RealitySummary: Exploring On-Demand Mixed Reality Text Summarization and Question Answering using Large Language Models

ANONYMOUS AUTHOR(S)

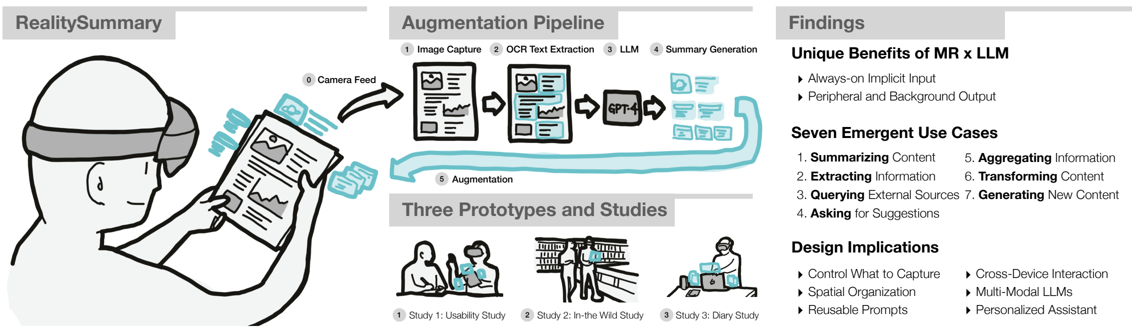

Fig. 1. We developed RealitySummary, an on-demand mixed reality reading assistant, over three iterations and conducted three different studies. Through this evolution, we highlight three key findings: 1) unique benefits of MR and LLMs, 2) emergent use cases of MR reading assistants, and 3) design implications for mixed reality reading assistants.

Large Language Models (LLMs) are gaining popularity as tools for reading and summarization aids. However, little is known about their potential benefits when integrated with mixed reality (MR) interfaces to support everyday reading assistants. We developed RealitySummary, an MR reading assistant that seamlessly integrates LLMs with always-on camera access, OCR-based text extraction, and augmented spatial and visual responses in MR interfaces. Developed iteratively, RealitySummary evolved across three versions, each shaped by user feedback and reflective analysis: 1) a preliminary user study to understand reader perceptions (N=12), 2) an in-the-wild deployment to explore real-world usage (N=11), and 3) a diary study to capture insights from real-world work contexts (N=5). Our findings highlight the unique advantages of combining AI and MR, including an always-on implicit assistant, minimal context switching, and spatial affordances, demonstrating significant potential for future LLM-MR interfaces beyond traditional screen-based interactions.

CCS Concepts: • **Human-centered computing** → **Mixed / augmented reality**.

Additional Key Words and Phrases: Mixed Reality; Large Language Models; Augmented Reading; In-the-Wild Study; Diary Study

ACM Reference Format:

Anonymous Author(s). 2024. RealitySummary: Exploring On-Demand Mixed Reality Text Summarization and Question Answering using Large Language Models. In *Proceedings of CHI*. ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

In his pioneering work [62], Pierre Wellner introduced the concept of AR-enhanced documents, advocating for a future where computer interfaces should augment our physical environment rather than restricting interactions to flat rectangle screens. Since then, HCI researchers have explored augmented reading experiences by leveraging AR and MR interfaces [9, 32, 46]. However, insights into the *real-world, everyday use* of these interfaces remain largely unexplored. Most existing systems rely on pre-prepared information and preprocessed responses, limiting their applicability beyond controlled lab environments.

This paper provides empirical study contributions to understanding the benefits and limitations of mixed reality reading assistants in *real-world* and *everyday* contexts. We developed RealitySummary, an MR reading assistant powered by large language models (LLMs) to ensure applicability and reliability for uncontrolled real-world environments. By leveraging an always-on camera, OCR-based text extraction, question-answering capabilities, and augmented spatial and visual LLM responses within MR interfaces, we deployed this system beyond lab settings, uncovering unique insights that have been previously unexplored.

RealitySummary evolved through three iterative versions, each shaped by user feedback and reflection. First, we developed the initial prototype of RealitySummary (**Study 1: RS-Doc**) on the Hololens 2 and conducted a preliminary user study (N=12) to understand reader perceptions and assess the usability of on-demand mixed reality reading assistants with AI-generated summarization. Based on the findings, we developed a second version of the system (**Study 2: RS-Wild**) on the Apple Vision Pro, enabling an in-the-wild study (N=11) to gather insights on real-world usage. Finally, we created the third iteration (**Study 3: RS-Diary**) to conduct a diary study (N=5), which explored how readers adapt to MR reading assistants over an extended period.

These iterations and studies collectively highlight the unique benefits of integrating LLMs with MR for everyday assistance. For instance, the implicit input and spatial output modalities provide a clear advantage over traditional explicit inputs and screen-based outputs, allowing readers to stay focused on their tasks with minimal context switching. The always-on camera functionality also enables new, previously unconsidered use cases, such as enhancing not only documents but also posters, signage, nutrition labels, emails, and even restaurant menus in everyday environments with LLM capabilities. Additionally, MR's spatial and tangible affordances open up novel ways to interact with LLMs, including summarizing entire bookshelves or multiple books on a table rather than just individual pages held in hand. These insights demonstrate how the combined MR and LLM systems can effectively support long-term reading and knowledge work in daily activities.

Finally, our paper contributes:

(1) RealitySummary, an iteratively developed mixed reality reading assistant that uses large language models for on-demand content summarization and augmentation.
(2) Observations and opportunities drawn from our usability study (N=12), in-the-wild study (N=11), and diary study (N=5), which highlight the potential benefits of mixed reality document enhancements.
(3) Lessons learned from the three iterations of our system.

2 RELATED WORK

We summarize prior work on reading augmentation and content summarization with a focus on interactive approaches.

2.1 Reading Augmentation

Document Enhancement. Human-computer interaction researchers have explored ways to augment physical documents with AR-enhanced content. Examples include *HoloDoc* [32], *Replicate and Resuse* [22], and *PaperTrail* [46], all of which laid the groundwork for augmenting physical paper by superimposing dynamic virtual content onto printed documents. Prior works have demonstrated these concepts on different platforms. For example, *Pacer* [33], *MagPad* [64], *Tecahable Reality* [40], and *Dually Noted* [44] use mobile AR for text augmentation, whereas *Replicate and Resuse* [22] and *Wiki-TUI* [63] leverage head-mounted displays to overlay information. Other examples such as *DigitalDesk* [62], *DocuDesk* [17], *Affinity Lens* [53], *Matulic et al.* [37, 38], and *QOOK* [67] use projection mapping for document augmentation. Finally, cAR [25] leveraged transparent displays for document augmentation to support active reading. At their core, such augmented reading research [4, 20, 24, 29] aims to improve cognitive load [10], knowledge accumulation [15, 66], and spatial visualization [51], given a long-standing consensus that readers prefer physical paper over screen-based reading [50, 55].

However, none of the existing systems have achieved an *on-demand* document enhancement. Here, we define *on-demand* as the ability to dynamically and interactively enhance documents without requiring prior manual preparation or pre-processing. For example, prior works like *HoloDoc* [32], *PaperTrail* [46], and *Affinity Lens* [53] did not extract text directly from the given document; instead, content was prepared in advance and loaded based on attached fiducial markers. As a result, these systems are not fully adaptable and deployable to real-world scenarios. Some preliminary works have explored on-demand text analysis using machine learning, such as *Dually Noted* [44] analyzing document structure in real-time, *Augmented Math* [12] and *Augmented Physics* [21] extracting diagrams and math equations, and *SOCRAR* [52] using OCR to extract key information. However, none have yet achieved comprehensive and general-purpose document enhancement.

Screen-Based Reading Support Tools. Outside of the context of mixed reality, a wide range of projects and systems have explored reading support tools for screen-based interfaces. Examples including *LiquidText* [56], *texSketch* [54], *ScholarPhi* [23], *Chameleon* [35], and *Marvista* [8] developed on-screen reading support tools. Similar to our work, *Marvista* [8] helps users read online documents by providing reading aids including summaries, contextual prompt questions, and reading metrics for active reading. Other works studied active and close reading support like *Metatation* [39] for closed reading, *XLibris* [43] with free form annotation, *Matulic and Norrie* [36] to support navigation, and *GatherReader* [26] with information gathering. Other works also support literature review and reference search for science papers [7, 28, 41]. Lastly, works like *Explorable Explanations* [58] and *Potluck* [34] try to enhance static documents with interactive content to improve reading comprehension through exploration. While these past works are limited to screen-based interfaces, this paper explores the augmented reading design for AR-based interfaces.

2.2 Content Summarization

Automatic Summarization. In the field of natural language processing, several works have employed automatic summarization to support readers [16, 65]. These approaches provide various summarization methods, including abstractive and extractive summaries generated from document photos [2], real-time summarization of user writing for quick text iterations [13], enhancing reading comprehension through automatic summarization [8], and *responsive text summarization* for adapting document summaries to screen sizes ranging from large displays to small watch faces [31]. With the recent advances in large language models like GPT-4, such automatic summarization tasks have become more accurate, robust, and accessible. For example, Goyal et al. [19] showed that GPT-3 summaries were preferred by humans over those generated by fine-tuned models. Various types of summarization have been explored using GPT-3, such as opinion summarization [3], news summarization [19], and medical dialogue summarization [11] while others have built custom frameworks for feeding long texts into GPT-3 to elicit more finely-controlled summarizations [42].

