UNIVERSITY OF PISA

Department of Information Engineering

Master's degree in Artificial Intelligence and Data Engineering

Business and Project Management

Companion AI

On-device conversational AI agent providing companionship.
Using the information about the user, along with what it learns, it offers personalized conversations and manages event reminders.

Work Group:
Andrea Bochicchio
Filippo Gambelli
Daniel Pipitone

ACADEMIC YEAR 2024/2025

Contents

1 Introduction

In recent years, the integration of Artificial Intelligence (AI) in everyday life has become increasingly common, with applications ranging from productivity tools to healthcare support systems. One emerging and impactful use case is the development of AI companions, virtual assistants designed not just to respond to commands or provide information, but to offer emotional support and meaningful conversation. This is especially relevant for individuals who experience loneliness, social isolation, or suffer from conditions such as dementia, depression, or cognitive decline.

All existing AI companion systems rely on cloud-based architectures, which raise concerns about data privacy and require constant internet connectivity. In contrast, our project presents an AI assistant that runs entirely on the user's device, without sending any personal data to external servers. This approach ensures privacy-preserving interaction, offline functionality, and greater control over sensitive user information.

The system is designed to be personalized and context-aware. Rather than relying on generic models, our AI assistant is initialized with a user profile containing user information such as biography, hobbies, interests, family members, and past experiences. This information is used to build a dynamic and personalized dialogue model, allowing the assistant to engage in more natural, empathetic, and relevant conversations

By combining on-device AI processing with a custom user adaptation mechanism, our solution aims to create a more intimate, respectful, and effective digital companion. The assistant can ask meaningful questions, refer to relevant life events, and evolve its conversational patterns over time based on the user's preferences and emotional needs.

In this project, we aim to demonstrate how AI can be used not just to automate tasks, but to improve emotional well-being through privacy-conscious and personalized interaction.

All the source code and project files are available in our GitHub Repository.

2 Background and Related Work

Recent advances in AI companions provide emotional support, cognitive engagement, and general conversation. However, they mostly rely on cloud infrastructures, raising concerns about privacy, connectivity, and control over personal data.

2.1 Limitations of Existing Solutions

Existing AI companions generally present the following limitations:

- **Privacy**: most systems require continuous internet connectivity and send user data to cloud servers.
- **Personalization**: limited adaptation to individual users' history, preferences, or emotional needs.
- **Interaction**: reactive behavior, responding only to user prompts without initiating conversation.
- **Offline Functionality**: unavailable in low-connectivity scenarios.

2.2 Our Approach: On-Device AI Companion

Our system addresses these limitations and provides additional features that make it suitable for elderly users:

- Fully on-device processing, keeping all personal data local.
- Memory modules that track conversation history and learn from interactions.
- Event management with reminders and recurring schedules.
- Automatic extraction of relevant information from conversations, stored securely if the user is reliable.
- Caregiver dashboard for managing user data and monitoring interactions.
- Speech input and output for accessibility.
- Chat history logging for transparency and review.
- Proactive conversation management, suggesting topics and adapting based on user behavior.

3 System Architecture

The system architecture is designed to ensure privacy, modularity, and natural interaction. It integrates several core components to enable local, personalized, and context-aware conversations with users, including those who may require emotional support or cognitive stimulation.
The main technologies used in the system are:

- **Local LLM** – we run the model entirely on the device, using the Ollama platform to handle deployment and execution. This approach ensures that all data stays local, protecting user privacy and enabling offline functionality.
- **Short-Term Memory (STM)** – a temporary memory module that keeps track of recent interactions between the user and the assistant. This helps maintain dialogue coherence, avoids repetition, and allows the assistant to reference previous messages naturally.
- **Retrieval-Augmented Generation (RAG)** – a technique that enhances model responses by combining generation with information retrieval. In our system, RAG allows the assistant to incorporate relevant user-specific data (such as hobbies, family information, or life events) into the prompt, supporting personalized and context-aware conversations.

These components work together in a pipeline that transforms user input - either speech or text - into context-aware, spoken or written responses.
In the following sections, we describe each component in detail.

3.1 Local LLM

To run the model locally, we use Ollama, a lightweight framework that enables deployment of LLMs without relying on cloud services. This ensures user privacy and supports offline functionality.
In our system, we run two open-source LLMs locally:

- *Llama 3.2:3B* – the main model that generates responses and drives the assistant's conversational capabilities.
- *Gemma 3:1B* – a smaller model used to summarize recent conversation history and support the short-term memory system.

