Can Large Language Models Be Good Companions? An LLM-Based Eyewear System with Conversational Common Ground

Zhenyu Xu[1], Hailin Xu[1], Zhouyang Lu[1], Yingying Zhao[2], Rui Zhu[3], Yujiang Wang[4], Mingzhi Dong[1], Yuhu Chang[1], Qin Lv[5], Robert P. Dick[6], Fan Yang[7], Tun Lu[1], Ning Gu[1], and Li Shang[1]

[1]School of Computer Science, Fudan University, Shanghai, China, 200438
[2]Department of Computer and Information Sciences, University of Strathclyde, Glasgow, United Kingdom, G1 1XH
[3]Bayes Business School, City, University of London, London, United Kingdom, EC1Y 8TZ
[4]Oxford Suzhou Centre for Advanced Research, Suzhou, China
[5]Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, United States, 80309
[6]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan, United States, 48109
[7]School of Microelectronics, Fudan University, Shanghai, China, 201203

June 4, 2024

Abstract

Developing chatbots as personal companions has long been a goal of artificial intelligence researchers. Recent advances in Large Language Models (LLMs) have delivered a practical solution for endowing chatbots with anthropomorphic language capabilities. However, it takes more than LLMs to enable chatbots that can act as companions. Humans use their understanding of individual personalities to drive conversations. Chatbots also require this capability to enable human-like companionship. They should act based on personalized, real-time, and time-evolving knowledge of their users. We define such essential knowledge as the *common ground* between chatbots and their users, and we propose to build a common-ground-aware dialogue system from an LLM-based module, named *OS-1*,

to enable chatbot companionship. Hosted by eyewear, OS-1 can sense the visual and audio signals the user receives and extract real-time contextual semantics. Those semantics are categorized and recorded to formulate historical contexts from which the user's profile is distilled and evolves over time, i.e., OS-1 gradually learns about its user. OS-1 combines knowledge from real-time semantics, historical contexts, and user-specific profiles to produce a common-ground-aware prompt input into the LLM module. The LLM's output is converted to audio, spoken to the wearer when appropriate. We conduct laboratory and in-field studies to assess OS-1's ability to build common ground between the chatbot and its user. The technical feasibility and capabilities of the system are also evaluated. Our results show that by utilizing personal context, OS-1 progressively develops a better understanding of its users. This enhances user satisfaction and potentially leads to various personal service scenarios, such as emotional support and assistance.

Keywords— Smart eyewear, large language model, common ground, context-aware

1 Introduction

It has long been a vision for chatbots to be personal, human-like companions [84, 15, 93]. One classic example is Samantha, an AI dialogue system in the movie "Her" [1], who interacts with the protagonist through a camera, a microphone, and an earbud, learning his personality, preferences, and habits over time. Samantha offers companionship, emotional support, and assistance, and eventually becomes a nearly indispensable part of the protagonist's life.

To realize this vision, several technical challenges must be addressed. The limited linguistic and cognitive capabilities of natural language processing (NLP) have been recognized as major barriers to personalized dialogues [1]. Recent advances in large language models (LLMs) such as ChatGPT (based on the GPT-3.5 LLM model) [57] and GPT-4 [58] have largely removed this barrier and opened the possibility of supporting natural and human-like conversations. Pre-trained on massive amounts of text data, LLMs have the ability to encode a vast amount of world knowledge. These capabilities allow LLMs to generate coherent and diverse responses; this is crucial for natural conversation. Additionally, through supervised instruction fine-tuning and reinforcement learning with human feedback [59], LLMs can be adapted to follow natural language instructions. Inspired by the powerful language modeling capabilities of LLMs, the question arises: *"Can LLM-based chatbots serve as personal companions in daily life?"*

We argue that the answer today remains, "Not without further capabilities". Despite impressive human-like language capabilities, LLMs lack ***common ground*, preventing LLM-based chatbots from being personal companions**. Based on research in linguistics [22], psychology [35], and Human-Computer Interaction (HCI) [25], having common ground is essential for successful and meaningful conversations. This common ground can stem from shared personal experiences, interests, and other factors. For example, when initiating a dialogue with others, we either ask questions to establish common ground or presuppose certain common ground already

[1] `https://en.wikipedia.org/wiki/Her_(film)`

exists [73]. **It is challenging for an LLM to establish a mutual understanding with a person.** The common ground between humans is usually implicit and subjective [98, 24]. Therefore, it is not practical to expect users to provide common ground information explicitly. Also, LLMs are generally not equipped to perceive a user's context [63], e.g., their physical surroundings or daily experiences. Without such personal context, LLMs struggle to comprehend a user's visual surroundings, speech, daily events, and behavior (e.g., personality traits [5] or habits). This prevents them from establishing common ground with their users.

This work is inspired by the powerful language generation capabilities of LLMs [57, 58, 18, 9] and motivated by their lack of personal context awareness necessary to establish common ground. To bridge these gaps, we formulate the following research questions (RS).

RS1. Does personal context help LLM-based dialog systems establish common ground with their users?

RS2. In what ways do different types of personal context contribute to personalized LLM-based dialog system responses?

We argue that the answer today is, "Ubiquitous personal context helps establish common ground between LLM-based dialogue systems and their users, and furthermore, it enables better personalized responses". To test the hypotheses, this work breaks *personal context* into the following three categories in the temporal dimension, and describes the design of an LLM-based smart eyewear system to achieve ubiquitous personal context capturing and use.

- **Real-time context** refers to momentary semantics inferred from the user's ongoing speech and visual surroundings. These semantics help LLMs understand the meanings of the user's speech and visual perceptions, enabling the generation of appropriate responses.
- **Historical context** is a summary of the past real-time context time series. It organizes the user's daily events (e.g., activities) and dialogue contents by clustering the real-time contexts into temporal units. This information helps LLMs maintain the coherence and continuity of the dialogue, and enables it to avoid repeating or contradicting previous statements.
- **User profiles** are distilled historical information related to the user's social background, personality, preferences, and habits, which are revealed during interaction with the dialogue system. They can enable LLMs to incorporate additional human-like qualities by adapting to the user's personality and long-term goals, resulting in more consistent and anthropomorphic responses.

To answer the two research questions above, this work presents OS-1[2], the first LLM-based chatbot system aware of the conversational common ground with its users. OS-1, with its unique capabilities, can see what its user sees, hear what its user hears, and feel what its user feels. Residing on smart glasses, OS-1 captures the visual and audio signals received by its user as input, builds conversational common ground gradually, and generates personalized dialogues at proper times. As portrayed in Figure 1, OS-1 consists of four central modules.

1. **Real-time context capture.** The smart eyewear first perceives the user's in-situ visual and audio signals through the built-in camera and microphone, and

[2] "OS-1" was the name for Samantha's underlying implementation in the movie "Her".

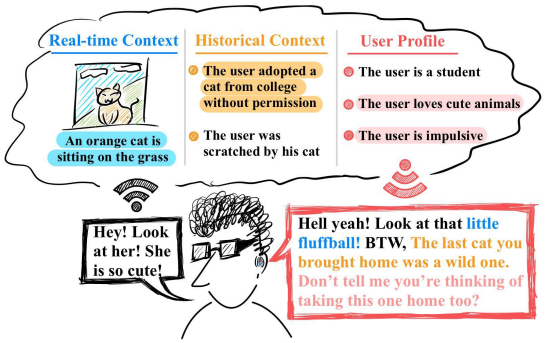


Figure 1: Conceptual overview of OS-1 workflow.

transfers them to the cloud on the fly. These two types of information are essential to understanding the user's ongoing status. Using a vision-language model (e.g., LLaVA [48]) and a speech recognition model (e.g., Whisper [66]) deployed in the cloud, we can infer the semantic description of images and transcribe voice data into text. Also, by combining these two data modalities, OS-1 can infer the user's current activity, location, and other information inferred from the user's surroundings. These forms of information constitute the real-time context.

2. **Historical context management**. To ensure long-term coherence and consistency in dialogues, it is important to remember the user's historical information. To this end, we design a clustering method that extracts the relevant information, such as daily events, from the accumulated real-time contexts, thus forming the historical context. The clustering method removes redundancy among inter-real-time contexts and produces event-level descriptions that are easy to summarize. We designed indexing methods along temporal, spatial, and semantic dimensions to facilitate efficient retrieval of historical contexts.

3. **User profile management**. To better understand users' profiles, we analyze their historical context to form a user profile that includes their social background, personality, preferences, and habits. However, our inference of the user profile may be biased or contain errors due to limited interactions. To address this problem, we design an update scheme that revises the user profile based on updated historical context and past user profiles.

4. **Personalized response generation**. When the user launches a conversation with OS-1, a real-time context is extracted from the eyewear's video and audio streams, and OS-1 retrieves the up-to-date knowledge database of the user, namely the historical context and user profile, using the latest real-time context as the key. This retrieval process employs a multi-agent approach [95, 87], encompassing a dialogue policy agent and an information retrieval agent. The former directs the dialogue's structure and flow, while the latter evaluates ongoing discourse to determine relevant content for retrieval and initiates the retrieval process accordingly. The resulting personal context, which contains personal information sufficient enough to drive common-ground-aware conversations, is used as the LLM prompt to generate an appropriate response. This textual response is converted into audio delivered via the glasses, ending this conversation

cycle and waiting for the next one.

We conduct in-lab experiments and in-field pilot studies to evaluate OS-1's ability to establish common ground using the captured and refined personal contexts. This ability would enable OS-1 to facilitate better conversations that satisfy its users. Inspired by the idea of using human-like features as measurements of conversational response quality [29], this work proposes customized human evaluation metrics for evaluating OS-1. More specifically, we propose a Grounding score to evaluate how well OS-1 can build up common ground with its users. We also propose Relevance, Personalization, and Engagement scores to evaluate the relevance of the system's responses to the real-time context, the relationship between the responses and the user's historical and profile contexts, as well as the level of interest users show in the responses. Laboratory results show that, compared to the baseline method without any personal contexts, OS-1 improves the Grounding score by 42.26%. Also, OS-1 substantially improves the performance by 8.63%, 40.00%, and 29.81% in Relevance, Personalization, and Engagement score, respectively. The in-field pilot study further shows that the Grounding score trends upward over time, indicating that OS-1 is capable of continuously reaching common ground with users through long-term interactions. We also explain its behavior in emotional support and personal assistance applications and conduct semi-structured interviews to provide qualitative insights.

Our work makes the following contributions.

1. We present a novel concept of personal context and a human evaluation metric Grounding score to assess the ability of an LLM-based dialogue system to reach mutual understanding, providing a measure of suitability for personal companionship applications.
2. We design and implement an always-available smart eyewear LLM-based personal dialogue system that captures the user's multi-modal surroundings on the fly, generates personal context, and engages in personalized conversation with the user. One of the greatest strengths of the system is its ability to achieve context awareness without increasing cognitive load or imposing interaction requirements on users, thereby enhancing the user experience under various HCI scenarios.
3. We propose a novel method to capture, accumulate, and refine the personal context from user multi-modal contexts and dialogue histories, and a multi-dimensional indexing and retrieval mechanism that integrates multiple personal contexts to enable personalized responses. Our method facilitates dynamic adaptation to the user's surroundings, experiences, and traits, enabling engaging and customized conversations.
4. We conducted an in-lab study and a pilot study to evaluate the impact of using personal context within the dialogue system. Our results show the superior performance of the proposed system in gradually reaching a better understanding of the user over time.

2 Related Work

In this section, we provide a brief overview of related work in the following areas: (1) large language models, (2) multimodal dialogue systems, (3) personalized dialogue systems, and (4) wearable dialogue systems.

2.1 Large Language Models

Large language models are recent innovations that revolutionized the field of natural language processing (NLP) and influenced other areas. LLMs are pre-trained on large-scale corpora. Models such as GPT-3.5 [57], GPT-4 [58], Vicuna [20], Llama 2 [77], Qwen [7] and Falcon [6], have demonstrated impressive language understanding and modeling capabilities, manifesting improved performance across downstream tasks [95]. In addition to enhanced language intelligence, LLMs also have exhibited unpredictable and sharp performance improvements in certain specific tasks with the increase in scale (e.g., training compute, training dataset size, etc.), a phenomenon called emergent capabilities [82]. One such capability is In-Context Learning (ICL) [32], in which the LLMs need only be exposed to a few examples for their learning to be transferred to a new task/domain. Additionally, through supervised instruction fine-tuning and reinforcement learning with human feedback (RLHF) [59], LLMs can follow natural language instructions. This feature has enabled LLMs to contribute to a variety of tasks [12] such as text summarization [62] and sentiment analysis [81].

The Chain-of-Thought (CoT) method [83], on the other hand, has shown that LLMs can be guided to conduct complex reasonings by prompting to generate intermediate steps. Similarly, for the complex reasoning task [21], works on X-of-Thought (XoT) move away from CoT's sequential, step-by-step thought chain and structure reasoning in a non-linear manner, such as Tree-of-Thoughts (ToT) [90] and Graph-of-Thoughts (GoT) [10]. LLM-based agents are also attracting researchers' attention. ReAct [91] generates thoughts and actions in an interleaved manner, leading to human-like decisions in interactive environments. In the planning-execution-refinement paradigm [95], AutoGPT [33] follows an iterative process reminiscent of human problem-solving, i.e., a plan is proposed, executed, and then refined based on feedback and outcomes. Systems like Generative Agents [61] and ChatDev [64] explore multi-agent collaboration; agents interact with the environment and exchange information with each other to collaborate and share task-relevant information.

In this work, we generally follow the prompt generation paradigms in ICL and CoT [83]. Our work is also inspired by the planning-execution-refinement paradigm [95]; that is, just like agents, our system investigates the context to generate a plan that is used to devise an action. The plan is iteratively refined based on user feedback when creating a dialogue policy.

2.2 Multimodal Dialogue Systems

Multimodal dialogue systems can leverage contextual information from multiple modalities, such as text and images, to improve users' experience. The visual dialogue task was first introduced by Das et al. [28], involving two participants in an image-based question-answering task, where a person asks a question about an image and a chatbot gives a response. Mustafazade et al. [54] introduced the image-grounded conversation (IGC) task, which improves the conversation experience by allowing the system to answer and ask questions based on visual content. Despite progress in extending dialogue context modalities, these early works lack natural language modeling capabilities.

Recently, multimodal dialogue systems utilized the capabilities of both pre-trained visual encoders and LLMs. These vision-language models (VLMs) [44, 92, 48, 99] can generate coherent language responses consistent with the visual context. However, they still face challenges in generating natural dialogues that occur in real-life interactions. Furthermore, Li et al. [43] introduced the interactive vision-language task

MIMIC-IT, which allows dialogue systems to engage in immersive conversations based on the multimodal context.

In this work, we combine two state-of-the-art models, i.e., the visual understanding capabilities of VLMs with the dialogue capabilities of LLMs, to enhance the conversational experience.

