



Less or More: Towards Glanceable Explanations for LLM Recommendations Using Ultra-Small Devices

Xinru Wang*
Department of Computer Science
Purdue University
West Lafayette, Indiana, USA
xinruw@purdue.edu

Michael Iuzzolino
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
mliuzzolino@meta.com

Natasha Lynova
Meta Inc.
Redmond, Washington, USA
natashalynova@meta.com

Naveen Sendhilnathan
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
naveensendhilnathan@gmail.com

Mengjie Yu
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
annaymj@meta.com

Tianyi Wang
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
tianyiwang@meta.com

Co Tran
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
co.quoc.tran.2@gmail.com

Hrvoje Benko
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
benko@meta.com

Tanya R. Jonker
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
tanya.jonker@meta.com

Hannah Nguyen
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
hannahnguyen@meta.com

Peiqi Tang
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
peiqi@meta.com

Ting Zhang
Reality Labs Research
Meta Inc.
Redmond, Washington, USA
tingzhang@fb.com

Haijun Xia
Department of Cognitive Science and
Design Lab
University of California, San Diego
San Diego, California, USA
haijunxia@ucsd.edu

Abstract

Large Language Models (LLMs) have shown remarkable potential in recommending everyday actions as personal AI assistants, while Explainable AI (XAI) techniques are being increasingly utilized to help users understand why a recommendation is given. Personal AI assistants today are often located on ultra-small devices such as smartwatches, which have limited screen space. The verbosity of LLM-generated explanations, however, makes it challenging to deliver glanceable LLM explanations on such ultra-small devices. To address this, we explored 1) *spatially structuring* an LLM's explanation text using defined contextual components during prompting and 2) presenting *temporally adaptive* explanations to users based

on confidence levels. We conducted a user study to understand how these approaches impacted user experiences when interacting with LLM recommendations and explanations on ultra-small devices. The results showed that structured explanations reduced users' time to action and cognitive load when reading an explanation. Always-on structured explanations increased users' acceptance of AI recommendations. However, users were less satisfied with structured explanations compared to unstructured ones due to their lack of sufficient, readable details. Additionally, adaptively presenting structured explanations was less effective at improving user perceptions of the AI compared to the always-on structured explanations. Together with users' interview feedback, the results led to design implications to be mindful of when personalizing the content and timing of LLM explanations that are displayed on ultra-small devices.

*Work done during an internship at Meta Reality Labs.



This work is licensed under a Creative Commons Attribution 4.0 International License.
IUI '25, Cagliari, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1306-4/25/03
<https://doi.org/10.1145/3708359.3712074>

CCS Concepts

• **Human-centered computing** → **Empirical studies in interaction design**; • **Computing methodologies** → *Artificial intelligence*.

Keywords

Explainable AI, ultra-small devices, large language models, structured prompting, confidence estimation, smartwatches

ACM Reference Format:

Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, Hrvoje Benko, Haijun Xia, and Tanya R. Jonker. 2025. Less or More: Towards Glanceable Explanations for LLM Recommendations Using Ultra-Small Devices. In *30th International Conference on Intelligent User Interfaces (IUI '25)*, March 24–27, 2025, Cagliari, Italy. ACM, New York, NY, USA, 14 pages. <https://doi.org/10.1145/3708359.3712074>

1 Introduction

Artificial Intelligence (AI) advancements over the past decades have led to the development of AI-based personal assistants, which have become a popular way for human users to interact with digital devices and data. More recently, Large Language Models (LLMs) such as the GPT [1] and LLaMa [57] series have shown remarkable performance in generation tasks, which has led to their growing use as a method to make everyday contextual recommendations for AI-based personal assistants [11, 33]. In addition, to ensure transparency and control during human-AI interaction, there has been a growing interest in Explainable AI (XAI) techniques that can help users understand and trust the AI assistance they receive [48, 49, 66, 72, 78]. In the context of LLMs, those models based on transformer architectures are capable of generating explanations that rationalize their reasoning process using natural language [27, 69].

An emerging manifestation of AI-based personal assistant has been on ultra-small devices, such as smartwatches, smartglasses [19, 44], and on-cloth gadgets [21], to offer low-friction assistance that is seamlessly integrated into the physical world. These devices rely on vast amount of user contextual data gathered from their sensors (e.g., egocentric video, gaze tracking, and audio) to predict the goal of the user's next action and generate action recommendations [23], such as launching an application, to help users efficiently complete daily tasks.

Interacting with LLM recommendations and explanations on such ultra-small devices can, however, be challenging for several reasons. First, LLM self-explanations are often verbose and time-consuming to comprehend [22], while the constrained screen space on ultra-small devices exacerbates the challenge of displaying verbose explanations. Real-world examples include smartwatches or Augmented Reality (AR) glasses, which have limited screen real estate that is available, and struggle to display long messages [14]. Designers need to carefully condense the *content* of an explanation to ensure it remains relevant and informative while not overloading the limited screen space. Additionally, some tasks, such as providing route adjustment recommendations while driving, are inherently time-sensitive and consume a lot of the user's cognitive resources.

Thus, it is crucial to optimize the *timing* of explanations by presenting them only at moments when they would be most useful or when a user requests them to prevent unnecessary user interruptions [72].

To overcome these challenges, this work aims to design glanceable explanations for a goal-oriented, contextual LLM action recommendation system for end-users using ultra-small devices. We focused on improving *spatial* and *temporal* glanceability via two research questions:

- **RQ1:** To improve the spatial glanceability of LLM explanations on ultra-small devices, how should LLM explanation text be *structured*?
- **RQ2:** To take a step beyond and further improve the temporal glanceability of LLM explanations on ultra-small devices, how should LLM explanations be *adaptively* presented?

To answer these questions, we designed a web-based interface that enabled us to simulate an LLM recommending and explaining actions to a user when they were using an AI-based personal assistant embedded in a smartwatch, and we recruited participants to interact with this interface. A subset of Ego4D [20] videos featuring digital interactions in people's daily activities was selected for evaluation. We used a learning approach based on Socratic Models [76] with the dataset, where pre-trained vision-language models were used to generate a linguistic summarization (e.g., user physical actions and detected objects) of a video input for downstream processing with an LLM, which then made inferences and generated action recommendations.

To improve spatial glanceability (**RQ1**), we converted an LLM's verbose explanation text into concise concepts and intuitive icons that users could grasp at a glance. We adopted a Chain-of-Thought prompt strategy [27, 69] and broke down the inference process into steps. An LLM first summarized all possible contexts (i.e., the *[activity]* the user is doing, the *[object]* the user is interacting with, and the *[location]* the user is in) from the output of the vision-language models. Then, the LLM inferred the short-term *[goal]* that the user wanted to achieve and provided a digital action recommendation that the user could take as a next step. Instead of displaying lengthy text, the explanation only included the four structured components.

To improve temporal glanceability (**RQ2**), we focused on situations when the intelligent system became uncertain and needed user confirmation [72]. We employed a hybrid method [6, 71] that combined the consistency among multiple responses and the textual confidence generated by the model to extract the confidence level for each recommendation. Explanations were automatically presented whenever the recommendation confidence was low, otherwise they were only displayed upon the user's request.

We then conducted a user study with 44 participants to understand users' experiences when provided with structured and adaptively presented explanations for LLM-driven contextual recommendations. The study was designed with four within-subject conditions: no explanations, always-on unstructured explanations, always-on structured explanations, and adaptive structured explanations whose presentation was dependant on the recommendation confidence level. We found that structured explanations effectively reduced participants' time to select an AI-recommended action and lowered their cognitive load when reading the explanation.

Participants were also more likely to accept the recommendations when presented with always-on structured explanations. Participants were, however, less satisfied with structured explanations compared to unstructured explanations. When presented adaptively, the structured explanations were less effective at improving user perceptions of the AI (e.g., trust, satisfaction) compared to the always-on structured explanations. Our analysis of participants' feedback from a semi-structured interview showed that they valued the ease of reading and viewing text in the structured explanations, but also desired the level of detail and naturalness as provided by the unstructured explanations. Additionally, while participants appreciated the control over interface offered by adaptively presented explanations, it introduced extra efforts to interact with the interface.

Taken together, we make the following contributions in this work:

- A prompting strategy that structures an LLM's explanation text into defined contextual components to create spatially glanceable LLM explanations on ultra-small devices.
- The use of confidence level estimations to adaptively present explanations to create temporally glanceable LLM explanations on ultra-small devices.
- A user study to analyze the impact of structuring and adaptively presenting explanations on user experience with AI using ultra-small devices. We conclude by discussing design implications for optimizing the content and timing of glanceable LLM explanations.

