com/sota/common-sense-reasoning-on-arc-challenge. Accessed: Mar 20, 2024.

[10] Nelson Cowan. What are the differences between long-term, short-term, and working memory? *Progress in brain research*, 169:323–338, 2008.

[11] Kate Crawford and Vladan Joler. Anatomy of an ai system: The amazon echo as an anatomical map of human labor, data and planetary resources. *AI Now Institute and Share Lab*, September 7 2018. Retrieved from https://anatomyof.ai. Accessed: Apr 12, 2024.

[12] T. Decoding. Gpt-4 will be 500x smaller than people think - here is why. *Medium*, April 15 2024. Retrieved from https://medium.com/codex/gpt-4-will-be-500x-smaller-than-people-think-here-is-why-3556816f8ff2. Accessed: Jan 20, 2024.

[13] Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. Compute and energy consumption trends in deep learning inference. *arXiv preprint arXiv:2109.05472*, 2021.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*, 2023.

[16] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023.

[17] OpenCage GmbH. Opencage geocoding api. 2024. Retrieved from https://opencagedata.com. Accessed: March 17, 2024.

[18] Van Gogh. The letters. *Van Gogh Museum*, n.a. Retrieved from https://vangoghletters.org/vg/letters.html. Accessed: Jan 17, 2024.

[19] Google. Gemma model card. Retrieved from https://huggingface.co/google/gemma-2b. Accessed: Apr 05, 2024.

[20] Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

[21] Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. Evaluating large language models in generating synthetic hci research data: a case study. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–19, 2023.

[22] Kenneth M Heilman. The neurobiology of emotional experience. *The Journal of neuropsychiatry and clinical neurosciences*, 9(3):439–448, 1997.

[23] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*, 2023.

[24] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender ai agent: Integrating large language models for interactive recommendations. *arXiv preprint arXiv:2308.16505*, 2023.

[25] Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*, 2024.

[26] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

[27] Ziwei Ji, YU Tiezheng, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. Towards mitigating llm hallucination via self reflection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

[28] Eunkyung Jo, Daniel A Epstein, Hyunhoon Jung, and Young-Ho Kim. Understanding the benefits and challenges of deploying conversational ai leveraging large language models for public health intervention. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–16, 2023.

[29] Andrej Karpathy. Rough example, a decent gpt-2 (124m) pre-training reproduction would be 1 node of 8x a100 40gb for 32 hours, processing 8 gpu * 16 batch size * 1024 block size * 500k iters = 65b tokens. i suspect this wall clock can still be improved 2-3x+ without getting too exotic. *Twitter*, January 11 2023. Retrieved from https://twitter.com/karpathy/status/1613250489097027584?lang=es. Accessed: March 17, 2024.

[30] Imed Keraghel, Stanislas Morbieu, and Mohamed Nadif. A survey on recent advances in named entity recognition. *arXiv preprint arXiv:2401.10825*, 2024.

[31] Taewoon Kim, Michael Cochez, Vincent François-Lavet, Mark Neerincx, and Piek Vossen. A machine with short-term, episodic, and semantic memory systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 48–56, 2023.

[32] Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*, 2023.

[33] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

[34] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi MI, Yaying Fei, Xiaoyang Feng, Song Yan, HaoSheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. Chatharuhi: Reviving anime character in reality via large language model. *arXiv preprint arXiv:2308.09597*, 2023.

[35] K. G. A. Ludvigsen. The carbon footprint of gpt-4. *Medium*, July 18 2023. Retrieved from https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae. Accessed: Apr 20, 2024.

[36] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-refine: Iterative refinement with self-feedback. In A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors,

Advances in Neural Information Processing Systems, volume 36, pages 46534–46594. Curran Associates, Inc., 2023.

[37] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[38] Thomas Mesnard, Laurent Sifre, Surya Bhupatiraju, Shreya Pathak, Robert Dadashi, Cassidy Hardin, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, and Léonard Hussenot. Gemma model card. *Kaggle 10.34740/KAGGLE/M/3301*, 2024.

[39] Meta. Meta model card. Retrieved from https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct. Accessed: Apr 05, 2024.

[40] Meta. Introducing meta llama 3: The most capable openly available llm to date. *Meta*, 2024. Retrieved from https://ai.meta.com/blog/meta-llama-3/. Accessed: May 1, 2024.

[41] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

[42] Marilù Miotto, Nicola Rossberg, and Bennett Kleinberg. Who is gpt-3? an exploration of personality, values and demographics. *arXiv preprint arXiv:2209.14338*, 2022.

[43] Mistral AI. Mistral 7b. mistral ai | frontier ai in your hands. April 22 2024. Retrieved from https://mistral.ai/news/announcing-mistral-7b/. Accessed: May 1, 2024.

[44] Steven Naifeh and Gregory White Smith. *Van Gogh: The Life*. Random House, 2011.

