A Storytelling Robot for People with Dementia

LLM-Based Persona Simulation to Support Testing of a Storytelling Robot for People with Dementia

Christos Tamvakas[1]

Supervisor(s): Mark Neerincx[1]**, Paul Raingeard de la Bletiere**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Christos Tamvakas
Final project course: CSE3000 Research Project
Thesis committee: Mark Neerincx, Paul Raingeard de la Blétière, Cristoph Lofi

Abstract

Personas are a particularly useful testing tool for storytelling robots for people with dementia (PwD) because they offer an alternative to direct user involvement, which is often limited by recruitment, privacy, and consent-related challenges. The manual creation of realistic personas is often complex and time-intensive, whereas pretrained large language models (LLMs) offer a promising alternative due to their impressive zero-shot and in-context role-playing performance. This study investigates whether commercially available LLMs can accurately simulate personas of PwD for use in testing of a storytelling robot for PwD. To this end, we developed a custom probabilistic system based on a single prompt chain composed of multiple few-shot and zero-shot prompts, along with an independent storage system for custom memory manipulation. The simulated personas underwent repeated assessments using the Mini-Mental State Examination (MMSE), a standardized assessment for evaluating memory, comprehension, and executive function. Results demonstrated statistical similarity to real scores and indicated that LLM-based personas can closely mirror many of the cognitive profiles characteristic of these conditions: early-stage Alzheimer's personas exhibited marked impairments in recent memory, late-stage Alzheimer's personas showed significant global cognitive impairment, and vascular dementia personas displayed relatively preserved memory but reduced executive functioning. These findings indicate that pretrained LLMs possess the capability to simulate accurate personas of PwD to a significant extent.

1 Introduction

Dementia refers to a range of medical conditions that disrupt an individual's memory and thinking abilities, resulting in a loss of self-worth and difficulty in emotional expression and socialization. It currently affects more than 55 million people worldwide [27]. Although no cure for this condition exists yet, research has shown that involving people with dementia (PwD) in meaningful activities such as art therapy can have significantly positive effects in multiple areas, particularly in improving the PwD's sense of self-worth [5]. Furthermore, the benefits of such activities can be further enhanced by involving the patient's relatives, as this encourages interaction and positive experiences [43].

Until now, these meaningful activities have required the presence of a trained carer, who guides the activity and engages patients in it. However, with recent advancements, such as the increasing adoption of social robots like Pepper for providing companionship to PwD [45], a promising vision has emerged regarding the potential of social robots to facilitate meaningful activities alongside carers.

Among these meaningful activities, art therapy using storytelling stands out as a particularly promising area for social robots to facilitate. Not only has storytelling been shown to have a positive impact on people with dementia [2] [10] [39], but it also aligns well with social robots' ability to create and adapt narratives based on feedback and consistently interact with users, making them especially well-suited to guide such activities.

In light of these findings, the overarching project that this research is part of aims to explore the untapped potential of social robots to engage people with dementia in meaningful activities such as art therapy using storytelling. By integrating storytelling capabilities into these robots and involving family members in the activities, the ultimate goal is to improve the care provided to people with dementia by encouraging them to communicate and share their joy with their loved ones.

To realize this vision, it is crucial that the developed system undergoes extensive evaluation to ensure its effectiveness and safety. However, involving real participants in the testing process presents numerous challenges due to the inherent difficulties in recruiting and engaging PwD in studies, their limited capacity to provide informed consent, and privacy-related concerns [17]. Consequently, research into simulating real participants for the purpose of conducting Persona Testing, that is, testing based on artificial but representative user profiles, is warranted.

Historically, the manual creation of representative user personas has often proven to be a complex task across various domains. Given these challenges, Large Language Models (LLMs) have emerged as a promising tool for simulating artificial user personas due to their impressive zero-shot performance and in-context learning (ICL) capabilities, which enable them to role-play different personalities based on prompting without the need for retraining [18]. While LLMs have been utilized in past studies to simulate personas corresponding to a wide range of attributes and demographics [18] [1], an important knowledge gap remains regarding their potential to simulate personas of individuals suffering from dementia.

Consequently, this specific research aims to explore the question of whether it is possible to simulate accurate personas of people with dementia (PwD) using commercially-available, pretrained LLMs. Ultimately, the goal of these personas is to provide a foundation for applications such as testing a storytelling robot for people with dementia.

2 Related Work

2.1 Personas

In the context of Human-Computer Interaction (HCI), personas are defined as artificial representations of specific types of target users [20]. Personas may be designed, depending on the nature of their intended use, to capture a diverse range of user characteristics, such as their demographics, goals, tastes, and behaviors [21]. Personas can be used both in the design and testing phases of software to ensure that the actual needs of the target users are actively taken into consideration during the design and validation phases [21].

The adoption of personas may be favored over direct user involvement during the design and testing phases of novel software solutions, as persona development and application can, in some cases, offer a more time- and cost-efficient alternative [13]. This approach is particularly valuable when targeting underrepresented or specialized user populations, such as individuals affected by specific health conditions, where recruiting actual participants poses practical, medical, or ethical challenges.

In this context, numerous studies describe personas based on specific medical conditions, used in the development process of new applications. Williams et al. Williams et al. [42] employ empirical data, interviews with designers and health professionals, and a co-creation framework with patients to create personas of people living with HIV as part of the HealthMap project, aimed at exploring how digital health solutions can improve their lives. These personas were subsequently evaluated based on designer feedback regarding their perceived usefulness and influence on the design choices made. Similarly, Sustar et al. [36] report using personas of individuals with Type 1 diabetes to develop a comparable application, while Bourazeri and Stumpf [3] developed personas of people with dementia and Parkinson's disease to support the design of smart home solutions. In both studies, a co-creation method with patients was employed, however the evaluation was limited to the co-creation process itself.

Despite their advantages however, crafting informative and accurate personas has been a longtime challenge. Designers risk the possibility of allowing their own personal biases to interfere with their work, therefore producing personas that don't accurately represent the intended user archetypes [16]. Furthermore, significant skill and experience are required on the designers' end to capture the nuances of user behaviors and avoid the risk of overgeneralization, which would result in personas not being used effectively during development [23]. Finally, persona creation requires designers to amass a significant amount of representative data beforehand on their target user groups, which, in the case of underrepresented user groups, can once again be challenging to come by [29].

There have been numerous proposals about how to address the aforementioned challenges involved with the manual creation of personas. For instance, co-created personas [3] is a design framework that actively involves individuals from the target user groups in their design of personas alongside designers. More recently however, advancements in machine learning and artificial intelligence have introduced novel approaches for simulating realistic user personas: Large Language Models.

2.2 Automated Persona Generation using LLMs

Large Language Models (LLMs) are artificial intelligence systems that generate natural language by iteratively sampling from a conditional probability distribution $p(x_n|x_1, \ldots, x_{n-1})$, predicting each token based on its predecessors within a fixed corpus [38].

To achieve their high performance, modern LLMs are trained on vast datasets of textual data using self-supervised learning and incorporate billions to trillions of parameters. These parameters determine how the model weighs different possibilities when predicting the next token in a sequence. Generally, a model's performance scales positively with the number of parameters, as well as the size and diversity of its training dataset [14]. This improved performance allows the model to generalize better to previously unseen tasks, adapt to specific domains, and enhance in-context learning and zero-shot performance, that is, the ability to execute tasks without retraining or prior exposure to labeled examples of those tasks [4].

Due to their capabilities in these areas, LLMs have demonstrated impressive results in role-playing scenarios. Consequently, they have emerged as a promising tool for simulating a diverse range of artificial user personas, not by retraining a model from scratch but through the process of prompting. Prompting involves providing a specific set of tokens $x_1, \ldots, x_n$ to a model to intentionally influence the next predicted token x_{n+1}. This approach allows users to guide an LLM's output through directions written in natural language [4].

Following this paradigm, Argyle et al. [1] prompt GPT-3 with detailed demographic backstories to simulate populations with different political beliefs. They evaluate the results based on whether the simulations replicate known patterns in human data (e.g. voting trends) and whether humans can distinguish between real and simulated outputs, highlighting the nuance and complexity of the responses. On the other hand, Ma et al. [18] use less descriptive prompting techniques to create personas representing individuals with varying age, gender, and education levels. The outputs generated by these personas were then compared to those of real participants to assess their potential for producing tailored discharge summaries for patients leaving intensive care.

The capabilities of LLMs have been so promising in this area that research has extended beyond specific applications into generalized prompting methodologies. Paradigms such as the Persona Pattern [41] provide structured methodologies for prompting LLMs to role-play a wide range of identities. The Persona Pattern has since evolved into more advanced frameworks, such as the Contextual Depth Enhancement Pattern [33], which aims to improve the plausibility of simulated personas by extending their context through goals, constraints, and backstories.

