# Demonstrating *PANDALens*: Enhancing Daily Activity Documentation with AI-assisted In-Context Writing on OHMD

Nuwan Janaka
nuwanj@u.nus.edu
Smart Systems Institute, Synteraction Lab
National University of Singapore
Singapore

Runze Cai*
runze.cai@u.nus.edu
Synteraction Lab, School of Computing
National University of Singapore
Singapore

Shengdong Zhao*
shengdong.zhao@cityu.edu.hk
Synteraction Lab, School of Creative Media & Department of Computer Science,
City University of Hong Kong
Hong Kong, China
National University of Singapore
Singapore

David Hsu
dyhsu@comp.nus.edu.sg
Smart Systems Institute, School of Computing
National University of Singapore
Singapore

## ABSTRACT

We introduce *PANDALens*, a **P**roactive **AI** **N**arrative **D**ocumentation **A**ssistant built on an Optical See-Through Head-Mounted Display that transforms the in-context writing tool into an intelligent companion during daily activities. *PANDALens* observes multimodal contextual information from user behaviors and the environment to detect interesting moments and elicit contemplation. It also employs Large Language Models to transform such multimodal information into coherent narratives with significantly reduced user effort. *PANDALens* was iteratively designed through a formative study identifying the user requirements. We verify its utility in a real-world travel scenario in improving writing quality and travel enjoyment while minimizing user effort.

## CCS CONCEPTS

• **Human-centered computing → Ubiquitous and mobile computing systems and tools**; **Empirical studies in interaction design**.

## KEYWORDS

HMD, smart glasses, AI, large language model, multimodal information, Human-AI collaborative writing, in-context writing, travel blog

**ACM Reference Format:**

Nuwan Janaka, Runze Cai, Shengdong Zhao, and David Hsu. 2024. Demonstrating *PANDALens*: Enhancing Daily Activity Documentation with AI-assisted In-Context Writing on OHMD. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3613905.3648644

*Corresponding Authors.



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



ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3648644

## 1 INTRODUCTION AND RELATED WORK

The development of technology has improved our capacity to document our daily life experiences. Such experience documentation serves various purposes, including preserving memories that support Recollecting, Reminiscing, Retrieving, Reflecting, and Remembering intentions (the 5Rs) [32], as well as sharing experiences [25].

Consider Jane, a traveling enthusiast who likes to capture and share travel moments on social media. Unlike traditional lifelogging cameras that record entire journeys but need extensive post-editing efforts to identify interesting moments [4, 22], AI-driven lifelogging streamlines this process by detecting and extracting intriguing events [4, 9] for Jane. However, it might overlook essential moments due to limited accuracy [22] and may miss Jane's personal expression within the context. Jane could document her reflections post-trip, but details might be lacking due to memory decay [20]. One approach to prevent memory loss is in-context writing, a process of documenting experiences as they unfold, embedding the writer's immediate, vivid thoughts, feelings, and reflections on the instant moments [20, 21]. For example, *LiveSnippets* [20] allows Jane to take photos of interesting moments with short comments in context using her smartphone. However, this method isn't without flaws: 1) Although commonly used in travel [34], smartphones are more like reactive tools requiring hands-occupied and heads-down interaction [17, 37], challenging users to capture fleeting moments. 2) Without context-related guidance, users may tend to make short and superficial comments [20], compromising in-context writing quality and needing extra effort for post-editing.

This leads us to our research question and design goal: *How can we support high-quality, personalized documentation in everyday activities (e.g., travels) but with seamless interaction during users' primary tasks (e.g., travels) and minimum post-editing efforts*?