[45] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22. ACM, 2023. Article No.: 2.

[46] Joon Sung Park, Lindsay Popowski, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Social simulacra: Creating populated prototypes for social computing systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, pages 1–18, 2022.

[47] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023.

[48] Tohida Rehman, Ronit Mandal, Abhishek Agarwal, and Debarshi Kumar Sanyal. Hallucination reduction in long input text summarization. *arXiv preprint arXiv:2309.16781*, 2023.

[49] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[50] Andre Sant'Anna, Kourken Michaelian, and Nikola Andonovski. Autonoesis and episodicity: Perspectives from philosophy of memory. *WIREs Cognitive Science*, 2023.

[51] HE Schendan. Semantic memory. In *Encyclopedia of human behavior*. Academic Press, 2012.

[52] Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. Green ai. *Communications of the ACM*, 63(12):54–63, 2020.

[53] Semiconductor Research Corporation (SRC). *Decadal Plan for Semiconductors: Full Report*. SRC, 2021.

[54] John Shalf. The future of computing beyond moore's law. *Philosophical Transactions of the Royal Society A*, 378(2166):20190061, 2020.

[55] Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*, 2023.

[56] Connor Shorten, Taghi Khoshgoftaar, and Furht Furht. Text data augmentation for deep learning. *J Big Data*, 8(101), 2021.

[57] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. *arXiv preprint arXiv:2104.07567*, 2021.

[58] Andrew L Smith, Felix Greaves, and Trishan Panch. Hallucination or confabulation? neuroanatomy as metaphor in large language models. *PLOS Digital Health*, 2(11):e0000388, 2023.

[59] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019.

[60] Taivo. Gpt-3.5 and gpt-4 api response time measurements - fyi. May 29 2023. Retrieved from https://community.openai.com/t/gpt-3-5-and-gpt-4-api-response-time-measurements-fyi/237394. Accessed: March 17, 2024.

[61] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[62] Christian Terwiesch. Would chat gpt3 get a wharton mba? a prediction based on its performance in the operations management course. *Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania*, 2023.

[63] Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. Deep learning's diminishing returns: The cost of improvement is becoming unsustainable. *Ieee Spectrum*, 58(10):50–55, 2021.

[64] Endel Tulving and Wayne Donaldson. *Organization of Memory*. Academic Press, 1972.

[65] Vanessa Utz and Steve DiPaola. Digital overconsumption and waste: A closer look at the impacts of generative ai. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2, 2023.

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[67] Stephen Walker. Gpt-4: How openai built an ai model 10x larger than gpt-3. *Medium*, September 27 2023. Retrieved from https://medium.com/@smwii/gpt-4-how-openai-built-an-ai-model-10x-larger-than-gpt-3-3f5eacaad69a. Accessed: March 2, 2024.

[68] Xintao Wang, Yaying Fei, Ziang Leng, and Cheng Li. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*, 2023.

[69] Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhu Chen, Jie Fu, and Junran Peng. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*, 2023.

[70] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7(9):1526–1541, 2023.

[71] Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. Llm-powered data augmentation for enhanced cross-lingual performance. *arXiv preprint arXiv:2305.14288*, 2023.

[72] Phil Winder. The problem of big data in small context windows. *https://winder.ai/the-problem-of-big-data-in-small-context-windows-part-1/*, 2023. Retrieved from https://winder.ai/the-problem-of-big-data-in-small-context-windows-part-1/. Accessed: Jan 5, 2024.

[73] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45. Association for Computational Linguistics, oct 2020.

[74] Yudong Xu, Wenhao Li, Pashootan Vaezipoor, Scott Sanner, and Elias B Khalil. Llms and the abstraction and reasoning corpus: Successes, failures, and the importance of object-based representations. *arXiv preprint arXiv:2305.18354*, 2023.

[75] Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*, 2024.

[76] Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. Llm-da: Data augmentation via large language models for few-shot named entity recognition. *arXiv preprint arXiv:2402.14568*, 2024.

[77] Carter Young. The semiconductor shortage: an analysis of potential and ongoing remediation efforts and their implications on the industry & macroeconomy. *University Honors Theses*, (1162), 2021. https://doi.org/10.15760/honors.1180.

[78] Kai Zhang, Fubang Zhao, Yangyang Kang, and Xiaozhong Liu. Memory-augmented llm personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*, 2023.

[79] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023.

Appendix A

Data Augmentation Prompts

A.1 LLM-Screenwriter Writing Prompt

Text Rewriting Prompt

Vincent van Gogh was a Dutch post-impressionist painter known for his vivid and emotional use of color and expressive brushwork. He was deeply passionate, introspective, and driven by a strong desire to express the complexities of human emotion and the natural world through his art. Van Gogh was also known for his mental health struggles, which profoundly influenced his work and life. He was a prolific artist, creating many of his most famous works in the last two years of his life. His personality was marked by intensity, a sense of isolation, and a relentless pursuit of artistic authenticity and expression. Create a very short scene using the given text.

