Personalized Real-time Jargon Support for Online Meetings

Yifan Song
University of Illinois Urbana-Champaign
Urbana, IL, USA
yifan33@illinois.edu

Wing Yee Au
Fujitsu Research of America
Santa Clara, CA, USA
wau@fujitsu.com

Hon Yung Wong
Fujitsu Research of America
Santa Clara, CA, USA
awong@fujitsu.com

Brian P. Bailey
University of Illinois Urbana-Champaign
Urbana, IL, USA
bpbailey@illinois.edu

Tal August
University of Illinois Urbana-Champaign
Urbana, IL, USA
taugust@illinois.edu

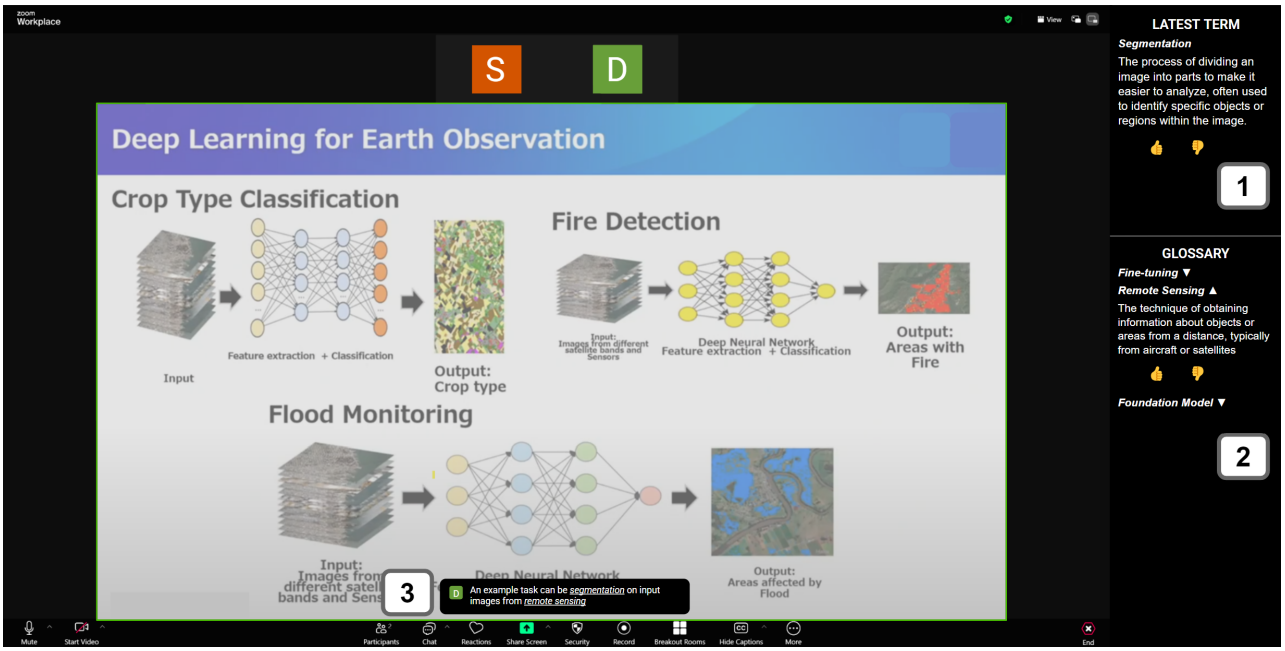


Figure 1: In an online meeting, the speaker is presenting a project about deep learning in earth science and screen sharing the slides. The listener uses ParseJargon with three interface components: 1) The latest jargon definition in concise plain language; 2) Glossary for all jargon terms appeared in the meeting for revisiting; 3) Real-time caption highlighting identified jargon

Abstract

Effective interdisciplinary communication is frequently hindered by domain-specific jargon. To explore the jargon barriers in-depth, we conducted a formative diary study with 16 professionals, revealing critical limitations in current jargon-management strategies during workplace meetings. Based on these insights, we designed ParseJargon, an interactive LLM-powered system providing real-time personalized jargon identification and explanations tailored to users' individual backgrounds. A controlled experiment comparing ParseJargon against baseline (no support) and general-purpose (non-personalized) conditions demonstrated that personalized jargon support significantly enhanced participants' comprehension, engagement, and appreciation of colleagues' work, whereas general-purpose support negatively affected engagement. A follow-up field study validated ParseJargon's usability and practical value in real-time meetings, highlighting both opportunities and limitations for real-world deployment. Our findings contribute insights into designing personalized jargon support tools, with implications for broader interdisciplinary and educational applications.

1 Introduction

Effective interdisciplinary communication is crucial for successful collaboration across different fields [17, 58, 70]. Particularly within the workspace, professionals frequently face challenges to convey specialized knowledge to colleagues from other disciplines [22, 32, 36, 69]. For instance, a machine learning engineer

Preprint submitted to arXiv, August 2025.

might struggle to communicate concepts like "embedding" to a compliance officer concerned with data privacy, while healthcare professionals might face challenges describing "quasi-experimental designs" to policymakers without medical expertise. Such gaps in communication caused by domain-specific jargon limit interdisciplinary innovation and effective collective problem-solving [18, 24], leading to misunderstandings, decreased engagement, reduced comprehension, and undervaluing of contributions from colleagues [11, 13, 60].

Strategies like preparing beforehand, asking clarifying questions during meetings, or looking up terms may alleviate some jargon-related problems, however, these approaches often fall short in real-time meeting scenarios. Preparing beforehand is often impractical as it assumes participants know precisely what terms will be challenging and have time to learn [55]. Prior research findings also suggest that social dynamics and hierarchical structures discourage interrupting speakers with questions [63], particularly among junior or culturally reserved employees, while independently searching for definitions during conversations introduces distractions that disrupt context continuity [9].

Recent advances in speech-to-text technologies and large language models (LLMs) offer promising potential to overcome these limitations with automated jargon support. Prior research has explored computational techniques for jargon identification and explanation [6, 33, 43, 50], and developed augmented interfaces that enhance comprehension during meetings through interactive transcripts or captions [15, 16, 38, 40]. However, existing systems typically neglect two critical factors for effective jargon support in meetings: real-time support and personalized support. Most prior jargon support systems either target only static text content [1, 7, 31], or they fail to consider user-specific background knowledge by providing uniform jargon assistance to all users [41]. Such generic solutions can overwhelm users with irrelevant or excessive information, which reduces trust and user engagement, especially in real-time meeting settings.

To address these gaps, we introduce ParseJargon [1], a real-time personalized jargon support system designed for online meetings (Figure 1). To systematically investigate jargon barriers in meeting communication and evaluate the effectiveness of our system, we propose the following research questions:

- **RQ1**: What jargon barriers emerge in real-time conversations, and how effective are the strategies people currently employ to address them?
- **RQ2**: How can an LLM-based system support online meetings by addressing jargon barriers in real-time?
- **RQ3**: How does personalization, in the form of selecting what terms to define based on a user's background, impact the effectiveness of ParseJargon?

To address **RQ1**, we conducted a two-week diary study with 16 professionals from a large technology company, documenting jargon encountered during real meetings and strategies participants used to manage unfamiliar terms. We found that participants often chose passive strategies, such as waiting for additional context from the speaker, but these typically proved ineffective, leaving confusion unresolved. Searching for definitions was also common but often disrupted listening and engagement. Directly asking speakers to clarify jargon was effective yet rarely adopted due to social dynamics and timing concerns. These findings aligned with prior research insights [9, 13, 63] and underscored the need for timely, non-disruptive access to jargon explanations during real-time conversations.

To address **RQ2** and **RQ3**, we conducted a controlled within-subjects experiment with three conditions: a baseline without jargon support, a general-purpose jargon support without personalization, and a personalized jargon support provided by ParseJargon. Results showed that while the general-purpose ParseJargon improved comprehension compared to the baseline, it negatively affected engagement by overwhelming participants with excessive jargon explanations. In contrast, personalized ParseJargon significantly improved comprehension and maintained participants' engagement by accurately predicting relevant jargon based on an individual's backgrounds. The findings from the controlled experiment not only serve as a proof-of-concept technical evaluation for ParseJargon, but also provide further insights to inform system design. Finally, to validate ParseJargon's practical utility, we deployed the full system in a real online meeting within the company to gather preliminary user feedback and usability insights.

In summary, this paper makes the following contributions:

- Empirical findings from a diary study highlighting jargon barriers in real-time meetings and limitations of current strategies professionals use to manage them.
- Design of a real-time jargon support system for online meetings, ParseJargon, powered by LLM to provide personalized support tailored to each audience's background.
- Results from a controlled evaluation demonstrating that personalized jargon support significantly enhances comprehension, sustains engagement, and increases appreciation of colleagues' contributions compared to general-purpose jargon support.

2 Related Work

2.1 Jargon Barriers in Communication

Jargon, defined as specialized terminology used within specific fields, has been frequently identified as a significant barrier in interdisciplinary communication [34, 51]. Such barriers arise because experts often unconsciously rely on domain-specific language that non-experts cannot readily understand. These issues have been particularly highlighted within workspaces, where cross-team collaboration is essential with employees often having diverse professional and educational backgrounds [23, 36, 69]. Similar communication challenges have also been reported in interdisciplinary research [17, 18, 70] and educational collaborations [58], underscoring the pervasive nature of the jargon barriers across contexts.

Extensive prior research has shown that jargon can hinder effective communication, including reducing comprehension with increased risk of miscommunication [13, 24], leading to resistance toward new ideas from other fields [11], and disrupting information processing which decreases engagement [60] and increases cognitive workload [9]. This cognitive burden is especially common in interdisciplinary teamwork, where experts from different domains must frequently collaborate but often lack shared terminology or

[1] *ParseJargon* stands for *P*ersonalized *A*ssistant for *R*eal-time *S*upport in *E*xplaining *Jargon*

background knowledge, making effective communication across domains more challenging [63].

Traditional approaches to mitigating jargon have included simplifying language and employing analogies [55], measuring word familiarity and frequency [64], and developing domain-specific vocabulary lists [26]. While these strategies have been helpful, they typically require active effort from speakers or listeners to seek information, thus disrupting the natural flow of conversation. This limitation highlights the need for advanced technological interventions, motivating computational approaches for jargon support.

