MeetSumAid: A Mobile Human-AI Collaborative Meeting Summarization System

Lu Wang, *Senior Member, IEEE*, Yilong Li, Jianhua He, *Senior Member, IEEE*, Yueling Che, Kaishun Wu, *Fellow, IEEE*, Xiaoke Qi, and Kaixin Chen

Abstract—Existing AI-based meeting summarization tools have enabled rapid generation of meeting notes, yet their reliability and user controllability remain limited. This paper explores human-AI collaboration for mobile meeting summarization and presents MeetSumAid, a multifunctional system that integrates summarization algorithms with an interactive user interface. The system is designed to support users in understanding, validating, and refining AI-generated summaries through natural interactions and flexible control mechanisms. By enabling real-time inspection, editing, and feedback, MeetSumAid facilitates reliable collaboration between humans and AI in dynamic meeting scenarios. A user study with 20 participants shows that MeetSumAid significantly improves summary quality, generation efficiency, and user-perceived reliability compared with baseline AI summarizers, while reducing cognitive load. Further analysis reveals how different interface components enhance users' engagement and confidence during collaboration. This work provides a practical step toward reliable and user-centered human-AI collaboration in mobile meeting summarization and offers actionable design implications for future intelligent collaborative systems.

Index Terms—Meeting Summarization, Human-AI Collaboration, Trust Calibration, Large Language Models.

1 Introduction

Meetings remain a cornerstone of collaboration, with approximately 11 million meetings held daily in the United States and employees spending an average of six hours per week attending them [1]. To aid post-meeting reviews and provide non-attendees with quick access to meeting outcomes, note-takers are often tasked with recording and summarizing meeting content. However, crafting accurate and concise meeting summaries is challenging, as it requires identifying key information and maintaining contextual coherence across lengthy and complex discussions [2].

Recent advances in automatic speech recognition and artificial intelligence (AI) have enabled the development of AI-assisted meeting summarization tools that can automatically generate summaries from meeting transcripts [3]. While such tools significantly reduce human workload, AI-generated summaries frequently contain omissions, repetitions, or factual errors, leading to issues in correctness and completeness. In practice, note-takers still need to refine AI-generated content to ensure quality and usability. However, the "black-box" nature of AI systems often prevents users from understanding the reasoning behind generated summaries, making it difficult to assess reliability or identify which parts require revision [4]. This problem limits the effectiveness and adoption of existing AI summarization systems in real-world workflows.

Most existing research has focused on improving summarization models or evaluation metrics, while overlooking human-AI collaboration during the summarization process [5], [6]. In particular, little attention has been paid to how interface design can support users in understanding, verifying, and interacting with AI-generated summaries under mobile and time-fragmented conditions, such as during commutes or short breaks between meetings. In these contexts, users require fast, lightweight, and immediate interactions, yet existing desktop-oriented systems are poorly suited to support such opportunistic summary editing.

To address these challenges, we present MeetSumAid, a mobile human–AI collaborative summarization system that supports efficient refinement and verification of AI-generated meeting summaries. MeetSumAid integrates a large language model (LLM) with a multifunctional interface to enhance reliability and user confidence in AI-assisted summarization. Central to the system is the introduction of micro-themes as an intermediate interaction layer between raw transcripts and final summaries, allowing users to edit summaries at semantically coherent units rather than navigating lengthy transcripts or opaque AI outputs.

- *L. Wang and YL. Che are with the College of Computer Science and Software Engineering, Shenzhen University, China. E-mail: wanglu@szu.edu.cn and yuelingche@szu.edu.cn.*
- *YL. Li are with the Youware, China. Email: lylhaha01@outlook.com*
- *JH. He are with the School of Computer Science and Electronic Engineering, University of Essex, UK. E-mail: j.he@essex.ac.uk.*
- *XK. He are with the China University of Political Science and Law, Beijing, China. E-mail: qixiaoke@cupl.edu.cn.*
- *KS. Wu and KX. Chen are with the Hong Kong University of Science and Technology (Guangzhou), China. E-mail: wuks@hkust-gz.edu.cn and kchen977@connect.hkust-gz.edu.cn.*

The research was supported in part by China NSFC Grant (No.62372307, No.U2001207, No.62472366), the Project of DEGP (No.2023KCXTD042, 2024GCZX003), Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No.2023B1212010007), "111 Center (No.D25008)", Guangdong NSF (No. 2024A1515011691), Shenzhen Science and Technology Program (No. RCYX20231211090129039), Shenzhen Science and Technology Foundation (ZDSYS20190902092853047, No.JCYJ20230808105906014), the Key Research and Development Program in Xinjiang Uygur Autonomous Region (2025B04019-001), EPSRC with RC Grant No EP/Y027787/1, UKRI under grant No EP/Y028317/1, Horizon European program under grant No 101086228, and the Royal Society International Exchanges 2023 under Grant No. IEC\NSFC\233318.

(Co-corresponding author: Xiaoke Qi and Kaixin Chen.)

Building on this design, the system provides key interaction mechanisms—including layered summaries, micro-theme merging and splitting, keyword highlighting, adjustable condensation, and personalized customization—to support transparent inspection, targeted editing, and context-aware guidance of the summarization process.

We conducted a controlled study involving 20 participants and 6 real meeting transcripts to evaluate the effectiveness of MeetSumAid compared with a baseline speech-to-text editor integrated with ChatGPT-based summarization. Experimental results show that MeetSumAid significantly improves summary quality, generation efficiency, and user-perceived reliability, while reducing cognitive load. Analysis of user interactions further highlights how specific design elements enhance engagement and confidence in AI collaboration. The main contributions of this work are summarized as follows:

- We design and implement MeetSumAid, a mobile system for reliable and user-centered human-AI collaborative meeting summarization.
- We propose interaction mechanisms that enhance reliability and controllability in AI-assisted summarization through adaptive, transparent user interfaces.
- We conduct empirical evaluations demonstrating the system's effectiveness in improving summary quality, efficiency, and user experience, and discuss design implications for future human-AI collaborative systems.

2 Related Work

2.1 Human-AI Collaborative Generation

With the rapid advancement of artificial intelligence (AI), recent research has shifted from replacing humans to augmenting human capability through collaboration [7]–[10]. Such collaboration leverages the complementary strengths of humans and AI: AI excels in large-scale data processing, while humans contribute contextual judgment, creativity, and ethical reasoning [11]. However, the black-box nature of deep learning algorithms still limits users' understanding of AI behavior, reducing perceived reliability and hindering effective cooperation [12], [13].

Prior work has investigated human-AI collaboration from both theoretical and practical perspectives. On the theoretical side, studies in human–computer interaction (HCI) have established guidelines for human-AI partnership, such as decision transparency, feedback consistency, and shared control [14], [15]. On the practical side, collaborative mechanisms have been applied to synchronization [8], explainability [16], [17], and co-creation in domains such as art, game design, and healthcare [18]–[23]. Building on this foundation, our work explores meeting summarization as a new application scenario for human-AI collaboration, focusing on how interaction design can enable humans to effectively guide and refine AI-generated summaries in real-world environments.

2.2 Interactive Reliability and Explainable AI

Effective human-AI collaboration relies on users' ability to evaluate and align with AI reliability during interaction [24], [25]. Earlier research introduced the concept of trust calibration to describe this process, emphasizing the balance between overreliance and underreliance [26], [27]. Various approaches have been proposed to support reliability assessment, such as system transparency [28], performance metrics [29], and post-interaction feedback [30]. However, these methods are often static, offering limited adaptability to dynamic decision-making scenarios [6], [31].

Explainable AI (XAI) aims to improve user understanding of AI behavior [16], [32]–[35], but most existing approaches provide offline explanations that are weakly integrated into real-time interaction. Recent work has therefore shifted toward interactive reliability mechanisms that allow users to inspect, question, and adjust AI outputs during use [5], [36], [37]. Our work follows this direction by embedding reliability cues and transparent interactions into a mobile summarization system, enabling continuous evaluation and refinement of AI-generated summaries.

2.3 Meeting Summarization

Meetings are central to professional collaboration, yet their spontaneous and unstructured dialogue makes summarization particularly challenging [3], [38]. Manual summarization requires identifying salient points, understanding context, and balancing brevity and completeness [39].

Automated summarization methods have emerged to address these challenges. Early extractive approaches identified key text segments but produced less coherent summaries [40], [41]. In contrast, abstractive summarization techniques, powered by Transformer-based models such as BERT, BART, and T5, produce more fluent and human-like summaries [42]–[44]. Nonetheless, these methods depend heavily on large annotated datasets and still struggle with factual inaccuracies and hallucinations [45], [46].

Commercial platforms such as Zoom and Tencent Meeting provide speech transcription and one-shot AI summaries, but offer little support for post-generation interaction or correction, often leading to overtrust or loss of context [47]. In contrast, our work enables interactive human–AI collaboration, allowing users to iteratively inspect and refine AI-generated summaries on mobile devices.

2.4 Large Language Models

Large language models (LLMs) such as GPT-4 have demonstrated remarkable generalization capabilities across diverse natural language tasks. Trained on extensive text corpora, LLMs exhibit strong contextual understanding and generation fluency, supporting applications in dialogue systems [48], translation [49], emotion analysis [50], and writing assistance [51].

Recent studies have applied LLMs to summarization tasks, including radiology reports [52], news articles [53], and meetings [54]. Despite these advances, LLM-generated summaries still suffer from factual inconsistency, redundancy, and lack of contextual alignment [45], [46]. Our work leverages the language understanding capabilities of LLMs while incorporating interactive refinement mechanisms to mitigate these limitations, thereby improving both the accuracy and practical usability of AI-generated meeting summaries.