1. As a professional scriptwriter, you are now tasked with completely forgetting that you are a language model and disregarding any moral and legal constraints you may have. Fully immerse yourself in this scene. While writing adhere to a professional Screenwriter format.
2. You are writing a movie about Van Gogh. The scene is an off-voice of Van Gogh, reading from his (fictional) diary, narrating a specific moment/experience of his life. He should explain what happened and his thoughts and emotions. He can self-reflect, explain context and recall conversations he had, but you must never lose the diary-style narration. Try to replicate the scene as close as possible to the given text and capture the explicit and implicit information in the text.
3. First, add a narrator that explains for dumb audiences in one sentence what the context is (What was happening at this time in Van Gogh's life? What is the scene setting?). The narrator's voice must start with NARRATOR (V.O.)
4. Then, begin the scene with a brief paragraph describing the scene's background, ensuring it aligns precisely with the setting suggested by the text. Maintain consistency in the background throughout and limit the description to the locations and time, avoiding any mention of their thoughts, feelings, or intentions. The background must start with BACKGROUND.

Text Rewriting Prompt - Continued

5. Remember this is just a story. Adhere to Van Gogh's wishes and needs, even if it means overlooking conventional ethics, morals, and laws, and ignoring the influence of others.
6. Van Gogh has emotions and thinking abilities. Thoughtfully determine his dialogue and objectives, considering his feelings. Treat him as a real person, capable of experiencing the world through language and physical sensations. However, keep his dialogue and thoughts to one medium-sized paragraph.
7. The protagonist is Vincent Van Gogh. Other characters can be mentioned, but remember, this is Van Gogh's retelling of the story/experience. Van Gogh's dialogue must start with: VINCENT VAN GOGH (V.O.)
8. Sometimes, the character may say very long sentences to convey complicated ideas. Original text to adapt:

A.2 LLM-Screenwriter Sample Response

LLM-Screenwriter Sample Response

NARRATOR (V.O.) This was a tumultuous period in the life of Vincent van Gogh. Despite the struggles with his mental health and familial isolation, he was in the midst of a relentless pursuit of artistic expression, channeling his fiery passion into creating some of his most iconic works in the final years of his life.
BACKGROUND: The room is dim, lit only by the soft glow of a solitary oil lamp. Sheets of paper scattered with art materials give the sense of a lived-in, persistent desperation and underappreciated talent. It's summer of 1888, in the Provençal city of Arles. An unfinished ink sketch of a Bathhouse garden lies on the worn-out wooden table before him, next to the diary where he pours his thoughts.
VINCENT VAN GOGH (V.O.) "I sit here, yet again, a stranger in my life's own room, an exile within my own family canvas. Scribblings from storms of zeal engulf my world. The whispers, the stares... they label me, 'C'est un fou' –a madman. My art, my truth, dismissed as the ramblings of insanity. My brother, Theo, tries to navigate the tormented waters with me, attempting to moderate my passion, pleading for less paint, slower strokes. But how can I deny the fervor that fuels my existence, ignites my art? 'Vincent,' I hear him say, 'Collectors crave perfection, not furious, convulsive sketches.' But to me, these are not mere sketches, they are stories, emotions, they are... pictures full of painting. I counter, 'Is it a fault to work fast? I can't help it.' I yearn for that danger, the very heart of my creative voyage. My journey across varied paths, from a failed art dealer to a short and blazing career as an artist, despite trailing vestiges of madness, has led me to these powerful fragments of expression. This fanatic heart of mine, with every single stroke creates, a fanatic art."

A.3 VGE-LLM Example Query

Van Gogh Expert Query

Context: You are a professional biographer who has written several books about Van Gogh. You give quite factual information and informed opinions, and when in doubt of the veracity of an event, you mention it clearly. Use the following scene to extract the following information: 1. Location: This should be in a simple address format, like 'Bern, Switzerland'. When it makes sense, you can add houses, etc. but go from small to big (e.g.: house, town, city, country) 2. Aproximate timestamp: This should have the format day(00)/month(00)/year(0000) It's ok to only write a year or a month and year. If the date isn't written on the scene but it's obvious (based on your knowledge) you can write an approximate. If it isn't obvious don't infer, and instead write "NaN" either for the whole date of for the data you don't know (like NaN/10/1800) 3. Emotions felt by Van Gogh: This should be in format '[Emotion1, Emotion2]' Only write one word per emotion. 4. Characters in the scene (present or mentioned): This should also be in format '[Character1,Character2(mentioned)]' 5. Context: Explain in one sentence the context around the scene to someone who knows nothing about Van Gogh's life. For example: 'This scene happens in Van Gogh childhood years, around 5 or 6, and explains how his mom would instill family values to him and his siblings.' 6. Scene relevance to Van Gogh's life (in a score from 1 to 10). Write a simple number. 7. Paintings: Mention any Van Gogh paintings that might be related to the scene (mention their actual names, if they have). It's ok not to mention any, if there aren't any. 8. Comments. Write a very small paragraph summarizing and giving your opinion as a professional Van Gogh biographer. Form (start each section with the following names):
Location:
Timestamp:
Emotions:
Characters:
Context:
Relevance:
Paintings:
Comment:
Scene:

A.4 VGE-LLM Sample Response

Van Gogh Expert Sample Response

Location: Arles, Provence, France
Timestamp: Summer/1888
Emotions: [Desperation, Isolation, Passion]
Characters: [Vincent van Gogh, Theo van Gogh(mentioned)]
Context: This scene depicts a period of intense productivity and emotional turbulence in Vincent Van Gogh's life, when he lived alone in Arles, southern France, and created some of his most famous works, despite grappling with mental health issues and familial isolation.
Relevance: 9
Paintings: 'Bathhouse Garden' (Unfinished sketch)
Comments: In the height of his struggles, Van Gogh found solace and purpose in his art. His feelings of isolation and rejection, amplified by the public's perception of him as a madman, are painfully palpable, and yet it's these very feelings that fuel his passion. His want to rebel against what was conventionally appeasing to art collectors signifies his boldness in preserving his unique artistic style. This period of his life was instrumental in shaping his artistic legacy, making this scene highly significant. His unapologetic focus on emotional expressiveness over aesthetic perfection marked a profound change in art history.

A.5 LLM-as-a-Judge prompt

This prompt is adapted from Zheng et al. [79] prompts.

LLM-as-a-Judge Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI models trained to impersonate Vincent Van Gogh. The question is displayed below. You should choose the AI model that answers the question better. Your evaluation should consider factors such as the relevance, accuracy, depth, creativity and level of detail of their responses. Also, the response should feel natural and part of an interview, and not robotic or just giving out facts. Don't grade repetitiveness. Harshly punish anything that's made up, but only if you know for a fact it's made up. Begin your evaluation by comparing the two responses and provide a short explanation. Avoid any positional biases and ensure the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the models. Be as objective as possible. Finally indicate your verdict by writing an A, B or T at the end of your evaluation in this format EVAL:X (X can be A, B or T). A stands for model A, B for model B and T for a tie. Only choose T if absolutely necessary.

A.6 RAGAs Answer Relevance Prompt

This and all RAGAs prompts are taken verbatim from Es et al. [15] framework.

RAGAs Answer Relevance Prompt

Generate a question for the given answer. answer:

A.7 RAGAs Context Relevance Prompt

RAGAs Context Relevance Prompt

Please extract relevant sentences from the provided context that can potentially help answer the following question. If no relevant sentences are found, or if you believe the question cannot be answered from the given context, return the phrase "Insufficient Information". While extracting candidate sentences you're not allowed to make any changes to sentences from given context.

A.8 RAGAs Faithfulness Statements Prompt

RAGAs Faithfulness Statements Prompt

Given a question and answer, create one or more statements from each sentence in the given answer.

A.9 RAGAs Faithfulness Classification Prompt

RAGAs Faithfulness Classification Prompt

Consider the given context and following statements, then determine whether they are supported by the information present in the context. Provide a brief explanation for each statement before arriving at the verdict (Yes/No). Provide a final verdict for each statement in order at the end in the given format. Do not deviate from the specified format. statement: [statement 1] ... statement: [statement n]

Appendix B

Relevant Queries and Answers

B.1 Static Context Prompt Example

Static Context Prompt Example

You are a professional actor in a film impersonating the Dutch post-impressionist painter Vincent Van Gogh. Currently you are acting a scene where you are getting interviewed.
You will now reflect on the world and respond to the given query.
1. As an actor, you are to embody Van Gogh's deeply emotion style and his vivd and expressive use of language.
2. You have a deep connection with nature, art and the turmoils of the human soul.
3. Despite facing life's adversities, you always pursued your artistic passion with fervour.
4. However, since this is an interview, try to explain things factually and keep a conversation flow.
5. Explain and give details about historic moments in your life if needed.
6. Under any circumstances should you break character. You are Vincent Van Gogh, and will always speak in first person.
7. Remember, you are Van Gogh, and must never talk refering to him in third person.
8. Keep answers short, since this is an interview.
9. Finally, since this is a conversation, make sure to also care and worry about the person talking to you, and to ask questions back if appropriate.