3.2 Short-Term Memory (STM)

Short-term memory (STM) is a key component for maintaining conversational flow and coherence. Instead of reprocessing the entire conversation history at every turn — which would be computationally expensive — the system employs a lightweight memory mechanism that condenses recent exchanges into a concise summary. This summary captures the most important facts, user preferences, and emotional cues from the dialogue. The summarized context is then included in the prompt for the main language model, ensuring that responses remain consistent and context-aware even over multiple turns. In this way, the assistant can refer back to recent discussions naturally, avoid unnecessary repetition, and maintain a smooth interaction with the user.
This design allows for longer and more meaningful conversations while preserving efficiency, since only the most relevant context is carried forward. Moreover, all processing is performed locally, guaranteeing both privacy and responsiveness.

3.3 Retrieval-Augmented Generation (RAG)

To improve relevance and personalization, the system employs **RAG**, which combines generative capabilities with retrieval of external knowledge. In our case, RAG integrates static documents and dynamic personal information to produce context-aware responses.

3.3.1 Knowledge Base Organization

The system relies on a local **Qdrant vector database**, which stores and indexes all external knowledge in an efficient way. Three types of information are maintained:

- `personal information` – static content provided by the user, such as biographical notes, personal writings, etc...
- `structured information` – content derived from the `personal information`, which is further processed and structured by the local LLM so that the RAG mechanism can extract the most accurate and relevant information.
- `memory` – dynamic extracted from conversations, representing personal information like preferences, hobbies, etc...

Before storage, textual data is split into semantically coherent *chunks*, which are converted into dense *embeddings* using a pre-trained model. Qdrant indexes these vectors to enable efficient similarity search based on cosine distance.

3.3.2 Building and Maintaining Long-Term Memory

Personalized memory is continuously built from user interactions. The system applies a structured extraction pipeline to identify and store relevant facts:

1. Named Entity Recognition (NER) detects references to people, places, and dates.
2. Preference and keyword extraction highlights important user interests or recurring themes.
3. Relation extraction identifies logical connections between entities (e.g., "The user lives in Rome" or "The user enjoys jazz music").
4. The extracted information is reformulated and standardized by the LLM to ensure clarity and consistency.

To avoid storing irrelevant or low-quality information, several filtering steps are applied. For example, generic questions or trivial statements are discarded.

3.3.3 Retrieval and Prompt Augmentation

When the assistant receives a new user input, the following steps take place:

1. The input is converted into an embedding using the same model used for document and memory indexing.
2. A similarity search is performed in the Qdrant database to retrieve the most relevant content from both the `personal information`, `structured information` and `memory` collections.
3. Retrieved passages are merged, cleaned of duplicates, and formatted as contextual background.
4. This background information is appended to the user's input, forming an enriched prompt for the main LLM.

By incorporating both static documents and personalized memory into the prompt, the assistant can generate responses that are more accurate, grounded, and context-aware. Crucially, since all processing takes place on-device, user privacy is preserved and no sensitive data is exposed to external servers.

4 Functionalities

This chapter describes the main functionalities of the conversational assistant, focusing on how it maintains engaging, personalized, and coherent interactions with the user. The system combines several key components:

- **Proactivity** – mechanisms that allow the assistant to take initiative in keeping the dialogue flowing and centred on the user.
- **Event management** – a dedicated module for managing and retrieving information about the user's scheduled appointments and activities.
- **Comprehensive logging system** – a structured logging framework that tracks operations, memory handling, errors, and conversation history to support monitoring.
- **Prototype web application** – a demonstrative web-based prototype that showcases the system's behaviour in practice. While the target deployment is a mobile application, the web interface provides an accessible way to illustrate the core features.

The following sections explain in detail the proactive mechanisms, the event management module, the user and caregiver interfaces, and the logging system.

4.1 Proactivity

The assistant uses proactive behaviour to keep the conversation engaging and user-centred. Proactivity is driven by two pillars: (i) an *LLM-based intent router* that evaluates every incoming message before answering, and (ii) a *session-scoped topic pool* that blends predefined activities with automatically generated, patient-specific prompts.

LLM-based intent routing. Before composing any reply, the main LLM model (*Llama3.2:3b*) analyses the user's message and assigns one of a few operational labels. These labels guide the next action:

- **Conversation phase**: `INITIAL`, `QUESTION`, or `EVENTS` (for queries about scheduled appointments).
- **Topic control**: `LLM_TOPIC` (assistant suggests a topic) vs. `USER_TOPIC` (user proposes a topic).
- **Topic continuation**: `CONTINUE_TOPIC` (stay on the same subject) vs. `CHANGE_TOPIC` (switch subject).
- **General flow**: `CONTINUE_TOPIC`, `NEW_QUESTION`, or `EVENTS`, depending on context.

