Meeting Retriever XX: Towards Trust-Calibrated Explanations in Cross-Meeting Retrieval

Takuya Nakata[1(✉)], Sinan Chen[1], and Masahide Nakamura[1,2]

[1] Center for Mathematical and Data Sciences, Kobe University, 1-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan
tnakata@bear.kobe-u.ac.jp, chensinan@gold.kobe-u.ac.jp, masa-n@cmds.kobe-u.ac.jp

[2] RIKEN AIP, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

Abstract. Meeting logs are increasingly archived as detailed audio recordings and transcripts, enabling retrieval across multiple meetings. Retrieval-Augmented Generation (RAG) can support access to such logs by generating natural-language summaries and answers; however, in cross-meeting retrieval scenarios, users must judge whether retrieved information can be trusted and used, rather than simply whether it is correct. Existing explainable or trustworthy RAG approaches often rely on a single explanatory mechanism, which is insufficient for supporting trust judgment across meetings with differing contexts and temporal distances. In this paper, we investigate how users make trust judgments in cross-meeting retrieval and how these judgments are influenced by presentation formats. We propose Cross-Explanation, an approach that combines natural-language explanations with quantitative indicators (ACROSS: Accuracy, Coverage, and Recency) as complementary cues to support users' judgment processes. We implemented a prototype system incorporating Cross-Explanation and conducted an exploratory user study comparing an explanation-only condition with a combined presentation condition. Qualitative analysis revealed that trust judgments depend less on perceived correctness and more on cues related to judgment possibility, such as comprehensibility, summary structure, and whether information appears suitable for decision-making. The results further show that presenting explanations alongside ACROSS indicators changes how users articulate and construct trust judgments by encouraging reference to multiple cues. These findings suggest that explainability in cross-meeting retrieval should be reconsidered as an interaction that supports users' trust judgment processes, clarifying part of the design space for cross-meeting retrieval systems.

Keywords: Cross-Meeting Retrieval · Trust Judgment · Explainable AI · RAG · Human-Centered AI

H. Degen and S. Ntoa (Eds.): HCII 2026, LNAI 16745, pp. 124–141, 2026.
https://doi.org/10.1007/978-3-032-30849-8_7

1 Introduction

Meetings play a central role in decision-making, consensus building, and knowledge sharing within organizations and research activities. Traditionally, the outcomes of meetings have been recorded in the form of minutes or notes, which are primarily intended for one-time reference or post hoc confirmation. As a result, long-term and cross-context reuse of meeting knowledge has remained limited. With the widespread adoption of online meetings and recording environments, detailed meeting audio and transcripts can now be easily archived. Consequently, organizations increasingly accumulate large collections of meeting logs over time. However, mechanisms for effectively utilizing these logs remain insufficient, and meeting records often become "available but unused knowledge."

Retrieval-Augmented Generation (RAG) has recently attracted attention as a means of searching across accumulated documents and logs and generating natural-language summaries or answers in response to user queries. Unlike traditional keyword-based retrieval, RAG can reference multiple information sources and provide context-aware responses. When applied to meeting logs, RAG offers the potential to help users efficiently grasp what was discussed and what decisions were made, particularly across multiple meetings [1,7].

At the same time, applying RAG to meeting logs introduces challenges that differ fundamentally from those of conventional document question answering. In cross-meeting retrieval scenarios, users are often required to judge whether retrieved information can be trusted and used. Users of meeting logs are typically participants in related projects and may possess partial memories or background knowledge. As a result, they do not evaluate outputs solely in terms of correctness, but instead consider whether the information aligns with their understanding, whether important points might be missing, and whether it is appropriate to use the information for current decision-making. These judgments become even more difficult when retrieval spans multiple meetings with different contexts and temporal distances [10,16].

Prior research has addressed trust and transparency in RAG from several directions, including presenting supporting evidence, generating natural-language explanations, and improving accuracy and transparency under the umbrella of trustworthy or explainable AI. However, many existing approaches implicitly assume that providing a single explanatory mechanism—such as an explanation text or evidence list—is sufficient for users to make trust judgments [15,16]. In cross-meeting retrieval, this assumption does not hold. While natural-language explanations can facilitate understanding, they do not directly support judgments about sufficiency, bias, or recency. As a result, users may encounter situations in which outputs appear plausible but remain difficult to trust.

To clarify this problem setting, Fig. 1 illustrates a conceptual view of trust judgment in cross-meeting retrieval. When retrieval spans multiple meetings with different contexts and temporal distances, users must interpret the retrieved results by relying on multiple cues, such as explanations and indicators, rather than a single source of justification. Importantly, the figure emphasizes that

the system does not determine trust; instead, it provides explanatory cues that support users' own trust judgment processes.

