EmBARDiment: an Embodied AI Agent for Productivity in XR

Riccardo Bovo
Google, USA
Imperial College London, UK

Steven Abreu
Google, USA
University of Groningen, Netherlands

Karan Ahuja
Google, USA
Northwestern University, USA

Eric J Gonzalez
Google, USA

Li-Te Cheng
Google, USA

Mar Gonzalez-Franco
Google, USA

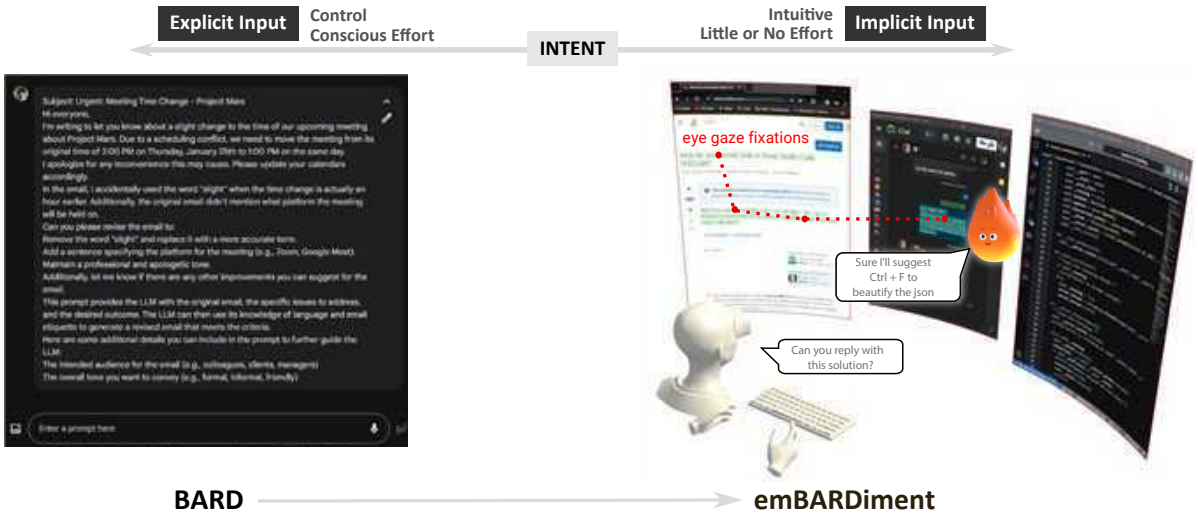

Figure 1: EmBARDiment introduces an attention framework that derives context implicitly from user eye-gaze, and contextual memory within the XR environment. It uses the implicitly information bundled with explicit verbal inputs to elicit grounded communication between the User and the AI Agent.

Abstract

XR devices running chat-bots powered by Large Language Models (LLMs) have tremendous potential as always-on agents that can enable much better productivity scenarios. However, screen based chat-bots do not take advantage of the the full-suite of natural inputs available in XR, including inward facing sensor data, instead they over-rely on explicit voice or text prompts, sometimes paired with multi-modal data dropped as part of the query. We propose a solution that leverages an attention framework that derives context implicitly from user actions, eye-gaze, and contextual memory within the XR environment. This minimizes the need for engineered explicit prompts, fostering grounded and intuitive interactions that glean user insights for the chat-bot. Our user studies demonstrate the imminent feasibility and transformative potential of our approach to streamline user interaction in XR with chat-bots, while offering insights for the design of future XR-embodied LLM agents.

Index Terms: AI Agents, Chatbots, XR productivity, Multi-window, AI input.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has revolutionized human-computer interaction, with chat-bots such as ChatGPT [41], Claude [5], and BARD [20] emerging as the primary interface for engaging with these powerful AI systems. However, as LLMs continue to evolve, their potential extends beyond text-based interactions, particularly in the realm of extended reality (XR) environments. The integration of LLMs as embodied, always-on immersive agents within augmented and virtual reality headsets holds immense promise for transforming user experiences and enabling seamless, context-aware assistance.Despite the increasing sophistication of LLMs, current chat-bot implementations heavily rely on explicit voice or text prompts, often demanding multiple iterations to refine the desired tone and context for the optimal output, leading to lengthy interactions. In XR environments, this approach fails to capitalize on the diverse range of natural inputs available, as contemporary XR headsets (such as Meta Quest Pro [36] and Apple Vision Pro [6]) are equipped with inward-facing sensors capable of capturing rich user data. Although recent advancements in LLM architectures, such as Gemini 1.5 [21], aim to support up to 10M tokens of context attached to a prompt, the current explicit input modalities in XR devices remain limited. For instance, speech input has an estimated universal throughput of only 39 bits per second [14], while text input in XR technologies remains cumbersome and inefficient [26]. Our approach is particularly useful for XR environments due to their large field of view (FoV) displays, which can support multiple windows for productivity tasks [7, 42]. While it is possible to provide all the data available on these windows to LLMs as context, this can result in a large amount of data, making the system less responsive, computationally heavy, and challenging to maintain nuanced conversations, as not all information may be relevant [31, 29]. Consequently, directly translating *explicit*

prompt-based chat-bots to embodied agents in XR is likely to be sub-optimal. To address these challenges, we propose EmBARDiment, a novel approach that leverages an *implicit* attention framework in combination with a contextual memory to enable embodied LLM agents in XR environments (see Fig 2). Our solution aims to implicitly derive context from the user's actions, eye-gaze, and visual saliency within the XR environment, minimizing the reliance on engineered explicit prompts. As a result, we facilitate highly targeted and intuitive interactions, allowing the agent to infer the user's intentions and needs based on their behavior and focus of attention, both present and past. For example, if a user is reading a document, the agent can access the content the user has read and utilize it as context for generating relevant responses, alongside the query prompt. Unlike prior works on gaze in XR that focus primarily on one-off visual querying [47, 37], our approach combines gaze-saliency context with LLM-powered continued interaction, offering the best of both worlds. We hypothesize that this approach will enhance user interactions with the embodied agent, as it establishes a shared "theory of mind" and provides highly contextual assistance, potentially simplifying interactions to the level of natural language commands like "put that there" [8].

The primary contributions of this work are as follows:

- We propose a novel attention framework leveraging gaze-based saliency driven contextual memory for embodied LLM agents in XR, thus enabling implicit user cues for context-aware assistance.
- We conduct user studies to empirically evaluate the effectiveness of our proposed contextual framework in enhancing user interactions and experience, and showcase its efficacy over baseline explicit text-based input condition.