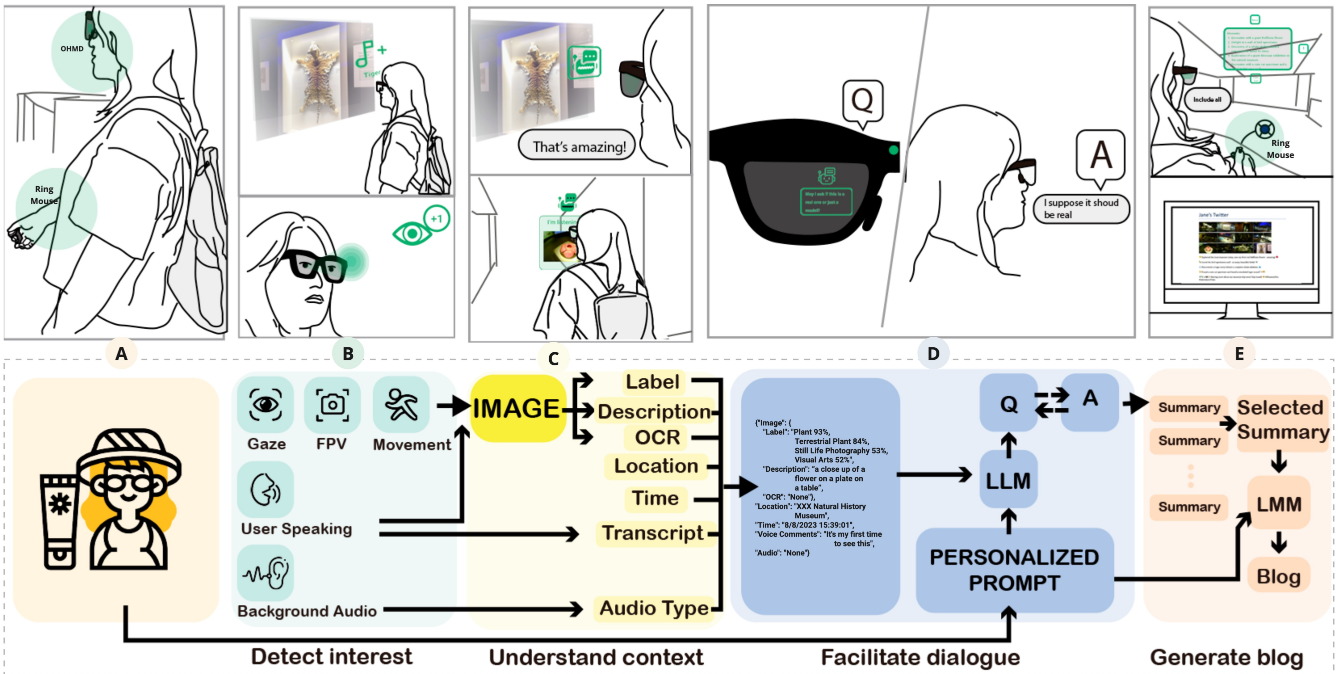


**Figure 1: (A) A user travels with *PANDALens*, an AI-assisted in-context writing tool equipped with an Optical See-Through Head-Mounted Display (OHMD) and a ring mouse. (B) The system leverages various modalities to detect the user's interests during travel, such as potential interesting audio (e.g., music sounds in a demonstration) and gaze patterns (e.g., looking at objects). (C) Detecting interests, the system displays icons (e.g., with auto-captured images) and prompts the user to comment verbally. It then transcribes this comment and combines it with other data such as image, audio, time, and location to assemble the contextual data. (D) Using the contextual information, the system formulates context-specific questions in the user's preferred style with a Large Language Model (LLM). The user can then respond to these questions. LLM also creates a summary of the moment, which can be refined based on the user's feedback. (E) Post-trip, the user can activate *PANDALens* using the ring mouse to automatically generate travel blogs. A list of recorded moments is displayed for the user to choose from. Once selected, the system drafts a travel blog that mirrors the user's unique style.**

We introduce *PANDALens* (**P**roactive **AI** **N**arrative **D**ocumentation **A**ssistant, see usage at **Figure 1**), a proof-of-concept, AI-assisted in-context writing system on Optical See-Through Head Mounted Displays (OST-HMD, OHMD, augmented-reality smart glasses). The wearable heads-up platform [37] reduces the efforts in moment capture by leveraging AI to observe multimodal context information [1] [11, 27] from user natural behaviors (e.g., gaze, movement, voice) and environments (e.g., objects in egocentric view and ambient audio), subsequently offering moment capture suggestions proactively. Users can respond to these suggestions via natural voice dialogue or subtle ring interactions [7, 31]. It leverages mixed-initiative interactions [1, 2, 15] to reduce interference and utilizes Large Language Models (LLMs) [5, 26] for document co-creation [10, 30]. To elicit detailed user expressions and facilitate intelligent dialogues, the LLM is used to interpret the multimodal contextual information of the captured moment and generate context-related [23] questions. To enhance the quality of the final documentation, the integrated LLM utilizes contextual information with detailed user expressions to craft the narratives progressively, minimizing user editing efforts. For detailed evaluation, please refer to our original paper, *PANDALens* [6].

[1]In our context, multimodal information refers to visual, audio, spatial, and temporal data of the user and environment.

## 2 *PANDALENS* SYSTEM

In this section, we first depict the usage scenarios of *PANDALens*, then detail its primary features.

### 2.1 Usage Scenarios of *PANDALens*

Consider Jane, the previously mentioned traveler, as a keen participant at CHI'24 in Hawaii, exploring the interactive sessions dedicated to the latest advancements in human-computer interaction. As Jane enters the venue, a cutting-edge robotic exhibit catches her attention. Drawn by the exhibit's intriguing design, Jane approaches for a closer inspection. Sensing Jane's focused interest (e.g., prolonged gaze), *PANDALens* automatically captures an image of the robotic exhibit. Recognizing that Jane is fully engaged with the exhibit, *PANDALens* waits to offer comment suggestions until she has finished her observation and proceeds to the next showcase.

Delighted by the robotic design, Jane mentions to *PANDALens* that this is her first encounter with such an advanced robotic system. With this feedback, and by analyzing the captured image and location data, *PANDALens* inquires for more specifics, asking, "Do you want to try to interact with the interactive robot?" Jane responds, "Yes, let me try. It's fascinating to see the technology in person!"

