

# Multi-Agent AI System for Adaptive Cognitive Training in Elderly Care

Isabel Ferri-Molla<sup>a</sup>, Jordi Linares-Pellicer<sup>b</sup>, Carlos Aliaga-Torro<sup>c</sup>  
and Juan Izquierdo-Domenech<sup>d</sup>

Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València (UPV) Camí de Vera,  
s/n, 46022, Valencia, Spain  
{isfermol, jorlipel, calitor, juaizdom}@upv.es

**Keywords:** Multi-Agent System, Artificial Intelligence, Large Language Models, Cognitive Training, Elderly Care.

**Abstract:** The accelerated ageing of the global population presents significant societal and healthcare challenges, particularly concerning cognitive decline in older adults. This paper introduces a multi-agent system designed to stimulate and preserve cognitive abilities in elderly users through personalized exercises tailored to their needs. The proposed system integrates a suite of specialized AI agents: Teacher, Critic, Conciliator, Performance Evaluator, and Psychologist, each fulfilling specific roles to generate, validate, and adapt cognitive exercises collaboratively. The system establishes a self-correcting feedback loop that mitigates errors and reduces hallucinations through multi-agent consensus mechanisms by employing specialized LLM-based agents for generation, critique, evaluation, and emotional assessment. This approach enhances inference depth and ensures the generation of reliable exercises and dynamic feedback. Two interaction modes, voice-based and text-based, are implemented using state-of-the-art speech recognition and synthesis technologies, enhancing accessibility for users with varying preferences and abilities. A user study evaluated the system's usability and effectiveness. Results indicate that the multi-agent architecture enhances cognitive engagement and provides a personalized user experience. The system demonstrated efficacy in addressing diverse cognitive needs, highlighting its potential as an adaptable tool for cognitive training in elderly care.

## 1 INTRODUCTION

Global ageing represents one of society's most significant challenges in the coming decades. As life expectancy increases worldwide, the demographic shift towards an older population introduces significant economic, social, and healthcare implications. This growing population segment requires innovative approaches to address age-related challenges, such as physical decline, cognitive deterioration, and increased dependency while fostering well-being and active participation in society. Among these challenges, preserving independence, enhancing quality of life, and maintaining cognitive functionality emerge as top priorities.

In response to these needs, technology has become crucial in developing solutions to support older adults. Advances in digital tools, assistive technologies, and artificial intelligence (AI) enable new ap-

proaches to care, rehabilitation, and the prevention of age-related decline. Technology not only facilitates better care for dependent individuals but also empowers older adults to lead healthier and more autonomous lives. Solutions such as telehealth platforms, wearable devices, and smart home systems are transforming traditional care models by offering remote monitoring, personalised support, and early detection of health-related issues (Wu et al., 2024), (Ferguson et al., 2021).

A key area of focus in this technological evolution is cognitive health. Cognitive decline is one of the most significant concerns associated with ageing, as it directly impacts memory, decision-making, and overall brain functionality. Cognitive exercises are a proven strategy for maintaining mental sharpness and mitigating the effects of ageing on the brain. However, traditional approaches to cognitive stimulation often need more personalisation and adaptability, limiting their efficacy and user engagement. This is where AI emerges as a transformative tool.

With the capabilities of AI, it is now possible to design and implement cognitive exercises tai-

<sup>a</sup> <https://orcid.org/0009-0008-3608-9891>

<sup>b</sup> <https://orcid.org/0000-0002-3315-1716>

<sup>c</sup> <https://orcid.org/0009-0001-0426-886X>

<sup>d</sup> <https://orcid.org/0000-0003-0076-7001>

lored to the unique needs of individuals. AI systems can assess users' cognitive abilities, adapt exercises in real-time, and provide personalised feedback that enhances engagement and effectiveness (Garcia-Betances et al., 2015). AI systems can also offer a deeper level of user interaction by incorporating emotional intelligence. These systems can also analyse the emotional tone of users, adjusting interactions dynamically to prevent frustration and foster a supportive, positive experience (Picard, 2010).

Despite the promising potential of AI-driven solutions, challenges remain. Large language models (LLMs) and foundational AI models, integral to many of these tools, have limitations. These models are prone to errors, such as generating inaccurate or irrelevant information, a phenomenon commonly referred to as "hallucinations." In the context of healthcare and cognitive support for older adults, such errors could undermine trust, effectiveness, and safety. Addressing these challenges requires robust solutions that enhance the reliability and accuracy of AI systems (Moor et al., 2023).

Agentic systems offer a promising approach to overcoming these limitations. By leveraging a collaborative framework of specialised AI agents, tasks can be distributed across domains such as medicine, psychology, education, and system coordination. Each agent contributes expertise to a specific aspect of the user experience, ensuring that outputs are accurate, coherent, and highly personalised. This distributed architecture not only mitigates the risk of errors but also enhances the overall effectiveness of the intervention. For example, psychological agents can determine the user's emotional states during exercises and adjust the interface accordingly, while multi-agent system components ensure integration and coordination of all functionalities.