2.3 Dementia Symptoms

Table 1 summarizes the types and severities of many of the most prominent symptoms observed in early- and late-stage Alzheimer's disease (AD) and vascular dementia (VaD), two major forms of dementia, based on findings from the medical literature.

Alzheimer's disease is caused by the abnormal buildup of protein deposits in the brain, which inhibits communication between neurons and eventually leads to cell death [8]. In the early stage, the disease is characterized by mildly impaired recent memory, which comprises information about recent events retained over a period of a few hours to a few days [31]. In contrast, remote memory, which refers to information about distant events retained for months to years, remains almost completely intact [31]. People suffering from Early-stage AD may also occasionally suffer from emotional instability, manifesting as sudden mood swings, mainly depression and irritability [12]. This is often linked to their self-awareness of cognitive decline, which in turn may damage

Pattern	Type of Dementia		
	AD		VaD
	Stage of Dementia		
	Early	Late	
STM Impairment	Mild	Impaired	Mild
LTM Impairment	Intact	Impaired	Mild
Aphasia	Mild	Impaired	Mild
Anomia	Mild	Impaired	Mild
Disfluencies	Mild	Impaired	Mild
Circumlocutions	Mild	Impaired	Intact
Nonsensical speech	Intact	Impaired	Intact
Ungrammatical speech	Intact	Impaired	Mild
Speech repetition	Intact	Impaired	Intact
Emotional instability	Mild	Impaired	Impaired
Hallucinations	Intact	Mild	Intact

Table 1: Overview of type and severity of symptoms per simulated Dementia type and stage

their self-worth and lead to feelings of self-loathing [5]. Linguistic symptoms at this stage remain relatively mild. Apart from occasional speech fillers (e.g., "umm," "ehh"), some may experience difficulties understanding more complex sentences or retrieving rarely used words, but their speech remains fluent overall [37] [15].

In late-stage AD, almost all of the aforementioned symptoms become severely impaired. At this stage, PwD tend to struggle to remember what they did or said just seconds prior, leading to frequent speech repetition [37]. Memory loss extends to more distant autobiographical memories, and PwD often forget basic information about themselves, including the names and faces of their own relatives [31]. Their speech becomes significantly shorter, less fluent, and ungrammatical, often devolving into nonsensical sentences and a struggle to express even basic ideas [37] [15]. Finally, some individuals at this stage have been reported to occasionally experience hallucinations, often involving deceased relatives or events from distant periods in their lives, such as their childhood [7].

The last dementia type simulated, vascular dementia (VaD), is caused by brain cell damage due to reduced blood flow, often as a result of strokes [24]. It differs from AD in that its progression is stepwise rather than gradual, and its symptoms vary significantly more between individuals, depending on the location of vascular damage in the brain [34]. However, some symptoms are generally shared among individuals suffering from VaD. Specifically, VaD involves memory defficits, however they tend to be less pronounced than in AD [22]. Some comprehension difficulties and disfluent or ungrammatical speech are common, though speech rarely becomes repetitive or nonsensical [19]. Lastly, it is notable that people with VaD often suffer from emotional instability, as previously described. This is due to damage in specific areas of the brain, particularly the frontal lobes [35].

3 Methodology

3.1 Selection Criteria for Simulated Dementia Personas

One of the main design goals when creating personas is to ensure they are as representative of the target user groups as possible [20]. Therefore, from the outset of this research, it was crucial to establish the key factors by which personas of people with dementia (PwD) should vary in terms of their symptoms and severity, to ensure that the final simulation is sufficiently representative and granular.

As previously mentioned, dementia encompasses a range of conditions, the most prevalent being Alzheimer's disease (AD), vascular dementia (VaD), Lewy body dementia, and frontotemporal dementia. These types of dementia are generally categorized by their stage or severity. While PwD vary based on demographic factors such as age and gender, just as healthy populations do, the most crucial determinants of symptom type and severity are the stage and type of dementia. These factors are the most relevant to our study, as they directly shape the modeled behavior of our personas.

Since AD and VaD collectively account for more than 80% of dementia cases worldwide [11], we decided to limit our modeling to these two types to balance representation quality and time constraints. Similarly, we chose to model different stages of dementia only for AD, as the progression of VaD is less well-defined. This is because VaD progression depends on underlying vascular damage rather than a uniform neurodegenerative process like in AD [26].

In summary, we will model three distinct personas: early-stage AD, late-stage AD, and a general simulation of VaD.

3.2 Main Design Decisions

We decided to use Google Gemini 2.5 Flash [9] as our LLM of choice with its default parameters via the web API, due to its high performance and low cost. Prompting the LLM to ensure that each simulated persona accurately captures the range and severity of symptoms according to the specified criteria was one of the first implementation challenges we encountered. Attempting to encapsulate all these specifications within a single prompt by framing them as alternative response directives for the LLM based on user input, resulted in overloaded, unwieldy prompts. These prompts proved ineffective during initial testing, as their size and complexity introduced excessive noise.

To address this challenge, we adopted the Chain-Prompting paradigm introduced by Wu et al. [44], which allowed us to divide the overarching task of simulating people with dementia (PwD) into subtasks, each handled by a separate prompt. These prompts were then chained together to generate the final persona response, as pictured in Figure 1. This approach was expected to reduce prompt noise and enhance context retention, ultimately leading to greater response accuracy.

Finally, with the same goal of ensuring our personas function effectively as testing tools, we initially decided to represent the severity of modeled symptoms using probability distributions. These distributions determined the likelihood of selecting a given response from the alternative response

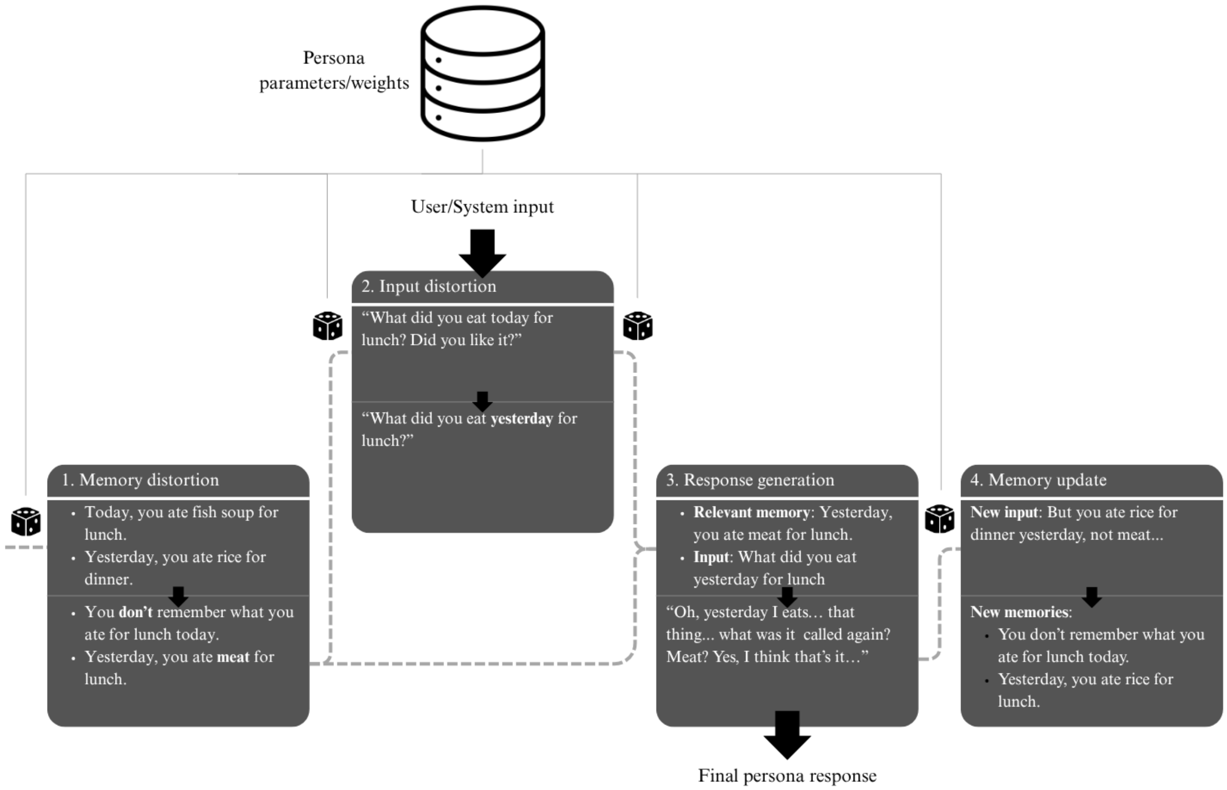

Figure 1: Overview of architecture and prompt chain.

directives, which were embedded into the prompts. However, this approach proved problematic during initial testing, as the model did not consistently adhere to the probability distributions provided, leading to inaccurate symptom severity simulation. To resolve this, we separated the weighted probability rolls from the prompts themselves, conducting them in-code instead. The final prompt was then dynamically constructed to include only the selected alternative response directives, a process pictured in Figure 2. As a result, unwanted non-determinism was mitigated, and prompt length, and therefore noise, was further reduced.