As an enthusiast of Augmented Reality (AR), when Jane discovers a section dedicated to AR experiences, *PANDALens* detects the AR smart glasses in Jane's field of view. Given Jane's interests, it captures a snapshot and displays a 'like icon', inviting Jane's comments. Jane expresses her excitement about the diverse AR applications displayed. *PANDALens* then inquires about her favorite AR experience. However, Jane's attention shifts to an immersive virtual reality (VR) presentation. Ignoring the question, which fades away, Jane seeks the perfect angle to capture the VR experience in a CAVE. After snapping a photo through a subtle gesture, Jane shares her thoughts on the potential of VR with *PANDALens.* On revisiting the AR section later, *PANDALens* refrains from auto-capturing or suggesting comments to avoid repetition. Later, intrigued by an interactive AI art installation, Jane moves closer, prompting *PANDALens* to take photos. An invitation for comments appears when detecting the ambient sounds associated with the installation. Jane continues to explore the conference with *PANDALens* as her digital companion.

Following the demo session, Jane uses ring interactions to prompt *PANDALens* to compile a blog post. *PANDALens* presents a selection of recorded moments for Jane to highlight in the narrative. After her selection, *PANDALens* crafts a blog detailing Jane's interactivity experiences at CHI'24, enriched with personal insights and standout moments. Although Jane values detailed narratives, she wishes to share her journey on Twitter. Hence, she requests *PANDALens* to adapt the content into a Twitter-friendly format, incorporating emojis for emphasis. After refining the content based on Jane's feedback, *PANDALens* finalizes the narrative and images, transferring them to Jane's laptop for sharing on social media.

## 2.2 Key Features of *PANDALens* System

As demonstrated in the above usage scenarios, the interaction flow of *PANDALens* encompasses three stages: (1) Capturing Interested Moments: Using a mixed-initiative interface seamlessly merges AI-driven and human-initiated actions. (2) Context-Related Questions Generation: *PANDALens* presents context-related queries by leveraging the multimodal information extracted from the captured moments. (3) Final Narratives Generation: After travel, *PANDALens* offers users the autonomy to select their favored captured moments. It then generates a draft document and enables revisions based on user preferences. In the following, we introduce the functions of three major components in the ***Final*** *System*[2].

### 2.2.1 *Mixed Initiation for Moment Capture: AI Initiation.*

We incorporated a set of modalities tailored for travel scenarios as a proof of concept to detect users' situational and personal interests [28, 33]. We also designed strategies to mitigate false positive suggestions and information overwhelming from the AI assistant.

*Multimodal Analyzer for User Interest Detection.* The system processes various modalities in real-time and concurrently to detect the two types of interests. To identify situational interests, the system recognizes positive sentiments, including joy and surprise, in user verbal expressions, given travelers mainly report positive experiences in travel blogs [8].

For personal interest detection, the system monitors two types of context information from the environment that match user preferences, objects within the FPV and background audio, to discern visual and auditory preferences. To quickly assess user visual and auditory preferences from a wide range of categories, we ask LLM to "Create an interactive questionnaire to narrow down two lists related to interest detection for the COCO dataset and MediaPipe Audio classification, based on the user's travel preferences." Based on users' answers, the LLM consequently formulates two lists of potential options, allowing users to narrow down their choices further.

Additionally, two triggers are utilized to detect both situational and personal interest, with optimization based on pilot testing results. The first, Gaze Fixation, is detected when eyes remain focused on a small area, deviating no more than 4.91 degrees for at least 1 second. The second trigger, "Zoom-In", activates when users approach an object closely while looking at it. This intent is identified by the target object size increases by 10% in two consecutive FPV frames.

*Interaction Design.* As depicted in Figure 2 (AI Initiative-Detect Interest), the AI suggests moment capture upon detecting user interest. Users can confirm such interest by verbally commenting, which triggers the system to auto-record. To overcome the uncertainty in AI decisions, we follow the mixed-initiative guidelines [2, 15]. If the user ignores the suggestion, it gradually vanishes, or users can dismiss it manually (by pressing the center button on the ring mouse).

Moreover, to mitigate distraction from AI suggestions, we adopted the following designs: (1) We utilized the principles of "matching attentional draw with utility" [13, 36] for notifications. For instance, audio notifications attract more attention [12]. Thus, initial invitations of moment commenting are only conveyed through subtle visual feedback (e.g., icons [7, 16, 19]), and audio notifications are only enabled after confirming the user's interest. (2) To facilitate user concentration on primary activities while still being subtly aware of digital alerts, we situated the visual notifications within peripheral vision [7, 16, 18]. Additionally, we employed higher inter-line text spacing [38] to enhance the text readability during mobility, as shown in Figure 1 and Figure 2. (3) To prevent users from being bombarded with notifications, we limit the frequency of sending the same type of notification. Specifically, we set a minimum interval for the same type of suggestions within a similar FPV ($threshold = base\_threshold\ (15s) + (FPV\_similarity)^2 \times threshold\_factor\ (200s)$). (4) To ensure users remain engaged in the present experience, certain notifications are deferred [3] to less interesting moments. For instance, gaze fixation-based suggestions are deferred until a transition [14] during the trip.