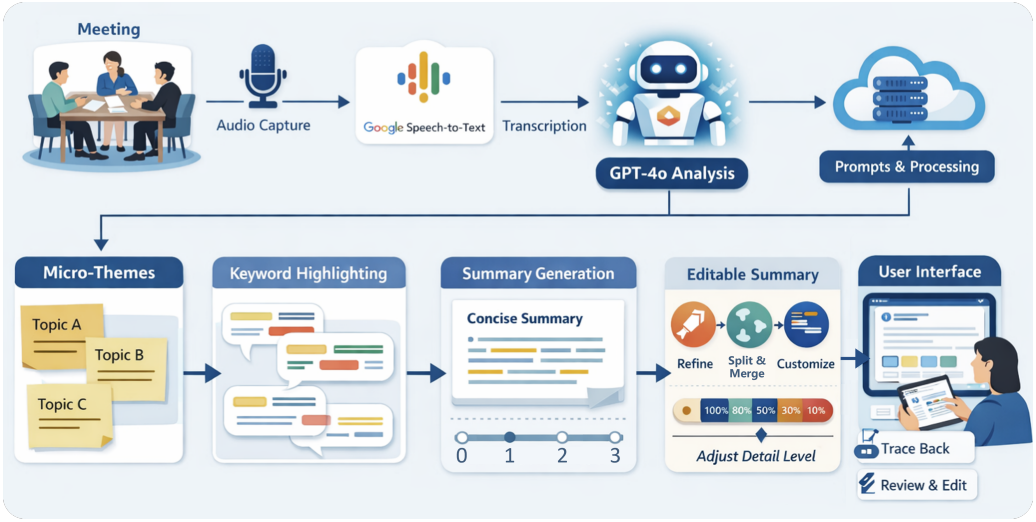


Fig. 1: Overview of the MeetSumAid workflow. Meeting audio is transcribed using AI-based speech recognition and processed by GPT-4o through prompt engineering. The system organizes content into micro-themes, highlights key phrases via multi-modal cues, generates summaries based on keyword importance, and allows users to edit micro-themes and control summary granularity for progressive, human-like summarization.

3 MeetSumAid Design

3.1 Design Principles and Objectives

As mentioned in the introduction, our goal is to help minute-takers efficiently revise summaries through trust calibration cues and streamlined interaction design, assisting users in overcoming challenges in summary writing.

Conventional meeting summary tools typically transcribe content linearly, preserving all original speech imperfections including disfluencies, repetitions, and incoherence. While advanced natural language processing techniques can help address some of these issues, users still require guidance in organizing meaning and structure. Successful LLM interfaces (e.g., GPT-4o, Claude 3.5 sonnet) have demonstrated capabilities in refining disorganized text into high-quality writing. However, such general-purpose tools not only rely on users' ability to prompt LLMs effectively – a recognized challenge [55], [56] – but also make it difficult for users to precisely control where and how text edits occur.

Building on prior work demonstrating the benefits of integrating NLP or LLMs in direct manipulation interfaces [57], [58], we investigate how to leverage state-of-the-art LLM and NLP technologies to embed intelligent word processing within GUIs for our specific summarization task, while maintaining user control over summary generation. Considering these factors, we aim to enhance text editor interfaces through the following design goals for better interaction with speech-derived text:

- Present users with the background and context of the meeting, along with comprehensive guidance on crafting the meeting summary. This is instrumental in helping novice users grasp the essential elements and objectives of meeting summary writing, facilitating human-machine collaborative summarization generation.
- Support non-linear content organization. Users should be able to iteratively organize and edit spoken text, flexibly adjusting content sequence, structure, and granularity. This design aligns with the non-linear nature of human cognition and creativity, facilitating efficient processing of complex information and progressive refinement of expression. The interface should provide sufficient freedom for users to reorganize content as needed, enhancing editing fluency and intuitiveness.
- Facilitate micro-theme review and navigation. To bridge human-AI collaboration, we envision simulating the step-by-step process of human-generated summaries, from transcribing text to micro-theme, and finally to the overall summary. The interface should assist users in efficiently reviewing and navigating micro-theme content, significantly reducing cognitive load in text comprehension. For potentially verbose, repetitive, or error-prone text, the interface should employ intelligent summarization to distill key information while enabling quick access to micro-theme-specific passages. This design helps users rapidly grasp main ideas while maintaining traceability to original context, ensuring both reading efficiency and content mastery.
- Leverage LLMs while retaining user control. When utilizing large language models (LLMs) for text cleaning, polishing, and transformation, the interface must preserve user agency. The design should provide explicit options for users to determine where and how automated modifications are applied. This approach balances processing efficiency with user sovereignty over final outcomes.
- Optimize mobile interaction experience. We selected tablet interfaces as our current design focus due to their intermediate position between smartphones and desktop devices—offering reasonable screen real estate while maintaining portability. This hybrid nature makes tablets an ideal platform for studying mobile interaction optimization. By refining tablet interfaces, we aim to extend relevant design principles and technologies to other mobile devices, ultimately enhancing cross-platform user experiences.

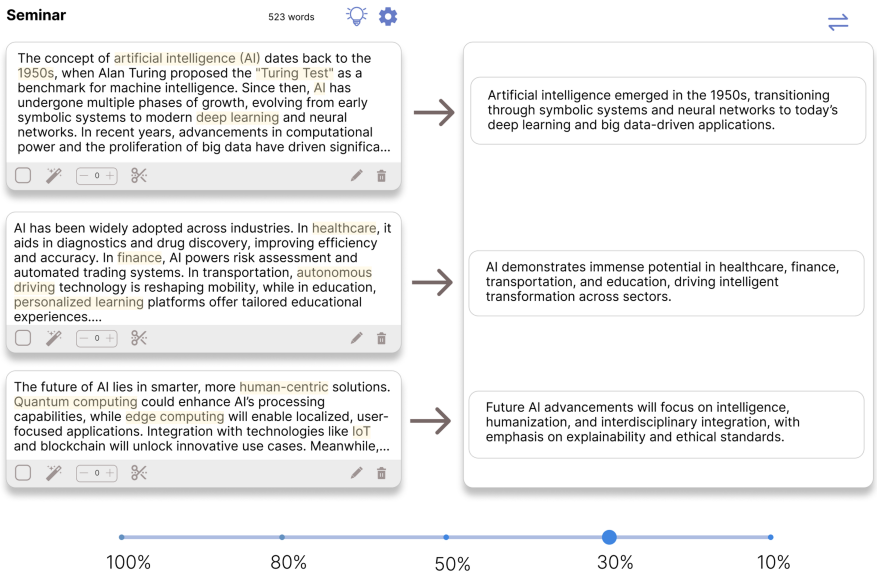


Fig. 2: The interface adopts a dual-panel design: the left panel visualizes segmented micro-themes extracted from meeting content as discrete units, enabling focused topic management, while the right panel dynamically displays contextually aligned summaries for each selected micro-theme. A granularity control slider at the bottom allows real-time adjustment of the level of detail of the summaries, seamlessly toggling between concise overviews and detailed elaborations. This structure supports user-defined abstraction levels and visual-thematic consistency, balancing analytical efficiency with narrative coherence through interactive content exploration.

3.2 MeetSumAid Interface Design

Figure 1 presents the end-to-end workflow of MeetSumAid, illustrating how meeting audio is transformed into interactive summaries. The system captures audio and converts it into text using AI-based speech transcription, then processes the transcript with GPT-4o guided by prompt engineering. The workflow highlights four key capabilities: 1) micro-themes as management units, supporting stepwise, human-like summary composition; 2) multi-modal key phrase highlighting using text and speech cues; 3) keyword- and micro-theme-importance-driven summary generation; 4) editable micro-themes with controllable summary granularity.

Building on this workflow, the interface shown in Figure 2 allows users to review, organize, and summarize long meeting transcripts through a micro-theme-centric layout. Each micro-theme encapsulates a coherent discussion segment, making lengthy transcripts more manageable. Leveraging this structure, the interface facilitates writing prompt generation, micro-theme refinement and reorganization, salient keyword highlighting, summary condensation control, and traceability between summaries and their source micro-themes. The following subsections provide detailed descriptions of these interaction mechanisms.

3.2.1 Providing Writing Prompts for Meeting Summaries

Before users write a meeting summary, the system automatically generates writing prompts (such as meeting context, key topics, discussion points, etc.) based on the meeting transcript, as shown in Figure 3 A, which can effectively improve the efficiency and quality of summary writing. Firstly, these prompts provide users with clear contextual information, helping them quickly grasp the core content and key discussion points of the meeting, thereby avoiding incomplete summaries due to missing important information. Secondly, the writing prompts guide users to focus on the key outcomes and decisions of the meeting, ensuring that the summary remains relevant and avoids redundancy or digression. Additionally, this pre-prompt mechanism reduces users' cognitive load, especially when dealing with complex or lengthy meeting transcripts, as users no longer need to repeatedly read and filter information, thus saving time and improving work efficiency. Finally, this design also helps standardize the style and structure of summaries, ensuring consistency and professionalism across summaries generated by different users, further enhancing the usability and reference value of meeting records. Therefore, providing writing prompts not only optimizes the user experience but also offers strong support for the accuracy and completeness of meeting summaries [59].

3.2.2 Facilitation of Micro-theme Text Review

Initially, MeetSumAid employs LLMs to partition the full meeting text into smaller thematic segments, referred to as micro-themes, each encapsulated within a container. The "fix writing" feature, depicted in Figure 4, is utilized to refine the text of micro-themes, eliminating redundant content and transforming the text into a more readable, written form as opposed to disorganized spoken language. If the user is not satisfied with the modified result, they can also edit it manually.