## 2 LITERATURE REVIEW

The fast growth of the global elderly population has led to an increased prevalence of dementia and age-related cognitive decline, posing significant societal and healthcare challenges (Gates and Valenzuela, 2010). Cognitive decline often manifests in deterioration of memory, attention, language, and executive functions, undermining independence and quality of life. Consequently, interventions that slow or mitigate cognitive deterioration have gained substantial importance. Traditional cognitive training programs, ranging from paper-and-pencil exercises to computerized brain games, have shown promise in improving or maintaining cognitive functions in older adults (Kelly

et al., 2014) (Kueider et al., 2012). However, the efficacy of these interventions remains a topic of debate, as evidence for their long-term cognitive benefits is mixed and often limited by methodological concerns (Simons et al., 2016). Moreover, these conventional approaches typically offer limited adaptability, personalization, and scalability, reducing their long-term engagement and overall efficacy.

Recently, AI has emerged as a powerful tool to enhance dementia care and cognitive training methodologies. AI-driven analytics have significantly improved early detection and diagnosis of cognitive impairments. Machine learning (ML) and Deep learning (DL) models can integrate socio-demographic data, neuroimaging results, speech patterns, and longitudinal cognitive assessments to identify early markers of dementia, thus enabling timely interventions (Graham et al., 2020) (Li et al., 2015). These predictive models enhance diagnostic accuracy and refine treatment planning, guiding clinicians toward tailored strategies that may delay disease progression and improve patient outcomes.

Beyond detection, AI-driven personalization has revolutionized the delivery of cognitive interventions. Adaptive systems that learn from user performance and preferences over time are gradually replacing traditional one-size-fits-all methods. This adaptability is particularly evident in multi-agent AI systems, where multiple specialized agents collaborate to design, validate, and adapt cognitive exercises (Faraziani and Eken, 2024). Such frameworks employ agents for teaching, criticizing, evaluating, and providing psychological support—each contributing distinct expertise to produce a cohesive, user-centric intervention. The real-time analysis of user behaviour and progress allows dynamic adjustments to difficulty levels, content type, and emotional tone, ensuring that training remains consistently challenging yet accessible.

A critical innovation in this area is the integration of LLMs and advanced speech technologies into these systems. Voice-based interfaces and Natural language understanding (NLU) facilitate more inclusive engagement, especially for individuals with sensory or motor limitations (Aghajani et al., 2021). By simplifying user-system interactions, voice-based platforms can reduce barriers to participation and improve adherence, allowing elderly users to access and benefit from cognitive exercises. As AI advances, these multimodal interfaces, combining voice, text, and visual cues, enhance user autonomy and foster greater long-term participation in cognitive maintenance programs (Faraziani and Eken, 2024).

Multi-agent AI systems also increasingly incorporate emotional and affective computing techniques.

Emotional well-being is tightly linked to cognitive health, and interventions that address affective states may offer more holistic benefits. AI-enabled robots and embodied virtual agents, for instance, integrate cognitive training tasks with simulated social interactions to meet both the cognitive and emotional needs of older adults (Gochoo et al., 2020). These socially assistive robots can detect and respond to changes in the user's mood or frustration levels, providing encouragement, empathy, and motivational support. By attending to the user's emotional state, AI systems can sustain user engagement over extended periods, potentially enhancing intervention adherence and outcomes.

Moreover, AI is expediting research advancements in elderly cognitive care. Automated literature review agents, leveraging Natural language processing (NLP) and ML-based summarization techniques, can efficiently synthesize vast amounts of research findings (Sami et al., 2024). These systems help researchers stay current with the rapidly evolving evidence base and accelerate the development of new, more effective cognitive intervention strategies. Meta-analyses and systematic reviews generated with AI support ensure practitioners and policymakers can access the latest, high-quality evidence, guiding better-informed decision-making in clinical and community settings.

Despite these technological strides, current solutions often target isolated aspects of cognitive support. For example, some systems focus on exercise generation without adequately addressing emotional or motivational support, while others emphasize detection and monitoring without delivering actionable training content (Castro et al., 2023). Integrating multiple AI capabilities into a single cohesive framework remains a key challenge, including personalisation, emotional adaptation, and multimodal interaction. Additionally, best practices for orchestrating agent collaboration and mitigating conflicting outputs are still evolving. There is a clear need for comprehensive multi-agent architectures that can seamlessly synchronize these functionalities, thereby improving user experience, accessibility, and training efficacy.

This paper addresses these gaps by presenting a novel multi-agent system for adaptive cognitive exercise design in elderly care. Building on the advances discussed in the literature, our system incorporates specialized agents, such as the Teacher, the Critic, the Conciliator, the Performance evaluator, and the Psychologist, within the CrewAI (crewAI, 2024) framework to deliver holistic, user-tailored interventions. The system enables sophisticated agent collaboration by leveraging LLM-based interfaces, speech recog-

niton, and emotional adaptation strategies while remaining accessible and engaging. This integrated approach is poised to advance the state of the art in AI-driven cognitive training, offering a more comprehensive, responsive, and human-centered solution for an increasingly urgent public health challenge.

### 3 SYSTEM STRUCTURE

The proposed multi-agent system is designed to assist older adults in stimulating their cognitive abilities through exercises that focus on, but are not limited to, six distinct areas: attention, memory, mental calculation, language, logical reasoning, and general practice.