### 2.2.2 *Mixed Initiation for Moment Capture: Human Initiation.*

We incorporate human initiation using subtle ring interaction [7, 31] to complement AI initiation, especially when AI might not detect user

[2] The ***Final*** *System* was developed after one iteration with user testing [6].

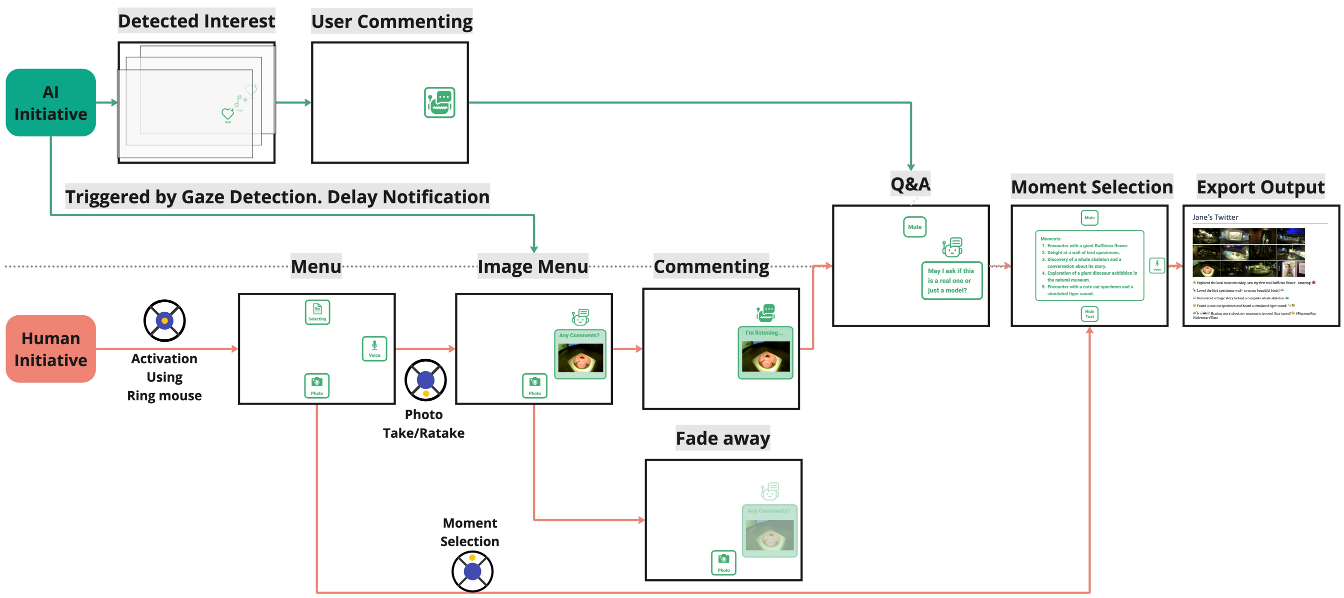


**Figure 2: Interaction flow of *PANDALens* system. It includes both AI-initiative and Human-Initiative interactions. The ring mouse for human-controlled interaction is also shown (yellow dots presenting button clicks). Note: Icons are re-scaled to make the figure clear.**

interest. Adopting the attention-maintaining interface design of *ParaGlassMenu* [7, 18], our design enables users to remain engaged in their travel activities while leveraging their peripheral vision for menu manipulation. By default, the menu is hidden to reduce disruptions. As illustrated in Figure 2 (Human Initiative-Menu), users can activate the menu by pressing the center button on the ring mouse. Following natural spatial mapping guidelines [24] to minimize cognitive effort, users can utilize the up, down, and right, buttons to generate final writing, take photos, or record voice comments, respectively. Moreover, proficient users can snap photos directly via the ring mouse's down button as a shortcut, bypassing menu activation. Options for photo retakes are provided, enabling refining captures. Once a photo is taken, the system displays a notification consisting of an image and comment invitation. Mirroring the AI initiation process (sec 2.2.1), the system anticipates user voice commenting and fades the notification if left unattended after 8 seconds.

*2.2.3 Processing of Interested Moments.* Upon users confirming interesting moments via comments, *PANDALens* transcribes the user's voice into text. As shown in Figure 3 (Contextual Information), these transcriptions are then sent to the LLM, enriched with additional contextual modalities in textual formats using various AI models (detailed in Appendix A-Table 1), including image descriptions, visual objects' labels, text recognized from images (OCR), timestamps, location, and background audio category. This facilitates 1) presenting context-relevant questions to users for inspiration and 2) creating a concise moment summary that eventually contributes to the final narratives.

*Context-Related Questions for Inspiration.* Leveraging the aforementioned multimodal context, the LLM employs a predefined prompt to pose context-specific questions tailored to the user's preferred style (e.g., 'question links to memories'). User preferences regarding question formats are pre-configured (Appendix A-Figure 3-green highlighted parts) and summarized by another LLM model, which first queries users for their preferred style and offers examples for decision support when user preferences are unclear (e.g., a question style example provided by the LLM is: *Specific and Detailed: "Can you describe the flavors and aromas of your coffee? How did they contribute to your overall experience?"*).