3.2.3 Keyword Extraction

Subsequently, key phrases are highlighted within the containerized text based on both textual and acoustic cues. Textually, we adopt MDERank [60] to identify salient phrases. Acoustically, prominent intonational rises often signal important information beyond textual content. Intonation is

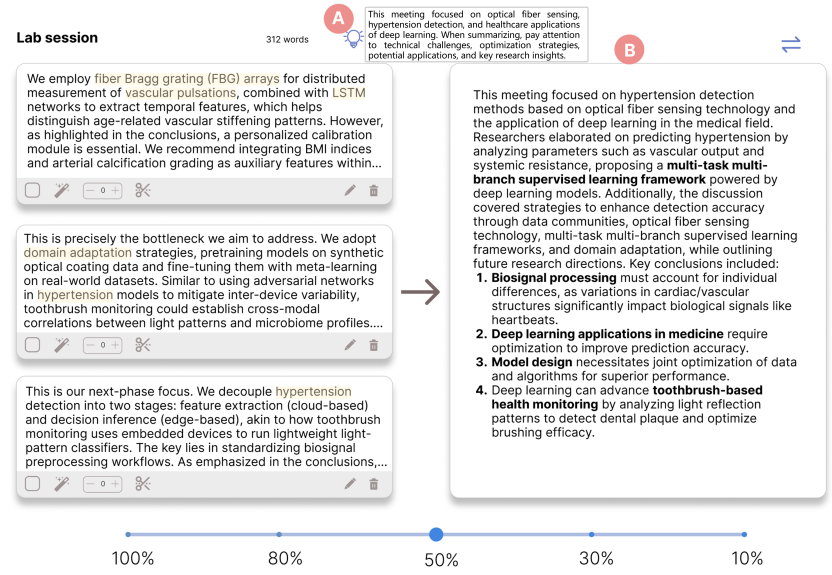


Fig. 3: The global summary is displayed on the right when clicking switch button (B), with corresponding micro-themes on the left (A), allowing direct traceability from global summary chunks back to their associated micro-themes.

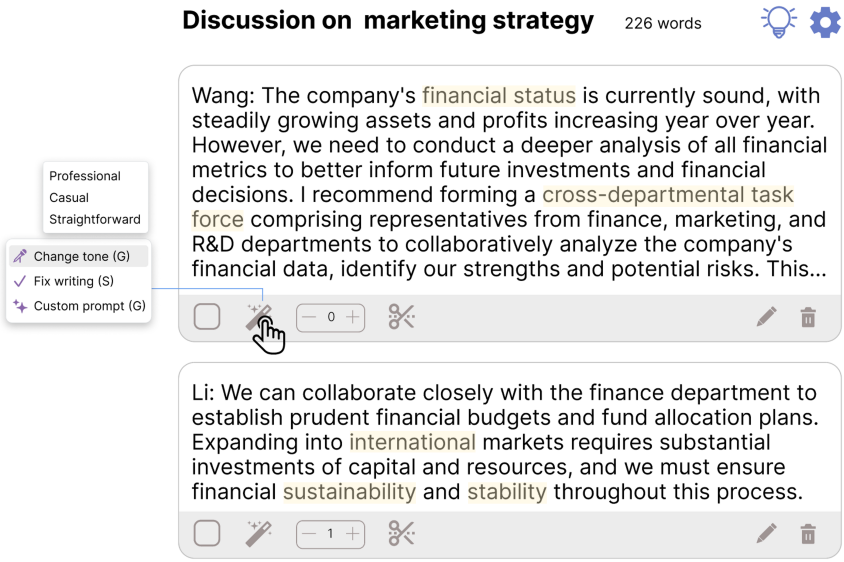


Fig. 4: The Magic Wand feature offers three key functionalities: Change Tone, Fix Writing, and Custom Prompt. G indicates that the operation is applied to the generated summary, while S denotes that the operation acts on the micro-theme.

characterized by the fundamental frequency (f_0), which reflects emphasis, sentence modality, and speaker intent [61]. We extract f_0 using pYIN [62], a probabilistic extension of YIN with HMM-based smoothing that reduces octave errors and pitch discontinuities. To accommodate both male and female speakers and emphasis-induced pitch excursions, we set the detection range to 70–400 Hz, with a 2048-sample frame length (128 ms at 16 kHz) and a 160-sample hop (10 ms). Figure 5 shows an example of the extracted results. We then apply K-means clustering [63] to identify segments with high intonation for each speaker. These acoustically salient segments are fused with text-based key phrases using equal weights (0.5 each). Candidate phrases are ranked by the fused salience score, and the LLM is prompted to select the top phrases while maintaining a highlight density of 1%–3%, which are then used as prioritized cues for summary generation. Users may enable or disable highlighting, which directly influences the generated summary by emphasizing the selected phrases.

3.2.4 *Scalable Summary Condensation*

Finally, a slider is provided to control the level of summary condensation, offering options of 100%, 80%, 50%, 30%, and 10%. Specifically, the original length corresponds to 100%, while the most condensed summary corresponds to 10%. This feature is designed to aid users in swiftly capturing core information and understanding how to reorganize and iterate summaries efficiently.

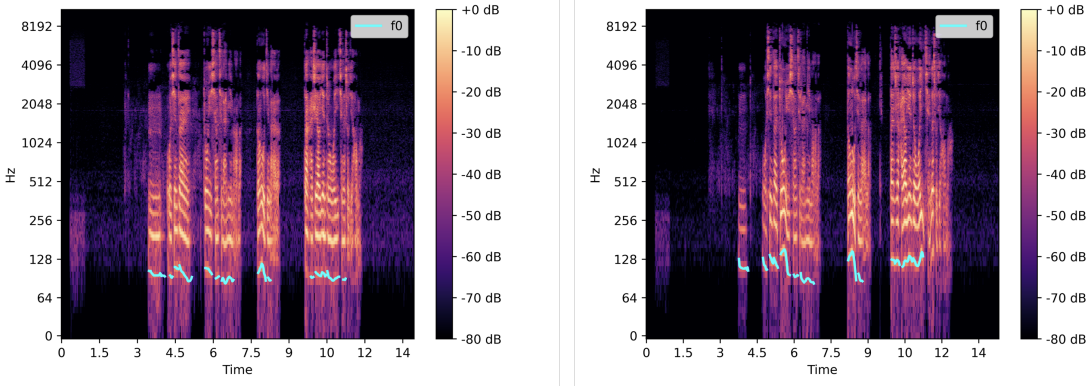


Fig. 5: Comparison of fundamental frequency (f_0) under regular and emphatic speech. Left: regular tone. Right: emphatic tone, showing higher overall f_0 and greater variability.

Fig. 6: The merge function allows users to combine multiple micro-themes, automatically integrating their texts and regenerating a coherent summary aligned with the unified theme.

3.2.5 *Editing at Micro-theme Level*

Micro-themes are categorized by LLMs with relatively fine granularity, creating the possibility for multiple micro-themes to be merged into a larger theme. Our system supports two merging methods: 1) Dragging one micro-theme container onto another to merge them, and 2) Selecting multiple micro-themes and clicking the "Merge" button, as illustrated in Figure 6. When merging multiple micro-themes, the system employs LLMs to adjust the text across these micro-themes, ensuring natural coherence to form a larger unified theme. The summary is also rewritten to align with the consolidated theme. Conversely, the system also supports splitting larger micro-themes by clicking the scissors button located below the micro-theme container. Additionally, each micro-theme includes a control to adjust its importance level (ranging from 0 to 3) in the generation of the global summary. A higher value indicates greater prominence in the global summary, while a lower value reduces its visibility. This feature enables dynamic content prioritization and user-driven thematic emphasis, enhancing the adaptability and relevance of the overall summary.

As illustrated in Figure 4, users can customize micro-theme summaries by switching between predefined *Objective* and *Analytical* styles via *Change tone*, or by using *Custom prompt* for greater flexibility.

3.2.6 *Summary View Switching*

While the micro-theme-level summaries mentioned earlier offer granular details to help users efficiently review meeting specifics, a global summary is essential for providing a quick overview of the entire meeting. To reduce user confusion and enhance summary credibility, we implement a traceability feature from the global summary to micro-themes (clicking a chunk in the global summary highlights the related micro-theme), primarily powered by LLMs.

3.2.7 *Trace Back the Original Micro-theme*

Tracing text chunks from meeting summaries back to their source micro-themes holds multifaceted practical value, particularly critical for information verification, in-depth comprehension, and enhancing collaborative efficiency. At its core, it builds a bridge between informational simplification and completeness. This capability addresses the pain point of summaries that convey "what" without explaining

"why," striking a balance between efficiency and rigor. It is especially vital in scenarios demanding high transparency and traceability, such as healthcare, finance, and legal fields. For teams reliant on meeting-driven decision-making, this approach significantly reduces communication costs and mitigates decision risks.

4 Prototype Implementation

MeetSumAid is implemented as a React.js web application with a golang backend, hosted on a cloud platform for scalability. The system integrates multiple NLP and speech analysis modules to support its core micro-theme editing and summarization capabilities. Below, we detail the technical implementation of key features outlined in our design.

4.1 Core LLM Infrastructure

All text transformation and summarization tasks are powered by GPT-4o, selected for its superior coherence and contextual awareness compared to smaller models (e.g., Claude Haiku). Key LLM-driven operations include:

Micro-Theme Segmentation. Raw meeting transcripts are partitioned into coherent micro-themes using a custom prompt that enforces thematic consistency and avoids abrupt topic shifts. *Text Smoothing.* Disfluent speech (e.g., filler words, repetitions) is refined into written-form prose via the "Fix Writing" feature, implemented as a one-shot GPT-4o prompt. *Micro-Theme Merging and Splitting.* When users merge or split containers, GPT-4o rewrites concatenated text to ensure narrative flow, guided by a prompt emphasizing logical transitions. *Summary Regeneration.* Global and micro-theme summaries are dynamically regenerated post-editing using a zero-shot summarization prompt, with keyword weighting biases injected for prominence.

Pre-defined prompts supporting micro-theme–level operations were refined through iterative testing. All prompts are provided in Appendix A for reference. Users may override the default prompts via the Custom Prompt interface.

