Same Meeting, Different Minutes: A Controlled Cross-Platform Evaluation of Commercial AI Meeting Summarizers

Anonymous ACL submission

Abstract

Increasingly, workplaces are using intelligent agents to take notes and provide summaries during in-person and virtual meetings. Despite widespread adoption of AI, meeting summarization across commercial platforms, systematic quality evaluation remains limited. This study presents the first cross-platform analysis of meeting summarization performance, comparing native-platform summarizers (Google Meet, Zoom, and Microsoft Teams), agent-based summarizers (Otter.ai and Fireflies.ai), and general-purpose LLMS (ChatGPT, Llama, and Claude), all trained on identical meeting content. Our novel methodology uses automated methods (Rouge, BLEU, METEOR, BERTScore, and LSA) to score multiple runs of generated summaries. This standardized evaluation framework lays the groundwork for future investigations of summary quality under varying conditions, such as changes in signal-to-noise ratio and speaker volume. Our work establishes the first benchmark for commercial meeting summarization tools, revealing significant performance disparities and providing actionable insights for both platform users and researchers.

1 Introduction

In an effort to improve workplace efficiency, people are increasingly using artificialss intelligence (AI)- powered virtual agents to summarize meetings in many contexts. This recent change may be a direct result of an uptick in company meetings that an average worker is expected to attend, leading to an increased likelihood of confusion about diverse meeting topics covered in a single week and a greater need for organizational support to remain productive (Allen et al., 2022). Whether it be note-taking-specific AI agents (e.g., Otter.ai) or more general-purpose AI models (e.g., ChatGPT), workers are more likely to use such aides to reduce menial post-meeting tasks, including finalizing meeting notes, tracking topics covered, and documenting action items (Asthana et al., 2025; Wang et al., 2024).

As AI Meeting summarizers have become increasingly commonplace, there is a necessity for more formalized evaluations of the accuracy of generated summaries (Laskar et al., 2023). Evaluation is critical, as relying on inaccurate summaries that misrepresent the actual meeting can have serious implications, including omitting crucial information, disproportionate emphasis on certain topics, failing to credit the correct person for a contribution, drawing incorrect conclusions, hallucinating meeting content, and changing the meaning of an utterance. Decisions made using inaccurate information can have costly consequences for all workplace environments. Therefore, it is imperative to empirically evaluate meeting summarization tools.

This work presents an initial benchmark of commercial meeting summarizers and examines how summary accuracy varies across them. We used BERTScore, Recal-Oriented Undertstudy for Gisting Evaluation (ROUGE), Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), and Latent Semantic Analysis (LSA) to provide comprehensive quantitative scores for all meeting summaries (Zhang et al., 2019; Lin, 2004; Papineni et al., 2002; Banerjee and Lavie, 2005; Steinberger et al., 2004). We aim to understand the accuracy tradeoffs associated with choosing meeting summarizers. Accordingly, we created a framework to run a tightly controlled experiment, generating meeting summaries using 8 popular platforms, agents, and general-purpose LLMs. Our experimental results indicate that summary quality differs substantially across platforms. However, Llama produced the strongest overall alignment with the human reference summaries, ranking first or second across all metrics. These results underscore the need for additional studies on summary quality analysis. Fire-

flies.ai scored the lowest across all automatic metrics. The automated metrics disagreed in informative ways. For example, Zoom Smart Assistant scored poorly on ROUGE-L and BLEU but highly on METEOR and LSA. This finding suggests that some systems may preserve latent topical content while diverging from reference wording and structure.

The contributions of this preliminary investigation are manifold:

- We introduce a reusable controlled evaluation framework for comparing commercial AI meeting summarization tools under identical meeting conditions.
- We provide a cross-platform benchmark spanning native meeting platforms, third-party meeting agents, and transcript-to-summary LLM baselines
- We quantify cross-system performance and run-to-run variability to characterize summary tendencies and stability across repeated runs of the same meeting.
- We openly release the script, audio files, generated summaries, human reference summary, and prompts to support further replication and extension.

