Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation

Wazeer Zulfikar
MIT Media Lab
Cambridge, USA
wazeer@mit.edu

Samantha Chan
MIT Media Lab
Cambridge, USA
swtchan@media.mit.edu

Pattie Maes
MIT Media Lab
Cambridge, USA
pattie@media.mit.edu

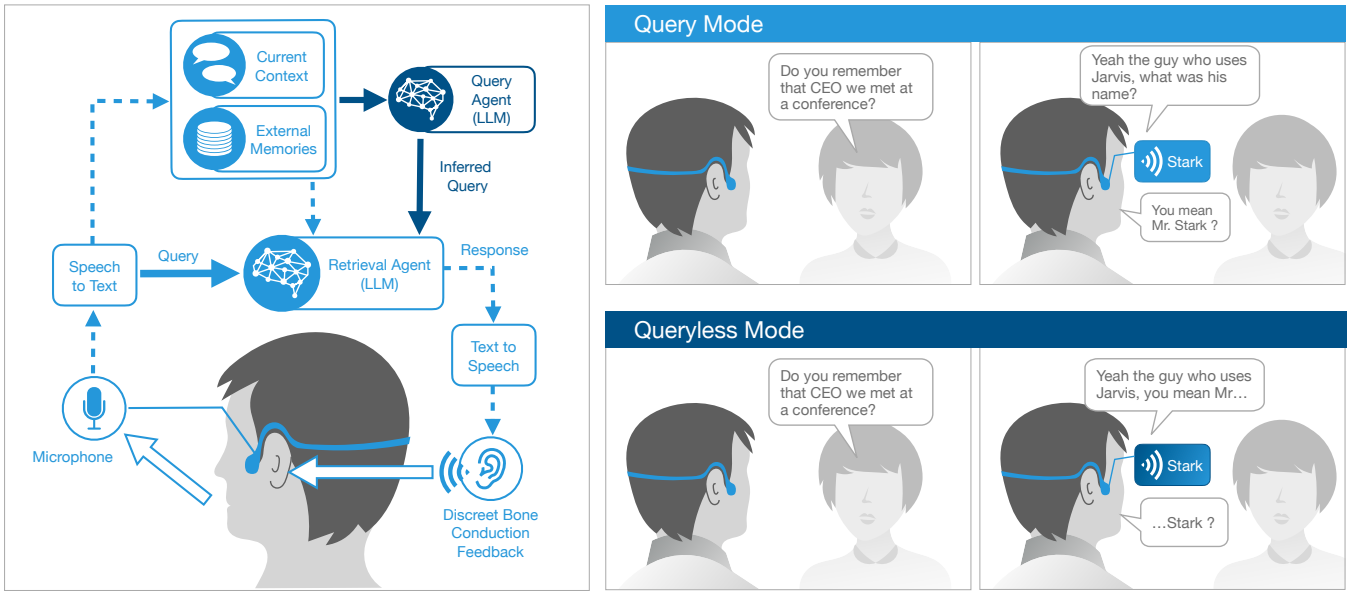

Figure 1: Architecture of Memoro and its two interaction modes. (Left) System architecture of the memory assistant. (Right) Two interaction modes: (1) *Query Mode* where the user can ask contextual questions (2) *Queryless Mode* where the user can request predictive assistance and skip query formation. In both modes, responses are discreetly played back to the user using a bone conduction headset.

ABSTRACT

People have to remember an ever-expanding volume of information. Wearables that use information capture and retrieval for memory augmentation can help but can be disruptive and cumbersome in real-world tasks, such as in social settings. To address this, we developed Memoro, a wearable audio-based memory assistant with a concise user interface. Memoro uses a large language model (LLM) to infer the user's memory needs in a conversational context, semantically search memories, and present minimal suggestions. The assistant has two interaction modes: Query Mode for voicing queries and Queryless Mode for on-demand predictive assistance, without explicit query. Our study of (N=20) participants engaged in a real-time conversation, demonstrated that using Memoro reduced device interaction time and increased recall confidence while preserving conversational quality. We report quantitative results and discuss the preferences and experiences of users. This work contributes towards utilizing LLMs to design wearable memory augmentation systems that are minimally disruptive.

CCS CONCEPTS

• **Human-centered computing** → **Interaction techniques**; **Natural language interfaces**; *Empirical studies in HCI*.

KEYWORDS

memory assistant, large language models, voice interfaces, context-aware agent, minimal interfaces

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642450

ACM Reference Format:
Wazeer Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642450

1 INTRODUCTION

Memory plays an essential role in people's lives, whether in communication, learning, decision-making, or maintaining relationships [4, 43]. However, memory is imperfect and error-prone due to factors such as lack of sleep, stress, and divided attention [55, 62]. Furthermore, neurological disorders related to memory loss, such as dementia, are rising as populations in many parts of the world grow older [52].