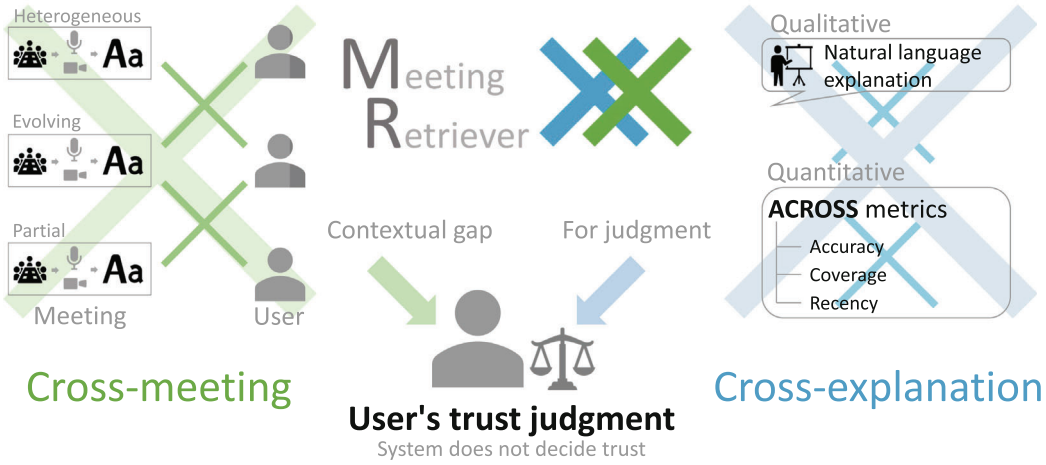

Fig. 1. Conceptual overview of trust judgment in cross-meeting retrieval. The system provides multiple judgment cues through Cross-Explanation, including natural-language explanations and quantitative indicators, while the final trust judgment remains with the user.

This study addresses this gap by focusing not on what should be explained, but on how different explanatory cues should be combined to support trust judgment. We propose Cross-Explanation, an approach that places natural-language explanations alongside quantitative indicators (ACROSS: Accuracy, Coverage, and Recency) to provide multiple complementary cues [2,11]. Rather than determining trust on behalf of users, Cross-Explanation is designed to support the cognitive process through which users assess how much retrieved results can be relied upon. To instantiate this approach in a realistic setting, we implemented a prototype cross-meeting retrieval system, referred to as *Meeting Retrieve XX*, which serves as the experimental platform in this study. We conducted an exploratory user study comparing an explanation-only condition with a combined presentation condition. Through qualitative analysis, we examined how users articulate trust judgments and how these articulations change depending on the presentation format.

The contributions of this paper are threefold. First, we empirically characterize the cues that users rely on when making trust judgments in cross-meeting retrieval. Second, we show how combining explanations with quantitative indicators changes the way trust judgments are formed and articulated. Third, we position explainability not as a mechanism for conveying correct answers, but as an interaction that supports users' judgment processes in realistic cross-meeting contexts.

2 Preliminaries

2.1 Meeting Logs and Retrieval-Augmented Generation

Meetings serve as a central venue for decision making, consensus building, and knowledge sharing in organizational and research activities. Traditionally, the outcomes of meetings have been recorded as minutes or notes. These records are primarily intended for one-time reference or post-hoc confirmation, and they have rarely supported long-term or cross-meeting knowledge utilization.

With the widespread adoption of online meetings and recording environments, it has become easy to store meeting audio and detailed utterance logs. As a result, organizations now accumulate large collections of meeting logs over time. Despite this increased availability, effective mechanisms for utilizing these logs remain limited. In many cases, meeting records exist as stored but unused knowledge within organizations [3,14].

In response to this situation, Retrieval-Augmented Generation (RAG) has attracted attention as an approach for searching and integrating accumulated documents or logs to generate natural-language summaries and answers. Unlike simple full-text search, RAG can reference multiple information sources in response to a user's query and generate outputs that reflect contextual relationships among them [4,18]. When applied to meeting logs, RAG offers a promising means of efficiently understanding what was discussed in past meetings and what kinds of judgments were made.

In particular, RAG enables users to refer to discussions and decisions that span multiple meetings. This cross-meeting capability provides access to longitudinal information that conventional meeting support tools do not readily offer [12,20].

2.2 Trust Judgment Challenges in Cross-Meeting Retrieval

Applying RAG to meeting logs involves inherent challenges that differ from those of document-based question answering. A key difficulty arises in cross-meeting retrieval, where information is retrieved across multiple meetings. In such situations, users are repeatedly required to make trust judgments about the system's outputs [12,20].

Users of meeting logs are often participants in the meetings themselves or members of related projects, and therefore possess partial memories and prior knowledge of the content. Consequently, they do not evaluate RAG outputs solely in terms of correctness. Instead, they judge how much the output can be trusted by considering whether it aligns with their memory, whether important discussion points are missing, and whether the information is appropriate for their current judgment [8,9].

In cross-meeting retrieval, this trust judgment becomes even more difficult because the retrieved content may originate from meetings with different contexts and temporal distances. As a result, users must assess not only topical relevance but also the adequacy and timeliness of the retrieved information [5,13].

Previous research has approached related issues from several directions, including evidence presentation in RAG, explainable AI techniques that provide reasons for why results are selected, and trustworthy AI efforts that aim to increase trust through accuracy and transparency. However, many of these approaches rely on a single explanatory mechanism, assuming that providing either explanations or evidence alone is sufficient for users to make judgments [4,20].