The script, audio files, codebase, annotations, and protocol are publicly available at the anonymized Open Science Framework linked here: osf.io.

2 Related Work

Meeting summarization sits at the intersection of two types of summarizations, dialogue and speech, and workplace collaboration. Summarization in this context is especially challenging, since meetings, by nature, are spoken, multi-party, temporally extended, and interactional. They can include interruptions, false starts, topic shifts, references to prior context, speaker-specific commitments, and decisions that gradually emerge rather than being subsumed into a single sentence. Some meetings could have unproductive conclusions, rendering future action plans null, something that modern-day summary models continue to struggle to identify. These context-specific properties make meeting summarization challenging for transcription systems and summarization models (Feng et al., 2022; Rennard et al., 2023; Retkowski et al., 2025).

Evaluating meeting summaries is a complex task that may lead to various complications. For example, there is no definitive gold standard for summarization due to human summarizers often prioritizing different aspects of a summary, including the level of detail, the information emphasized, the phrasing of the same content, and, most importantly, the intended purpose of the summary (Fabbri et al., 2021; Subbiah et al., 2025). Reference-based metrics remain useful, but they measure only part of summary quality. ROUGE captures lexical overlap, BLEU captures n-gram precision, METEOR incorporates alignment and lexical variation, and BERTScore estimates contextual semantic similarity (Lin, 2004; Papineni et al., 2002; Banerjee and Lavie, 2005; Zhang et al., 2020). These metrics are informative, but they can miss errors that matter in meetings, such as incorrect speaker attribution, unsupported decisions, and incomplete action items.

Prior work has made substantial progress on LLM-based meeting summarization and summary evaluation; however, we have yet to determine the reliability of commercial meeting summarizers. Recently, Laskar et al. have evaluated LLMs and prompting strategies for summarization, but this work does not compare deployed meeting platforms or third-party meeting agents (Laskar et al., 2023, 2024). In addition, Kirstein et al. showed that automatic metrics often fail to capture meeting-specific errors (Kirstein et al., 2024). Kirstein et al. also improved LLM-based meeting summaries using multi-source context and personalization. Despite these works motivating richer summary evaluation, they do not evaluate commercial systems under matched input conditions.

There are many commercial meeting summarizers. We selected Zoom, Google Meet, Microsoft Teams, Otter.ai, and Fireflies.ai because they represent the tool categories most likely to be encountered in workplace use. We selected ChatGPT, Claude, and Llama as summary generators as they represent common approaches to summarizing transcripts outside the meeting platform.

3 Methods

The present work evaluates whether commercially available meeting summarization systems produce materially different summaries for the same meeting. Accordingly, we developed a framework to evaluate meeting summarization systems across multiple platforms and "meeting summarizing

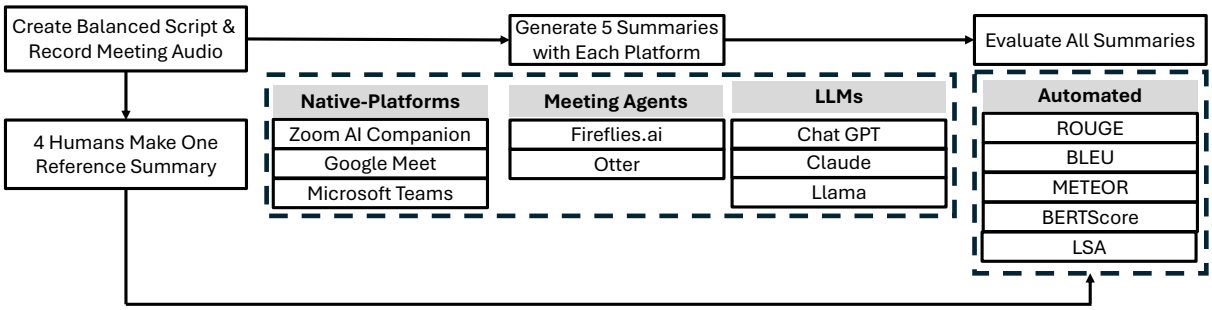

Figure 1: An overview of our research approach.