Our approach prioritizes multi-agent orchestration and LLMs over explicit symbolic or hybrid methods. This enables greater adaptability and scalability, reducing reliance on predefined rules or rigid knowledge representations.

The system leverages a multi-layered team of dedicated agents, each tasked with specific roles such as exercise generation, critique, evaluation, and emotional assessment, to continuously refine both content and user interactions. This approach simulates a "System 2"-like reasoning process, enabling the generation of high-quality exercises, real-time evaluation of user performance, and emotional monitoring to adapt future exercises to the user's needs.

Crew AI, a Python-based open-source framework designed to orchestrate multi-agent AI systems, has been employed to orchestrate the multi-agent system. It facilitates the creation of teams of autonomous agents, each with specific roles, tools, and objectives, collaborating efficiently to perform complex tasks.

Each agent assumes a specific role, fostering communication and coordination between agents to share information and delegate tasks efficiently. Its modular design facilitates the integration of customised tools and LLMs, enabling adaptation to diverse applications and environments, including the dynamic generation of exercises tailored to the user's profile. Furthermore, it automates workflows by managing task allocation and agent interactions, optimising processes and enhancing operational efficiency.

Using LLMs to generate cognitive exercises offers multiple benefits supported by recent research (Kasneci et al., 2023). These models enable personalisation tailored to each user's profile and cognitive level, enhancing learning effectiveness. Additionally, their generative capabilities allow the creation of a wide variety of exercises, avoiding repetition and maintaining the user's engagement (Ichien et al., 2024). LLMs can

also evaluate responses and provide immediate feedback, fostering autonomous learning and real-time error correction (Letteri and Vittorini, 2024). Moreover, they can simulate human cognitive processes (Aher et al., 2023) and design exercises that stimulate specific functions such as memory, attention, and planning. Lastly, LLMs can generate high-quality educational content, offering detailed explanations and solutions for complex exercises, facilitating understanding challenging concepts, and enriching educational material (Sarsa et al., 2022; Moore et al., 2023).

However, despite their increasing precision, LLMs may exhibit hallucinations or provide unreliable information that aligns differently from the specific context sought. Hallucinations refer to generating content without factual grounding in the training data. This phenomenon occurs when the model, without sufficient information, attempts to probabilistically complete patterns without a genuine understanding of the content, resulting in fabricated or incorrect responses.

The lack of adaptation to specific contexts arises from the limitations of some models in adequately interpreting the environment or the user's particular needs. This constraint may lead to irrelevant, incoherent, or poorly adjusted responses that do not meet the requirements of the given scenario.

In the realm of creating and evaluating cognitive exercises, it is essential to implement mechanisms that control the outputs of language models to mitigate inaccuracies before they reach the end user. Such inaccuracies can compromise the quality of the exercises generated, the interpretation of user performance data, and, consequently, the system's utility in adapting future exercises or enhancing the accuracy and feedback provided.

In this context, multi-agent systems emerge as a highly recommended solution (Shinn et al., 2024). These systems enable the development of architectures involving agents with specific roles working in a coordinated manner. Furthermore, it is possible to establish groups of agents sharing the same role, interacting with one another to reach a consensus before presenting the information to the user. For instance, in generating a cognitive exercise, multiple agents powered by language models can collaborate, refine the content, and validate its quality, ensuring that the final output is optimised to meet the user's needs.

This multi-agent system approach offers two primary benefits. Firstly, it enhances the robustness of the generated content by reducing errors and ensuring accuracy through a consensus-driven process. Secondly, it provides significant flexibility to adapt to each user's specific cognitive and individual needs,

delivering a personalised and adequate experience.

At the start of execution, the system gathers basic information about the user, such as their name, age, cognitive state, and specific needs. This information forms the user's specific context and is shared with the various agents to be considered when performing their respective tasks.

Based on the structure provided by the CrewAI technology, an agent in our system will possess a role, a goal, a backstory, and a language model to generate appropriate responses.

The agent's role defines its function within the group, while its goal represents the individual objective that guides its decision-making process. The backstory provides contextual information about the agent, shaping its interactions and behaviour.

Each agent will execute tasks, which are specific assignments the agent is responsible for completing. These tasks will be defined with a detailed description, the agent assigned to execute them, the expected output upon task completion, and the relevant context.

The crews will also be established once the agents and tasks have been defined. A crew is a group of agents collaborating to solve a set of tasks. Each crew will have its own parameters, such as the agents involved in the collaboration, the tasks to be addressed, and the type of process to be followed.

The decision-making processes of the presented multi-agent system are structured around sequential task execution, leveraging consensus-driven mechanisms and precise coordination protocols. Each task is broken into smaller, manageable subtasks, which are assigned to agents sequentially to ensure an efficient flow of operations. At every step, agents interact iteratively to validate and refine outputs through consensus mechanisms. For example, the critic agent evaluates the outputs from the teacher agent against established benchmarks. If conflicts arise, the conciliator agent mediates to synthesize improvements, ensuring robust and coherent results. Tasks are executed in a predefined sequence, where the output of one agent becomes the input for the next. This structure minimizes redundancy and ensures a linear progression, enhancing both speed and accuracy. For instance, after the teacher agent generates an exercise, it is passed to the critic agent for review and then to the performance evaluator agent for assessment.