Interactive Summarization. Researchers have also explored interactive summarization approaches based on context and users' needs. For example, *Hoeve et al.* [57] studied Q&A interactions to assist users with a document-centric question-answering paradigm, finding that information-seeking was a common query. *Semantic Reactor* [18], developed by Google, enables users to ask questions about documents in natural language by fine-tuning BERT [14] using the *SQuAD dataset* [48]. *VERSE* [59] is another Q&A system that allows users to ask specific types of questions about a document but is limited by its inability to summarize. *Hoeve et al.* [57] identified various question types asked by participants regarding documents, with factual/summarization questions being the most frequent. Some question types include document-related, factoid, mechanical (answerable by rule-based methods), factual, yes/no, navigational, and summary questions. Voice assistants like Cortana, Alexa, Siri, VERSE [59], Firefox Voice [6], and Pushpak [27] offer efficient query modalities, as speech input has been found to be significantly faster than manual input, especially for mixed reality devices [1, 49].

Reality-Based Information Retrieval. Many recent works explored visual reality-based querying of information [5]. For example, GazePointAR [30] explored long term use of pronoun disambiguation in visual question answering in mixed-reality. G-Voila [60] explored gaze as a medium for question answering in everyday scenarios. While prior works have either focused on document augmentation with manual preparation [32] or on general mixed-reality LLM-based QnA [30], our goal is to build on this foundation by developing and exploring on-demand and real-time reading augmentation within everyday Mixed-Reality settings.

3 OVERVIEW OF OUR RESEARCH APPROACH

To uncover insights into on-demand reading assistants, we employed a three-phase iterative design process to investigate reader interactions and experiences with MR reading assistants. These explorations included a usability study on the initial prototype (Study 1: RS-Doc), followed by an in-the-wild study on a second prototype (Study 2: RS-Wild), and an author-conducted diary study using a third prototype (Study 3: RS-Diary). Each phase was designed to build upon the insights gained from the previous round, allowing for continuous refinement of the design and functionality.

4 RS-DOC: DESIGN EXPLORATION AND USABILITY STUDY WITH THE FIRST PROTOTYPE

We implemented the first version of our system (RS-Doc), which focused on in-place summarization for physical and digital documents, on a Microsoft Hololens 2. In this section, we present initial observations from our experiences developing and evaluating RS-Doc as well as new design opportunities highlighted by those explorations.

4.1 Implementation

System Design. RS-Doc captures documents using the HoloLens 2's internal camera and extracts text with Google Cloud OCR. The extracted text is processed using GPT-4 via OpenAI's API to identify key information, while spaCy's named entity recognition (NER) module assists in identifying important entities. The system employs Vuforia for markerless image tracking, and the Mixed Reality Toolkit 3 (MRTK3) SDK for content rendering. Readers interact with the system through voice commands processed via Google's speech-to-text API, allowing them to modify or query visualizations.

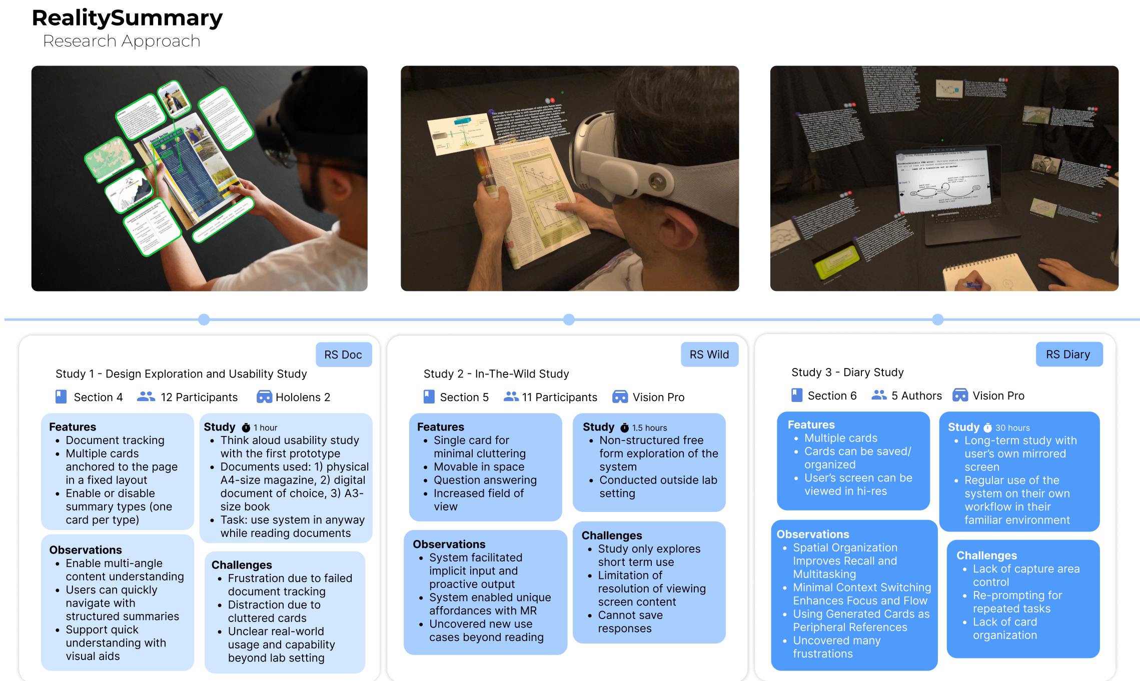

Fig. 2. Overview of our research approach spanning an evolution of our system across three different versions.

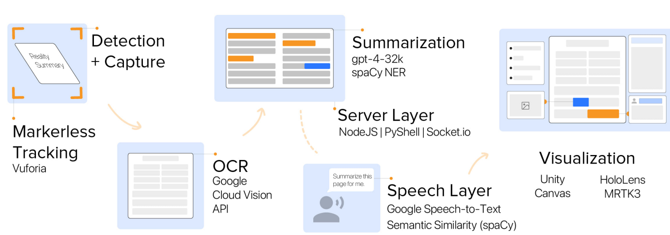

Fig. 3. RS-Doc Hololens 2 architecture.

The system captures images periodically for OCR, with an option for manual activation, and tracks documents by generating dynamic image targets, ensuring accurate document tracking and augmented content rendering.

Enhancement Types. To design and implement our system, we identified five key categories of AI-generated content enhancements—**summarize**, **compare**, **augment**, **extract**, and **navigate**—building on our preliminary elicitation study (see supplementary materials) as well as prior work, such as *HoloDoc* [32], *PaperTrail* [46], *DuallyNoted* [44], *QOOK* [67], and *Digital Desk* [32]. These categories informed both the conceptual framework and the system features. **Summarize** condenses text for quick reference, ranging from broad overviews to personalized summaries. This is implemented by sending the current page's text to GPT-4 for concise summaries, limited to 150 words. **Compare** transforms data into visual formats by parsing documents. Our system provides the compare feature by generating *comparison tables* for documents discussing multiple related themes, entities, or concepts, as well as *timelines* for documents describing sequences of events. **Augment** enhances content with external data. For instance, *HoloDoc* [32] and *PaperTrail* [46] augment papers with external content, such as references, videos, and figures. Similarly, our system augments documents by providing *information cards* about people and places using Google Search, Image APIs, and Mapbox, based on keywords identified by spaCy. **Extract** enables readers to pull and store elements like keywords and quotes. For example, *DigitalDesk* [62] allows users to extract content from documents for copy-and-paste functionality. RS-Doc implements this feature by generating *keyword lists* that highlight the most central named entities extracted. Finally, **Navigate** improves document navigation through features like a table of contents or bookmarks. For instance, *QOOK* [67] allows users to add bookmarks to specific sections. RS-Doc supports navigation by generating a *table of contents*, created by the LLM based on the document.

Interactions. After launching RS-Doc, a reader can trigger summaries for an in-view document by saying "give me summaries". The system then captures an image of the page currently in view, creates a dynamic image target, and generates summary cards of all types relevant to that page. Summaries are displayed around the page using a *fixed layout* (for example, timeline summaries always appear at the top right of the page and information cards always appear to the left) and remain locked to the page if either the document or reader moves. For example, if a user is reading a magazine about aerial drones, the system automatically detects important keywords such as people, places, and phrases from the page. The system then generates an image, a comparison table, a summary, and information about key people and keywords. Readers can also show or hide specific cards using voice commands ("hide keywords", "show timeline", etc.). Note that the system only supports one card of each type at a time. The user can query for a card multiple times, which will re-prompt the LLM (each card type has its own fixed prompt which the user cannot change).

The decision to lock content to the document was motivated by the limitations noted in prior work like *Holodoc* [32], where the participants expressed frustration and preferred the content to follow their views. Additionally, prior research, such as *Dually Noted* [44], indicates that the participants preferred not to obstruct the view of the page itself.

Document Tracking and Content Generation Performance. To characterize the technical performance of our prototype, we assessed document tracking reliability and system latency. To test tracking, we collected 60 seconds of tracking data each for 20 diverse documents (10 physical, 10 on-screen—including posters, newspapers, patents, and books) recording what fraction of time the document was correctly registered. Tracking was largely reliable for documents with visuals (M=92%, SD=5.4%) but less so for text-only documents (M=64%, SD=3.7%) which had fewer distinct features. Performance varied with lighting (better in well-lit conditions), target size (larger images performed better), and motion (sudden movements caused temporary tracking loss).