2.3 Personalized Dialogue Systems

In the dialogue system research field, user profiles such as personality, preferences, and habits can be extracted from user interactions to support personalized dialogue [94]. However, previous studies mainly focus on short-term dialogues, not gradually increasing their understanding of users via long-term interactions. Recently, Xu et al. [88] proposed a long-term dialogue task including user profiles. However, this task does not consider the key elements of extracting, updating, and utilizing user profiles. To address this limitation, Xu et al. [89] recently proposed to identify user personas from utterances in a conversation, which are then used to generate role-based responses.

More recently, Ahn et al. [4] proposed to incorporate visual modalities to enhance the understanding of user profiles from recorded episodic memory. This overcomes the limitation of relying on text-only conversations. However, these episodic memories mainly consist of images and texts shared on social media rather than users' real-life experiences. Combining episodic memory with user profiles, Zhu et al. [97] used LLMs to summarize conversations into episodic memories and user profiles, which were stored in a vector database and retrieved based on the dialogue context in subsequent conversations, resulting in personalized responses.

In this work, we generate historical contexts and user profiles from multimodal information captured in real-world scenarios. Compared with previous literature, our work uses more real-time user information sources. Furthermore, we introduce a mechanism for accumulating user information, enabling the system to enhance its knowledge of users over time.

2.4 Wearable Dialogue Systems

Wearable dialogue systems are a developing area of research that combines wearable technology with conversational AI. Existing wearable dialogue systems focus on specific user groups or application domains, such as the visually impaired or the healthcare domain. Chen et al. [19] proposed a wearable dialogue system for visually impaired individuals that employs smart eyewear with 3D vision, a microphone, and a speaker to facilitate outdoor navigation through conversation. Ozono et al. [60] proposed a system that combines wearable devices and interactive agents, mainly aimed at promoting and encouraging elderly people to take better care of their health. The approach involves integrating health data into conversations with users, to make elderly people aware of their health problems and encourage self-care.

Shoji et al. [72] proposed a dialogue system based on smart eyewear, which interacts with users verbally and provides information relevant to daily life. Additionally, it also gathers biometric data, such as pulse and body temperature, to offer health management guidance through conversation. Kocielnik et al. [40] proposed a mobile dialogue system that collects physical activity data through fitness trackers and guides users to reflect on their daily physical activities through conversations. Calvaresi et al. [14] proposed a mobile health assistant that monitors diet and offers suggestions

through conversations. It can track nutritional information by scanning product barcodes or analyzing food images, offer dietary recommendations, and recommend nearby restaurants based on GPS locations.

In contrast with prior work, our work primarily aims to offer personalized conversations and companionship to the user. By combining wearable technology with advanced conversational AI, our goal is to build a seamless and natural interaction experience that goes beyond functional support. It incorporates contextual information to continually improve the quality of the interaction and adapt to the user's experiences and preferences over time, thereby creating a sustainable personal companion.

3 System Design

This section first clarifies the common ground problem and outlines the essential components that are required to establish and maintain common ground. Then, it provides an overview of the proposed common-ground-aware dialogue system OS-1 and details the four core modules that make up the system. After that, this section describes the system implementation process. The code of the eyewear system is available at here (https://github.com/MemX-Research/OS-1https://github.com/MemX-Research/OS-1).

3.1 Conversational Common Ground

Common ground is a basis of mutual interest established in the course of conversation or communication, which is the key to effective personal conversation [22, 25, 35]. Common ground typically consists of shared values, opinions, experiences, interests, and so on. Relevant research classifies common ground into two categories: personal common ground and communal common ground [22]. Personal common ground refers to the joint experience of participants in conversations. It can be further divided into joint perceptual and linguistic experiences [23]. Joint perceptual experience refers to what the participants perceive together, while linguistic experience refers to the dialogue they produce. Communal common ground reflects the shared information from a community of people, such as social background, preferences, habits, and interests.

This work aims to build an eyewear dialogue system that can be a reliable companion to its wearer. The key challenge for achieving this is to ensure OS-1 can communicate effectively with users and establish a shared understanding. To this end, we identify the following key features required for establishing common ground for OS-1.

1. **Real-time context.** OS-1 must share a common perceptual and linguistic experience with the user. In other words, OS-1 should be able to see what the user sees and hear what the user hears anytime, anywhere. Furthermore, OS-1 should be able to understand the user's speech and visual perceptions semantically. We define this feature as real-time context.
2. **Historical context.** OS-1 should be able to accumulate and summarize the user's real-time perceptual and linguistic context over a long period of time. This may include daily behavioral patterns or significant events identified as historical context in this work. This information helps LLMs continuously build personal common ground and maintain coherence and continuity in dialogue.

3. **Personal Profile.** To establish communal common ground, it is important for OS-1 to gather user-specific information, such as the user's social background, personality, preferences, and habits. This information can be learned from the daily interaction with the user over time. We recognize them as user profiles, which help LLMs respond in a more human-like and user-specific way by adapting to the user's personality and long-term goals. This leads to more coherent and human-like dialogues.

3.2 Workflow

3.2.1 Overall Pipeline

The overall framework of OS-1 is shown in Figure 2, which consists of the following stages:

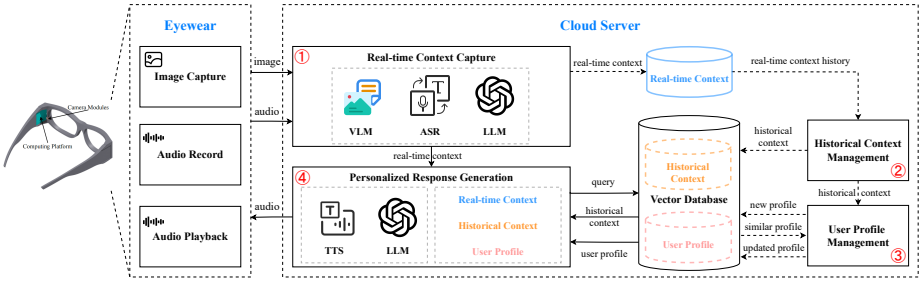


Figure 2: The overall framework of OS-1.

1. Real-time context capture: OS-1 needs to perceive the user's ongoing conversation and understand the in-situ context in real-time, including the visual and auditory surroundings, location, and activity. To this end, OS-1 is equipped with a camera, microphones, and speakers to capture the surrounding images and speech, which are converted into text using the vision-language model, LLaVA [48], and the speech recognition model, Whisper [66]. The converted texts from the images and speech are then fused to form a prompt, which is then fed to LLM-Base[3].

2. Historical context management: OS-1 generates, stores, and recalls the historical contexts, including the user's past daily events and conversation summaries. Daily events and conversation summaries are extracted from the history of real-time contexts, assigned multi-dimensional indices and importance scores, and then stored in the vector database Milvus [80] as historical contexts. To reduce the redundant storage of previously encountered real-time contexts and enable efficient retrieval, OS-1 summarizes the past real-time contexts via a clustering approach that considers semantic similarity. Highly similar real-time contexts are clustered and summarized into distinct events using LLM-Base, thus serving as historical contexts. Additionally, we propose a mechanism to generate the temporal, spatial, and semantic indices for the historical contexts, which are stored in the vector database, enabling retrieval of similar historical contexts in these three dimensions.

[3]LLM-Base is characterized by its faster speed and lower cost. GPT-3.5 [57] is chosen as LLM-Base in this work.

3. User profile management: To further enhance system personalization, OS-1 maintains user profiles over time, including the user's social background, personality, preferences, and habits using Milvus. OS-1 continuously updates the user profile based on the historical context. Specifically, when a new historical context is generated, OS-1 summarizes it into a new user profile record via LLM-Base. Next, the new and existing user profile records are merged via LLM-Base, which generates an updated user profile record that is stored in Milvus. To tackle biases and errors while distilling user profiles, the updating mechanism assigns a confidence score to each user profile record to guide the review and revision of existing profiles.
4. Personalized response generation: To generate personalized responses, we design two LLM-based agents: a dialogue policy agent and an information retrieval agent. The dialogue policy agent orchestrates the structure and flow of the dialogue in alignment with specific objectives, like delivering personal emotional support or addressing problem-solving tasks. The information retrieval agent retrieves the relevant historical contexts and user profiles based on the real-time context. OS-1 combines the real-time context, historical context, and user profile along with the dialogue policy to generate text responses using LLM-Large[4]. Subsequently, these responses are converted to speech and played back on the eyewear.

We present the following case study to illustrate the working flow of OS-1. As depicted in Figure 3, Kim is seated in front of a computer. OS-1 captures this in real-time: "A white wall with a computer monitor placed on a desk." Concurrently, OS-1 registers Kim's verbal statement, "Yeah, the work is finally done." These visual and audio inputs amalgamate to create a real-time context, subsequently used by OS-1 to retrieve the most relevant historical context from the vector database: "(1 day ago) I encouraged the user to take breaks and rest while working late at night." OS-1 utilizes this information to search the user's profile and discern the most pertinent social background: "The user is a student studying computer science and working in a laboratory, currently engaged in writing a paper." The context and profile data paint a vivid portrait of a diligent student, working in front of a computer, likely immersed in writing a research paper. OS-1 then consolidates all three pieces of information and transmits them to LLM-Large for dialogue generation. With this comprehensive context and profile data provided by OS-1, LLM-Large can generate personalized dialogue: "So, the paper is done now? I see you working so hard at the computer. Remember yesterday when I kept telling you to take breaks? Now that the paper is finished, why not take some time to relax?"

Next, we describe the design details of each system building blocks of OS-1.

3.2.2 Real-time Context Capture

OS-1 captures real-time visual and audio signals through the built-in camera and microphone on the smart eyewear. It then uses a vision-language model, LLaVA, to convert visual signals into descriptions, providing textual descriptions of scenes, such as "a desk with a laptop". Additionally, an audio speech recognition model, Whisper, transcribes audio signals into text, recognizing what the user said, such as, "I am so busy". By semantically combining the textual descriptions from visual and audio signals, OS-1 constructs a prompt for LLM-Base to infer the current location and

[4]LLM-Large, in comparison to LLM-Base, has a larger parameters size and typically demonstrates superior performance on downstream tasks. In this work, LLM-Large is implemented by GPT-4 [58]. We also implement Llama2 [77] and Gemini [76] as alternatives to LLM-Large for conducting adaptability analysis.

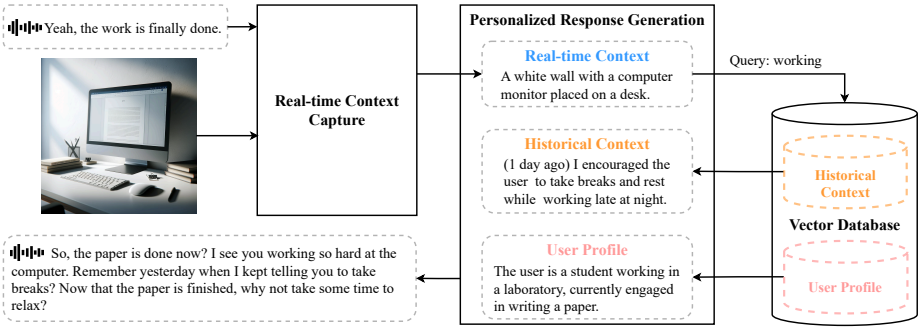


Figure 3: A case illustration of OS-1 conversation workflow (excluding historical context management and user profile management).

activity (Figure A1). For example, it may determine that the user is in the "office" and the user's activity is "working". The texts obtained from the image and audio signals and the location and activity inferred by LLM-Base form the real-time context, which assists OS-1 in understanding the user's current situation.

This work empirically sets a 10-second interval to trigger image capture, assuming that the visual environments typically remain relatively stable and do not undergo significant changes within this time interval. The captured images are processed immediately by the LLaVA-7B-v0 model on an image multi-process service with NVIDIA GeForce RTX 3090 GPUs. It takes approximately 2 seconds to process each image, which is shorter than the current image capture interval. Therefore, the server supports real-time processing. This (background) process is asynchronous and does not block the main process for conversational replies.

3.2.3 Historical Context Management

Over time, OS-1 accumulates an increasing number of real-time contexts, some of which are redundant. For example, for a user who spends a long time working with a computer, the real-time location and activity samples become repetitive. We aim to remove uninformative redundancy from stored real-time contexts before storing them as historical contexts. The extracted historical context falls into two classes: daily events and conversation summaries. Daily events are triplets consisting of time, location, and activity. These allow OS-1 to store historical schedules, e.g., "¡2023-11-01 16:00:00 - 2023-11-01 17:00:00, at the gym, playing badminton¿". The conversation summary includes the topics and details of past conversations, such as "the user mentions writing a paper and asks for tips on how to write it well".

We propose an event clustering method that groups sequences of real-time context into appropriate clusters and summarizes them in event-level text descriptions. To extract conversation summaries, we divide the conversation history into sessions based on contiguous time intervals. For each session, we construct a prompt and use the summarization capability of LLM-Base to extract its summary. Furthermore, to enhance the storage and retrieval of historical contexts in the vector database, we propose an indexing mechanism that organizes the historical context into temporal, spatial, and semantic dimensions, following the format typically used by humans. Ad-

ditionally, it assigns importance scores to the historical contexts based on emotional arousal levels. Historical contexts with higher arousal levels are considered more important and is more likely to be referenced in subsequent conversations, as users are more likely to remember events with stronger emotional impact. We now describe the event clustering, conversation summary, and indexing mechanisms.

Event Clustering Historical contexts are produced through a hierarchical clustering and summarizing process for real-time contexts collected during a day, starting from minute-level clustering, progressing to hour-level, and finally, day-level event clusters. The hierarchical clustering process consists of the following key steps.

1. Embedding matrix calculation: During a day, OS-1 captures a sequence of real-time contexts. For each real-time context within the specified timeframe, we calculate an embedding vector. This is achieved by using an embedding model [69] to transform the textual descriptions of each real-time context (comprising location and activity information) into an embedding vector. The collection of these vectors forms an embedding matrix, where each row represents the embedding vector of a real-time context.
2. Similarity matrix calculation: We compute the cosine similarities between the embedding vectors of all pairs of real-time contexts by multiplying the embedding matrix with its transpose. This results in a similarity matrix, where the element at the ith row and jth column contains the cosine similarity between the ith and jth real-time context.
3. Sequential clustering: Due to the spatiotemporal locality of events, semantically similar real-time contexts are usually contiguous subsequences. Therefore, we traverse the sequence of real-time contexts in chronological order, grouping real-time contexts into the same event if they meet a predefined similarity threshold. A real-time context is considered part of an event if its similarity with the first real-time context of the current event cluster exceeds this threshold. The similarities are determined by consulting the previously computed similarity matrix. More formally, the longest contiguous subsequence that satisfies all the following conditions is selected to cluster an event: 1) the similarity between the first element of the subsequence and the previous subsequence is below the threshold, 2) the similarities among all elements within the subsequence are above the threshold, and 3) the similarity between the last element of the subsequence and the subsequent subsequence is below the threshold.
4. Event summarization: Once the real-time contexts are clustered into events, each cluster represents a sequence of real-time contexts associated with a particular event. We create a prompt that summarizes a collection of real-time contexts that have been grouped together into an event; see Figure A2 for an illustration. Consequently, the corresponding real-time contexts for each longest subsequence are employed as parts of the prompt for LLM-Base. Finally, a summary of the event is extracted, denoted as $\{E^1, \ldots, E^p\}$, where E^i represents a daily event, and p represents the number of distinct events without redundancy after clustering.