To balance inspiration and potential distraction, we limit the number of questions posed for each moment to two, as suggested by users. Regarding the notification modality of these queries, our system integrates both automatic and manual toggling between auditory and visual feedback to ensure a balance between noticeability and minimal distractions. Automatic modality toggles are environmentally dependent; for instance, a scene with many nearby individuals in the FPV prompts auditory rather than visual feedback to preserve the user's visual focus. Concurrently, manual modality adjustments using the ring mouse, such as muting or unmuting notifications, are also available.

*Prompt Design: Processing Interesting Moments for High-Quality Questions and Final Narratives.* To ensure a comprehensive understanding of user travel experiences, interactions with LLM maintain the chat memories, including previous contextual and Q&A details in the same travel session (See Appendix A-Figure 3, Interaction flow with LLM). However, two primary issues were encountered during LLM data processing: 1) the LLM asked irrelevant questions due to overlooking important context that contains unclear or erroneous information (e.g., voice comments with errors like 'Seeshell Potoms' ['Seashell Patterns']), and 2) it produced unsatisfactory final narratives from lengthy, unstructured chat histories (e.g., voice transcription errors preserving in final narratives while user elaborations on questions are missing). To mitigate these challenges, we iteratively refined the prompts for LLM (Appendix A-Figure 3).

To address the first challenge, the refined prompts require the LLM to correct inaccuracies using multimodal information before generating context-relevant questions (detailed in the Authoring Mode Task Description in Appendix A-Figure 3). This approach reduced unsatisfied questions and enabled a more accurate understanding of the environment and user intentions. For example, instead of ignoring 'Seeshell Potoms,' the refined prompt enabled the LLM to accurately understand it with the museum's multimodal context and inquire about captivating aspects of the seashell pattern.

We adopt an approach similar to Chain-of-Thought [35] to address the second challenge. Rather than prompting the LLM to generate final narratives directly from an unstructured chat history, the prompt first instructs the LLM to craft a summary for each distinct moment, accompanied by every in-situ question generation. These summaries can be dynamically enriched or corrected based on users' responses regarding specific moments. For example, a moment summary first accurately recorded a plant name as 'Rafflesia' instead of 'Raising' from voice transcription. Then, it updated details on how the plant's structure enables its regeneration after a fire disaster using user responses to questions. Ultimately, the LLM model generates the final narrative using these refined momentary summaries.

*2.2.4 Generation of Final Writing.* Post-travel, users can compose their travel blogs by selecting which captured moments to incorporate (see Figure 2, Moment Selection and Export Output). The LLM provides a concise summary for each recorded moment, facilitating users in choosing different moment combinations for diverse narratives. After moment selection, the LLM crafts the complete narrative based on a predetermined personalized prompt (Appendix A-Figure 3-green highlighted parts)[3]. Recognizing that preferences may change over time, users can modify the writing style or other narrative adjustments through voice commands. Ultimately, the system offers the final draft narrative in Microsoft Word format, facilitating various sharing options, including social media posts. In addition, to satisfy the comprehensive reviewing needs, the system attaches all the moment summaries to the end of the documentation.

## 3 CONCLUSION

We explored the integration of OHMD interactions with a proactive AI assistant equipped with a multimodal context analyzer and the LLM pipeline. This facilitates in-context writing during travel, transforming a passive tool into a travel companion. We have open-sourced this project at https://github.com/Synteraction-Lab/PANDALens, and welcome contributions from the community to expand its usage scenarios. Future work could focus on developing a general AI assistant capable of processing multimodal contexts and auto-generating documents in various application scenarios, such as creating news reports or summarizing presentations at a conference.

[3]Mirroring the approach for setting question preferences, user preferences for writing styles are preconfigured using an LLM with a separate prompt.

## ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore, under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-016). It is also supported in part by the Ministry of Education, Singapore, under its MOE Academic Research Fund Tier 2 programme (MOE-T2EP20221-0010), and by a research grant #22-5913-A0001 from the Ministry of Education of Singapore. Additionally, the CityU Start-up Grant also provides partial support. Any opinions, findings and conclusions, or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation or the Ministry of Education, Singapore.

We extend our gratitude to the staff at the Lee Kong Chian Natural History Museum, Singapore, especially Dr. Tan Swee Hee, for their invaluable assistance with our user studies.

## REFERENCES

[1] J.E. Allen, C.I. Guinn, and E. Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications* 14, 5 (1999), 14–23. https://doi.org/10.1109/5254.796083

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Christoph Anderson, Isabel Hübener, Ann-Kathrin Seipp, Sandra Ohly, Klaus David, and Veljko Pejovic. 2018. A Survey of Attention Management Systems in Ubiquitous Computing Environments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 2 (July 2018), 1–27. https://doi.org/10.1145/3214261

[4] M. Blum, A. Pentland, and G. Troster. 2006. InSense: Interest-Based Life Logging. *IEEE MultiMedia* 13, 4 (2006), 40–48. https://doi.org/10.1109/MMUL.2006.87