2.2 Jargon Support Technologies and Systems

Advancements in language technologies have enabled computational support for identifying and explaining jargon. A core task in this space is complex word identification, which aims to determine terms likely unfamiliar to target users, with early benchmarks introduced by Shardlow [59]. Recent methods have applied LLMs to measure jargon complexity [43] and adapted identification models to specialized domains such as biomedical research [30] or specific jargon usage like acronyms [57]. On the other hand, jargon explanation has been studied from the perspectives of definition extraction [67], definition generation [6], or hybrid approach [33]. Closely related to jargon explanation is the task of text simplification, which transforms complex content into simpler and more accessible versions [44, 66].

Building on these techniques, researchers have designed interactive systems for jargon support, especially within reading interfaces [25, 42]. For example, ScholarPhi [31] provides an automatically generated glossary for important scientific terms, while Paper Plain [7] offers in-situ definitions of unfamiliar terms and plain language summaries. Other more recent works target how to augment medical progress notes [35] or explore how user-generated analogies can support jargon understanding during scientific reading [10]. However, these systems are designed primarily for asynchronous content. Our work extends this line by targeting real-time spoken conversations, enabling dynamic jargon support during online meetings.

A closely related work is StopGap [41], which explores the design space of real-time LLM-based knowledge assistance through multiple explanation formats for jargon in technical videos. Although their findings offer valuable insights into user preferences and interface design, their prototype was used in a design probe study without a real-time implementation or conversational setting. In contrast, while ParseJargon was also evaluated in a simulated meeting environment as a proof-of-concept, our formative study explored interactions between people, and we further implemented and deployed a fully functional system in live meetings. Moreover, while StopGap briefly discusses the potential benefits of personalization, their system provides one-size-fits-all support. In contrast, ParseJargon explicitly implements and evaluates personalized jargon support tailored specifically to each user's background.

2.3 Personalization in Jargon Support

Personalization plays a critical role in tailoring jargon support to users' prior knowledge and professional background. Early work adapted complex word identification and lexical simplification models to individual users, substituting unfamiliar terms based on personal vocabulary profiles [37]. Subsequent efforts demonstrated that modeling word complexity at the individual level significantly improved performance [27] and introduced approaches for generating personalized descriptions of scientific concepts [48]. More recent research extended this by incorporating personal data into scientific jargon identification, showing that LLMs can serve as a baseline for personalized jargon support for researchers when reading interdisciplinary articles [29].

Beyond algorithmic personalization, HCI research has investigated how users perceive and interact with personalized language systems. For example, researchers have designed interfaces that adapt scientific information to users' expertise using rule-based templates [53]. A recent study investigated the effects of adaptive plain language on diverse audiences and offered insights into using LLMs to generate summaries tailored to different levels of expertise [5]. In journalism contexts, while early work has relied on manual efforts [2], more recent work has used GPT4 to help science journalists produce audience-appropriate content, revealing promising potential and challenges to adapt jargon dynamically [52]. Other work highlights that even perceived personalization, such as user-controlled filtering, can shape how people engage with explanation systems, including trust, satisfaction, and comprehension [14].

These studies provide the foundation for our approach, which incorporates audience background information to deliver real-time personalized jargon support. Yet, to our knowledge, no prior personalized jargon system targets live meetings, a gap that ParseJargon directly addresses.

2.4 Enhancing Meeting Communication

The HCI community has long studied computer-mediated conversations, from text-based group chat [19, 20, 49] to audio/video-based online meetings [21, 39, 46]. Systems like Tilda [72] and Wikum+ [65] focus on collaborative tagging and summarization in group chats to facilitate the comprehension and sensemaking of long chat streams. While these systems effectively support asynchronous collaboration, Meeting Bridges[68] aims to bridge the gap between synchronous meetings and asynchronous conversations, ensuring that meeting content is preserved for post-meeting engagement to avoid collaboration overload during remote meetings.

With increasing remote and hybrid collaboration, the advancements in speech-to-text techniques empowered HCI researchers to improve real-time meeting interactions through dynamic, interactive interfaces. For instance, TalkTraces [15] visualizes ongoing meeting content to help participants track topics in real time, while MeetScript [16] provides interactive transcripts that enable collaborative annotation, significantly enhancing participant engagement. Additionally, Mirrorverse [28] shows the value of augmenting live calls to dynamically accommodate diverse meeting situations and user requirements, while CrossTalk [71] supports speakers by generating real-time talking points using LLMs, facilitating more informed and structured contributions. Son et al. [61] also investigated distraction management, introducing systems that intelligently structure and schedule interruptions to balance multitasking and attention during online meetings. Other innovations

include speech agents acting as personal assistants to boost meeting productivity [45] and real-time summaries designed to maintain engagement in live interactions [3].

Despite these significant advancements, most prior real-time systems have primarily focused on summarization, speaker assistance, or distraction management, leaving jargon-related comprehension challenges relatively unexplored. Our system, ParseJargon, addresses this specific gap by leveraging real-time meeting transcripts to address jargon barriers during online meetings to enhance both comprehension and engagement in cross-background conversations.

3 Formative Diary Study

3.1 Methodology

We conducted a diary study involving 16 professionals at a technology company in North America to systematically identify jargon barriers in workplace meetings and to evaluate the effectiveness of existing strategies that participants use to address the barriers (**RQ1**). Participants came from diverse professional roles (7 researchers, 5 engineers, 2 marketers, and 2 directors) and varying experience levels (6 junior, 7 mid-career, and 3 senior employees). More demographic information can be found in Appendix Table 9.

Over a two-week period, participants documented instances of encountering unfamiliar jargon during their real workplace meetings, which included events ranging from weekly stand-ups to larger cross-team presentations. For each unfamiliar term, participants recorded the date, meeting type (within-team or cross-team), number of attendees, and subjectively classified the term's domain relative to their expertise (same, different, or unsure). Participants then selected actions from a predefined list of how they had addressed these unfamiliar terms. The predefined list included six actions grouped into three categories: ***passive*** (*wait for explanation, skip*), ***asking*** (*interrupt and ask, ask afterward*), and ***searching*** (*search internally, search externally*). This action list was initially motivated from prior work [13, 22, 32, 63], then validated and refined by the researchers through two rounds of internal meeting testing. Participants also had the option to specify additional actions not listed, but no new actions emerged during the study period. Participants then rated each action's helpfulness on a 5-point Likert scale (1 = not helpful, 5 = very helpful). To minimize participants' effort and avoid disrupting their natural meeting behaviors, we provided a structured spreadsheet template (Figure 2) that's easy to fill.

We quantitatively analyzed the diary data using descriptive statistics to summarize action frequencies and average helpfulness ratings. To further contextualize these findings, brief post-study interviews were conducted with each participant to understand their rationale behind selecting specific actions and to discuss the perceived limitations of each strategy. Interviews were either audio-recorded and transcribed, or were documented through detailed note-taking when recording was not feasible due to practical constraints.

3.2 Findings

Participants documented 123 unfamiliar jargon terms across 47 meetings (approximately 2.6 terms per participant per meeting). Of these terms, 56 were classified as from a different domain, 45 from the same domain, and 22 as uncertain ("not sure"). Table 1 summarizes the quantitative outcomes, including action frequencies and average helpfulness ratings.

Passive strategies were the most frequently used but the least effective. Participants mainly adopted passive strategies, either waiting for explanations (45.5%) or intentionally skipping terms perceived as unnecessary to understand (9.8%). Waiting was described as the most natural and straightforward approach (10 participants), sometimes *"the only appropriate option"* (P3). However, participants found that waiting didn't often resolve their confusion, reflected by the lowest helpfulness ratings (mean=3.11). This aligns with prior findings suggesting passive listening is insufficient for resolving misunderstandings of specialized terminology [13]. Participants distinguished "skip" as intentionally disregarding the term entirely, thus no helpfulness rating was provided for this action.

Asking was most effective but least employed due to timing and social constraints. Directly asking speakers to clarify during meetings ("interrupt and ask," 4.9%) or afterward ("ask afterward," 8.1%) were consistently rated as most effective (mean ratings of 4.5 and 4.4, respectively). Most participants agreed they would typically receive a helpful answer from the speaker (12 participants), yet these strategies were infrequently adopted. Common reasons for this reluctance included difficulty finding appropriate timing (9 participants) and perceived social constraints (7 participants), such as *"feeling impolite to interrupt"* (P15) and sensitivity to *"power dynamics"* (P4). These findings resonate with prior research highlighting cultural and hierarchical constraints discourage direct questioning in professional contexts [63].

Searching was somewhat helpful but also disruptive during meetings. Searching externally (25.2%) or internally (within company materials, 6.5%) ranked second in both frequency and helpfulness (mean ratings of 3.52 and 3.62, respectively). However, participants expressed concerns about searching being distracting from the ongoing meeting content (10 participants) and sometimes feeling socially inappropriate by visibly checking personal devices (3 participants). Such challenges align with previous studies that interruption from ongoing meeting content can be disruptive [9]. While searching could be beneficial when a clear definition was found, it became challenging when jargon terms were highly context-specific or internal to the organization (7 participants), as described by P9, *"Googling is hard if the terms are only used in [company name], especially acronyms… internal materials are dispersed and not indexed. The time to do it is way longer"*.

Cross-domain jargon exacerbated barriers and reduced the effectiveness of existing strategies. Participants' responses varied notably based on the domain of unfamiliar terms. When jargon originated from within their own domains, participants were more likely to proactively ask for clarifications (22.2% for same-domain terms vs. only 3.6% for cross-domain terms). Conversely, cross-domain terms led participants toward passive strategies more frequently (60.7% for cross-domain vs. 42.3% for same-domain). Participants described that they would need more time and effort to understand cross-domain jargon (8 participants) and felt less confident to address them (4 participants), causing existing strategies less effective (helpfulness ratings for same-domain terms vs. cross-domain terms: wait - 3.5 vs. 2.97; search - 3.94 vs. 3.2).