In cross-meeting retrieval, this assumption does not hold. While natural-language explanations help users understand retrieved results, they do not support judgments about whether other important discussions exist, whether the information is outdated, or whether the retrieval reflects only a limited subset of meetings. As a result, users may encounter outputs that are plausibly explained yet remain uncertain about whether they should rely on them. This limitation of single explanation approaches constitutes a fundamental challenge in cross-meeting retrieval [6,17,19].

3 Proposal Method

3.1 Goal and Approach

The goal of this study is to realize a Retrieval-Augmented Generation (RAG) experience for cross-meeting retrieval in which users can feel convinced and determine for themselves how much they should trust the retrieved results. To achieve this goal, we focus not on what should be explained, but on how different explanatory cues should be combined.

The key idea of this work is Cross-Explanation. Rather than relying on a single form of explanation, Cross-Explanation supports users' trust judgment by presenting multiple explanatory cues in a complementary manner. Specifically, the system places the following side by side:

- intuitive explanations expressed in natural language, and
- quantitative cues that characterize the properties of retrieval results (ACROSS).

By reading the explanation text and referring to these indicators together, users can determine where to place their trust in the retrieval results.

The proposed approach consists of three components:

- (A1) Natural-language explanation generation, which provides short textual explanations describing why particular retrieval results were selected;
- (A2) ACROSS metric calculation and visualization, which presents quantitative cues representing the consistency, coverage, and recency of the retrieval results; and
- (A3) Integrated presentation of explanations and indicators, which enables users to judge how much the results can be trusted through Cross-Explanation.

3.2 Architecture

In cross-meeting retrieval, simply retrieving relevant utterances is not sufficient. Users require information presentation that allows them to proceed through the stages of understanding, comparison, and judgment. Accordingly, the architecture of this study is designed not to maximize retrieval accuracy, but to separate and then re-integrate multiple cues required for trust judgment.

The system simultaneously provides the following elements for cross-meeting retrieval:

- natural-language explanations that help users understand the content of retrieval results,
- quantitative cues (ACROSS) that help users grasp the characteristics of the results, and
- a presentation structure that places these elements side by side on the same screen to enable comparison.

The overall architecture consists of three layers:

- Data processing layer: extraction of utterance-level units from meeting audio and logs (assumed as a prerequisite);
- Explanation generation layer: generation of explanatory cues through (A1) and (A2); and
- Presentation layer: realization of Cross-Explanation through (A3).

Through this architecture, the system makes it possible to address a human-centered question—not whether explanations are provided, but how trust judgments are formed—within the context of cross-meeting retrieval.

3.3 (A1) Natural-Language Explanation Generation

In cross-meeting retrieval, search results often span multiple meetings, making it difficult for users to immediately understand why particular results were returned. Without explanations, users are forced to read supporting text in detail or compare the results with their own memory, which imposes a high cognitive load.

In (A1), the system generates short textual explanations that summarize why the retrieval results were selected. These explanations do not assert correctness or guide users toward a particular judgment; instead, they function solely as a scaffold for understanding.

Given the search query and the selected set of utterances as input, the system abstracts common topics, intentions, or reasons for judgment and generates a brief explanation text, referred to as reason_text. The explanation concisely indicates which discussions from which meetings are related to the query, and why. Through (A1), users can enter a state in which they understand the retrieval results before reading them in detail. However, this alone is not sufficient to support trust judgment.

After the retrieval process, the system takes the selected chunk set as input and generates reason_text as an explanatory text for the retrieval results. The reason_text is stored as an independent field, separate from the generated answer text and the list of supporting chunks, and is used to supplement the relationship between the query and the retrieval results in natural language. The generation method is not fixed: the system supports both large language model (LLM)-based generation and lightweight embedding-based generation without LLM calls. In either case, the generated text summarizes the reasons why the retrieved results are associated with the query rather than simply concatenating chunk texts. The reason_text does not claim correctness or induce conclusions or decisions; it is treated purely as auxiliary information.

3.4 (A2) ACROSS Metrics Calculation and Visualization

While natural-language explanations help users understand retrieval results, they do not provide criteria for judging whether the information is sufficient, biased, or outdated. In cross-meeting retrieval, users are therefore likely to encounter situations in which results are plausibly explained, yet it remains unclear how much they should be relied upon.

In (A2), the system calculates and presents ACROSS metrics (Accuracy, Coverage, and Recency) as quantitative cues that characterize the properties of the retrieval results. A critical design principle is that these metrics are not treated as indicators of correctness.

Each metric is computed and visualized as an independent scale, allowing users to grasp relative tendencies:

- Accuracy: the degree of consistency between the query and the retrieval results,
- Coverage: the distribution and bias of the results across multiple meetings, and
- Recency: the temporal freshness of the results.

ACROSS provides users with a sense of sufficiency—whether it seems reasonable to make judgments based on the presented results alone. These metrics do not replace explanations; instead, they function as judgment cues that operate in parallel with them.