We also measured the average latency of three components: AR rendering, network communication, and API responses. AR rendering averaged 100 ms on HoloLens, while network communication via socket.io averaged 40 ms, indicating responsive performance. API callbacks included GPT-4 requests (2.4 s) and our NLP module using spaCy (670 ms). Since we limited responses to a short response (e.g., 150 words for a summary), each GPT-4 latency remains relatively short. These latencies remained largely consistent across our two later prototypes.

4.2 Study Method

Using the RS-Doc prototype, we conducted an initial usability study involving 12 participants (P1-P12, 6 male and 6 female, aged 19-35) recruited via word-of-mouth and compensated $15 for their one-hour participation. The study aimed to evaluate the system's usability, explore participants' preferences for different document enhancement strategies, and

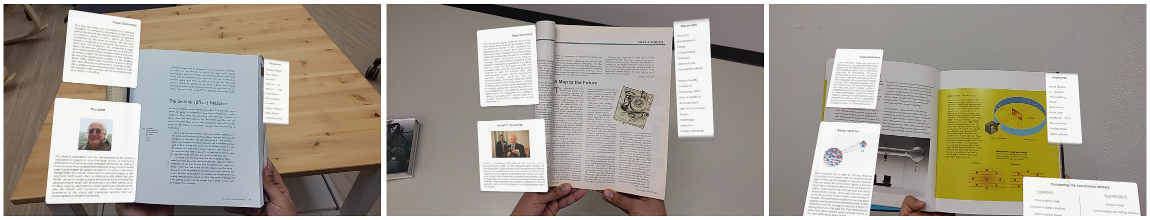

Fig. 4. Examples from our RS-Doc Hololens implementation.

identify limitations for future iterations with real-world use in mind. We conducted the study in a lab environment, where we first introduced participants to the HoloLens 2 and then provided an overview of the system and the set of document enhancements. Participants used our system to explore a variety of materials and media: 1) an A4-sized magazine and an A3-sized book in physical format, and 2) a digital document of their choice. The physical magazine and book contained the same reading content, while the digital documents were selected by participants to reflect their diverse preferences, including blog posts, course assignments, news articles, Wikipedia entries, and digital books. This study aimed to investigate participants' interactions and preferences across different reading materials and formats, rather than focusing on standardized tasks; therefore, we did not explicitly measure or control for difficulty levels. We gave the participants the task to read and understand the documents provided to them and utilize RealitySummary in whichever way they see fit and use whatever combination of enhancements best suited the document.

Throughout the study, we encouraged participants to think aloud and documented their experiences. Following the tasks, we also conducted semi-structured interviews and a Likert scale questionnaire, including a System Usability Scale (SUS). We then performed a thematic analysis using transcripts of both the think-aloud sessions and interviews, from which we identified and refined key themes.

4.3 Results

Overall, participants responded positively to RealitySummary, with most participants agreeing that the prototype was *easy to use* (M=4.3/5, SD=0.98), *easy to understand* (M=4.5/5, SD=0.67), and *well-integrated with the task* (M=4.3/5, SD=0.77) with 5-point Likert-scale questionnaire. Similarly, nearly all participants reported that the summary approaches provided by RealitySummary *helped with comprehension* (M=4.6/5, SD=0.514), *were relevant* to the document content (M=4.7/5, SD=0.65), *improved comprehension speed* (M=4.4/5, SD=0.79), and *augmented their reading experience* (M=4.7/5, SD=0.65). Finally, the System Usability Scale (SUS) returned an overall composite score of *71*, suggesting a reasonable level of usability for an early-stage prototype.

4.4 RS-Doc Observations

Our analysis identified several themes related to participants' experience in comprehending and navigating documents. We also explore participants' comparisons of the visual and text-based summarization techniques as well as preferences for various augmentation techniques.

Page-Level Summary and Keyword Lists Sufficiently Support Reading. Since our RS-Doc system can only capture the currently visible page, our summaries are inherently focused on page-level content rather than the entire book. Interestingly, when asked about their experiences, several participants expressed appreciation for this feature. P5 reported, *"I like the page-by-page summary—having a summary for each page does help make it feel less dense compared* *to a summary of the whole document*." Similarly, P6 emphasized, "*I like that it (the system) gives you keywords for each page and not for the whole document.*"

Structured Overviews can Facilitate Efficient Document Navigation. Both structured overviews and in-context highlights provide useful entry points that make it easy to navigate documents from a top-down perspective, even within individual pages. Four participants (P3, P4, P6, P8) specifically mentioned using structured navigation elements like timelines to better understand the structure and flow of the text, with P4 noting that *"the timeline helped me organize and navigate my thoughts very quickly while reading the text*." In fact, ten participants reported that the system allowed them to focus their reading by taking a "top-down" approach—first examining the high-level ideas and terms, then transitioning into reading the lower-level content as needed. P8 highlighted, "*[the enhancements] give you a foundation before you actually jump inside [the document].*" This approach enabled participants to prioritize their reading, allowing them to "*focus on things that you're more unsure about*" (P1).

Visual Aids, Engagement, and Memorability. Both visual and text-based enhancements offer distinct benefits to readers—visual aids provide quick, high-level overviews, while text summaries capture more complex details. Some participants preferred text-based features for gaining a deeper understanding of specific topics, while others found that visual elements improved their comprehension speed, engagement, and retention. All 12 participants rated the visual summarization techniques as helpful, with eight specifically highlighting the usefulness of images and bio cards. Several participants (P1, P4, P6, P7, P9, P10, P11) noted that visual augmentations allowed them to grasp the document's content more quickly compared to reading text alone, with P11 stating, "*Seeing the visuals helped me grasp what's happening in the document right away*." Participants (P7, P8) also emphasized that visuals enhanced their engagement with the material, with P7 commenting, *I'm not a big fan of reading, so having visuals helps me engage with the content*." Additionally, participants highlighted the potential of visuals to improve memorability, with P8 noting that for historical documents, people cards helped *put a face to a name*." Similarly, P7 observed that while text-based summaries aided understanding, "*graphical elements help me remember it for a longer time*."

4.5 Opportunities for the Second Iteration

Building on these observations, we identified a number of opportunities to explore in our second iteration.

***Reader-Controllable Placement Rather Than Document Tracking*.** An intriguing finding that might contradict previous work is the role of document tracking. For on-demand document enhancement, where dedicated markers or image targets cannot be prepared in advance, unreliable document tracking often leads to frustration. For example, while *HoloDoc* [32] and *Dually Noted* [44] relied on dedicated markers or pre-defined image targets, enabling much more reliable document tracking, our workflow cannot rely on these methods. Instead, we need to rely on a complex and unreliable workflow like image extraction. Moreover, document tracking performed well only with visually rich documents. Participants expressed frustration when dealing with text-heavy documents, particularly under challenging conditions such as poor lighting, hand occlusion, or extreme viewing angles. Therefore, while we acknowledge the benefits of document tracking, it is worth exploring reader-controlled placement as a less frustrating alternative.

***Layouts and Visual Clutter*.** Several participants (P7, P10, P11) noted that around-document augmentation helped them focus on their reading but emphasized that the number of displayed cards and their placements should be adjusted based on the user's needs and context. Displaying too many cards at once can lead to visual clutter and divert attention from the document itself. We observed that participants might become overly reliant on the generated content rather than engaging directly with the reading material. Whether positive or negative, 11 out of 12 participants reported that they could understand the document without actually reading it, relying entirely on the contextual information provided by RealitySummary. P9 even stated, *"I didn't even read the document, but I understand what it's about."* Currently, the system automatically generates various types of cards when relevant, but allowing users to reduce the number of visible cards based on their preferences could help manage visual clutter and maintain focus.

Querying and Contextual Flexibility. Participants highlighted the potential for enhancing the system by enabling queries on the generated summaries, which could expand its adaptability to broader use cases and contexts. The current fixed prompt approach for generating different types of cards suffices for general reading but could be further refined to address more specific reader questions. Several participants (P1, P3, P5) envisioned a more conversational system where they could ask targeted questions beyond the preset summary types. While participants found general summaries and visual information cards broadly useful and adaptable to varying contexts, they also identified opportunities to enhance aids with specific queries for timelines, comparison tables, and keyword lists that better align with their content.

Overall, the controlled nature of the study limited its ability to assess broader applicability and real-world uses. While the initial findings are encouraging, they do not fully reflect how AR summarization and question-answering tools might perform in diverse, unstructured environments. As a result, transitioning to in-the-wild studies is essential to understanding how these augmentations might work in everyday scenarios.

5 RS-WILD: IN-THE-WILD STUDY WITH THE SECOND PROTOTYPE

To explore these kinds of real-world use cases, we developed a second prototype (RS-Wild) that allowed us to study summarization and question-answering in-the-wild. To address performance issues with the Hololens 2 and support more free-ranging use, we instead implemented our RS-Wild prototype with Apple Vision Pro. This provided higher rendering quality for cards, as well as a broader field of view. Building on observations from our prior study, we also simplified the interface—replacing the suite of document-anchored cards from RS-Doc with a single reader-positioned card that readers could dynamically update using spoken queries.

5.1 Implementation

Our RS-Wild implementation for the Vision Pro follows a similar approach to the Hololens but with a few key adjustments. Since the Vision Pro does not yet allow direct access to the camera feed, we implemented a workaround for text extraction. First, we screencast the Vision Pro to an external iPad, which is then connected to a MacBook Air to stream the feed via QuickTime Player's mirroring feature. A web-based interface on the Mac captures the live stream from the Vision Pro. Using the captured feed, we employed Google OCR to extract text, which was then processed by GPT-4 via the OpenAI API. After querying the LLM, data was transmitted from the web-based interfaces to the Vision Pro via Firebase's real-time database. Voice commands were handled via a separate web interface (using the Web Speech API and React.js and hosted on Glitch) loaded on the reader's smartphone.