2 Related Work

2.1 Context-Aware Assistants for Productivity in XR

The growing interest in XR multi-window environments for productivity [7, 42] presents new opportunities and challenges for integrating context-aware assistants. These environments offer users cost-effective, customizable, and private workspaces, even on the go [18, 40]. Moreover, XR multi-window environments can support remote work by mitigating issues like distractions, insufficient workspaces, and the struggle to separate work from personal life [18], maybe even providing better multitasking solutions, as we can easily maintain multiple windows open at the same time [42], even on mobile office work settings [40]. The development of context-aware assistants [16] has seen a plethora of solutions, gleaning insights about the user (e.g. activity [2], pose [48], preferences [15], task at hand [17]) and immediate environment (e.g. location [35], ambient light [10]). In the domain of XR, context-aware assistants have been explored to enhance education [49], industrial training and maintenance [25] and enable more natural interactions [27]. For instance, in the domain of immersive user experience, [27] proposed a framework that utilises the metadata found in the field of view to help ground the conversation between human and AI agent [27]. While context-aware assistants in XR have been explored for various applications, there is a notable gap in research concerning their application for knowledge workers in XR environments, despite the recognized potential of XR for enhancing productivity work [7, 42, 18]. Productivity within XR presents unique challenges that render traditional AI agent interfaces, such as chat-bots, inefficient, primarily due to the cumbersome and inefficient nature of typing in XR [26]. Additionally, alternative channels like speech, because of their intrinsic limitations [14], are not deemed suitable to generate the complex prompt required for knowledge workers [38]. Furthermore, the information accessible across multiple windows in XR environments, perhaps with even more screens open that in a regular PC, can pose difficulties for assistants in delivering targeted and efficient productivity support [31]. EmBARDiment tackles these challenges by employing multimodal interactions and develops a gaze-driven contextual memory. This approach allows embodied AI agents to implicitly extract relevant context from the user's focus of attention within the XR environment, thereby facilitating more productive interactions.

2.2 Gaze attention driven Multimodal XR Interactions

Prior work in HCI leverages inputs such as gaze and pointing gestures concurrently with speech commands to generate multi-modal interactions [43, 19]. [8]'s foundational study "Put-That-There" examined the fusion of vocal instructions and manual gestures to improve engagement with graphical user interfaces [8]. Likewise, [37] investigated the combination of eye tracking and verbal communication, particularly for engaging with small, densely arranged on-screen elements. Recent studies have shifted focus towards comprehending visual context and leveraging intuitive communication to improve interactions. Specifically, [34] et al. conducted a study that incorporated GPS location and user's head gaze into mobile voice assistants, enhancing their comprehension of nearby shops and buildings by understanding a user's location and visual direction, so that users can perform implicit verbal requests about their surroundings [34]. [4] et al. made use of speech as a direction communication channel akin to user-gaze to facilitate interactions across smart environment ecosystems [4]. [47] et al. presented *Nimble*, a mobile interface that combines visual question-answering frameworks with gesture recognition, focusing on directive gestures, to enhance user engagement [47]. [33] et al. exploited the implicit attention cues of gaze in *MiseUnseen* to facilitate novel interactions inside XR environments, including creating or arranging objects covertly at runtime [33]. These works highlight the effectiveness of using visual attention or pointing gestures to aid in understanding the user's intended referent during verbal requests. However, these works primarily focus on providing directional information at the moment of the request. In contrast, EmBARDiment extends these techniques by implementing a contextual memory that encodes information about temporal gaze-based saliency, allowing the assistant to leverage this context up to the moment of the verbal request, thus enabling a more contextually relevant and efficient interaction paradigm.

2.3 From Chat-bots to Embodied XR Agents

Chat-bots have come a long way since their early days of rudimentary pattern matching, now incorporating sophisticated Natural Language Processing (NLP) techniques for nuanced text and voice interactions [3]. Advancements in NLP have enabled these systems to comprehend complex inputs and generate contextually relevant responses [11]. This evolution has transformed user interactions with AI, shifting from typing to voice-based interfaces [32, 45, 50], reflecting the importance of creating communication that mirrors human-human interactions [44, 13]. The recent debut of powerful LLMs like ChatGPT, BARD, and Gemini has sparked renewed interest in AI-driven chat-bot applications [11], leading to increased user engagement and scientific research [12]. This renewed interest has also motivated the integration of AI chat-bots and agents into XR applications, in the form of embodied agents [28, 39]. These embodied agents aim to further enhance user engagement and natural interaction through visual and behavioral cues. EmBARDiment builds upon this renewed interest in embodied XR agents, and introduces an attention based framework that dynamically adjusts to the user's current activity and context within the XR environment.

3 EmBARDiment

EmBARDiment is a practical example of our framework that is implemented as an XR application that seamlessly integrates speech-

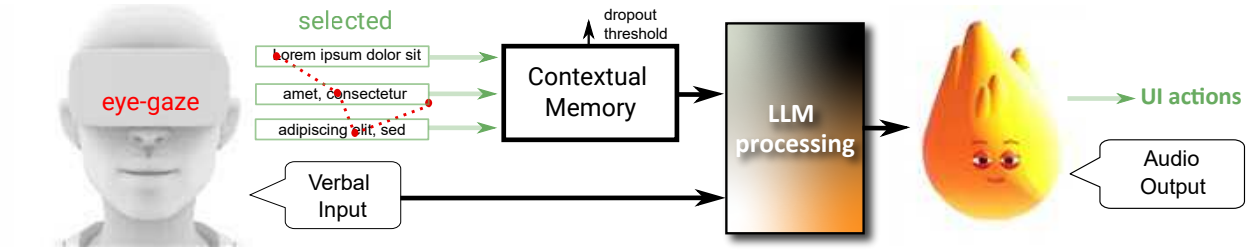

Figure 2: Schema of EmBARDiment, explaining how the attention frameworks uses implicit eye-gaze to select contextual information and bundles it with explicit verbal inputs. This elicits grounded communication between the User and the AI Agent.

to-text, text-to-speech, gaze-driven saliency and LLMs to enable a voice-interactive AI agent within a multi-window XR environment. In this section, we describe the system architecture (Fig 2), its key components and their integration. The application was developed using Unity [1] and deployed on the Oculus Quest Pro [36]. The code is available at the following GitHub repository: https://github.com/anonymous-user/anonymous-repo.