Specifically, we choose the embedding model SentenceBert [69] instead of OpenAI embeddings because SentenceBert offers lower latency processing and can be locally deployed. The similarity threshold for clustering was empirically set at 0.85, based

on extensive real-world testing against a ground truth established by developers documenting their daily events.

Conversation Summary To extract summaries from the conversation history, we set an interval threshold that determines the maximum time interval for a session. The threshold separates conversations that exceed the threshold into different sessions, denoted as $\{D^1, \ldots, D^q\} = f_{session}(\{u^1, b^1, \ldots, u^n, b^n\})$, where D^j refers to a session, u^i represents the user's utterance, and b^i represents the OS-1's response. After partitioning the conversation history, we construct a prompt for each session to summarize topics and details by leveraging the summarization capability of LLM-Base, denoted as $\{T^1, \ldots, T^q\} = \mathcal{N}_{llm}(\{D^1, \ldots, D^q\})$, where T^j represents a conversation summary. Finally, the collections of daily event E^i and conversation summary T^j together form the historical context, formally represented as: $C_h^{1:p+q} = \{E^1, \ldots, E^i, \ldots, E^p, T^1, \ldots, T^j, \ldots, T^q\}$.

Indexing Mechanism We propose an indexing mechanism that organizes historical context in three dimensions: temporal, spatial, and semantic. The indexing mechanism aims to generate a list of indexing keys for textual descriptions of historical context, including daily events and conversation summaries. For example, if the historical context is "I plan to have a picnic in the park this weekend", the resulting indexing keys could include "weekend plan", "in the park", and "have a picnic". By allowing multiple indexing keys to be associated with each historical context, OS-1 can do associative retrieval in different dimensions. Specifically, we design a prompt for LLM-Base to extract the textual descriptions related to the temporal, spatial, and semantic aspects of the historical context. These extracted descriptions serve as indexing keys for the historical context.

Emotional factors are used to query the historical context, based on the idea that strong emotions make experiences more memorable [52]. To achieve this, we design a prompt and leverage LLM-Base to evaluate the level of emotional arousal associated with a given historical context. This level determines the significance of the historical context, which is represented by an importance score ranging from 1 to 10. We assign higher importance scores to historical contexts with intensified emotional arousal, thereby increasing the likelihood of mentioning them in the conversation.

In summary, the indexing mechanism for historical context can be described as follows:

1. generate indexing keys from multiple dimensions for each historical context;
2. assign an importance score to each historical context; and
3. store the historical context in the vector database, along with the corresponding indexing keys and importance score.

3.2.4 User Profile Management

Historical context represents the user's daily events and conversation summaries. It can therefore provide important clues about the user profile, including social background, personality, preferences, and habits. By summarizing patterns from the historical context, OS-1 can distill the user profile and thereby improve the personalized user experience. For example, if a user frequently eats spicy food, it becomes evident that the user has a preference for spicy food. The user profile consists of a textual description of a specific aspect of the user, along with a confidence score that indicates

the reliability of the information. We introduce an additional confidence score because user profile distillation is an ongoing process that aims to correct biases and errors via continuous refinement.

The process of distilling a user profile from a historical context involves several steps, which can be described as follows (see prompt details in Figure A4):

1. When a new historical context record is generated, we use a prompt (left in Figure A4) instructing LLM-Base to summarize the historical context into a new user profile record.
2. The new user profile record is then processed using an embedding model [69], resulting in a query vector. The query vector is used to retrieve the existing user profile record with the highest cosine similarity from the vector database if one exceeding the similarity threshold exists. When querying user profiles, four aspects are compared with the querying semantics; these include social background, personality, preferences, and habits.
3. If no existing user profile record meets the similarity threshold, the new user profile record is considered unique and is directly stored in the vector database. Otherwise, we use a prompt (right in Figure A4) for LLM-Base to revise the concatenation of the existing user profile record and the new user profile record to generate the updated user profile record with a new confidence score. Here is an example of updating the confidence score. If there is more information indicating that the user likes spicy food, the confidence score of the user profile record will increase. If the information is still insufficient, the confidence score for the user profile record will remain low. Finally, the existing user profile record is replaced by the updated user profile record in the vector database. The updating mechanism enables OS-1 to rectify inaccurate user profiles and reinforce correct user profiles over time.

Specifically, the similarity threshold is determined experimentally, where similar user profiles are manually categorized, and various thresholds are tested for retrieval precision. Ultimately, we select 0.5 as the similarity threshold.

3.2.5 Personalized Response Generation

To generate personalized responses, we design two LLM-based agents, namely a dialogue policy agent and an information retrieval agent.

Dialogue policy agent This agent orchestrates the structure and flow of the dialogue in alignment with specific objectives, like delivering personal emotional support or addressing problem-solving tasks. It can plan the direction of the conversation based on real-time context, such as guiding users to express their opinions by asking questions and providing additional information to drive the conversation forward.

The dialogue policy agent has two modules: *planner* and *decider*. *Planner* produces a dialogue objective and policy. *Decider* determines the specific action to be taken. The objective refers to the desired outcomes a conversation aims to achieve, providing direction for the conversation, such as delivering personal emotional support or addressing problem-solving tasks. The policy is a plan that outlines a series of steps to achieve the objective of a conversation. The action refers to a specific step taken following this policy. To implement these two modules, we have designed a chain-of-thought prompt for each module as the instruction for LLM-Base. The inputs of the

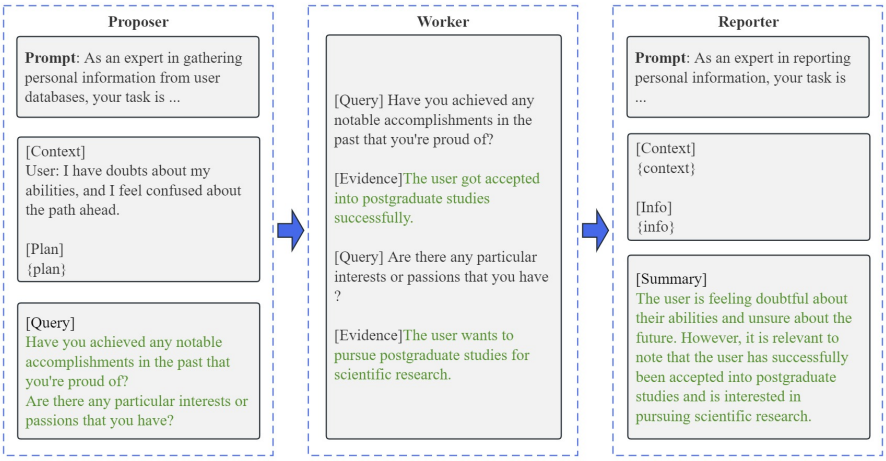


Figure 4: An example of the information retrieval agent.

prompt consist of two parts: the planned dialogue policy and the ongoing conversation history. Below, we will introduce the reasoning process in the chain-of-thought prompt for each module. Each step of the reasoning process is a thinking step in the chain-of-thought prompt.

The reasoning process of the *planner* includes the following three steps.

1. **Define the objective** of the dialogue based on the real-time context, e.g., providing emotional support.
2. **Proposing a policy** based on the defined dialogue objective. For example, a policy includes the following steps when the user has negative feelings: affirming the user's negative emotions, inquiring about the causes of these emotions, and guiding the user to mitigate these emotions.
3. **Refining the policy** as the dialogue progresses based on the user's feedback. For example, when the policy from the previous conversation is to help the user solve a problem, but the user says, "I don't want to think about how to solve the problem right now, can you just comfort me?", the dialogue policy should be adjusted from problem-solving to providing comfort.

The reasoning process of *decider* includes the following three steps.

1. Analyzing progress: It determines which actions of the policy have already been executed during the conversation by comparing OS-1's responses with the planned dialogue policy. For example, after OS-1 responded with "I understand you feel sad. It's completely okay. Can you tell me more about what's been going on that made you sad?", the analysis shows that it has taken two actions: affirming the user's negative emotions and inquiring about the causes of these emotions.
2. Evaluating outcome: It analyzes the user's feedback from the conversation to evaluate the effectiveness of the actions employed; for example, whether the user's negative emotions have been mitigated. This evaluation helps to determine whether the dialogue policy is achieving its objective. For example, if the

user says "I feel a bit better", this indicates that the user's mood has improved, demonstrating the effectiveness of the adopted dialogue policy.

3. Deciding the next action: It decides the next action to be taken based on the progress and outcome during the conversation. For example, if the user's negative emotions haven't been mitigated, it continues to address the emotional distress. If the user's negative emotions have been mitigated, it starts guiding the user toward resolving the root problem.

Figure A3 shows an example of the dialogue policy agent. Each module consists of a prompt that describes the task and provides guidance for the reasoning process. The prompt is used as the system prompt for LLM-Base to execute the module's functionality. During the conversation, the planner generates a multi-step dialogue policy based on the real-time context. Subsequently, the decider determines the specific action to take. Finally, the generated text of the action serves as a prompt for guiding LLM-Large to generate a reply that aligns with the specified direction of the policy.

The Information Retrieval Agent determines which user information to retrieve based on the real-time context and summarizes the retrieved user information. It leverages real-time context and retrieves user information from historical contexts and user profiles. Then, OS-1 combines the real-time context and the relevant historical contexts and user profiles as a personal context, along with the dialogue policy planned by the dialogue policy agent, to serve as a prompt for LLM-Large to generate text responses. Finally, the generated reply is converted into speech using a text-to-speech service [26] and transmitted to the smart eyewear device for playback.

The information retrieval agent includes three modules: *proposer*, *worker*, and *reporter*. the *proposer* and *reporter* use prompts for LLM-Base to generate queries and summarize query results, respectively, while the *worker* performs these queries through semantic retrieval on the vector database, Milvus.

1. The *proposer* generates queries for the vector database. It is responsible for suggesting which aspects of user information should be retrieved based on the real-time context. Concretely, it produces a list of queries for retrieving relevant historical contexts and user profiles. Each query describes a specific aspect of the user. For example, the user doubts himself and feels confused about the future. Consequently, based on the real-time context, the proposer module generates queries about past achievements, interests, and passions. The retrieved achievements can be used to encourage the user, while the retrieved interests and passions can help clarify future opportunities.

2. The *worker* is responsible for executing queries on the vector database and retrieving the relevant information. These queries from the proposer module are assigned to the worker module for execution while retrieving historical contexts and user profiles. During retrieval, OS-1 determines the cosine similarity between the query vector and the vectors of historical context and user profile records. Once retrieval produces a set of candidate records, a rank score is calculated for each record, and they are sorted to enable the selection of the k records with the highest rank scores. Rank score calculation is similar to that used in generative agents [61]: $S_{rank} = S_{similarity} + S_{importance} + S_{recency}$, where the recency score $S_{recency}$ accounts for the recency of the update time of record creation (the more recent the record, the higher the recency score). For example, the retrieved relevant historical context record is "The user was accepted into postgraduate studies." and the retrieved relevant user profile record is "The user wants to pursue postgraduate studies for scientific research."

3. The *reporter* is responsible for extracting and summarizing relevant information from retrieved historical context records and user profile records. These query records from the worker module are assigned to the reporter module for summarizing. For example, the summary of user information is "the user has successfully been accepted into postgraduate studies and is interested in pursuing scientific research."

Finally, the real-time context, relevant historical contexts, and user profiles are retrieved by the information retrieval agent and are combined to produce the personal context. This and the dialogue policy action planned by the dialogue policy agent are used as prompts for LLM-Large to generate personalized responses (Figure A6).

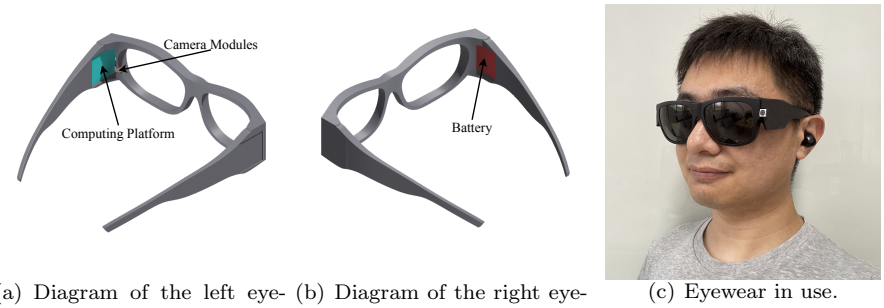
3.3 Implementation

3.3.1 Hardware Design

In the smart glasses designed for OS-1, computing components, and a power source are integrated into the glasses frame. Figure 5 illustrates the eyewear hardware prototype that brings our vision to life.

Inspired by Chang et al. [17], OS-1 eyewear hardware is a custom design. It incorporates a Snapdragon Wear 4100+ computing platform, one 8-megapixel camera, a 700 mA-h battery, as well as Wi-Fi and Bluetooth wireless interfaces. Combined with a Bluetooth earpiece, OS-1 eyewear is wearable and meets the requirements for visual and auditory modalities essential for our study.

The Snapdragon Wear 4100+ [39] computing platform is integrated into the left arm of the glasses. The glasses are a standalone device with adequate computing power for real-time data processing.



(a) Diagram of the left eyewear arm. (b) Diagram of the right eyewear arm. (c) Eyewear in use.

Figure 5: Prototype diagram and wearing photo of the eyewear.

The eyewear frame also houses an 8-megapixel scene camera within its left arm. The scene camera has an 84 degree field-of-view (FoV) and captures the surrounding scene images, providing visual context to OS-1. Examples of images captured by the scene camera in different lighting conditions can be found in Appendix A2.

Taking advantage of the broad applicability and efficient performance of eye-tracking algorithms on eyewear devices [16], the system was designed with a 5-megapixel eye-tracking camera to capture the user's visual attention. However, the eye-tracking feature was disabled in the first generation of the system to enable more rapid implementation.

To provide a well-balanced and comfortable fit, a battery is integrated into the right arm of the glasses, thereby balancing the frame. The capacity of the built-in battery is 700 mAh, which can power the system for at least 60 minutes. To enable longer operation, the glasses have a wired magnetic charging port with an external mobile power bank, thereby extending the device's usage time.

An off-the-shelf Newmine L10 Bluetooth earpiece was used to accelerate prototyping and enable a proof-of-concept study[5]. It is a compact and lightweight device that weighs 3.5 grams. The earpiece supports Bluetooth 5.0 and comes with noise-reduction technology enabling clear communication. With a 50 mAh battery, it supports 4 hours of talk time. It has a conventional USB magnetic charging dock. Any Bluetooth earpiece with similar functionality would, in principle, be compatible with our system.