Interactions. RS-Wild allows readers to request summaries and pose free-form questions about any text visible in their environment by clicking on a mic button on the smartphone interface and asking a question or clicking summaries button. This then triggers the OCR and LLM pipeline. Responses to the most recent request are displayed on a single reader-positioned card which can be placed anywhere in the environment. This added flexibility allowed viewers to recreate any of the individual cards available in RS-Doc as well as pose new unanticipated questions and request new summary types. Using a single manually-positioned card with a transparent background also allowed readers to position that response or summaries in more useful locations in their environment—placing it both close to and apart from documents as appropriate for their current task.

5.2 Study Method

To explore question-answering and summarization in-the-wild we recruited a new cohort of 11 participants (W1-W11, 7 male and 4 female; ages 19-32 years) from the local community. We compensated participants $25 for the 1.5-hour study, which included a one-hour experiment and a 30-minute interview. The participants were first called to our lab, briefed on the procedure, and given the opportunity to pick one or multiple places around the university campus (such as a library, cafeteria, store, hallway, classroom, or various labs) to explore with the system.

Before the experiment, we collected participants' demographic information and reading habits, which we then input into the system as prior knowledge for ChatGPT. Participants familiarized themselves with the Apple Vision Pro by completing the eye-tracking calibration setup. Unlike structured tasks or prepared reading materials, participants were encouraged to use the system with any text they encountered during the study, including documents, books, papers, posters, online news on mobile devices, and other materials in their environment. Participants were encouraged to explore various locations at a local university, rather than using spaces prearranged by the experimenter. The locations they explored included the library, cafeteria, store, hallway, classroom, and research labs. Throughout the one-hour experiment, an experimenter accompanied the participants, observing their usage and encouraging them to think aloud about their experiences. After the experiment, we conducted semi-structured interviews, followed by a Likert scale questionnaire similar to the one used in Study 1. Both the interview and questionnaire took approximately 30 minutes. As in the previous usability study, we analyzed participants' questionnaire responses and reflections to identify high-level trends and opportunities for augmented reading tools.

5.3 Results

We gathered feedback from participants on various aspects, including the reading experience (immersiveness, understanding, distraction), perceptions of the system (content and query comprehension, trustworthiness, and privacy concerns), and overall usability (content placement and comfort using the MR headset). In general, participants reported feeling more *immersed and interested* while reading (M=3.5/5, SD=1.04), and they perceived the system as *effectively interpreting their queries* (M=3.3/5, SD=1.00). Their responses on the *content placement* within the system were also positive overall (M=3.4/5, SD=1.29).

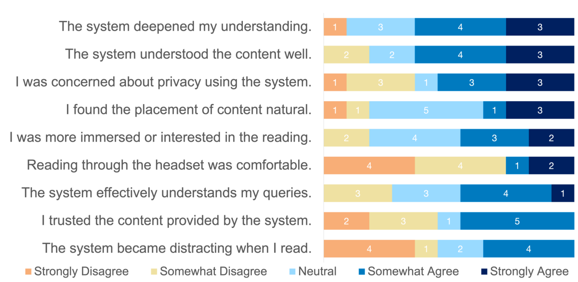

Fig. 5. In-the-wild study questionnaire responses.

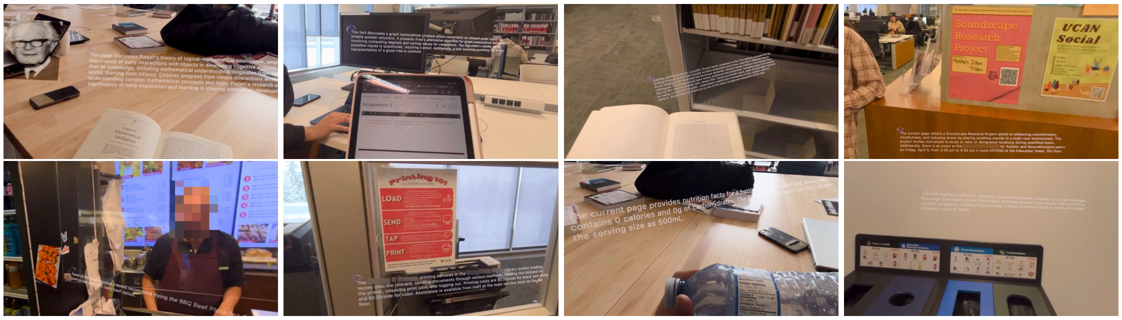

Fig. 6. Snapshots captured from the Vision Pro during our in-the-wild study.

Participants had varied opinions on whether the system became *distracting* throughout the reading process (M=2.6/5, SD=1.37) and whether they tended to *trust the accuracy and reliability* of contents provided by the system (M=2.8/5, SD=1.25). *Privacy concerns* were also noted by the majority (6/11), and discomfort using the Vision Pro headset was prevalent among the respondents (8/11). These findings are detailed in Figure 5.

5.4 RS-Wild Observations

Observations from the in-the-wild deployment highlighted a wider range of use cases for augmented reading as well as some of the practical trade-offs associated with real-world use.

Diverse Reading Content. Participants engaged with a variety of books (spanning textbooks, encyclopedias, fiction, and non-fiction), a variety of other types of documents (including academic papers, handouts, and class assignments), and diverse digital content accessed through smartphones and laptops (like social media posts, online news articles, and e-commerce websites). Beyond conventional reading materials, participants also used our system with everyday objects and text in the environment, such as restaurant menus, printer instructions, product nutrition labels, furniture assembly guides, posters, signage on trash bins, and advertisements. Notably, some participants even applied our system to their handwritten notes from lectures.

Incorporating Spatial and Tangible Interactions. Participants appreciated the unique affordances offered by mixed reality interfaces, such as the opportunities for spatial and tangible exploration. In a library setting, for example, participants could naturally navigate through bookshelves (*spatial exploration*), physically pick up a book (*tangible interaction*), and browse through its pages to automatically receive summaries.W5 and W10 even leveraged our system to provide summaries or recommendations for entire collections of books on a bookshelf, not just content from a single page. Furthermore, W1 demonstrated the capability to synthesize information across multiple books on a table, requesting summaries or connections between various books and pages. Such observations highlight how mixed reality interfaces enable spatial and tangible interactions that go beyond the limitations of traditional screen-based interfaces, which can only capture and summarize visible content.

Creative Uses of AI-Generated Responses. We observed several unique and unexpected use cases that took advantage of the LLM's capacity for free-form questioning and answering. For instance, when reading a physics textbook, W8 asked the system to generate ten quizzes to help him understand quantum mechanics, transforming the textbook into a more active and personalized learning tool. Other common usages included translations (W4, W5, W6), learning aids (W4, W8), and personalized content recommendations (W5, W10). For example, W8 applied the system to his handwritten calculus notes to find answers to particular questions. Beyond traditional reading applications, the AI's capabilities supported practical decision-making; W4 compared nutritional information of products on store shelves and menu items to determine healthier options, while W5 used it for guidance on waste disposal and printer operations. The LLM's ability to process a broad spectrum of inquiries clearly expanded the system's applicability and utility.

Proactive Summaries vs Reader-Driven Question Answering. Our system offers both 1) proactive assistance through automatic summary generation and 2) reader-driven assistance where readers manually ask questions. Participants shared their insights on these two functionalities by comparing the benefits and limitations. Readers appreciated how proactive summaries could often provide a useful initial overview of unfamiliar topics. Preferences for proactive summaries varied with reading habits. For example, infrequent readers like W6 and W10 found these proactive summaries helpful in deciding whether to read the content or understand it without reading the full text. On the other hand, W10, who is a more frequent reader, found the constant summary updates unnecessary. In contrast, reader-driven question answering allowed for a more focused engagement on specific topics. For instance, W1 enjoyed the "rapid-fire" questioning with a physics textbook, noting that it encouraged them to ask more questions than they would with traditional ChatGPT. W2 valued how questioning was woven into reading, preserving thought flow without the disruption of toggling between reading and ChatGPT interfaces. Overall, participants favored the reader-driven feature due to its applicability to diverse situations, yet many appreciated having both options for different purposes: proactive features for implicit skimming and reader-driven for addressing more specific questions.

5.5 Opportunities for the Third Iteration

Overall, the design of the second prototype enhanced its applicability and stability for real-world use. Compared to the earlier Hololens 2 prototype, the Vision Pro's superior rendering quality and broader field of view provided a better reader experience. Additionally, the simple system design greatly improved usability despite the absence of document tracking. While RS-Wild showed great potential, we also identified several feature-level opportunities for improvement.

Sustained History and Recall. Participants reported that the minimalist design effectively reduced visual clutter. However, it also limited the system's overall utility. Although readers appreciated the simple interface, many envisioned options to save or revisit previously-generated responses. Several participants suggested that a history feature or the ability to store cards for future reference would substantially enhance the reader experience.

Opportunities for Enhanced Screen Viewing for Extended Use. Participants highlighted the potential for improving the experience of viewing phone and computer screens through the Vision Pro, particularly for longer viewing periods. While the current approach of enlarging text worked well for short-term tasks like quickly reviewing an assignment, participants envisioned improvements that could better support extended work sessions, especially when interacting with computing devices. Displaying the reader's screen visually in high resolution would further align the system with the demands of realistic, work-related tasks in everyday contexts.