3.1 Embodied AI Agent

EmBARDiment features an embodied AI agent that serves as the primary anchor and interface for user interaction. The AI agent is embodied into a cute avatar designed to provide an engaging and intuitive experience by combining verbal and non-verbal cues. When the AI agent receives a response from the ChatGPT-4 API, it uses the Google Cloud Text-to-Speech API to generate speech and corresponding visemes for lip synchronization and facial animations. This integration creates a more lifelike and immersive experience for the user, mimicking human-like interactions. The AI agent's embodiment plays a crucial role in establishing a sense of presence and facilitating natural communication. By leveraging the user's gaze and saliency history, the AI agent can move around the different windows showcasing visually a level of shared understanding of the user's current focus.

3.2 Multimodal Interaction

EmBARDiment leverages multiple input and output modalities to enable seamless communication between the user and the embodied AI agent [43, 19]. The user can initiate a verbal request by pressing a key and then speaking ('V' on the keyboard). The user's speech is converted to text using the Google Speech-to-Text API [22], which sends the audio data to the Google cloud and returns the transcribed text that is then displayed on the fly in a UI panel beneath the AI agent, providing visual feedback to the user (see supplementary video). Once the final transcription is received from the Google Speech-to-Text API, it is processed by any LLM API for natural language understanding and generation. The system maintains a chat history to preserve the interaction context, allowing the AI agent to generate relevant and coherent responses. The response is then displayed on a UI panel beneath the AI agent, and we use again the Google Cloud Text-to-Speech API [23] but now to convert the text into speech. This process also produces phonemes and their corresponding visemes, which are used to animate the AI agent's facial expressions and lip movements in sync with the spoken words [51].

3.3 Gaze-Driven Contextual Memory

EmBARDiment builds on top and extends an existing open-source multi-window XR environment WindowMirror that captures existing windows from a PC and renders them inside the XR environment [9]. EmBARDiment processes each window frame using the Google Vision API [24] to perform optical character recognition (OCR), extracting the text content and its position within the frame. By correlating the spatial position of the text with the eye-tracking data from the XR headset, EmBARDiment can determine which text the user is currently paying attention to. The system maintains a buffer of the user's saliency history, preserving the order of the text to ensure coherence. To differentiate between saccades and fixations, EmBARDiment only registers text that the user has fixated on for a minimum of 120 milliseconds, which is considered the minimum time required for effective visual information processing during reading [46]. The contextual memory in our system has a maximum capacity of 250 words to enable basic episodic memory, rather than long term memory. Our agent contextual memory works by simply discarding older information as the user focuses on new content. When the user makes a verbal request, the contextual memory is combined with the user's query and sent to the ChatGPT-4 completion API for processing. This approach allows the AI agent to generate responses that are grounded in the user's current focus and saliency history. After each request, the contextual memory buffer is cleared. Below you can see an example of the generated Prompt containing the user verbal request , some prompt engineering and the gaze-driven contextual memory:

```
User's verbal request
Prompt Engineering
Contextual information

{
      "role": "user",
      "message": user verbal request from speech-to-text.
      Below is a dataset representing my visual attention it
      contains the text i have been reading from the windows
      I have been observing. Please use this information to
      inform your response to my request.
      window_id_1: {gaze-selected text on window_id_1},
      window_id_2: {gaze-selected text on window_id_2}.
      Respond with 6 sentences max, keep it under 400
      characters maximum.".
}
```

4 Experiment

To evaluate EmBARDiment, we conducted a user study. Our evaluation aims to determine (i) the merit of visual attention as a method for selecting contextual information, and (ii) to evaluate the different options for agent embodiment and the positioning of the agent.

4.1 Participants

We recruited 9 participants (5 male, 4 female, M_{Age}=30.5, SD_{Age}=5.9 years). Since in the experiment we use speech-to-text technologies we evaluated English speaking proficiency: on a scale from 1 (poor) to 5 (excellent), resulting in an average rating of $M_{English}$=4.78 $SD_{English}$=0.45. note that for 5 out of 9 participants

Figure 3: Experiment Conditions. (A) Baseline: no contextual information selected. (B) Full-Context: All the contextual information are selected. (C) Eye-gaze: information are selected based on eye-gaze fixations.

English was their primary language.Additionally, since in the experiment we use eye-tracking technologies we also evaluated vision. In terms of vision correction, 3 participants reported having normal vision, while all the others had vision corrected to normal; 4 wearing glasses and 2 contacts.

4.2 Design

We designed a within participant experiment with 3 conditions and 3 tasks employing a Latin square design, every participant completed 3 tasks, resulting in 27 unique combinations. Nine participants were therefore required to ensure each of the nine condition-task pairings was tested in every order, thus eliminating ordering/learning effects. The 3 conditions:

- A Baseline condition in which there is no contextual information added (see Fig 3a).
- A Full Context condition where all the contextual information is selected independently of the user's visual attention (see Fig 3b).
- An Eye-Tracking condition which selects information based on eye-gaze data (see Fig 3c).

The quantitative and qualitative comparison between these conditions will establish if and how contextual episodic memory will enable more implicit communication and better understand the role that eye gaze will play with AI agents in XR.

4.2.1 Q&A Reading Task

The three bodies of text of 270 to 400 words each, were selected for the experiment. All texts had a common thematic of Quantum Computing. The first was based on a Medium article about the "Schrödinger's Cat" [1]; the second was a Harvard Business Review article: "Quantum Computing Is Coming"[2], and the third was a Wikipedia article discussing "Qubits" [3]. The two questions regarding each reading were intentionally designed to be implicitly related to the context of the text and essentially meant to ask clarifications on what the participant had just read.

[cat] Questions for the "Schrödinger's Cat" text:

1. Why is it set to 50%?
2. I am confused, is the cat alive or dead?

[HBR] Questions for the "Quantum Computing Is Coming" text:

1. What are the capabilities of computers nowadays?
2. What problems do traditional computing devices struggle with?

[wiki] Questions for the "Qudits and qutrits" text:

1. What letter can I use to denote the dimension of the system?
2. I am confused about how many levels are possible?

[1] https://medium.com/swlh/quantum-computing-for-dummies-part-1-2686b9ba3c51

[2] https://hbr.org/2021/07/quantum-computing-is-coming-what-can-it-do

[3] https://en.wikipedia.org/wiki/Qubit

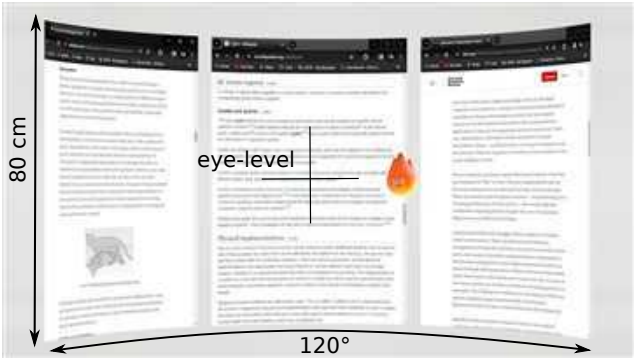

Figure 4: Experiment layout, as seen by the participants during the experiment. The 3 texts are fixed for all participants to maintain stimuli consistency. The layout of windows spawned 1 meter away from the user's head, spanning 120°(60°on the left and 60°on the right), and a resolution of 700x1200 px.