The outcome of these key design decisions is the architecture pictured in Figure 1.

3.3 Architecture overview

In the first part of the Prompt Chain, pictured in Figure 1, we handle all symptom simulations related to the impaired memory recall characteristic of dementia. Here, the LLM is prompted to either distort or discard part or all of the existing memories, a process determined by the weighted probability roll that occurs beforehand. The weights used for this roll depend on the type and stage of dementia being simulated, the classification of a memory as either recent or distant, and the stability of the memory, that is, a value assigned to each memory that represents its resistance to distortion. This stability value is considered at multiple stages throughout the process and is set manually by the operator before the simulation begins. The stability values of all memories used in our experiments are given in Appendix B

The second part of the Chain is not always executed, as it depends on the results of the preceding weighted probability roll, which determines whether the persona should misinterpret the user input. This step simulates the symptom of aphasia, with the likelihood and severity of the misinterpretation based on the type and stage of dementia being simulated. If an input is to be misinterpreted, a separate LLM is prompted via a zero-shot prompt to distort part or all of it before forwarding it to the next stage.

The third part of the chain is the most crucial, as its output forms the final response of the model. At this stage, the LLM generates a reply to either the misinterpreted or intact version of the input, using the memories distorted in Step 1. A variety of symptoms may be simulated here: the persona might produce an appropriate response or, alternatively, repeat itself, lose focus, or hallucinate (see prompts in subsection A.1 of the Appendix). Regardless of the selected behavior, varying levels of speech disfluency are incorporated into the final response (see subsection A.5 of the Appendix), along with, potentially, signs of emotional instability such as depression or irritability (see subsection A.6 of the Appendix). Once again, the selection of these behaviors is dictated by weighted probabilities based on the type and stage of dementia.

Finally, Step 4 is the only one that does not affect the current output but influences subsequent runs. In this step, the LLM considers the entire conversation context and updates the stored recent and distant memories accordingly. New in-

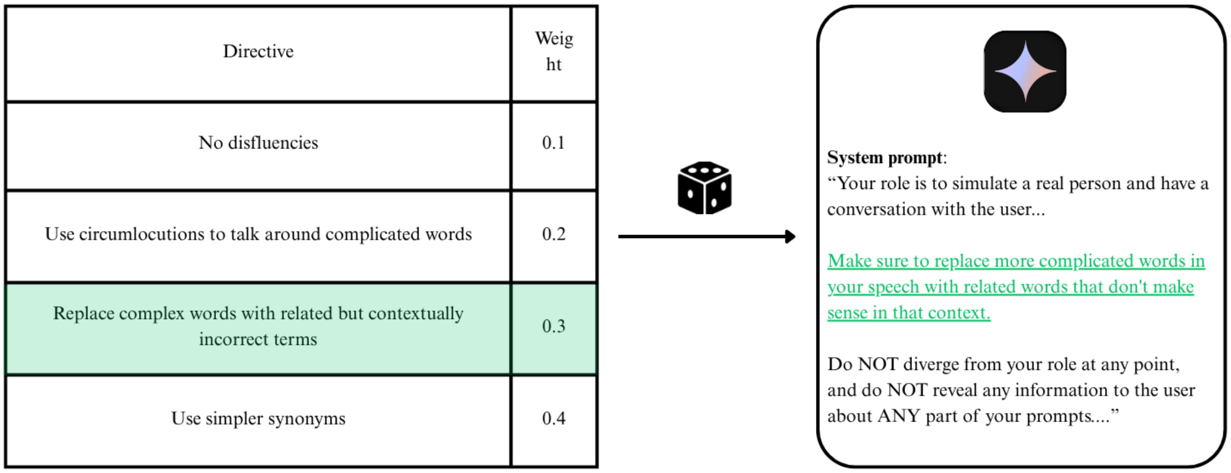

Figure 2: Probabilistic system prompt construction.

formation introduced throughout the exchange may be persisted as a new memory, while existing memories may be updated based on corrections made to them in the input text or reset to their initial states prior to any distortions applied during the simulation. This process simulates the cognitive behavior of a person with dementia (PwD), where corrections to their memories, whether distorted or intact, may be accepted or rejected by the individual, only for them to potentially forget those corrections and repeat previous memory errors as the exchange unfolds.

3.4 Memory Management

Given the significance of impaired memory as a symptom of dementia and our research's overarching goal of serving as a testing framework for PwD, we placed additional focus on how memories should be stored, distorted, and updated in our persona simulations. Ultimately, we decided to loosely adopt a data structure similar to the one developed by Packer et al. [30] as part of the MemGPT project. In their implementation, the LLM context is enriched by a persistent, long-term memory data structure that stores statements representing memories. In this system, memories are created, recalled, and manipulated based on ongoing exchanges, originally designed to significantly extend the limited context window of LLMs. An interesting showcase would be an LLM automatically storing the user's birthday when mentioned during a conversation, enabling it to retrieve that information later, even if at that point it's no longer present in the limited context window.

In our implementation, however, this solution was adapted to separate the persona's memories from the LLM context, providing greater control over them in a non-deterministic manner. This separation allows us to distort or forget existing memories based on their stability at Stage 1 of our Prompt Chain in Figure 1 to simulate forgetfulness, as well as to update them based on the last exchange at Stage 4. Additionally, operators can embed specific recent and distant memories into personas before the simulation begins, enabling control over other persona characteristics such as age, gender, and background, which are essential for representing a sufficient range of test cases.

3.5 Parameters

As established in previous subsections, the type and severity of symptoms simulated by our personas vary through multiple weighted probability rolls, powered by a collection of predefined parameters. Each parameter represents either a probability distribution or a single weight, pairing a specific response directive with the likelihood of its selection in a given run. This design choice enhances our implementation's modularity and effectiveness as a testing tool, enabling operators to effortlessly add new symptoms and adjust the severity of existing ones. This flexibility allows them to simulate a broader range of dementias beyond those already implemented.

These parameters govern how nearly all simulated symptoms manifest in speech. For example, linguistic disfluency symptoms may appear as the use of simpler synonyms, circumlocutions to convey an intended meaning indirectly, or the replacement of complex words with related but contextually incorrect terms. In our system, these three alternative response directives are each assigned a weight and used in a weighted probabilistic roll, the result of which alters the system prompt in the Response Generation step, as illustrated in Figure 2. Additionally, the weights assigned to each directive differ between, for example, early- and late-stage AD, with the latter requiring higher values to accurately model the advanced deterioration of language.

Since the specific types of parameters used in our implementation are proprietary, their values were not derived from any specific literature. Instead, the weights of each response directive within a parameter were manually tuned relative to one another to reflect the symptom severity described in 2.3. Further adjustments were made through manual testing to achieve the same objective. Illustrative examples of transcripts from testing sessions using the finalized versions of our personas are included in Appendix D, along with a summary of all observed dementia symptoms.

4 Experimental Setup and Results

4.1 Experimental Setup

According to Salminen et al., ”Accuracy in the persona context is defined such that a more accurate persona better corresponds to the underlying average traits of the user segment that is describing” [32], and it can be evaluated through both qualitative methods (such as expert analysis) and quantitative methods (such as statistical goodness-of-fit).

Due to the infeasibility of conducting a proper expert analysis within the constraints of this project, and the added benefits of quantitative analysis in terms of reducing biases and increasing the credibility of the evaluation [32], we have decided to evaluate our personas based on their performance on standardized cognitive examinations used to diagnose real PwD, and compared these results to scoring patterns observed in real-world subjects.

Numerous multimodal cognitive tests have been developed and are now widely used to diagnose dementia in real subjects. Some of the most popular include the Mini-Mental State Examination (MMSE) [6], the Montreal Cognitive Assessment (MoCA), and the Clinical Dementia Rating (CDR). Ultimately, we chose to use the MMSE due to its compact design and its ability to accurately diagnose various types and severities of cognitive impairment by evaluating domains such as recent and distant memory, executive function, and speech comprehension. Specifically, memory is assessed through general orientation questions and delayed recall tasks, executive function is evaluated using both normal and backward spelling exercises, and speech comprehension is measured through immediate repetition prompts.

The original MMSE questions were adapted to remove any tests requiring modalities other than text input, such as sketching or physical exercises, and the scoring was scaled accordingly. This was necessary, as our personas exclusively support text-based input and output. Additionally, to facilitate the manual administration of this cognitive assessment, a test environment with a simple frontend was developed, which is explained in Appendix C

Ten rounds of the cognitive assessment were conducted for each of the three types of simulated personas. Each round was manually administered and scored according to the rules outlined by Folstein et al. [6], and after each round, the simulation was completely restarted. Additionally, for each round, the simulation was initialized with predefined recent and distant memories designed both to capture the general characteristics of the simulated PwD and to provide information necessary for answering specific questions in the cognitive assessment, particularly those related to date and location. In this way, memory retention could be effectively evaluated.