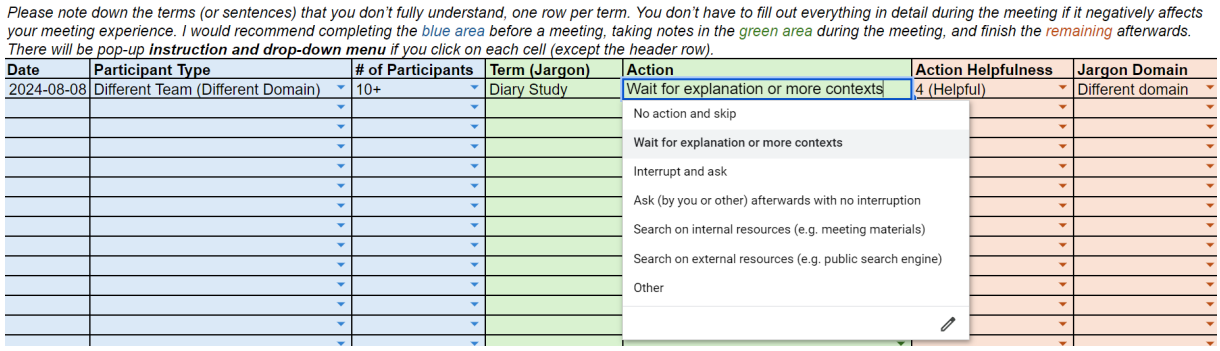


Figure 2: A screenshot of the diary study spreadsheet template with an example jargon of "diary study"

Table 1: Diary study results: the number of occurrences (with percentage) and the average helpfulness rating of each action (out of 5) - passive strategy is most popular but least effective while asking is effective but least employed.

		Overall		Same Domain		Different Domain	
Strategy	**Action**	*N (%)*	Helpfulness	*N (%)*	Helpfulness	*N (%)*	Helpfulness
Asking	Interrupt and ask	6 (4.9%)	4.5	5 (11.1%)	4.6	0 (0%)	-
	Ask afterward	10 (8.1%)	4.4	5 (11.1%)	4.2	2 (3.6%)	4.5
Searching	Search internally	8 (6.5%)	3.62	2 (4.4%)	4.0	6 (10.7%)	3.33
	Search externally	31 (25.2%)	3.52	14 (31.1%)	3.93	14 (25%)	3.14
Passive	Wait for explanation	56 (45.5%)	3.11	16 (35.6%)	3.5	29 (51.8%)	2.97
	Skip	12 (9.8%)	-	3 (6.7%)	-	5 (8.9%)	-

Such findings provide a strong motivation for personalized jargon support tailored to individuals' domain-specific backgrounds to alleviate cognitive overload, which is also suggested by Guo et al. [29] that personalized jargon identification is crucial.

These findings collectively underscore persistent barriers posed by jargon in cross-background conversations, clearly indicating user needs for non-disruptive, timely, and personalized jargon support integrated directly into meeting workflows.

3.3 Design Considerations

Based on insights from the diary study and prior literature, we derived two key design considerations, which directly guided the design and implementation of our system:

Automatic Real-time Jargon Support: Participants frequently relied on passive strategies (e.g., waiting for potential explanations) that were ineffective, or disruptive strategies (e.g., independently searching for definitions) that caused distraction. Directly asking was effective but rarely used due to social constraints such as politeness concerns and power dynamics. Thus, the system should automatically identify jargon and provide immediate explanations without manual intervention, enabling seamless access without disrupting the meeting flow.

Personalized Jargon Identification and Explanation: Compared to jargon terms from a user's familiar domain, it would be helpful to more precisely identify cross-domain jargon, which significantly increased cognitive load and reduced the effectiveness of existing strategies. Consequently, the system should deliver personalized jargon predictions tailored to individual users' backgrounds, enhancing relevance and minimizing cognitive overload.

4 System Design

Building on the design considerations identified through our formative diary study (§3), we developed ParseJargon, a real-time system that automatically identifies and explains jargon tailored to individual participants' backgrounds during online meetings. In this section, we first illustrate how the system supports users through an example scenario (§4.1), followed by a description of the system interface and backend (§4.2), and conclude with technical implementation details (§4.3).

4.1 Example Usage Scenario

Consider a scenario in which a researcher presents a project involving deep learning applications in earth science to a business team responsible for product development (Figure 1). The audience lacks expertise in both machine learning and earth science, making it challenging for them to follow key technical terms such as "segmentation" or "remote sensing", despite the speaker's effort to explain these terms briefly. The business team members are reluctant to interrupt the speaker or independently search for definitions, fearing social discomfort and potential distraction to miss other important points, leading to persistent confusion and potentially undervaluing the presented research.

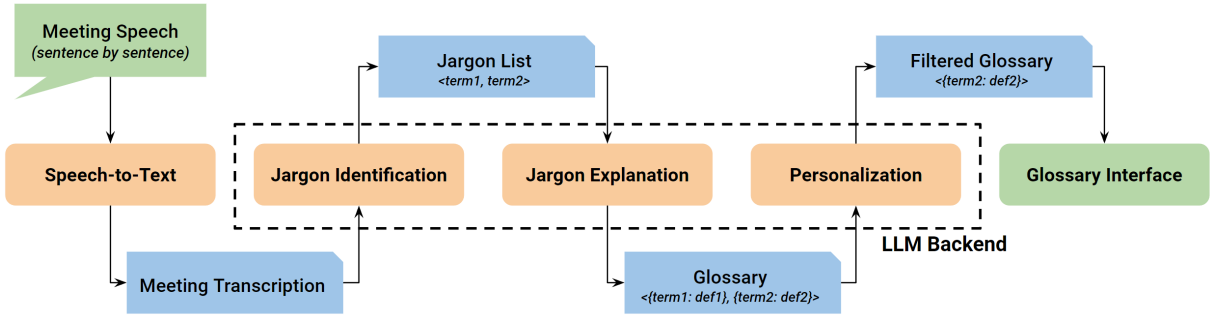


Figure 3: System Architecture Flowchart: from speech input to personalized glossary output

Now with ParseJargon, the business team has real-time access to automatically generated explanations of unfamiliar jargon directly within their meeting interface. As the speaker presents, terms like "segmentation" and "remote sensing" are identified as jargon based on each audience's specific background and appear in a glossary sidebar next to the main meeting window, offering concise and accessible definitions. Audience members no longer need to actively search for terms or hesitate about interrupting the flow; instead, they seamlessly access essential explanations. Note that the system works for any audio/video meeting scenarios, with or without screen sharing. With ParseJargon, the team maintains focus, enhances their comprehension, and can fully appreciate the value and implications of the presented research by fully understanding the technical difficulties.

4.2 System Architecture

Our system architecture (Figure 3) consists of two main components: an LLM-powered backend for real-time jargon identification, explanation, and personalization, and a frontend user interface that displays jargon definitions seamlessly within the meeting environment.

4.2.1 Backend Technology. The backend leverages the OpenAI GPT model to perform three interconnected tasks: *jargon identification*, *jargon explanation*, and *personalization*. These tasks are executed through prompting techniques to ensure both efficiency and effectiveness. Prompts and parameters are provided in Appendix.

- **Jargon Identification & Explanation:** ParseJargon first fetches the live transcription generated by the service provided in online meeting platforms. Upon receiving the transcription, our backend identifies potential jargon and generates concise plain-language definitions for each term using LLM. This process uses a single combined prompt, analyzing each sentence of the meeting transcript sequentially. Each identified jargon term is defined only once throughout the meeting.
- **Personalized Filtering:** To tailor jargon support to individual audience members' expertise, the system applies a second filtering step. Using a separate prompt, the system assesses each participant's professional or educational background (provided via text-based user profiles, e.g., *"I am a quantum computing researcher and hold a Physics PhD."*). It then removes any identified jargon terms that the user is likely already familiar with based on their background. This personalized filtering significantly reduces unnecessary cognitive load, presenting users only with definitions they are likely to need.

We found that this two-step approach of first identifying terms and then filtering for those relevant to a user was more effective than a single prompt for reducing the number of terms that the user may already know.

4.2.2 Interface Design. The user interface of ParseJargon (Figure 1) integrates seamlessly into standard online meeting platforms. The primary components of the interface include:

- **Real-time Captions with Highlighted Jargon:** The live captions are generated by the transcription service of the online meeting platform. ParseJargon highlights the identified jargon terms for easy recognition.
- **Latest Jargon Term Definition:** The definition for currently identified jargon (latest term) appears in real-time, enabling users to quickly glance at explanations. Users can provide feedback by indicating their preference for each identified jargon term ("like" or "dislike"). The preference list is then added to the system backend via the personalization filtering prompt to iteratively refine their user profiles and improve future personalization accuracy.
- **Persistent Glossary Sidebar:** All identified jargon terms from the meeting accumulate in a persistent glossary list, allowing users to revisit terms and definitions at any point during the meeting.

The persistent glossary and user feedback mechanisms were introduced as iterative design improvements informed by findings from the controlled experiment (detailed in §5.3). Each latest term is displayed for at least 7 seconds (based on average reading speed [12]) or until the next jargon term is identified. If a new term is identified within 7 seconds, it is queued in the glossary list. This timing mechanism is informed by prior research [41] and user feedback from our controlled experiment, where there is no minimum term displaying interval and the maximum of simultaneously displayed terms is three.

4.3 Implementation Details

ParseJargon is implemented as a Chrome browser extension integrated into web-based meeting platforms (currently only supports Zoom). The frontend interface is developed with `React.js`. The backend server, developed using Python `Flask`, manages calls using the OpenAI API. In the full implementation for the field study, we use GPT-4o-mini[2] for its fast speed as we value low latency more in real-time experience. Conversely in the controlled experiment, to evaluate the effectiveness of our technology, we chose GPT-4o[3] for its superior accuracy, as response speed was not needed with pre-recorded videos (more information in the next section). The server is deployed on Heroku with a PostgreSQL database.

5 Controlled Experiment

To systematically evaluate ParseJargon's backend capabilities, we conducted a controlled experiment, aimed to assess whether the system improves meeting comprehension, engagement, and participants' perceived value of speakers' presentations, and how effective the personalized filter is. We made the following hypotheses, associated with **RQ2** and **RQ3**, respectively:

- **H1:** ParseJargon will enhance participants' comprehension, engagement, and perceived value of others' work during meetings.
- **H2:** Participants will find the personalized version of ParseJargon more helpful and less distracting compared to a non-personalized general-purpose version.

Our system depends on live captions provided by online meeting platforms' speech-to-text engines (e.g., Otter.ai for Zoom [54], Azure Speech-to-Text for Teams [8]). However, as these transcriptions are not always reliable [56], we chose to use pre-recorded videos with manually verified transcriptions in this controlled experiment. This allowed us to isolate and precisely evaluate the backend capabilities of ParseJargon as a proof-of-concept, without interference from transcription inaccuracies or latency variations. We conducted another supplementary lightweight field study (§6) to evaluate usability in real meeting scenarios.

5.1 Methodology

5.1.1 Participants and Presentation Preparation. We recruited seven interns from diverse teams within the same technology company as our diary study (no overlap with diary study or field study participants). This participant selection ensured minimal prior knowledge about each other's projects, creating authentic conditions for assessing jargon barriers. Participants had varied educational backgrounds (including Computer Science, Applied Mathematics, Physics, Civil Engineering, and Statistics) and diverse intern job roles. Detailed demographics are included in Table 2.