agents". Following established best practices in meeting summarization benchmarks, we employed a scenario-based approach that enables systematic comparison while controlling for content variability that could confound performance assessment ((Kirstein et al., 2024)). This tightly controlled experiment was designed to isolate the summarization system as the primary independent variable, as detailed in Figure 1. Accordingly, we created a scripted virtual meeting and recorded each speaker's audio stream. Each summarization system analyzed the same audio stream five times. The summary was generated and evaluated for errors. Our goal is to produce a standardized, reproducible cross-platform benchmark for meeting summaries under controlled input conditions.

3.1 Meeting Script Creation

We created a scripted workplace meeting in which four colleagues were planning a charity gala. This scenario was chosen because it replicated the problem-solving, competing ideas, decisions, responsibilities, urgency, and action items typically found in workplace meetings. To control for variability, the following script constraints were made:

- **Semi-Equal Number of Words Spoken:** To control for biases from conversation dominance, each speaker contributed 403 words to the scripted conversation.
- **Equal Number of Turns Taken:** To further control for biases from conversation dominance, each speaker had exactly 11 turns to speak in the scripted conversation.
- **Equal Position Rank:** To control for perceived power dynamics, each character in the script was at the same "level" with no one stated to be in charge or in a position of power.
- **Unisex Names:** To control for the possibility of gender bias, each character in the script was given a unisex name that was not distinctly male or female.

To enhance the meeting's believability, four trained actors from the School of Theatre and Dance were recruited to perform the scripted meeting. After a joint rehearsal to establish natural interaction patterns, actors were separated into individual rooms and equipped with identical Windows laptops configured with standardized Zoom settings (validated through preliminary recording tests). Participant display names set to the scripted speaker labels. During the virtual meeting to record the script, the individual speaker recordings setting was enabled, so each speaker's audio was saved locally on each laptop to facilitate testing with other meeting platforms. The script recording was conducted via Zoom with the internal transcription and Smart Assistant features activated.

3.2 Reference Summary Creation

Before generating any meeting summaries, we created a reference summary (or golden summary) to compare all generated summaries against. Four human summary writers followed a five-step protocol, using consensus to create the reference summary:

1. Listen to the meeting audio while following the script for reference.
2. Write individual summaries focusing on meeting purpose, logistics decisions, and next steps using specific participant names.
3. Read summaries aloud for group awareness
4. Designate a group leader for consensus facilitation
5. Collaboratively merge individual summaries into a consensus summary of approximately six sentences

This choice of protocol balances individual perspective with group consensus while maintaining standardized evaluation conditions.

3.3 Meeting Summary Generation

Table 1: Study Platform and Summarization Tools

Native-Platform	Agent	LLM
Zoom	Fireflies.ai	ChatGPT
Google Meet	Otter	Claude
Microsoft Teams		Llama

Meeting summaries were created by 3 types of tools. As shown in Table 1, these include *Native-Platform*, in which the summarization tool is embedded in the meeting platform, *Agent*, in which the summarization tool is an independent agent that is invited to a virtual meeting, and *LLM* in which different LLMs that were not in attendance at the meeting later summarized the meeting. Each Summarization tool analyzed the same meeting 5 times.

3.3.1 Models

Zoom AI Companion. Zoom AI Companion is a native meeting-platform summarizer that can generate meeting summaries from Zoom meetings and make those summaries available after the meeting (Zoom Support, 2026). In our study, Zoom was evaluated as an end-to-end commercial meeting summarization system rather than as a transparent LLM. We used Zoom Workplace version 6.4.12, with the default AI Companion Setting.

Google Meet. Google Meet's *take notes for me* feature can automatically capture meeting notes in Google Docs and provide a post-meeting recap link to the organizer (Google Meet Help, 2026). We evaluated Google Meet as a native meeting platform summarizer. We used Google Workspace with up-to-date Google Meet and with Gemini 2.5 Pro.