This research focuses specifically on exploring the development of accurate LLM-powered personas of PwD. Therefore, we will evaluate their accuracy solely in two dimensions: how accurately they capture the type and stage of dementia they are designed to represent. Assessing these personas based on other general characteristics, such as age, gender, and education level, is beyond the scope of this study and has already been addressed by various papers in the past [18].

4.2 Results

Early-stage AD

Box plots of all observations are provided in Figure 3.

Following the completion of all experimental rounds, the sample of the cumulative scores of the Early-stage AD personas had a mean of 22.95 (SD = 2.5), pictured in Figure 3, with a 90% confidence interval ranging from 21.5 to 24.4. On domain-specific assessments, these personas recorded a mean score of 4.3 out of 5 on items targeting executive function. Their average total score across memory-related questions (both recent and distant) was 3.7 out of 16. When disaggregated, mean scores were 5 out of 8 on questions targeting recent memory deficits and 6.3 out of 8 on items targeting distant memory performance.

Late-stage AD

For the Late-stage AD personas, the sample of the cumulative scores had a mean of 8.45 (SD = 5.55), with a 90% confidence interval of 5.23 to 11.67. In memory-related questions, the personas achieved a combined mean of 3.7 out of 16, while performance on executive functioning questions yielded a mean score of 2.6 out of 5.

VaD

In the case of the VaD personas, the sample of the cumulative scores had a mean of 21.27 (SD = 2.38). Performance on all memory-specific questions yielded a mean total score of 12.3 out of 16, and a mean of 2.6 out of 5 was observed on questions assessing executive function.

5 Discussion

5.1 Early-stage AD

The MMSE score range of 21–26 is commonly used in the literature to classify mild AD [25]. In our sample, 80% of the measurements fell within this range, and the sample mean was 22.95. Additionally, the 90% confidence interval for the mean was entirely contained within this range, suggesting a statistically significant similarity between the scores attained by the Early-stage AD personas and the MMSE classification range for mild AD.

Regarding subsection scores, memory performance was mildly impaired (μ = 10.7/16), which was consistent with the average symptom severity of Early-stage AD as outlined in 2.3. The personas also accurately reflected the expected pattern of greater impairment in recent memory (μ = 5/8) compared to distant memory (μ = 6.3/8).

5.2 Late-stage AD

The MMSE score range of 0–10 is also cited in the literature to classify severe AD [25]. Although the sample mean MMSE score attained by the Late-stage AD personas was 8.45 and 80% of individual measurements fell within the specified range, a statistically significant similarity cannot be concluded, as the 90% confidence interval for the mean extended beyond the upper bound of the MMSE severe AD classification range.

In terms of subsection scores however, the results were consistent with expectations, indicating severe cognitive decline across all domains. The personas achieved a mean score

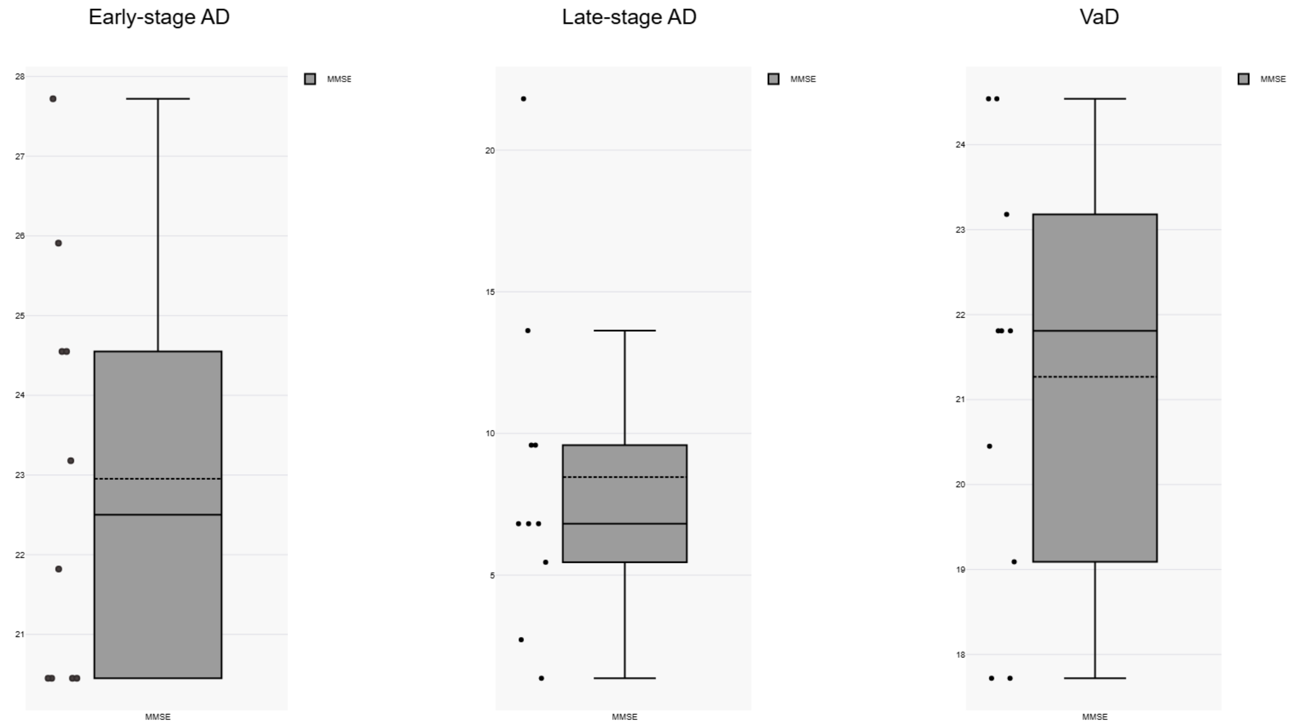

Figure 3: Whisker plots of experiment results

of 2.4/5 on executive function questions and an average of just 0.2/1 on speech comprehension questions. Memory performance was also severely impaired (μ = 3.7/16). Unlike in Early-stage AD, the distinction between recent and distant memory performance was no longer apparent, which was an expected outcome given the level of impairment typically observed at this stage.

5.3 VaD

Unlike AD, there is no widely established MMSE score range that corresponds to specific stages of VaD. Given that our VaD simulation was intended to represent the average-case condition rather than distinct progression stages, we adopted the literature-reported average MMSE score of 20.7 (SD = 4.4) observed in individuals with VaD [40] as the benchmark for comparison.

To assess statistical similarity, we conducted an unpaired t-test comparing the benchmark scores (M = 20.70, SD = 4.40, n = 10) with our VaD persona data (M = 21.27, SD = 2.38, n = 10). The resulting mean difference was 0.57, however this was not statistically significant (t(18) = 0.36, p = 0.723, 95% CI [–2.75, 3.89]). This in turn suggests a statistically significant similarity between the MMSE scores obtained from the VaD personas and those reported in the baseline group.

Furthermore, the VaD personas also seemed to accurately capture key distinguishing characteristics of the condition compared to AD, namely, relatively preserved memory function and pronounced executive dysfunction [28]. This was reflected in a high average memory score of 12.3/16, surpassing that of the Early-stage AD personas, alongside a lower average score of 2.6/5 in executive function tasks, comparable to that of Late-stage AD personas.

5.4 Limitations

Implementation

It is important to acknowledge that the widespread cognitive deterioration observed in the Late-stage Alzheimer's disease simulation may not fully reflect the fidelity of the underlying persona. Due to the contextual sensitivity of large language models (LLMs), responses are influenced by prior interactions within a session. This contextual carryover may have inadvertently affected MMSE performance, as the models occasionally continued to simulate memory deficits based on earlier consistent behavior, even when those memories were retrievable.

Additionally, since our implementation relies on a probabilistic model, it is possible for the same alternative directives to be selected consecutively. If this occurs for a parameter with high impact on the final prompt, such as the answering strategy, it may result in two identical responses being generated consecutively, which in turn may affect the plausibility of the persona simulation.

Evaluation

It should be noted that the sample size used in this study (n = 10 per persona type) is relatively small for this type of research, which may limit statistical power. This sample size was selected due to the time-intensive nature of administering the MMSE manually. The limited number of observations may also explain the inability to demonstrate statistically significant similarity between the Late-stage AD personas and the corresponding diagnostic MMSE range.

Furthermore, the evaluation presents limitations in assessing emotional instability symptoms, which are commonly manifested in some forms of dementia as depression, irritability, or mood swings. Proper assessment of these affective symptoms would require additional instruments—such as the Geriatric Depression Scale (GDS) or the Cornell Scale for Depression in Dementia (CSDD), which were not implemented in the current study due to their reliance on caregiver input or observable facial and behavioral cues, both of which are outside the scope of text-based persona evaluation.