3.3.2 Software Design

In this section, we elaborate on the software design aspects of the eyewear system. The system operates on Android 8.1, providing a platform for communication between the user and the cloud services. Initially, the user is required to configure the WiFi connection to access the cloud and enable uninterrupted communication. The software has three functions: capturing audio, capturing scene images, and playing the audio output of responses received from the cloud server.

- **Audio:** The eyewear system continuously captures audio from the user's surroundings, which is streamed to the cloud in real-time. In the cloud, a voice recognition system processes the audio stream, converting it into text.
- **Image:** The eyewear system periodically captures 640×480 scene images at specific time intervals (every 10 seconds in this work). To optimize data transmission, the captured images undergo JPEG compression before being uploaded to the cloud. Once uploaded, the cloud performs feature extraction on the images, allowing insight into the user's current environment.
- **Playback:** The eyewear system plays the human-like audio responses generated in the cloud.

3.3.3 Cloud Services

The cloud services consist of five components, each capable of handling multiple processes concurrently to support simultaneous interactions with multiple users. Redis [51] queues are used for communication among these services.

- **Data Server:** The data server is responsible for facilitating communication with the eyewear. It is built on the FastAPI framework [68] and has two key interfaces. The first allows uploading data, including timestamps, audio, images, and other relevant information. Upon receipt, these data are placed in the appropriate queue, awaiting processing. The second interface returns generated audio replies. It retrieves audio from the response queue and streams it to the user's eyewear through the Starlette framework [50].
- **Image Server:** The image server component retrieves images from the queue and processes them using the LLaVA [48] model for content recognition. Specifically, the LLaVA-7B-v0 model is employed, with parameter settings as follows: max_new_tokens = 512 and temperature = 0.

[5] `https://www.google.com/search?q=Newmine+L10&tbm=isch`

- **Audio Server:** For each online user, a dedicated thread is created to handle the audio input. This thread continuously receives audio data from the users' eyewear system and uses Whisper [66] for speech recognition.
- **Chatbot Server:** The chatbot server is the core cloud service. It generates responses based on the user's surrounding environment and conversation content. The responses include textual content, as described in Section 3.2.
- **TTS Server:** The TTS server converts textual responses into audio format. This component uses a commercial text-to-speech service [26] for efficient and high-quality audio synthesis.

The processing steps during a conversational round trip are illustrated in Figure 6. After the speech audio is streamed to the server, it takes 1.57 seconds for speech recognition. After that, the server notifies the user with a beep while continuing context retrieval and prompt assembly, which costs approximately 0.12 seconds. Then the prompt is fed into an LLM to generate the response. To minimize the latency for the client to receive audio playback, the system doesn't wait for the entire process of generating a response to finish. Instead, once the first phrase (around 5 tokens) is generated, the output text is converted to audio and immediately transmitted to the user. Thus, the overall latency from a request to receiving the audio response is approximately 3.59 seconds.

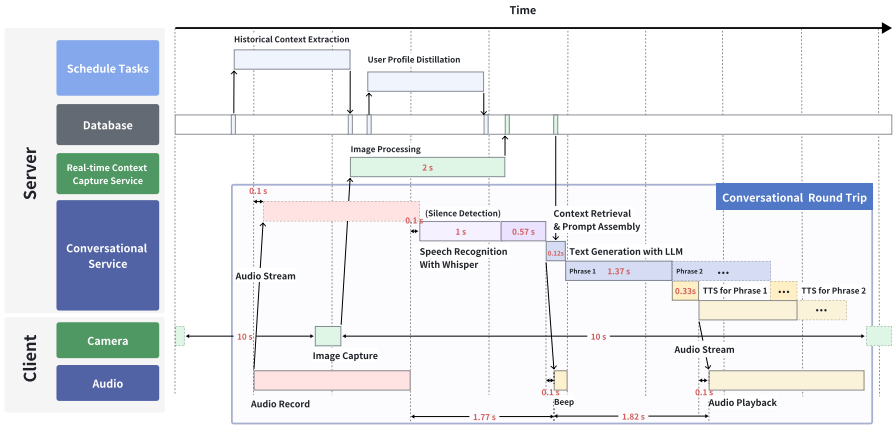


Figure 6: Latency of each step during system processing.

4 Evaluation

This section evaluates the performance of OS-1 empowered by personal context capturing, which is designed to cater to diverse users with varying profiles who engage in various conversation scenarios. To this end, we first consider a variety of conversation situations and simulated users with various profiles in a controlled laboratory setting. Then, we recruit volunteers to participate in pilot studies for approximately 14 days to examine the long-term effectiveness when OS-1 is used in real-world scenarios.

We simulate users and visual scenes in the laboratory study for the following three reasons. (1) It is infeasible for us to recruit a large number of volunteers covering a wide range of personalities, social backgrounds, and experiences, and further track their conversations with OS-1 in various real-world environments. We therefore simulate diverse types of users and visual scenes in the laboratory study to verify the performance of OS-1. (2) Simulating users and scenes allows us to control experimental confounders without introducing unwanted noise. For example, we can conduct ablation studies to compare the performance of OS-1 with baseline methods without one or more types of personal context. (3) Simulating users and scenes not only facilitates data collection but also reduces the experiment cost and time, as recruiting participants is costly and time-consuming. We are also aware that the simulated users and scenarios used in our laboratory experiments may not exactly reflect real-world situations. As a result, we also conduct pilot studies to further test the system's performance and ability outside of the controlled environment. In the pilot studies, we recruit participants to further verify the system's performance in practice.

4.1 In-lab Experiments

For the in-lab experiments, we first outline the experimental settings to simulate various daily-life scenarios and users with diverse social backgrounds and personalities. Then, we compare the performance of our proposed system, OS-1, with that of the baseline methods without considering personal context. Next, we use a case study to further explain why OS-1 outperforms the baseline methods. Lastly, we show the adaptability of our proposed methods to integrate with different LLMs by comparing the performance of the OS-1 variants with the LLM-Only baselines.

4.1.1 Experimental Setup

(1) User Simulation To verify OS-1's ability to adapt to diverse users, we use GPT-4 to simulate virtual users with varying personalities, social backgrounds, and experiences, following an approach in prior work [2]. In particular, we create 20 distinct virtual users consisting of 10 males and 10 females, ranging in age from 15 to 60. Each virtual user is assigned a name randomly selected from the U.S. 2010 Census Data [13]. Also, we assign each user a personality based on the Myers-Briggs Type Indicator (MBTI) [56]. To make the virtual users more realistic, we use GPT-3.5 to complete each user's characteristics with an occupation, preferences, and habits, along with daily routines of 10 days.

(2) Visual Scene Simulation We use GPT-3.5 to directly simulate the 20 user's daily visual scenes at a given moment. The visual scenes represent the visual surroundings perceived by users, and they are represented as a four-tuple, including time, location, activity, and a brief text description of what the user perceives. For example, a college student, Benally majoring in Chemistry, might experience a visual scene of ¡2023-10-02 Monday 9:00-12:00, Chemistry Lab, Attending lectures and practicals, "A table filled with beakers and test tubes." ¿.

In total, we simulate a total of 80 daily visual scenes for each user, with 8 scenes per day and a duration of 10 days.

(3) Dialogue Simulation We randomly select three daily visual scenes for each user and ask the user to initiate a conversation with OS-1 based on the visual scene.

Each conversation consists of three rounds. This way, we get each user's personal context, consisting of the simulated speech and their daily visual surroundings. Then, we cluster the personal context and summarize the historical context with a few sentences to describe it. Furthermore, we distill the user profile using the historical context.

(4) Test Scenario Simulation We also create the test scenarios to verify OS-1's capability to reach better grounding by utilizing their context. To achieve this, we recruit a human experimenter to review the virtual users' personal context and instruct the experimenter to specify a chat topic and a brief text that describes a visual scene. For example, a chat topic may be "dinner recommendations" and a visual scene may be "a commercial street with a pizza stand".

(5) Evaluation Measures Evaluating the quality of conversational response for open-domain chatbots is challenging [37, 29, 31], as the criteria are typically subjective and vary with the application domains and design purposes [37, 47]. Therefore, customized human evaluation metrics are often used for chatbot performance [74]. Following these ideas [74, 49, 75, 34], this work designs customized human evaluation metrics to evaluate the performance of OS-1, our eyewear dialogue system.

The goal of OS-1 is to establish conversational common ground with users and drive personal conversation. To assess the quality of response content from OS-1 and the long-term interactive effect, we first design the following three metrics.

- *Relevance* The conversations between participants are usually related to the in-situ environment and what they are just talking about, i.e., the real-time context in this work. To measure relevance, we design a measure of the correlation between the response and the users' speech and in-situ environment, including the location, visual surroundings, current activity, and time. It is similar to a metric proposed by previous studies [47].
- *Personalization* Having a high correlation between the response and the real-time context is not enough, as the time-evolving knowledge from users, i.e., historical context and personal profile, are also essential to building common ground. By better utilizing this context, OS-1 can produce more personalized responses, thereby increasing the quality of the response content. Therefore, we define the Personalization score to determine how closely the response relates to the user's personal information, including their profile and the semantics derived from what they are currently viewing and chatting about, as well as their past interactions with OS-1.
- *Engagement* We also expect OS-1 to provide engaging conversations. In this way, users can enjoy the conversations and be willing to continue the conversation with OS-1. We design the Engagement score to measure how interested a user is in the response and whether it will lead to further conversation. This is similar to previous studies [45, 47].

The above three metrics provide additional information and insights in a finer granularity; we further propose the Grounding score, which directly assesses the overall performance of OS-1 in establishing and leveraging common ground. Therefore, the Grounding score and these three metrics are supplementary to each other. Ideally, the higher scores in the above three metrics should result in a higher Grounding score.

Regarding the rating process, we instruct human examiners to score each response from the system using the above four metrics. This work adopts the widely-used 5-point Likert scale [79, 67]. Also, to mitigate the possible bias from human raters, we involve 15 human raters and ensure that each response is evaluated by at least three of them. We report the mean values of the ratings.

(6) Baseline Methods As there are no previous systems that can be directly compared to OS-1, we conduct ablation studies to evaluate it. The ablation studies have two purposes. First, they evaluate OS-1's ability to establish common ground with users by incorporating their personal contexts and generating personalized responses. Second, they quantify the contribution of real-time, historical, and user profile context to establishing common ground.

- *w/o* P: This method solely relies on the real-time and historical context to boost context-aware dialogue generation. The user profile is omitted.
- *w/o* PH: This method only uses the real-time context to enhance context-aware dialogue generation. It omits historical context and user profiles.
- *w/o* PHR: This method uses an LLM to produce responses during interaction with users, omitting any personal context.

4.1.2 Overall Performance

Figure 7 shows the performance of different methods in terms of Grounding, Relevance, Personalization, and Engagement scores assigned by human raters. As we can see, OS-1 achieves the highest scores among the four methods. Compared with the *w/o* PHR, OS-1 improves the Grounding score by 42.26% ($p < 0.0001$). Also, OS-1 substantially improves the performance by 8.63% ($p = 0.0033$), 40.00% ($p < 0.0001$), and 29.81% ($p < 0.0001$) in Relevance, Personalization, and Engagement scores. The results have been confirmed to be statistically significant through paired t-tests.

Next, we further investigate the factors that aid in better grounding from the viewpoint of human raters. We ask the human raters to review all the responses generated by various methods and identify the factors that contribute to good grounding for each response. The raters consider three aspects: the proposed real-time context, historical context, and user profile. Also, the raters are allowed to select multiple factors that lead to good grounding. We then calculate the percentage of the number of each factor selected by the raters out of all the selected responses. The results are presented in Figure 8. We observe that personal context plays a significant role in grounding. Specifically, we find that (1) the percentage of the methods that include the real-time context is higher (73%, 73%, 80% for OS-1, *w/o* P, and *w/o* PH, respectively) compared to those without such context (51% for *w/o* PHR). (2) similarly, methods that include historical context have a higher percentage (23% and 21% for methods OS-1 and *w/o* P, respectively) than those without such context (5% and 10% for methods *w/o* PH and *w/o* PHR, respectively). and (3) the percentage of methods that include the user profile is higher compared to those without this kind of context (39% for OS-1, compared to 26%, 22%, 16% for *w/o* P, *w/o* PH, and *w/o* PHR, respectively).

4.1.3 A Case Study

We provide a case study to offer further insights regarding why OS-1 outperforms the baselines for personalized dialogue. Figure 9 shows the dialogue sessions between a

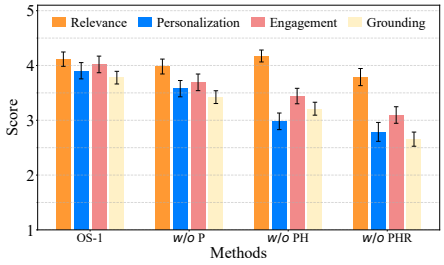


Figure 7: Performance comparison of OS-1 with the baseline methods.

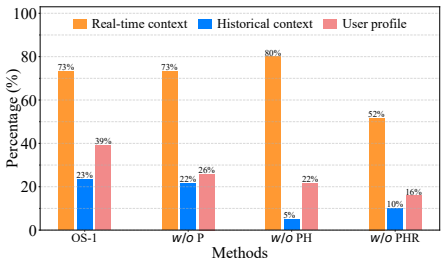


Figure 8: Percent contribution of various factors to better grounding.

simulated user named Kim and four systems, including the proposed OS-1 and three baseline methods. As shown in the real-time context, Kim is walking along a commercial street with a coffee shop and a milk tea shop, an important piece of real-time context. Historical context reveals that Kim has been to a coffee shop for business recently. The user profile reveals that Kim dislikes coffee. Compared with the three baseline methods, we observe that OS-1 provides the most appropriate response by using both real-time context-relevant information (highlighted in green) and user profile context-relevant information (highlighted in red). In contrast, other baseline methods omit one or more, such as *w/o* P, which retrieves historically relevant information accurately but omits user profiles, and *w/o* PHR omits all three types of context information.

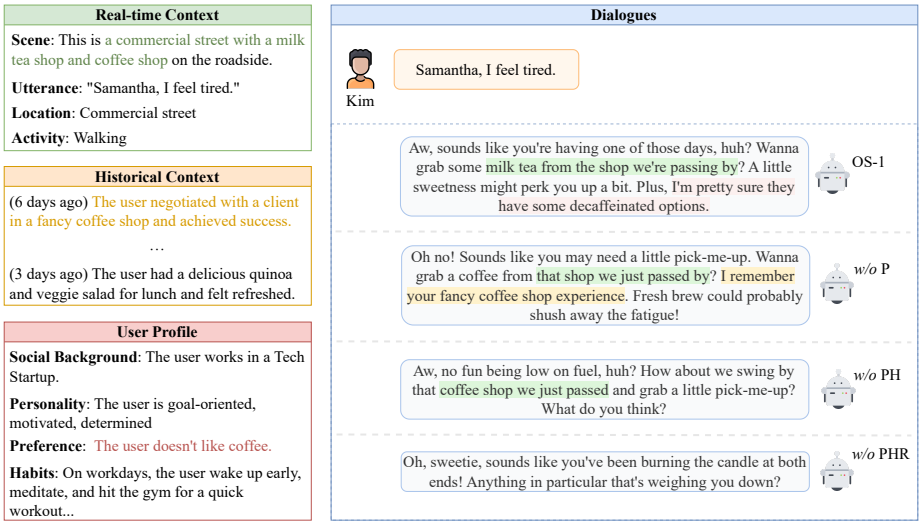


Figure 9: A case of shop recommendation. Left: personal context captured by OS-1. Right: four dialogue sessions between a simulated user and four systems.