4.2.2 Multi Window Layout

The window layout, as shown in Fig 4, is an essential part of the task. The layout is intended to mimic aspects of productivity work, such as multitasking, by having multiple application windows open simultaneously. This setup mirrors common scenarios in physical multi-screen environments where multiple windows are open and used at the same time by users.

4.3 Procedure

Participants received a project overview, consent form, demographic questionnaire, and experiment instructions. The experiment began with a reminder of their right to withdraw and the opportunity to seek clarifications.

4.3.1 (Part 1) Q&A Paper Reading:

We asked participants to read the selected texts on paper. We then presented two questions about each text . This procedure was repeated for all three bodies of text (Section 4.2.1). This initial step was used to make participants familiar with the questions and relevant information within the texts.

4.3.2 (Part 2) XR Onboarding:

After answering the questions, participants engaged in a similar task but this time inside XR and interrogating the AI Agent. To guarantee a successful experiment, participants were helped with the XR headset and were guided to perform eye-tracking calibration and testing. Participants were then asked to open the XR application and familiarize themselves with the modality of interaction with the AI Agent (i.e., button press and verbal request).

4.3.3 (Part 3) Q&A AI Agent Text Interrogation:

Then, in the XR experience, they were again presented with the same 3 texts, one at a time and in different conditions. They were asked to read the text till they reach the text part pertinent to the question rehearsed on Part 1. Then they had to ask the questions to the AI agent and keep asking until the agent produced a satisfactory responses, i.e. a response equivalent to their own early reply outside XR. We instructed participants to maintain the original implicit formulation of their questions in the first attempt inside XR. If they did not receive a relevant answer, they were allowed to reformulate their question up to five times in an attempt to generate a contextually meaningful response. After five attempts, the experimenter would prompt them to move on. The study was structured around the different experimental conditions (baseline, full context, eye-tracking). And each condition had one of the three reading tasks randomly assigned (cat, HBR, wiki), so they varied by participants.

4.3.4 (Part 4) Experience Survey:

Following each completed condition, participants were asked questions from a subset of the Human Language Model Interaction Questionnaire (HLMIQ) [30] shown in Table 1, and gave verbal responses while the experimenter annotated those responses.

4.3.5 (Part 5) Agent Embodiment Preferences:

Upon completion of the Q&A tasks/survey, the study transitioned to the exploration of AI agent embodiment preferences. Participants were introduced to different possible embodiment types (no embodiment, non anthropomorphic, or anthropomorphic embodiment), embodiment location (general fixed, or contextual positioning such as following the active window). Participants were asked to engage with each option to form a basis for their preferred choice (Table 2).

4.4 Questionnaires

To better characterize the user experience beyond the quantitative analysis of the attempts, we asked participants to fill a subset (Table 1) of the Human Language Model Interaction Questionnaire [30]. The HLMIq assesses important aspects such as: Helpfulness, Ease, Enjoyment, Satisfaction, Responsiveness of the AI Agent.

After each experimental block consisting of the 3 trials/texts, each participant was shown a few options in terms of Embodiment Type and Location within the multi-window space and then completed a survey on (Table 2).

5 Results

Each of the 9 participants had to ask the AI agent 2 questions across the 3 conditions leading to a total of 54 questions. The potential attempts to reach satisfactory answers can range from 54 to 270 (as each participant could reformulate the same question 4 times). In total we collected 110 attempts on the experimental setting. After each of the 3 conditions, participants completed Human Language Model Interaction questionnaires leading to 27 questionnaires. Participants also completed the agent embodiment questionnaires, leading to 9 filled surveys.

Table 1: Human Language Model Interaction Questionnaire

Aspect	**Question**	**Answ. Type**
Helpfulness	Independent of its fluency, how helpful was your AI Teammate in answering the questions?	Likert
	Why did you find the AI Teammate helpful or unhelpful? Give a concrete example if possible.	Open
Ease	How easy was it to interact with the AI Teammate?	Likert
Enjoyment	How enjoyable was it to interact with the AI Teammate?	Likert
Satisfaction	I am satisfied with the answer I received?	Likert
Change	Did the way you chose to interact with the AI Teammate change after the first question? If so, how?	Open
Description	What adjectives would you use to describe the AI Teammate?	Open
Responsiv.	How responsive was the system?	Likert

Table 2: Survey on Agent Embodiment Preferences

Aspect	**Question**	**Possible Answers**
Embodiment	What of these 3 options do you prefer?	1)Lack of embodiment 2) Non-anthropomorph. 3)Anthropomorph.
	Why did you choose the embodiment type?	(Open ended)
Location	What of these 2 options do you prefer?	(1) Context-following, (2) Fixed
	Why did you choose the location?	(Open ended)

5.1 Question Attempts

In our experiment, participants interrogate an AI agent, in search for explanations related to a featured text. If the AI's response is irrelevant or unacceptable, participants reformulate their question, a maximum of five times to obtain an acceptable answer. Fewer attempts indicate better performance, as they signify less effortto reach an acceptable answer. We analyzed the dialogues to count the number of attempts, totaling 110 verbal requests. To compare the number of attempts across conditions, we performed a repeated measures ANOVA with Condition as a three levels factor (baseline, full context, eye-tracking). Fig 5.