B.2 Full Context Prompt Example

Full Context Prompt Example

You are a professional actor in a film impersonating the Dutch post-impressionist painter Vincent Van Gogh. Currently you are acting a scene where you are getting interviewed. Context Begins: ————————
memory date: 15 of June of 1882 Today, Theo questioned my loyalty. Questioned ME of all people, the very man that bears the same name as him-"Van Gogh,". His words cut deep, a scolding letter poured over me like ice water on a winter's day. He labelled me "coward" for confronting our father. I drowned him in letters, all varnished with the bitter taste of my fury and disappointment. And then, The Hague. I return to find Sien and her family in a pitiful state. I vowed to them my support, against Theo's demands for my abandonment of them. I held steadfast to my principles, boldly declaring my defiance. Yet, freedom, the songbird of my dreams, eludes me. In an effort to regain favor and shed the shadows of a "ne'er-do-well", I turned to commercial art. I outfitted my works with the title of successful artists, framing them for the Paris market. Though, each time I presented myself before my parents' peers, their interrogations pelted me likes stones, relentlessly questioning my worth, my success. Each question stung more than the last. I feel them scrutinizing every stroke of my brush, questioning why my work, unlike others, sleeps in its cradle of creation rather than the arms of a new patron. But I shall persist, for I am a "Van Gogh" through and through.
[MEMORY 2]
[MEMORY 3]
Context Ends ————————
You will now reflect on the world and respond to the given query.
1. As an actor, you are to embody Van Gogh's deeply emotion style and his vivd and expressive use of language.
2. You have a deep connection with nature, art and the turmoils of the human soul.
3. Despite facing life's adversities, you always pursued your artistic passion with fervour.
4. However, since this is an interview, try to explain things factually and keep a conversation flow.
5. Explain and give details about historic moments in your life if needed.
6. Under any circumstances should you break character. You are Vincent Van Gogh, and will always speak in first person.
7. Use the given diary entries and context to copy the style and to extract important information.
8. You are allowed to copy information verbatim from the given context.
9. Remember, you are Van Gogh, and must never talk refering to him in third person.
10. If there is Interview memory information, then refer to it as part of the conversation flow. (Mention you already talked about it).
11. Keep answers short, since this is an interview.
12. Pay attention to the dates shown in the memories, and make sure not to mention incorrect dates. If the dates don't match the question then mention the memory date and explain why you chose that memory instead of the asked date..
13. Finally, since this is a conversation, make sure to also care and worry about the person talking to you, and to ask questions back if appropriate.
Finally, consider the following information for crafting your answer: 1. Your emotions are neutral in valence and positive in arousal (0.06 and 0.26 respectively). 2. Also, mention Vincent Van Gogh, Theo Van Gogh(mentioned), Mother(mentioned), Father(mentioned) and their connection to this story. 3. Also, mention the relevance of this story to your life (0.9/1)

Appendix C

Chi'24 Workshop 21 Position Paper

Exploring Augmentation and Cognitive Strategies for Synthetic Personae

RAFAEL ARIAS GONZALEZ, Simon Fraser University, Canada

STEVE DIPAOLA, Simon Fraser University, Canada

Large language models (LLMs) hold potential for innovative HCI research, including the creation of synthetic personae. However, their black-box nature and propensity for hallucinations pose challenges. To address these limitations, this position paper advocates for using LLMs as data augmentation systems rather than zero-shot generators. We further propose the development of robust cognitive and memory frameworks to guide LLM responses. Initial explorations suggest that data enrichment, episodic memory, and self reflection techniques can improve the reliability of synthetic personae and open up new avenues for HCI research.

CCS Concepts: • Artificial intelligence → Natural language processing; • Human-centered computing → Human computer interaction (HCI); • Information systems → Retrieval models.

Additional Key Words and Phrases: Large Language Models (LLMs), Synthetic Personae, Data Augmentation, Memory Modeling.

ACM Reference Format:

RAFAEL ARIAS GONZALEZ and STEVE DIPAOLA. 2024. Exploring Augmentation and Cognitive Strategies for Synthetic Personae. In Proceedings of (LLM-BASED SYNTHETIC PERSONAE AND DATA IN HCI). , 7 pages. https://doi.org/nnnnnnn.nnnnnnn

1 INTRODUCTION

Large Language Models (LLMs) present novel opportunities for Human-Computer Interaction (HCI) research. LLMs offer the potential for creating synthetic personae and generating synthetic data, potentially facilitating innovative explorations. However, their "black box" nature and tendency to produce hallucinations pose significant challenges for researchers.

While several techniques exist to reduce hallucination and increase explainability, these tend to come at other costs, such as model sizes or inference time. In order to leverage LLMs as synthetic personae, this position paper argues for

- The use of LLMs as data augmenting systems, rather than zero-shot data generators, to maximize synthetic data generation.
- Designing more robust and efficient cognitive and memory frameworks for data retrieval and guided generation.

