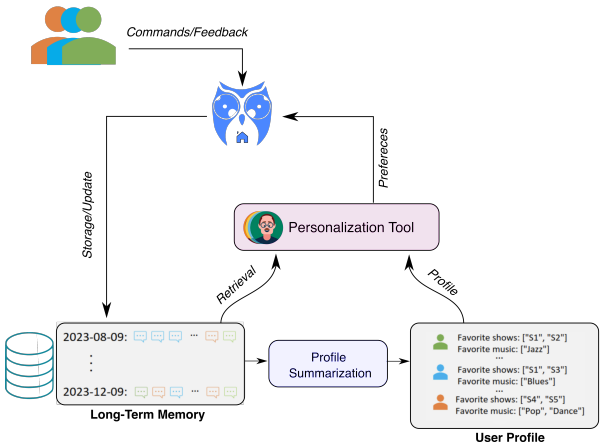

Figure 3: An overview of the personalization module.

integrates several other tools for external grounding, including a TV schedule search tool, a weather tool, and an email and calendar tool. We do not discuss these in detail because tools of this type have been well explored in the literature, and are readily available in libraries such as LangChain.

5.1 Personalization

This section introduces a personalization tool that enables the system to adapt and respond in a manner that is consistent with its user's preferences. The personalization tool, illustrated in Figure 3, is composed of two main sub-components: (i) the long-term memory that stores users' interactions, and (ii) the user profiler that constructs a hierarchical understanding of users' preferences.

5.1.1 Long-Term Memory

Long-term memory [1], shown in Figure 3, stores information about the user's past interactions and behavior. The memory records commands or feedback in chronological order and is used for retrieval by the agent to make decisions that are better aligned with the individuality of its user. Similar to information retrieval techniques for LLM augmentation summarized in [19], the long-term memory is used to retrieve memories relevant to the user query in order to augment the in-context information available to the personalization tool. Each entry in the memory is encoded using a dense retrieval embedding model (e.g., [20]). The vector representations are then indexed and stored in a vector database. In this work, we use the MiniLM embedding model [21].

5.1.2 User Profiler

The user profile, shown in Figure 3, provides a high-level summary of the interactions between the users and the agent to build a dynamic and holistic understanding of the users' preferences. We adopt a hierarchical approach, first proposed in [22], to build the users' profiles. The user profiler starts by splitting all the entries in the long-term memory by day and generates daily summaries to capture all the nuances of the users' preferences. Next, all the daily summaries are aggregated into a single global overview serving as the user profile. Our choice of a hierarchical approach is motivated by two main reasons: (i) scalability: as the long-term memory grows with time, a hierarchical approach is highly scalable because it is amenable to MapReduce-style processing [23], (ii) information loss: directly generating a concise summary from the long-term memory involves a long-context prompt. LLMs' ability to successfully retrieve and identify relevant information within the input context is known to degrade as the length of the input context increases [24].

5.1.3 Personalization tool

The personalization tool (depicted in Figure 3) combines the long-term memory and the user profiler to allow the SAGE agent to query them with questions about the user's preferences. When queried with a question about the user's preferences, this tool first encodes the question using the dense retrieval embedding model and similar memories are retrieved from the long-term memory, using cosine similarity in the embedding space as the distance metric. Next, the tool constructs an LLM prompt consisting of the retrieved memories, the user profile, and the question. Finally, it queries the LLM with the prompt and returns the response. The user profile and the