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165

[6] Runze Cai, Nuwan Janaka, Yang Chen, Lucia Wang, Shengdong Zhao, and Can Liu. 2024. PANDALens: Towards AI-Assisted In-Context Writing on OHMD During Travels. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3613904.3642320

[7] Runze Cai, Nuwan Janaka, Shengdong Zhao, and Minghui Sun. 2023. ParaGlassMenu: Towards Social-Friendly Subtle Interactions in Conversations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 721, 21 pages. https://doi.org/10.1145/3544548.3581065

[8] Lalith Chandralal, Jennifer Rindfleish, and Fredy Valenzuela. 2015. An Application of Travel Blog Narratives to Explore Memorable Tourism Experiences. *Asia Pacific Journal of Tourism Research* 20, 6 (2015), 680–693. https://doi.org/10.1080/10941665.2014.925944

[9] Yuhu Chang, Yingying Zhao, Mingzhi Dong, Yujiang Wang, Yutian Lu, Qin Lv, Robert P. Dick, Tun Lu, Ning Gu, and Li Shang. 2021. 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 (June 2021), 56:1–56:23. https://doi.org/10.1145/3463509

[10] Elizabeth Clark, Anne Spencer Ross, Chenhao Tan, Yangfeng Ji, and Noah A. Smith. 2018. Creative Writing with a Machine in the Loop: Case Studies on Slogans and Stories. In *23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 329–340. https://doi.org/10.1145/3172944.3172983

[11] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, Cheng Peng, Chris Sweeney, Cole Wilson, Dan Barnes, Daniel DeTone, David Caruso, Derek Valleroy, Dinesh Ginjupalli, Duncan Frost, Edward Miller, Elias Mueggler, Evgeniy Oleinik, Fan Zhang, Guruprasad Somasundaram, Gustavo Solaira, Harry Lanaras, Henry Howard-Jenkins, Huixuan Tang, Hyo Jin Kim, Jaime Rivera, Ji Luo, Jing Dong, Julian Straub, Kevin Bailey, Kevin Eckenhoff, Lingni Ma, Luis Pesqueira, Mark Schwesinger, Maurizio Monge, Nan Yang, Nick Charron, Nikhil Raina, Omkar Parkhi, Peter Borschowa, Pierre Moulon, Prince Gupta, Raul Mur-Artal, Robbie Pennington, Sachin Kulkarni, Sagar Miglani, Santosh Gondi, Saransh Solanki, Sean Diener, Shangyi Cheng, Simon Green, Steve Saarinen, Suvam Patra, Tassos Mourikis, Thomas Whelan, Tripti Singh, Vasileios Balntas, Vijay Baiyya, Wilson Dreewes, Xiaqing Pan, Yang Lou, Yipu Zhao, Yusuf Mansour, Yuyang Zou, Zhaoyang Lv, Zijian Wang, Mingfei Yan, Carl Ren, Renzo De Nardi, and Richard Newcombe. 2023. Project Aria: A New Tool for Egocentric Multi-Modal AI Research. arXiv:2308.13561 [cs.HC]

[12] Niaja Farve, Tal Achituv, and Pattie Maes. 2016. User Attention with Head-Worn Displays. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems - CHI EA '16*. ACM Press, Santa Clara, California, USA, 2467–2473. https://doi.org/10.1145/2851581.2892530

[13] Jennifer Gluck, Andrea Bunt, and Joanna McGrenere. 2007. Matching Attentional Draw with Utility in Interruption. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '07)*. Association for Computing Machinery, New York, NY, USA, 41–50. https://doi.org/10.1145/1240624.1240631

[14] Joyce Ho and Stephen S. Intille. 2005. Using context-aware computing to reduce the perceived burden of interruptions from mobile devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '05*. ACM Press, Portland, Oregon, USA. https://doi.org/10.1145/1054972.1055100

[15] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 159–166. https://doi.org/10.1145/302979.303030

[16] Nuwan Janaka. 2023. Minimizing Attention Costs of Visual OHMD Notifications. IEEE Computer Society, 136–140. https://doi.org/10.1109/ISMAR-Adjunct60411.2023.00036

[17] Nuwan Janaka, Jie Gao, Lin Zhu, Shengdong Zhao, Lan Lyu, Peisen Xu, Maximilian Nabokow, Silang Wang, and Yanch Ong. 2023. GlassMessaging: Towards Ubiquitous Messaging Using OHMDs. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 3 (Sept. 2023). https://doi.org/10.1145/3610931

[18] Nuwan Janaka, Chloe Haigh, Hyeongcheol Kim, Shan Zhang, and Shengdong Zhao. 2022. Paracentral and near-peripheral visualizations: Towards attention-maintaining secondary information presentation on OHMDs during in-person social interactions. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3491102.3502127

[19] Nuwan Janaka, Shengdong Zhao, and Shardul Sapkota. 2023. Can Icons Outperform Text? Understanding the Role of Pictograms in OHMD Notifications. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 575, 23 pages. https://doi.org/10.1145/3544548.3580891