Our in-the-wild study was inherently limited by its short-term, public use — with each session lasting only one hour per participant. As a result, participants did not have the opportunity to fully engage with the system in their personal workflows. For instance, due to privacy concerns related to screen recording, participants rarely used our system for more personalized content, such as emails, messages, or calendars. Furthermore, since participants were encouraged to explore various scenarios, they tended to experiment broadly rather than focusing on any specific task in depth. While the in-the-wild study provided valuable insights into real-world use cases, it only captured short-term interactions in public spaces. There is still much to learn about the sustained, regular use of reading assistants in daily life, particularly in how readers interact with information over time, revisit previous queries, or support ongoing tasks.

6 RS-DIARY: DIARY STUDY WITH THE THIRD PROTOTYPE

To explore more realistic longer-term use cases, we developed a third prototype variant (RS-Diary) incorporating feedback from the in-the-wild study. This version (Figure 7) introduced the ability to create, retain, and organize multiple information cards, along with the ability to view or mirror the content from external devices directly to the Vision Pro.

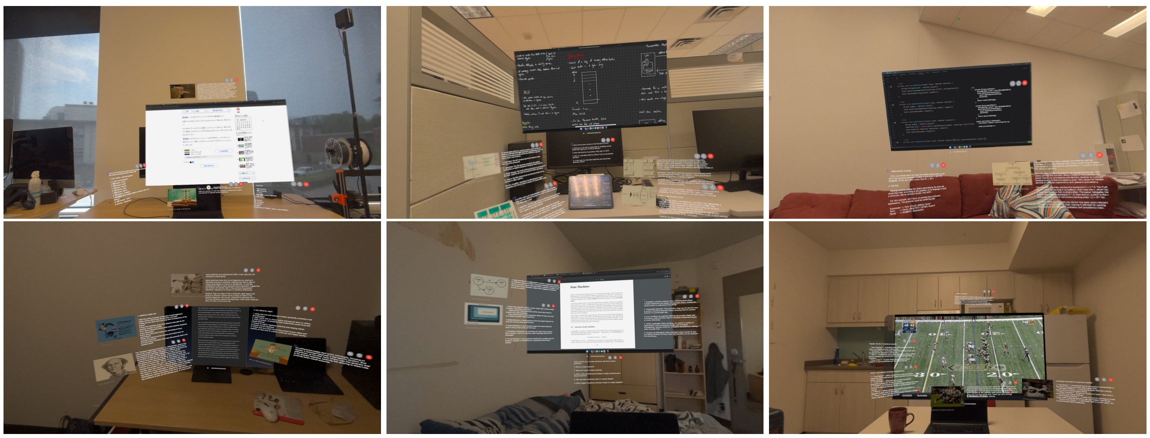

Fig. 7. Snapshots captured from the Vision Pro during our diary study.

6.1 Implementation

RS-Diary builds on the earlier Vision Pro implementation with two key improvements: 1) Readers can now add multiple draggable cards on-demand by either making a new query or clicking an "add" button (which spawns a new summary card based on the document content currently in front of them). They can also spatially anchor these cards anywhere in their immediate environment and can re-select existing cards to update them with new prompts or text content form the environment. These cards are stored as JSON objects in a real-time database. 2) The system allows readers to mirror their device screen to the Vision Pro for high-resolution viewing. This functionality—implemented through a WebRTC video call or via Mac Virtual Display on the Vision Pro—lets readers view their computer or tablet screens at full resolution in the MR environment and also allows RS-Diary to more accurately capture and process on-screen content. This increased fidelity allowed readers to perform a wide variety of everyday reading, coding, browsing, and watching activities for extended periods without the legibility and eye strain experienced with video passthrough.

6.2 Study Method

Five authors (A1-A5[1]) used the system over the course of 3 weeks, integrating it into their day-to-day activities and accumulating 30 total hours of use. Due to physical fatigue and battery limitations of the Vision Pro, each session

[1]The first, fourth, fifth, seventh, and sixth authors of this submission.

lasted less than 2 hours. The authors independently documented their use cases, insights, frustrations, thoughts, and reflections while using RS-Diary, which we later thematically coded. During the study, each participating author recorded their observations via a phone microphone, following a think-aloud process. Additionally, the first author conducted follow-up interviews with A2-A5 at the end of the study to gain deeper insights into their experiences.

6.3 Results

Throughout the diary study, participants used the system in various settings, including labs, kitchens, outdoor areas, and offices, though the majority of usage took place at desks in office environments. They engaged with a wide range of reading materials, such as academic papers, news articles, code, emails, calendars, Google Scholar, classroom assignments, lecture slides, textbooks, and YouTube videos. Participants quickly adapted the system to fit a variety of personal, academic, and technical tasks. For example, A4 used the assistant not only to summarize email content but also to gather higher-level insights, such as identifying how many emails were sent by a particular sender. A2 utilized the system to support programming assignments, where they used the system to break down complex coding problems into manageable steps and provided detailed explanations of specific code snippets. A2 also leveraged the system to create learning aids, such as generating to-do lists and quiz questions based on problem descriptions. The ability to spatially organize tasks and concepts allowed him to effectively manage his workload, maintain focus, and dynamically track his progress. Additionally, in his language learning efforts, A2 used the assistant to generate contextual vocabulary lists from Japanese reading materials, which helped him focus on the most relevant and frequently used words. A5 employed it to extract a step-by-step sequence of keyboard commands from an article on rebooting a MacBook. Overall, the diary study highlighted the system's flexibility in supporting diverse tasks, from email management and coding to technical troubleshooting, language learning, and academic research.

6.4 RS-Diary Observations

From these experiences, participants shared various insights, frustrations, and potential future improvements. Our diary study also provided insight into the use of an always-on reading assistant in more realistic work-oriented settings.

Quick Summary Responses Meet Many Immediate Needs. Participants valued the system's ability to deliver quick, concise responses, which were often sufficient for immediate clarifications. A2, for instance, appreciated how the system provided brief answers when they needed a rapid comparison of philosophical concepts, like distinguishing between act and rule utilitarianism. In many cases, just a high-level summary of the page in view was enough to satisfy their needs. A4 found this method to be a useful shortcut, as pressing the virtual "add" button was faster and less disruptive than speaking and often provided exactly the information they were seeking. These simple, proactive summaries were particularly beneficial for tasks that required quick comprehension and minimal interruption.

Using Generated Cards as Peripheral References. Participants highlighted the advantages of cards that could exist in the periphery of their current workspace. For instance, A2 used the system as a reading aid while learning Japanese. Instead of frequently opening a dictionary, which disrupted the reading flow, they generated vocabulary flashcards around the document. This provided the needed language support without interrupting the reading and learning process. A4 also noted that they would often cluster and group these generated cards thematically to the sides of the display for quick reference and then delete them once they were no longer needed. Similarly, A1 used the system to "pin" generated code versions and key concepts around their workspace, allowing for easy access without breaking focus and enabling them to quickly revisit important points as needed.

Spatial Organization for Recall and Multitasking. Perhaps the biggest emergent pattern in the study was how all participants leveraged the system's spatial organization features, similar to arranging post-it notes on a board. Both A1 and A2 utilized this capability to group related concepts together. For example, A2 clustered cards on CPU architecture, placing related topics like buses and registers nearby, much like organizing post-its by theme. This spatial arrangement enabled quick access and improved recall, as readers could visually track where specific information was placed. Similarly, A1 pinned cards documenting different problem-solving methods across their workspace, ensuring easy access when needed. These organizational structures supported memory retention and task management by allowing readers to arrange information in a way that aligned with their thought processes.

6.5

.

.

.

7 LESSONS LEARNED FROM THREE STUDIES

Drawing on results from across our prototyping and evaluations, we close by showcasing the unique benefits of combining MR and AI tools, enumerating potential use cases, and highlighting opportunities for future research.

7.1 Key Insights from the Three Studies

Minimal Context Switching Enhances Focus and Flow. Participants across all three studies appreciated the integration of LLM into the MR interfaces, in part, because it helped them minimize context switching, allowing them to maintain greater focus. Several participants reported that the system's automatic text extraction feature helped them stay in a flow state. For example, A2 noted that the interface allowed them to concentrate on their work without the mental disruption of switching between tasks or manually inputting context, even when using a computer screen. In particular, members of the author team in the RS-Diary study remarked on how the seamless workflow kept them engaged, enhancing productivity and reducing cognitive load. A1 highlighted how the system functioned like a "teacher" by offering real-time assistance without requiring them to step away from their tasks. Participants also remarked that on-demand support boosted their efficiency during complex tasks like coding or writing, where focus is critical.

A key benefit of this workflow is its support for *implicit inputs*, which dramatically lower the barrier to asking questions of LLMs. This automatic, continuous content capture broadens the range of applications beyond what participants initially envisioned. For instance, W4 mentioned they had never thought to type a question into ChatGPT when looking at a product label. Even for common reading tasks, they appreciated that the summary was automatically generated and updated in the background as they flipped through pages in a library. Overall, the always-on, implicit assistant made it effortless to switch context and maintain a flow state.

Potential Risks of AI-Driven Augmented Reading. Participants also expressed several concerns about potential risks, notably privacy issues due to the possibility of constant surveillance. Despite this concern, most participants expressed interest in continue to use these kinds of systems, provided they could control when it was active. For example, W1 expressed the desire to capture all text throughout her entire life, allowing for an extensive searchable archive of everything she has encountered. Such use cases, however, also pose social privacy questions—particularly when systems might record other individuals or their content.