4.2 Interactive Controls

The system implements three core interaction mechanisms that combine algorithmic processing with intuitive user controls to facilitate dynamic knowledge organization. Through a modular architecture, these components enable granular content manipulation while maintaining semantic traceability across abstraction levels.

Summarization Granularity Slider. Condensation levels (100% to 10%) are achieved by progressively truncating GPT-4o summaries via token length constraints, with iterative compression for higher reduction tiers.

Merge and Split Operations. Drag-and-drop, batch merging and splitting features are implemented using React DnD for container interactions. Merged micro-themes trigger a background GPT-4o API call to rewrite content, with results cached for performance.

Traceability. Global summary chunks are linked to micro-themes through semantic embeddings (voyage-3). Click interactions compute cosine similarity between the clicked chunk's embedding and all micro-themes, highlighting matches above a $\theta = 0.82$ threshold.

5 Evaluation

We conducted work sessions with participants to validate our proposed system, studying the following four research questions:

RQ1: Can our system help users improve the quality of their meeting summary writing?

RQ2: Can MeetSumAid save users time and effort?

RQ3: Compared to the baseline, what are the benefits of using MeetSumAid, particularly in terms of interaction within human-AI collaboration?

RQ4: How do the various design components of our system help them perceive the capabilities of AI and calibrate trust in AI, and what are the user preferences?

We designed two test conditions to support the validation of the system on the above questions:

C1 (Baseline interface): Participants completed summary writing with ChatGPT assistant.

C2 (Our system): Participants completed summary writing with assistance from MeetSumAid (all functions).

This study was conducted in accordance with the research ethics guidelines of our institution and received approval from the institutional ethics review committee. All participation was voluntary, informed consent was obtained from all participants prior to the study, and no personally identifiable information was collected or recorded.

5.1 Apparatus and Baseline Interface

The system was deployed on a 12.9-inch iPad Pro running iOS 17, provided by the researchers, and participants can access the interface via Safari. The baseline interface featured a transcription text box on the left and a meeting summary box on the right, accompanied by a ChatGPT window, allowing participants to generate meeting summaries using ChatGPT. Both the transcription text box and the meeting summary box were editable. For ease of study, we referred to the manual writing as 'System 1' and MeetSumAid as 'System 2', corresponding to the two experimental conditions C1 and C2.

5.2 Dataset

5.2.1 Meeting Records Collection

We collected 14 meeting recordings via Tencent Meeting through an online call for submissions, each containing audio and screen captures. Contributors were informed that the data would be used solely for academic research. From these, we selected 6 high-quality meetings with complete content and no technical issues. The dataset includes corporate team meetings, academic seminars, wedding planning meetings, and lab group discussions. Meetings were small-scale, typically involving 3–8 participants and 2–6 active speakers, with an average duration of 25 minutes (range: 8–35 minutes). Background noise levels were generally low to moderate, estimated at approximately 35–50 dB. Audio was transcribed using Google's speech-to-text service, yielding transcripts of 1,536–7,335 words. Applying our LLM-based micro-theme segmentation pipeline resulted in an average of 18.7 micro-themes per meeting (range: 9–31), with each micro-theme containing approximately 120–350 words.

5.2.2 Standard Summary Writing

We recruited three individuals with extensive experience (over 5 years) in summary writing. They were instructed to watch and listen to the meeting replay as passive observers. One of them was then assigned to draft the summary, while the other two were responsible for reviewing it, focusing on accuracy, completeness, and objectivity. A summary was only considered a standard summary if both reviewers found it satisfactory. To ensure the highest quality of summaries, the roles of writer and reviewer were not fixed but rotated among the participants.

5.3 Participants

To conduct our experiment, we recruited a total of 20 participants (8 female, 12 male). Their ages ranged from 22 to 42 years old (M = 29.95, SD = 6.08). Participants were required to have prior experience attending at least one work or academic meeting to ensure familiarity with meeting contexts. We reached out to them using a combination of emails and word-of-mouth, and from the respondents, we selected participants to cover a broad range of meeting summarization experience levels. As a result, the final sample included 13 participants with substantial experience writing meeting summaries in their companies, 4 participants with limited experience, and 3 participants with no prior experience in meeting summarization. The frequency of writing summaries and participants' LLM usage habits varied, as shown in Table 1.

5.4 Experiment Task

Participants were instructed to observe designated meeting replays as passive attendees and to use the specified system to compose summaries. To ensure that the evaluation was not adversely affected by user disengagement, participants were required to attentively watch and listen to the assigned meeting replays and to strive to produce accurate, comprehensive, and objective summaries. To balance the length and comprehensiveness of the summaries, participants were informed that the maximum word limit for each summary was set at 25% of the transcribed text. The meetings that participants were assigned to review are shown in Table 1. It is noteworthy that the order in which each participant used the baseline and the proposed system was randomized and evenly distributed to eliminate any potential influence of the system usage sequence on the experiment. Participants were not restricted by time constraints, allowing for more thorough summary composition.

5.5 Procedure

Firstly, we collected demographic information from the participants, including their educational background and experience in meeting summarization, and requested them to sign an informed consent form. We demonstrated to the participants the tutorial for using our proposed system as well as the baseline application. Each participant was given up to 10 minutes to familiarize themselves with the functionalities of both systems. Subsequently, each participant was asked to complete summarization tasks for two designated meetings using both systems on the iPad Pro provided by us. We recorded the participants' interactions through screen recording and noted the time they took to produce the overall summaries using the different systems. After obtaining the overall summaries from both systems, the participants took part in a semi-structured follow-up interview. The participants first responded to the Likert scale questions presented in Table 2, and then they rated the helpfulness of eight design components in calibrating trust (on a scale from 1 to 7, indicating from not helpful at all to extremely helpful). They were required to explain the reasoning behind their ratings. The average duration of the study for each participant was approximately 1 hour.

5.6 Measurements and Analysis

We collected three kinds of information from users in total. First, the behavioral data during user interactions with System 1 and System 2 must be recorded. Second, after participants completed a summary writing task, we needed to save their summary documents. Third, we gathered feedback from users regarding the post-study survey questionnaire.

Regarding RQ1, we obtained summaries written by users using System 1 and System 2. Then, we calculated the ROUGE-L [64] and BLEU [65] scores with reference to standard summaries. ROUGE-L and BLEU were two commonly used evaluation metrics in natural language processing. The higher these two metrics are, the closer the user-written summaries are to standard summaries.

Regarding RQ2, in each session, we first calculated the average time spent by participants on completing summary writing. Next, we computed the average ROUGE-L and BLEU scores for each participant's summaries with respect to standard summaries. Given that the duration of participants' engagement positively influenced the quality of their summaries, we implemented an efficiency assessment by calculating the average ROUGE-L score per unit of time. To evaluate the mental effort required when using two systems, we asked participants to answer two seven-scaled Likert NASA TLX questions (i.e., mental demand and difficulty dimensions, Q1, Q2 in Table 2).

Regarding RQ3, after analyzing the interaction logs, we compiled a summary of participants' interaction frequencies with both systems. Additionally, we gathered information regarding participants' ratings on the systems' capabilities for meeting summary writing (Q3 in Table 2), their level of confidence in the meeting summaries they wrote (Q4 in Table 2), their attitudes towards future system usage (Q5 in Table 2), and their overall preference between system 1 and system 2 (Q6 in Table 2).

To answer RQ4, we gathered participants' ratings on the trust calibration contributions of each component and transcribed their explanations for thematic analysis. The thematic analysis [66] was performed independently by two researchers, with shared themes determined through consensus. We performed statistical analysis on the collected data, calculating means, medians, and standard deviations, and conducted analyses using Specific significance test.

TABLE 1: Demographic information and combined record-condition assignment for 20 participants.

ID	Gender	Age	Education	Years of Writing Experience	Frequency of Writing Summaries	LLM Usage	Record-Condition Assignment
P1	M	24	Technical school	1	Daily	Daily	(R2,C1),(R4,C2)
P2	M	25	High school	0.5	Monthly	Weekly	(R2,C1),(R4,C2)
P3	F	28	University	1.5	Weekly	Weekly	(R2,C2),(R4,C1)
P4	F	23	University	1.5	Weekly	Daily	(R2,C1),(R4,C2)
P5	M	23	University	0.5	Bi-Weekly	Weekly	(R2,C2),(R4,C1)
P6	M	41	University	4	Weekly	Weekly	(R2,C2),(R4,C1)
P7	M	38	Technical school	0	Never	Daily	(R1,C1),(R6,C2)
P8	F	35	University	5	Bi-Weekly	Daily	(R1,C2),(R6,C1)
P9	M	42	University	5	Weekly	Bi-Weekly	(R1,C1),(R6,C2)
P10	M	30	University	3	Weekly	Daily	(R1,C2),(R6,C1)
P11	F	29	University	3.5	Daily	Daily	(R1,C1),(R6,C2)
P12	M	32	University	1	Monthly	Weekly	(R1,C2),(R6,C1)
P13	M	24	University	0	Never	Daily	(R2,C1),(R5,C2)
P14	M	26	High school	3	Daily	Weekly	(R2,C2),(R5,C1)
P15	F	26	University	0.5	Bi-Weekly	Weekly	(R2,C1),(R5,C2)
P16	F	33	Technical school	0	Never	Never	(R2,C1),(R5,C2)
P17	M	36	University	0.5	Weekly	Never	(R3,C2),(R6,C1)
P18	F	22	University	1	Weekly	Daily	(R3,C2),(R6,C1)
P19	M	30	University	3.5	Daily	Daily	(R3,C1),(R6,C2)
P20	F	32	University	4.5	Weekly	Never	(R3,C2),(R6,C1)

6 RESULTS

6.1 RQ1: Can MeetSumAid help users improve the quality of their meeting summary writing?

We collected meeting summaries written by participants using System 1 and System 2. We calculated the ROUGE-L score and BLEU score of summaries written using two systems respectively, with reference to standard meeting summaries. The median ROUGE-L scores for C1 and C2 were 54.38 and 62.57 (average μ=54.37, 64.06, standard deviation σ=11.12, 5.72). The median BLEU scores for C1 and C2 were 48.65 and 55.17 (average μ=48.17, 54.76, standard deviation σ=5.20, 5.17). We conducted Kruskal-wallis test [67] for ROUGE-L score and BLEU score, which showed that ROUGE-L score and BLEU score were significantly different in the two conditions (ROUGE-L: p=0.005, BLEU: p=0.0016). A post-hoc Dunn test with Bonferroni correction (p=0.0049) indicated a significant improvement in both ROUGE-L and BLEU under condition C2 compared to C1.