[20] Hyeongcheol Kim, Shengdong Zhao, Can Liu, and Kotaro Hara. 2020. LiveSnippets: Voice-Based Live Authoring of Multimedia Articles about Experiences. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (Oldenburg, Germany) *(MobileHCI '20)*. Association for Computing Machinery, New York, NY, USA, Article 31, 11 pages. https://doi.org/10.1145/3379503.3403556

[21] Rebecca Krosnick, Fraser Anderson, Justin Matejka, Steve Oney, Walter S. Lasecki, Tovi Grossman, and George Fitzmaurice. 2021. Think-Aloud Computing: Supporting Rich and Low-Effort Knowledge Capture. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3411764.3445066

[22] Amel Ksibi, Ala Saleh D. Alluhaidan, Amina Salhi, and Sahar A. El-Rahman. 2021. Overview of Lifelogging: Current Challenges and Advances. *IEEE Access* 9 (2021), 62630–62641. https://doi.org/10.1109/ACCESS.2021.3073469

[23] Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive Conversational Agents in the Post-ChatGPT World. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Taipei, Taiwan) *(SIGIR '23)*. Association for Computing Machinery, New York, NY, USA, 3452–3455. https://doi.org/10.1145/3539618.3594250

[24] Donald A. Norman. 2013. *The design of everyday things* (revised and expanded edition ed.). Basic Books.

[25] Blaine A. Price, Avelie Stuart, Gul Calikli, Ciaran Mccormick, Vikram Mehta, Luke Hutton, Arosha K. Bandara, Mark Levine, and Bashar Nuseibeh. 2017. Logging You, Logging Me: A Replicable Study of Privacy and Sharing Behaviour in Groups of Visual Lifeloggers. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 22 (jun 2017), 18 pages. https://doi.org/10.1145/3090087

[26] Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

[27] Valentin Radu, Catherine Tong, Sourav Bhattacharya, Nicholas D. Lane, Cecilia Mascolo, Mahesh K. Marina, and Fahim Kawsar. 2018. Multimodal Deep Learning for Activity and Context Recognition. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 157 (jan 2018), 27 pages. https://doi.org/10.1145/3161174

[28] K. Ann Renninger and Suzanne Hidi. 2015. *The Power of Interest for Motivation and Engagement.* Routledge, New York. https://doi.org/10.4324/9781315771045

[29] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[30] Jeba Rezwana and Mary Lou Maher. 2022. Designing Creative AI Partners with COFI: A Framework for Modeling Interaction in Human-AI Co-Creative Systems. *ACM Transactions on Computer-Human Interaction* (Feb. 2022). https://doi.org/10.1145/3519026

[31] Shardul Sapkota, Ashwin Ram, and Shengdong Zhao. 2021. Ubiquitous Interactions for Heads-Up Computing: Understanding Users' Preferences for Subtle Interaction Techniques in Everyday Settings. In *Proceedings of the 23rd International Conference on Mobile Human-Computer Interaction (MobileHCI '21)*. Association for Computing Machinery, New York, NY, USA, Article 36, 15 pages. https://doi.org/10.1145/3447526.3472035

[32] Abigail J. Sellen and Steve Whittaker. 2010. Beyond total capture: a constructive critique of lifelogging. *Commun. ACM* 53, 5 (May 2010), 70–77. https://doi.org/10.1145/1735223.1735243

[33] Paul J Silvia. 2006. *Exploring the psychology of interest.* Psychology of Human Motivation.

[34] Dan Wang, Zheng Xiang, and Daniel R. Fesenmaier. 2014. Adapting to the mobile world: A model of smartphone use. *Annals of Tourism Research* 48 (2014), 11–26. https://doi.org/10.1016/j.annals.2014.04.008

[35] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://openreview.net/forum?id=_VjQlMeSB_J

[36] Xuhai Xu, Anna Yu, Tanya R. Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, Te-Yen Wu, Junxiao Shen, Ting Zhang, Narine Kokhlikyan, Fulton Wang, Paul Sorenson, Sophie Kim, and Hrvoje Benko. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–30. https://doi.org/10.1145/3544548.3581500

[37] Shengdong Zhao, Felicia Tan, and Katherine Fennedy. 2023. Heads-Up Computing Moving Beyond the Device-Centered Paradigm. *Commun. ACM* 66, 9 (aug 2023), 56–63. https://doi.org/10.1145/3571722

[38] Chen Zhou, Katherine Fennedy, Felicia Fang-Yi Tan, Shengdong Zhao, and Yurui Shao. 2023. Not All Spacings Are Created Equal: The Effect of Text Spacings in On-the-Go Reading Using Optical See-Through Head-Mounted Displays. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 720, 19 pages. https://doi.org/10.1145/3544548.3581430

# A IMPLEMENTATION

The *PANDALens* system is developed with the OHMD, XREAL Air[4], for a near-eye display and uses the Pupil Core add-on for gaze detection and FPV streaming. Implementation can be found at https://github.com/Synteraction-Lab/PANDALens. Built on a TKinter-based UI and a Python backend, it seamlessly handles the real-time capture and concurrent processing of various context data and user interaction. Due to computational constraints, our choice of context analysis models aimed to balance performance and efficiency, especially in mobile scenarios without constant power sources. We employed the GPT3.5-Turbo-16K model as the primary LLM to generate context-specific questions and structure narratives. Few-shot prompts (i.e., Auto Mode Selection in Figure 3) enabled LLM to discern whether to generate questions, compile a moment selection list, or create a full blog entry