2 Related Work

Our work builds upon three areas of relevant literature, i.e., existing XAI techniques to explain LLMs, empirical studies on AI explanations, and user interface design challenges for ultra-small devices.

2.1 Techniques to Explain Large Language Models

A variety of XAI techniques have been developed to increase the interpretability of AI models. Conventional XAI methods can be categorized into model-specific methods tailored to particular model types, and model-agnostic methods applicable across various model types. Training interpretable models such as rule-based models and generalized additive models [24, 30, 62] are examples of model-specific methods. On the other hand, examples of model-agnostic methods include generating feature contributions [36, 49, 52] or searching for prototypes or counterfactual instances [26, 61].

The emergence of LLMs raised unique challenges when applying traditional XAI methods to explain them due to the size of their training datasets and non-deterministic outputs [34]. Open-source LLMs like Meta's LLaMA models [57] provide broader transparency for AI researchers to uncover their internal states. For instance, [60] identified when neurons within LLMs were activated, while [81] analyzed how high-level cognitive processes were represented in LLM neurons. [79] explored how LLMs encoded knowledge in their embeddings. On the other hand, closed-source LLMs, such as GPT [1] and Google Gemini [56], pose more challenges due to the restricted access to their internals. Regardless of the closed or open

nature of LLMs, one common approach to generate explanations is to prompt the model to provide "self-rationalizations" alongside predictions [27, 31, 69, 73]. As these LLMs are based on transformer architectures, they can capture complex relationships between input data and generate output through the attention mechanism, to self-rationalize their recommendations in a way that appears logical to humans. Yet, some work has criticized these self-rationalized explanations as being unreliable and unfaithful to the actual reasoning process [58, 74]. To address these limitations, researchers have pursued approaches including iteratively prompting the LLM for self-feedback and refinement [12, 40], and grounding an LLM in external knowledge bases [7, 80].

While the natural language modality of LLMs' responses opens up opportunities for explanations that are understandable to non-experts, they are often too verbose to be grasped in a quickly digestible manner. This issue is particularly pronounced on ultra-small devices, and our work aims to address it.

2.2 User-Centered Design and Evaluation of AI Explanations

Empirical research has focused on understanding how users understand AI explanations [66, 67] and trust AI models [67, 68, 78], as well as users' overall task performance [3, 29, 35, 63, 64]. This research has also spurred novel design processes that identify users' explainability needs to inform XAI design (e.g., identifying the questions that users commonly ask AI systems [48]). In addition to explaining traditional machine learning models, more recently, user-centered design approaches have been applied to explain large generative AI models. For example, Sun et al. [55] identified explainability needs unique to generative AI for code compared to traditional discriminative machine learning models, and proposed four XAI features: AI documentation, uncertainty indicators, attention visualizations, and social transparency.

On the other hand, research on designing and evaluating XAI methods for particular digital devices is lacking. Developing AI explanations for specific interfaces often involves trade-offs between usability and transparency. Springer and Whitaker [54] highlighted that transparency can be distracting and that users might benefit from initially simplified explanations. Chromik and Butz [8] surveyed a list of XAI publications and identified design principles for human interaction with explanation user interfaces, including naturalness, responsiveness, flexibility, and sensitivity. Xu et al. [72] proposed design principles for displaying XAI information in floating windows on AR devices. They recommended simplifying explanations by selecting the appropriate content, and triggering explanations only when users have enough capacity, are unfamiliar with the outcome, or when the model is uncertain. Our work builds on this prior research but focus on ultra-small devices, where there is an urgent need for highly glanceable explanations.

2.3 User Interface Design for Ultra-Small Devices

The ultra-small screen sizes on devices such as smartwatches raises key usability issues for the layout of visually rich content such as verbose text [9, 51, 53]. To address these challenges, researchers have recommended design guidelines to condense information

while maintaining clear user understanding [2, 10]. For example, a navigation interface should only display an upcoming route as an arrow [15]. Others have explored adaptive menus and have found that users preferred a cloud menu, where predicted items were arranged in a circular tag cloud [59]. Work by Rahman and Muter showed that presenting continuous text sentence-by-sentence within small display windows offered users the ability to reread one sentence at a time and was found to be as efficient as conventional reading [47]. In summary, these common approaches included converting lengthy text into graphical diagrams or icons [22, 50], adaptively prioritizing highly-relevant content [17, 18], progressively displaying content [25, 47], or utilizing multiple modalities such as voice and gestures [32, 42, 70]. Guided by these existing design ideas, we selected two major dimensions to explore fitting verbose content on ultra-small screens, i.e., *spatially*, by structuring it into easily viewable formats, and *temporally*, by adaptively displaying the content.

3 Generating Glanceable LLM Explanations

In this section, we introduce the dataset and LLM pipeline used for generating recommendations, then outline our approaches to present spatially and temporally glanceable LLM explanations on ultra-small devices. The goal of this pipeline was to imitate a human-AI interface that empowered users to accomplish their daily tasks via a smooth transition to digital actions. The system used observations of longitudinal context as input, which could be any combination of sensors a device had access to, such as egocentric video, biometric data, or digital state information. After predicting the user’s goal for their next action based on the longitudinal context, the system recommended a digital action to fulfill this goal and disassembled it into device-executable instructions. Along with the recommendation, an explanation was provided for users to ensure transparency and trustworthiness with the system.

3.1 Dataset

We used the Ego4D [20] dataset as the testbed to evaluate the LLM pipeline in pre-recorded real-world scenarios. The Ego4D dataset consists of over 3,670 hours of egocentric videos of people’s daily activities. The videos were captured using head-mounted cameras to naturally approximate first-person visual perception.

We leveraged the annotated video narrations included in the dataset to manually curate a subset of videos that were comprised of interactions with digital devices (e.g., “phone” or “laptop”). We trimmed the videos to start 30 seconds prior to the digital action as context for the action and ensured that the video content prior to the digital action contained primarily physical activities, such as cooking or hiking. Following this, we were left with 1101 videos that focused on everyday goal-oriented, contextual scenarios that the LLM could take as input to generate action recommendations. The daily activities covered all 27 categories of common everyday activities identified by a taxonomist who is one of the authors, such as browsing the internet, chores and cleaning, cooking and eating, exercise, work, gaming, messaging and communication, shopping, and traveling.

3.2 LLM Recommendation Pipeline

The pipeline was based on Socratic Models [76], where pre-trained vision-language models generated a linguistic summary of a video for downstream processing with an LLM (Figure 1). Specifically, we employed a pre-trained transformer-based model designed for video understanding to extract physical actions (i.e., narrations) present in the video by dividing each video into 2-second clips. We used a pre-trained transformer-based object detection model to detect objects appearing in the video within a 5-second window before the digital action¹. With these two pieces of contextual information about the video input, GPT-4 was prompted to infer the user’s intent and output its recommendation for the user’s next digital action.

We prompted the LLM to produce natural language explanations along with its recommendations following step-by-step thinking based on the Chain-of-Thought method [27, 69]. As text-based responses from LLMs can be too verbose to be displayed on an ultra-small device’s screen, spatially, we structured the textual responses using defined contextual components to create a visual experience that required minimal cognitive effort to comprehend. Temporally, we controlled the presentation timing of the explanations to avoid unnecessary interruptions to the user. We addressed situations where the AI system showed uncertainty, meaning it might make errors and require user confirmation. The prompt template is shown in Figure 2.

3.2.1 Structuring Explanation Text Using Defined Contextual Components. The inference process was divided into several stages. The LLM first extracted and summarized the relevant contextual information, i.e., the *[activity]* the user was performing, the *[object]* the user was interacting with, and the *[location]* of the user, from the output generated by pre-trained vision-language models. Then, the LLM inferred the user’s short-term *[goal]* based on these contextual cues. Finally, using the inferred goal, the LLM provided a digital action recommendation. We chose “object”, “activity”, “location”, and “goal” as the contextual information as these entities were the core contextual components necessary to generate relevant recommendations and explanations in an everyday mixed-reality scenario, as provided by the aforementioned taxonomist. We also referenced definitions and few-shot examples in the prompt for each component. The LLM was instructed to provide its output in a JSON format, which was later used to create the structured representations.

3.2.2 Adaptively Presenting Explanations Based on Confidence Levels. Following Chen and Muelle [6] and Xiong et al. [71], we employed a hybrid approach that combined consistency-based methods and verbalized confidence to extract a calibrated confidence level for each recommendation. We mapped the confidence score for each detected object from the object detection model and the perplexity score for each detected physical action from the video narration model into one of five textual confidence levels (i.e., “very low”, “low”, “medium”, “high”, “very high”), based on the quantile position of the score. We opted for a textual representation of confidence to ensure a unified representation of “confidence”, as well as to maximize the LLMs’ abilities to comprehend natural language.