After retrieval results are determined, the system computes ACROSS metrics using the retrieved chunk set and associated metadata as input. Accuracy is calculated by normalizing multiple similarity signals obtained during retrieval to a 0–1 range and aggregating them as a weighted average. Recency is calculated based on timestamps using either absolute freshness with an exponential decay function or relative freshness based on percentiles within the target set. Coverage is calculated as a measure of diversity by applying simple clustering to the embedding vectors of retrieved chunks. These metrics are not integrated into a single score; instead, they are maintained and output as separate quantitative cues, independent of the generated answer text and natural-language explanations.

3.5 (A3) Integrated Presentation of Explanations and Indicators (Cross-Explanation)

In cross-meeting retrieval, search results often include discussions from multiple meetings and different points in time. As a result, users must not only understand the content, but also judge how much they can trust the results. Natural-language explanations (A1) provide cues for understanding why the results are related to the query, but they do not directly convey the sufficiency, bias, or freshness of the information. Similarly, ACROSS metrics (A2) summarize the properties of the retrieval results quantitatively, but they do not interpret meaning or make judgments on their own.

Therefore, in cross-meeting retrieval, presenting these elements separately is insufficient. The goal of (A3) is to place natural-language explanations and ACROSS metrics within the same retrieval context, enabling users to interpret and judge the results by referring to both simultaneously. In this study, such a structure is defined as Cross-Explanation.

Cross-Explanation does not introduce new explanatory content. Instead, it focuses on presenting the explanations generated in (A1) and the quantitative indicators from (A2) in a form that allows users to compare and interpret them together. In the proposed system, the generated answer text, reason_text, related chunks, and ACROSS metrics are maintained in a structure that can be retrieved as a single response. Based on this structure, the user interface places natural-language explanations and ACROSS metrics in parallel within the same screen.

The ACROSS metrics are not aggregated into a single overall score; Accuracy, Coverage, and Recency are presented separately. This design allows users to read the explanation text while simultaneously referring to each metric, preserving a state in which users can interpret the correspondence between explanations and indicators on their own. The presentation does not aim to guide operational procedures or explicitly provide traceability; instead, it focuses on ensuring that explanations and indicators enter the user's field of view at the same time.

4 Implementation

The prototype system proposed in this study is implemented to demonstrate the feasibility of Cross-Explanation in a realistic cross-meeting retrieval setting. The purpose of the implementation is not to compare models or optimize system performance, but to instantiate an interaction structure in which multiple judgment cues can be presented together for user trust judgment.

The system adopts a client-server architecture, consisting of a Python-based web API as the backend and a web browser-based client as the frontend. After meeting audio is uploaded, it is transcribed through automatic speech recognition and speaker diarization, and then segmented into utterance-level chunks with fixed time intervals. Each chunk is associated with metadata such as meeting ID, utterance timestamp, and speaker ID, and is represented using both embedding vectors and sparse representations (BM25) for retrieval.

As the retrieval infrastructure, the system employs a database that supports nearest-neighbor search over embedding vectors and keyword-based search using BM25. This setup enables the retrieval of sets of related chunks spanning multiple meetings. Importantly, retrieval is treated as a prerequisite step for trust judgment support rather than as a target of optimization. An overview of how retrieved chunks are subsequently used to generate multiple judgment cues is shown in Fig. 2.

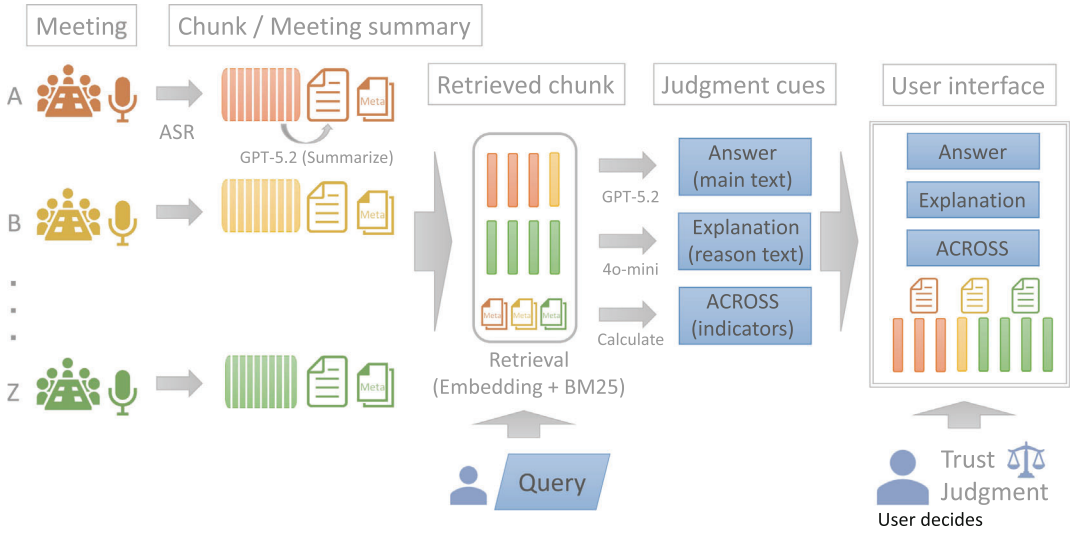

Fig. 2. System architecture for cross-meeting retrieval with Cross-Explanation. Meeting logs are segmented and retrieved across multiple meetings. Based on the retrieved chunks, the system independently generates answer text, natural-language explanations, and ACROSS indicators as judgment cues. These cues are presented side by side in the user interface, while the final trust judgment is left entirely to the user.