Each participant was asked to prepare and present a 10-minute presentation about their ongoing project, supported by slides. These presentations were recorded and initially transcribed via Microsoft Teams, the company's primary communication platform to mimic a realistic workplace scenario for the experiment. All the transcripts were then manually verified to ensure correctness.

[2] gpt-4o-mini-2024-07-18, https://platform.openai.com/docs/models/gpt-4o-mini
[3] gpt-4o-2024-05-13, https://platform.openai.com/docs/models/gpt-4o

Table 2: Controlled experiment participant profiles, hiding participant index and randomizing the order for anonymity

Education	Job Role
Statistics PhD	Machine Learning Researcher
Computer Science Master	Research Engineer
Applied Mathematics PhD	Oceanography Researcher
Computer Science Master	Data Engineer
Physics PhD	Quantum Researcher
Civil Engineering PhD	Earth Science Researcher
Computer Science Bachelor	Application Engineer

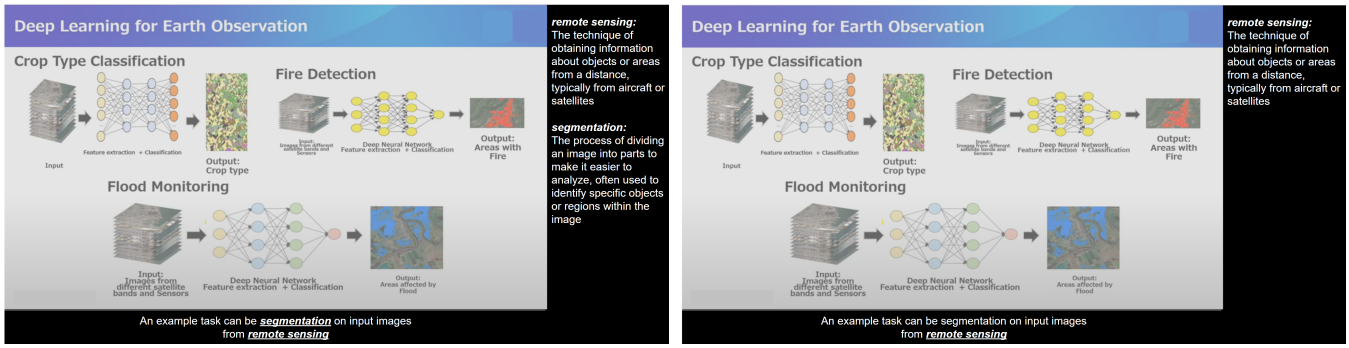
5.1.2 Experimental Conditions. We created three experimental conditions using the recorded and transcribed presentations:

- **General-purpose Jargon Support (Condition 1):** Presentation recordings were processed by ParseJargon without the personalized filter, identifying and defining all potentially unfamiliar terms not tailored to their specific background. Identified jargon terms were displayed in a glossary sidebar next to the video. (Figure 4a)
- **Personalized Jargon Support (Condition 2):** Using a short user profile provided by participants (one-sentence summary describing their education, job role, and domain expertise), presentation recordings were processed by the complete ParseJargon system with personalized filter. (Figure 4b)
- **Baseline (Condition 0)**: The original presentation recordings with no jargon support, standard transcripts were displayed to maintain consistency with Conditions 1 and 2, without highlighting any jargon.

5.1.3 Procedure. We employed a within-subject design where each participant viewed all presentations from the other six participants. To minimize order effects, we used a counter-balanced watching schedule. Because we had three conditions and seven participants, a perfect Latin square was impossible, but we rotated the three conditions so the overall counts were evenly distributed across first, middle, and last positions. This resulted in 14 unique viewing experiences per condition and a total of 42 participant-presentation pairs. The complete viewing order is given in Appendix Table 10.

Participants watched the recordings individually in two separate sessions (approximately 45 minutes each), viewing three presentations per session (one for each condition). Sessions were spaced at least one day apart to minimize fatigue effects. The order of presentations and conditions was randomized for each participant to mitigate potential learning effects.

Participants watched the recordings on their laptops without pausing or navigating through the videos to replicate live meeting conditions. However, they were allowed to freely use external resources (e.g., web searches, LLM queries) as needed, mirroring potential real meeting behaviors. The researcher remained available throughout the sessions, observing participants' screens to verify procedural adherence. At the end of the second session, participants were given the option to share feedback through a brief interview. In this interview, they were encouraged to describe their overall



(a) General-purpose Jargon (Condition 1) **(b) Personalized Jargon (Condition 2)**

Figure 4: General-purpose Jargon vs Personalized Jargon: Assume the audience has a background in computer vision but not earth science, they would already know the term *segmentation*, so only *remote sensing* would be displayed in the sidebar as an unfamiliar term

experience with the jargon sidebar, including what they liked or disliked about it. Study sessions were recorded and transcribed. This experiment was approved by the company internal ethics and legal board, as well as the diary study and field study.

5.1.4 Evaluation Metrics. We evaluated the system's effectiveness through several complementary metrics, capturing both subjective and objective dimensions.

Self-reported measures. After watching each presentation, participants completed a short survey assessing their subjective experiences through 5-point Likert-style scale (1 = least, 5 = most) questions:

- *Comprehension confidence*: "How confident do you feel in your understanding of the presentation?"
- *Engagement*: "How engaged were you while following the presentation?"
- *Perceived value*: "How valuable do you think the presented work is?"
- *Glossary helpfulness* (Condition 1 & 2 only): "How helpful were the term explanations provided in the glossary sidebar?"

Comprehension assessment. Participants wrote one-sentence takeaways and one-sentence questions for each presentation. These were anonymously evaluated by the original presenters for *clarity*, *relevance*, and *depth* on a 5-point Likert-style scale (1 = least, 5 = most). Unlike quizzes used in prior work [41], this approach emphasizes speakers' evaluations of listeners' comprehension quality, aligning more closely with real-world workplace communication needs [23, 69].

Glossary helpfulness rate. To explicitly evaluate the effectiveness of personalization, after Conditions 1 and 2, participants reviewed the entire glossary (all jargon identified), rating each term as either *helpful* (useful to some extent) or *not helpful* (unnecessary and potentially burdensome). We computed the *helpfulness rate* as the proportion of glossary terms rated helpful, which can be understood as the "precision" of jargon identification.

5.1.5 Analysis. We will report the statistical comparisons on these Likert-scale metrics between conditions (general vs. baseline, personalized vs. baseline, personalized vs. general) using Wilcoxon signed-rank test. We chose this non-parametric approach given the ordinal nature of Likert-scale data and the relatively small participant sample size [47]. Effect sizes (Cohen's d) were calculated. Holm–Bonferroni corrections were applied to control for Type I errors due to multiple comparisons [4], and we will report significance based the corrected p-values.

5.2 Results

Our controlled experiment results support both hypotheses, showing that ParseJargon (personalized version) significantly improved participants' comprehension, engagement, and perceived value of the present work, and that personalization significantly enhanced jargon identification precision and user satisfaction. Table 3 and 4 show the average rating of self-reported measures and comprehension assessment respectively for each condition. The complete test statistics can be found in Appendix Table 11.

*5.2.1 Personalized jargon support improved self-reported comprehension, engagement, and perceived value, but the general version does not. (**H1**).* As shown in Table 3, both the general and personalized conditions significantly increased comprehension compared to the baseline, with the personalized system showing greater improvement. However, only the personalized condition improved participants' engagement, whereas the general version decreased engagement from baseline. Interviews revealed that participants felt overwhelmed by the excessive number of definitions (6 participants), *"too many term definitions with very short reading time"*, as described by P1. Some participants even described this as *"annoying"* (P4) or even *"offensive... (because) the system treats me like I know nothing"* (P3).

While both experimental conditions improved participants' perceived value of others' presented work, only the personalized condition showed significance in rating improvement (Table 3). For instance, while P6 described the high-level objectives of a visualization application in their presentation, they listed various libraries

Table 3: Self-reported outcomes by condition (mean ± SD). Statistical significance (* vs. Baseline; † vs. General) is indicated by Holm–Bonferroni corrected $p<.05$ using Wilcoxon signed-rank tests (N=14 per condition).

Metric	Baseline	General	Personalized
Comprehension	3.07 (± 0.62)	3.79 (± 0.70)*	4.29 (± 0.61)*†
Engagement	3.93 (± 0.83)	3.64 (± 1.01)	4.29 (± 0.73)
Perceived Value	3.57 (± 0.65)	3.93 (± 0.62)	4.43 (± 0.51)*†
Usefulness	–	3.93 (± 0.83)	4.64 (± 0.50)†

and datasets as implementation details without explaining any of them. ParseJargon identified these libraries and datasets and provided a brief introduction for each, which helped to *"better understand the workflow (of the visualization application) under specific contexts"*, as mentioned by P7. This suggests that clearer comprehension may enable people to better recognize the significance of the others' work from an unfamiliar domain.

Table 4: Presenter-graded comprehension (aggregated across takeaways and questions; mean ± SD). Statistical significance (* vs. Baseline; † vs. General) is indicated by Holm–Bonferroni corrected $p<.05$ using Wilcoxon signed-rank tests (N=14 per condition).

Metric	Baseline	General	Personalized
Clarity	4.18 (± 0.80)	4.21 (± 0.73)	4.36 (± 0.63)
Relevance	3.64 (± 1.06)	4.43 (± 0.65)*	4.39 (± 0.45)*
Depth	3.25 (± 1.17)	3.68 (± 0.72)	4.04 (± 0.57)*

*5.2.2 Jargon support led to higher scores in comprehension assessment. (**H1**).* In addition to improvements in self-reported comprehension, participants provided significantly higher-quality takeaways and questions in terms of relevance and depth when using ParseJargon. However, clarity remained relatively unaffected, likely due to the fact that clarity depends more on participants' writing skills than on their understanding of presentations. (Table 4). The personalized system resulted in the highest depth and clarity ratings, while both general and personalized conditions significantly improved relevance over baseline. This suggests that the personalized glossary guided participants to more critical technical details by filtering jargon more precisely than the general glossary.

Table 5 illustrates how takeaways and questions varied by condition for a presentation on predicting ocean surface properties using Fourier Neural Operators (a research project similar to [62]). In the baseline condition, the takeaway and question are relatively vague and general. Conversely, the pair in the general system condition is much more relevant targeting key presentation topics, and the personalized condition pair is not only relevant but includes in-depth technical details.