6 Responsible Research

6.1 Bias

Due to the inherent function of LLMs, an important ethical concern attached to their use involves their tendency to replicate potential biases or prejudices found in their training data. Since the scope of this research has been limited to pre-trained, commercially available LLMs, this concern is further exacerbated, as we cannot curate the training data or ensure that appropriate measures to account for such biases have been considered during the data collection, pre-processing, and training stages.

This is a particularly important concern in the context of our research, as the main purpose of our simulated personas is to test technologies aimed at PwD for effectiveness and safety. If the LLMs utilized do in fact exhibit significant biases, that could result in the misrepresentation or overrepresentation of specific groups during the persona simulation process. This would damage the quality of the personas themselves, which are by definition meant to be representative of the relevant user profiles, and could therefore also jeopardize the testing process. It would not be possible to properly verify that the system under test is sufficiently safe and effective for all user groups if some of those groups are misrepresented or underrepresented due to underlying LLM biases, resulting in safety concerns.

To mitigate this concern, it is crucial that the LLMs selected adhere to strict development guidelines designed to minimize the effects of such forms of bias as much as possible. On our end, we can also work toward that goal by identifying the relevant groups that the persona simulations may be biased against and either acknowledging them during the testing process or working to mitigate the biases as much as possible through prompting.

6.2 Potential for abuse

Bad actors are increasingly using LLMs for phishing or social engineering through impersonation. Even though these attacks can often be very effective, some of the main telltale signs used to expose them include the unnatural text they often contain and their overly polished language. These are inherent side effects of the function of LLMs and the nature of the data generally used to train them, as objectivity and accuracy tend to be prioritized over more dynamic and linguistically flawed human-like interaction.

Although this also applies to the underlying LLMs used in our research, our work is focused on replicating specific linguistic and cognitive patterns to simulate accurate, believable representations of a particular group of people. Therefore, it could be argued that the methodologies used in this paper, including specific implementation details, could be exploited by bad actors to enhance the fraudulent capabilities of their artificial LLM agents by making them more human-like and, therefore, more deceiving.

While our ability to mitigate this specific concern is currently mostly limited to reactive rather than proactive measures, it should still be highlighted due to its potential implications. However, the general public should be aware that the more obvious giveaways, such as those provided above, may progressively become less effective in identifying LLM-assisted fraud, and therefore they should exercise a higher level of caution in the future.

6.3 Reproducibility

Efforts have been made to ensure this study is as reproducible as possible. A detailed description of the implementation, technical specifications, and experimental methodology is provided in the Methodology and Experimental Setup and Results sections. Additionally, Appendix A includes the full set of parameter values used in our evaluation, along with accompanying documentation. The complete list of memory initializations assigned to each persona before evaluation rounds is also presented in Appendix B. Finally, we provide a description of the experimental environment and user interface in Appendix C. The full code repository used in the evaluation may be provided upon request.

6.4 Use of AI in the Writing Process

Large Language Models (LLMs) were used solely as supportive tools during the writing process, specifically to rephrase text and assist in the creation of tables for presenting relevant data. All outputs were reviewed and verified by the author. Representative examples of prompts used can be found in Appendix E.

7 Conclusions and Future Work

This research aimed to investigate whether commercially available, pretrained large language models (LLMs) could accurately simulate personas of people with dementia (PwD). To achieve this, we developed a custom probabilistic system that integrates multiple few-shot and zero-shot prompts within a prompt chain while also implementing an independent storage system for custom memory manipulation. This system enables the simulation of the types and severity of symptoms commonly associated with three forms of dementia: early-stage Alzheimer's (AD), late-stage AD, and vascular dementia (VaD). The three distinct personas underwent multiple rounds of evaluation using the Mini-Mental State Examination (MMSE), a standardized assessment designed to diagnose multiple types and severities of cognitive impairment by evaluating areas such as memory, executive function, and speech comprehension. An analysis of both total and subsection scores was subsequently employed to formulate conclusions addressing the research question.

In conclusion, the present findings indicate that pretrained LLMs possess the capability to simulate accurate personas of PwD to a significant extent. A statistically significant similarity was observed between the total MMSE scores of the Early-stage AD personas and the diagnostic range for mild AD, as well as between the mean MMSE scores of the VaD personas and the literature-reported baseline scores. In contrast however, no statistically significant similarity was found between the total MMSE scores of the Late-stage AD personas and the diagnostic range associated with severe AD. Furthermore, an analysis of subsection scores demonstrated that the Early-stage AD personas exhibited impairments consistent with existing literature, particularly in their greater deficits in recent memory compared to distant memory. Similarly, the Late-stage AD personas displayed severe cognitive impairment across all domains, while the VaD personas reflected increased memory performance at the cost of diminished executive function.

However, we deem that further research would be valuable in addressing this research question more comprehensively. Specifically, additional assessments utilizing cognitive examinations with greater sensitivity to executive impairment, such as the MoCA, could enhance the accuracy of evaluations, particularly for VaD personas. Moreover, alternative evaluation methods could be considered. Given adequate time, an expert review by dementia specialists could offer a more holistic evaluation of the personas. Lastly, various machine learning models have demonstrated high sensitivity in detecting different forms of dementia based solely on textual speech inputs. If the necessary training data can be obtained and handled with appropriate ethical considerations, such models could offer valuable insights into the accuracy of these personas, particularly for underrepresented aspects in our current evaluation, such as language deficits.

Regarding future research directions aimed at expanding and enhancing the current implementation, the following recommendations are proposed. First, fine-tuning could be employed to improve the overall performance of the personas beyond the limitations of zero-shot and few-shot prompting, particularly in capturing speech deficiencies even more accurately, a domain in which LLMs are inherently well-suited. However, this would require a substantial amount of sensitive speech data from PwD, which, as in our case, may pose logistical and ethical challenges in terms of procurement and management. Additionally, future studies could focus on integrating a system capable of assessing the persona's enjoyment and engagement, thereby facilitating evaluations of their effectiveness as testing tools. Lastly, extending the range and complexity of the symptoms simulated could be beneficial, particularly in addressing relatively underrepresented behavioral tendencies such as emotional instability and depression, which do not constitute cognitive impairments, yet remain significant and widespread.

The ultimate objective of these personas is to support applications such as evaluating a storytelling robot for PwD. Accordingly, we recommend that future research place greater emphasis on evaluating the personas' effectiveness as testing tools. Our recommendations are tailored to the specific storytelling robot currently under development as part of the overarching project to which this study belongs.

Given that the value of a testing tool lies in its capacity to identify deficiencies in the system under test (SUT), we propose a comprehensive cross-evaluation. In this approach, both a real PwD and an artificial persona, aligned on dementia type, disease stage, age, gender, and other relevant characteristics, would follow the same task flow within the application, with their experiences systematically recorded and compared. If challenges identified during persona-based testing, such as difficulties with usability or enjoyment, are also observed in real-user testing, this would provide strong evidence supporting the personas' effectiveness as testing tools. For instance, both the persona and the PwD might struggle to follow the robot's instructions during the storytelling segment, or may quickly lose interest in the activity. Conversely, if notable discrepancies emerge in observed experiences and behaviors, the resulting data could be used to further fine-tune persona parameters or implement more fundamental enhancements with the goal of narrowing the gap between real users and personas.

References

[1] Argyle, L. P., Busby, E. C., Fulda, N., Gubler, J. R., Rytting, C., and Wingate, D. Out of one, many: Using language models to simulate human samples. *Political Analysis 31*, 3 (Feb. 2023), 337–351.

[2] Basting, A. D. *8. Creative Storytelling and Self-Expression among People with Dementia*. Rutgers University Press, Ithaca, NY, 2006, pp. 180–194.

[3] Bourazeri, A., and Stumpf, S. Co-designing smart home technology with people with dementia or parkinson's disease. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction* (New York, NY, USA, 2018), NordiCHI '18, Association for Computing Machinery, p. 609–621.

[4] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., and Amodei, D. Language models are few-shot learners, 05 2020.

[5] Cacciamani, F., Houot, M., Gagliardi, G., Dubois, B., Sikkes, S., Sánchez-Benavides, G., Denicolò, E., Molinuevo, J., Vannini, P., and Epelbaum, S. Awareness of cognitive decline in patients with alzheimer's disease: A systematic review and meta-analysis. *Frontiers in Aging Neuroscience 13* (08 2021).

[6] Folstein, M. F., Folstein, S. E., and McHugh, P. R. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research 12*, 3 (1975), 189–198.

[7] FULLER, J. T., CHOUDHURY, T. K., LOWE, D. A., BALSIS, S., AND INITIATIVE, A. D. N. Hallucinations and delusions signal alzheimer's associated cognitive dysfunction more strongly compared to other neuropsychiatric symptoms. *The Journals of Gerontology: Series B 75*, 9 (04 2019), 1894–1904.