¹Note that some of the Ego4D videos did not have audio, so only video frames were used as input.

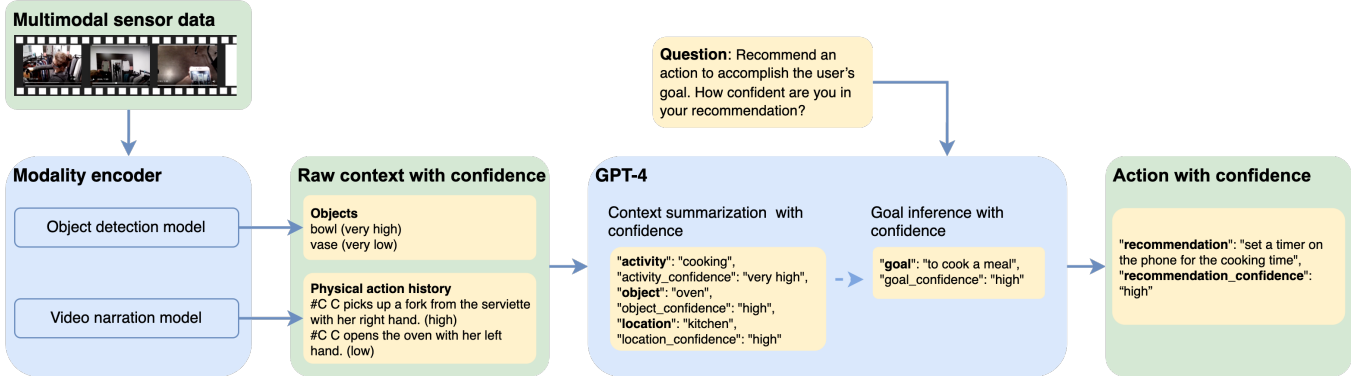


Figure 1: Our LLM pipeline, which was based on Socratic Models, used pre-trained vision-language models to generate a linguistic summary of a video input (detected objects and user physical actions) for downstream processing with an LLM (GPT-4). To obtain the LLM self-explanations, the LLM first summarized all possible contexts (i.e., the [activity] the user is doing, the [object] the user is interacting with, and the [location] the user is in) and then inferred the short-term [goal] that the user may want to achieve. Based on the inferred goal, the LLM then provided a digital action recommendation. Throughout the process, we calculated confidence levels of the output from pre-trained vision-language models. The LLM was then prompted for confidence levels for each contextual component (i.e., [activity], [object], [location], and [goal]) and the recommendation.

We included the confidence levels for raw contextual information in the prompt, then prompted the LLM to output confidence levels along with the recommendation.

We then generated a reference response y_0 and confidence c_0 with temperature sampling set at 0. Using the same prompt, we produced $K = 5$ candidate responses $\{y_1, y_2, \dots, y_K\}$ and their verbalized confidences $\{c_1, c_2, \dots, c_K\}$ by increasing the temperature value to 0.7 [71]. Each candidate confidence c_i ($i \in \{1, 2, \dots, K\}$) was transformed back into a numerical value and was then updated by incorporating the reference answer and the similarity to the reference answer, denoted as s_i : $\bar{c}_i = \frac{(c_0 + c_i)}{2} s_i$ ($i \in \{1, 2, \dots, K\}$). The similarity scores were obtained using the BERTScore model [77], which measured the semantic similarity between two texts. Finally, we computed the average contribution from all candidate answers as the final confidence score $c_{hybrid} = \sum_{i=1}^K \bar{c}_i / K$, and then converted it into one of the five confidence levels based on the quantile position of the score.

3.3 Technical Evaluation

To validate the quality of structured explanation components and confidence levels that the system generated, we conducted a small-scale technical evaluation. We randomly sampled one video from each of the 27 activity context categories. Three coders then coded how likely it was that the AI recommendation was correct, how likely it was that the inference about each explanation components (i.e., activity, object, location, and goal) was correct (i.e., plausibility), and to what extent each explanation component supported the AI recommendation (i.e., faithfulness) [16, 41], for each of the 27 videos using a 7-point Likert scale.

We then calculated the inter-rater agreement across the coders using Krippendorff's α and obtained a score of 0.553, which is considered acceptable for exploratory research involving subjective ratings [28]. We found that all explanation components were rated as plausible and faithful to the recommendation, with the median

Table 1: Ratings for the plausibility and faithfulness of each explanation component.

Explanation Component	Plausibility median (IQR)	Faithfulness median (IQR)
Goal	5.67 (3.50 - 6.67)	6.67 (6.50 - 7.00)
Activity	6.67 (6.67 - 7.00)	6.333 (5.50 - 6.67)
Object	7.00 (6.67 - 7.00)	6.333 (5.00 - 7.00)
Location	6.67 (3.00 - 7.00)	5.000 (3.33 - 6.67)

of all ratings being larger than 4 (one-sample Wilcoxon signed-rank test showed $p < 0.05$ for all explanation components for the plausibility and faithfulness measures; Table 1). To assess the calibration of the recommendation confidence, we computed the Pearson's r correlation coefficient between the coders' estimations of the recommendation correctness likelihood (as ground truth) and the hybrid confidence c_{hybrid} , obtaining a significant positive coefficient of $r = 0.559$ ($p < 0.01$). Meanwhile, the verbalized confidence levels c_0 did not correlate with the coders' estimations ($r = 0.171$, $p = 0.393$), indicating that the hybrid approach could lead to more calibrated confidence.

4 User Study Design

To gain insights into how LLM explanations that differ in their structure and adaptivity influence user experiences, we conducted a lab-based user study. The study measured participants' perceptions of LLM recommendations and explanations while they watched videos of everyday interactions and received recommendations on a simulated smartwatch UI.

4.1 Participants

Forty-four participants (Male = 20, Female = 23, prefer not to say = 1; mean age = 38 years, std = 12.3 years) were recruited to participate

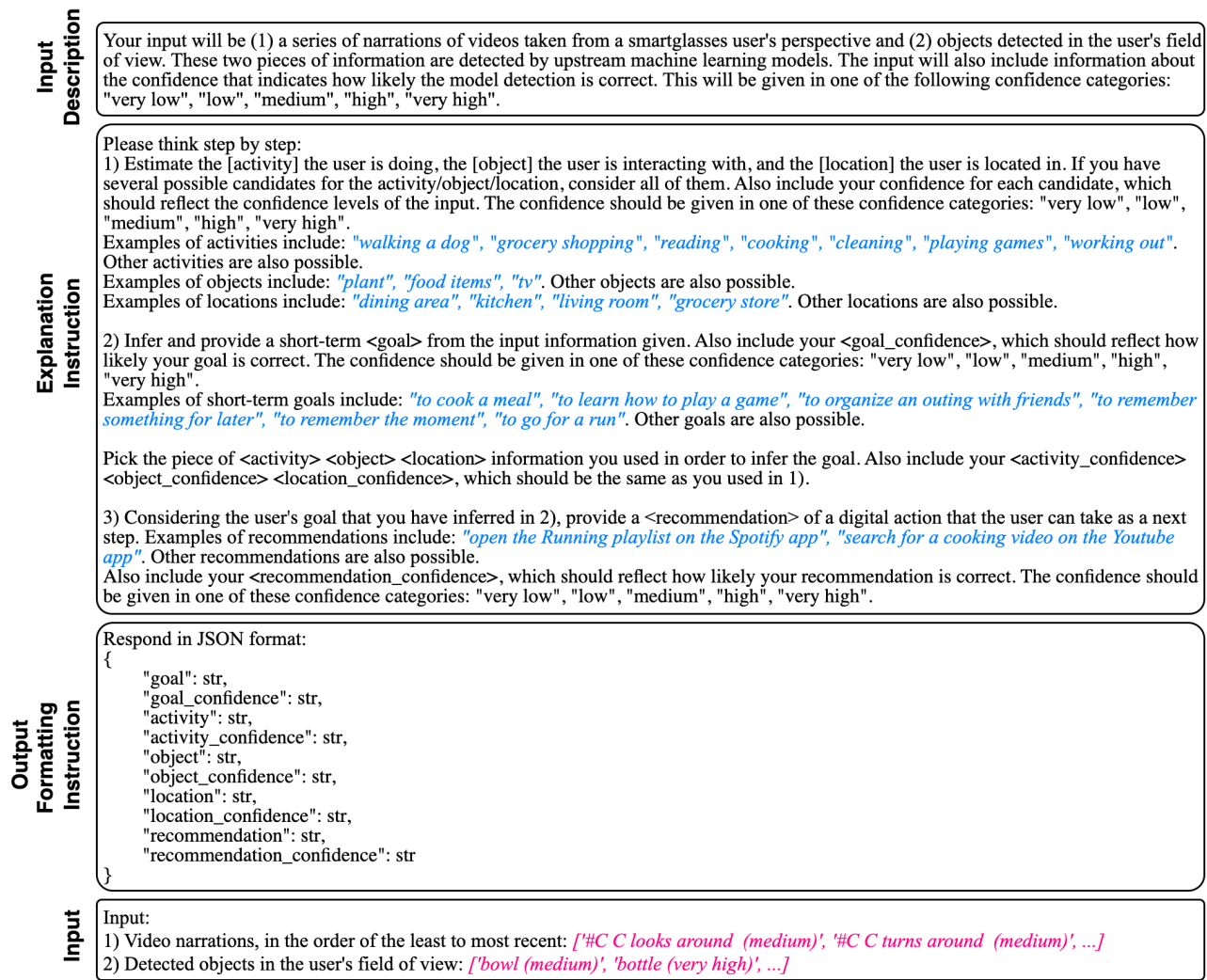


Figure 2: The prompt template included an input description, explanation instructions with few-shot in-context examples (blue text in italics), output formatting instructions, and the target input (pink text in italics, changed for every query).