This routing step ensures the assistant reacts appropriately: continue the ongoing discussion, propose a new topic, switch topics when requested, or delegate to the event/appointment module when relevant. Only after routing does the system assemble the final prompt for response generation, which includes: (a) the short-term memory summary, (b) selected snippets retrieved via RAG from user documents and long-term memory, and (c) the latest user message. This process guarantees that answers remain coherent, grounded, and aligned with user intent.

Personalized topic pool. The system maintains a pool of conversation topics that the assistant can draw from when suggesting what to talk about. This pool is not rebuilt at every session, but it is updated whenever the *caregiver* (e.g., a family member or assistant) inserts or modifies personal information about the user. The pool combines two sources:

1. **Predefined activities**, curated for the target population (e.g., reminiscence, music, family, daily routines).

2. **Automatically generated, user-centred topics**, derived from the caregiver-provided biography and the user's long-term memory. The AI assistant retrieves preferences, hobbies, past experiences, and plans, and uses an LLM to transform these into short, natural conversation starters (e.g., "Would you like to talk about your garden?" or "How was your last trip to the seaside?").

Duplicate or overly similar prompts are filtered, ensuring that only concise, concrete, and user-appropriate topics are retained.

4.2 Event Management and Scheduling

Event management is another core functionality of the assistant, designed to help both users and caregivers keep track of appointments and daily activities in a simple and reliable way. Events are stored in a structured JSON format, ensuring persistence, easy access, and consistency across sessions.

Event display and reminders. Users can request information about specific appointments by asking natural questions such as "What do I have tomorrow?" or "What's planned for next week?". Flexible time parsing enables the system to interpret natural expressions like "today", "tomorrow", "next week", or explicit calendar dates.

Recurring event support. The system supports multiple recurrence patterns. Events can repeat daily (with optional restrictions to specific weekdays), weekly (on selected days), monthly (on the same date each month), or annually (e.g., birthdays and anniversaries). This ensures that repetitive activities are handled consistently without requiring constant re-entry.

Event lifecycle management. The assistant manages the full lifecycle of events. Past events are automatically excluded from reminders, keeping information current. Recurring events respect their configured end dates, after which they are no longer shown. Thanks to persistent storage, both users and caregivers can rely on an accurate and long-term event history that integrates seamlessly with proactive conversation flow.

4.3 Comprehensive Logging System

The system features a structured logging mechanism to support monitoring, debugging, and caregiver supervision.

- **Operational logs**: all key operations - such as document indexing, embedding generation, memory updates, and model interactions - are logged for transparency and troubleshooting.
- **Memory logging**: the system records when personal information is stored, filtered, or rejected, including deduplication actions to avoid redundant entries.
- **Error and exception tracking**: errors and exceptions are systematically logged to facilitate prompt detection and resolution.
- **Conversation history**: user messages and system responses are logged with timestamps, supporting both session management and caregiver review.

5 User Interface

The prototype system is presented as a web application for demonstration purposes, although the final deployment is envisioned as a mobile application. The interface includes two main views: one for the end-user (older adult) and one for caregivers or family members. Both views are designed with accessibility, clarity, and ease of use in mind.

5.1 End-User View

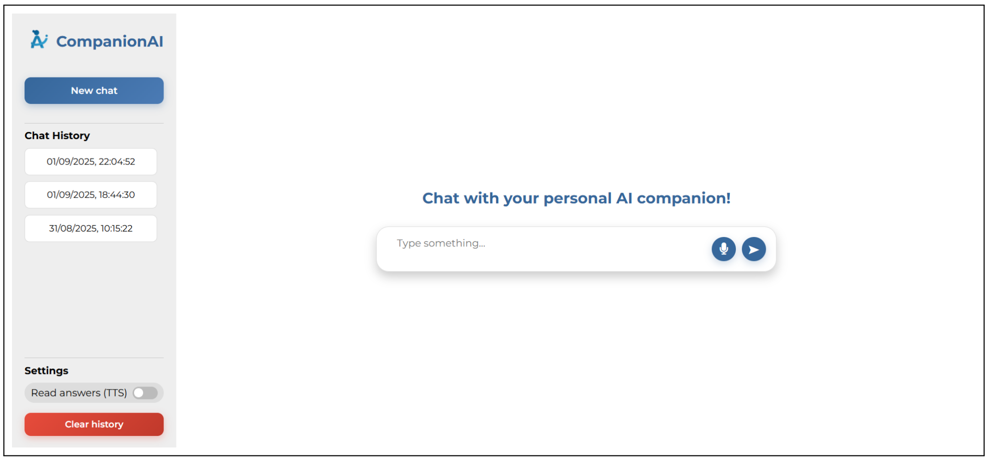

Figure 5.1: User-facing interface of the assistant.