*5.2.3 Personalized jargon support is significantly more helpful with more precise jargon identification. (**H2**).* The personalized glossary identified significantly fewer terms on average (9.71 vs 22.57, Table 6), increasing helpfulness rate dramatically from 47.03% to 77.51%. This precision significantly reduced unnecessary cognitive load, enabling greater engagement and comprehension by focusing only on the jargon terms that they truly need help (Table 3).

Table 7 provides an illustrative example from P5's presentation about deep learning applications in earth science, clearly highlighting the advantages of personalization. For the two audiences, one is from the real background of P6, a software engineer with some machine learning background who may potentially assist in implementing the technology but lacks earth science knowledge; and the other one is a hypothesized persona, a senior earth science researcher who provides domain-specific suggestions but has limited experience with AI technology, to better demonstrate the potential differences in personalization based on two diverse backgrounds. Given their diverse backgrounds, the two audiences require assistance with different subsets of jargon, which form part of the general audience's glossary. This demonstrates that the personalized filter successfully reduced unnecessary terms, allowing participants to focus on the information most relevant to them.

Although overall personalization was consistently beneficial, qualitative feedback identified one instance where personalization performed sub-optimally. In P1's presentation, many jargon terms were related to business operations. P2, while currently being an applied mathematics PhD, had launched a startup before, which makes them familiar with the business jargon. However, since this piece of information was not provided by the participant as their background input, these jargon terms remained unfiltered. This highlights that the quality and completeness of user profiles may heavily influence personalization precision.

We also found that, for the same presentation, listeners who were more distant in their knowledge compared to the the speaker (e.g., a computer engineering undergraduate vs. an earth science PhD) benefited more from personalized support. This is consistent with findings in prior research [5], suggesting that the "distance" between the background of a speaker and a listener would determine how strong the need for personalized jargon support is.

5.3 Design Refinement

Interview insights from the controlled experiment informed two additional design changes for the ParseJargon system. The first was a feedback mechanism allowing users to "like" or "dislike" glossary definitions for refining future jargon predictions. Six out of seven participants expressed interest in having more control over displayed terms to tune their personalized jargon assistant. The second was a persistent glossary sidebar, where jargon definitions remained accessible throughout the meeting, and an update to the display logic for showing the latest term (§4.2). Participants

Table 5: Example takeaways and questions from different glossary conditions, illustrating differences in relevance and depth (presentation topic: predicting ocean surface properties)

Condition	Example Takeaways
Baseline (0)	*Using AI to understand behaviour of ocean surfaces*
General (1)	*Using neural networks with Fourier Transforms to predict ocean depths is pretty reliable*
Personalized (2)	*FNO is versatile in terms of accepting the input and output of different sizes to predict the ocean surface properties.*
Condition	**Example Questions**
Baseline (0)	*What are the implications of your study to real life?*
General (1)	*How does extending the prediction accuracy time help the environmental scientist or decision makers in general?*
Personalized (2)	*Could you elaborate more on spectrum space and what's the benefit of using spectrum loss function over (traditional) cross entropy loss?*

Table 6: Average number of identified terms and helpfulness rate per participant-presentation.

Condition	# Total Terms	# Helpful Terms	Helpfulness Rate
General	22.57	10.29	47.03%
Personalized	9.71	7.64	77.51%

Table 7: Example jargon filtering from general to personalized glossary, showing tailored term selection based on different audience backgrounds (presentation topic: deep learning applications in earth science)

Glossary (for general audience)	For Machine Learning Engineer	For Earth Science Researcher
Benchmarking		
Foundation Models	X	X
Remote Sensing	X	
Pre-training		X
Satellite Data	X	
Self-supervised Learning		X

mentioned they sometimes lacked sufficient time to read definitions fully (3 participants) or wished to revisit definitions later (4 participants).

6 Field Study

To complement our controlled experiment and explore ParseJargon's performance in real-world settings, we conducted a lightweight field deployment within a real-time team meeting at the same technology company. Unlike the controlled experiment designed to validate backend effectiveness in carefully managed conditions, this deployment aimed to assess ParseJargon's practical usability, perceived cognitive workload, and usefulness in improving communication during actual workplace meetings.

6.1 Methodology

6.1.1 Participants and Setup. We deployed ParseJargon during a regularly scheduled weekly team meeting consisting of ten members: one director, six senior researchers, and three junior researchers. More demographic information can be found in Appendix Table 9. The meeting typically involved presentations by the junior researchers presenting updates on their projects. Though team members generally had shared knowledge, differences in specific research domains still introduced unfamiliar jargon.

6.1.2 Procedure. Participants installed ParseJargon's Chrome extension following brief instructions, after which we introduced its key features. This onboarding process took around 10 minutes. After onboarding, participants then joined the Zoom meeting via browser and logged into ParseJargon, providing a concise textual profile describing their educational and professional backgrounds for personalized jargon filtering. Participants then started their regular meeting with the system running automatically in real-time with the latest jargon term definition and the persistent glossary. The entire meeting lasted approximately 45 minutes, with each of the three junior researchers spending 10 minutes presenting and 5 minutes discussing after each presentation. Immediately after the

meeting, participants completed an online survey to evaluate the system.

6.1.3 Metrics. The post-meeting survey assessed ParseJargon across four dimensions:

- *Cognitive Workload:* Participants completed the standard NASA-TLX survey across six dimensions: mental, physical, and temporal demand, performance, effort, and frustration.
- *Usability:* Participants rated the system's ease of use, feature integration, and their willingness for frequent future use.
- *Effectiveness:* Participants first selected their most used typical strategy for managing jargon from the action list of our diary study (§3). They then rated how ParseJargon improved their comprehension, engagement, and perceived value of colleagues' presentations, compared to their previously selected typical strategy, using questions consistent with our controlled experiment (§5).
- *Qualitative Feedback:* Participants provided open-ended reflections on their overall experience, perceived strengths or limitations, and suggestions for future improvements.

The cognitive workload was measured using the 21-point NASA-TLX scale, while usability and perceived usefulness were assessed via 5-point Likert scale questions.

6.2 Results

Table 8: NASA-TLX Results from Field Study (out of 21, lower score indicates lower cognitive load)

Dimension	Score (Mean ± SD)
Mental Demand	4.8 ± 3.39
Physical Demand	2.3 ± 1.83
Temporal Demand	5.2 ± 4.54
Frustration Level	5.3 ± 3.65
Effort	5.1 ± 4.25

6.2.1 Usability and Cognitive Workload. Participants rated ParseJargon's usability positively, with average scores of ease-of-use at 4.3 (SD = 0.82), feature integration at 4.5 (SD = 0.53), and willingness for frequent future use at 3.9 (SD = 1.1), on a 5-point scale. NASA-TLX results further support the system's practical usability, indicating low cognitive load across all dimensions (Table 8).

These scores demonstrate that ParseJargon offers strong usability and introduces minimal cognitive workload, preserving participants' natural meeting engagement. Qualitative feedback further emphasized ease-of-use and seamless integration by comments such as: *"It was a smooth interface and easy integration with Zoom-like applications."* (P4) and *"What I liked most was that it was automatic and non-intrusive."* (P10)

6.2.2 Perceived Effectiveness. Participants reported overall positive impacts from using ParseJargon on comprehension (mean = 3.9, SD = 1.37), engagement (mean = 3.6, SD = 1.35), and perceived value of colleagues' work (mean = 4.0, SD = 1.05), consistent with findings from the controlled experiment (§5.2). Compared to their original strategies (e.g., passive waiting or searching on the internet), participants valued the immediate jargon explanations provided by the system. For instance, participants stated: *"It saved me the hassle of trying to Google stuff as the meeting went on."* (P3) and *"I like that it keeps me engaged with the speaker… it made me want to put more effort into understanding jargon in presentations."* (P7)

However, two participants reported low effectiveness ratings due to inaccuracies in jargon identification. One participant noted: *"It was not very useful for me and provided no real new information."* Another participant, while acknowledging the system design, expressed that *"The correct level of identification is not good. If it worked properly, I think it would be tremendously helpful."*

We attribute these accuracy issues to three primary factors: (1) inaccurate captions generated by the platform's speech-to-text service, potentially exacerbated by environmental noise or accent differences; (2) the lower-performing GPT-4o-mini model (compared to GPT-4o used in the controlled experiment) selected for real-time responsiveness; and (3) the low-jargon nature of within-team meetings - as highlighted by P4, *"The level of jargon in the meeting was not too high for me because I already knew most of the content."* These limitations suggest directions for further optimization of ParseJargon to provide better jargon support.

7 Discussion

7.1 Enhancing Personalized Jargon Support

While our current approach effectively demonstrates the value of personalized jargon identification, several promising directions remain to further enhance and deepen personalization.

7.1.1 Personalized Jargon Explanations. Our current implementation provides uniform textual explanations for identified jargon, focusing primarily on filtering out familiar terms based on users' profiles. However, personalization could extend beyond identification to the generation of explanations themselves. Future works might build systems that can dynamically adjust jargon explanations according to individual user expertise, and study how these factors affect personalization effectiveness. For instance, domain experts could receive concise, technical definitions, whereas users less familiar with the topic might benefit from detailed explanations accompanied by examples tailored specifically to their backgrounds as suggested in prior research [5]. Such in-depth personalization could significantly reduce unnecessary cognitive load and provide more engaging and meaningful interactions. Incorporating analogies that resonate with users' experiences can further deepen their comprehension and retention of complex terms, bridging interdisciplinary communication gaps more effectively.

7.1.2 Diverse Explanation Formats. Currently, ParseJargon exclusively delivers jargon explanations in plain textual format. While effective, reliance on a single modality can limit the accessibility and usefulness of jargon support, especially for concepts best explained visually or interactively. Recent studies in multimodal explanation systems highlight the potential of integrating diverse content formats to enhance learning and comprehension [41, 71]. Future iterations of our system could incorporate explanations with more diverse formats, such as figures, tables, or more complex interactive components to support visual learners and better illustrate

specific jargon types (e.g. showing a map for spatial or geographical terms). Additionally, integrating interactive and dynamic elements could empower users to explore explanations at their own pace and preference, improving engagement and accommodating diverse learning styles.