### 3.1 System Agents

The system comprises various types of agents, each with distinct functionalities and roles, which are enumerated as follows:

### 3.1.1 Teacher Agent

The Teacher agent (TA) is responsible for generating exercises tailored to the user. These exercises are designed through prompting to be short, dynamic, and specific, focusing on stimulating key cognitive skills such as attention, memory, mental arithmetic, language, and logical reasoning. Each exercise will be adaptable to the user's profile and needs, prioritising an appropriate level of difficulty to ensure completion within one minute. Tasks will vary, including identifying distinct elements in a sequence, performing simple calculations, completing logical patterns, ordering words or lists, and quick memory exercises.

### 3.1.2 Performance Evaluator Agent

The primary goal of the Performance evaluator agent (PEA) is to assess the user's responses to the exercises objectively. This agent compares the user's response with the correct solution generated internally to determine whether the response is correct. Additionally, it provides specific and concise feedback based on the user's context, the response given, and the characteristics of the exercise.

### 3.1.3 Critic Agent

The Critic agent (CA) evaluates the responses generated by other agents to ensure they are appropriate and fulfill the requested task. Its main objective is to guarantee the quality and relevance of responses within the system by employing an analytical and critical approach. In this case, it validates the exercises created by the TA to ensure the final output is coherent and adapted to the specific user context. Furthermore, it reviews the corrections made by the agent to verify that the feedback is accurate and suitably tailored to the user's specific characteristics.

### 3.1.4 Conciliator Agent

The Conciliator agent (CoA) is responsible for creating an improved version of a cognitive exercise, using the original one and the suggested improvements as a foundation. This agent mediates between the initial exercise design and the critiques received, ensuring adjustments are made to enhance quality, effectiveness, and suitability. Its role is crucial in preventing infinite loops between the CA and TA. During the creation of exercises, the debate between these agents may prolong the process; therefore, after several interactions, the CoA is introduced to provide a resolution based on the exercises created by the TA and the critiques made by the CA.

### 3.1.5 Psychologist Agent

The Psychologist agent (PsA) analyses the user's emotional state based on their exercise performance and the information provided. This agent acts as an empathetic observer, identifying emotions such as frustration, interest, satisfaction, or others that may arise during the user's interaction with the system. Feedback on these emotions is stored in a document.

## 4 METHODOLOGY

The system begins by creating an initial context for the user based on key information such as their name, age, and specific cognitive characteristics. Using this context, the TA is tasked with designing an exercise to train and enhance the user's cognitive abilities.

The integration of AI in cognitive training for elderly users raises critical ethical concerns. Privacy is a key issue, as the sensitive nature of personal and health-related data makes it susceptible to misuse if not properly safeguarded. To address this, the system prioritizes data privacy and security by collecting only essential user information, such as name, age, and cognitive characteristics. Importantly, the name is not transmitted externally and is used only within the system to personalize the experience. Age and cognitive information are anonymized before being sent to the model, ensuring data protection.

Once the exercise has been generated, an interaction is initiated between two agents: the TA and the CA. The CA receives the designed exercise and the user's context, conducting a detailed evaluation to determine whether the exercise is coherent and adheres to the established criteria. At this stage, the CA may either approve or reject the exercise. If the exercise is rejected, the TA must create a new one, considering the observations and suggestions provided by the CA.

If the exercise is approved following the initial interaction between the TA and CA, the system will present it to the user for completion. Conversely, if the exercise is rejected again after the TA generates a revised exercise and undergoes a new critique by the CA, the CoA is introduced into the process. This agent analyses both the revised exercise created by the TA and the observations made by the CA, ultimately producing a final proposal that harmonises both perspectives while still considering the user's specific characteristics.

Through this iterative cycle, the three agents collaborate to achieve consensus, ensuring the exercise is high quality, coherent, and tailored to the user's needs. After completing this process, the system will deliver

the final exercise to the user, allowing them to proceed and complete it.

Once the user provides their response to the exercise, the PEA takes action. As primary inputs, this agent receives the exercise prompt, the user's response, and the previously defined user context. Based on this information, the PEA determines whether the user's response is correct or incorrect. Additionally, it tailors its explanation, the justification of its evaluation, and its manner of expression to align with the user's specific needs, ensuring that the feedback is clear, understandable, and relevant to their profile.

Once the PEA delivers its verdict, it will share it with the CA. The CA is provided with all the pertinent information: the exercise prompt, the user's context, the user's response, and the evaluation generated by the PEA. Using this data, the CA conducts a thorough analysis to determine whether the evaluation provided by the PEA is appropriate, well-founded, and in line with the previously established criteria.

Following the consensus reached between both agents, a final evaluation of the exercise is generated and presented to the user. This evaluation not only aims to inform the user about the correctness of their response but also to provide a clear, well-justified, and personalised explanation to help them better understand the exercise and reflect on their performance.