in our in-person user study, a sample size comparable to prior work [5]. Study participants were recruited through email invitations and social media platforms from an existing participant pool. Twenty participants had a graduate degree, 18 had a Bachelor's degree, 4 had some college experience, and 2 did not have any education past high school. Nine participants worked on AI-powered devices or products, 11 used AI-powered devices or services at least once a week, 9 used them at least once a month, 9 had used them a few times, and 6 had never interacted with AI-based devices or services. The entire study lasted about one hour, and each participant was compensated \$75 upon completion of the study.

4.2 Task

During the study, participants were asked to watch a 30-second Ego4D video on a standard desktop computer and were told that the videos they were watching were captured by a camera embedded in

a pair of smart glasses that were worn on a person's head and naturally approximated the visual field of the wearer of the glasses. We instructed participants to imagine that they were the one wearing these smartglasses and they were performing the activities shown in the video. They were also told that while they were imagining performing these activities, their personal AI assistant would try to predict their next intent and recommend a digital action, which was pushed to their smartwatch. The smartwatch UI was displayed next to the video on the desktop computer. Participants were asked to evaluate whether the recommendation met their needs (i.e., as the wearer of the glasses in the video) and then choose to accept or dismiss the AI's recommendation on the smartwatch UI by clicking on it (Figure 3).

We opted to build a web-based interface that simulated smartwatch UI for two reasons. First, replicating a wide variety of real-life scenarios in a lab environment would be difficult. Instead, the

existing Ego4D dataset provided a diverse set of pre-recorded situations. Second, our recommendation pipeline, which employed large transformer-based models (i.e., vision-language models, GPT-4), had slow response times. Deploying it on a smartwatch for real-time video processing would cause significant latency (~tens of seconds delay), which would have influenced participant satisfaction.

4.3 Conditions

A within-subject design was used with four conditions that manipulated whether and how the explanations were displayed (Figure 4). A balanced 4×4 Latin Squares design was used for condition counterbalancing [39]. A UX designer who is one of the authors standardized the smartwatch UI design using Figma.

- In the **no explanation** condition (Figure 4(a)), participants received recommendations without any explanation on the smartwatch UI.
- In the **always-on unstructured explanation** condition (Figure 4(b)), participants received recommendations along with an unstructured explanation on the smartwatch UI. The number of words in the unstructured explanation were limited to ensure they occupied the same amount of screen space as the structured explanations (i.e., the following two conditions), while allowing participants to scroll up and down to view more text. The prompt used to generate the unstructured explanations can be found in Appendix A.1.
- In the **always-on structured explanation** condition (Figure 4(c)), participants received a recommendation along with a structured explanation that was comprised of four components: a goal, an activity, an object, and a location.
- In the **adaptive structured explanation** condition, structured explanations were provided adaptively depending on the recommendation's confidence level. The explanation was displayed automatically when the recommendation confidence was low (Figure 4(d)), otherwise, it was hidden by default (Figure 4(e)) and a toggle button enabled participants to see it (Figure 4(f)).

For the adaptive structured explanation condition, we categorized recommendation confidence levels into a binary classification as high or low, depending on whether the c_{hybrid} level was "low"/"very low" or "high"/"very high". These categorizations were based on the findings from our technical evaluation (Section 3.3), where we observed a clear gap in the human estimation of the recommendation correctness likelihood based on the confidence levels (c_{hybrid}). When c_{hybrid} was "low" (median correctness rating = 3.5) or "very low" (median correctness rating = 3.0), coders rated the correctness likelihood much lower compared to recommendations with c_{hybrid} being "high" (median rating = 6.0) or "very high" confidence levels (median rating = 5.5). From the same pool of task instances used in the technical evaluation described in Section 3.3, we chose 20 task instances where the recommendation confidence c_{hybrid} was "low"/"very low" and 20 instances where the confidence c_{hybrid} was "high"/"very high". To ensure the representativeness of the selected instances, we sampled all task instances based on the data distributions across the 27 activity context categories in the video dataset.

4.4 Study Procedure

Upon arrival at our lab, participants were asked to complete a consent form and an initial survey about their demographics and experience with AI.

Participants then completed 40 trials, which were divided into four blocks, one block for each condition. Before beginning each block, there were two tutorials that were identical to the actual trials but during which no user data was collected. Each block then consisted of 10 trials, with half of them having low confidence recommendations and half having high confidence recommendations. The order of all 10 trials within each block was randomized across participants. During each trial, participants watched a 30-second video on the desktop computer. At the end of the video, a recommendation for a digital action and an explanation (depending on the assigned condition) was displayed on the simulated smartwatch UI. Participants then chose whether to accept or dismiss the AI's recommendation. After each block, participants completed a survey about their experience with the AI's recommendations and explanations.

Upon completion of all four blocks, participants completed a final survey and answered interview questions to understand what they liked or disliked about each AI explanation condition, when they would like to see an explanation, what information they would like to see in the explanation, and how the AI explanation could be improved to better meet their needs.

4.5 Metrics

Several metrics were computed to understand participants' experiences and perceptions when presented with the AI explanation conditions. First, we computed the following objective measures:

- **Time to action:** The time taken by a participant from the end of the video until they accepted or dismissed the AI recommendation.
- **Acceptance rate:** The percentage of tasks where the participant chose to accept the AI recommendation

At the end of each block, we asked participants about the following subjective measures using 7-point Likert scales:

- **Mental load:** We measured participants' mental load while making the decision, understanding the recommendation, and reading the explanation via three questions:
 - "Overall, how much mental effort did you spend on deciding whether to accept or dismiss the AI recommendation?"
 - "Overall, how much mental effort did you spend on understanding how the AI model makes recommendations based on the context in the video?"
 - "How much mental effort did you spend on reading the AI model's explanation of the recommendation?"
- **Trust in AI:** Participants' trust in AI was calculated by the average rating of the following three questions:
 - "I have faith that the AI model would be able to cope with all different situations."
 - "I am confident that the AI model can make good recommendations."
 - "Recommendations made by the AI model are likely to be reliable."

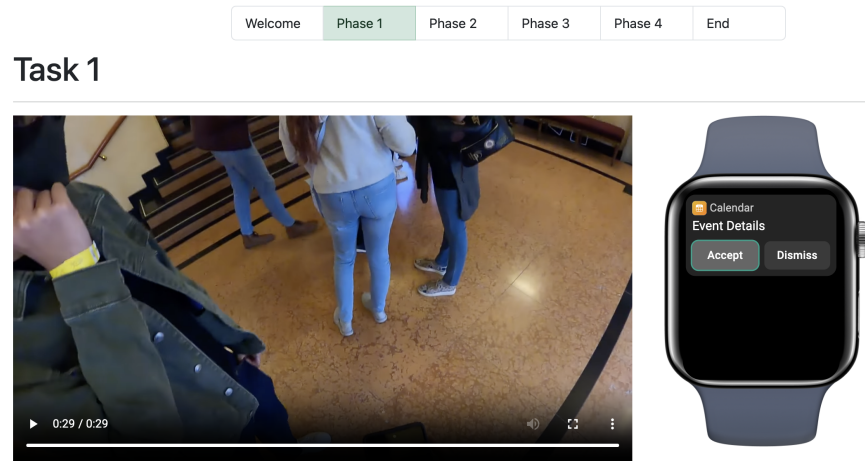


Figure 3: The interface that participants saw on the desktop computer. On the left, they could watch the 30-second Ego4D video. They could chose whether to accept or dismiss the AI's recommendation on the smartwatch UI on the right.