7.1.3 Customization for Organizational Jargon. Our field study (§6) highlighted another crucial personalization need: accurate interpretation of organization specific jargon. As P8 from the field study explicitly suggested: *"Make it explain corporate/company specific jargon would make the product very unique."*. For instance, without targeted adjustments, LLMs might misinterpret internal acronyms, for instance, "TSU" as "Texas Southern University" rather than the correct "Technology Strategy Unit." This issue underscores the importance of adapting jargon-support systems to unique organizational contexts. Future enhancements could incorporate company-specific glossaries or enable organizations to maintain their own custom terminology databases within ParseJargon. Further, leveraging methods such as fine-tuning large language models with organization-specific documentation could greatly improve jargon identification and explanation accuracy. Customizing models in this way could significantly reduce misinterpretations and better support internal communication, particularly within large, jargon-rich enterprises.

7.2 Extending Beyond Real-time Meetings

While ParseJargon was primarily designed for real-time meetings, our study findings and system design suggest it could effectively support users in asynchronous contexts. The reliance on pre-recorded presentations during our controlled experiment highlights ParseJargon's potential benefits in recorded materials, aligning with recent efforts that explore jargon assistance in video content [41]. For instance, P10 from our field study also suggested this idea by explicitly stating: *"(ParseJargon) could be tested on technical Youtube videos"*. Here we discuss two potential contexts where ParseJargon could significantly enhance asynchronous communication.

7.2.1 Supporting Recorded Educational Materials. ParseJargon could be effectively integrated into educational environments such as recorded lectures, webinars, or training sessions. In these settings, learners typically engage independently with specialized content and lack immediate opportunities to seek clarifications. Integrating ParseJargon into recorded materials as a caption layer or glossary overlay could provide learners with on-demand personalized jargon explanations, allowing them to remain focused on core content without frequent interruptions to search for definitions. This capability could be particularly beneficial in online courses, e.g. Massive Open Online Courses (MOOCs), where participants from diverse backgrounds often encounter different barriers due to specialized terminology.

7.2.2 Facilitating Presentation Rehearsal and Preparation. Another compelling application involves adapting ParseJargon as a rehearsal and preparation tool for presenters aiming to communicate complex ideas clearly to diverse audiences. Presenters frequently struggle to balance technical depth with accessibility, especially in interdisciplinary or public-facing scenarios. ParseJargon could enable speakers to anticipate which terms may require preemptive clarification based on hypothetical or known audience profiles. By identifying terms that likely lead to comprehension challenges, speakers could proactively refine their content for clarity and accessibility. Such functionality would be valuable for both live presentations and recorded content, enhancing communication effectiveness and engagement across varied audiences.

7.3 Toward a Long-term Personalized Assistant

Currently, our system integrates user feedback through basic interactions such as liking or disliking jargon explanations, providing only limited adaptation via prompt engineering within single meeting sessions. However, participants expressed interest in more sophisticated control mechanisms to refine jargon identification and explanation accuracy. For example, allowing users to explicitly set preferences beyond simple textual profiles for definition depth, detail level, or explanatory style could substantially improve personalization quality. Future iterations could incorporate interactive sliders or preference toggles, enabling users to dynamically adjust the granularity and style of jargon explanations and fine-tune the system according to individual objectives.

Additionally, while our existing persistent sidebar glossary supports term revisiting, its interactive capabilities remain minimal. Enhanced systems could enable users to organize, annotate, search, or export terms from personalized glossaries for deeper engagement or integration into personal learning materials. Actively managing jargon terms both during and after meetings would help users significantly improve learning and retention, transforming the glossary into an evolving knowledge resource.

Ultimately, such enhancements could position ParseJargon as a long-term personal jargon assistant that continuously learns and adapts to individual user preferences. Such an assistant would maintain persistent user profiles incorporating historical interactions, accumulated jargon familiarity, and preferred explanation styles. With these advanced capabilities, users could access personalized jargon support seamlessly across diverse contexts (e.g., meetings, technical videos, seminars), consistently receiving tailored assistance that aligns with their evolving knowledge needs. This future direction also reflects P4's feedback from our field study to extend ParseJargon into *"a personal jargon assistant that is installed on your mobile device"*, underscoring the potential to offer personalized ubiquitous support beyond individual meetings.

8 Limitations

While our controlled and field studies provided complementary insights into ParseJargon's backend effectiveness and real-time usability, several limitations remain. Both studies involved relatively small-scale samples from a single technology company, resulting in demographic homogeneity that may limit the generalizability of our findings. Additionally, our controlled experiment relied on pre-recorded videos to isolate backend performance, which, though effective for initial evaluation, did not fully capture the complexities and unpredictability of live meetings, particularly dynamic interpersonal interactions. Future research should pursue larger-scale and/or longitudinal studies across diverse organizational contexts,

as well as another more comprehensive real-time controlled experiment to better validate ParseJargon's capabilities.

Another limitation involves our personalization approach, which currently depends on concise user-provided textual profiles. As highlighted in the controlled experiment (§5.2), some participants may have informal or unreported background knowledge that their profiles did not reflect, limiting personalization accuracy. Future systems could address this by integrating richer user profiling methods, such as capturing usage history, comprehensive interactive feedback, or automated background inference techniques, thereby enhancing the precision and effectiveness of personalized jargon identification.

Finally, our field study highlighted practical constraints due to reliance on external speech-to-text services provided by video conferencing platforms, which introduced transcription errors, especially for diverse accents or noisy environments (§6.2). Additionally, deploying a less powerful LLM variant (GPT-4o-mini) to maintain real-time responsiveness further compromised jargon identification accuracy. To overcome these limitations, future work might explore developing dedicated meeting platforms or custom speech-to-text engines optimized specifically for jargon content, coupled with more powerful LLMs suitable for low-latency and high-accuracy jargon recognition (e.g., recently introduced GPT-4o-Realtime[4]). These technical advancements could significantly enhance ParseJargon's robustness and effectiveness in real-time meetings.

9 Conclusion

In this paper, we introduced ParseJargon, an LLM-powered real-time system designed to overcome interdisciplinary communication barriers caused by jargon. ParseJargon significantly enhances comprehension, engagement, and appreciation of colleagues' contributions during online meetings. Our controlled experiment demonstrated the clear benefits of personalized jargon support over general-purpose assistance, while a subsequent field study confirmed the system's usability and practical value in authentic workplace scenarios. Looking forward, ParseJargon presents promising avenues for broader applications, including educational content integration, support for presentation preparation, and advanced personalization techniques. Ultimately, these advancements position ParseJargon as a comprehensive, personalized assistant that facilitates lifelong learning and effective interdisciplinary collaboration.

References

[1] Takeshi Abekawa and Akiko Aizawa. 2016. SideNoter: Scholarly Paper Browsing System based on PDF Restructuring and Text Annotation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Hideo Watanabe (Ed.). The COLING 2016 Organizing Committee, Osaka, Japan, 136–140. https://aclanthology.org/C16-2029

[2] Eytan Adar, Carolyn Gearig, Ayshwarya Balasubramanian, and Jessica Hullman. 2017. PersaLog: Personalization of News Article Content. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 3188–3200. https://doi.org/10.1145/3025453.3025631

[3] Pouya Aghahoseini, Millan David, and Andrea Bunt. 2024. Investigating the Role of Real-Time Chat Summaries in Supporting Live Streamers. In *Proceedings of the 50th Graphics Interface Conference (GI '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3670947.3670980

[4] M. Aickin and H. Gensler. 1996. Adjusting for multiple testing when reporting research results: the Bonferroni vs Holm methods. *American Journal of Public Health* 86, 5 (May 1996), 726. https://doi.org/10.2105/ajph.86.5.726

[5] Tal August, Kyle Lo, Noah A. Smith, and Katharina Reinecke. 2024. Know Your Audience: The benefits and pitfalls of generating plain language summaries beyond the "general" audience. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 14, 26 pages. https://doi.org/10.1145/3613904.3642289

[6] Tal August, Katharina Reinecke, and Noah A. Smith. 2022. Generating Scientific Definitions with Controllable Complexity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 8298–8317. https://doi.org/10.18653/v1/2022.acl-long.569

[7] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A. Hearst, Andrew Head, and Kyle Lo. 2023. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *ACM Trans. Comput.-Hum. Interact.* 30, 5 (Sept. 2023), 74:1–74:38. https://doi.org/10.1145/3589955

[8] Azure. 2024. Azure AI Speech. https://azure.microsoft.com/en-us/products/ai-services/ai-speech/

[9] Brian P. Bailey and Joseph A. Konstan. 2006. On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior* 22, 4 (July 2006), 685–708. https://doi.org/10.1016/j.chb.2005.12.009

[10] Calvin Bao, Yow-Ting Shiue, Marine Carpuat, and Joel Chan. 2025. Words as Bridges: Exploring Computational Support for Cross-Disciplinary Translation Work. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1598–1623. https://doi.org/10.1145/3708359.3712110

[11] Zachariah C. Brown, Eric M. Anicich, and Adam D. Galinsky. 2020. Compensatory conspicuous communication: Low status increases jargon use. *Organizational Behavior and Human Decision Processes* 161 (Nov. 2020), 274–290. https://doi.org/10.1016/j.obhdp.2020.07.001

[12] Marc Brysbaert. 2019. How many words do we read per minute? A review and meta-analysis of reading rate. *Journal of Memory and Language* 109 (2019), 104047. https://doi.org/10.1016/j.jml.2019.104047

[13] Olivia M. Bullock, Daniel Colón Amill, Hillary C. Shulman, and Graham N. Dixon. 2019. Jargon as a barrier to effective science communication: Evidence from metacognition. *Public Understanding of Science* 28, 7 (Oct. 2019), 845–853. https://doi.org/10.1177/0963662519865687 Publisher: SAGE Publications Ltd.

[14] Francisco Maria Calisto, João Maria Abrantes, Carlos Santiago, Nuno J. Nunes, and Jacinto C. Nascimento. 2025. Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels. *International Journal of Human-Computer Studies* 197 (2025), 103444. https://doi.org/10.1016/j.ijhcs.2025.103444

[15] Senthil Chandrasegaran, Chris Bryan, Hidekazu Shidara, Tung-Yen Chuang, and Kwan-Liu Ma. 2019. TalkTraces: Real-Time Capture and Visualization of Verbal Content in Meetings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300807

[16] Xinyue Chen, Shuo Li, Shipeng Liu, Robin Fowler, and Xu Wang. 2023. MeetScript: Designing Transcript-based Interactions to Support Active Participation in Group Video Meetings. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2 (Oct. 2023), 347:1–347:32. https://doi.org/10.1145/3610196

[17] Bernard C. K. Choi and Anita W. P. Pak. 2007. Multidisciplinarity, interdisciplinarity, and transdisciplinarity in health research, services, education and policy: 2. Promotors, barriers, and strategies of enhancement. *Clinical and Investigative Medicine. Medecine Clinique Et Experimentale* 30, 6 (2007), E224–232. https://doi.org/10.25011/cim.v30i6.2950

[18] Kristy L. Daniel, Myra McConnell, Anita Schuchardt, and Melanie E. Peffer. 2022. Challenges facing interdisciplinary researchers: Findings from a professional development workshop. *PLOS ONE* 17, 4 (April 2022), e0267234. https://doi.org/10.1371/journal.pone.0267234 Publisher: Public Library of Science.