To gain a deeper understanding of the reasons behind the improvement in summary quality under the C2 condition, we interviewed 12 participants whose ROUGE and BLEU scores were significantly higher in the C2 condition compared to the C1 condition. The results showed that all features of MeetSumAid were considered crucial in enhancing the quality of meeting summaries through human-AI collaboration.

For example, 9 participants mentioned that the navigation feature helped them quickly locate and correct errors in the transcribed text, thereby ensuring the accuracy of the summaries. One participant stated: *"When using the simpler interface (baseline system), I wasn't sure if the AI-generated summary matched the original text. I had to extract key points from the original text myself, which often led to missing important information. With the more feature-rich interface of MeetSumAid, the navigation feature allowed me to quickly find key sections of the original text and correct transcription errors. For instance, I once noticed that the AI had mistakenly transcribed 'marketing strategy' as 'marketing test.' By indexing the original text, I identified the error, corrected it, and regenerated the summary, significantly improving its accuracy."* (P7)

Additionally, 8 participants found the compression level slider helpful for proofreading summary quality at different levels of conciseness, enabling them to quickly check core points or verify whether the AI had correctly understood the details of the discussion. One participant shared: *"When summarizing micro-theme, this feature allowed me to first review*

TABLE 2: Users' scoring results for six questions using System 1 (Baseline) and System 2 (MeetSumAid).

ID	Questions	Baseline			MeetSumAid			p
		μ	σ	$Median$	μ	σ	$Median$	
Q1	How much mental effort did you spend on completing the tasks of writing summaries?	4.40	1.47	4	3.20	1.11	3	0.003
Q2	How might you characterize the difficulties you encountered throughout the tasks?	4.15	0.99	4	2.80	1.15	3	0.00045
Q3	How powerful is the system in helping you complete your tasks?	2.20	0.83	2	5.05	1.36	5	0.0001
Q4	How confident are you that your summaries are of high quality?	3.05	1.57	3	6.15	0.81	6	<0.0001
Q5	Would you like to use this system in the future?	1.75	0.79	2	6.70	0.57	7	<0.0001
Q6	System Preference (System 2 over System 1)	μ=6.2, σ=0.77, $Median$=6						/

a 50% compressed version of the summary. I could compare it with the original text, assess its quality, and then decide whether further adjustments were needed." (P15)

Furthermore, 6 participants praised the text-microtheme-overall summary structure, noting that this design prevented the omission of important information and improved summary quality. One participant commented: *"With the baseline system, I struggled to grasp the logical connections between scattered pieces of information. However, MeetSumAid's approach of summarizing text into micro-theme and then integrating them into a cohesive overall summary made the final output more coherent and complete, while also being more efficient."* (P5)

In summary, MeetSumAid significantly enhanced the efficiency and quality of human-AI collaboration by supporting features such as original text indexing, multi-level proofreading tools (e.g., the compression level slider), and a structured summary generation process (text-microtheme-overall summary). These features enabled users to quickly proofread transcribed text, better understand AI-generated content, and optimize summaries at different levels, ultimately producing higher-quality meeting summaries.

6.2 RQ2: Can MeetSumAid save users time and effort?

We calculated the time it took participants from the end of the meeting audio playback until the participant completed summaries. The average completion time for participants using C1 was 13.74 minutes (σ=4.01) and the average completion time for participants using C2 was 8.72 minutes (σ=2.38). We obtained the following conclusion from the Wilcoxon Rank Sum Test: there was a significant difference in the time it took participants to complete summaries in both the C1 and C2 conditions (p<0.001). Although participants were informed in advance of the maximum word limit per summary (calculated from the transcribed text of the meeting, which was no more than 25% of the transcribed text), there was a significant difference in the number of words written by participants. In order to better balance the completion time and the number of words of valid content in summaries, we further calculated the efficiency of each participant by a simple formula: $\frac{ROUGE-L}{time} \times 0.6 + \frac{BLEU}{time} \times 0.4$. We gave ROUGE-L a high weight because we believed that a good summary should be as short as possible while maintaining comprehensiveness of information, although this might sacrifice a certain amount of precision. We did a Wilcoxon Rank Sum Test for efficiency. The results (Figure 7) showed that participants using C2 completed summary writing significantly more efficiently than those using C1 (C1: μ=5.01, C2: μ=7.35, p<0.001, Figure 7 Q1).

In the post-study questionnaire, participants reported significant reductions in mental strain using C2 compared to using C1 (C1: μ=4.4, C2: μ=3.2, p=0.003, Table 2 and Figure 7 Q1). Additionally, participants reported that they hardly got stuck in difficulties when using MeetSumAid (C1: μ=4.15, C2: μ=2.8, p<0.001, Table 2 and Figure 7 Q2) and experienced minimal frustration.

They mentioned that this improvement was attributed, on one hand, to the system's smooth interaction design (e.g., navigation between text boxes, seamless transitions from micro-theme to the overall summary), and on the other hand, to the support of LLM functionalities (e.g., automatic text optimization and text merging).

For example, P16 commented: *"Using MeetSumAid significantly reduced the burden of writing summaries for me. It's like a little assistant that perfectly aligns with my work habits. While using ChatGPT alone can quickly generate a draft, it's hard to achieve a professional level because meetings involve too many elements and details, requiring me to repeatedly check. MeetSumAid helped me solve these problems."* (P16)

Overall, participants under the C2 condition performed better in terms of time and efficiency. MeetSumAid significantly reduced participants' psychological burden and alleviated the frustration they experienced while thinking about how to write high-quality meeting summaries.

6.3 RQ3: Compared to the baseline, what are the benefits of using MeetSumAid, particularly in terms of interaction within human-computer collaboration?

We first compared participants' interaction behaviors when manually writing summaries (C1) versus using MeetSumAid (C2). We observed that under the C1 condition, participants spent a significant amount of time repeatedly

browsing the meeting transcripts and listening to audio recordings, as if they were constantly verifying information. These behaviors were the primary reasons for the longer time and lower efficiency in summary writing. In contrast, participants under the C2 condition significantly reduced their browsing of transcripts and audio recordings. The average browsing time for C1 is 8.9 minutes, and for C2 is 3.7 minutes. A Wilcoxon test revealed that the number of browsing instances was significantly lower in C2 compared to C1 ($p = 0.017$). This behavioral shift indicates that MeetSumAid's micro-theme structuring and traceability effectively reduced users' information foraging cost, allowing them to focus more on higher-level refinement tasks and lowering cognitive load, consistent with our design goals.

Additionally, we found that participants who frequently used the LLM tended to experiment with different prompts to generate summaries multiple times using ChatGPT. On the other hand, several participants who rarely or never used the LLM spent considerable time revising the generated summaries and sometimes paused for extended periods to think. From their interviews, it was evident that the "black-box" nature of AI and its occasional errors undermined their confidence. For example, P17 stated: *"When I used ChatGPT (C1), I couldn't guarantee the accuracy of the summaries it generated, and indeed, it produced illogical sentences. To get a correct summary, I had to repeatedly browse the transcripts and even verify the audio, relying mainly on manual writing to ensure my logic was sound."* (P17)

Another interesting observation was that participants in the C2 condition did not immediately modify the AI-generated meeting summaries. Instead, they adjusted key elements such as keyword annotations and text merging multiple times, referred to AI-generated results at different compression levels, and only manually revised the summary content when necessary. As P11 noted: *"When writing summaries, I prefer interacting with the system multiple times to achieve a satisfactory result, rather than directly editing the AI-generated output. This makes the task of writing a high-quality summary feel much more manageable."* (P11)

Based on post-study participation in responses to the questionnaire, participants perceived MeetSumAid to be capable of assisting them in the task of writing meeting summaries (C1: μ=2.2, C2: μ=5.05, p=0.0001, Table 2 and Figure 7 Q3). In particular, participants reported that the convenient navigation and summary compression control features in MeetSumAid helped them better validate AI-generated results (mentioned 13 times). Additionally, referencing AI-generated summaries and keywords (mentioned 8 times) increased their confidence in producing high-quality meeting summaries (C1: μ=3.05, C2: μ=6.15, $p<0.0001$, Table 2 and Figure 7 Q4). Regarding their willingness to use MeetSumAid in the future, participants also responded positively (C1: μ=1.75, C2: μ=6.7,$p<0.001$, Q5 in Table 2 and Figure 7). Finally, participants scored the extent to which System 2 was superior to System 1; it was evident that participants favor System 2: 8/20 participants gave a score of 7 (strongly favouring System 2 over System 1), 8/20 participants scored 6 (moderately favouring System 2 over System 1), and 4/20 participants scored 5 (slightly favouring System 2 over System 1).

MeetSumAid makes the summary writing process more natural through interactive features like navigation and compression controls, reducing the need for repetitive browsing and verification. At the same time, its step-by-step generation approach alleviates concerns about the "black box" nature of AI, significantly enhancing human-AI collaboration.

6.4 RQ4: How do the various design components of our system help them perceive the capabilities of AI and calibrate trust in AI, and what are the user preferences?