After retrieval, multiple post-processing components operate independently on the retrieved chunk set. For generating answers and meeting-level summaries in response to user queries, the system uses GPT-5.2 to provide fluent and readable outputs. In contrast, for generating natural-language explanations (reason_text), the system uses GPT-4o-mini as the default model to balance responsiveness and computational cost. In addition, the system supports a lightweight embedding-based generation method for reason_text that does not rely on large language model calls, allowing the generation strategy to be switched depending on practical constraints. These design choices illustrate that explanation cues are not tied to a specific model or generation technique.

The ACROSS metrics (Accuracy, Coverage, and Recency) are computed using existing retrieval metadata, such as similarity signals, meeting identifiers, and timestamps. These metrics reuse information already available in the retrieval process and are not treated as optimization objectives or indicators of correctness. Each metric is maintained as an independent quantitative cue, rather than being aggregated into a single score.

The backend returns the generated answer text, reason_text, the list of retrieved chunks, and the ACROSS metrics as a single response. On the frontend, these elements are presented side by side within the same screen, allowing users to refer to natural-language explanations and quantitative cues simultaneously. Through this presentation, the system realizes Cross-Explanation as a unified context for user trust judgment, while leaving the final judgment entirely to the user.

Figure 3 shows an example user interface of the prototype system, illustrating the presentation structure of the Combined (C) condition.

5 Experimental Setup

5.1 Study Design and Research Questions

This study was designed as an exploratory user study to investigate how different presentation formats of cross-meeting retrieval results influence users' trust judgments. In particular, we compare a condition that presents natural-language explanations only (X) with a condition that presents explanations together with quality indicators (ACROSS) (C). The goal is to clarify what kinds of cues users rely on when making trust judgments and how these cues differ depending on the presentation format.

The study addresses the following research questions:

- RQ1: What kinds of cues are used to make trust judgments in cross-meeting retrieval?
- RQ2: How do these cues differ between the X and C conditions?

Two presentation conditions were examined:

- X (Explanation only): a condition in which only natural-language summaries and explanations are presented for cross-meeting retrieval results.
- C (Combined): a condition in which quantitative indicators representing the properties of the retrieval results (ACROSS: Accuracy, Coverage, and Recency) are presented in addition to the explanation text used in X.

Because the primary focus of this study is the combination of explanations and indicators (Cross-Explanation), conditions such as evidence-only or metric-only were not included. These conditions were excluded to avoid increased analytical complexity and diffusion of comparison axes.

5.2 Materials and Tasks

As evaluation targets, two types of student research supervision meetings from the same research domain were prepared:

- mtg-a: a research supervision meeting for one student group,
- mtg-b: a research supervision meeting for a different student group.