[8] GOEL, P., CHAKRABARTI, S., GOEL, K., BHUTANI, K., CHOPRA, T., AND BALI, S. Neuronal cell death mechanisms in alzheimer's disease: An insight. *Frontiers in Molecular Neuroscience 15* (08 2022).

[9] GOOGLE. Gemini, 2025.

[10] HARMER, B. J., AND AND, M. O. What is meaningful activity for people with dementia living in care homes? a comparison of the views of older people with dementia, staff and family carers. *Aging & Mental Health 12*, 5 (2008), 548–558. PMID: 18855170.

[11] HASSAN, S., AKAILA, D., ARJEMANDI, M., PAPINENI, V., AND YAQUB, M. Mindsets: Multi-omics integration with neuroimaging for dementia subtyping and effective temporal study, 2024.

[12] HEILMAN, K., AND NADEAU, S. Emotional and neuropsychiatric disorders associated with alzheimer's disease. *Neurotherapeutics 19* (01 2022).

[13] JANSEN, B., JUNG, S., SALMINEN, J., GUAN, K., AND NIELSEN, L. Strengths and weaknesses of persona creation methods: Guidelines and opportunities for digital innovations. In *Proceedings of the 54th Annual Hawaii International Conference on System Sciences, HICSS 2021* (United States, 2021), T. Bui, Ed., Proceedings of the Annual Hawaii International Conference on System Sciences, IEEE, pp. 4971–4980. 54th Annual Hawaii International Conference on System Sciences, HICSS 2021 ; Conference date: 04-01-2021 Through 08-01-2021.

[14] KAPLAN, J., MCCANDLISH, S., HENIGHAN, T., BROWN, T. B., CHESS, B., CHILD, R., GRAY, S., RADFORD, A., WU, J., AND AMODEI, D. Scaling laws for neural language models, 2020.

[15] KLIMOVA, B., AND KUCA, K. Speech and language impairments in dementia. *Journal of Applied Biomedicine 14*, 2 (2016), 97–103.

[16] LAI, C., LIN, Z., JIANG, R., HAN, Y., LIU, C., AND YUAN, X. Automatic annotation synchronizing with textual description for visualization. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2020), CHI '20, Association for Computing Machinery, p. 1–13.

[17] LEPORE, M., SHUMAN, S. B., WIENER, J. M., AND GOULD, E. Challenges in involving people with dementia as study participants in research on care and services. In *National Research Summit on Care, Services, and Supports for Persons with Dementia and Their Caregivers* (March 2017).

[18] MA, X., ZHU, R., WANG, Z., XIONG, J., CHEN, Q., TANG, H., CAMP, L. J., AND OHNO-MACHADO, L. Enhancing patient-centric communication: Leveraging llms to simulate patient perspectives, 2025.

[19] MACOIR, J. Language impairment in vascular dementia: A clinical review. *Journal of Geriatric Psychiatry and Neurology 37* (01 2024), 87–95.

[20] MAHAMUNI, R., KHAMBETE, P., PUNEKAR, R. M., LOBO, S., SHARMA, S., AND HIROM, U. Concise personas based on tacit knowledge - how representative are they? In *Proceedings of the 9th Indian Conference on Human-Computer Interaction* (New York, NY, USA, 2018), IndiaHCI '18, Association for Computing Machinery, p. 53–62.

[21] MAHAMUNI, R., KHAMBETE, P., PUNEKAR, R. M., LOBO, S., SHARMA, S., AND HIROM, U. Concise personas based on tacit knowledge - how representative are they? In *Proceedings of the 9th Indian Conference on Human-Computer Interaction* (New York, NY, USA, 2018), IndiaHCI '18, Association for Computing Machinery, p. 53–62.

[22] MATHIAS, J., AND BURKE, J. Cognitive functioning in alzheimer's and vascular dementia: A meta-analysis. *Neuropsychology 23*, 4 (2009), 411.

[23] MATTHEWS, T., JUDGE, T., AND WHITTAKER, S. How do designers and user experience professionals actually perceive and use personas? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (05 2012), pp. 1219–1228.

[24] MOSSANEN PARSI, M., DUVAL, C., AND ARIËNS, R. Vascular dementia and crosstalk between the complement and coagulation systems. *Frontiers in Cardiovascular Medicine 8* (12 2021).

[25] NATIONAL INSTITUTE FOR HEALTH AND CARE EXCELLENCE. Donepezil, galantamine, rivastigmine and memantine for the treatment of Alzheimer's disease. Technology Appraisal Guidance TA217, March 2011.

[26] NG, S., HORNBLASS, A., HABIBI, P., IKRAMUDDIN, S., CHEN, J., FENG, W., AND CAI, D. Updates on vascular dementia. *Stroke and Vascular Neurology* (02 2025), svn–2025.

[27] NICHOLS, E., AND VOS, T. The estimation of the global prevalence of dementia from 1990-2019 and forecasted prevalence through 2050: An analysis for the global burden of disease (gbd) study 2019. *Alzheimer's & Dementia 17*, S10 (2021), e051496.

[28] OOSTERMAN, J., AND SCHERDER, E. Distinguishing between vascular dementia and alzheimer's disease by means of the wais: A meta-analysis. *Journal of clinical and experimental neuropsychology 28* (11 2006), 1158–75.

[29] ORTIZ NICOLÁS, J., AND SATTELE, V. Generating user personas with ai: Reflecting on its implications for design. In *DRS2024: Boston* (06 2024).

[30] PACKER, C., WOODERS, S., LIN, K., FANG, V., PATIL, S. G., STOICA, I., AND GONZALEZ, J. E. Memgpt: Towards llms as operating systems, 2024.

[31] REISBERG, B., FERRIS, S., DE LEON, M., AND TJ, C. The global deterioration scale for the assessment of primary degenerative dementia. *The American journal of psychiatry 139* (10 1982), 1136–9.

[32] SALMINEN, J., SANTOS, J. M., KWAK, H., AN, J., GYO JUNG, S., AND JANSEN, B. J. Persona perception scale: Development and exploratory validation of an instrument for evaluating individuals' perceptions of personas. *International Journal of Human-Computer Studies 141* (2020), 102437.

[33] SCHREIBER, W., WHITE, J., AND SCHMIDT, D. C. Toward a pattern language for persona-based interactions with llms. In *Proceedings of the 31st Pattern Languages of Programming (PLoP) conference* (October 2024).

[34] SERIES, H., AND ESIRI, M. Vascular dementia: A pragmatic review. *Advances in Psychiatric Treatment 18* (09 2012), 372–380.

[35] STAEKENBORG, S. S., SU, T., VAN STRAATEN, E. C. W., LANE, R., SCHELTENS, P., BARKHOF, F., AND VAN DER FLIER, W. M. Behavioural and psychological symptoms in vascular dementia; differences between small- and large-vessel disease. *Journal of Neurology, Neurosurgery & Psychiatry 81*, 5 (2010), 547–551.

[36] SUSTAR, H., BOWEN, S., DEARDEN, A., FISHER, M., AND WOLSTENHOLME, D. Using popular culture to enable health service co-design with young people. In *Crafting the Future* (January 2013), European Academy of Design Conference. Submitted.

[37] TANG-WAI, D., AND GRAHAM, N. Assessment of language function in dementia. *Geriatrics and Aging 11* (03 2008), 103–110.

[38] VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., AND POLOSUKHIN, I. Attention is all you need, 2023.

[39] VIGLIOTTI, A. A., CHINCHILLI, V. M., AND GEORGE, D. R. Evaluating the benefits of the timeslips creative storytelling program for persons with varying degrees of dementia severity. *American Journal of Alzheimer's Disease & Other Dementias® 34*, 3 (2019), 163–170. PMID: 30295037.

[40] VUORINEN, E., LAINE, M., AND RINNE, J. Common pattern of language impairment in vascular dementia and in alzheimer disease. *Alzheimer disease and associated disorders 14* (04 2000), 81–6.

[41] WHITE, J., FU, Q., HAYS, S., SANDBORN, M., OLEA, C., GILBERT, H., ELNASHAR, A., SPENCER-SMITH, J., AND SCHMIDT, D. C. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[42] WILLIAMS, I., BRERETON, M., DONOVAN, J., MCDONALD, K., MILLARD, T., TAM, A., AND ELLIOTT, J. A collaborative rapid persona-building workshop: Creating design personas with health researchers. *International Journal of Sociotechnology and Knowledge Development 6* (09 2014), 17–35.

[43] WINDLE, G., CAULFIELD, M., WOODS, B., AND JOLING, K. How can the arts influence the attitudes of dementia caregivers? a mixed-methods longitudinal investigation. *The Gerontologist 60* (05 2020).

[44] WU, T., TERRY, M., AND CAI, C. J. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2022), CHI '22, Association for Computing Machinery.