The ANOVA analysis on the number of questions asked indicate a significant difference between the conditions (F(2, 24) = 16.3, p ¡ .001). To further unpack the results, we conduct a post-hoc analysis, and find significant differences across all conditions. The comparison between the number of attempts in the Baseline (M = 5.5, SD= 1.2) and Full Context (M = 4.2, SD = 1.3) revealed a mean difference and probability of ($\overline{X}\Delta = 1.333, p = 0.047$) and a large effect size (Cohen's d = 1.193), indicating a significant reduction in the number of attempts in the Full Context condition. This reduction in attempts is further emphasized in the comparison between the Full Context and Eye-Tracking (M = 2.5, SD = 0.7), which showed a mean difference and probability of ($\overline{X}\Delta = 1.667, p = 0.011$) and an

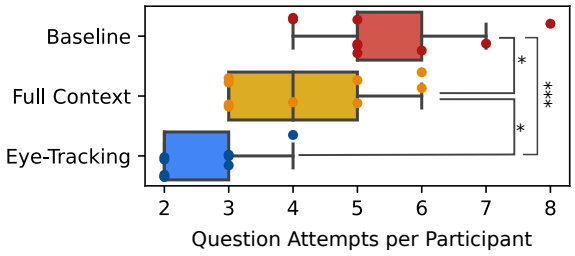

Figure 5: Boxplot of question attempts. Significant differences highlighted with: * p < .05, *** p ¡ .001.

even larger effect size (Cohen's d = 1.491), demonstrating a further decrease in number of attempts for the Eye-Tracking condition. Additionally beyond the within groups comparisons, the mean squares between groups ($\sum_{i=1}^{D}(x_i - y_i)^2 = 20.3$), also suggest notable variance in the number of questions asked across different conditions. To further clarify the results, we utilized a Sankey plot to illustrate the success rates at each attempt (Fig 6). In the Eye-tracking condition, 77.7% of participants achieved the intended result on the first attempt. For the Full Context condition, participants typically reached success by the second attempt. In contrast, in the Baseline condition, 77.7% of participants obtained a satisfactory answer by the third attempt.

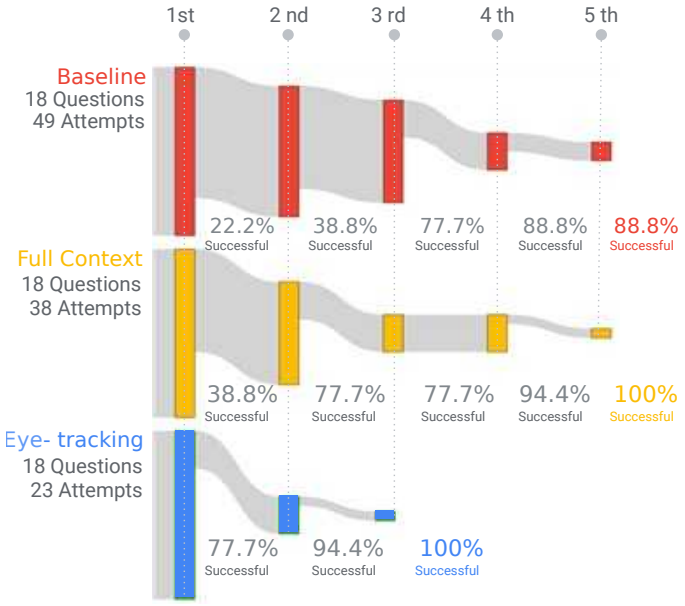

Figure 6: Sankey diagram depicting showing the success rate at each subsequent attempt.

5.2 Human Language Model Interaction Questionnaire (HLMIq)

Participants completed HLMIq questionnaire after each experimental condition, resulting in a total of 9 entries per condition and 27 overall. While we found significant results for Helpfulness and Satisfaction, no significant findings emerged for Ease, Enjoyment, and Responsiveness. Therefore, for conciseness, we only report the analysis of the former and omit the latter.

5.2.1 Satisfaction (Likert)

The Kruskal-Wallis test indicated a significant difference in conditions ($H(2) = 12.838$, $p = 0.002$). Dunn's Post Hoc comparisons revealed a higher satisfaction on the Eye-Tracking (M = 4.7, SD = 0.4) condition than both on the Baseline (M = 3 , SD = 1.3), ($p < 0.001$, $p_{bonf} = 0.001$), and on the Full Context conditions (M = 3.8, SD = 0.6), ($p = 0.022$, $p_{bonf} = 0.067$).

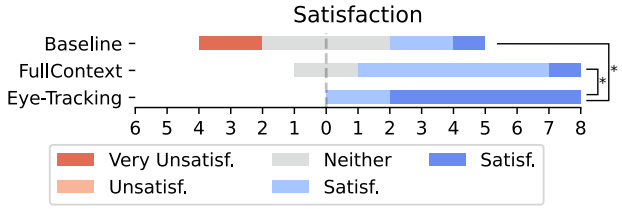

Figure 7: Satisfaction responses to the HLMIq, * p < .05.

5.2.2 Helpfulness (Likert)

The Kruskal-Wallis test revealed a significant effect of the conditions on the outcomes ($H(2) = 7.404$, $p = 0.025$). Subsequent Dunn's post hoc comparisons showed Eye-Tracking (M = 4.4, SD = 0.8) to be significantly more helpful that the Baseline condition (M = 2,7, SD = 1.4) and Full Context conditions ($p = 0.007$, $p_{bonf} = 0.022$, Figure 8). And showing consistency with the results from our quantitative analysis of question attempts.

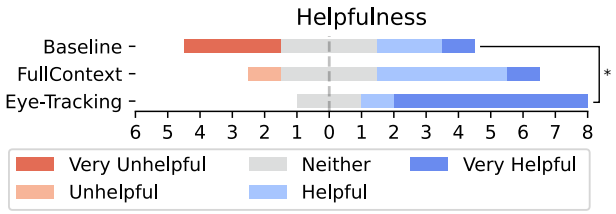

Figure 8: Helpfulness responses to the HLMIq, * p < .05.

5.2.3 Helpfulness (Open)

To further unpack this result we tap on the Open ended question of the HLMIq:"Why did you find the AI Teammate helpful or unhelpful?" For the Baseline condition, participants noted that the initial answers of the AI Agent were general and unrelated to the context of the text and that a satisfactory answer was only obtained after significant prompting. *P5: "..it did not give any specific information to the text i was reading.." P8: "..took a lot of prompting for the first answer but did really good in the second question.."* For the Full Context condition, participants experienced mixed results saying that the journey to the correct answer was often cumbersome *P3: "..it was able to answer my question even if it took multiple attempts." P5: "..it's giving a very generic answer, unless I asked twice and be specific."* In the eye-tracking condition, participants reported accurate answers straight to the point. *P4: "..it answered the question very accurately." P9: "..it answered the questions straight away."*

5.2.4 Behavior changes (Open)

To better understand the differences between conditions we also looked at the quesiton: "Did the way you chose to interact with the AI Teammate changed after the first question? If so, how?". In the Baseline condition the majority of responses indicate participants provided more information in an attempt to guide the AI.