Additionally, participants also mentioned the risk of decreased motivation and skill development. For example, W6 mentioned that relying too heavily on the system for translations could reduce the motivation to learn new languages. Similarly, W11 worried that immediate access to answers for every question might contribute to a broader intellectual complacency. Overall, participants wanted the flexibility to choose when and how the system would be used to mitigate these risks.

Trust of AI-Generated Content. Recognizing the known shortcomings or current LLMs, participants also approached the system's responses with caution, often verifying the answers with the reading material and highlighting the necessity of consulting additional sources. There was a general agreement among participants in all studies that they would not fully trust the system's responses to be complete or correct. Conversely, we noted some intriguing discussions regarding how participants perceived our system, either as a mere tool or more like a personal companion. While most participants engaged with the system as a conventional chatbot, some regarded it more as a personal companion than a mere tool. Notably, W3 and W5 described the system as a friend for discussing and inquiring about their readings. The system's ability to provide always-on, immediate responses facilitates this personification.

7.2 Unique Benefits of Combining MR and AI

We identified two unique benefits of combining MR with AI: 1) implicit input and 2) peripheral output.

Always-On Implicit Input. The always-on camera in MR enables implicit input, distinguishing it from the explicit input required in desktop interfaces, such as opening an app, typing queries, or copy-pasting contextual text. Participants in in-the-wild and diary studies appreciated the system's ability to capture information without requiring explicit effort. This always-on input allows for hands-free exploration in both spatial and temporal contexts, such as scanning entire bookshelves as a query or utilizing the history of flipped pages. Implicit input reduces cognitive load and minimizes context switching, as readers are not required to actively specify what information should be used for the LLM's context.

Peripheral and Background Output. We found that MR's spatial output allowed information to be presented in a peripheral, background manner. In contrast, other output modalities, such as audio or screen-based displays, often require the reader's foreground attention to consume generated content, disrupting focus and interrupting workflow. With MR and peripheral outputs, however, AI-generated content becomes less intrusive, as it can be easily ignored if not needed, simply by choosing not to look at it. For example, participants in the diary study often appreciated this feature for learning aids, such as understanding textbooks or accessing always-on vocabulary references when learning a foreign language. This peripheral output is a unique aspect of the integration of MR and AI. We believe this capability would promote more *calm* human-AI interactions in the future, reducing disruptions to focused attention by combining implicit input with unobtrusive background support.

7.3 Emergent Use Cases for Mixed Reality Text Augmentation

Both the in-the-wild and diary studies revealed a variety of interesting emergent use cases for MR text augmentation. To give a more comprehensive view of how readers might with these systems, we identify seven distinct approaches:

Summarizing Content. The primary use case was to summarize content. This was the most commonly observed behavior across various contexts. Participants summarized a wide range of texts, including research papers, blog posts, news articles, books, and instructional guides.

Extracting Information. Participants also utilized the system to extract key information, which can be translated into a more consumable format. This aligns with design elicitation results, which suggest tables of contents, keyword lists, timelines, and comparison tables. For example, one participant extracted vocabulary flashcards to assist with foreign language reading, while another participant created a table of keyboard shortcut commands from a troubleshooting guide to use as a reference. In these examples, the consumable format, such as tables or vocabulary lists, serves as helpful references for their tasks.

Querying External Sources. Participants frequently used extracted text as a query to retrieve additional information from external sources. Echoing the information types identified in our elicitation study, participants queried specific names or keywords from their reading materials. For instance, one participant retrieved details about players, teams, and standings from a sports-related article, while another fact-checked claims on-demand when reading news articles.

Asking for Suggestions. Participants frequently asked for recommendations or suggestions based on the current context—for instance, asking for book recommendations when looking at a bookshelf. Outside the context of reading, one participant asked for advice on which trash bin can be used based on disposal instructions, while others inquired about healthier menu options or products. They also asked for recommendations from external sources, including asking for related research papers when reading academic articles.

Aggregating Information. An interesting use case involved using the system to aggregate high-level information. Unlike simple information extraction, aggregation often involves operations like filtering and counting. For example, some participants used the system to count emails based on specific filters, while others filtered and counted first-author papers from a researcher's Google Scholar page. This allowed readers to perform quick, approximate queries that would normally require more complex searches.

Transforming Content. Participants also used the system to transform their reading content into interactive, actionable outputs. For instance, some participants applied our system to generate quizzes from class notes or textbooks, while others decomposed high-level assignments into step-by-step to-do lists. One participant asked the system to create reminder cards that remained spatially persistent. Participants also use our system to translate from one language to another. Unlike information extraction, these tasks involved converting content into actionable, interactive outputs.

Generating New Content. Finally, participants generated new content based on existing text. In one example, a participant requested multiple variations of a written paragraph, which they then organized spatially. Others used the system to create code based on instructions or to debug existing code snippets.

These emergent use cases highlight novel ways of interacting with AI-generated information, allowing readers to engage with individual snippets and convert them into dynamic outputs. They also illustrate how participants adapted the assistant to handle different types of content, which suggests a broad space of possible designs for such systems.

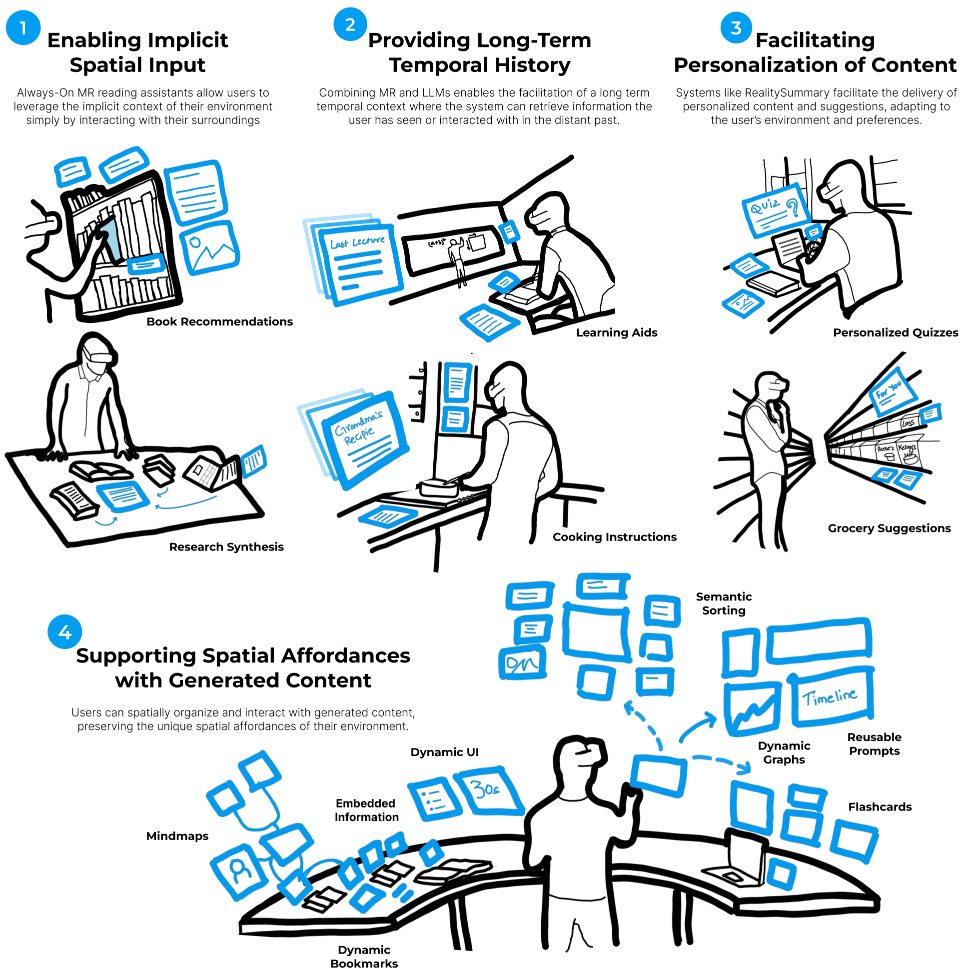

Fig. 8. For future opportunities, we summarize possible new interactions enabled by mixed-reality text augmentation through our findings from three studies: 1) Enabling implicit spatial input like browsing and comparing entire bookshelves for recommendations or walking around tables with documents spread out and viewing connections and details among them. 2) Providing long-term history such as connecting past lectures to current topics for better recall and understanding or revisiting meaningful past information like grandmother's recipes and connecting them to present activities for providing instructions. 3) Personalizing content such as generating custom quizzes or providing grocery recommendations based on health status. 4) Supporting spatial affordances like flexible sorting, reusable prompts, mindmaps, dynamic bookmarking, or embedding information on physical objects.

7.4 New Interactions and Future Opportunities for Always-On MR Reading Assistant

As both MR and LLM technologies continue to evolve rapidly, considerable opportunities exist for tools that more richly integrate them. Based on our three studies, we identify several key opportunities for future research and design.