[4]https://www.xreal.com/air

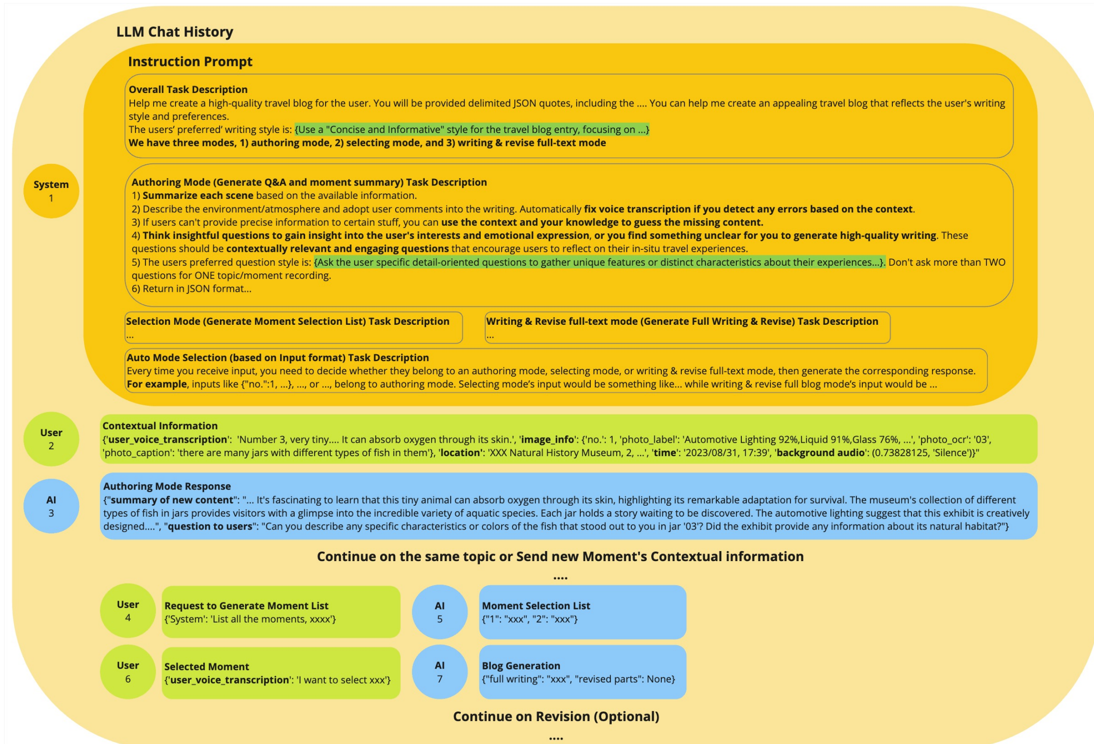


**Figure 3: LLM Chat History for the *PANDALens* system. 'System' represents the initial prompts directing the LLM's tasks. In the Instruction Prompt, sections highlighted in green are customized parts tailored to individual preferences and generated by another LLM. 'User' and 'AI' signify the inputs and outputs within the LLM dialogue, respectively, with message sequencing indicated numerically. Note: Some details are redacted to conserve space and preserve anonymity.**

**Table 1: System Components and Associated Technologies.**

| Component | Description | Associated Technologies/Tools |
|---|---|---|
| ***PANDALens*** | Main system developed for the application. | Python 3.9 |
| **OHMD UI** | Interfaces built on a laptop for near-eye display. | Tkinter |
| **Pupil Core** | Facilitates gaze detection and FPV video streaming. | Socket connection in Python with Pupil Capture App |
| **Multimodal Analyzer** | Analyzes multimodal context data concurrently and integrates contextual information in JSON format. | 1. Object Detection & OCR: YOLO v8, Google Cloud Vision API<br>2. Image Description: BLIP-large on Hugging Face<br>3. FPV Similarity: OpenCV<br>4. Audio Classification: MediaPipe<br>5. Voice Transcription: Whisper<br>6. Tone Analysis: Emotion English DistilRoBERTa-base model<br>7. Location: Geopy, Geocoder.<br>8. Time: Python's Datetime. |
| **LLM Model** | Processes context data to provide questions and assist writing. | GPT3.5-Turbo-16K (temperature value: 0.3) |
| **Prompt Engineering** | Ensures efficient task performance and seamless integration. | 1. Clear and Specific Instructions,<br>2. Few-shot prompts,<br>3. JSON formatted responses,<br>4. Chain-of-Thought approach |

based on the input format, and additional prompt engineering techniques [29, 35] were utilized to enhance its output. Our system compresses chat history into summaries to address the LLM's token limitations, facilitating longer documentation sessions. Comprehensive implementation details can be referred to in Table 1 and https://github.com/Synteraction-Lab/PANDALens.

Received 17 January 2024; accepted 16 February 2024