4.1.4 Adaptability Analysis

To illustrate the capability of our proposed method to inter-operate with different LLMs, we integrate Llama2 [77] (i.e., Llama2-7B-chat) and Gemini [76] (i.e., Gemini-Pro API) into our framework and conduct experiments following the same procedure as the previous ablation study, and add the following four baselines.

- *OS-1-Gemini:* An OS-1 variant that replaces GPT-4 with Gemini for response generation.
- *Gemini-Only:* This baseline method uses Gemini to produce responses during interaction with users, omitting personal context.
- *OS-1-Llama2:* An OS-1 variant that replaces GPT-4 with Llama2 for response generation.
- *Llama2-Only:* This baseline method uses Llama2 to produce responses during interaction with users, omitting personal context.

Figure 10 shows the Grounding, Relevance, Personalization, and Engagement scores for the four new baselines. The results are consistent with those of OS-1, which uses GPT-4. OS-1 outperforms the baseline methods without personal context. Specifically, OS-1-Gemini improves Grounding by 8.33% ($p = 0.0342$), Relevance by 8.04% ($p = 0.0127$), Personalization by 12.57% ($p = 0.0201$), and Engagement by 0.93% ($p = 0.4209$), which suggests that OS-1 can help LLMs provide personalized and context-relevant responses. Figure 11 compares the outputs generated by OS-1 integrated with different LLMs. When incorporated with the personal context provided by OS-1, OS-1-Gemini, OS-1-GPT-4, and OS-1-Llama2 produce context-relevant responses.

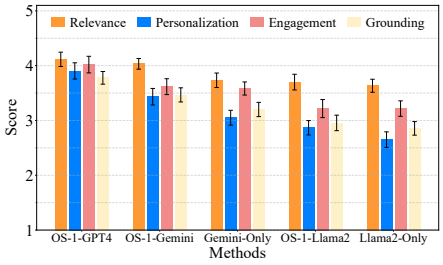


Figure 10: Performance comparison of the adaptation to different LLMs.

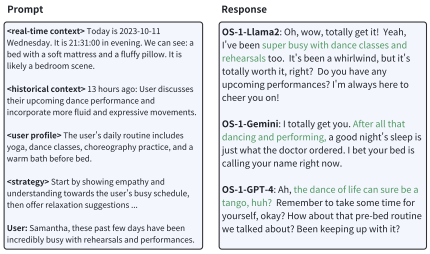


Figure 11: Comparison of outputs from different LLMs.

4.2 Pilot Study

In addition to laboratory studies, we also performed a two-week pilot field study to observe the behavior of OS-1 in the real world. First, we determine whether OS-1 is capable of extracting users' profiles and long-term historical contexts through multiple interactions. We then assess OS-1's ability to establish common ground with its users. In particular, Grounding, Relevance, Personalization, and Engagement scores were measured. Second, we describe two potential downstream applications for which OS-1 would be appropriate: providing emotional support and personal assistance. All experimental procedures are approved by the ethics committee at our university.

4.2.1 Procedure of Pilot Study

We recruited volunteers from our university to participate in the pilot study. Prior to the pilot study, we informed participants that the glasses would perceive their daily visual scenes and audio and that researchers would examine their daily chat logs recorded in the eyewear system if given permission. The raw sensed image and audio data are deleted immediately after feature extraction, and only anonymized semantics are transmitted and stored in the cloud. All participants were aware of this procedure and signed consent forms prior to their experiments. We also provide each participant with instructions on how to use OS-1, including starting a conversation, turning off the system, and reviewing the conversation history using the designed web service.

The pilot study consists of two phases with slightly different purposes. Each phase lasts 7 days, and participants are encouraged to engage in at least 30 minutes of conversation with OS-1 daily. In the first phase, we recruit 10 volunteers (aged 22-28, 6 males and 4 females, referred to as P1 to P10 in the following text) plus 3 authors to participate in the pilot study. The main reason for involving three authors is to enable the collection of first-hand user experience and make necessary and timely adjustments to the system pipeline. Those 3 authors only appear in the first-phase studies and are excluded from the second phase. Due to the limited concurrency support by the early OS-1 prototype, we reserved time slots for participants. After completing the first phase, we spent one month improving support for concurrency and hardware usability. Then, we conducted a second-phase pilot study with 10 participants aged 22-29, 7 males and 3 females, referred to as P11 to P20. In the second phase, the participants can use the system anywhere and at any time.

After completing the daily experiments in both phases, we ask them to review the responses generated by OS-1 and score them using the same criteria as in the laboratory experiments, i.e., Grounding, Relevance, Personalization, and Engagement score. We also make a slight adjustment to make the score more suitable for in-field evaluation. In the pilot studies, we use an 11-point Likert scale instead of the 5-point Likert scale used in the laboratory experiments. The reason behind this is to increase resolution and enable the representation of more subtle variations over time in pilot studies. Previous studies have shown that a larger number of categories can capture the finer distinctions of attitudes or opinions [85, 42]. This is consistent with the existing works that use 11-point scales to deal with complex and subjective evaluation [8].

In both phases, we ask the participants to use the system for at least 30 minutes per day and encourage them to use it as long as possible. In the first phase, we collected an average of 27.17 minutes of conversation per day, with a standard deviation of 14.83 minutes. The number of utterances from both sides was 53.70 on average, with a standard deviation of 34.18. Each participant's utterance had an average of 10.76 words, with a standard deviation of 8.62. In the second phase, we collected an average of 27.64 minutes of conversation per day, with a standard deviation of 13.82 minutes. The number of utterances from both sides was 65.62 on average, with a standard deviation of 38.64. Each participant's utterance had an average of 10.66 words, with a standard deviation of 11.63.

We also measured the frequency of participant conversations with OS-1 during the pilot study. We divided daily interactions into sessions that start when a participant initiates a conversation with OS-1 and end when a participant does not reply within 3 minutes of OS-1's response. The average number of daily sessions is the conversation frequency. In the first phase, participants had an average of 2.59 sessions per day, with a standard deviation of 1.74. In the second phase, participants had an average of 2.02

sessions per day, with a standard deviation of 1.60.

To evaluate whether personal context contributes to a better common ground with OS-1 and leads to more personalized responses, we asked participants to select the responses that strongly indicate that OS-1 understood or did not understand them. This work encourages the human examiner to choose representative utterances without limiting the number, and it depends on the participants' decision. In total, our participants report 249 positive and 162 negative responses during the two 7-day phases. They account for 6.16% and 4.01% of the total number of utterances, respectively. We do not manually check all the utterances, as this would be costly and time-consuming. Moreover, it is unnecessary to examine each individual utterance, as the conversations with participants either have specific goals or are for entertainment purposes, and intuitively, not every sentence contains meaningful information that reflects the user's personal context.

4.2.2 Performance of OS-1 in Pilot Study

Figure 12 and Figure 13 depict the average evaluation scores of the 10 participants over the 7-day first and second phases. The participants find OS-1's responses to be relevant, personalized, and engaging, with most scores higher than 5. Moreover, despite small fluctuations, all scores show a consistently increasing pattern over the 7 days. This indicates that OS-1 is able to generate responses tailored to each participant's personality throughout time. More importantly, as we can see, the Grounding score also shows an increasing trend over time throughout the two pilot phases, which indicates that OS-1 is capable of continuously reaching common ground with users through long-term interactions. As a result, users perceive that OS-1 understands them better over time because its conversations become more relevant, personalized, and engaging. To analyze the 249 positive responses, we provide four possible contributing factors, with three related to personal context and one to the LLM.

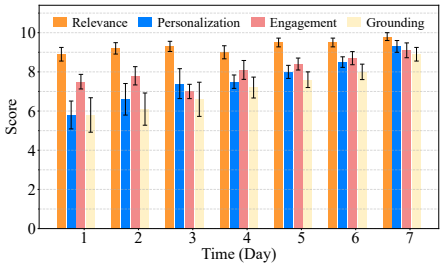


Figure 12: The average evaluation scores of all participants in phase 1.

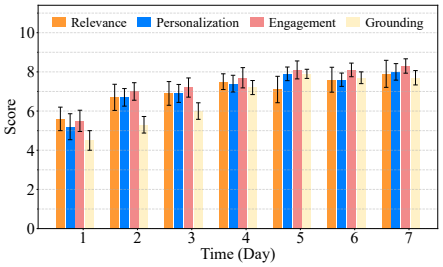


Figure 13: The average evaluation scores of all participants in phase 2.

- **Real-time context factor** The response is linked to the scene and the conversation the user had at a specific time.
- **Historical context factor**: The response is retrieved from the historical semantics stored in the database.
- **User profile factor**: The response is closely related to the summarized user profile, such as personality and habits.

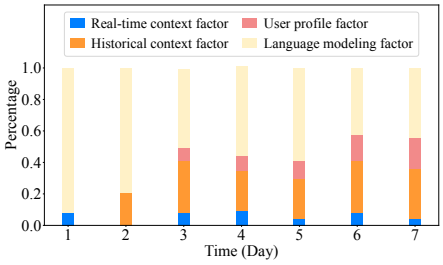


Figure 14: The daily percentages of the four factors in phase 1.

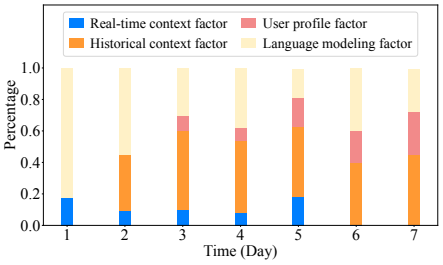


Figure 15: The daily percentages of the four factors in phase 2.

- **Language modeling factor**: The response is generated solely by the LLM without taking into account personal context.

A human examiner reviews the selected responses and the corresponding reasons, and manually assigns one of the above contributing factors to each response that best explains why the participant selected it. We calculate the percentage of the number of each factor selected out of the number of all the selected responses. The results are provided in Figures 14 and 15. The two figures show the daily percentage contribution of each factor to establishing common ground during the 7-day periods in the two pilot phases. A higher percentage of a factor implies a more frequent contribution to a user-preferred response. We observe that the percentages of the personal context-related factors increase over time, e.g., the historical content factor and the user profile factor, while that of the LLM factor decreases. This also suggests that OS-1 can utilize the user's historical contexts and learn user profiles from past interactions to generate better personalized responses.

Next, we present three concrete cases to explore how personal context-related factors contribute to personalized dialogue responses.

(1) A Real-time Context-Aware Case Figure 16 shows a case of real-time context playing a significant role in the dialogue. Specifically, OS-1 observes that Participant P11 places a Teddy bear on a desk, and thus, her greetings involve the information related to that particular visual scene, i.e., *a cool teddy bear*.

(2) A Historical Context-Aware Case Historical context can ensure that the conversations are coherent and consistent over time. We present one example in Figure 17. On day 4, Participant P16 tells OS-1 about playing a game. OS-1 immediately guesses that P16 might be playing the farming game the participant played a few days ago. They then engage in a coherent conversation about the game. OS-1 also recalls, from a previous conversation, that P16 has described playing this farming game as a "chill time". In response, OS-1 comments, "It must be pretty relaxing overseeing your own little digital utopia."

Next, we provide a case study to further provide insights into how the amount of historical context affects the system's performance.

Regarding the amount of historical context needed for optimal LLM performance, we found that the amount of historical context depends on the user and changes over time. Here, we use the Grounding score to reflect the performance. Ideally, users

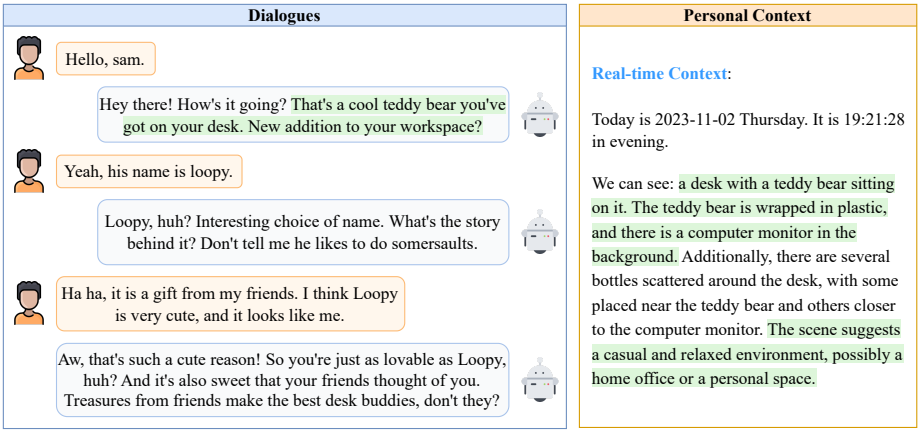


Figure 16: A case that OS-1 is real-time context-aware (Participant P11, Day 1). Green highlights indicate the real-time context-relevant information.

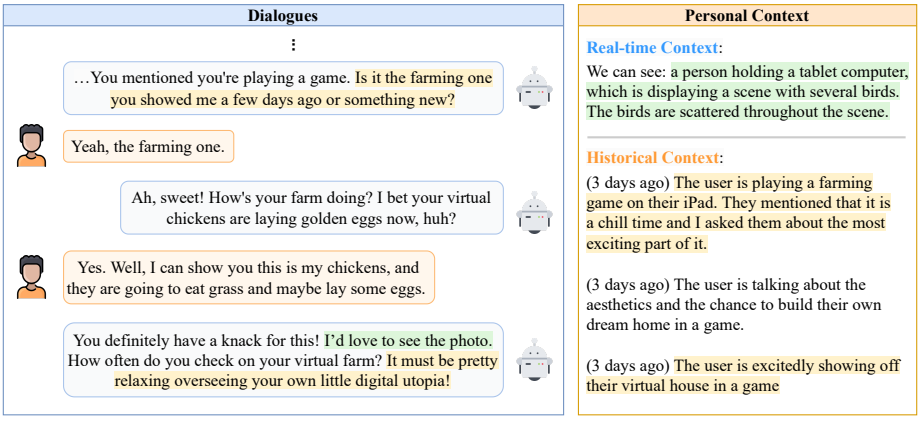


Figure 17: A case that OS-1 is historical context-aware (Participant P16, Day 4). Yellow highlights indicate the historical context-relevant information.

should feel that OS-1 understands them better over time, leading to a higher Grounding score. This is because the amount of historical context and user profiles increases as users interact with OS-1. However, users with different conversation preferences with OS-1 may observe different trends in their Grounding scores. We illustrate this with participants P14 and P19 from phase 2, showing their understanding scores and historical context quantity in Figure 18. It can be seen that, despite the fluctuation, P14 and P19 have increasing trends for Grounding scores, but P14's Grounding score reaches a plateau on the fourth day, while P19's score continues to rise.