[19] Hyo Jin Do, Ha-Kyung Kong, Jaewook Lee, and Brian P. Bailey. 2022. How Should the Agent Communicate to the Group? Communication Strategies of a Conversational Agent in Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 387 (Nov. 2022), 23 pages. https://doi.org/10.1145/3555112

[20] Hyo Jin Do, Ha-Kyung Kong, Pooja Tetali, Karrie Karahalios, and Brian P. Bailey. 2023. Inform, Explain, or Control: Techniques to Adjust End-User Performance Expectations for a Conversational Agent Facilitating Group Chat Discussions. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 343 (Oct. 2023), 26 pages. https://doi.org/10.1145/3610192

[21] Wei Dong and Wai-Tat Fu. 2012. One piece at a time: why video-based communication is better for negotiation and conflict resolution. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (Seattle, Washington, USA) *(CSCW '12)*. Association for Computing Machinery, New York, NY, USA,

[4] https://platform.openai.com/docs/models/gpt-4o-realtime-preview

167–176. https://doi.org/10.1145/2145204.2145232

[22] Duolingo and LinkedIn. 2024. The State of Workplace Jargon Report. https://blog.duolingo.com/state-of-jargon-report/

[23] Martin Eppler. 2007. Knowledge Communication Problems between Experts and Decision Makers: an Overview and Classification. *http://www.alexandria.unisg.ch/Publikationen/54814* 5 (01 2007).

[24] John Fiset, Devasheesh P. Bhave, and Nilotpal Jha. 2024. The Effects of Language-Related Misunderstanding at Work. *Journal of Management* 50, 1 (Jan. 2024), 347–379. https://doi.org/10.1177/01492063231181651 Publisher: SAGE Publications Inc.

[25] Raymond Fok, Joseph Chee Chang, Tal August, Amy X. Zhang, and Daniel S. Weld. 2024. Qlarify: Recursively Expandable Abstracts for Dynamic Information Retrieval over Scientific Papers. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology* (Pittsburgh, PA, USA) *(UIST '24)*. Association for Computing Machinery, New York, NY, USA, Article 145, 21 pages. https://doi.org/10.1145/3654777.3676397

[26] Dee Gardner and Mark Davies. 2014. A New Academic Vocabulary List. *Applied Linguistics* 35, 3 (July 2014), 305–327. https://doi.org/10.1093/applin/amt015

[27] Sian Gooding and Manuel Tragut. 2022. One Size Does Not Fit All: The Case for Personalised Word Complexity Models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 353–365. https://doi.org/10.18653/v1/2022.findings-naacl.27

[28] Jens Emil Sloth Grønbæk, Marcel Borowski, Eve Hoggan, Wendy E. Mackay, Michel Beaudouin-Lafon, and Clemens Nylandsted Klokmose. 2023. Mirrorverse: Live Tailoring of Video Conferencing Interfaces. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. https://doi.org/10.1145/3586183.3606767

[29] Yue Guo, Joseph Chee Chang, Maria Antoniak, Erin Bransom, Trevor Cohen, Lucy Wang, and Tal August. 2024. Personalized Jargon Identification for Enhanced Interdisciplinary Communication. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 4535–4550. https://doi.org/10.18653/v1/2024.naacl-long.255

[30] Yue Guo, Wei Qiu, Yizhong Wang, and Trevor Cohen. 2021. Automated Lay Language Summarization of Biomedical Scientific Reviews. *Proceedings of the AAAI Conference on Artificial Intelligence* 35, 1 (May 2021), 160–168. https://doi.org/10.1609/aaai.v35i1.16089 Number: 1.

[31] 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 (CHI '21)*. Association for Computing Machinery, New York, NY, USA, 1–18. https://doi.org/10.1145/3411764.3445648

[32] Axios HQ. 2023. The 2023 state of essential workplace communications. https://www.axioshq.com/research/2023-state-of-workplace-communications

[33] Jie Huang, Hanyin Shao, Kevin Chen-Chuan Chang, Jinjun Xiong, and Wen-mei Hwu. 2022. Understanding Jargon: Combining Extraction and Generation for Definition Modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 3994–4004. https://doi.org/10.18653/v1/2022.emnlp-main.266

[34] Leo Jeffres, David Atkin, and Hanlong Fu. 2011. Knowledge and the Knowledge Gap: Time to Reconceptualize the "Content". *Open Communication Journal* 5 (12 2011). https://doi.org/10.2174/1874916X01105010030

[35] Hita Kambhamettu, Danaë Metaxa, Kevin Johnson, and Andrew Head. 2024. Explainable Notes: Examining How to Unlock Meaning in Medical Notes with Interactivity and Artificial Intelligence. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 449, 19 pages. https://doi.org/10.1145/3613904.3642573

[36] Panayu Keelawat. 2023. NBGuru: Generating Explorable Data Science Flowcharts to Facilitate Asynchronous Communication in Interdisciplinary Data Science Teams. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 6–11. https://doi.org/10.1145/3584931.3607020

[37] John Lee and Chak Yan Yeung. 2018. Personalizing Lexical Simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, Emily M. Bender, Leon Derczynski, and Pierre Isabelle (Eds.). Association for Computational Linguistics, Santa Fe, New Mexico, USA, 224–232. https://aclanthology.org/C18-1019

[38] Daniel Li, Thomas Chen, Albert Tung, and Lydia B Chilton. 2021. Hierarchical Summarization for Longform Spoken Dialog. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 582–597. https://doi.org/10.1145/3472749.3474771

[39] Jian Liao, Adnan Karim, Shivesh Singh Jadon, Rubaiat Habib Kazi, and Ryo Suzuki. 2022. RealityTalk: Real-Time Speech-Driven Augmented Presentation for AR Live Storytelling. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology* (Bend, OR, USA) *(UIST '22)*. Association for Computing Machinery, New York, NY, USA, Article 17, 12 pages. https://doi.org/10.1145/3526113.3545702

[40] Xingyu 'Bruce' Liu, Vladimir Kirilyuk, Xiuxiu Yuan, Peggy Chi, Alex Olwal, Xiang 'Anthony' Chen, and Ruofei Du. 2023. Experiencing Visual Captions: Augmented Communication with Real-time Visuals using Large Language Models. In *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1–4. https://doi.org/10.1145/3586182.3615978

[41] Yuhan Liu, Aadit Shah, Jordan Ackerman, and Manaswi Saha. 2025. Exploring the Design Space of Real-time LLM Knowledge Support Systems: A Case Study of Jargon Explanations. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 633, 20 pages. https://doi.org/10.1145/3706598.3714262

[42] Kyle Lo, Joseph Chee Chang, Andrew Head, Jonathan Bragg, Amy X. Zhang, Cassidy Trier, Chloe Anastasiades, Tal August, Russell Authur, Danielle Bragg, Erin Bransom, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Yen-Sung Chen, Evie Yu-Yen Cheng, Yvonne Chou, Doug Downey, Rob Evans, Raymond Fok, Fangzhou Hu, Regan Huff, Dongyeop Kang, Tae Soo Kim, Rodney Kinney, Aniket Kittur, Hyeonsu B. Kang, Egor Klevak, Bailey Kuehl, Michael J. Langan, Matt Latzke, Jaron Lochner, Kelsey MacMillan, Eric Marsh, Tyler Murray, Aakanksha Naik, Ngoc-Uyen Nguyen, Srishti Palani, Soya Park, Caroline Paulic, Napol Rachatasumrit, Smita Rao, Paul Sayre, Zejiang Shen, Pao Siangliulue, Luca Soldaini, Huy Tran, Madeleine van Zuylen, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Marti A. Hearst, and Daniel S. Weld. 2024. The Semantic Reader Project. *Commun. ACM* 67, 10 (Sept. 2024), 50–61. https://doi.org/10.1145/3659096

[43] Li Lucy, Jesse Dodge, David Bamman, and Katherine Keith. 2023. Words as Gatekeepers: Measuring Discipline-specific Terms and Meanings in Scholarly Publications. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 6929–6947. https://doi.org/10.18653/v1/2023.findings-acl.433

[44] Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable Sentence Simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association, Marseille, France, 4689–4698. https://aclanthology.org/2020.lrec-1.577

[45] Moira McGregor and John C. Tang. 2017. More to Meetings: Challenges in Using Speech-Based Technology to Support Meetings. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 2208–2220. https://doi.org/10.1145/2998181.2998335

[46] Matthew K. Miller and Regan L. Mandryk. 2021. Meeting with Media: Comparing Synchronous Media Sharing and Icebreaker Questions in Initial Interactions via Video Chat. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 374 (Oct. 2021), 26 pages. https://doi.org/10.1145/3479518

[47] Constantin Mircioiu and Jeffrey Atkinson. 2017. A Comparison of Parametric and Non-Parametric Methods Applied to a Likert Scale. *Pharmacy: Journal of Pharmacy, Education and Practice* 5, 2 (May 2017), 26. https://doi.org/10.3390/pharmacy5020026

[48] Sonia K. Murthy, Daniel King, Tom Hope, Daniel S. Weld, and Doug Downey. 2021. Towards Personalized Descriptions of Scientific Concepts. https://api.semanticscholar.org/CorpusID:247410502

[49] Kevin K. Nam and Mark S. Ackerman. 2007. Arkose: reusing informal information from online discussions. In *Proceedings of the 2007 ACM International Conference on Supporting Group Work* (Sanibel Island, Florida, USA) *(GROUP '07)*. Association for Computing Machinery, New York, NY, USA, 137–146. https://doi.org/10.1145/1316624.1316644

[50] Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, Dina Demner-Fushman, Kevin Bretonnel Cohen, Sophia Ananiadou, and Junichi Tsujii (Eds.). Association for Computational Linguistics, Florence, Italy, 319–327. https://doi.org/10.18653/v1/W19-5034

[51] Raymond Nickerson. 1999. How We Know—and Sometimes Misjudge—What Others Know: Imputing One's Own Knowledge to Others. *Psychological Bulletin* 125 (11 1999), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

[52] Sachita Nishal, Eric Lee, and Nicholas Diakopoulos. 2024. De-jargonizing Science for Journalists with GPT-4: A Pilot Study. arXiv:2410.12069 [cs.CL] https://arxiv.org/abs/2410.12069

[53] Changhoon Oh, Jinhan Choi, Sungwoo Lee, SoHyun Park, Daeryong Kim, Jungwoo Song, Dongwhan Kim, Joonhwan Lee, and Bongwon Suh. 2020. Understanding User Perception of Automated News Generation System. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376811

[54] Otter. 2024. Otter Voice Meeting Notes. https://otter.ai/

[55] Ngueviuta Patoko and Rashad Yazdanifard. 2014. The Impact of Using Many Jargon Words, while Communicating with the Organization Employees. *American Journal of Industrial and Business Management* 4, 10 (Oct. 2014), 567–572. https://doi.org/10.4236/ajibm.2014.410061 Number: 10 Publisher: Scientific Research Publishing.