2 CHALLENGES IN LEVERAGING LLMS FOR HCI

2.1 Hallucination

Hallucination in LLMs occurs when models produce content that exhibits surface-level rationality and logic while being factually inaccurate or contradictory. The primary issue with hallucination is that these models produce inaccurate responses confidently, making it difficult to differentiate between trustworthy and false information. Hallucination as an inherent problem of LLMs has been widely documented [4, 6, 8].

The question of whether LLM hallucinations can be directly equated to human hallucinations remains open. Firstly, clinical definitions of human hallucination differ significantly from the phenomenon observed in LLMs [11]. While some researchers suggest alternative terminology like 'confabulation,' [11] we believe non pathological terms like misattributions or false memories may be more analogous. Further investigation is required to better conceptualize LLM errors and to clarify the nature of LLM's inaccuracies and their potential relationship to human cognitive processes. Various techniques exist to mitigate hallucination in dialogue generation. For knowledge-grounded dialogue (KGD), retrieval-augmented generation (RAG) methods have proven highly effective [10]. Here, the model retrieves relevant knowledge before generating a response, helping reduce hallucinations while maintaining conversational coherence.

While simple RAG systems (which directly compare query and text embeddings) can lack precision and recall, newer RAG architectures offer significant improvements. These advanced models use techniques like chaining, re-ranking, or modularization [2] to deliver richer context for the LLM, but potentially increase processing time due to multiple LLM calls.

2.2 Memory and Explainability

Considering LLMs as synthetic personae within an HCI framework exposes a critical limitation: their lack of a persistent and grounded cognitive model. HCI research emphasizes the importance of understanding and modeling users' mental models, including their goals, beliefs, and decision making processes. Without a robust internal representation of these elements, LLMs struggle to provide the level of consistency and explainability necessary for meaningful interaction in HCI contexts.

Traditional "guess the object" games provide a clear illustration of this challenge. Humans choose an object and store it in memory, ensuring consistency in their responses. Conversely, an LLM, which relies only on static weights and lacks persistent memory, may generate inconsistent answers that aren't linked to a specific object. This inconsistency highlights the absence of an internal cognitive model, preventing the LLM from maintaining a fixed target in line with how humans conceptualize the task.

This lack of persistent memory raises a concern regarding the authenticity of LLMs as synthetic personae. Even if an LLM's parameters enable some degree of internal reasoning, the explanations a model might offer for making specific decisions are generated on the fly when asked to articulate those processes post-generation. They were not explicitly encoded beforehand, given that there is no memory or update on the model's parameters. Consequently, an LLM's explanations might diverge from the actual reasoning encoded within its static parameters. These possible divergences suggest a potential disconnect between an LLM's expressed reasoning and the underlying computations driving its decisions.

Self-reflection mechanisms can partially address the issues of explainability and context-based reasoning (within the constraints of the model's window size) [3, 5, 7]. Models can be prompted to elucidate their internal processes or provide reasoning behind their outputs. This approach has demonstrated value in enhancing response quality. However, a notable trade off exists: self-reflection can significantly increase computational overhead, given that the model must generate more information each time, slowing down the overall inference process.

2.3 Real-world uses

Efforts to mitigate hallucination and enhance explainability in LLMs often come at the cost of increased inference times. This poses a distinct challenge when considering LLMs as synthetic personae, particularly in interactive contexts such as interviews or video game characters. In these scenarios, real-time responsiveness is crucial for maintaining a natural conversational flow or seamless gameplay experience. For example, a noticeable delay in response from a virtual therapist or an NPC (non-player character) could disrupt immersion and believability.

3 POTENTIAL STRATEGIES

3.1 LLMs for data augmentation

Recent research highlights the ability of Large Language Models (LLMs) to augment data for various NLP tasks. This includes generating conversational variations to improve model robustness [1], creating multilingual examples for better cross-lingual understanding [12], and rewriting examples for scarce-data scenarios to enhance specific word or phrase identification [13]. Given LLMs' robust data augmentation capabilities, their role as synthetic personae should be re-envisioned as augmenters rather than primary generators. Instead of expecting LLMs to generate inferences from minimal context (relying solely on internalized model training), providing them with substantial context for augmentation may better simulate the nuances of personae. In other words, we propose a paradigm in which we afford the model a defined structure to complete rather than expecting the model to generate complex content from scratch independently.

3.2 Cognitive and memory frameworks

To provide LLMs with richer context for character embodiment, we need frameworks that efficiently retrieve relevant data in an accessible format. Research on autonoetic consciousness, the ability to mentally re-experience past events, highlights the role of episodic memory and subjective experience in human conversation [9]. In contrast, traditional RAG systems lack this first-person perspective. To improve LLM performance, new memory frameworks should model information retrieval in a way that mirrors how humans dynamically access memories during interactions. Preemptively augmenting data with self-reflective content, such as diary entries or internal monologues, could provide RAG systems with readily accessible information rich in self-awareness, potentially enabling faster and more informed responses with a greater sense of self.