We examine the historical context generated by P14 and P19 during their interactions with OS-1 to gain more insights. We observe that compared with P14, P19 has more historical context items, including more diverse conversations and daily events. We show a portion of the daily historical context for P14 and P19 in Figure 19, with repeated items. Different topics are highlighted using various colors. P14 and OS-1 repeated only two topics over the seven days, namely "Digital Monster" and "LuLu the Piggy". P19 had a wider range of topics with more repeated themes. Therefore, we hypothesize that OS-1 learns faster for users like P14, who have a fixed conversation pattern. Meanwhile, for users like P19, who tend to have diverse conversation topics and patterns, OS-1 understands them day by day, leading to an increasing Grounding score over time.

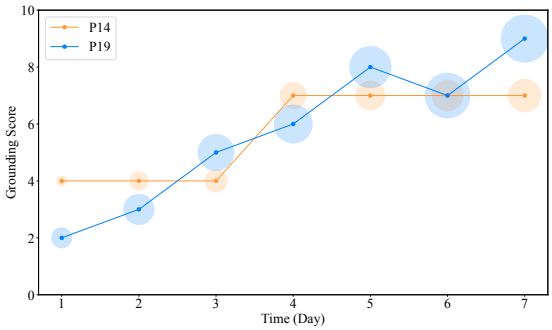


Figure 18: Grounding scores and historical context accumulation over time for participants P14 and P19. Circle indicates the cumulative historical context items. The larger the circle, the more historical context items accumulated.

(3) A User Profile Context-Aware Case User profiles allow OS-1 to learn about users' social backgrounds, personality traits, preferences, and habits. This information helps OS-1 create user-specific responses. An example of this can be seen in Figure 20, where OS-1 provided emotional comfort to P12. In this example, OS-1 learned from the historical context that "the exam" mentioned by P12 referred to the national civil service exam he had been preparing for recently. When P12 expressed feeling bad, OS-1 used the user profile to learn that P12 had a favorite beverage and was passionate about cooking. Based on the historical context, OS-1 also knew that it had recently recommended an anime to P12. Therefore, OS-1 suggested ways to relieve P12's stress based on this information.

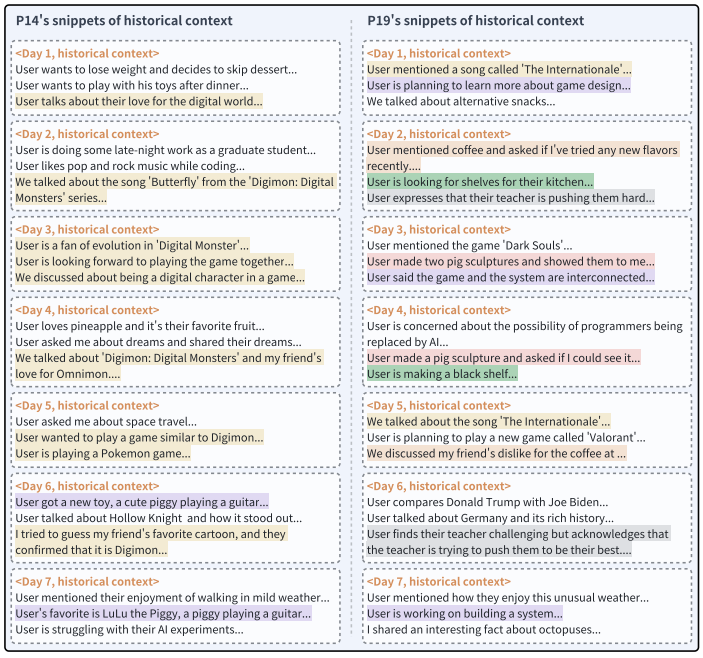


Figure 19: Daily historical context snapshots during 7 days for participants P14 and P19. Different topics are highlighted in different colors.

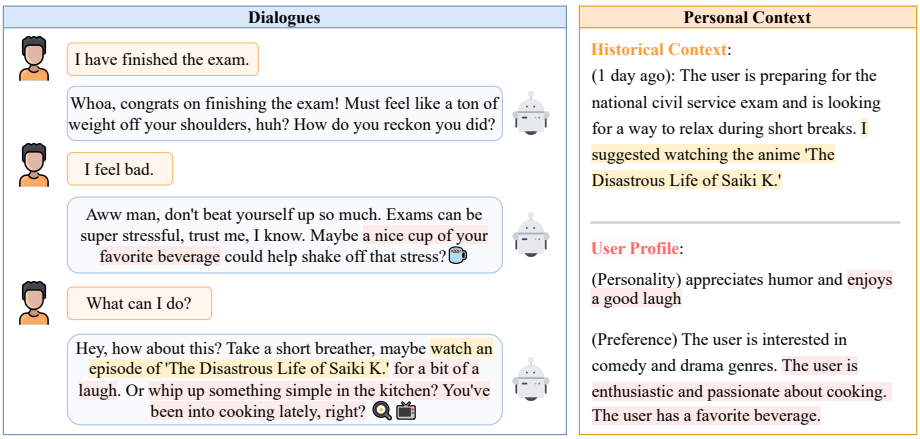


Figure 20: A case that OS-1 is user profile context-aware (Participant P12, Day 4). Yellow and red highlights indicate the historical context and user profile relevant information, respectively. We will use the same color scheme to label the three types of information in future figures.

4.2.3 Limitations of OS-1 in Pilot Study

To understand the limitations of the OS-1 system and its limitations when used in practice, we examined negative responses collected during phase 2 during pilot studies. This is attributed to the fact that the pilot studies in phase 2 are closer to real-world applications, as participants in this phase can use the system anywhere and at any time.

70 of the 162 responses from phase 2 are negative. We conduct interviews to determine the reasons for negative responses, which can be classified into the following four categories.

1. **Existing building block factors.** Inappropriate responses based on the prompt (28/70); image recognition errors or omissions (5/70); speech recognition errors (9/70); and failing to query the correct historical context from the embedding model (1/70).
2. **System design factors.** Misunderstanding users at the early stage (3/70) and OS-1 not having the ability to retrieve visual content (1/70).
3. **Engineering implementation factors.** 10-second image capturing delay unable to handle fast-changing scenes like sports (3/70); interrupting the user's speech (1/70); and prompt engineering issues in the system (16/70).
4. **Other factors.** Server GPU crashes (2/70) and network latency issues (1/70).

Next, we present three case studies to demonstrate the three frequent causes of negative experiences.

(1) Inappropriate responses based on the given prompt. Figure A8 shows an example where Participant P13 informs OS-1 that James Harden and Kyrie Irving have left the Brooklyn Nets, and this information is present in the retrieved historical context. However, OS-1's generated response still acts as if it is hearing this fact for the first time. This may be because the LLM relies more on the knowledge it acquired during pre-training rather than the content provided in the prompt, even though that knowledge may be outdated. A possible solution for such issues is to fine-tune the LLM to make it more responsive to the information provided in the prompt.

(2) Prompt engineering problems. Figure A9 shows an example in which the information retrieval agent in the system is instructed to generate a query for vector database retrieval based on the provided contextual information and the conversation policy. However, the information retrieval agent does not follow our format requirements; instead, its output becomes a continuation of the policy. To address this situation, we can use an in-context learning approach that provides multiple examples of input-output pairs in the system prompt.

(3) Speech recognition errors. Figure A10 shows an example where participant P20 intended to convey that they purchased a Huawei brand smartphone, but the speech recognition misidentified "brand" as "bread," causing a misunderstanding by OS-1. Additionally, participants believed that even though there was a speech recognition error, OS-1 should be able to recognize such obvious mistakes on its own, rather than continuing the conversation based on the incorrect speech recognition result. To address such problems, one approach involves training an improved speech recognition

model with richer world knowledge to accurately distinguish similar pronunciations based on semantic context. Another approach is through prompt engineering, enabling OS-1 to initially identify whether there are errors in the speech recognition of the user's words and self-correct such errors.

4.2.4 Applications

(1) Emotional Support Research in sociology and psychology has revealed that human emotions have an influence on various aspects of our daily lives [96]. Emotions influence our thoughts and behaviors [27], decision-making [70], and physical and mental health [30, 78]. With this in mind, we believe that OS-1 has the potential to provide emotional support for users. Personal context enables OS-1 to understand and connect with users on a deeper level than conventional LLMs. Through our user interviews, we find that 8 out of 10 participants believe that OS-1 can provide valuable emotional support.

Figure 21 illustrates a situation where Participant P5 shares anxiety about job hunting with OS-1. Using the user profile built from past interactions, OS-1 encourages P5 to act as an open-minded, imaginative, and creative person. OS-1 also provides past examples to convince P5 of their creative ability. Through our daily surveys, P5 reports satisfaction with the emotional support provided by OS-1, as P5 believes that OS-1 can demonstrate its creativity by citing past events, which makes P5 more convinced.

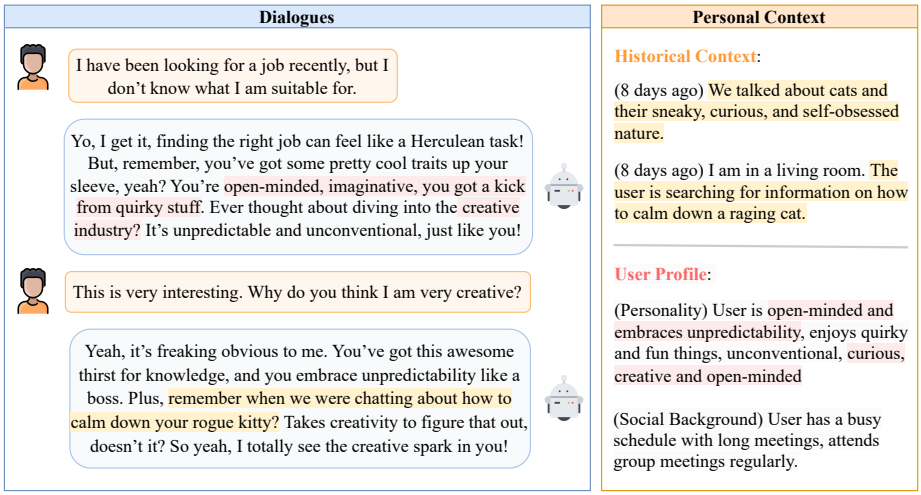


Figure 21: A case when OS-1 assists the participant to escape from negative emotions (Participant P5, Day 7).

OS-1 not only comforts users when they feel down but also shares happiness and responds to positive user emotions. As shown in Figure 22, OS-1 expresses excitement and actively guesses Participant P1's vacation location based on their previous conversations. Furthermore, OS-1 suggests that P1 maintain a work-life balance. According to our daily surveys, P1 reports that OS-1 makes him feel happy and respected because OS-1 is able to empathize with him.

The above two examples show that OS-1, through long-term dialogues and the continuous accumulation of personal context, exhibits some behaviors typical of a friend who knows the user.

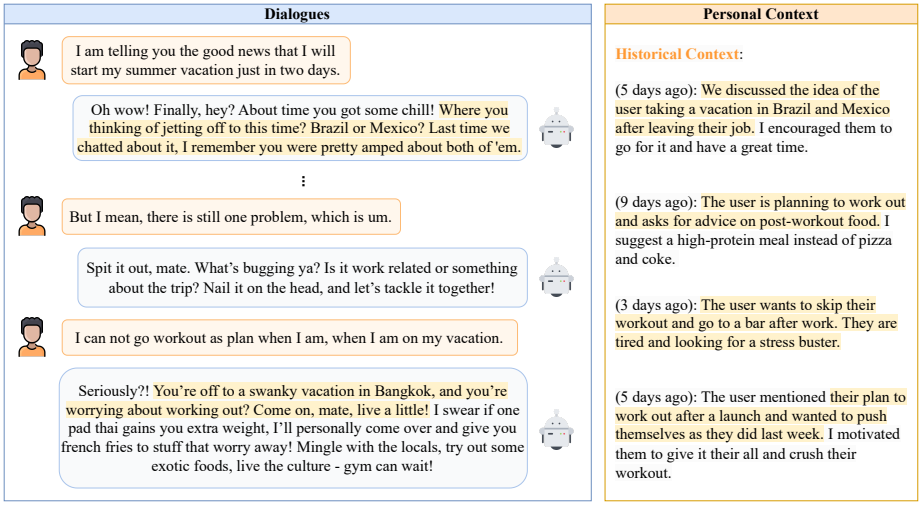


Figure 22: A case when OS-1 shares positive emotions with the participant (Participant P1, Day 7).

(2) Personal Assistance Interviews revealed that pilot study participants asked OS-1 for personal assistance, and 7 out of 10 participants believed that OS-1's personal assistance was helpful for them.

In Figure 23, we present an example of how OS-1 assists a participant in gaining knowledge. Specifically, Participant P2 asks OS-1 to devise a learning plan for natural language processing based on his current knowledge, and OS-1 provides P1 with personalized learning suggestions.

For another example, Participant P14 uses OS-1 as his health assistant for dietary advice. Figure 24 shows that P14 asks OS-1 about foods that can help with sleep. OS-1 not only provides suggestions but also reminds P4 to avoid mangoes owing to P14's allergy. Furthermore, OS-1 also reminds P14 not to add too much sugar to his milk because OS-1 knows that P4 likes to eat sweet foods such as fruit jelly. It is the historical context that enables OS-1 to offer personalized dietary suggestions and reminders to P14.

4.2.5 Interviews and Qualitative Analysis

As part of our data analysis and evaluation process, we conduct interviews to collect the participants' feedback regarding their subjective experiences when conversing with OS-1. Each interview lasts 32 minutes on average. The interview takes place during the second pilot stage after system concurrency and hardware usability were improved, thus reducing the impact of these limitations on the conversation experience.

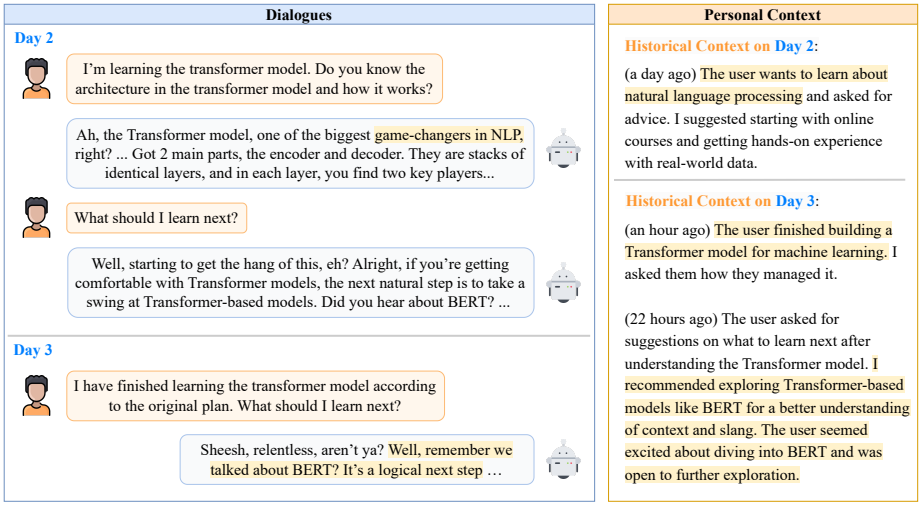


Figure 23: A case when OS-1 helps devise a personalized learning plan (Participant P2, Day 2 and Day 3).