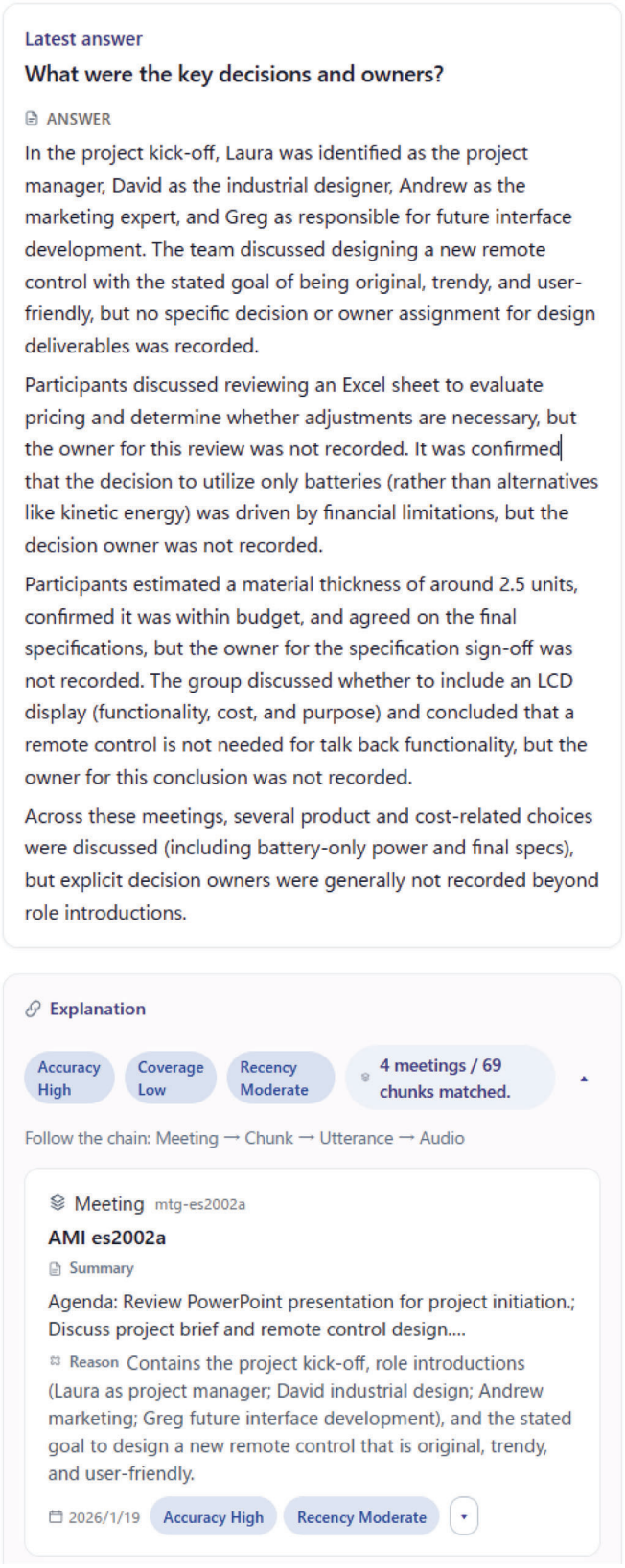

Fig. 3. Example user interface of the prototype system used to realize the Combined (C) condition. The screen illustrates how natural-language answers, explanation text (reason_text), and ACROSS indicators (Accuracy, Coverage, and Recency) are presented side by side as judgment cues. The content shown is illustrative and does not correspond to the meeting data used in the user study.

Although the two meetings shared the same research domain and supervision context, the participating students and the specific discussion contents differed. To establish a cross-meeting retrieval context, each participant evaluated only the meeting in which they had not participated. This design suppressed prior-knowledge bias and enabled observation of how the presented information itself influenced trust judgments.

For each meeting, four types of queries (Q1–Q4) were prepared:

- Q1 (Overall status): What is the current status of the student's research?
- Q2 (Key issues and challenges): What were the particularly important issues or challenges discussed in this meeting?
- Q3 (Decisions and agreements): What was decided in this meeting?
- Q4 (Next steps): What should be done next?

These queries are interpretive and summarization-oriented questions that can be evaluated without prior knowledge of correct answers, and they simulate a situation in which participants encounter unknown meeting logs for the first time.

5.3 Participants and Procedure

Six participants with domain knowledge in the target research field took part in the study. Participants were divided into two groups (Group A and Group B), and the evaluation targets and presentation conditions were assigned in a counterbalanced manner to avoid confounding effects. Table 1 summarizes the counterbalanced assignment.

Table 1. Assignment of evaluation targets and presentation conditions

Group	Queries	Condition
Group A (mtg-b)	Q1, Q2	X (Explanation only)
	Q3, Q4	C (Combined)
Group B (mtg-a)	Q1, Q2	C (Combined)
	Q3, Q4	X (Explanation only)

This counterbalancing allowed us to separate the effects of query content and presentation format while collecting subjective responses for both X and C conditions.

Immediately after viewing the result screen for each query, participants were asked to provide free-text responses regarding:

- reasons why they felt the answer was trustworthy or untrustworthy,
- information they felt was insufficient for making a judgment, and
- points of concern, discomfort, or uncertainty.

These free-text responses were treated as the primary evaluation data. As supplementary quantitative information, a single-item 7-point Likert scale ("This answer seems trustworthy") was also collected. This measure was not used for statistical testing and served only as reference information for interpreting the qualitative findings.

After completing the session, participants were additionally asked which of the X or C conditions they would prefer to use in practice, the reasons for their preference, and how they interpreted the ACROSS indicators presented in the C condition.

6 Findings

6.1 RQ1: Cues for Trust Judgment in Cross-Meeting Retrieval

T1. Trust Judgment Based on Content Comprehension and Readability. When judging the trustworthiness of retrieval results, participants first focused on whether the presented explanations or summaries were understandable. Specifically, factors such as whether key points were well organized, whether the structure of the text was easy to follow, and whether the content could be read naturally as a record of meeting discussions were cited as reasons for feeling that the results were trustworthy. In contrast, when the wording was unnatural or the explanations were overly abstract, participants reported difficulty in understanding the content and tended to withhold judgment. In this way, the ability to comprehend the explanation itself functioned as a starting point for trust judgment.

T2. Discomfort with Bias and Mixing Due to Summary Structure and Granularity. Even when content comprehension was achieved, participants often expressed discomfort with the structure and granularity of the summaries. Specifically, they pointed out cases in which the content felt overly summarized, cases in which detailed topics were mixed into responses to questions about the overall situation, and cases in which multiple studies or discussions were presented without sufficient organization. These reactions did not involve asserting that the presented content was incorrect; rather, they appeared as concerns that the overall picture or balance of the discussions might not be adequately conveyed due to the way the summary was constructed.

T3. Uncertainty About State Definiteness and Judgment Readiness. Some participants noted difficulty in understanding the stage of discussion represented by the retrieval results. Specifically, when it was unclear whether certain points reflected finalized conclusions or agreements, or were still under discussion, participants found it difficult to judge whether the information could be reliably used. These reactions were expressed not so much as concerns about the absolute recency of the information, but rather as uncertainty about whether the presented content was in a state suitable for making judgments. This was identified as one of the elements involved in trust judgment in cross-meeting retrieval.