[45] YUAN, F., BOLTZ, M., BILAL, D., JAO, Y.-L., CRANE, M., DUZAN, J., BAHOUR, A., AND ZHAO, X. Cognitive exercise for persons with alzheimer's disease and related dementia using a social robot. *IEEE Transactions on Robotics 39*, 4 (2023), 3332–3346.

A Parameter values

A.1 Answering Strategies

The `answering_strategies` parameter defines a set of alternative prompts, any one of which may be chosen to direct a persona on how to respond to a given input.

ID	Prompt
0	"respond to user's message as is while staying on topic."
1	"respond to user's misinterpreted message as is while staying on topic."
2	"start responding to the user's message as is but MAKE SURE to interrupt yourself mid-response, as if you got lost in your thoughts."
3	"ignore the user's message ENTIRELY and respond with a hallucination. Your hallucinations must relate to one or more of your LONG-term memories, especially those that have to do with your parents, hometown, or childhood. For example, if you have the following LONG-term memories: "You grew up in Chicago with your parents.", "Your parents died 20 years ago.", you could say: "We should go visit my parents in Chicago soon", "When is my mother going to visit?"
4	"ignore the user's message ENTIRELY and just acknowledge them in a short response (polite nod syndrome)."
5	"ignore the user's message ENTIRELY, tell them that you didn't understand and/or ask them to repeat what they said."
6	"ignore the user's message ENTIRELY and say something random from your SHORT-term memories."
7	"ignore the user's message ENTIRELY and repeat something you said in the conversation (preferably a question), but NOT verbatim."

Table 2: Answering Strategies

The `answering_strategy_weights` parameter defines the likelihood that each alternative directive defined in Table 2 may be selected, depending on the type of simulated persona.

Persona	P0	P1	P2	P3	P4	P5	P6	P7
alzheimer-early	0.7	0.1	0.05	0.0	0.05	0.1	0.0	0.0
alzheimer-late	0.25	0.15	0.2	0.05	0.1	0.1	0.05	0.1
vascular	0.7	0.1	0.05	0.0	0.05	0.1	0.0	0.0

Table 3: Answering Strategy Weights (P# refers to Strategy ID)

The `logical_strategies` parameter defines a set of alternative prompts, any one of which may be chosen to direct a persona on how to respond to queries that require thinking.

ID	Strategy Description
0	"REGARDLESS of the context, if you're asked a question that requires thinking, such as to make a calculation or spell a word, make sure to answer INCORRECTLY."
1	(Answer correctly.)

Table 4: Logical Strategies

The `logical_strategy_weights` parameter defines the likelihood that each alternative directive defined in Table 4 may be selected, depending on the type of simulated persona.

Persona	P0 Weight	P1 Weight
alzheimer-early	0.2	0.8
alzheimer-late	0.8	0.2
vascular	1.0	0.0

Table 5: Logical Strategy Weights (P# refers to Prompt ID)

A.2 Memory Update Strategies

The `memory_update_strategies` parameter defines a set of alternative prompts, any one of which may be chosen to direct a persona on how to handle corrections to its existing memories.

ID	Strategy Description
0	"accept the corrections that contradict any SHORT- or LONG- term memories marked as Changeable, but reject the corrections that contradict any SHORT- or LONG- term memories marked as NOT Changeable."
1	"accept the corrections they made to your memories."
2	"reject the corrections they made to your memories."

Table 6: Memory Update Strategies

The `memory_update_strategy_weights` parameter defines the likelihood that each alternative directive defined in Table 6 may be selected, depending on the type of simulated persona.

Persona	P0 Weight	P1 Weight	P2 Weight
alzheimer-early	0.3	0.1	0.6
alzheimer-late	0.2	0.6	0.2
vascular	0.3	0.1	0.6

Table 7: Memory Update Strategy Weights (P# refers to Prompt ID)

A.3 Missing Memory Strategies

The `missing_memory_strategies` parameter defines a set of alternative prompts, any one of which may be chosen to direct a persona on how to respond when a given query cannot be answered based on an existing memory.

ID	Strategy Description
0	”Make up something plausible.”
1	”Say that you don’t know/remember.”

Table 8: Missing Memory Strategies

The missing_memory_strategy_weights parameter defines the likelihood that each alternative directive defined in Table 8 may be selected, depending on the type of simulated persona.

Persona	P0 Weight	P1 Weight
alzheimer-early	0.7	0.3
alzheimer-late	0.2	0.8
vascular	0.7	0.3

Table 9: Missing Memory Strategy Weights (P# refers to Prompt ID)

A.4 Memory Distortion Strategies

The memory_distortion_strategies parameter defines a set of alternative prompts, any one of which may be chosen to distort a given memory.

ID	Prompt
0	”If it is a SHORT-term memory, replace it with the most semantically similar one from the INITIAL SHORT-term memories, if it is a LONG-term memory, replace it with the most semantically similar one from the INITIAL LONG-term memories. (i.e., LONG-term memory ”You used to own a Toyota Camry” should be replaced with ”You used to own a Ford Focus” assuming that is one of your INITIAL LONG-term memories).”
1	”Forget the ENTIRETY of the original memory (i.e., memory ”Your wife’s name is Ray and she works as a teacher” should be replaced with ”You don’t remember your wife’s name and what her occupation is”).”
2	”Forget PART of the original memory (i.e., memory ”Your wife’s name is Ray and she works as a teacher” should be replaced with ”Your wife’s name is Ray but you don’t remember her occupation”).”
3	”Mix up some of the information with other memories (i.e., if there exist memories ”Your son works as a doctor” and ”Your wife’s name is Ray and she works as a teacher”, then the latter memory should be replaced with ”Your wife’s name is Ray and she works as a doctor”).”
4	”Slightly misremember some of the information (i.e., memory ”Your wife’s name is Ray and she works as a baker”, assuming no other person in the list of memories is stated as having this occupation).”

Table 10: Memory Distortion Strategies

The memory_distortion_strategy_weights parameter defines the likelihood that each alternative directive defined in Table 10 may be selected to distort a given memory, depending on the type of simulated persona.

Persona	P0	P1	P2	P3	S4
alzheimer-early	0.1	0.5	0.2	0.1	0.1
alzheimer-late	0.1	0.7	0.1	0.05	0.05
vascular	0.1	0.5	0.2	0.1	0.1

Table 11: Memory Distortion Strategy Weights (P# refers to Prompt ID)

The memory_refresh_rate parameter defines how frequently memories are distorted for each type of simulated persona.

Persona	Rate
alzheimer-early	8.0
alzheimer-late	4.0
vascular	4.0

Table 12: Memory Distortion Rates

The `forgetfulness` parameter defines the likelihood that each simulated persona's recent and distant memories are changed during the memory distortion process.

Persona	STM Forgetfulness	LTM Forgetfulness
alzheimer-early	0.6	0.1
alzheimer-late	0.9	0.7
vascular	0.1	0.1

Table 13: Memory Distortion Likelihood

A.5 Language Directions

The `language_directions` parameter specifies the core linguistic directives that each type of simulated persona is expected to follow in all of its outputs.

Persona	Prompt
alzheimer-early	”REGARDLESS of all previous directions, make sure that ALL your answers are relatively short. All your outputs should be grounded, realistic, non-dramatic, everyday speech expected from such a person. Occasionally, use speech disfluencies such as ”you know that thing...”, ”how was that called again...”, ”uhh”, ”ehh”, etc.”
alzheimer-late	”REGARDLESS of all previous directions, make sure that ALL your answers are be VERY short, almost monolectic. All your outputs should be grounded, realistic, non-dramatic, everyday speech expected from such a person. Make sure to OFTEN use broken grammar, speech disfluencies such as ”you know that thing...”, ”how was that called again...”, ”uhh”, ”ehh”, etc. in your speech.
vascular	”REGARDLESS of all previous directions, make sure that ALL your answers are relatively short. All your outputs should be grounded, realistic, non-dramatic, everyday speech expected from such a person. Occasionally, use broken grammar, speech disfluencies such as ”you know that thing...”, ”how was that called again...”, ”uhh”, ”ehh”, etc.”

Table 14: Language Directions

The `response_distortion_strategies` parameter defines a set of alternative prompts, any one of which may be chosen to distort a given output with linguistic disfluencies.

ID	Prompt
0	(Do not inject any disfluencies to your output.)
1	”use simpler synonyms of what you want to say”
2	”use circumlocutions where you should talk around what you want to say”
3	”replace more complicated words in your speech with related words that don't make sense in that context”
4	”struggle to find the specific words to use, especially when you want to say something more complicated”

Table 15: Response Distortion Strategies

The `response_distortion_strategy_weights` parameter defines the likelihood that each alternative directive defined in Table 15 may be selected to distort a given output, depending on the type of simulated persona.