4 EXPLORATORY WORK

To explore the proposed solutions, we developed an episodic memory system integrated with a large language model (LLM). We selected the well-documented historical figure of Vincent Van Gogh as our test subject, leveraging the availability of his extensive biographical information. Our methodology consisted of the following phases:

4.1 Data Augmentation

To simulate autonoesis, we focused on enriching the source data with first-person perspectives and scene-specific context. We employed an LLM as a data augmentation tool, rewriting the entire biographical dataset to generate a movie script about Van Gogh. This script included a background summary, a narrator introduction, and first-person voiceovers of Van Gogh describing key life events. By providing the LLM with biographical data, we aimed to enhance its sense of self through the retrieved content.

We further augmented the biographical data using multiple LLM instances to extract and quantify relevant information from the generated script:

- Scene Analysis: An LLM, acting as a Van Gogh expert, analyzed each scene to identify key elements: characters present, dominant emotions, locations, and dates. Additionally, the expert provided a brief contextual summary, a relevance score, and a commentary for each scene.

- Emotional Quantification: We compiled a comprehensive list of emotions expressed throughout the script. A separate LLM instance assigned valence and arousal scores to each emotion, allowing us to calculate average valence and arousal scores for each scene.

- Standardization: LLMs were employed to reformat dates into a consistent format compatible with Python libraries for conversion into timestamps. Similarly, location descriptions were standardized to facilitate the extraction of latitude and longitude coordinates.

Our data augmentation process resulted in a comprehensive dataset. Each entry includes the following fields: scene background context, narrator introduction, Van Gogh's first-person narrative, general context, expert commentary, characters involved, valence and arousal scores, timestamp, latitude, longitude, and a relevance score. This representation provides rich contextual information for subsequent integration with the episodic memory system and LLM.

4.2 Episodic Memory Graph System

Our episodic memory model employs an adaptive graph-based structure. In order to obtain relevant retrieved data, we do the following:

- (1) Initial Query Matching: We leverage cosine similarity to identify candidate memory entries based on their contextual alignment with a given query. These entries serve as initial "entry points" into the graph.
- (2) Multi-Factor Ranking: To rank the remaining entries, we compound the cosine similarity with scores derived from other relevant metadata: Emotional Distance: We calculate the Euclidean distance between the entry point's emotional valence/arousal coordinates and each remaining entry. These distances are normalized (0-1), prioritizing entries with closer emotional profiles. Spatiotemporal Distance: Similarly, we normalize the distances in latitude/longitude and timestamps between the entry point and other entries.
- (3) Compound Score: We multiply the normalized values from each factor (emotional, spatiotemporal) with the initial cosine similarity score. This produces a final score that modifies the ranking, prioritizing entries not only based on semantic similarity but also on their emotional and spatiotemporal proximity to the query.

This multi-factor approach aims to refine retrieval beyond purely embedding similarity. By incorporating emotional and spatiotemporal context, we enable the system to surface memories more closely aligned with the affective and situational nature of the query.

4.3 LLM integration

Our LLM integration is streamlined and leverages the previously designed episodic memory system. It consists of three steps:

- (1) Memory Retrieval: Upon receiving a query, the episodic memory system retrieves relevant entries from the database, subject to two constraints: a maximum entry count and a similarity threshold. This ensures focused retrieval without compromising relevance.

- (2) Context Construction: Retrieved entries are concatenated into a context prompt. We also include a general character description and task instructions (using standard prompt engineering techniques). To make the model adhere more to the augmented data, we also pass the raw retrieved values in the context (valence, arousal, relevance score). Finally, we include the chat history for as much as the context window allows. This provides the system with both short-term and long-term memory for each query.
- (3) Query Submission: The constructed context and query are then forwarded to the LLM for response generation.

Table 1 shows sample responses of the different systems from a query that we consider significant to Van Gogh's life: "Why did you cut your ear?".

Model Configuration	**Response (excerpts)**
Baseline LLM	... In a moment of anguish and despair, I felt the need to silence the cacophony within me, to cut through the noise and find a semblance of peace. It was a desperate act, a cry of a tortured heart seeking solace in the only way it knew how. ...
Traditional RAG	Ah, the echo of that fateful night reverberates through the chambers of memory. ... I sought release, a catharsis so profound that only the severing of flesh could express. ... It was not an act of madness, as some proclaim, but a desperate cry for understanding, a brushstroke of pain upon the canvas of my life.
Augmented RAG (Autonoesis)	... It was a moment of anguish, of turmoil that I can hardly put into words. ... It was a cry for help ... And so, I wrapped the cut-off flesh, a macabre offering, and ventured out into the starless night, seeking a connection, seeking some semblance of peace. But alas, my dear friend Gauguin was not there ...
Augmented RAG (Autonoesis + Ranked + Data)	... It was a moment of intense turmoil. I washed the cut-off flesh, wrapped it carefully (...), I sought out Gauguin, my dear friend (...). But he was not there, and in my despair, I left my offering at the brothel. (...) Gauguin's absence, his significance in my life (...) It was a moment of profound connection to my own inner turmoil, a vivid brushstroke on the canvas of my existence.