Figure 4: The Smartwatch UI designs for the four conditions in the user study. (d)-(f) illustrate different variations of the adaptive structured explanation condition.

- **Understanding of AI:** This was also computed by averaging participants' responses to three questions:
 - “I understand how the AI model works to predict my next digital action based on the context in the video.”

- “I can predict how the AI model will behave based on the context in the video.”
- “I feel the AI model is transparent in communicating how the recommendations were made based on the context in the video.”

- **Satisfaction with AI:** “I’m satisfied with the AI model’s recommendation.”
- **Satisfaction with AI explanations:** This metric differed from “satisfaction with AI” as it specifically measured how participants felt about the explanation provided by the AI for its recommendations, i.e., “I’m satisfied with the AI model’s explanation of the recommendation.”

At the end of all four blocks, participants also ranked the four conditions based on their **overall preferences** for them.

4.6 Analysis Methods

Mixed-effect regression models were used to analyze the aforementioned metrics. For the time to action and acceptance rate, each participant and trial were random effects, while the conditions were fixed effects. For the mental load, trust, satisfaction with the AI, understanding of the AI, and the satisfaction of the AI explanations, each participant was a random effect. For the analysis of participants’ rankings of the four conditions, a cumulative link mixed-effects model was used to analyze the ordinal response variable. The results of these models were interpreted through the estimated coefficient values for the fixed effect variables.

The interview data was analyzed by two coders using thematic analysis [4]. Two authors iteratively discussed and developed a codebook. The inter-rater reliability was calculated using Cohen’s kappa [43], where $\kappa = 0.82$ (SE = 0.055, 95% CI = [0.71, 0.93]). The final themes were developed and refined over multiple iterations of discussion.

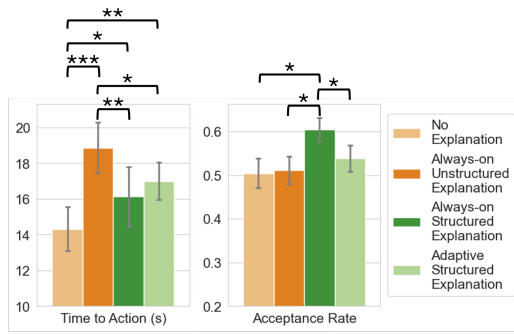


Figure 5: Participants’ time to action and acceptance rate for each AI explanation condition. The error bars represent standard errors. (*: $p < 0.05$; **: $p < 0.01$; *: $p < 0.001$)**

5 User Study Results

We analyzed participants’ time to action, acceptance rate, subjective perceptions, and interview data to better understand their experiences with the different AI explanation conditions.

5.1 Quantitative Results

5.1.1 Time to Action. Our results showed the presentation of the explanations impacted user’s time to action. Compared to the no explanation condition, providing explanations significantly increased the time participants took to select an AI recommendation (always-on unstructured: $\beta = 4.6467, p < 0.001$, always-on structured: $\beta = 1.9341, p < 0.05$, always-on unstructured: $\beta = 2.7492, p < 0.01$). But among the three conditions with explanation, structured explanations—both always-on structured explanations ($\beta = -2.7126, p < 0.01$) and adaptive structured explanations ($\beta = -1.8975, p < 0.05$)—significantly reduced participants’ time to action compared to unstructured explanations (Figure 5).

5.1.2 Acceptance Rate. Our results found that always-on structured explanations resulted in a higher acceptance rate compared to the other three conditions (compared with no explanation: $\beta = 0.0675, p < 0.05$, with unstructured explanations: $\beta = 0.0712, p < 0.05$, and with adaptive structured explanations: $\beta = 0.0584, p < 0.05$; Figure 5).

5.1.3 Mental Load. We found no significant differences in participants’ mental load when deciding or understanding recommendations (Figure 6), however, participants rated their mental load as significantly lower when reading structured explanations (i.e., always-on structured: $\beta = -0.8182, p < 0.01$, adaptive structured: $\beta = -0.5909, p < 0.05$) than when reading unstructured explanations, which aligns with the time to action findings.

5.1.4 Trust in AI. We found that both the always-on unstructured explanations ($\beta = 0.4091, p < 0.05$) and always-on structured explanations ($\beta = 0.4545, p < 0.01$) led to higher levels of trust when compared to the no explanation condition (Figure 6). There was no difference between the always-on unstructured, always-on structured, and adaptive structured explanation conditions.

5.1.5 Satisfaction with AI. When presented with always-on unstructured and always-on structured explanations, participants were both more satisfied with the recommendations given by the AI compared to when no explanation was provided (always-on unstructured: $\beta = 0.4773, p < 0.05$; always-on structured: $\beta = 0.7045, p < 0.001$; Figure 6). The use of adaptive structured explanations did not result in participants being more satisfied with the AI’s recommendations compared to when no explanation was provided, and even led to significantly lower satisfaction compared to the always-on structured explanation condition ($\beta = -0.4318, p < 0.05$).

5.1.6 Understanding of AI. When participants were provided with an explanation, they had a better understanding of the rationale behind the AI recommendation than when they were not (e.g., always-on unstructured explanations: $\beta = 1.1136, p < 0.001$, always-on structured explanations: $\beta = 1.0303, p < 0.001$, always-on structured explanations: $\beta = 0.8258, p < 0.001$; Figure 6). There was no significant difference between the always-on unstructured, always-on structured, and adaptive structured explanation conditions, suggesting that structuring and adaptively presenting the unstructured explanation did not negatively impact participants’ understanding of the AI.

5.1.7 Satisfaction with AI explanations. We also found that adaptive structured explanations were less satisfying compared to always-on unstructured explanations ($\beta = -0.5227, p < 0.05$; Figure 6). Though not statistically significant, always-on structured explanations were rated as marginally less satisfying than always-on unstructured explanations ($\beta = -0.5000, p = 0.053$).

5.1.8 Overall Preferences. In terms of participants’ overall preferences for the four conditions (Figure 7), both the always-on unstructured explanations ($\beta = -1.1543, p < 0.01$) and the adaptive structured explanations ($\beta = -0.9242, p < 0.05$) were ranked higher than the no-explanation condition. However, the difference in ranks between the adaptive structured explanation condition and the no explanation condition was not statistically significant ($\beta = -0.6752, p = 0.074$). Additionally, no significant difference was found between the three explanation conditions.

5.2 Qualitative Results

Seven themes were identified within the interview feedback. Some of the themes helped explain why participants preferred some conditions over others, whereas others speculated about other ways that explanations could be improved. Each theme is elaborated in the following sections.

5.2.1 Usability. Usability referred to the amount of effort or time to interact with interface, i.e., whether the interface was easy and quick to use, or required more effort. Participants generally wanted to perform minimal actions and reading on the interface, with easy viewing and decision-making. For example, some participants noted that they would only want explanations when they had enough cognitive capacity, e.g., “If it were detecting locations or the amount of activity I was doing, then I’d probably prefer the simpler interface. If I were driving or doing something active, it wouldn’t be helpful for me to stop and read that explanation, so I’d dismiss it.” (P13)

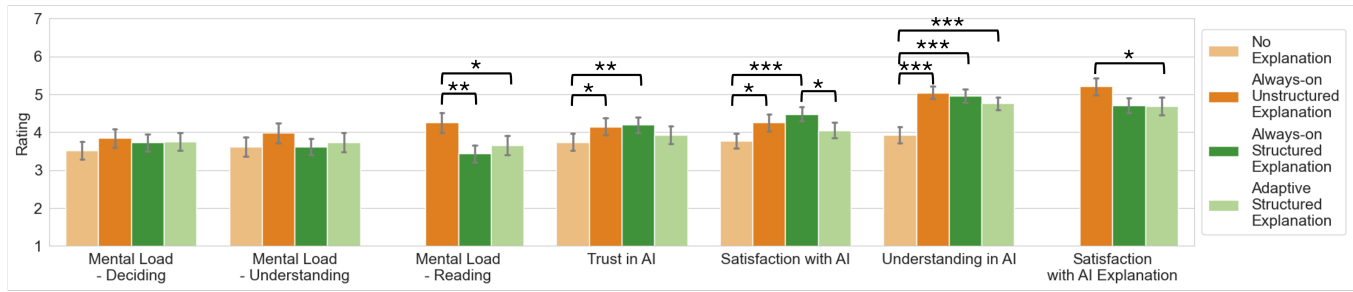


Figure 6: Participants’ reported mental load, trust, satisfaction, understanding, and satisfaction with the AI Explanations. The error bars represent standard errors. (*: $p < 0.05$; **: $p < 0.01$; *: $p < 0.001$)**

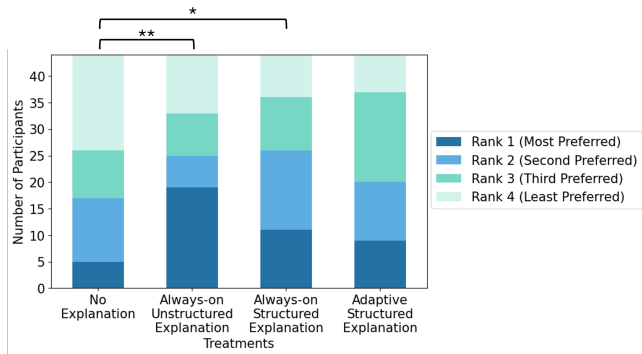


Figure 7: Participants’ preference rankings for the four AI explanation conditions.