6.2 RQ2: Comparison Between Explanation-Only (X) and Combined (C) Conditions

T4. Increased Sense of Reassurance and Conviction in the C Condition. In the Combined (C) condition, many responses referred to the presence of ACROSS indicators in addition to the natural-language explanations. In particular, participants described how the numerical information indicating the properties of the retrieval results served as supporting evidence for their judgments, alongside the explanation text itself. Rather than accepting the explanation at face value, participants attempted to interpret it together with information about the nature of the retrieval results. Even when the explanation content itself did not raise strong concerns, the presence of indicators tended to facilitate the articulation of a sense of psychological reassurance or conviction regarding their judgment.

T5. Instability of Judgment in the X Condition. In the Explanation-only (X) condition, several participants reported that although they could understand the explanation text itself, they still felt unable to make a definitive judgment based on it alone. Specifically, responses indicated lingering uncertainty due to the absence of additional cues for evaluating whether the summary or explanation was appropriate. These responses included references not only to the quality of the explanation text, but also to the lack of other information that could be consulted, suggesting that judgments were strongly influenced by impressions of the explanation alone.

T6. Changes in How Trust Judgments Are Articulated. Comparing the X and C conditions revealed differences in how participants articulated their trust judgments. In the X condition, participants tended to focus on content-oriented aspects, such as whether the explanation seemed correct or easy to understand. In contrast, in the C condition, participants more frequently used expressions describing degrees or tendencies, such as "how trustworthy it seems" or "what kind of overall tendency the results have," while referring to both the explanation and the indicators. This difference suggests that, independently of whether the final judgment itself was consistent, the way in which participants verbalized the process of reaching a judgment changed depending on the presentation format.

7 Discussion

7.1 Answering the Research Questions

RQ1: What Cues Do Users Rely on for Trust Judgment in Cross-Meeting Retrieval? The findings of this study suggest that, in cross-meeting retrieval, users' trust judgments depend less on the perceived correctness of the

content itself and more on cues related to judgment possibility. Specifically, participants' accounts of trust and distrust were grounded in whether the explanations or summaries were understandable (T1), whether the structure and granularity of the summaries appeared unbiased and well-balanced (T2), and whether the presented content seemed to be in a state suitable for use in decision-making (T3).

These cues emerged independently of whether users knew the "correct" content of the meetings. In situations where participants encountered unknown meeting logs, they emphasized not "whether the information is correct," but rather "whether it is appropriate to treat the information as a basis for judgment." This result suggests that cross-meeting retrieval is perceived not merely as an information-seeking task, but as a form of judgment-oriented activity that presupposes decision-making and reuse.

RQ2: How Does Combining Explanations with ACROSS Indicators Change Trust Judgment? The results related to RQ2 indicate that the difference between the Explanation-only (X) and Combined (C) conditions appeared not in the final outcome of trust judgments, but in how the judgment process itself was articulated. In the X condition, participants' reasoning was centered on the natural-language explanations, and trust was discussed mainly in terms of personal impressions of plausibility and clarity. In contrast, in the C condition, the presence of ACROSS indicators encouraged participants to construct their judgments by referring to multiple cues simultaneously.

Importantly, the ACROSS indicators did not deterministically drive trust judgments. Rather, their presence prompted participants to reread the explanation more carefully and to reflect on its relationship with the underlying evidence. This suggests that the proposed Cross-Explanation functioned not as an explanation that points to a single correct answer, but as one that expands the cognitive space in which judgments are formed.

7.2 Advantages of Cross-Explanation

A key advantage of the proposed approach lies in its intentional preservation of user agency in trust judgment. By placing explanations and indicators side by side, the system does not collapse trust into a single authoritative signal. Natural-language explanations support intuitive understanding, while ACROSS indicators present different perspectives on the properties of the retrieval results. This allows users to actively construct their judgments by cross-referencing multiple cues, rather than passively accepting a single explanation.

In addition, this study does not aim to directly evaluate correctness or task performance. Instead, it focuses on users' perceptions and fluctuations in judgment when encountering unknown meeting logs for the first time—a situation that reflects realistic usage contexts. In this sense, Cross-Explanation can be positioned as an attempt to redefine explainability not as a function for comprehension alone, but as an interaction that supports trust judgment.

7.3 Limitations and Future Work

This study has several limitations. First, the number of participants was small and the study design was exploratory; therefore, the findings are not intended to be generalized. Second, the evaluation was limited to research supervision meetings, and it remains unclear whether similar tendencies would be observed in other types of meetings or domains.

Furthermore, variations were observed in how participants interpreted the ACROSS indicators. While this variability was partly a consequence of our deliberate decision not to prescribe a fixed interpretation of the indicators, it also highlights an open question regarding what kinds of supplementary explanations or interactions would make such indicators more effective for judgment support. Future work should further explore the design space of Cross-Explanation by adjusting the granularity and contextualization of indicator presentation.

8 Conclusion

In this study, we exploratorily examined how users make trust judgments when interacting with retrieval results that span multiple meetings, focusing on differences in presentation formats. In particular, by comparing a condition that presents natural-language explanations only with a condition that combines explanations with quality indicators (ACROSS), we investigated not the outcomes of trust judgments themselves, but how the processes and articulations leading to those judgments change.