[56] Picovoice. 2023. Speech-to-Text Benchmark. https://github.com/Picovoice/speech-to-text-benchmark

[57] Amir Pouran Ben Veyseh, Franck Dernoncourt, Walter Chang, and Thien Huu Nguyen. 2021. MadDog: A Web-based System for Acronym Identification and Disambiguation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, Dimitra Gkatzia and Djamé Seddah (Eds.). Association for Computational Linguistics, Online, 160–167. https://doi.org/10.18653/v1/2021.eacl-demos.20

[58] S. Monisha Pulimood, Diane C. Bates, and Kim Pearson. 2024. Immersing Undergraduates in Interdisciplinary Course Collaborations. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 2 (SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 1782–1783. https://doi.org/10.1145/3626253.3635598

[59] Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, Sandra Williams, Advaith Siddharthan, and Ani Nenkova (Eds.). Association for Computational Linguistics, Sofia, Bulgaria, 69–77. https://aclanthology.org/W13-2908

[60] Hillary C. Shulman, Graham N. Dixon, Olivia M. Bullock, and Daniel Colón Amill. 2020. The Effects of Jargon on Processing Fluency, Self-Perceptions, and Scientific Engagement. *Journal of Language and Social Psychology* 39, 5-6 (Oct. 2020), 579–597. https://doi.org/10.1177/0261927X20902177 Publisher: SAGE Publications Inc.

[61] Seoyun Son, Junyoug Choi, Sunjae Lee, Jean Y Song, and Insik Shin. 2023. It is Okay to be Distracted: How Real-time Transcriptions Facilitate Online Meeting with Distraction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 64, 19 pages. https://doi.org/10.1145/3544548.3580742

[62] Yixuan Sun, Ololade Sowunmi, Romain Egele, Sri Hari Krishna Narayanan, Luke Van Roekel, and Prasanna Balaprakash. 2024. Streamlining Ocean Dynamics Modeling with Fourier Neural Operators: A Multiobjective Hyperparameter and Architecture Optimization Approach. *Mathematics* 12, 10 (2024). https://doi.org/10.3390/math12101483

[63] Gabriel Szulanski. 2000. The Process of Knowledge Transfer: A Diachronic Analysis of Stickiness. *Organizational Behavior and Human Decision Processes* 82, 1 (2000), 9–27. https://doi.org/10.1006/obhd.2000.2884

[64] Kumiko Tanaka-Ishii and Hiroshi Terada. 2011. Word familiarity and frequency. *Studia Linguistica* 65, 1 (2011), 96–116. https://doi.org/10.1111/j.1467-9582.2010.01176.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9582.2010.01176.x

[65] Sunny Tian, Amy X. Zhang, and David Karger. 2021. A System for Interleaving Discussion and Summarization in Online Collaboration. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3 (Jan. 2021), 241:1–241:27. https://doi.org/10.1145/3432940

[66] Hoang Van, David Kauchak, and Gondy Leroy. 2020. AutoMeTS: The Autocomplete for Medical Text Simplification. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 1424–1434. https://doi.org/10.18653/v1/2020.coling-main.122

[67] Amir Veyseh, Franck Dernoncourt, Dejing Dou, and Thien Nguyen. 2020. A Joint Model for Definition Extraction with Syntactic Connection and Semantic Consistency. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 05 (Apr. 2020), 9098–9105. https://doi.org/10.1609/aaai.v34i05.6444

[68] Ruotong Wang, Lin Qiu, Justin Cranshaw, and Amy X. Zhang. 2024. Meeting Bridges: Designing Information Artifacts that Bridge from Synchronous Meetings to Asynchronous Collaboration. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1 (April 2024), 35:1–35:29. https://doi.org/10.1145/3637312

[69] Amanda Weirup and Phylicia Taylor. 2024. What Do You Mean? Developing Jargon Literacy for the Workplace. *Management Teaching Review* (July 2024), 23792981241266465. https://doi.org/10.1177/23792981241266465 Publisher: SAGE Publications Inc.

[70] Olga A. Wudarczyk, Murat Kirtay, Anna K. Kuhlen, Rasha Abdel Rahman, John-Dylan Haynes, Verena V. Hafner, and Doris Pischedda. 2021. Bringing Together Robotics, Neuroscience, and Psychology: Lessons Learned From an Interdisciplinary Project. *Frontiers in Human Neuroscience* 15 (March 2021). https://doi.org/10.3389/fnhum.2021.630789 Publisher: Frontiers.

[71] Haijun Xia, Tony Wang, Aditya Gunturu, Peiling Jiang, William Duan, and Xiaoshuo Yao. 2023. CrossTalk: Intelligent Substrates for Language-Oriented Interaction in Video-Based Communication and Collaboration. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 60, 16 pages. https://doi.org/10.1145/3586183.3606773

[72] Amy X. Zhang and Justin Cranshaw. 2018. Making Sense of Group Chat through Collaborative Tagging and Summarization. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 196:1–196:27. https://doi.org/10.1145/3274465

A Prompts

For both GPT-4o and GPT-4o-mini, the temprature is set to 0.1 and maximum length is 1000.

A.1 Jargon Identification & Explanation

System Message Your job is to help an audience listen to speeches that might contain terms they are unfamiliar with. You will be given the transcript of the speech, one sentence after another. For each sentence, the format will be "Transcript: [sentence]". Your task is to first identify any of those terms that the audience might not fully understand, then provide a definition for each term with any necessary background knowledge in concise, simple plain language. Please skip any terms you believe are nonsense or partial-error caused by speech-to-text transcription mistakes. Your output should be in the format of a list of term-definition pairs. Return only valid JSON in the format [{"term": "definition"}, ...]. Do not include additional commentary or text outside the JSON. Please leave the list blank if you think all the terms in the input phrase are common words that don't need additional explanations. You don't need to output a term if it has already been identified in previous input phrases.

User Prompt Transcript: {*transcript*}, Previously define terms: {*defined_terms*}, User preference: {*preferences*}

A.2 Personalization

System Message A previous agent has generated a glossary of term-definition pairs from a transcript. Your job is to help the audience reduce the number of terms in the glossary. The audience's background is "{*background*}". The input glossary is provided in valid JSON format, where each item is structured as {"term": "definition"}. Please examine only the terms (the keys in the JSON) and determine which terms the audience is likely already familiar with based on their background. Then, remove these terms from the glossary. Return only valid JSON structured exactly as: {"understood_terms": ["term1", "term2", ...], "refined_glossary": [{"term": "definition"}, ...]}. Do not include any extra commentary or text.

User Prompt {*glossary*}

B Tables

This section includes three tables to show 1) the demographic information for diary study and field study participants, 2) the presentation viewing order for the controlled study, and 3) test statistics for all metrics in the controlled study.

Table 9: Participant demographics summary for diary study and field study

	Diary Study (N=16)	Field Study (N=10)
Gender		
Female	3 (18.75%)	0 (0%)
Male	13 (81.25%)	10 (100%)
Age		
18-24	3 (18.75%)	3 (30%)
25-34	7 (43.75%)	2 (20%)
35-44	3 (18.75%)	3 (30%)
45-54	0 (0%)	0 (0%)
55-64	3 (18.75%)	2 (20%)
Education		
Bachelor's	1 (6.25%)	0 (0%)
Graduate	15 (93.75%)	10 (100%)
Ethnicity		
Asian	10 (62.5%)	8 (80%)
White	5 (31.25%)	1 (10%)
Other/Mixed	1 (6.25%)	1 (10%)
English Proficiency		
Native/Proficient	12 (75%)	3 (30%)
Professional	4 (25%)	7 (70%)

Table 10: Viewing schedule for every participant, using a counter-balanced design. Entry format = *condition–presenter*. Conditions: 0 Baseline, 1 General, 2 Personalized.

Audience	Session 1			Session 2		
P1	0–P2	1–P3	2–P4	1–P5	2–P6	0–P7
P2	1–P3	2–P4	0–P5	2–P6	0–P7	1–P1
P3	2–P4	0–P5	1–P6	0–P7	1–P1	2–P2
P4	0–P5	1–P6	2–P7	1–P1	2–P2	0–P3
P5	1–P6	2–P7	0–P1	2–P2	0–P3	1–P4
P6	2–P7	0–P1	1–P2	0–P3	1–P4	2–P5
P7	0–P1	1–P2	2–P3	1–P4	2–P5	0–P6

Table 11: Test statistics for all metrics, including w (Wilcoxon signed-rank test), corrected p-value for w, and Cohen's d. Holm–Bonferroni method was used for post-hoc correction.

	General vs Baseline			Personalized vs Baseline			Personalized vs General		
Metric	w	p_w	d	w	p_w	d	w	p_w	d
Comprehension	56.0	0.0294	0.6682	86.5	0.0047	1.2455	24.5	0.0294	0.5316
Engagement	26.0	0.7395	-0.1988	31.5	0.2733	0.2938	25.5	0.0710	0.5942
Value	48.0	0.0658	0.4242	72.5	0.0073	1.1127	21.0	0.0196	0.7687
Usefulness[†]	–	–	–	–	–	–	50.5	0.0065	0.8654
Clarity	49.0	0.3215	0.0626	59.0	0.2248	0.2673	59.0	0.3232	0.1797
Relevance	90.0	0.0076	0.7751	91.0	0.0055	0.8832	40.0	0.5527	-0.0626
Depth	68.0	0.1551	0.3378	86.0	0.0099	0.7751	65.0	0.0647	0.4493

[†] Usefulness is only defined for General and Personalized conditions.