Finally, the system interacts with the user once again to gather their emotional feedback regarding the exercise. At this stage, the user is asked how they felt while completing the exercise, both in terms of its composition and its level of difficulty. This emotional feedback, combined with the user's context and their response to the exercise, is used to generate a detailed analysis of their emotional state based on their performance, which can subsequently be used to adapt future exercises, ensuring they are better aligned with the user's cognitive abilities and emotional well-being. The figure 1 presents a diagram illustrating the system architecture and internal interactions.

## 5 EXPERIMENTAL APPROACH

In terms of user interaction with the system, two versions have been developed: one where the interaction is conducted through text and another where it is conducted through voice.

In the voice interaction version, advanced speech recognition and synthesis technologies enable a natural and accessible voice-based interaction. The Deepgram API powers both speech recognition and syn-

thesis functionalities (Deepgram, 2024), ensuring a robust and efficient implementation. These tools allow the user's voice to be transformed into text and, in turn, generate responses in auditory format, creating a continuous flow of communication between the user and the system. This approach enables interaction for individuals who may have reading or writing difficulties, which is common among older users.

Speech recognition is crucial in capturing the user's spoken words and converting them into text. This process leverages DL models to achieve high accuracy even under challenging acoustic conditions, such as noisy environments or interacting with users with varied accents. The ability to process speech in real-time ensures that interactions remain immediate and dynamic. Moreover, these models transcribe individual words and interpret complete phrases, taking into account the context and intent behind the message.

On the other hand, speech synthesis transforms the text generated by the system into clear and natural audio. This not only facilitates the comprehension of instructions and feedback but also enhances the user experience by incorporating human-like intonations and rhythms.

The interaction flow begins by informing the user that the exercise is being generated. The exercise is then read aloud, and the system awaits the user's input, which is spoken directly into the system. The user's voice is converted into text and processed by intelligent agents to generate a response. This response is subsequently transformed into audio through speech synthesis, providing clear and personalised communication to the user. This process enables the user to engage in an interactive experience while enhancing the system's usability and accessibility, allowing individuals with varying abilities to interact effectively with the technology.

Moreover, both voice and text interaction models have utilised language models to enhance the agents' ability to comprehend and generate appropriate responses. For this purpose, GPT-4o was selected (Achiam et al., 2023), as it is one of the state-of-the-art LLMs. To ensure that the performance of the LLM did not compromise the system's effectiveness, we consulted an LLM leaderboard, specifically the Chatbot Arena LLM Leaderboard (Chiang et al., 2024). This leaderboard is based on an open platform that evaluates and ranks AI models through user votes, using the Bradley-Terry model to generate real-time rankings. Consequently, we selected the top-ranked LLM to evaluate the multi-agent system architecture properly.

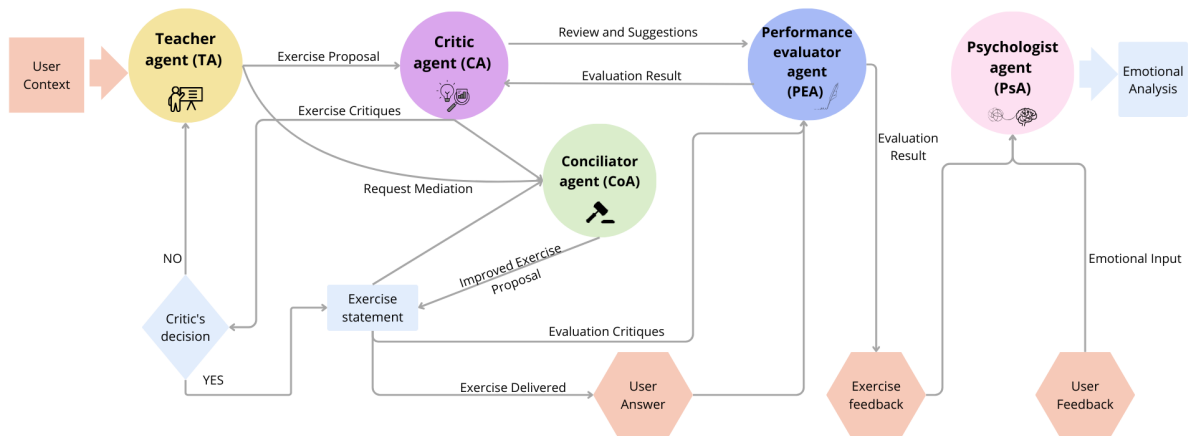


Figure 1: Overview of the multi-agent system architecture.

## 6 SYSTEM EVALUATION

The practical experimentation of the system was conducted with a group of 20 users aged between 54 and 83 years, most of whom were members of a senior university program. These users had a certain level of familiarity with technology, although not specifically with systems like the one being tested. During the tests, users were tasked with completing an entire interaction cycle designed to evaluate the system's performance.

The group of 20 users was divided into two subgroups of 10 participants each. One subgroup performed the tests using voice interaction, while the other used text-based interaction. Efforts were made to ensure that the average age of participants in both groups was similar. The objective of this approach was not only to test the multi-agent system but also to assess how the different forms of interaction could influence users' perceptions of the system.

The experimental process followed a similar structure for both groups.

Firstly, the system generated a personalised exercise tailored to each user's specific characteristics, which had been previously introduced. This level of personalisation ensured that the generated tasks were relevant and appropriately suited to the user's profile.