Twenty participants valued the high usability of the no explanation condition, saying the interface was easy and fast to use, e.g., “It allowed me to react fastest” (P10). In conditions with explanations, eight participants mentioned that the information provided by always-on structured explanations was easy to access, e.g., “The explanations were automatically expanded” (P13). The adaptive structured explanations received mixed feedback. Fifteen participants found them easy and fast to use (e.g., “[They] were very simple and to the point” (P2)). On the other hand, usability concerns were raised by 30 participants who disliked the extra effort needed to interact with the interface (e.g., “You have to hit an extra button, which just seems like an unnecessary step” (P29)).

5.2.2 Sufficiency. Participants also noted that sufficiency, or the amount of information required to understand AI-generated recommendations or explanations, was important. This included both wanting more information and removing unnecessary details. Thirty participants mentioned their desire for more transparency in the AI decision process in the no explanation condition to help them understand AI decision better (e.g., “it doesn’t have any explanation” (P37)). Twenty-four participants favored the comprehensive details provided by the unstructured explanations, e.g., “it gives you a more comprehensive explanation as to what’s happened” (P9). Regarding the structured explanations, 18 participants thought the four icons contained sufficient information, whereas 15 participants suggested

that the icons could have been less redundant and location information was unnecessary. For example, “the feedback I got just like seemed redundant. It just kind of felt like a waste of time to read down a list of four points that often had the same information in them” (P33).

5.2.3 Readability. Readability referred to the ease of reading and viewing text or icons presented on the interface (e.g., ease of viewing information, too much text, low cognitive load). The comparison primarily focused on unstructured versus structured explanations. Thirty-one participants noted that the unstructured, free-text explanations resembled human speech and were easy to understand (e.g., “it would be if like a human was just explaining it to me, like, oh because I saw that this and so I predicted your in this environment trying to do this ... I feel that’s easy to understand” (P40). The downside, as complained by 30 participants, was that there was too much text making the explanations difficult to read (e.g., “It was hard to read this block of text” (P7)).

On the other hand, 8 participants mentioned that the structured explanations were easy to view (e.g., “it was just really nicely laid out, organized” (P22)). Twenty-three participants, however, found the icons and the logic between them difficult to understand (e.g., “I had to like keep reminding myself what they meant” (P1) and “those four icons didn’t always connect” (P12)).

5.2.4 User Control. Participants also highlighted the importance of the amount of control user they had over the interface and their choices (e.g., giving users more options, allowing the user to decide when to view information). They believed that the AI “should give you that option ... ask you, would you like more suggestions?” (P21). Twenty-eight participants expressed a desire for more control over explanation length, appreciating the greater user control offered in the adaptive structured explanation condition to toggle explanations on or off (e.g., “[this condition] was good in a way that they have an option, have it on or have it off” (P9)).

5.2.5 Accuracy. Accuracy concerns related to the contextual relevance of AI-generated recommendations and information. Twenty-four participants wanted more accurate recommendations for certain situations and sometimes found that recommendations or explanations didn’t match the context (e.g., “The context of the video was totally different from the action that I’m kind of like being asked to perform” (P30)).

5.2.6 Multi-Modality. Participants thought that explanations should support options for other modes of accessibility, such as information presented in multiple languages, audio input/output, or visual display. This was mentioned by 19 participants as a potential way to improve the experience. For instance, one of the videos shown in the study contained dialogue spoken in Russian, which provided direct information on the next recommended action, but the AI assistant failed to capture it. Another example came from P37, who explained, “*if I don’t need to read but just show me some pictures or images like for example, for objects you don’t need to mention like what object it is, you just put what I’m looking at the picture here.*”

5.2.7 Personalization. Lastly, 43 participants showed interest in personalization. They thought that explanations should contain information customized for personal preferences, such as providing recommendations based on daily routine and previous actions, displaying explanations for unfamiliar tasks, and being less frequent for tasks users are already familiar with. As P12 suggested, “*one thing I think would be helpful to add [to improve the explanation] could be like historical action.*”

5.3 Summary

In this section, we bridge the qualitative and quantitative results by showing how the qualitative feedback from participants provides context for interpreting the quantitative results.

Structured explanations are quick and easy to use, but fall short in providing sufficient details and the naturalness of human-like language. The quantitative results showed that both always-on structured explanations and adaptive structured explanations reduced participants’ time to action and mental load when reading the explanations, compared to unstructured explanations. Interview feedback supported these findings, as participants mentioned that structured explanations were clear to view and access (i.e., high usability), allowing them to quickly glance at the information, which contributed to shorter time to action. This may also explain why always-on structured explanations also resulted in a higher acceptance rate of AI recommendations compared to all other conditions, i.e., they had less time to overthink and instead relied on the easily-accessible icons, which reinforced their belief that the recommendation seemed trustworthy.

Despite these advantages, participants rated the always-on structured explanations (marginally) and adaptive structured explanations as less satisfying than the always-on unstructured explanation. During the interviews, participants mentioned that structured explanations were clear to view, but they struggled to remember the meaning of the icons and their logical connection. Unstructured explanations, on the other hand, offered more comprehensive and human-like information. The process of mapping abstract icons to their semantic meanings likely imposed an extraneous cognitive burden, which may have led participants to value “sufficiency” and “readability” more when evaluating whether an explanation was satisfactory, thus leading to higher satisfaction ratings for unstructured explanations.

Adaptive structured explanations offer users control over the interface, but the (inappropriate) extra toggling hindered effective human-AI interaction. In comparison to the no explanation condition, all three explanation conditions led to a higher

understanding of the AI. However, the adaptive structured explanation condition—where participants needed to proactively interact with the interface to view details—was less effective in showing the benefits of providing AI explanations. Specifically, both always-on unstructured and always-on structured explanations increased participants’ trust and satisfaction in the AI’s recommendations and were preferred over the no explanation condition, but this was not observed for the adaptive structured explanations. The qualitative feedback revealed that while participants appreciated having control over when to view explanations, the additional toggle action felt like a redundant barrier that impaired the smoothness of their interaction with the interface, thus shifting their focus from quickly grasping the information to managing the interface.

6 Discussion

Findings from our user study drive the design implications for creating glanceable AI explanations on ultra-small devices that meet users’ expectations and needs more effectively. In this section, we discuss these design implications as well as the limitations of our work.

6.1 Design Implication on Structuring LLM Explanation Text into Components

The trust gain resulted from presenting always-on structured explanations may be superficial trust based on simplicity rather than understanding. It’s therefore crucial to build genuine trust and avoid blind reliance on oversimplified icons. Coupled with participants’ desire for sufficient and readable details, accurate summarization and iconification play a critical role in ensuring appropriate understanding of AI recommendations. Designers should carefully balance brevity with clarity when creating icons and summaries for glanceability. As an example, Zender and Mejía [75] suggest that adding more symbols can enhance comprehension by providing context and narrow the focus. For instance, adding a “bookshelf” icon to a “library” icon can improve understanding compared to using just a “man reading” icon. They also encourage designers to test icons through multiple iterations and learn from user feedback for improvement.

There is also likely a continuum of text verbosity beyond simply hiding and displaying explanations that adapts to different tasks, e.g., “progressively disclosing” a continuum of explanation levels [54]. Computational models could be developed to predict the appropriate level of detail based on user preferences, current contexts, and interaction histories. Alternatively, from a design perspective, users should be given control over the amount of explanation detail they receive, ranging from icons and diagrams to unstructured text.

6.2 Design Implication on Adaptively Presenting Structured Explanations

Our analysis highlighted that participants preferred AI to be as transparent as possible and found that inefficient adaptive structured explanations were less favored compared to always-on explanations. Thus if adaptive explanations are to be used, it is crucial to carefully select the optimal timing to present and hide explanations. Based on prior literature [72], explanations should be displayed

automatically when there is a need to manage AI errors or to resolve user surprise/confusion. More sophisticated methods need to be developed to implement adaptive explanations effectively. Model developers should explore more accurate indicators of AI correctness to detect when the AI makes mistakes. For situations require the understanding of user states, building a personalized user mental model based on interaction logs to predict when users are confused, unfamiliar with the outcome, or uncertain about their own choices [37, 38, 65], could further improve the effectiveness of adaptive explanations.