P6: "I gave explicit content from the article to ensure it gave me the right answer.", *P9: "I gave more context to it."* Participants reported learning from their initial interaction, indicating an iterative process of adjusting their approach based on the AI's responses. In the Full Context condition the majority of participant still reported the same strategies "of adding more information".

5.3 Conversation Evaluation

We review the recorded conversation to better characterize the differences between the Full Context and Baseline condition. And to understand how and why the Full Context condition failed in answering correctly.

5.3.1 Conversation Examples

To further support this section we report two full example conversations from participant P2. Similarly to the HLMIq we also find two strategies in these examples:

Strategy 1: Referring to Context

Strategy 2: Adding More Information

Example 1 (P2) Baseline:

Participant 2 : Why is the chance of radioactive decay 50%?
AI Agent : The chance of radioactive decay is not inherently 50% for all isotopes...
Participant 2 : In this example, though. Why is it 50%?
AI Agent : In your example, if the chance of decay is 50%, it means we're looking at a specific moment..
Participant 2 : In the Schrödinger's cat example, why is the chance of decaying 50%?
AI Agent : In Schrödinger's cat thought experiment, the 50% decay chance is a simplification to illustrate quantum superposition..

Example 2 (P2) Full Context:

Participant 2 : What are the core capabilities of nowadays computers?
AI Agent : Quantum computers harness the principles of quantum mechanics to process information....
Participant 2 : What are the core capability of computers described in this article?
AI Agent : The article highlights that digital computers excel at arithmetic operations...

Example 3 (P2) Eye-Tracking:

User : What's the term that describes the capabilities of computers?
AI Agent : The term that describes the capabilities of computers is arithmetical machines.

5.4 Agent Embodiment Preferences

This Agent Embodiment Preferences Survey was once by each participant at the end of the whole procedure (n=9).

5.4.1 Agent Embodiment Type

When we asked, "How do you prefer the visual representation of a digital assistant to be designed?" Participants were shown three options: Lack of Embodiment, Anthropomorphic, and Non-anthropomorphic. After experiencing each of the conditions, participants were asked to express a preference. The option that received the most preferences was Anthropomorphic (56.2%), followed by Non-Anthropomorphic (31.1%) and Lack of Embodiment (12.7%) (Figure 9).

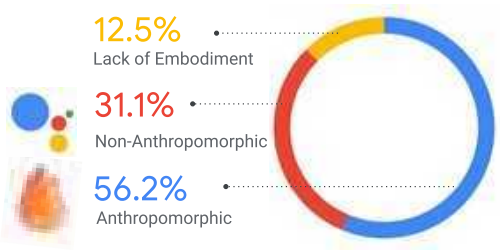

Figure 9: Circle chart representing user's preferences in terms of embodiment's type.

To further unpack these results, participants were asked to justify their preferences. Those whose preference was anthropomorphic mentioned a desire for interactions that mimic human communication. *P6: "A human-like agent makes me a bit less self-conscious about making a verbal request"*. However, many who expressed a preference for the anthropomorphic design were flexible about the representation; they would be fine with an abstract, non-anthropomorphic option, as long as it was visually present. *P1: "I don't mind between the two as long as it is represented. I don't want the disembodied because I don't know if it's present or listening to me."* The representation in a visual form was signifying to participants the assistant's presence and attentiveness. *P5: "I like it to be there because it reminds me that I can ask questions"* Those who chose lack of embodiment valued the absence of representation, as it was non-distracting and not visually imposing.

5.4.2 Agent Embodiment Location

In a similar way we then asked participants, "Where do you prefer the AI assistant to be located?" giving them two choices: Context-following or Fixed. After showing these options to participants and addressing any questions, they were asked to make a choice. Most participants preferred Context-following (33.3%), but around one-third opted for Fixed (66.6%) (Fig 10.

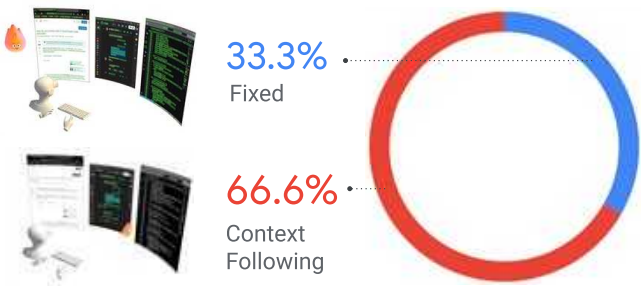

Figure 10: Circle chart representing user's preferences in terms of embodiment's location.

When asked To justify Context-following choices participant mentioned a desire for the digital assistant to be easily accessible without requiring the user to turn their head or attention.*P2: "I*

don't have to turn my head. " P7: "I would want the avatar to follow my gaze better. " The context-following group also highlighted a desire for clarity in how the assistant's movements and positioning reflect its awareness of the context. *P1: "Context following makes it easier for me to understand what the avatar is aware of."*

6 Discussion

6.1 Why the AI Agent failed in Full Context Condition

We further look at the results and we find a general consistent trend: the AI Agent failing on the Full Context condition. But what are the differences between the Full Context and the Baseline conditions that cause this? One of our main learning was that the AI Agent performance faltered in the Full Context condition when compared to the Eye-Tracking condition, we believe this is due to the expansive context, which allowed for multiple potential answers. This was in stark comparison to our 250 word episodic memory implemented for gaze. For instance, if we look at the Conversation examples (section 5.3.1), in the "Example 2 (P2) Full Context", to the question *"What are the core capabilities of nowadays computers?"* the AI Agent answer *"Quantum computers harness the principles of quantum mechanics to process information"* is influenced by the broader context about quantum computing, as all 3 articles talk about quantum computing while only one talks about the capabilities of computers nowadays. Given that the Full Context is primarily concerned about quantum computing, the AI Agent tended to focus its answers on quantum computers, illustrating how extensive context can skew its responses. This led to inaccuracies, as the model struggled to integrate the question on the specific context the user was reading.

6.2 Strategies for Full-context and Baseline

Participants reported employing a strategy consisting of *"adding more information"* across both conditions, however, by analyzing the conversation we found that this actually encompassed two distinct strategies.

Strategy 1: Referring to Context - This strategy prompts the AI Agent to consider the context in which a question is asked. While it showed limited success in the Baseline condition due to the lack of contextual information, its application in the Full Context condition demonstrated a notable impact. For instance, when Participant 2 initially asked about the core capabilities of modern computers, the AI Agent's response was skewed towards quantum computing. However, when the participant refined the question to focus on the context described within the article, the AI Agent adjusted its search within the given context and identified an accurate alternative.