Microsoft Teams. Microsoft Teams Recap and Microsoft 365 Copilot provide AI-generated meeting summaries and follow-up tasks from Teams meetings (Microsoft Support, 2026). We evaluated Microsoft Teams as a native meeting platform summarizer. We used Teams version 26093.311.4599.3126, Microsoft 365 Copilot Version 2.20260519.48.0 with default Recap settings.

Otter.ai. Otter.ai is a third-party meeting agent that can join meetings, transcribe them in real time, and provide automated live summaries and meeting summaries (Otter.ai Help Center, 2025). We evaluated Otter.ai Pro as an external meeting-agent summarizer.

Fireflies.ai. Fireflies.ai is a third-party AI meeting assistant that can generate meeting summaries, takeaways, next steps, and action items after calls (Fireflies.ai Guide, 2026).

ChatGPT. ChatGPT is a conversational LLM interface developed by OpenAI and designed to follow user instructions in dialogue (OpenAI, 2022).

Claude. Claude is a family of LLMs developed by Anthropic (Anthropic, 2026).

Llama. Llama is an open-weight family of large language models developed by Meta (Touvron et al., 2023; Dubey et al., 2024).

3.3.2 Native-Platform Summarizers

To analyze each Native Platform Summarizer, a meeting was created on the platform with four participants. In some cases, the platform was upgraded to access the summarization features. Each participant was mapped to a separate computer, and the participant was renamed to match the character whose audio was being played. After the meeting audio ended on all computers, summaries were created using the platform's default settings. We did not prompt for additional information concerning the meeting. Each platform evaluated the same meeting for 5 different runs. Each run's summary was collected for analysis.

3.3.3 Agents

In the case of Fireflies.ai and Otter.ai, each agent was invited to a virtual meeting in which the scripted meeting audio was played over 4 different computers in 4 different rooms (to avoid audio leakage). Each agent evaluated the same meeting for 5 different runs. Following each run, the default generated summary was collected for analysis.

3.3.4 LLMs

The recorded meeting transcript was provided to the general-purpose LLMs. LLM-based summaries were given the simple standardized prompt *"Summarize this meeting"*. This prompt was chosen to avoid model-specific optimization bias. Each LLM was prompted five times to account for the

non-deterministic nature of large language model output. All 5 summaries produced by each run was collected for analysis.

3.4 Evaluation Metrics

To evaluate the meeting summaries, we used comprehensive evaluation metrics and automated tools.

3.4.1 Automated

Automated metrics included ROUGE, BLEU, METEOR, BERTScore, and Latent Semantic Analysis (LSA).

Results

Table 2: Automated Metrics Performance Across Summarization Systems. Higher scores indicate better performance except for BLEU. All scores represent means across five model runs where applicable.

System	ROUGE-L	BLEU	METEOR	BERTScore	LSA
Zoom Internal	0.228	4	0.345	0.850	0.162
Zoom Smart Asst.	0.155	2	0.362	0.850	0.285
Google Meet	0.216	5	0.331	0.849	0.217
Microsoft Teams	0.230	5	0.321	0.872	0.278
Fireflies.ai	0.175	2	0.179	0.847	0.162
Otter	0.256	3	0.227	0.879	0.192
ChatGPT	0.205	3	0.378	0.855	0.218
Claude	0.204	4	0.335	0.861	0.204
Llama	0.292	7	0.372	0.874	0.284
Humans	1.000	100	1.000	1.000	1.000

In this section, we evaluate the results of multiple experiments to characterize the performance of each method used to produce meeting summaries.

In Table 2, the automated metrics performance shows the performance of the 10 summaries across 5 metrics. We note that, for example, the ROUGE and BLEU scores should be 1 or 100 for humans, as this is the reference summary.

Across the five trials run and then averaged, low ROUGE-L, BLEU, METEOR, and LSA scores were observed. ROUGE-L scores give insight into the lexical overlap between the human generated summary or the study's gold standard summary. A low score indicates there was much overlap between the two summaries. BLEU scores show how much the phrasing overlaps between the two and the low numbering shows overlap was not present overall. METEOR scores are low, showing that even when flexibility is provided in word matching, the reference summary and the AI-generated one are not very close. The BERTScore is fairly high, showing that when context is added to the evaluation metric, it can tell that the two summaries effectively contain the same meaning despite vastly different wording. The LSA score shows that the latent topic similarity is low between the reference and the AI summaries.