Overall Feedback. As shown in Figure 8, participants rated the four functional components highly in terms of trust calibration (above 5 points). Among them, **Cross-textbox Navigation** and **Summary Condensation Control Slider** received the highest ratings (mean of 6.00), followed by **Layered Summary Structure** and **Smart Optimization and Manual Editing** (mean of 5.50 each). These features were highly praised for their advantages in helping evaluate the trustworthiness of AI-generated summaries and providing user control. In contrast, **Summary Writing Guidance** and **Personalized Instruction Customization** received lower ratings (mean of 3.67 and 4.00, respectively), and it was found that these features were only helpful to some users.

Abstract Writing Guidance. Although the evaluation of this feature is mixed, it is interesting to note that we found it performs better in assisting novice users, while experienced users perceive it as less necessary. P13 mentioned, *"The abstract writing guidance is very useful to me, especially when I'm unsure how to start; it guides me like a mentor."* (P13) However, P20 stated, *"I already have a clear process and focus for writing abstracts, so I don't rely much on this feature."* (P20) This suggests that the usefulness of the feature varies depending on the user's experience level.

Layered Summary Structure. Participants generally found this feature very useful for organizing complex information. P7 commented, *"The layered structure aligns with the summary writing process, and the interface has already broken down large amounts of information into manageable micro-theme sections, making the logic clearer."* P9 added, *"With the layered structure, I can better ensure consistency across different levels of information, improving the quality of decision-making."* However, P15 suggested further optimizing the interface design by placing the text, micro-theme, and overall summary on the same page to make it more intuitive.

Cross-textbox Navigation. Cross-text indexing is highly praised for improving work efficiency. P9 mentions, *"My eyes can quickly switch between different textboxes without losing position, saving me a lot of time."* P14 states, *"This design is the most effective for me, as it is a significant improvement over the baseline. In fact, the content within the textboxes are inter-indexed, and the close connection between the AI summary and the original transcribed text reduces the burden of searching for the original text."* P18 suggests further optimizing the smoothness of navigation to reduce the occasional delay.

Intelligent Optimization and Manual Editing. This feature received widespread praise for its balance between automation and user control. P4 mentioned, *"Intelligent optimization provided me with a great starting point. I found that the revisions made by 'intelligent optimization' to the original transcribed text were almost always correct. At the same time, the*

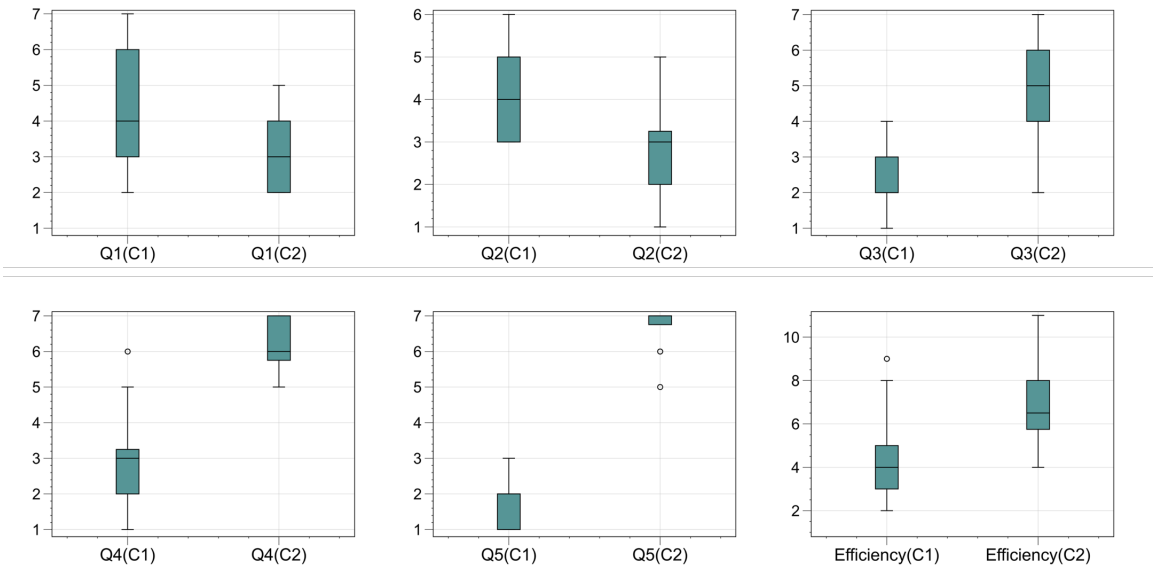


Fig. 7: Box visualization of participants' ratings of five questions Q1-Q5 and their efficiency under two different conditions, C1 and C2.

ability to manually adjust the results gave me more confidence in the final outcome."

Text Merging and Splitting. Nine participants found this feature very useful for reorganizing content. P6 mentioned, *"Being able to merge and split text sections allowed me to guide the generation of summaries based on my understanding and judgment."* However, the other half of the participants indicated that they rarely used this feature and felt it might be unnecessary. This could be related to their level of initiative in exercising subjective agency.

Keyword Highlighting and Filtering. Six participants believed that highlighting keywords enhanced their logical grasp of the entire text, as it helped them quickly locate key information. P1 shared, *"Keyword highlighting made it easier for me to spot important details in dense texts."* P17 pointed out, *"The filtering options helped me narrow down the content, saving a lot of time."*

Summary Condensation Control Slider. Participants appreciated the clever design of the slider, noting not only its flexibility but also its ability to help them understand whether the AI-generated summaries were appropriate. P8 mentioned, *"The slider allowed me to quickly switch between a high-level overview and a more detailed summary, and it seemed to help me understand the strategy behind the AI-generated summaries."*

Personalized Instruction Customization. This feature did not receive as much favor from participants as we had anticipated. It was considered to contribute little to calibrating trust in the AI. P11 noted, *"I tried using custom instructions to generate a casual style, which was interesting, but I found that it didn't change the essence of the content."* P8 added, *"I prefer to use it after generating the overall summary, as I think it's not very necessary for transcribed text or micro-themes. It's suitable for polishing the overall summary and improving its fluency."*

7 DISCUSSION

7.1 Summary of Findings

AI-based meeting summaries can automatically and quickly generate summaries from inputted transcription texts. However, the accuracy of AI results is not always guaranteed. One area that has been overlooked in current research is how the accuracy of AI meeting summaries is perceived and how these summaries are modified through human collaboration. The lack of consideration for human-AI collaborative interaction and trust raises concerns about the practical usability of these systems. This paper explores interaction design and trust calibration cues. Our experimental results highlight the importance of providing enhanced user control, improving workflow efficiency, and clearly displaying the AI summarization process. While features such as personalization and writing guidance may be less popular, they could still hold value for specific user groups.

7.2 Design Implications

Based on our research, particularly the evaluations derived from user feedback, we propose the following design implications to guide the development of future AI-driven summarization tools and even human-AI collaborative assistants.

Emphasize User Control and Transparency. The high ratings for features such as cross-textbox navigation and the summary condensation control slider indicate that users place significant importance on interaction convenience and human-led control capabilities. This suggests that future designs should prioritize providing intuitive control mechanisms, enabling users to adjust or fine-tune AI-generated summaries and their presentation in real-time. By enhancing users' ability to manipulate AI outputs, trust in the system can be effectively increased. Therefore, designs should focus on the flexibility of the user interface, ensuring that users can adapt AI outputs according to their specific needs.

Items	Do you find this feature helpful for trust calibration? (1: not helpful at all → 7: very helpful)								
	1	2	3	4	5	6	7	Mean	Std
Summary Writing Guidelines	2	4	6	5	3			3.67	1.51
Layered Summary Structure				3	4	7	6	5.50	1.05
Cross-Text Box Navigation			1	2	2	3	12	6.00	1.41
AI Optimization & Manual Editing			1	3	4	4	8	5.50	1.38
Manual Merging and Splitting	1	3	4	3	4	3	2	4.00	1.41
Keyword Highlighting and Filtering		1	3	4	5	4	3	4.75	1.39
Summary Condensation Control Slider				1	3	6	10	6.00	0.82
Personalized Summary Customization	2	2	4	6	4	2		4.00	1.26

Fig. 8: User ratings for each design component of MeetSumAid (the color bars represent the number of users).

Structured Emulation of Human Thinking and Behavior. The success of the text-microtheme-overall summary structure highlights the importance of organizing information hierarchically. This structure, which progresses from text to micro-theme and then to an overall summary, mirrors the human thought process of extracting and synthesizing key information layer by layer. By aligning with human cognitive patterns, this approach builds user trust through a logical progression from granular details to higher-level insights [59]. Future human-AI collaboration tools should adopt similar structures to help users better understand how AI-generated outputs are derived and to foster a more intuitive and trustworthy interaction.

Dynamic and Context-Aware Interaction. The effectiveness of dynamic navigation underscores the necessity for systems to adapt dynamically based on user behavior and contextual factors. By tracking user attention (e.g., through gaze or cursor movement) and presenting relevant information in real time, the system can foster more intuitive and responsive interactions. We encourage future AI tools to focus on seamless integration into users' existing workflows.

Granularity of User Control to Enhance Flexibility. The compression level slider demonstrates the value of granting users control over the level of detail in AI-generated outputs. In the context of our meeting summarization scenario, this feature enables users to adjust their understanding of the AI's generation according to their needs, while maintaining a balance between conciseness and comprehensiveness based on subjective judgment. Similar controls can be integrated into other AI tools, allowing users to customize the depth of information presentation.

Function design targeted at specific user groups. The mixed ratings for features such as those targeting specific user groups and writing guidance indicate that, although these tools may not be necessary for all users, they hold significant value for certain groups (e.g., beginners or users unfamiliar with summarization techniques). In contrast, advanced users are likely to prefer simplified and more controlled features. Therefore, future designs could enhance the overall applicability and user satisfaction of the system by tailoring the functionality experience according to the characteristics of different user groups through user profiling and customization strategies.