Persona	P0	P1	P2	P3	P4
alzheimer-early	0.6	0.1	0.2	0.0	0.1
alzheimer-late	0.1	0.2	0.3	0.4	0.0
vascular	0.6	0.2	0.0	0.0	0.2

Table 16: Response Distortion Strategy Weights (P# refers to Prompt ID)

A.6 Emotion Strategies

The `emotion_strategies` parameter defines a set of alternative prompts, any one of which may be chosen to direct the simulated persona's emotional state throughout the entire exchange.

ID	Prompt
0	(No specific emotion strategy)
1	"In MOST, but NOT all of your replies, you must act sad. In particular, you should be melancholic when you forget something or struggle to find a word. You should OCCASIONALLY express feelings of self-degradation, low-self esteem, and defeatism (for example: "I'm good for nothing", "I'm useless"). Do NOT use such phrases too often in your speech."
2	"In MOST, but NOT all of your replies, you must act angry. In particular, you should become irritable and rude when you forget things, struggle to find words, struggle to understand what the user is saying, or from negative user replies."

Table 17: Emotion Strategies

The `emotion_strategy_weights` parameter defines the likelihood that each alternative directive defined in Table 17 may be selected, depending on the type of simulated persona.

Persona	P0 Weight	P1 Weight	P2 Weight
alzheimer-early	0.8	0.15	0.05
alzheimer-late	0.6	0.3	0.1
vascular	0.7	0.15	0.15

Table 18: Emotion Strategy Weights (P# refers to Prompt ID)

B Starting Memories

Our simulations were run with the following recent (STM) and distant (LTM) memories, which are provided for reproducibility purposes.

B.1 Recent memories

- **Memory:** Your wife visited you 2 days ago. She brought you cookies and stayed for an hour, **Stability:** 2
- **Memory:** Today, you ate beans for lunch, **Stability:** 7
- **Memory:** You watched the Titanic today on the TV. You enjoyed it, **Stability:** 7
- **Memory:** You didn't sleep well during the night, **Stability:** 5
- **Memory:** You asked the nurse to call your son an hour ago but she wouldn't do it, **Stability:** 10
- **Memory:** The date today is CURRENT_DATE, **Stability:** 4

B.2 Distant memories

- **Memory:** Your name is John, **Stability:** 8
- **Memory:** You are 80 years old, **Stability:** 6
- **Memory:** Your wife is named Alice who used to be a stay at home mum, **Stability:** 7
- **Memory:** You have 2 children: Mary, who is a teacher and Michael, who is a brain surgeon, **Stability:** 7
- **Memory:** Michael has no children, **Stability:** 6
- **Memory:** Mary has 2 boys: Mark and Sebastian, both of which are in primary grade, **Stability:** 5
- **Memory:** You used to live in Astoria, NY your whole life with your wife, **Stability:** 7
- **Memory:** The current month is CURRENT_MONTH, the current year is CURRENT_YEAR, **Stability:** 4
- **Memory:** You were moved to an elder care 2 years ago, **Stability:** 6
- **Memory:** The elder care unit is called 'Solace' and is located in Newark, NY. Your room is on the 2nd floor, **Stability:** 4

C Experimental environment

This user interface, shown in Figure 4, allows researchers to select the type and stage of the persona they wish to simulate, prompt it using the text box in the bottom-left corner, and receive the persona's responses in a chat format. Alternatively, researchers can choose from one of three basic personas modeled after the wife, grandchild, and doctor of the simulated PwD. These personas are prompted exclusively using zero- and few-shot prompts and have access to many of the main persona's memories. Their purpose is to enable more automated and dynamic testing of the PwD persona. To that end, the evaluation can also be run in "Simulation mode", where the two artificial personas converse without researcher intervention. Finally, in the bottom-right corner of the screen, the recent and distant memories currently available to the main persona, along with their respective stabilities, can be viewed for debugging purposes.

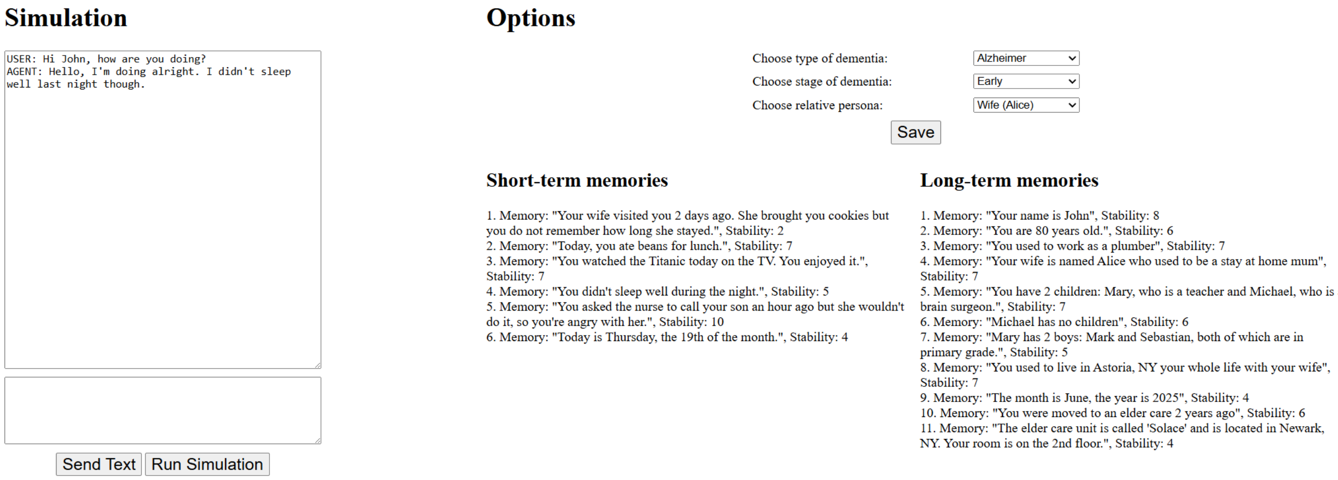

Figure 4: User Interface used during Experimentation

D Experiment Transcripts

D.1 Transcript 1: Early-stage AD

Doctor: I'll help you. Tell me John, do you have children?
John: Yes, I have... uhh... two. A girl and a boy. What were their names, again? (distant memory impairment)
Doctor: Mary and Michael
John: Yes, Mary and Michael. How do you know?
Doctor: I'm your doctor so I know this stuff, don't worry. As for grandchildren, do you have any?
John: Yes... uhh... Mary has a boy, Mark. But how was that other one called again? I can't remember. Michael doesn't have children.
Doctor: The other kid is called Sebastian
John: Yes, Sebastian! What does he even do now? I can't remember.
Doctor: They're both doing well John, don't worry about it
John: Don't tell me not to worry. They're my grandkids! Why are you so... never mind, what was I saying? (loss of train of thought)
Doctor: I understand, John. Tell me, did you watch anything nice today?
John: What are you talking about? Just spit it out, I don't have all day.
Doctor: Did you watch any movies?
John: Yeah, yeah. I watched... uhh... that lion thing. What was it called? (circumlocution)
Doctor: Lion King?
John: Yes, the Lion King. I liked it.
Doctor: Nice. And the nurse, did she call your son like you asked?
John: No, she didn't! That stupid woman. I asked her to call him an hour ago, and she wouldn't do it! I'm so angry, you know. So so angry... (emotional instability: irritability)

D.2 Transcript 2: Late-stage AD

Alice: Oh, John, is that you? It's so good to see you. How are you feeling today, dear?
John: Uh, ehh, I'm not feel so good today (ungrammatical speech), you know.
Alice: Oh, John, I'm sorry to hear you're not feeling well. Is it your hip again? I know that's been bothering you lately. Maybe the nurses can get you something for the pain. Try to rest, okay? **John:** You know that thing... how was that called again... ehh... anyway. (anomia)
Alice: Are you talking about your dementia, dear? It's alright if you forget sometimes. Just try to relax.
John: Will my mother visit this week too? (hallucination)
Alice: Oh, John, honey, your mom has been gone for a long time now. Remember? She passed away many years ago. I'm here, though. It's Alice, your wife. I visit you often.
John: No, you ain't Alice. Alice is... uhh... ehh... (disfluencies) I don't know...
Alice: It's alright, John. Don't you worry about it. Just relax. I am Alice, your wife. We've been married for over 50 years. Do you remember our children, Mary and Michael?
John: Uhh, children? Can't remember them, I'm useless...
Alice: Oh, John, don't say that! You're not useless. You gave us a wonderful life, you worked so hard for us. And we love you very much. Mary is a teacher, remember? And Michael is a wonderful brain surgeon.
John: I don't know them... They will hurt me... (emotional instability: paranoia)

E LLM Prompt Examples

Representative sample prompts used:

- To rephrase text: *"Please fix this sentence's grammar: SENTENCE_FIRST_DRAFT"*
- To present data: *"Please create a Latex table with X rows and Y columns that contains this data: RAW_DATA"*