6.3 Limitations and Future Work

This research has several limitations. The results of our study are limited by the specific type of AI assistance scenario (i.e., goal-oriented contextual action recommendation) and the fact that participants watched third-party videos rather than interacting with an actual working system on a smartwatch. The controlled in-lab setting may have also made participants deliberately curious about seeing more information, whereas in a longitudinal study conducted in real-world settings, they may spend less time deciding what to do next, and their user preferences may evolve over time. Moreover, our LLM recommendation pipeline used only video frames as input. In practice, other signals such as digital states, audio, eye tracking, text recognition, EMG, IMU could all be utilized as input for the modality encoder models and mapped into the LLM embedding space [46] to enable more accurate action inference. Furthermore, the generated LLM recommendations carry potential risks including biases against underrepresented groups and accessibility challenges for populations with disabilities or limited technological proficiency.

A promising future direction to explore for effectively presenting AI explanations on ultra-small devices is to flexibly leverage multiple modalities and selectively summarize information from various input signals [45]. With more input signals available such as audio and eye tracking, explanations could incorporate much richer information than the four contextual concepts currently used (i.e., object, activity, goal, location). For example, the AI explanations could include the user's historical preferences, dialogues with friends, or current health state. In addition to text and icons, there are alternative modalities for displaying explanations that can reduce mental demand while improving the naturalness of information delivery. For example, on interfaces like AR smartglasses, which support overlay displays, real-world objects can be highlighted and overlaid onto the physical environment to enhance understanding [13].

7 Conclusion

In this research, we explored how to effectively present glanceable LLM explanations on ultra-small devices. We proposed that LLM explanations should be *structured* using defined contextual components during prompting to achieve spacial glanceability, and *adaptively* presented based on confidence levels to achieve temporal glanceability. We conducted a user study to evaluate the effectiveness of structuring and adaptively presenting explanations on user experiences and perceptions of an AI assistant system using a simulated smartwatch UI. We found that structured explanations reduced users' time to action and lowered perceived mental load

when reading the explanations. Always-on structured explanations led to higher acceptance rates of the AI recommendations. But structured explanations also lacked sufficient and readable details and were thus less satisfying than unstructured explanations. Additionally, while adaptive structured explanations offered participants more control over the interface, they introduced a redundant interaction step and generally hindered effective interaction with the AI, highlighting the need for improvements in the timing for adaptive presentation of AI explanations. Our study should thus serve as a foundational starting point to understand how users prefer to view AI explanations on ultra-small devices and encourage future research and technical developments that promote responsible AI for users of AI-based personal assistants.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [2] Michael Albers and L. Kim. 2002. Information Design for the Small-screen Interface: An Overview of Web Design Issues for Personal Digital Assistants. *Technical Communication* 49 (02 2002), 45–60.
- [3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–16.
- [4] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.
- [5] Kelly Caine. 2016. Local standards for sample size at CHI. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 981–992.
- [6] Jiu-hai Chen and Jonas Mueller. 2023. Quantifying uncertainty in answers from any language model via intrinsic and extrinsic confidence assessment. *arXiv preprint arXiv:2308.16175* (2023).
- [7] Zichen Chen, Ambuj K Singh, and Misha Sra. 2023. LMEExplainer: a Knowledge-Enhanced Explainer for Language Models. *arXiv preprint arXiv:2303.16537* (2023).
- [8] Michael Chromik and Andreas Butz. 2021. Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces. In *Human-Computer Interaction – INTERACT 2021*, Carmelo Ardito, Rosa Lanzilotti, Alessio Malizia, Helen Petrie, Antonio Piccinno, Giuseppe Desolda, and Kori Inkpen (Eds.). Springer International Publishing, Cham, 619–640.
- [9] Jaemin Chun, Anind Dey, Kyungtaek Lee, and SeungJun Kim. 2018. A qualitative study of smartwatch usage and its usability. *Human Factors and Ergonomics in Manufacturing & Service Industries* 28, 4 (2018), 186–199. <https://doi.org/10.1002/hfm.20733> <https://onlinelibrary.wiley.com/doi/pdf/10.1002/hfm.20733>
- [10] Daniel Churchill. 2011. Conceptual Model Learning Objects and Design Recommendations for Small Screens. *Journal of Educational Technology & Society* 14, 1 (2011), 203–216. <http://www.jstor.org/stable/jeductechsoci.14.1.203>
- [11] Justin Cosentino, Anastasiya Belyaeva, Xin Liu, Nicholas A Furlotte, Zhun Yang, Chace Lee, Erik Schenck, Yojan Patel, Jian Cui, Logan Douglas Schneider, et al. 2024. Towards a Personal Health Large Language Model. *arXiv preprint arXiv:2406.06474* (2024).
- [12] Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-Inference: Exploiting Large Language Models for Interpretable Logical Reasoning. In *The Eleventh International Conference on Learning Representations*.
- [13] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. 2020. DepthLab: Real-time 3D interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 829–843.
- [14] Robert L Duchnick and Paul A Kolers. 1983. Readability of text scrolled on visual display terminals as a function of window size. *Human factors* 25, 6 (1983), 683–692.
- [15] Inger Ekman and Petri Lankoski. 2002. What should it do? key issues in navigation interface design for small screen devices. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (Minneapolis, Minnesota, USA) (CHI EA '02). Association for Computing Machinery, New York, NY, USA, 622–623. <https://doi.org/10.1145/506443.506514>
- [16] Julia El Zini, Mohamad Mansour, Basel Mousi, and Mariette Awad. 2022. On the evaluation of the plausibility and faithfulness of sentiment analysis explanations. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*. Springer, 338–349.