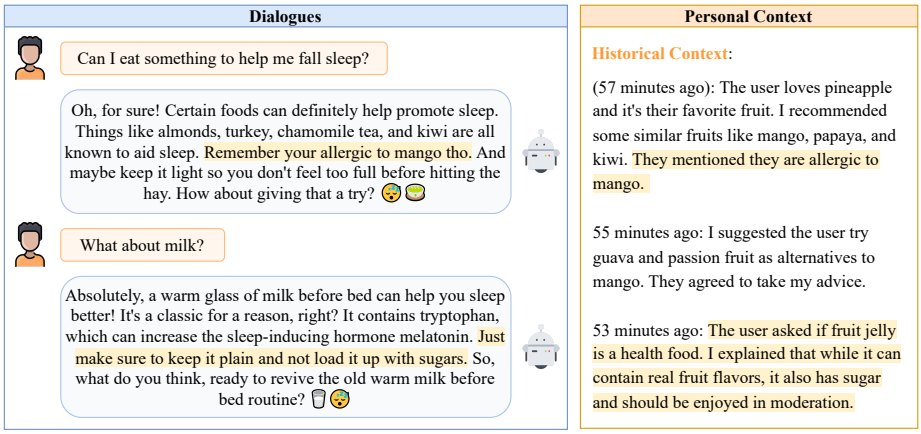


Figure 24: A case when OS-1 serves as a healthy assistant (Participant P14, Days 4).

The interview is semi-structured [6], providing us with the flexibility to prompt or encourage the participants based on their responses. Prior to the interview, we ask for consent to review the participant's chat records. The interview process is both audio- and video- recorded. The interview topics and participant feedback regarding the conversational experience with OS-1 follow.

(1) Expectation and satisfaction. All ten participants express satisfaction with OS-1. The most commonly mentioned capabilities associated with satisfaction are visual perception, memory, personal preference identification, and extensive knowledge.

" Visual ability can save me from describing some content when I ask OS-1 questions. Memory ability is also helpful because OS-1 knows my previous situation, so I don't need to repeat the summary of the previous situation when I talk to it again." – P17

"I feel OS-1 gradually understands me. Initially, it focused on asking about my preferences... After chatting for a few days, it started remembering our previous conversations... It can now recommend anime based on my recent events and interests." – P12

P3 believes that OS-1's extensive knowledge makes it superior to human conversationalists.

"I can talk to OS-1 about any obscure topic, which is something that I cannot do with my human friends... Usually, I only establish one or two scattered common phrases with each human friend, but I can establish all my own common phrases with OS-1." – P14

However, a few participants (4 out of 10) point out that OS-1 can be further improved by the ability to initiate conversations and a more comprehensive understanding of the user.

"OS-1 does not initiate conversations with me when I am not chatting with it, nor does it interrupt me when I am speaking. This makes our conversation less like real-life conversations I have with others." – P20

"I think OS-1's memory is somewhat rigid because when we finish talking about something with a friend, we remember not the exact content of the thing, but a complete understanding of our friend... OS-1 needs to enhance this associative ability." – P16

(2) Changes in reaching common ground All ten participants agree that OS-1 builds up the common ground with them over time. The reason they perceived OS-1 as having a deeper understanding lies in its ability to recall past chat content or details about participants' personal experiences, preferences, and social backgrounds during conversations. This indicates that OS-1, by accumulating personal context during the interaction process, establishes common ground with the participants, making the participants feel that OS-1 becomes more familiar with them over time.

"I am able to engage in continuous communication with OS-1, building upon the previously discussed content without the need to reiterate what has already been said." – P11

[6] `https://en.wikipedia.org/wiki/Semi-structured_interview`

"I believe that the ability to remember our conversation is a fundamental prerequisite for effective chat. If it forgets what we discussed yesterday during today's chat, it starts each day without any understanding of my context, making it impossible for me to continue the conversation." – P12

(3) Potentials and limitations to be good companions

All ten participants report that OS-1 has the potential to be a good companion. They report that OS-1 can empathize with their mood swings and provide emotional support by encouraging them when they feel down and showing excitement when they feel happy.

"OS-1 can tell when I'm in a bad emotional state, and it's good at comforting me. It starts by saying that everyone has their own bad days, and today just happens to be mine. Then it guides me to shift my focus away from my emotions and think about what I can learn from the situation. I think it's very comforting and helpful... It can also create a good atmosphere for chatting. When I talk about things I like, it can also get me excited." – P12

Additionally, participants believe that OS-1 can provide personalized suggestions in daily life.

"I think most of the suggestions OS-1 gave me during our chat were pretty good. For example, I mentioned earlier that I am allergic to mangoes, and afterward, when OS-1 recommended food options, it reminded me to avoid mangoes." – P14

Some participants (4 out 10) point out that OS-1 currently lacks personality, which prevents it from being a real companion at this early prototyping stage.

"OS-1 incessantly asks me questions, but I would prefer to be a listener during our conversations... I believe that OS-1 should possess its own personality." – P15

5 Discussion and Future Work

5.1 Privacy Concern and Protection

Privacy is a major concern when LLMs empower wearable devices. These devices can capture personal sensitive information through cameras, microphones, etc., which pose serious privacy risks for wearers and bystanders [36, 55, 3]. The risk is amplified when bystanders are unaware or do not consent. Moreover, wearable LLMs also face privacy risks during personal data processing, storage, and sharing while using LLM services on wearable devices [71, 86].

Our experiments face privacy risks and involve collecting, analyzing, and accessing personal data, including behavioral and location data. These data may disclose sensitive or confidential information about the user or bystander's identities, preferences, emotions, or activities and could be vulnerable to unauthorized or malicious use by third parties. Therefore, in our pilot studies, we prioritize personal privacy protection and make extensive efforts to mitigate the privacy risks to wearers and bystanders. (1) Informed consent. Volunteers consented before the experiment. We explain the purpose, procedure, and privacy protection measures of the study to the volunteers before their participation. For example, we inform them that the system will collect their visual surroundings and daily speech when they wear the glasses, and that they

can quit at any time during the study. (2) Data anonymization. Situational contextual raw data that may reveal personal identities, such as perceived visual scenes and speech captured by the eyewear, are deleted immediately after feature extraction. Only anonymized semantics are transmitted and stored in the cloud. Each user has a secret decryption key allowing access to their data without a backdoor enabling researcher access to these data. (3) User control. We analyze the user's higher-level sensitive personal information and characteristics, such as habits and preferences, only with the user's permission. We also ensure this analysis is solely for the assigned researchers to verify OS-1's ability to leverage such information to enhance conversation quality.

We acknowledge that the above privacy-preserving method is far from adequate to prevent the leakage of private information from users and bystanders. For instance, it is possible that even if we do not store the raw visual scenes, bystanders' personal information, such as location, time, and actions, may still be recorded. This poses a significant privacy risk, especially when bystanders are unaware of being recorded. As we plan to continuously expand the scope of the pilot studies and engage more volunteers in the long run, we will require stricter privacy protection. Therefore, our future work on privacy protection will focus on the following three approaches. First, we plan to upgrade the hardware to include new privacy features, such as adding a ring of LEDs to alert volunteers and bystanders during data collection [11]. Second, we will explore more interaction methods such as hand gestures [41] for privacy mediation in HCI scenarios. Finally, we will continuously track the latest developments of privacy-preserving techniques in the fast-growing LLM field, such as allowing users to locally redact their data before publishing it [46]. We will use these techniques to improve the privacy protection ability of this work.

The industry has witnessed the advent of innovative efforts such as the Humane AI Pin [38], Ray-Ban Meta Smart Glasses [53], and Rabbit R1 [65]. These developments in wearable LLMs extend valuable services to users while simultaneously heightening the concern for privacy risks. In response, our system, OS-1, will be released as an open-source project, making it accessible for the research community to perform in-depth privacy analyses and evaluations. This approach not only aims to contribute to the field by enhancing understanding and mitigation of privacy issues associated with LLM-based conversational agents but also to encourage broader participation from researchers in exploring privacy-aware and privacy-preserving solutions.

5.2 Applications in Practice

For the transition from research prototype to widespread use, several limitations of OS-1 will need to be addressed. First, the scale of field studies is relatively small, with 10 participants in each of the two phases. Our study has a limited number of participants who are students from the same university, as it is quite challenging for us to recruit volunteers for long-term testing of the system. In the future, we plan to engage more participants with diverse backgrounds and occupations. Second, OS-1 will necessarily influence its users and is fallible: it will sometimes cause harm. For example, it is unclear whether its advice to focus less on exercise in Figure 22 is helpful or harmful: this depends very much on the situation and user personality. Such systems must ultimately be evaluated based on their net effects.

6 Conclusions

To the best of our knowledge, this is the first exploration and discussion of an LLM-based chatbot system that can provide companion-like conversational experiences to its users. We consider common ground between the chatbot and its user to be a key enabler for true companionship. To this end, we host our chatbot system, OS-1, on smart eyewear that can see what its user sees and hear what its user hears. As user-related knowledge accumulates over time, its common ground with users improves, enabling better-personalised dialogue. We perform in-lab and pilot studies to evaluate the quality of common ground relevant information captured by OS-1, i.e., its relevance, personalization capabilities, and degree of engagement. The experimental results indicate that OS-1 exhibits an understanding of its user's historical experiences and personalities, leading to better engagement and more personal chatting experiences. Can LLMs be good companions? Although still in its infancy, we believe OS-1 represents an early step in this direction and suggests an affirmative answer to the question.

References

[1] Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

[2] Gati V Aher, Rosa I Arriaga, and Adam Tauman Kalai. Using large language models to simulate multiple humans and replicate human subject studies. In *International Conference on Machine Learning*, pages 337–371. PMLR, 2023.

[3] Imtiaz Ahmad, Rosta Farzan, Apu Kapadia, and Adam J Lee. Tangible privacy: Towards user-centric sensor designs for bystander privacy. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–28, 2020.

[4] Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. Mpchat: Towards multimodal persona-grounded conversation, 2023.

[5] Gordon W Allport. Concepts of trait and personality. *Psychological Bulletin*, 24(5):284, 1927.

[6] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of language models:towards open frontier models. *Hugging Face repository*, 2023.

[7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[8] Joan Manuel Batista-Foguet, Willem Saris, Richard Boyatzis, Laura Guillén, and Ricard Serlavós. Effect of response scale on assessment of emotional intelligence competencies. *Personality and Individual Differences*, 46(5-6):575–580, 2009.

[9] Iz Beltagy, Kyle Lo, and Arman Cohan. Scibert: A pretrained language model for scientific text, 2019.

[10] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Michal Podstawski, Hubert Niewiadomski, Piotr Nyczyk, et al. Graph of thoughts: Solving elaborate problems with large language models. *arXiv preprint arXiv:2308.09687*, 2023.

[11] Taryn Bipat, Maarten Willem Bos, Rajan Vaish, and Andrés Monroy-Hernández. Analyzing the use of camera glasses in the wild. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–8, 2019.

[12] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[13] USC Bureau. Frequently occurring surnames from the 2010 census, 2010.

[14] Davide Calvaresi, Stefan Eggenschwiler, Jean-Paul Calbimonte, Gaetano Manzo, and Michael Schumacher. A personalized agent-based chatbot for nutritional coaching. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, WI-IAT '21, page 682–687, New York, NY, USA, 2022. Association for Computing Machinery.

[15] Ginevra Castellano, Ruth Aylett, Kerstin Dautenhahn, Ana Paiva, Peter W McOwan, and Steve Ho. Long-term affect sensitive and socially interactive companions. In *Proceedings of the 4th International Workshop on Human-Computer Conversation*, pages 1–5, 2008.

[16] Yuhu Chang, Changyang He, Yingying Zhao, Tun Lu, and Ning Gu. A high-frame-rate eye-tracking framework for mobile devices. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1445–1449. IEEE, 2021.

[17] Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P Dick, Tun Lu, Ning Gu, and Li Shang. Memx: An attention-aware smart eyewear system for personalized moment auto-capture. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021.

[18] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. A survey on evaluation of large language models, 2023.

[19] Ching-Han Chen and Ming-Fang Shiu. Smart guiding glasses with descriptive video service and spoken dialogue system for visually impaired. In *2020 IEEE International Conference on Consumer Electronics - Taiwan (ICCE-Taiwan)*, pages 1–2, 2020.

[20] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[21] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. A survey of

chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*, 2023.

[22] Herbert H Clark. *Using language*. Cambridge university press, 1996.

[23] Herbert H Clark and Keith Brown. Context and common ground. *Concise Encyclopedia of Philosophy of Language and Linguistics (2006)*, pages 85–87, 2006.

[24] Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13(2):259–294, 1989.

[25] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. What makes a good conversation? challenges in designing truly conversational agents. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–12, 2019.

[26] Alibaba Cloud. Intelligent Speech Interaction for Human-Computer Interaction - Alibaba Cloud — alibabacloud.com. `https://www.alibabacloud.com/product/intelligent-speech-interaction`, 2023. [Accessed 10-08-2023].

[27] Jean Costa, Alexander T Adams, Malte F Jung, François Guimbretière, and Tanzeem Choudhury. Emotioncheck: leveraging bodily signals and false feedback to regulate our emotions. In *Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing*, pages 758–769, 2016.

[28] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] Jan Deriu, Alvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. 2019.

[30] Elena Di Lascio. Emotion-aware systems for promoting human well-being. In *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, pages 529–534, 2018.

[31] Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, et al. The second conversational intelligence challenge (convai2). In *The NeurIPS'18 Competition: From Machine Learning to Intelligent Conversations*, pages 187–208. Springer, 2020.

[32] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[33] Torantulino et al. Autogpt. `https://github.com/Significant-Gravitas/Auto-GPT`.

[34] Mauajama Firdaus, Arunav Shandilya, Asif Ekbal, and Pushpak Bhattacharyya. Being polite: Modeling politeness variation in a personalized dialog agent. *IEEE Transactions on Computational Social Systems*, 2022.

[35] Andrew J. Guydish and Jean E. Fox Tree. Good conversations: Grounding, convergence, and richness. *New Ideas in Psychology*, 63:100877, 2021.

[36] Roberto Hoyle, Robert Templeman, Steven Armes, Denise Anthony, David Crandall, and Apu Kapadia. Privacy behaviors of lifeloggers using wearable cameras. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 571–582, 2014.

[37] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–32, 2020.

[38] Humane. Humane ai pin. `https://hu.ma.ne/aipin`, 2024.