Once the exercise was presented, participants completed the task using either text or voice interaction, depending on their assigned modality. Upon task completion, the system automatically generated a correction of the user's responses and provided detailed feedback on their performance. Subsequently, users were asked to share their thoughts on the exercise. The system produced a psychological profile based on the exercise outcomes, user feedback, and observed interaction patterns. This profile included

an analysis of the emotional state experienced during the process, identifying emotions such as satisfaction, frustration, or motivation.

Following the experimentation phase, both groups were given a questionnaire, with questions assessed using a 5-point Likert scale, where one indicated "strongly disagree" and five indicated "strongly agree."

The results for the subgroup that tested the system using text-based interaction are presented in Table 1, showing the mean and standard deviation for each question. Table 2 presents the results for the subgroup that used the system through voice-based interaction, measuring the participants' opinions and responses.

As the analysis presented in the preceding tables shows, the average score for text-based interaction is 4.63 out of 5, indicating a very positive overall opinion. In contrast, this value decreases to 4.10 out of 5 for voice-based interaction. While this still reflects a good predisposition towards using the system, it is evident that users who interacted with the system via text provided higher ratings.

Nevertheless, regarding whether users would utilise the system frequently, the average response in both cases exceeds 4 out of 5 on the Likert scale. This demonstrates a strong general acceptance of the system, regardless of the mode of interaction employed. It suggests that, although text-based interaction is rated slightly higher than voice-based interaction, both modes reveal a favourable user inclination.

In particular, the fact that the average responses for anticipated frequency of use remain above 4 in both cases highlights a positive perception and confidence in the system as a valuable and viable tool.

These findings imply that the system possesses significant potential for frequent and sustained adop-

Table 1: Descriptive results of the Likert scale questionnaire for the text interaction system.

Item	Mean	Standard Deviation
I find the system useful	4.83	0.41
I would use the system frequently	4.50	0.83
The system instructions were clear and easy to understand	4.67	0.52
I felt that the activities matched my abilities	4.50	0.83
I felt secure while using this system	4.00	0.63
I believe that with some practice, I could use this system on my own	4.83	0.41
I found it entertaining and pleasant to use this system	4.67	0.52
I believe the system can help keep my mind active	4.83	0.41

Table 2: Descriptive results of the Likert scale questionnaire for the voice interaction system.

Item	Mean	Standard Deviation
I find the system useful	4.67	0.52
I would use the system frequently	4.17	0.41
The system instructions were clear and easy to understand	4.00	0.89
I felt that the activities matched my abilities	3.83	0.75
I felt secure while using this system	3.83	0.75
I believe that with some practice, I could use this system on my own	3.83	0.75
I found it entertaining and pleasant to use this system	4.00	0.63
I believe the system can help keep my mind active	4.33	0.52

tion despite the differences in user experience between the interaction modes. However, further refinements should be made to enhance the user experience, particularly for voice-based interaction.

Furthermore, following user feedback, participants were asked open-ended questions regarding their overall experience and suggestions for system improvement. Among users of the text-based interaction, the majority highlighted that the initial difficulty of the exercises was somewhat high. This suggests the need for further adjustments to ensure a smoother learning curve, especially for older users who may require a more gradual progression in task complexity. Additionally, some participants mentioned that the correction feedback provided by the system was occasionally too lengthy, which could hinder their ability to process the information effectively.

Several notable challenges were identified for users who interacted with the system via voice. The exercises relied solely on auditory input, which created specific difficulties for users. In memory-based tasks, such as recalling words, participants found it challenging to retain information presented only through voice, as they lacked visual references to support their memory. While the voice interface was designed to be fully auditory, participants expressed that having to rely exclusively on spoken instructions made it more difficult to process and remember multiple steps or complex exercise requirements. A recurring suggestion was to complement the voice-based interface with a visual display of the text on the

screen, as this dual-modality approach would provide users with both auditory and visual cues, improving comprehension and accessibility.

To evaluate whether significant differences exist in the perception of interaction between two modes (text and voice), a Student's t-test was applied to compare the satisfaction means. The results showed that interaction via text received a significantly higher rating than voice. The analysis yielded a t-statistic of 3.19 and a p-value of 0.014, which allows us to affirm, with a 95% confidence level, that the observed difference is unlikely to have occurred by chance. These findings suggest that participants perceive interaction through text as a more positive and efficient experience.

Additionally, the coefficient of variation (CV) was calculated to analyse the responses' consistency and assess the perceptions' homogeneity. The results revealed that voice interaction exhibits more significant variability in user opinions than text mode. For instance, in the question "The system instructions were clear and easy to understand", the CV for text mode was 11.13%, whereas for voice it reached 22.25%. This difference suggests that users' experience with voice interaction is less consistent, possibly due to factors such as difficulties in auditory comprehension, the absence of visual support, or individual differences in the use of technology.