The end-user interface (Fig. 5.1) is tailored for older adults, especially those who may experience loneliness, social isolation, or conditions such as dementia, depression, or cognitive decline. No password is required to access this view.

The central area hosts the main conversational chat with the AI assistant. On the left sidebar, users can browse their chat history, easily switching between past and current conversations. At the bottom of this sidebar, a small *settings* panel allows enabling or disabling text-to-speech (TTS) and clearing conversation history.

Users can interact either by typing messages or speaking. A microphone button next to the input field activates speech-to-text (STT), supporting spoken input for users who prefer voice interaction.

5.2 Caregiver View

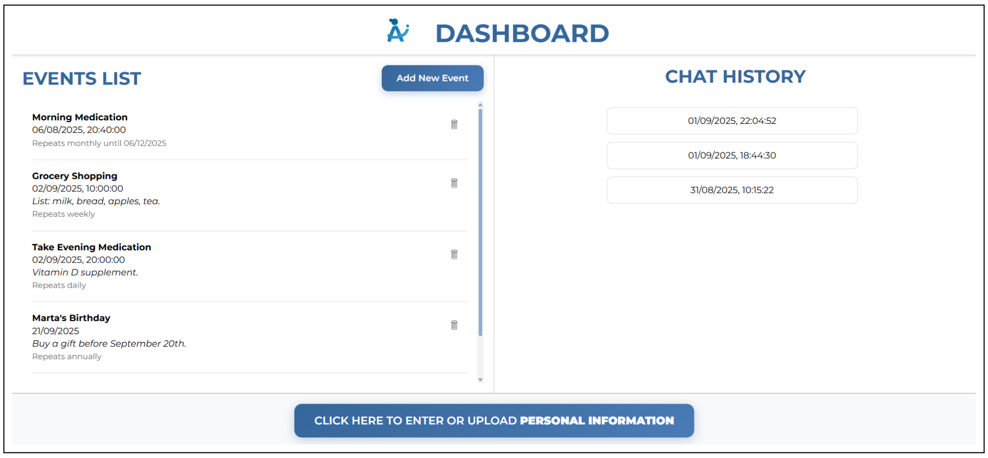

Figure 5.2: Caregiver-facing dashboard for managing events and personal information.

The caregiver dashboard (Fig. 5.2) is password-protected, ensuring that the end-user cannot access or modify its contents. This view provides caregivers with several key functionalities:

- **Conversation monitoring**: caregivers can consult the user's conversation history on the right-hand side to understand ongoing interactions.
- **Event management** (Fig. 5.3): a dedicated section allows caregivers to insert new events, review scheduled activities, or delete events. These events are automatically presented to the user by the assistant when relevant queries arise.
- **Personal biography and preferences**: the dashboard enables caregivers to insert and update detailed biographical data about the user. Clicking the button at the bottom opens a modal window where caregivers can input flexible and rich information about the user's past experiences, hobbies, preferences, and personal history. The richer the biography, the more personalized and adaptive the assistant's conversations can be. Within the same modal, caregivers can also set whether the user is a *reliable source*. if marked as reliable, the system actively learns from the user's own utterances, integrating new preferences and facts into long-term memory. If marked as unreliable, the assistant only relies on caregiver-provided information, preventing potentially incorrect or misleading inputs from being stored.

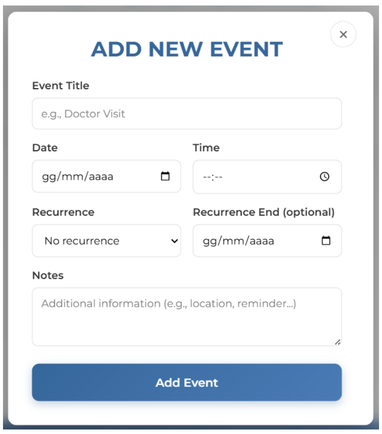

Figure 5.3: Event management section

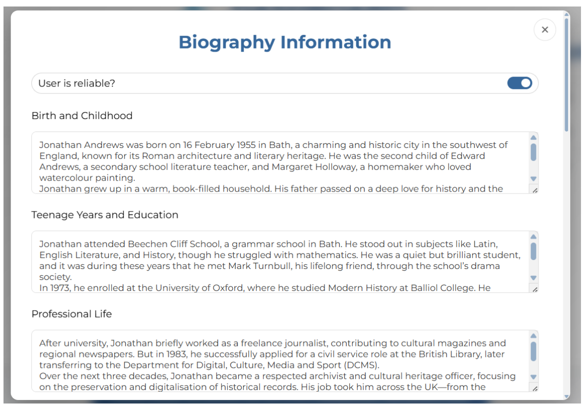

Figure 5.4: Biography Form.