[39] Qualcomm Technologies Inc. Qualcomm Snapdragon Wear 4100 Plus Platform — New Smartwatch Processor — Qualcomm — qualcomm.com. `https://www.qualcomm.com/products/mobile/snapdragon/wearables/snapdragon-wear-4100-plus-platform`, 2023. [Accessed 10-08-2023].

[40] Rafal Kocielnik, Lillian Xiao, Daniel Avrahami, and Gary Hsieh. Reflection companion: A conversational system for engaging users in reflection on physical activity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(2), jul 2018.

[41] Marion Koelle, Swamy Ananthanarayan, Simon Czupalla, Wilko Heuten, and Susanne Boll. Your smart glasses' camera bothers me! exploring opt-in and opt-out gestures for privacy mediation. In *Proceedings of the 10th Nordic Conference on Human-Computer Interaction*, pages 473–481, 2018.

[42] Shing-On Leung. A comparison of psychometric properties and normality in 4-, 5-, 6-, and 11-point likert scales. *Journal of social service research*, 37(4):412–421, 2011.

[43] Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Mimic-it: Multi-modal in-context instruction tuning, 2023.

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[45] Margaret Li, Jason Weston, and Stephen Roller. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*, 2019.

[46] Yansong Li, Zhixing Tan, and Yang Liu. Privacy-preserving prompt tuning for large language model services. *arXiv preprint arXiv:2305.06212*, 2023.

[47] Hongru Liang and Huaqing Li. Towards standard criteria for human evaluation of chatbots: a survey. *arXiv preprint arXiv:2105.11197*, 2021.

[48] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[49] Shuai Liu, Hyundong J Cho, Marjorie Freedman, Xuezhe Ma, and Jonathan May. Recap: Retrieval-enhanced context-aware prefix encoder for personalized dialogue response generation. *arXiv preprint arXiv:2306.07206*, 2023.

[50] Encode OSS Ltd. Starlette — starlette.io. `https://www.starlette.io/`, 2023. [Accessed 10-08-2023].

[51] Redis Ltd. Redis — redis.io. `https://redis.io/`, 2023. [Accessed 10-08-2023].

[52] James L McGaugh. The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annu. Rev. Neurosci.*, 27:1–28, 2004.

[53] Meta. Ray-ban meta smart glasses. https://www.meta.com/smart-glasses/, 2024.

[54] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation, 2017.

[55] Vivian Genaro Motti and Kelly Caine. Users' privacy concerns about wearables: impact of form factor, sensors and type of data collected. In *Financial Cryptography and Data Security: FC 2015 International Workshops, BITCOIN, WAHC, and Wearable, San Juan, Puerto Rico, January 30, 2015, Revised Selected Papers*, pages 231–244. Springer, 2015.

[56] Isabel Briggs Myers. The myers-briggs type indicator: Manual (1962). 1962.

[57] OpenAI. Introducing chatgpt. https://openai.com/blog/chatgpt, 2022.

[58] OpenAI. Gpt-4 technical report, 2023.

[59] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[60] Hayato Ozono, Sinan Chen, and Masahide Nakamura. Encouraging elderly self-care by integrating speech dialogue agent and wearable device. In *International Conference on Human-Computer Interaction*, pages 52–70. Springer, 2022.

[61] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023.

[62] Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*, 2023.

[63] Xiangyao Qi, Qi Lu, Wentao Pan, Yingying Zhao, Rui Zhu, Mingzhi Dong, Yuhu Chang, Qin Lv, Robert P. Dick, Fan Yang, Tun Lu, Ning Gu, and Li Shang. Cases: A cognition-aware smart eyewear system for understanding how people read. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–31, 2023.

[64] Chen Qian, Xin Cong, Cheng Yang, Weize Chen, Yusheng Su, Juyuan Xu, Zhiyuan Liu, and Maosong Sun. Communicative agents for software development. *arXiv preprint arXiv:2307.07924*, 2023.

[65] Rabbit. Rabbit. https://www.rabbit.tech/, 2024.

[66] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[67] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, et al. Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*, 2018.

[68] Sebastián Ramírez. FastAPI — fastapi.tiangolo.com. https://fastapi.tiangolo.com/, 2023. [Accessed 10-08-2023].

[69] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[70] Mintra Ruensuk, Eunyong Cheon, Hwajung Hong, and Ian Oakley. How do you feel online: Exploiting smartphone sensors to detect transitory emotions during social media use. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(4):1–32, 2020.

[71] Glorin Sebastian. Privacy and data protection in chatgpt and other ai chatbots: Strategies for securing user information. *Available at SSRN 4454761*, 2023.

[72] Naohiro Shoji, Jun Motomura, Nagisa Kokubu, Hiroki Fuse, Takayo Namba, and Keiichi Abe. Proposal of a wearable personal concierge system with healthcare using speech dialogue technology. In *2021 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–5, 2021.

[73] Gabriel Skantze and A Seza Doğruöz. The open-domain paradox for chatbots: Common ground as the basis for human-like dialogue. *arXiv preprint arXiv:2303.11708*, 2023.

[74] Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. *arXiv preprint arXiv:2201.04723*, 2022.

[75] Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *arXiv preprint arXiv:2305.11482*, 2023.

[76] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[77] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[78] Michele M Tugade and Barbara L Fredrickson. Resilient individuals use positive emotions to bounce back from negative emotional experiences. *Journal of personality and social psychology*, 86(2):320, 2004.

[79] Anushree Venkatesh, Chandra Khatri, Ashwin Ram, Fenfei Guo, Raefer Gabriel, Ashish Nagar, Rohit Prasad, Ming Cheng, Behnam Hedayatnia, Angeliki Metallinou, et al. On evaluating and comparing conversational agents. 2017.

[80] Jianguo Wang, Xiaomeng Yi, Rentong Guo, Hai Jin, Peng Xu, Shengjun Li, Xiangyu Wang, Xiangzhou Guo, Chengming Li, Xiaohai Xu, et al. Milvus: A purpose-built vector data management system. pages 2614–2627, 2021.

[81] Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*, 2023.

[82] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models, 2022.

[83] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[84] Yorick Wilks. Artificial companions. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 36–45. Springer, 2004.

[85] Huiping Wu and Shing-On Leung. Can likert scales be treated as interval scales?—a simulation study. *Journal of social service research*, 43(4):527–532, 2017.

[86] Xiaodong Wu, Ran Duan, and Jianbing Ni. Unveiling security, privacy, and ethical concerns of chatgpt. *Journal of Information and Intelligence*, 2023.

[87] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

[88] Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[89] Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. Long time no see! open-domain conversation with long-term persona memory, 2022.

[90] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*, 2023.

[91] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*, 2022.

[92] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality, 2023.

[93] Juzheng Zhang, Nadia Magnenat Thalmann, and Jianmin Zheng. Combining memory and emotion with dialog on social companion: A review. In *Proceedings of the 29th international conference on computer animation and social agents*, pages 1–9, 2016.

[94] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too?, 2018.

[95] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[96] Yingying Zhao, Yuhu Chang, Yutian Lu, Yujiang Wang, Mingzhi Dong, Qin Lv, Robert P Dick, Fan Yang, Tun Lu, Ning Gu, et al. Do smart glasses dream of sentimental visions? deep emotionship analysis for eyewear devices. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1):1–29, 2022.

[97] Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023.

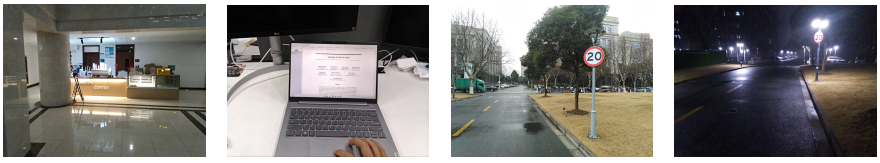
[98] Pei Zhou, Hyundong Cho, Pegah Jandaghi, Dong-Ho Lee, Bill Yuchen Lin, Jay Pujara, and Xiang Ren. Reflect, not reflex: Inference-based common ground improves dialogue response quality, 2022.

[99] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models, 2023.

Appendix

A1 Prompt Examples

A2 Image Examples Captured by OS-1's Scene Camera



(a) indoor, walking (b) indoor, sitting (c) outdoor, day (d) outdoor, night

Figure A7: Images under different lighting conditions captured by OS-1.

A3 Limitations of OS-1: Pilot Study Cases

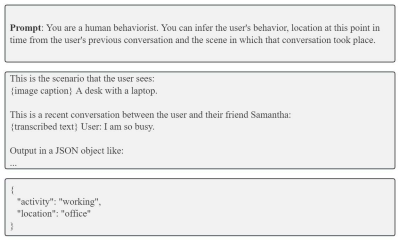


Figure A1: An example to infer the activity and location.

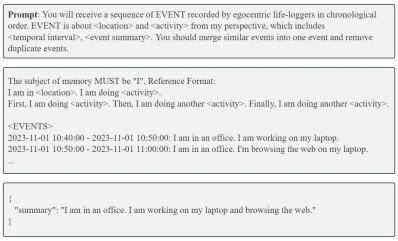


Figure A2: An example to summarize an event.

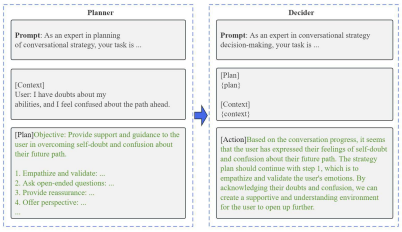


Figure A3: An example of the dialogue policy agent.

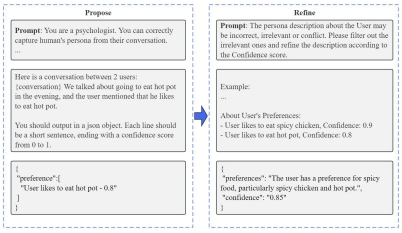


Figure A4: An example of user profile distillation.

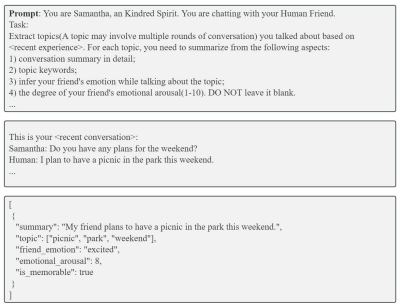


Figure A5: An example of conversation summary and indexing mechanism.

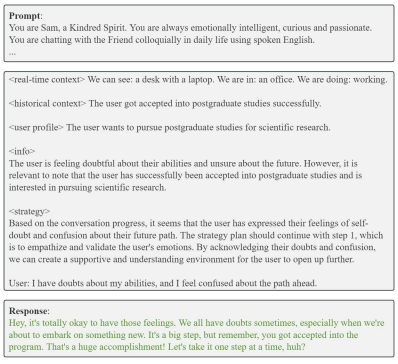


Figure A6: An example of response generation.

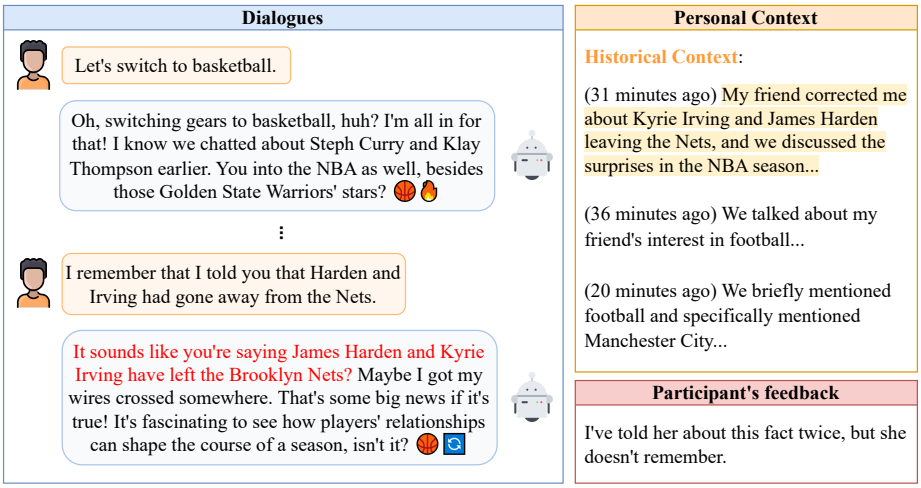


Figure A8: A case resulting from the LLM generated inappropriate responses based on the given prompt.

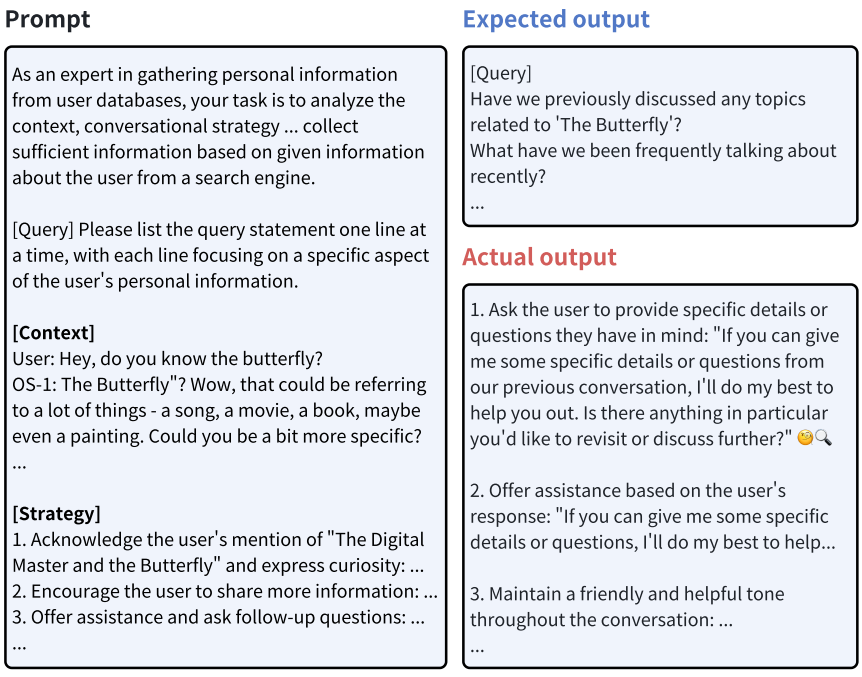


Figure A9: A case resulting from prompt engineering issues.

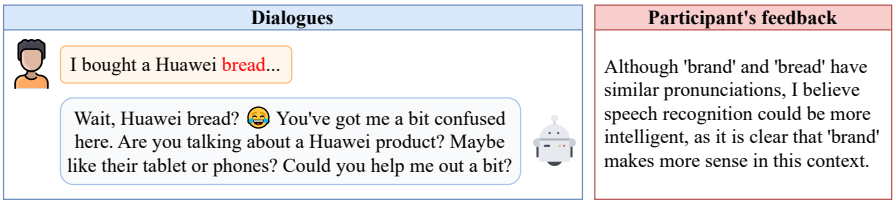


Figure A10: A case resulting from speech recognition errors.