New Interactions Enabled by Combining MR and LLMs. The combination of always on camera input with mixed-reality spatial output facilitates new ways of interacting with AI unique from traditional screen-based workflows such as 1) **Enabling Implicit Spatial Input**. RealitySummary uses the implicit context of the user's environment, enabling tasks not possible before, such as comparing entire bookshelves for recommendations or browsing in a grocery store to find the healthiest choices through natural interaction. In contrast, traditional LLMs require specific queries and context, making such natural interaction impossible. 2) **Providing Long-Term Temporal History** RealitySummary facilitates long-term temporal context by remembering and retrieving information the user has seen or interacted with in the distant past, enabling new AI use cases. For instance, a student reads a complex equation on a whiteboard but doesn't take notes; months later, RealitySummary recalls and integrates it into their assignment in the classroom. Unlike traditional LLMs that require users to provide information explicitly, RealitySummary captures implicit everyday interactions, ensuring nothing is forgotten. Its always-on system differs from previous MR reading augmentation by leveraging users' history for more relevant recommendations and augmentations. 3) **Facilitating Personalization of Content** Having the context of the user and their environment enables unique capabilities, such as returning personalized content and suggestions. For instance, many users have used RealitySummary to create tailored learning aids like quizzes, customized to their current context (e.g., an assignment) and personal preferences/experiences (e.g., their educational background as an undergraduate CS student). Compared to traditional screen-based LLMs or MR reading assistants, RealitySummary enables content personalization based on both the current view and the user's experiences, a level of contextual integration not previously seen. and lastly, 4) **Supporting Spatial Affordances** RealitySummary enables spatial interactions with generated content by allowing users to freely organize it in 3D space. Unlike traditional LLM interfaces, where information output is one-dimensional, RS leverages spatial arrangement for use cases (which the participants have tried) like dynamic bookmarking, peripheral references (e.g., flashcards), anchored information, or grouping cards by theme. A future version could automatically filter, sort, or group cards and information by themes or generate semantic mind maps using LLM capabilities.

Improving Organization and Filtering of Interactive Cards. Participants expressed the need for more efficient ways to manage the growing number of spatially organized content cards. While some diary study participants manually arranged these cards around documents, others suggested more advanced organizational tools. Features like sorting, snapping, filtering, or semantically organizing cards could simplify workflows and reduce clutter. Such capabilities would be particularly beneficial for long-term tasks, where the accumulation of content can become overwhelming. Integrating these organizational tools would enhance usability, supporting readers in managing dynamic and task-specific information over time, which enables better workspace management for daily tasks.

Balancing Reusable Prompts and On-Demand Custom Queries. Our studies highlighted the importance of accommodating both reusable prompts and custom queries. Each has its advantages: prepared prompts can generate complex outputs such as timelines or comparison tables, while custom queries offer readers flexibility in generating content. The diary study participants expressed a desire for dynamic interfaces that could quickly reuse prompts and adapt to different contexts. For example, some suggested having customizable buttons that automatically generate summaries or lists in a consistent style. Others mentioned the need for a more efficient way to reuse queries across tasks, such as repeatedly generating lists or comparisons for class topics. Participants also envisioned the system generate simple widgets or dynamic functionalities like a timer or a check list where the information becomes dynamic and interactive. This feedback highlights the importance of designing adaptive interfaces that allow readers to seamlessly reuse and modify their prompts, improving efficiency and overall satisfaction.

Facilitating Seamless Transitions Between Screen and Space. Participants also valued the ability to drag content cards from the system into their computer, such as pasting generated answers directly into coding tasks. Some participants also envisioned this feature being useful for writing tasks, allowing for a smooth transfer of generated content. Future systems should explore leveraging cross-device interactions between MR and other devices to facilitate seamless transitions, enhancing the workflow between screen-based and spatial interactions.

Expanding Beyond Text Input to Capture Images and Everyday Objects. Participants expressed a strong desire for the system to recognize not only text but also images. For example, the in-the-wild study revealed the limitations of inquiring about a diagram that the system could not recognize. Readers also anticipated broader applications beyond text recognition, such as generating recipes based on ingredients in the fridge or providing information on real-world objects. Enhancing the system with multi-modal LLM capabilities could address these needs, expanding its utility beyond text-based interactions.

Personalization Through Always-On History. Participants showed interest in an "always-on" history feature that would allow them to recall previous interactions with the system, creating a more personalized experience. They speculated that such a system could eventually serve as a personalized assistant, continuously capturing and understanding context through an always-on camera. Although current hardware limitations, such as battery life and physical fatigue, prevent continuous long-term use, we believe future MR interfaces will seamlessly integrate with AI to provide both context-aware and personalized assistance based on readers' daily activity histories.

Enabling Control and Awareness of Capture. Privacy and trust were among participants' top concerns throughout our studies. Participants wanted the flexibility to opt in and tell the system when to capture their environment. Future designs of such systems can take a privacy-first approach, where the reader is always aware of the system's intentions and input. While the system's implicit input offers significant benefits, it also raises privacy concerns due to the automatic capturing without the reader's constant awareness. Ensuring privacy from a system designer's perspective is an excellent starting point. Tools like Reframe [47] can assist in training designers to identify privacy threats effectively. However, addressing these threats by developing suitable mitigation techniques remains a significant challenge. Adopting an adaptive interface paradigm, where features are selectively enabled or disabled based on user preferences, offers a promising direction [45]. Future designs should give readers control over what data is captured. As also noted in the diary study, participants expressed a desire for more transparency and control over the captured content. Another approach could involve providing visual feedback, allowing readers to see and select the specific areas or information being captured, such as through gaze interaction, which would help mitigate privacy concerns and prevent the inclusion of irrelevant or personal content. Future systems should balance the benefits of implicit input with privacy considerations by offering greater control and visual feedback.

Transparent Generation. In addition, participants expressed a lack of trust in AI-generated content. System designers can allow for a transparent approach in recommending content where the system indicates its level of competence or surety of the answer or suggestion. This would make it clear to the reader and give them agency to utilize the content as they see fit. Hallucinations are an emergent property of LLMs, but many solutions like Chain of Thought prompting (CoT) [61] can help the LLM reason through and explain its answer in a more transparent context. Lastly, designers should be mindful of the mental effects and copyright infringements the LLM may produce and either correct or make the reader aware wherever applicable.

8 CONCLUSION

RealitySummary is the first on-demand MR reading assistant that combines OCR and LLM technologies to generate spatially-situated on-demand reading aids for both physical and digital documents. Through our empirical studies, we aimed to highlight new possibilities for on-demand, AI-powered reading support. This paper presented an iterative development of the system, along with key findings and lessons learned from three studies. We hope our work will inspire the HCI community to further explore the potential of integrating MR and AI in future applications.

REFERENCES

[1] Jiban Adhikary and Keith Vertanen. 2021. Text entry in virtual environments using speech and a midair keyboard. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2648–2658.

[2] Karim Benharrak, Florian Lehmann, Hai Dang, and Daniel Buschek. 2022. SummaryLens–A Smartphone App for Exploring Interactive Use of Automated Text Summarization in Everyday Life. In *27th International Conference on Intelligent User Interfaces*. 93–96.

[3] Adithya Bhaskar, Alexander R Fabbri, and Greg Durrett. 2022. Zero-Shot Opinion Summarization with GPT-3. *arXiv preprint arXiv:2211.15914* (2022).

[4] Mark Billinghurst, Hirokazu Kato, and Ivan Poupyrev. 2001. The magicbook-moving seamlessly between reality and virtuality. *IEEE Computer Graphics and applications* 21, 3 (2001), 6–8.

[5] Wolfgang Büschel, Annett Mitschick, and Raimund Dachselt. 2018. Here and now: Reality-based information retrieval: Perspective paper. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 171–180.

[6] Julia Cambre, Alex C Williams, Afsaneh Razi, Ian Bicking, Abraham Wallin, Janice Tsai, Chinmay Kulkarni, and Jofish Kaye. 2021. Firefox voice: an open and extensible voice assistant built upon the web. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[7] Joseph Chee Chang, Amy X Zhang, Jonathan Bragg, Andrew Head, Kyle Lo, Doug Downey, and Daniel S Weld. 2023. CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[8] Xiang'Anthony' Chen, Chien-Sheng Wu, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. Marvista: A Human-AI Collaborative Reading Tool. *arXiv preprint arXiv:2207.08401* (2022).

[9] Zhutian Chen, Wai Tong, Qianwen Wang, Benjamin Bach, and Huamin Qu. 2020. Augmenting static visualizations with paparvis designer. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[10] Kun-Hung Cheng. 2017. Reading an augmented reality book: An exploration of learners' cognitive load, motivation, and attitudes. *Australasian Journal of Educational Technology* 33, 4 (2017).

[11] Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*. PMLR, 354–372.

[12] Neil Chulpongsatorn, Mille Skovhus Lunding, Nishan Soni, and Ryo Suzuki. 2023. Augmented Math: Authoring AR-Based Explorable Explanations by Augmenting Static Math Textbooks. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–16.

[13] Hai Dang, Karim Benharrak, Florian Lehmann, and Daniel Buschek. 2022. Beyond Text Generation: Supporting Writers with Continuous Automatic Text Summaries. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–13.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2019).

[15] Andreas Dünser, Lawrence Walker, Heather Horner, and Daniel Bentall. 2012. Creating interactive physics education books with augmented reality. In *Proceedings of the 24th Australian computer-human interaction conference*. 107–114.

[16] Wafaa S El-Kassas, Cherif R Salama, Ahmed A Rafea, and Hoda K Mohamed. 2021. Automatic text summarization: A comprehensive survey. *Expert systems with applications* 165 (2021), 113679.

[17] Katherine M Everitt, Meredith Ringel Morris, AJ Bernheim Brush, and Andrew D Wilson. 2008. Docudesk: An interactive surface for creating and rehydrating many-to-many linkages among paper and digital documents. In *2008 3rd IEEE International Workshop on Horizontal Interactive Human Computer Systems*. IEEE, 25–28.