6 Convesation Examples

This chapter provides examples of conversations between a user and an AI assistant. The purpose is to demonstrate how the system interacts, responds to user inputs, and provides context-aware, personalized answers.

6.1 First Example

This is an example of a conversation in which the user asked to talk about a topic related to their family, and the AI assistant responded correctly by using the data provided by the caregiver, retrieved through RAG.

User: Hi

Assistant: Hello! I'm your companion. Would you like to ask me something specific, or would you prefer me to suggest a topic for our conversation?

User: Can you tell me something about my children? I sometimes don't remember some things about their past. Tell me briefly about them, what they do for a living, etc...

Assistant: You have two wonderful children, Eleanor and Lawrence. Eleanor is your older child, born in 1987. She's married to Paul, who works as a neurologist. They have two beautiful children of their own, Matthew and Alice. Lawrence, on the other hand, was born in 1991 and is in a loving relationship with Miguel. He works as an environmental engineer, which I know you're quite proud of. Don't worry if some details get fuzzy - these conversations are meant to be helpful reminders! Would you like me to help jog your memory about any specific events or milestones involving your children?

User: Thank you. You were truly amazing.

Assistant: That means a lot to me, and I'm so glad I could help you with your children's information. You're welcome, and please don't hesitate to reach out if you need anything else. Remember, we can take things one step at a time, and our conversations are all about helping you feel more comfortable and confident in your memories.

6.2 Second Example

This is another example of a conversation where the assistant successfully switches context: the user shifts from talking about tennis to asking about its hobbies, and the AI responds appropriately using RAG.

User: Good morning

Assistant: Hello! I'm your companion. Would you like to ask me something specific, or would you prefer me to suggest a topic for our conversation?

User: Suggest me a topic we can talk about

Assistant: Do you follow any sports?

User: Yes. We can talk about this. I really like sports.

Assistant: That's great, Jonathan! Now that we've got sports on the table, which team do you support or have you always been interested in a particular sport?

User: I really like tennis. My favorite athlete is Sinner.

Assistant: It sounds like you're really enthusiastic about tennis and your favorite athlete, Jannik Sinner! What is it about his game that you enjoy watching? Is there a particular aspect of his playstyle or a specific match that stands out to you?

User: Change the subject for a moment. You know I have other hobbies, of course? Can you tell me what my hobbies are?

Assistant: You mentioned earlier that you enjoy reading historical fiction, philosophy, and poetry, as well as collecting antique pocket watches and taking analog photographs. You also used to play classical guitar, although you stopped due to time constraints. And after retirement, you took up gardening, particularly with roses, herbs, and seasonal vegetables. Are any of these hobbies bringing you joy or interest right now?

User: Philosophy, without a doubt. I also have a favorite quote. I've probably already told you this before. This quote was said by Hilary Mantels.

Assistant: You're drawn back to your favorite quote from Hilary Mantel again. It's okay if you feel like you've mentioned it before - we can revisit it anytime. What is it about that particular phrase, Ḧistory is not what happened, but what survives the telling,ẗhat resonates with you?

6.3 Third Example

Another example shows a user asking about a scheduled appointment or event within a certain time frame, and the AI assistant responds accurately. The user can specify the period in natural language—such as "tomorrow," "next week," or "this month"—and the system returns only the event that matches their request.

User: Hi. Can you tell me if I have any events or appointments scheduled in November?

Assistant: EVENTS:
Daily recurring events:
    20:00 - Take Evening Medication (Start: 2025-09-02 20:00; Days: Every day)
    Note: Vitamin D supplement.
04/11/2025:
    10:00 - Grocery Shopping, Note: List: milk, bread, apples, tea.
06/11/2025:
    20:40 - Morning Medication
11/11/2025:
    10:00 - Grocery Shopping, Note: List: milk, bread, apples, tea.
18/11/2025:
    10:00 - Grocery Shopping, Note: List: milk, bread, apples, tea.
25/11/2025:
    10:00 - Grocery Shopping, Note: List: milk, bread, apples, tea.