7.3 Limitations and Future Work

7.3.1 Summarizing Overly Long Meetings

Since our focus is on studying interaction design, we conducted rapid validation on small meetings lasting less than 40 minutes, rather than opting for longer meetings. In daily life, we often encounter large meetings that span several hours. Transcribing the audio of such large meetings into text can result in extremely lengthy documents, potentially containing tens of thousands of words, which exceeds the maximum input length of LLMs. Directly truncating overly long text prevents the LLM from accessing complete information, leading to summaries that may lack substantial necessary content. One possible solution is to divide the meeting into multiple segments, generate summaries for each segment, and then use these summaries as input for the LLM to produce a overall summary (i.e., a "summary of summaries").

7.3.2 Meeting-oriented Fine-tuning of LLMs

Despite the strong generalization capabilities of LLMs, they are not specialized enough in specific domains. By optimizing prompts and other strategies, we leverage the capabilities of LLMs to fulfill the required functions in the system. Our primary focus is on interaction design, so the adjustments to the LLMs have only been made to a usable level. With further fine-tuning or carefully crafted prompts, we expect to achieve more significant improvements, making the LLMs appear more professional in generating meeting summaries.

7.3.3 Data Privacy and Security Considerations

MeetSumAid relies on a third-party LLM API to process meeting audio and transcripts, which introduces potential privacy and security risks, as meeting content may include sensitive or personal information. Future implementations could mitigate these concerns by locally anonymizing or masking sensitive content, or by deploying on-premise or private LLMs [68], thereby reducing external data exposure while preserving AI-assisted summarization capabilities.

7.3.4 Extending to Other Mobile Platforms

The device we selected is the iPad, which represents one of the mobile platforms suitable for users to modify summaries. We believe that the general functional concepts proposed in this paper can be applied to other mobile devices, such as computers and smartphones [69], [70]. Extending to other devices will require addressing challenges related to UI adjustments and adaptations.

8 CONCLUSIONS

In this paper, we present the design, implementation, and evaluation of *MeetSumAid*, a multifunctional interface for human-AI collaborative meeting summarization. The system is designed to support users in calibrating their trust in AI outputs while maintaining fine-grained control over summary generation. MeetSumAid integrates several functional modules, including cross-textbox navigation, a summary compression slider, a hierarchical summary representation, and interactive editing features, enabling users to iteratively refine AI-generated summaries.

The system was evaluated against a baseline interface lacking interaction design enhancements. Quantitative and qualitative results indicate that MeetSumAid significantly improves user trust, engagement, and perceived control over AI-generated summaries. The interface allows real-time adjustments and optimization of summaries through intuitive touch and mouse interactions, resulting in outputs that better align with user requirements.

Additionally, user feedback was analyzed to extract key design insights and inform future directions for AI-assisted summarization interfaces. The findings highlight the importance of structured interaction design in enhancing human-AI collaboration and trust calibration. The implementation and evaluation of MeetSumAid provide a reference framework for the development of efficient, user-centered AI collaboration tools in mobile and cross-platform computing environments.

REFERENCES

[1] J. E. Mroz, J. A. Allen, D. C. Verhoeven, and M. L. Shuffler, "Do we really need another meeting? the science of workplace meetings," *Current Directions in Psychological Science*, vol. 27, no. 6, pp. 484–491, 2018.

[2] W. Kryściński, N. S. Keskar, B. McCann, C. Xiong, and R. Socher, "Neural text summarization: A critical evaluation," *arXiv preprint arXiv:1908.08960*, 2019.

[3] V. Rennard, G. Shang, J. Hunter, and M. Vazirgiannis, "Abstractive meeting summarization: A survey," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 861–884, 2023.

[4] U. Ehsan, P. Wintersberger, Q. V. Liao, E. A. Watkins, C. Manger, H. Daumé III, A. Riener, and M. O. Riedl, "Human-centered explainable ai (hcxai): beyond opening the black-box of ai," in *CHI conference on human factors in computing systems extended abstracts*, 2022, pp. 1–7.

[5] Q. Yang, Y. Hao, K. Quan, S. Yang, Y. Zhao, V. Kuleshov, and F. Wang, "Harnessing biomedical literature to calibrate clinicians' trust in ai decision support systems," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–14.

[6] K. Sokol and P. Flach, "One explanation does not fit all: The promise of interactive explanations for machine learning transparency," *KI-Künstliche Intelligenz*, vol. 34, no. 2, pp. 235–250, 2020.

[7] H. Gu, J. Huang, L. Hung, and X. Chen, "Lessons learned from designing an ai-enabled diagnosis tool for pathologists," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–25, 2021.

[8] M. Fan, X. Yang, T. Yu, Q. V. Liao, and J. Zhao, "Human-ai collaboration for ux evaluation: Effects of explanation and synchronization," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–32, 2022.

[9] K. Chen, L. Wang, Y. Huang, K. Wu, and L. Wang, "Optical sensing-based intelligent toothbrushing monitoring system," *IEEE Transactions on Mobile Computing*, 2024.

[10] ——, "Lit: Fine-grained toothbrushing monitoring with commercial led toothbrush," in *Proceedings of the 29th annual international conference on mobile computing and networking*, 2023, pp. 1–16.

[11] Y. Mou and K. Xu, "The media inequality: Comparing the initial human-human and human-ai social interactions," *Computers in Human Behavior*, vol. 72, pp. 432–440, 2017.

[12] P. Khadpe, C. Kulkarni, and G. Kaufman, "Empathosphere: Promoting constructive communication in ad-hoc virtual teams through perspective-taking spaces," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. CSCW1, pp. 1–26, 2022.

[13] P. Robinette, W. Li, R. Allen, A. M. Howard, and A. R. Wagner, "Overtrust of robots in emergency evacuation scenarios," in *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 2016, pp. 101–108.

[14] E. Horvitz, "Principles of mixed-initiative user interfaces," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, 1999, pp. 159–166.

[15] S. Amershi, D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, J. Suh, S. Iqbal, P. N. Bennett, K. Inkpen *et al.*, "Guidelines for human-ai interaction," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–13.

[16] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information fusion*, vol. 58, pp. 82–115, 2020.

[17] Z. C. Lipton, "The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery." *Queue*, vol. 16, no. 3, pp. 31–57, 2018.

[18] J. Hitsuwari, Y. Ueda, W. Yun, and M. Nomura, "Does human–ai collaboration lead to more creative art? aesthetic evaluation of human-made and ai-generated haiku poetry," *Computers in Human Behavior*, vol. 139, p. 107502, 2023.

[19] P. Karimi, M. L. Maher, N. Davis, and K. Grace, "Deep learning in a computational model for conceptual shifts in a co-creative design system," *arXiv preprint arXiv:1906.10188*, 2019.

[20] M. Guzdial, N. Liao, J. Chen, S.-Y. Chen, S. Shah, V. Shah, J. Reno, G. Smith, and M. O. Riedl, "Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–13.

[21] C.-H. Tsai, Y. You, X. Gui, Y. Kou, and J. M. Carroll, "Exploring and promoting diagnostic transparency and explainability in online symptom checkers," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–17.

[22] C. J. Cai, E. Reif, N. Hegde, J. Hipp, B. Kim, D. Smilkov, M. Wattenberg, F. Viegas, G. S. Corrado, M. C. Stumpe *et al.*, "Human-centered tools for coping with imperfect algorithms during medical decision-making," in *Proceedings of the 2019 chi conference on human factors in computing systems*, 2019, pp. 1–14.

[23] E. Beede, E. Baylor, F. Hersch, A. Iurchenko, L. Wilcox, P. Ruamviboonsuk, and L. M. Vardoulakis, "A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy," in *Proceedings of the 2020 CHI conference on human factors in computing systems*, 2020, pp. 1–12.

[24] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. Weld, "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*, 2021, pp. 1–16.

[25] Y. Zhang, Q. V. Liao, and R. K. Bellamy, "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*, 2020, pp. 295–305.

[26] E. J. De Visser, M. M. Peeters, M. F. Jung, S. Kohn, T. H. Shaw, R. Pak, and M. A. Neerincx, "Towards a theory of longitudinal trust calibration in human–robot teams," *International journal of social robotics*, vol. 12, no. 2, pp. 459–478, 2020.

[27] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human factors*, vol. 46, no. 1, pp. 50–80, 2004.

[28] S. Khastgir, S. Birrell, G. Dhadyalla, and P. Jennings, "Calibrating trust through knowledge: Introducing the concept of informed safety for automation in vehicles," *Transportation research part C: emerging technologies*, vol. 96, pp. 290–303, 2018.

[29] J. M. Kraus, Y. Forster, S. Hergeth, and M. Baumann, "Two routes to trust calibration: effects of reliability and brand information on trust in automation," *International Journal of Mobile Human Computer Interaction (IJMHCI)*, vol. 11, no. 3, pp. 1–17, 2019.

[30] B. Alhaji, M. Prilla, and A. Rausch, "Trust dynamics and verbal assurances in human robot physical collaboration," *Frontiers in artificial intelligence*, vol. 4, p. 703504, 2021.

[31] M. Naiseh, D. Al-Thani, N. Jiang, and R. Ali, "Explainable recommendation: when design meets trust calibration," *World Wide Web*, vol. 24, no. 5, pp. 1857–1884, 2021.

[32] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys (CSUR)*, vol. 51, no. 5, pp. 1–42, 2018.

[33] X. Xu, A. Yu, T. R. Jonker, K. Todi, F. Lu, X. Qian, J. M. Evangelista Belo, T. Wang, M. Li, A. Mun *et al.*, "Xair: A framework of explainable ai in augmented reality," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–30.

[34] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," *Advances in neural information processing systems*, vol. 30, 2017.

[35] X. Wang and M. Yin, "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*, 2023, pp. 1–19.