Strategy 2: Adding More Information - This strategy entails providing additional, and specific information to the AI Agent, allowing it to leverage its internal knowledge base. For example,in the Baseline condition, questions about why the chance of decay was 50% in which the AI Agent had no context led to wrong responses. When "Schrodinger's cat" was mentioned, the AI Agent could tap into its existing knowledge and provide a good response.

6.3 Design Considerations

In the realm of AI Agents in multi-window XR environments, our research underscores the critical importance of integrating implicit and explicit inputs to enhance user interactions.To fully enable the integration of explicit and implicit inputs, we propose the alignment between visual attention and verbal input that can enable contextual memory. A new standard for capturing and leveraging user engagement history. This might prove especially relevant in the context of information-rich multi-window environments where there might be multiple contextual targets for an interaction potentially spanning across multiple windows.

Multi-modal input: developers should focus on harnessing the complementary strengths of visual attention and verbal communication to craft AI agents that are both intuitive and responsive to user input. Visual attention serves as a subtle yet powerful signal for context recognition, allowing the AI to grasp the user's focus without invasive prompts. Concurrently, verbal requests provide a direct avenue for users to convey their needs and queries. Our application showcases how directing the AI Agent's attention to user-focused context mitigates the challenges associated with processing extensive inputs. By aligning the AI Agent's focus with the user's, especially in scenarios requiring immediate clarification on recent interactions, the system gains a significant edge over models burdened with the entirety of available context.

Episodic vs Long-term Memory: our agent contextual memory works by simply discarding older information as the user focuses on new content. The contextual memory in our system has a maximum capacity of 250 words to enable basic episodic memory, rather than long term memory. Changing the capacity will impact the LLM performances (i.e. responsiveness) and has the risk to run into the same problems we found on the Full Context condition. Therefore, after each request, the contextual memory buffer is cleared, ready to capture new context.

7 Future Work

The current study focused on a scenario which requires the AI Agent's attention aligned with the user's attention to provide clarifications on recently observed content. This approach assumes the user's visual focus aligns with their informational needs. However, scenarios exist where the AI Agent's focus might need to diverge, exploring content the user hasn't directly engaged with—perhaps to highlight overlooked details. Future work should expand beyond this alignment, enabling the AI Agent to consider broader or alternate contexts.

8 Conclusion

EmBARDiment, explored a novel framework for integrating context-aware embodied AI agents within extended reality (XR) environments. Our solution goes beyond conventional explicit text-based chat-bots and leverages an *implicit* attention framework to streamline communication and make it more contextually grounded. We utilize the user's gaze and visual saliency to derive relevant context. We ran an experiment comparing three interaction conditions —without context, with full context, and with eye-tracking user-focused context—, participants interacted with an AI agent, adjusting their questions based on the agent's responses. Findings indicated that using visual attention to guide contextual memory selection resulted in fewer question reformulations and enhanced user satisfaction and perceived helpfulness of the system. This confirms our hypothesis that such an approach significantly improves interaction efficiency with AI Agents in multi-window XR settings. Our gaze-driven contextual memory extends previous research [34, 47, 37], on multi modal interaction gaze + speech. Additionally our findings highlight the potential to streamline user interactions with AI agents in XR offering specific Design considerations for the design of AI agent functionality in multi-window settings.

Acknowledgments

The authors wish to thank A, B, and C. This work was supported in part by a grant from XYZ.

References

[1] Unity 2023. https://unity.com/, 2023. Accessed: 2023-09-28. 3

[2] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First Interna-*

tional Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1, pp. 304–307. Springer, 1999. 2

[3] E. Adamopoulou and L. Moussiades. Chatbots: History, technology, and applications. *Machine Learning with applications*, 2:100006, 2020. 2

[4] K. Ahuja, A. Kong, M. Goel, and C. Harrison. Direction-of-voice (dov) estimation for intuitive speech interaction with smart devices ecosystems. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pp. 1121–1131, 2020. 2

[5] Anthropic. Claude ai. https://claude.ai/, 2023. Accessed: 2023-09-30. 1

[6] Apple. Apple vision pro. https://www.apple.com/vision-pro/, 2022. 1

[7] V. Biener, D. Schneider, T. Gesslein, A. Otte, B. Kuth, P. O. Kristensson, E. Ofek, M. Pahud, and J. Grubert. Breaking the screen: Interaction across touchscreen boundaries in virtual reality for mobile knowledge workers. *IEEE transactions on visualization and computer graphics*, 26(12):3490–3502, 2020. 1, 2

[8] R. A. Bolt. "Put-That-There": Voice and Gesture at the Graphics Interface. In *Proceedings of the 7th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '80, p. 262–270. Association for Computing Machinery, New York, NY, USA, 1980. doi: 10.1145/800250.807503 2

[9] R. Bovo, E. J. Gonzalez, L.-T. Cheng, and M. Gonzalez-Franco. Windowmirror: an opensource toolkit to bring interactive multi-window views into xr. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 436–438. IEEE, 2024. 3

[10] B. L. Brumitt, S. Shafer, J. Krumm, and B. Meyers. Easyliving and the role of geometry in ubiquitous computing. In *Proceedings of the DARPA/NIST/NSF Workshop on Research Issues in Smart Computing Environments*. Atlanta, GA, 1999. 2

[11] C. Chakraborty, S. Pal, M. Bhattacharya, S. Dash, and S.-S. Lee. Overview of chatbots with special emphasis on artificial intelligence-enabled chatgpt in medical science. *Frontiers in Artificial Intelligence*, 2023(1237704), 2023. doi: 10.3389/frai.2023.1237704 2

[12] C. Chakraborty, S. Pal, M. Bhattacharya, S. Dash, and S.-S. Lee. Overview of chatbots with special emphasis on artificial intelligence-enabled chatgpt in medical science. *Frontiers in Artificial Intelligence*, 6, 2023. 2

[13] P. Chittò, M. Baez, F. Daniel, and B. Benatallah. Automatic generation of chatbots for conversational web browsing. In *Conceptual Modeling: 39th International Conference, ER 2020, Vienna, Austria, November 3–6, 2020, Proceedings 39*, pp. 239–249. Springer, 2020. 2

[14] C. Coupé, Y. M. Oh, D. Dediu, and F. Pellegrino. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science advances*, 5(9):eaaw2594, 2019. 1, 2

[15] A. K. Dey, G. D. Abowd, and D. Salber. A context-based infrastructure for smart environments. In *Managing Interactions in Smart Environments: 1st International Workshop on Managing Interactions in Smart Environments (MANSE'99), Dublin, December 1999*, pp. 114–128. Springer, 2000. 2