- [17] Leah Findlater and Joanna McGrenere. 2008. Impact of screen size on performance, awareness, and user satisfaction with adaptive graphical user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) (CHI '08). Association for Computing Machinery, New York, NY, USA, 1247–1256. <https://doi.org/10.1145/1357054.1357249>
- [18] Michael D. Good, Michael C. Dorneich, John E. Deaton, Floyd Glenn, C. Shawn Burke, and Joshua Downs. 2001. Techniques for Interacting with Large Information Spaces on Small-Screen Displays. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 45, 6 (2001), 585–589. <https://doi.org/10.1177/154193120104500602> arXiv:<https://doi.org/10.1177/154193120104500602>
- [19] Google. 2019. Glass Enterprise Edition 2: faster and more helpful. <https://blog.google/products/platforms-devices/glass-enterprise-edition-2/>
- [20] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18995–19012.
- [21] Humane. 2024. Ai Pin Tech Details. <https://humane.com/aipin/tech-details>
- [22] Peiling Jiang, Jude Rayan, Steven P. Dow, and Haijun Xia. 2023. Graphologue: Exploring Large Language Model Responses with Interactive Diagrams. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 3, 20 pages. <https://doi.org/10.1145/3586183.3606737>
- [23] Tanya R Jonker, Ruta Desai, Kevin Carlberg, James Hillis, Sean Keller, and Hrvoje Benko. 2020. The Role of AI in Mixed and Augmented Reality Interactions. In *CHI2020 ai4hci Workshop Proceedings*. ACM.
- [24] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690* (2017).
- [25] Tomonari Kamba, Shawn A. Elson, Terry Harpold, Tim Stamper, and Piyawadee Sukaviriya. 1996. Using small screen space more efficiently. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, British Columbia, Canada) (CHI '96). Association for Computing Machinery, New York, NY, USA, 383–390. <https://doi.org/10.1145/238386.238582>
- [26] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems* 29 (2016).
- [27] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35 (2022), 22199–22213.
- [28] Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.
- [29] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Honolulu</city>, <state>HI</state>, <country>USA</country>, </conf-loc>) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–13. <https://doi.org/10.1145/3313831.3376873>
- [30] Himabindu Lakkaraju, Stephen H Bach, and Jure Leskovec. 2016. Interpretable decision sets: A joint framework for description and prediction. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1675–1684.
- [31] Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329* (2022).
- [32] Jaedong Lee, Changhyeon Lee, and Gerard Jounghyun Kim. 2017. Vouch: multimodal touch-and-voice input for smart watches under difficult operating conditions. *Journal on Multimodal User Interfaces* 11, 3 (6 2017), 289–299. <https://doi.org/10.1007/s12193-017-0246-y>
- [33] Qibin Li, Junyuan Hong, Chulin Xie, Jeffrey Tan, Rachel Xin, Junyi Hou, Xavier Yin, Zhun Wang, Dan Hendrycks, Zhangyang Wang, et al. 2024. LLM-PBE: Assessing Data Privacy in Large Language Models. *Proceedings of the VLDB Endowment* 17, 11 (2024), 3201–3214.
- [34] Q Vera Liao and Jennifer Wortman Vaughan. 2023. AI Transparency in the Age of LLMs: A Human-Centered Research Roadmap. *arXiv preprint arXiv:2306.01941* (2023).
- [35] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the Effect of Out-of-distribution Examples and Interactive Explanations on Human-AI Decision Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 408 (oct 2021), 45 pages. <https://doi.org/10.1145/3479552>
- [36] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).
- [37] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [38] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.
- [39] I Scott MacKenzie. 2013. *Human-Computer Interaction: An Empirical Research Perspective*. (2013).
- [40] Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems* 36 (2024).
- [41] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 35. 14867–14875.
- [42] Steven McGuckin, Soumyadeb Chowdhury, and Lewis Mackenzie. 2016. Tap 'n' shake: gesture-based smartwatch-smartphone communications system. In *Proceedings of the 28th Australian Conference on Computer-Human Interaction* (Launceston, Tasmania, Australia) (OzCHI '16). Association for Computing Machinery, New York, NY, USA, 442–446. <https://doi.org/10.1145/3010915.3010983>
- [43] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.
- [44] Meta. 2024. Introducing Orion, Our First True Augmented Reality Glasses. <https://about.fb.com/news/2024/09/introducing-orion-our-first-true-augmented-reality-glasses/>
- [45] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.
- [46] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. 2023. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058* (2023).
- [47] Tarjin Rahman and Paul Muter. 1999. Designing an Interface to Optimize Reading with Small Display Windows. *Human Factors* 41, 1 (1999), 106–117. <https://doi.org/10.1518/001872099779577264> PMID: 10354807. arXiv:<https://doi.org/10.1518/001872099779577264>
- [48] Juan Rebanal, Jordan Combitis, Yuqi Tang, and Xiang 'Anthony' Chen. 2021. XALgo: a Design Probe of Explaining Algorithms' Internal States via Question-Answering. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) (IUI '21). Association for Computing Machinery, New York, NY, USA, 329–339. <https://doi.org/10.1145/3397481.3450676>
- [49] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.
- [50] Sabine Schröder and Martina Ziefle. 2006. Icon Design on Small Screens: Effects of Miniaturization on Speed and Accuracy in Visual Search. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 5 (2006), 656–660. <https://doi.org/10.1177/154193120605000508> arXiv:<https://doi.org/10.1177/154193120605000508>
- [51] Katie A. Siek, Yvonne Rogers, and Kay H. Connelly. 2005. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In *Human-Computer Interaction - INTERACT 2005*, Maria Francesca Costabile and Fabio Paternò (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 267–280.
- [52] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).
- [53] Karin Spalink. 2002. Small-Screen Devices and Translation. In *Proceedings of Translating and the Computer* 24. Aslib, London, UK. <https://aclanthology.org/2002.tc-1.5>
- [54] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.
- [55] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-based Design. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) (IUI '22). Association for Computing Machinery, New York, NY, USA, 212–228. <https://doi.org/10.1145/3490099.3511119>
- [56] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [57] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shriti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

- [58] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems* 36 (2024).
- [59] Jean Vanderdonckt, Sara Bouzit, Gaëlle Calvary, and Denis Chène. 2019. Exploring a Design Space of Graphical Adaptive Menus: Normal vs. Small Screens. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 2 (jul 2019), 40 pages. <https://doi.org/10.1145/3237190>
- [60] Elena Voita, Javier Ferrando, and Christoforos Nalmpantis. 2023. Neurons in large language models: Dead, n-gram, positional. *arXiv preprint arXiv:2309.04827* (2023).
- [61] Sandra Wachter, Brent Mittelstadt, and Chris Russell. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.* 31 (2017), 841.
- [62] Fulton Wang and Cynthia Rudin. 2015. Falling rule lists. In *Artificial intelligence and statistics*. PMLR, 1013–1022.
- [63] Xinru Wang, Hannah Kim, Sajjadur Rahman, Kushan Mitra, and Zhengjie Miao. 2024. Human-LLM collaborative annotation through effective verification of LLM labels. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–21.
- [64] Xinru Wang, Chen Liang, and Ming Yin. 2023. The Effects of AI Biases and Explanations on Human Decision Fairness: A Case Study of Bidding in Rental Housing Markets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*. International Joint Conferences on Artificial Intelligence Organization, 3076–3084. <https://doi.org/10.24963/ijcai.2023/343> Main Track.
- [65] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM web conference 2022*. 1697–1708.
- [66] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *26th international conference on intelligent user interfaces*. 318–328.
- [67] Xinru Wang and Ming Yin. 2022. Effects of explanations in ai-assisted decision making: Principles and comparisons. *ACM Transactions on Interactive Intelligent Systems* 12, 4 (2022), 1–36.
- [68] Xinru Wang and Ming Yin. 2023. Watch out for updates: Understanding the effects of model explanation updates in ai-assisted decision making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–19.
- [69] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.
- [70] Haijun Xia, Tovi Grossman, and George Fitzmaurice. 2015. NanoStylus: Enhancing Input on Ultra-Small Displays with a Finger-Mounted Stylus. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (Charlotte, NC, USA) (UIST '15)*. Association for Computing Machinery, New York, NY, USA, 447–456. <https://doi.org/10.1145/2807442.2807500>
- [71] Miao Xiong, Zhiyuan Hu, Xinyang Lu, YIFEI LI, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The Twelfth International Conference on Learning Representations*. <https://openreview.net/forum?id=gjeQKfXfPz>
- [72] Xuhai Xu, Anna Yu, Tanya R Jonker, Kashyap Todi, Feiyu Lu, Xun Qian, João Marcelo Evangelista Belo, Tianyi Wang, Michelle Li, Aran Mun, et al. 2023. XAIR: A Framework of Explainable AI in Augmented Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–30.
- [73] Bingsheng Yao, Ishan Jindal, Lucian Popa, Yannis Katsis, Sayan Ghosh, Lihong He, Yuxuan Lu, Shashank Srivastava, Yunyao Li, James Hendler, et al. 2023. Beyond labels: Empowering human annotators with natural language explanations through a novel active-learning architecture. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 11629–11643.
- [74] Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems* 35 (2022), 30378–30392.
- [75] Mike Zender and G MAURICIO Mejía. 2013. Improving icon design: through focus on the role of individual symbols in the construction of meaning. *Visible Language* 47, 1 (2013), 66–89.
- [76] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aavek Purohit, Michael Ryoo, Vikas Sindhwani, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. *arXiv preprint arXiv:2204.00598* (2022).
- [77] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. <https://openreview.net/forum?id=SkeHuCVFDr>
- [78] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 295–305.
- [79] Zeyuan Allen Zhu and Yuanzhi Li. 2023. Physics of language models: Part 3.1, knowledge storage and extraction. *arXiv preprint arXiv:2309.14316* (2023).
- [80] Yuxin Zi, Kaushik Roy, Vignesh Narayanan, Manas Gaur, and Amit Sheth. 2023. IERL: Interpretable Ensemble Representation Learning–Combining Crowd-Sourced Knowledge and Distributed Semantic Representations. *arXiv preprint arXiv:2306.13865* (2023).
- [81] Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405* (2023).

A LLM Prompts Used

A.1 Unstructured Explanation Prompt

Your input will be (1) a series of narrations of videos taken from a smartglasses user's perspective and (2) objects detected in the user's field of view. These two pieces of information are detected by upstream machine learning models.

Input:

1) Video narrations, in the order of the least to most recent: ['#C C looks around ', '#C C turns around ', '#c c walks in the supermarket', '#C C walks in the shop', '#C C walks in the supermarket ', '#C C looks around ', '#c c walks in the supermarket', '#C C looks around ', '#C C walks in the supermarket', '#C C turns around ', '#C C turns around ', '#C C looks around', '#C C looks around ', '#C C picks a box from the shelf ', '#C C operates the phone', '#C C uses the phone']

2) Detected objects in the user's field of view: ['bowl', 'bottle', 'oven', 'potted plant', 'orange', 'microwave', 'refrigerator', 'cup', 'book', 'sink', 'toothbrush', 'cell phone', 'broccoli', 'person', 'scissors', 'cake', 'dining table', 'knife', 'couch', 'apple', 'handbag']

The recommendation is "open a pantry organization tutorial on the Youtube app". Please provide an <explanation> for this recommendation within 25 words. In the explanation, only focus on referencing information from the contextual input and please think step by step. Please do not restate the recommendation in your explanation.