Memory augmentation and information retrieval systems have been of key interest to the HCI community over the past several decades as tools to address these growing challenges. Since Vannevar Bush's conception of the Memex in 1945 [12], there has been extensive work on systems and devices to extend our memory [17, 18, 39, 60] such as lifelogging systems that continuously record the user's media and signals [34, 44, 69], and just-in-time information retrieval systems [7, 22, 36, 45, 60] that provide relevant information based on the user's context. While these wearable systems demonstrate the capabilities of users to retrieve vast amounts of information, limited research exists on designing interfaces that enable the retrieval of information in a minimally disruptive way when the user is already engaged in a primary task, which is often the case with wearables.

We define minimal disruption for a memory augmentation interface as (1) requiring minimal input from the user to request information, i.e., the input the user gives is short, and (2) providing minimal output, namely the suggestion or response provided by the augmentation system is the smallest amount of information that will give the user the information they need. The minimal disruption design consideration is critical for the usability of wearable memory augmentation systems [23], especially in social settings that are attention-demanding and where incidentally the highest number of memory lapses occur [51], such as conversations.

Therefore, an important challenge for the design of wearable memory augmentation systems is that of a seamless, user-friendly, and concise search interface [23] to keep disruption to the user's primary task minimal. Incorporating context awareness can reduce or, as we show in this paper, even completely eliminate the query input, allowing users to skip posing an explicit, comprehensive retrieval query, as the system can directly infer the user's specific memory needs. Recent developments in large language models (LLMs) have improved capabilities in understanding conversational context in natural settings [11, 70] and enable more flexible search queries using alternative phrases [42]. They also enable the shortening of answers [24] for succinct suggestions. This highlights the opportunity to leverage LLMs to design easy-to-use and minimally disruptive interfaces.

In this paper, we aim to answer the following research questions

- **RQ1.** How can we design a seamless wearable memory assistant using LLMs to reduce disruption to the primary task with minimal and effective input and output?
- **RQ2.** What are the effects of using the memory augmentation system during the primary task of a real-time conversation across metrics such as quality of conversation, performance, and task load?
- **RQ3.** How do context awareness and conciseness affect the system's usability, user perception and experience?

We developed a minimally disruptive audio-based wearable assistant, Memoro, that uses LLMs to aid the user in retrieving relevant information from previously recorded personal data through concise suggestions. Memoro continuously transcribes and encodes audio data from conversations the user engages in. The memory assistant has two modes of interaction for retrieval: Query Mode, where the user voices their natural language query, and Queryless Mode where the user is presented with a suggestion relevant to the current conversational context without having to explicitly query the system. Both modes provide minimal memory responses to the user (see Figure 1). In terms of hardware form factor, Memoro uses a light-weight, bone-conduction headset for unobstructed and private responses.

To study the use of Memoro and its two query modes in the context of a real-time conversation, we conducted a study with N=20 participants. We found that the use of Memoro increased their recall confidence while preserving conversational quality. We also conducted a technical evaluation to measure the conciseness of input and output and the accuracy of the system responses. Most participants (15 of 20) expressed a preference for Memoro over no system and baseline (system without context awareness and conciseness), with 10 participants favoring the Queryless Mode. Participants elaborated upon their preferences and reservations, allowing for future design considerations. The highest-rated condition, Query Mode, achieved a mean usability score of 80.0, which falls between the good and excellent range [5] and was significantly improved due to contextual awareness and conciseness as compared to the baseline. The goal of this paper is not to present a full-fledged memory augmentation system, but rather to evaluate whether LLMs can be used to make memory augmentation systems that are less disruptive.

In summary, the contributions of this paper are threefold:

(1) Design of a wearable memory assistant, called Memoro, focusing on minimal disruption to the user's primary real-world task by using conversational context and conciseness.
(2) Exploration of a query-less approach to eliminate query time and thereby increase seamless memory assistance by inferring the user's memory need.
(3) A within-subject user study showing that the proposed system has good usability and low interruption in a social task while preserving conversation quality and decreasing task load as compared to no system.

2 RELATED WORK

Our work is related to, and inspired by past work on wearable memory augmentation systems, context-aware agents in conversations, and large language models in virtual assistants.

2.1 Wearable Memory Augmentation Systems

Wearable memory augmentation has been a well-researched area since the 1990s when Mik Lamming coined the term "memory prosthesis" [41]. Since then, there have been various forms of memory augmentation systems, including reminder systems and lifelogging systems [14, 28, 29, 34, 41, 44, 59, 69]. Lifelogging devices continuously capture signals such as audio, video, and biosignals resulting