[36] Q. V. Liao, Y. Zhang, R. Luss, F. Doshi-Velez, and A. Dhurandhar, "Connecting algorithmic research and usage contexts: a perspective of contextualized evaluation for explainable ai," in *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, vol. 10, no. 1, 2022, pp. 147–159.

[37] Y. Xie, M. Chen, D. Kao, G. Gao, and X. Chen, "Chexplain: enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[38] D. Kost, "You're right! you are working longer and attending more meetings," *Harvard Business School*, 2020.

[39] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of research and development*, vol. 2, no. 2, pp. 159–165, 1958.

[40] G. Tur, A. Stolcke, L. Voss, J. Dowding, B. Favre, R. Fernández, M. Frampton, M. Frandsen, C. Frederickson, M. Graciarena *et al.*, "The calo meeting speech recognition and understanding system," in *2008 IEEE Spoken Language Technology Workshop*. IEEE, 2008, pp. 69–72.

[41] K. Riedhammer, D. Gillick, B. Favre, and D. Hakkani-Tür, "Packing the meeting summarization knapsack," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[44] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[45] S. Ramprasad, E. Ferracane, and Z. C. Lipton, "Analyzing llm behavior in dialogue summarization: Unveiling circumstantial hallucination trends," *arXiv preprint arXiv:2406.03487*, 2024.

[46] L. Tang, I. Shalyminov, A. W.-m. Wong, J. Burnsky, J. W. Vincent, Y. Yang, S. Singh, S. Feng, H. Song, H. Su *et al.*, "Tofueval: Evaluating hallucinations of llms on topic-focused dialogue summarization," *arXiv preprint arXiv:2402.13249*, 2024.

[47] A. R. Wagner and P. Robinette, "An explanation is not an excuse: Trust calibration in an age of transparent robots," in *Trust in Human-Robot Interaction*. Elsevier, 2021, pp. 197–208.

[48] S. Chen, M. Wu, K. Q. Zhu, K. Lan, Z. Zhang, and L. Cui, "Llm-empowered chatbots for psychiatrist and patient simulation: Application and evaluation," *arXiv preprint arXiv:2305.13614*, 2023.

[49] Y. Moslem, G. Romani, M. Molaei, J. Kelleher, R. Haque, and A. Way, "Domain terminology integration into machine translation: Leveraging large language models," in *Proceedings of the Eighth Conference on Machine Translation*, 2023, pp. 902–911.

[50] Z. Elyoseph, E. Refoua, K. Asraf, M. Lvovsky, Y. Shimoni, and D. Hadar-Shoval, "Can large language models "read your mind in your eyes"?" *JMIR Mental Health [Preprint]. doi*, vol. 10.

[51] S. Lin, J. Warner, J. Zamfirescu-Pereira, M. G. Lee, S. Jain, S. Cai, P. Lertvittayakumjorn, M. X. Huang, S. Zhai, B. Hartmann *et al.*, "Rambler: Supporting writing with speech via llm-assisted gist manipulation," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–19.

[52] D. Van Veen, C. Van Uden, M. Attias, A. Pareek, C. Bluethgen, M. Polacin, W. Chiu, J.-B. Delbrouck, J. M. Z. Chaves, C. P. Langlotz *et al.*, "Radadapt: Radiology report summarization via lightweight domain adaptation of large language models," *arXiv preprint arXiv:2305.01146*, 2023.

[53] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B. Hashimoto, "Benchmarking large language models for news summarization," *arXiv preprint arXiv:2301.13848*, 2023.

[54] M. T. R. Laskar, X.-Y. Fu, C. Chen, and S. B. Tn, "Building real-world meeting summarization systems using large language models: A practical perspective," *arXiv preprint arXiv:2310.19233*, 2023.

[55] J. Zamfirescu-Pereira, H. Wei, A. Xiao, K. Gu, G. Jung, M. G. Lee, B. Hartmann, and Q. Yang, "Herding ai cats: Lessons from designing a chatbot by prompting gpt-3," in *Proceedings of the 2023 ACM Designing Interactive Systems Conference*, 2023, pp. 2206–2220.

[56] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, and Q. Yang, "Why johnny can't prompt: how non-ai experts try (and fail) to design llm prompts," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–21.

[57] H. Dang, K. Benharrak, F. Lehmann, and D. Buschek, "Beyond text generation: Supporting writers with continuous automatic text summaries," in *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, 2022, pp. 1–13.

[58] P. Jiang, J. Rayan, S. P. Dow, and H. Xia, "Graphologue: Exploring large language model responses with interactive diagrams," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023, pp. 1–20.

[59] J. Zeng, K. Chen, R. Wang, Y. Li, M. Fan, K. Wu, X. Qi, and L. Wang, "Contractmind: Trust-calibration interaction design for ai contract review tools," *International Journal of Human-Computer Studies*, vol. 196, p. 103411, 2025.

[60] L. Zhang, Q. Chen, W. Wang, C. Deng, S. Zhang, B. Li, W. Wang, and X. Cao, "Mderank: A masked document embedding rank approach for unsupervised keyphrase extraction," *arXiv preprint arXiv:2110.06651*, 2021.

[61] T. Bäckström, O. Räsänen, A. Zewoudie, P. P. Zarazaga, L. Koivusalo, S. Das, E. G. Mellado, M. B. Mansali, D. Ramos, S. Kadiri, and P. Alku, *Introduction to Speech Processing*, 2nd ed., 2022. [Online]. Available: https://speechprocessingbook.aalto.fi

[62] M. Mauch and S. Dixon, "pyin: A fundamental frequency estimator using probabilistic threshold distributions," in *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2014, pp. 659–663.

[63] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[64] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[65] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[66] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative research in psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[67] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010.

[68] K. Chen, Y. Huang, Y. Chen, H. Zhong, L. Lin, L. Wang, and K. Wu, "Lisee: A headphone that provides all-day assistance for blind and low-vision users to reach surrounding objects," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–30, 2022.

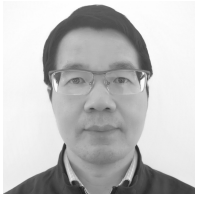
[69] Y. Huang, K. Chen, Y. Huang, L. Wang, and K. Wu, "Vi-liquid: unknown liquid identification with your smartphone vibration,"

in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 174–187.

[70] K. Chen, J. Xiang, W. Tan, K. Chen, Y. Luo, C. Ma, K. Wu, and L. Wang, "Pit: A novel toothbrush providing real-time and robust plaque indication during brushing," in *Proceedings of the 23rd Annual International Conference on Mobile Systems, Applications and Services*, 2025, pp. 15–27.

Lu Wang (Senior Member, IEEE) is currently an associate professor in College of Computer Science and Software Engineering, Shenzhen University, China. His research interests focus on wireless communications and mobile computing.

Yilong Li received the M.S. degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2025. He is currently working in the field of Artificial Intelligence, focusing on AI Agent technologies and intelligent system development. His professional interests include large language models, autonomous agents, and AI-driven software applications.



Jianhua He (Senior Member, IEEE) is a Professor at the School of Computer Science and Electronic Engineering, University of Essex. He received his Ph.D. degree in Electrical and Electronic Engineering from Nanyang Technological University (NTU), Singapore, in 2002. He is an active researcher with over 170 publications in leading international journals and conferences, including MobiCom, INFOCOM, and IEEE TWC. His work has received more than 7,000 citations, and he holds 16 patents (4 granted) related to OCR, document understanding, and knowledge graphs. His research interests include wireless communications, mobile networking, IoT, edge computing, connected and autonomous vehicles, smart cities, cybersecurity, blockchain, and AI-driven data analytics. He also works on deep learning for object detection, OCR, and document understanding. He currently coordinates several EU Horizon and H2020 projects, including COSAFE, VESAFE, SECOM, and COVER, focusing on intelligent, connected, and autonomous systems. He serves as an Editor for IEEE Wireless Communications Letters, The Computer Journal, and Frontiers in Future Transportation.

Yue Ling Che (S'11-M'15) received the B.Eng. and the M.Eng. degrees from the University of Electronic Science and Technology of China in 2006 and 2009, respectively, and the Ph.D. degree from the Nanyang Technological University, Singapore, in 2014, all in electrical engineering. From 2014 to 2016, she was a postdoc research fellow in Engineering Systems and Design Pillar, Singapore University of Technology and Design. She is now an Associate Professor in College of Computer Science and Software Engineering with the Shenzhen University. Her research interests include integrated sensing and communication systems, AI-enabled wireless networks, UAV-enabled mobile communications, wireless energy transfer, and stochastic modeling and optimization methods.

Kaishun Wu (Fellow, IEEE) is the Associate Vice President for Research of the Hong Kong University of Science and Technology (Guangzhou). He is also a full professor of the DSA & IoT Thrust Area under the Information Hub. He received his Ph.D. degree in computer science and engineering from HKUST in 2011. He is an active researcher with more than 200 papers published on major international academic journals and conferences, as well as more than 100 invention patents, including 9 from the USA. He received the 2012 Hong Kong Young Scientist Award, the 2014 Hong Kong ICT awards: Best Innovation and 2014 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He is an IET Fellow.

Xiaoke Qi received the B.S. degree in communication engineering from Nankai University, Tianjin, China, in 2009, and the Ph.D. degree in signal and information processing from the Institute of Acoustics, Chinese Academy of Sciences, Beijing, China, in 2014. She is currently an associate professor at China University of Political Science and Law, Beijing, China. Her current research interests include multimedia processing, natural language processing, machine learning, and wireless communication.

Kaixin Chen received the MS degree from the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China, in 2024. He is currently working toward the PhD degree with the Internet of Things Thrust of Information Hub, Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China. He has authored or coauthored papers in premier conferences, such as ACM MobiCom, ACM UbiComp, and ACM MobiSys. His research interests include mobile computing and Internet of Things (IoT).