## 7 SYSTEM LIMITATIONS

During the testing conducted throughout the experimental process, the system, leveraging LLMs for content creation and language comprehension along with agentic systems to fulfill various roles, showed promise according to most users. Multi-agent interactions were especially beneficial in regulating the outputs of LLMs, thereby reducing inaccuracies and hallucinations. Despite these advantages, our evaluation highlighted several limitations and proposed areas for improvement that could enhance system functionality and user experience. It was observed that the initial difficulty of the cognitive exercises could benefit from more precise adjustments. Implementing a progressive approach, starting with simpler exercises and gradually increasing complexity based on user performance and feedback, might be more effective. Such a strategy would provide a more user-friendly introduction to the system, especially for users with limited cognitive abilities, and support a more natural and adaptive learning curve. This would help mitigate frustration associated with complex initial exercises, facilitating a smoother transition to more challenging levels. Although initially designed to operate through verbal commands for integration with home assistants, there is a significant opportunity to enhance user interaction by adding a visual interface. This interface would display exercise text, allowing users to both listen to instructions and read them, thus improving comprehension and performance. This feature would be particularly beneficial for users who require visual reinforcement to process information more effectively.

The current evaluation involved a limited number of participants due to resource constraints. Despite these limitations, the insights gleaned provide a valuable foundation for understanding the system's initial usability and effectiveness. Moving forward, it would be beneficial to undertake a more comprehensive evaluation involving a broader user base to further validate and refine the system's capabilities, thus enhancing the robustness of our findings.

Additionally, considering ethical and normative aspects, it would be intriguing to develop an ethical agent in future iterations of this project. This agent will be designed to process regulatory documents, ensuring that the outcomes of interactions within the system, specifically, the final exercise generated by the TA after discussions with the CA and CoA, as well as the responses from the PEA and PsA, are in strict adherence to the applicable ethical standards and regulations. This approach not only reinforces the integrity of the system but also safeguards the interests

and rights of all participants involved.

The system is designed with a strong emphasis on user autonomy and trust. Users are fully informed about its purpose and capabilities and are explicitly asked for their consent before engaging with it, promoting transparency and confidence in the system. Furthermore, it is intended as a supportive tool for cognitive training rather than a replacement for professional medical judgment. Healthcare professionals retain ultimate responsibility for interpreting the system's output and making treatment decisions, recommending its use when it aligns with the specific needs of their patients.

These improvements and considerations are intended to refine the system's capabilities and ensure a high-quality user experience in future deployments.

## 8 CONCLUSIONS

The system presented in this article demonstrates how technology can assist and provide solutions with the potential to mitigate some of the problems that our society will inevitably face in the coming years, such as the ageing of the population and the cognitive decline experienced by individuals at advanced ages.

As shown in this article, the combination of language models with multi-agent systems allows us to harness the creative capabilities of these systems while mitigating issues such as hallucination through the interaction between agents.

The presented system demonstrates how integrating specialized agents with clearly defined roles, such as the TA, the CA, the CoA, the PEA, and the PsA, plays a fundamental role in creating precise, user-oriented cognitive exercises. This multi-agent, LLM-driven strategy combines generation, critique, evaluation, and emotional insights to form a self-correcting, layered process. By advancing toward more System 2-like reasoning, the system reduces hallucinations and ensures a structured, adaptive approach where user needs and emotional states take center stage. The additional computational effort is justified by consistently producing reliable, well-considered outcomes tailored to enhance cognitive engagement and effectiveness.

On the one hand, the TA serves as the primary guide and is responsible for generating the exercises. At the same time, the CA acts as a constructive evaluator of the exercise created, identifying areas for improvement and providing detailed feedback, which ultimately ensures higher quality in the proposed exercise.

The role of the CoA is crucial in integrating the

responses of both the TA and the CA, ensuring that the content remains comprehensible and appropriately tailored to the user's current level. In parallel, the PEA provides real-time correction of the user's responses, monitoring their performance and highlighting potential errors.

Finally, the PsA adds a unique value to the system by conducting an emotional analysis of the user during the completion of the exercises. This functionality allows for detecting the user's emotional state in response to the exercises, which can be highly useful for adapting both the difficulty level and the type of exercises presented.

This holistic approach, which integrates cognitive performance with emotional well-being, ensures a more inclusive, effective, and personalised user experience.

The interaction modes, voice and text, are thoughtfully designed to provide flexibility and adaptability, allowing users to engage with the system in the way that suits them best. Voice-based interfaces offer a hands-free, intuitive option, making them particularly helpful for individuals with limited literacy or motor challenges, such as older adults or those with physical impairments. Meanwhile, text-based interactions ensure clarity and precision, appealing to users who feel comfortable reading and typing, and offering a straightforward and efficient experience. By accommodating diverse abilities and preferences, these dual modes reflect a strong commitment to accessibility and inclusivity, creating opportunities for a wide range of users to interact with the system effectively.

Both systems received positive user evaluations, with average scores exceeding 4 out of 5 on the Likert scale questions. These results reflect a high level of acceptance of the system and a positive inclination towards its use, regardless of the interaction modality chosen.

In conclusion, this system represents a significant step forward in leveraging AI-driven multi-agent frameworks to deliver personalised, adaptive, and accessible cognitive training solutions, potentially enhancing cognitive engagement and emotional well-being in elderly users.