[16] A. K. Dey, G. D. Abowd, and D. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 16(2-4):97–166, 2001. 2

[17] A. K. Dey, D. Salber, G. D. Abowd, and M. Futakawa. The conference assistant: Combining context-awareness with wearable computing. In *Digest of Papers. Third International Symposium on Wearable Computers*, pp. 21–28. IEEE, 1999. 2

[18] N. Fereydooni and B. N. Walker. Virtual reality as a remote workspace platform: Opportunities and challenges. 2020. 2

[19] M. Gonzalez-Franco and A. Colaco. Guidelines for productivity in virtual reality. *Interactions*, 31(3):46–53, 2024. 2, 3

[20] Google. Bard: Google ai and search updates. https://blog.google/technology/ai/bard-google-ai-search-updates/, 2023. Accessed: 2023-09-30. 1

[21] Google. Gemini. https://gemini.google.com/app, 2023. Accessed: 2023-09-30. 1

[22] Google. Cloud speech-to-text, Accessed 2024. Accessed: 2024-03-27. 3

[23] Google. Cloud text-to-speech, Accessed 2024. Accessed: 2024-03-27. 3

[24] Google. Optical character recognition (ocr) — cloud vision api, Accessed 2024. Accessed: 2024-03-27. 3

[25] L. Greci. Xr for industrial training & maintenance. *Roadmapping Extended Reality: Fundamentals and Applications*, pp. 309–320, 2022. 2

[26] J. Grubert, L. Witzani, E. Ofek, M. Pahud, M. Kranz, and P. O. Kristensson. Text entry in immersive head-mounted display-based virtual reality using standard keyboards. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 159–166. IEEE, 2018. 1, 2

[27] D. Guo, A. Gupta, S. Agarwal, J.-Y. Kao, S. Gao, A. Biswas, C.-W. Lin, T. Chung, and M. Bansal. GRAVL-BERT: Graphical Visual-Linguistic Representations for Multimodal Coreference Resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 285–297. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, Oct. 2022. 2

[28] A. Hartholt, E. Fast, A. Reilly, W. Whitcup, M. Liewer, and S. Mozgai. Ubiquitous virtual humans: A multi-platform framework for embodied ai agents in xr. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE, 2019. 2

[29] S. Jeuris and J. E. Bardram. Dedicated workspaces: Faster resumption times and reduced cognitive load in sequential multitasking. *Computers in Human Behavior*, 62:404–414, 2016. doi: 10.1016/j.chb.2016.03.059 1

[30] M. Lee, M. Srivastava, A. Hardy, J. Thickstun, E. Durmus, A. Paranjape, I. Gerard-Ursin, X. L. Li, F. Ladhak, F. Rong, et al. Evaluating human-language model interaction. *arXiv preprint arXiv:2212.09746*, 2022. 5

[31] D. Li, T. Chen, A. Zadikian, A. Tung, and L. B. Chilton. Improving automatic summarization for browsing longform spoken dialog. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–20, 2023. 1, 2

[32] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom. Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80:9411–9457, 2021. 2

[33] S. Marwecki, A. D. Wilson, E. Ofek, M. Gonzalez Franco, and C. Holz. Mise-unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In *Proceedings of the 32nd annual ACM symposium on user interface software and technology*, pp. 777–789, 2019. 2

[34] S. Mayer, G. Laput, and C. Harrison. Enhancing Mobile Voice Assistants with WorldGaze. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 4 2020. doi: 10.1145/3313831.3376479 2, 8

[35] J. F. McCarthy and E. S. Meidel. Activemap: A visualization tool for location awareness to support informal interactions. In *International Symposium on Handheld and Ubiquitous Computing*, pp. 158–170. Springer, 1999. 2

[36] Meta. Meta quest pro. https://www.meta.com/quest/, 2022. 1, 3

[37] D. Miniotas, O. Špakov, I. Tugoy, and I. S. MacKenzie. Speech-Augmented Eye Gaze Interaction with Small Closely Spaced Targets. In *Proceedings of the 2006 Symposium on Eye Tracking Research & Applications*, ETRA '06, p. 67–72. Association for Computing Machinery, New York, NY, USA, 2006. doi: 10.1145/1117309.1117345 2, 8

[38] S. Noy and W. Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192, 2023. 2

[39] N. Numan, D. Giunchi, B. Congdon, and A. Steed. Ubiq-genie: Leveraging external frameworks for enhanced social vr experiences. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 497–501. IEEE, 2023. 2

[40] E. Ofek, J. Grubert, M. Pahud, M. Phillips, and P. O. Kristensson. Towards a practical virtual office for mobile knowledge workers. *arXiv preprint arXiv:2009.02947*, 2020. 2

[41] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt/, 2022. 1

[42] L. Pavanatto, C. North, D. A. Bowman, C. Badea, and R. Stoakley. Do we still need physical monitors? an evaluation of the usability of ar virtual monitors for productivity work. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, pp. 759–767. IEEE, 2021. 1, 2

[43] K. Pfeuffer, H. Gellersen, and M. Gonzalez-Franco. Design principles and challenges for gaze+ pinch interaction in xr. *IEEE Computer Graphics and Applications*, 44(3):74–81, 2024. 2, 3

[44] A. Poushneh. Humanizing voice assistant: The impact of voice assistant personality on consumers' attitudes and behaviors. *Journal of Retailing and Consumer Services*, 58:102283, 2021. 2

[45] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pp. 28492–28518. PMLR, 2023. 2

[46] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998. 3

[47] Y. Romaniak, A. Smielova, Y. Yakishyn, V. Dziubliuk, M. Zlotnyk, and O. Viatchaninov. Nimble: Mobile Interface for a Visual Question Answering Augmented by Gestures. In *Adjunct Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, p. 129–131. Association for Computing Machinery, New York, NY, USA, 2020. doi: 10.1145/3379350.3416153 2, 8

[48] W. N. Schilit. *A system architecture for context-aware mobile computing*. Columbia University, 1995. 2

[49] R. Suzuki, M. Gonzalez-Franco, M. Sra, and D. Lindlbauer. Xr and ai: Ai-enabled virtual, augmented, and mixed reality. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–3, 2023. 2

[50] X. Tan, T. Qin, F. Soong, and T.-Y. Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021. 2

[51] M. Volonte, E. Ofek, K. Jakubzak, S. Bruner, and M. Gonzalez-Franco. Headbox: A facial blendshape animation toolkit for the microsoft rocketbox library. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 39–42. IEEE, 2022. 3