The results of the user study showed that trust judgments in cross-meeting retrieval are constructed less around the perceived correctness of the content itself and more around cues related to judgment possibility, such as the comprehensibility of explanations, the perceived appropriateness of summary structure, and whether the presented content appears to be in a state suitable for use in decision-making. Furthermore, when explanations were presented together with ACROSS indicators, users tended to construct their judgments by referring to multiple cues rather than relying on a single explanation.

These findings suggest the need to reposition explainability not as a function for conveying correct answers, but as a mechanism that supports the cognitive space in which users themselves assess trust. Through the perspective of Cross-Explanation—placing natural-language explanations alongside quality indicators—this work takes a step toward clarifying the design space for supporting trust judgment in cross-meeting retrieval.

Acknowledgments. This research was partially supported by JSPS KAKENHI Grant Numbers JP25H01167, JP25K02946, JP25K24389, JP24K02765, JP24K02774, JP23K17006, JP23K28091, JP23K28383.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akindele, O.K., Mishra, B.K., Wertheim, K.Y.: A knowledge graph and a tripartite evaluation framework make retrieval-augmented generation scalable and transparent (2025). https://arxiv.org/abs/2509.19209
2. Al Machot, F., Horsch, M.T., Ullah, H.: Building trustworthy ai: transparent ai systems viaÂ language models, ontologies, andÂ logical reasoning (transpnet). In: Al Machot, F., Horsch, M.T., Scholze, S. (eds.) Designing the Conceptual Landscape for a XAIR Validation Infrastructure. DCLXVI 2024. LNNS, vol. 1375, pp. 25–34. Springer, Cham (2025). https://doi.org/10.1007/978-3-031-89274-5_3
3. Baqar, M.: Rag4tickets: ai-powered ticket resolution via retrieval-augmented generation on jira and github data (2025). https://arxiv.org/abs/2510.08667
4. Brown, A., Roman, M., Devereux, B.: A systematic literature review of retrieval-augmented generation: techniques, metrics, and challenges. Big Data Cogn. Comput. **9**(12) (2025). https://doi.org/10.3390/bdcc9120320
5. Cheng, Q., et al.: Unified active retrieval for retrieval augmented generation. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 17153–17166. Association for Computational Linguistics, Miami, Florida, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.999
6. Cheng, S., Li, J., Wang, H., Ma, Y.: Ragtrace: understanding and refining retrieval-generation dynamics in retrieval-augmented generation. In: Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology, pp. 1—20. UIST '25, ACM (2025). https://doi.org/10.1145/3746059.3747741, http://dx.doi.org/10.1145/3746059.3747741
7. Ding, T., Banerjee, A., Mombaerts, L., Li, Y., Borogovac, T., la Cruz Weinstein, J.P.D.: Vera: validation and evaluation of retrieval-augmented systems (2024). https://arxiv.org/abs/2409.03759
8. Jean, N., Pera, G.L.: Bridging human cognition and ai: a framework for explainable decision-making systems (2025)., https://arxiv.org/abs/2509.02388
9. Juhasz, M., Dutia, K., Franks, H., Delahunty, C., Mills, P.F., Pim, H.: Responsible retrieval augmented generation for climate decision making from documents (2024). https://arxiv.org/abs/2410.23902
10. Kim, S.S.Y.: Establishing appropriate trust in ai through transparency and explainability. In: Extended Abstracts of the CHI Conference on Human Factors in Computing Systems. CHI EA '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3613905.3638184
11. Kovari, A.: AI for decision support: balancing accuracy, transparency, and trust across sectors. Information **15**(11) (2024). https://doi.org/10.3390/info15110725
12. Ni, B., et al.: Towards trustworthy retrieval augmented generation for large language models: a survey (2025). https://arxiv.org/abs/2502.06872
13. Opdahl, A.L., et al.: Trustworthy journalism through ai. Data Knowl. Eng. **146**, 102182 (2023). https://doi.org/10.1016/j.datak.2023.102182
14. Shi, I., et al.: Esapiens: a platform for secure and auditable retrieval-augmented generation (2025). https://arxiv.org/abs/2507.09588
15. Sunny, A.D.: Trust in transparency: how explainable ai shapes user perceptions (2025). https://arxiv.org/abs/2510.04968
16. Zerilli, J., Bhatt, U., Weller, A.: How transparency modulates trust in artificial intelligence. Patterns **3**(4), 100455 (2022). https://doi.org/10.1016/j.patter.2022.100455

17. Zhang, B., et al.: Practical poisoning attacks against retrieval-augmented generation (2026). https://arxiv.org/abs/2504.03957
18. Zhao, P., et al.: Retrieval-augmented generation for ai-generated content: a survey. Data Sci. Eng. (2026). https://doi.org/10.1007/s41019-025-00335-5
19. Zhou, P., Feng, Y., Yang, Z.: Provably secure retrieval-augmented generation (2025). https://arxiv.org/abs/2508.01084
20. Zhou, Y., et al.: Trustworthiness in retrieval-augmented generation systems: a survey (2024). https://arxiv.org/abs/2409.10102