[18] Google. [n. d.]. Semantic Reactor. https://research.google.com/semanticexperiences/semantic-reactor.html. Accessed: 2023-03-18.

[19] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. 2022. News summarization and evaluation in the era of gpt-3. *arXiv preprint arXiv:2209.12356* (2022).

[20] Raphael Grasset, Andreas Dunser, and Mark Billinghurst. 2008. The design of a mixed-reality book: Is it still a real book?. In *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*. IEEE, 99–102.

[21] Aditya Gunturu, Yi Wen, Nandi Zhang, Jarin Thundathil, Rubaiat Habib Kazi, and Ryo Suzuki. 2024. Augmented Physics: Creating Interactive and Embedded Physics Simulations from Static Textbook Diagrams. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–12.

[22] Aakar Gupta, Bo Rui Lin, Siyi Ji, Arjav Patel, and Daniel Vogel. 2020. Replicate and reuse: Tangible interaction design for digitally-augmented physical media objects. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[23] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.

[24] Wahyu Nur Hidayat, Muhammad Irsyadul Ibad, Mahera Nur Sofiana, Muhammad Iqbal Aulia, Tri Atmadji Sutikno, and Rokhimatul Wakhidah. 2020. Magic Book with Augmented Reality Technology for Introducing Rare Animal. In *2020 3rd International Conference on Computer and Informatics Engineering (IC2IE)*. IEEE, 355–360.

[25] Juan David Hincapié-Ramos, Sophie Roscher, Wolfgang Büschel, Ulrike Kister, Raimund Dachselt, and Pourang Irani. 2014. cAR: Contact augmented reality with transparent-display mobile devices. In *Proceedings of The International Symposium on Pervasive Displays*. 80–85.

[26] Ken Hinckley, Xiaojun Bi, Michel Pahud, and Bill Buxton. 2012. Informal information gathering techniques for active reading. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1893–1896.

[27] Shradha Holani, Akashdeep Bansal, and Meenakshi Balakrishnan. 2019. Pushpak: Voice command-based ebook navigator. In *Proceedings of the 16th International Web for All Conference*. 1–2.

[28] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[29] Tereza Gonçalves Kirner, Fernanda Maria Villela Reis, and Claudio Kirner. 2012. Development of an interactive book with augmented reality for teaching and learning geometric shapes. In *7th Iberian Conference on Information Systems and Technologies (CISTI 2012)*. IEEE, 1–6.

[30] Jaewook Lee, Jun Wang, Elizabeth Brown, Liam Chu, Sebastian S Rodriguez, and Jon E Froehlich. 2024. GazePointAR: A Context-Aware Multimodal Voice Assistant for Pronoun Disambiguation in Wearable Augmented Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–20.

[31] Luis A Leiva. 2018. Responsive text summarization. *Inform. Process. Lett.* 130 (2018), 52–57.

[32] Zhen Li, Michelle Annett, Ken Hinckley, Karan Singh, and Daniel Wigdor. 2019. Holodoc: Enabling mixed reality workspaces that harness physical and digital content. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[33] Chunyuan Liao, Qiong Liu, Bee Liew, and Lynn Wilcox. 2010. Pacer: fine-grained interactive paper via camera-touch hybrid gestures on a cell phone. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2441–2450.

[34] Geoffrey Litt, Max Schoening, Paul Shen, and Paul Sonnentag. 2022. Potluck: Dynamic documents as personal software.

[35] Damien Masson, Sylvain Malacria, Edward Lank, and Géry Casiez. 2020. Chameleon: Bringing Interactivity to Static Digital Documents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[36] Fabrice Matulic and Moira C Norrie. 2012. Supporting active reading on pen and touch-operated tabletops. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*. 612–619.

[37] Fabrice Matulic and Moira C Norrie. 2013. Pen and touch gestural environment for document editing on interactive tabletops. In *Proceedings of the 2013 ACM international conference on Interactive tabletops and surfaces*. 41–50.

[38] Fabrice Matulic, Moira C Norrie, Ihab Al Kabary, and Heiko Schuldt. 2013. Gesture-supported document creation on pen and touch tabletops. In *CHI'13 Extended Abstracts on Human Factors in Computing Systems*. 1191–1196.

[39] Hrim Mehta, Adam Bradley, Mark Hancock, and Christopher Collins. 2017. Metatation: Annotation as implicit interaction to bridge close and distant reading. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 5 (2017), 1–41.

[40] Kyzyl Monteiro, Ritik Vatsal, Neil Chulpongsatorn, Aman Parnami, and Ryo Suzuki. 2023. Teachable reality: Prototyping tangible augmented reality with everyday objects by leveraging interactive machine teaching. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–15.

[41] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[42] Matt Payne. 2022. State of the Art GPT-3 Summarizer For Any Size Document or Format. https://www.width.ai/post/gpt3-summarizer.

[43] Morgan N Price, Bill N Schilit, and Gene Golovchinsky. 1998. XLibris: The active reading machine. In *CHI 98 conference summary on Human factors in computing systems*. 22–23.

[44] Jing Qian, Qi Sun, Curtis Wigington, Han L Han, Tong Sun, Jennifer Healey, James Tompkin, and Jeff Huang. 2022. Dually noted: layout-aware annotations with smartphone augmented reality. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.

[45] Shwetha Rajaram. 2024. Enabling Safer Augmented Reality Experiences: Usable Privacy Interventions for AR Creators and End-Users. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*. 1–8.

[46] Shwetha Rajaram and Michael Nebeling. 2022. Paper trail: An immersive authoring system for augmented reality instructional experiences. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–16.

[47] Shwetha Rajaram, Franziska Roesner, and Michael Nebeling. 2023. Reframe: An Augmented Reality Storyboarding Tool for Character-Driven Analysis of Security & Privacy Concerns. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–15.

[48] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822* (2018).

[49] Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James Landay. 2016. Speech is 3x faster than typing for english and mandarin text entry on mobile devices. *arXiv preprint arXiv:1608.07323* (2016).

[50] A. J. Sellen and R. H. R. Harper. 2002. *The Myth of the Paperless Office*. MIT Press.

[51] Brett E Shelton and Nicholas R Hedley. 2004. Exploring a cognitive basis for learning spatial relationships with augmented reality. *Technology, Instruction, Cognition and Learning* 1, 4 (2004), 323.

[52] Jannis Strecker, Kimberly García, Kenan Bektaş, Simon Mayer, and Ganesh Ramanathan. 2022. SOCRAR: Semantic OCR through Augmented Reality. In *Proceedings of the 12th International Conference on the Internet of Things*. 25–32.

[53] Hariharan Subramonyam, Steven M Drucker, and Eytan Adar. 2019. Affinity lens: data-assisted affinity diagramming with augmented reality. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.

[54] Hariharan Subramonyam, Colleen Seifert, Priti Shah, and Eytan Adar. 2020. Texsketch: Active diagramming through pen-and-ink annotations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[55] Craig S Tashman and W Keith Edwards. 2011. Active reading and its discontents: the situations, problems and ideas of readers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2927–2936.

[56] Craig S Tashman and W Keith Edwards. 2011. LiquidText: A flexible, multitouch environment to support active reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3285–3294.

[57] Maartje ter Hoeve, Robert Sim, Elnaz Nouri, Adam Fourney, Maarten de Rijke, and Ryen W White. 2020. Conversations with documents: An exploration of document-centered assistance. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 43–52.

[58] Bret Victor. 2011. Explorable explanations.

[59] Alexandra Vtyurina, Adam Fourney, Meredith Ringel Morris, Leah Findlater, and Ryen W White. 2019. Verse: Bridging screen readers and voice assistants for enhanced eyes-free web search. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*. 414–426.

[60] Zeyu Wang, Yuanchun Shi, Yuntao Wang, Yuchen Yao, Kun Yan, Yuhan Wang, Lei Ji, Xuhai Xu, and Chun Yu. 2024. G-VOILA: Gaze-Facilitated Information Querying in Daily Scenarios. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–33.

[61] 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.

[62] Pierre Wellner. 1991. The DigitalDesk calculator: tangible manipulation on a desk top display. In *Proceedings of the 4th annual ACM symposium on User interface software and technology*. 27–33.

[63] Chih-Sung Andy Wu, Susan J Robinson, and Alexandra Mazalek. 2007. WikiTUI: leaving digital traces in physical books. In *Proceedings of the international conference on Advances in computer entertainment technology*. 264–265.

[64] Ding Xu, Ali Momeni, and Eric Brockmeyer. 2015. Magpad: a near surface augmented reading system for physical paper and smartphone coupling. In *Adjunct Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 103–104.

[65] Xiaoyu Zhang, Jianping Li, Po-Wei Chi, Senthil Chandrasegaran, and Kwan-Liu Ma. 2023. ConceptEVA: Concept-Based Interactive Exploration and Customization of Document Summaries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.

[66] YanXiang Zhang, Li Tao, Yaping Lu, and Ying Li. 2019. Design of Paper Book Oriented Augmented Reality Collaborative Annotation System for Science Education. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*. IEEE, 417–421.

[67] Yuhang Zhao, Yongqiang Qin, Yang Liu, Siqi Liu, Taoshuai Zhang, and Yuanchun Shi. 2014. QOOK: enhancing information revisitation for active reading with a paper book. In *Proceedings of the 8th International Conference on Tangible, Embedded and Embodied Interaction*. 125–132.