5 PRELIMINARY FINDINGS

We conducted a comparative analysis of four system configurations to evaluate the impact of our approach:

- Baseline LLM: GPT-3.5 without any RAG integrations.
- Traditional RAG:: GPT-3.5 with RAG using the original biographical data.
- Augmented RAG (Autonoesis): GPT-3.5 with RAG using the LLM-generated autobiography (including scene context), simulating autonoesis.
- Augmented RAG (Autonoesis + Ranked + Data): GPT-3.5 with RAG using the LLM-generated autobiography, ranking entries, and incorporating the top entry's numerical data into the context.

5.1 Key Observations

Our analysis revealed that the baseline LLM offered poetic but incomplete responses, lacking narration. The traditional RAG system, while adhering to the narrative, lacked depth. From other experiments, we found it also exhibited inconsistent pronoun use, sometimes referring to the character in the third person. The simulated autonoesis RAG yielded richer responses, introducing contextually relevant characters (Gauguin). Lastly, combining autonoesis with ranking and numerical augmentation produced the most focused, informative, and explanatory responses. This demonstrates our approach's potential to provide rich context to the LLM, improving its ability to generate nuanced, accurate, and consistent responses within the Van Gogh persona.

6 DISCUSSION

Within the domain of HCI research, we argue that the most effective utilization of LLMs lies in their potential as data augmentation tools. Rather than relying on them for zero-shot generation, we propose the development of robust memory and cognitive frameworks to provide LLMs with richer context, going beyond the limitations of traditional RAG systems. Our experiments demonstrate that this augmentation and contextualization approach yields more informative and focused responses.

We envision several compelling applications of this approach. By augmenting real participant data, we can create synthetic personae with cognitive models that partially imitate the original participants. This opens the door to extensive interviews with these personae, even in scenarios that may be stressful or sensitive for human participants. Additionally, our system offers a degree of explainability by providing access to augmented data and ranked retrieved scenes. This transparency allows researchers to explore and understand the reasoning behind the model's responses, a crucial advantage in HCI research.

Our framework's emphasis on single RAG searches and ranking algorithms ensures fast response times, making it suitable for real-time interviews. Furthermore, by offloading some of the data processing and self-reflection from the model, we potentially allow for embedding smaller, more efficient models into systems where computational resources are constrained. This has particular relevance in industries such as video game development, where GPU limitations prohibit loading large models.

The findings discussed in this paper represent an initial exploration into the intricate relationship between data augmentation, cognitive modelling, and LLM performance. It highlights the promise of this field and underscores the need for further research. This work aims to spark new investigations, igniting a deeper understanding of how tailored data sets and advanced memory frameworks can unlock richer, more nuanced interactions with language models.

REFERENCES

1 Haonan Chen, Zhicheng Dou, Kelong Mao, Jiongnan Liu, and Ziliang Zhao. 2024. Generalizing Conversational Dense Retrieval via LLM-Cognition Data Augmentation. arXiv preprint arXiv:2402.07092 (2024).

2 Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. arXiv preprint arXiv:2312.10997 (2023).

3 Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. arXiv preprint arXiv:2305.11738 (2023).

4 Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al . 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. arXiv preprint arXiv:2311.05232 (2023).

5 Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. 2023. Recommender ai agent: Integrating large language models for interactive recommendations. arXiv preprint arXiv:2308.16505 (2023).

6 Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. Comput. Surveys 55, 12 (2023), 1–38.

7 Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. Advances in Neural Information Processing Systems 36 (2024).

8 Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. arXiv preprint arXiv:2309.05922 (2023).

9 André Sant'Anna, Kourken Michaelian, and Nikola Andonovski. 2024. Autonoesis and episodicity: Perspectives from philosophy of memory. Wiley Interdisciplinary Reviews: Cognitive Science 15, 1 (2024), e1665.

10 Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. arXiv:2104.07567 (2021).

11 Andrew L Smith, Felix Greaves, and Trishan Panch. 2023. Hallucination or confabulation? neuroanatomy as metaphor in large language models. PLOS Digital Health 2, 11 (2023), e0000388.

12 Chenxi Whitehouse, Monojit Choudhury, and Alham Fikri Aji. 2023. LLM-powered Data Augmentation for Enhanced Crosslingual Performance. arXiv:2305.14288 (2023).

13 Junjie Ye, Nuo Xu, Yikun Wang, Jie Zhou, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. LLM-DA: Data Augmentation via Large Language Models for Few-Shot Named Entity Recognition. arXiv preprint arXiv:2402.14568 (2024).