## ACKNOWLEDGEMENTS

This work is partially supported by Generalitat Valenciana, FPI grant CIACIF/2022/098 and CI-PROM/2021/077

## REFERENCES

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Aghajani, M., Ben Abdesslem, H., and Frasson, C. (2021). Voice emotion recognition in real time applications. In *Intelligent Tutoring Systems: 17th International Conference, ITS 2021, Virtual Event, June 7–11, 2021, Proceedings 17*, pages 490–496. Springer.
- Aher, G. V., Arriaga, R. I., and Kalai, A. T. (2023). Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR.
- Castro, C. B., Costa, L., Dias, C. B., Chen, J., Hillebrandt, H., Gardener, S. L., Brown, B. M., Loo, R., Garg, M., Rainey-Smith, S. R., et al. (2023). Multi-domain interventions for dementia prevention—a systematic review. *The Journal of nutrition, health and aging*, 27(12):1271–1280.
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., et al. (2024). Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.
- crewAI (2024). Crewai: Framework for orchestrating role-playing, autonomous ai agents.
- Deepgram (2024). Deepgram speech recognition platform.
- Faraziani, F. and Eken, Ö. (2024). Enhancing cognitive abilities and delaying cognitive decline in the elderly through exercise-based health management systems. *International Journal of Sport Studies for Health*, 7(2):13–22.
- Ferguson, C., Hickman, L. D., Turkmani, S., Breen, P., Gargiulo, G., and Inglis, S. C. (2021). “wearables only work on patients that wear them”: Barriers and facilitators to the adoption of wearable cardiac monitoring technologies. *Cardiovascular Digital Health Journal*, 2(2):137–147.
- Garcia-Betances, R. I., Jiménez-Mixco, V., Arredondo, M. T., and Cabrera-Umpiérrez, M. F. (2015). Using virtual reality for cognitive training of the elderly. *American Journal of Alzheimer's Disease & Other Dementias*, 30(1):49–54.
- Gates, N. and Valenzuela, M. (2010). Cognitive exercise and its role in cognitive function in older adults. *Current psychiatry reports*, 12:20–27.
- Gochoo, M., Vogan, A. A., Khalid, S., and Alnajjar, F. (2020). Ai and robotics-based cognitive training for elderly: A systematic review. In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)*, pages 129–134. IEEE.
- Graham, S. A., Lee, E. E., Jeste, D. V., Van Patten, R., Twamley, E. W., Nebeker, C., Yamada, Y., Kim, H.-C., and Depp, C. A. (2020). Artificial intelligence approaches to predicting and detecting cognitive decline in older adults: A conceptual review. *Psychiatry research*, 284:112732.

- Ichien, N., Bhatia, S., Ivanova, A., Webb, T., Griffiths, T., and Binz, M. (2024). Higher cognition in large language models. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.
- Kasneci, E., Seßler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., et al. (2023). Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.
- Kelly, M. E., Loughrey, D., Lawlor, B. A., Robertson, I. H., Walsh, C., and Brennan, S. (2014). The impact of cognitive training and mental stimulation on cognitive and everyday functioning of healthy older adults: a systematic review and meta-analysis. *Ageing research reviews*, 15:28–43.
- Kueider, A. M., Parisi, J. M., Gross, A. L., and Rebok, G. W. (2012). Computerized cognitive training with older adults: a systematic review. *PloS one*, 7(7):e40588.
- Letteri, I. and Vittorini, P. (2024). Exploring the impact of llm-generated feedback: Evaluation from professors and students in data science courses. In *International Conference in Methodologies and intelligent Systems for Technology Enhanced Learning*, pages 11–20. Springer.
- Li, F., Tran, L., Thung, K.-H., Ji, S., Shen, D., and Li, J. (2015). A robust deep model for improved classification of ad/mci patients. *IEEE journal of biomedical and health informatics*, 19(5):1610–1616.
- Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., and Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265.
- Moore, S., Tong, R., Singh, A., Liu, Z., Hu, X., Lu, Y., Liang, J., Cao, C., Khosravi, H., Denny, P., et al. (2023). Empowering education with llms-the next-gen interface and content generation. In *International Conference on Artificial Intelligence in Education*, pages 32–37. Springer.
- Picard, R. W. (2010). Affective computing: from laughter to iee. *IEEE transactions on affective computing*, 1(1):11–17.
- Sami, A. M., Rasheed, Z., Kemell, K.-K., Waseem, M., Kilamo, T., Saari, M., Duc, A. N., Systä, K., and Abrahamsson, P. (2024). System for systematic literature review using multiple ai agents: Concept and an empirical evaluation. *arXiv preprint arXiv:2403.08399*.
- Sarsa, S., Denny, P., Hellas, A., and Leinonen, J. (2022). Automatic generation of programming exercises and code explanations using large language models. In *Proceedings of the 2022 ACM Conference on International Computing Education Research-Volume 1*, pages 27–43.
- Shinn, N., Cassano, F., Gopinath, A., Narasimhan, K., and Yao, S. (2024). Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.
- Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., and Stine-Morrow, E. A. (2016). Do “brain-training” programs work? *Psychological science in the public interest*, 17(3):103–186.
- Wu, X., Freeman, S., Miyagi, M., Park, U., Nomura, K., and Ebihara, S. (2024). Comprehensive geriatric assessment in the era of telemedicine. *Geriatrics & Gerontology International*, 24:67–73.