4 Discussion & Conclusion

In this section, we will interpret the results.

Across automatic metrics, Llama consistently produced the strongest overall alignment with the human reference summaries, ranking first or second on every metric. Among commercial meeting tools, Microsoft Teams aso showed the most stable cross-metric performance, while Otter achieved the highest BERTScore and second-highest ROUGE-L, but performed less strongly on METEOR and LSA. Fireflies.ai consistently scored lowest across most metrics. The metrics also disagree in informative ways: Zoom Smart Assistant scored poorly on ROUGE-L and BLEU but highly on METEOR and LSA, suggesting that some systems may preserve latent topical content while diverging from reference wording and structure. The authors caution that since automatic metrics can miss or mask meeting-specific errors, such as incorrect references, hallucinations, and missing action items, these results should be paired with the planned human error coding and LLM-as-judge evaluation.

Low Rouge-L, BLEU, METEOR, and LSA scores across all summarizers when compared to human summarizers shows us that LLMs and AI notetakers within meeting platforms can capture meaning about the meeting they are added to or asked to summarize, however they will not take notes or summaries in the same a human might. Their wording for the same ideas will be much different to how a human will choose to word or phrase the same ideas.

Limitations

This study employs several methodological choices that, while introducing limitations, were necessary for rigorous comparative evaluation. Our use of a scripted meeting ensures consistent content and timing across all platforms, enabling fair comparisons by controlling for variables such as topic drift and uneven participation that could confound results.

The authenticity of our actors' performances, though carefully rehearsed, may not fully capture the nuances of genuine business meetings. However, since our primary goal was a comparative

assessment of platform capabilities rather than exact realism, the trained delivery and attention to the subtleties of human speech strengthened our ability to isolate differences in summarization quality from conversational noise or from an untrained person reading the script.

Our platform selection, while comprehensive, represents a subset of available tools due to practical constraints. We focused on widely adopted commercial solutions and leading AI systems, providing coverage of the primary market segments most relevant to current professional practice. Additionally, our controlled audio environment excluded common meeting challenges, such as frequent interruptions or poor connection quality, allowing us to establish optimal-condition baselines before introducing additional complexity variables.

5 Conclusion

This work is an exploratory first step in understanding how AI-generated summaries compare with human-generated summaries. In future work, the authors hope to expand upon this and include evaluation metrics such as LLMs as judges and evaluators.

References

Joseph A. Allen, Matt S. Thiese, Emilee Eden, and Sarah E. Knowles. 2022. Why am i so exhausted? *Journal of Occupational and Environmental Medicine*, 64(12):1053–1058.

Anthropic. 2026. Models overview. https://docs.anthropic.com/en/docs/about-claude/models. Accessed: 2026-05-26.

Sumit Asthana, Sagi Hilleli, Pengcheng He, and Aaron Halfaker. 2025. Summaries, highlights, and action items: Design, implementation and evaluation of an llm-powered meeting recap system. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–29.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Xiachong Feng, Xiaocheng Feng, and Bing Qin. 2022. A survey on dialogue summarization: Recent advances and new frontiers. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, pages 5453–5460.

Fireflies.ai Guide. 2026. Fireflies AI meeting summaries: View, customize, expand, regenerate. https://guide.fireflies.ai/articles/9547055509-Fireflies-AI-Meeting-Summaries-View-Customis Accessed: 2026-05-26.

Google Meet Help. 2026. Take notes for me in Google Meet. https://support.google.com/meet/answer/14754931. Accessed: 2026-05-26.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024. What's under the hood: Investigating automatic metrics on meeting summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6709–6723.

Md Tahmid Rahman Laskar, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2023. Building real-world meeting summarization systems using large language models: A practical perspective. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 343–352.

Md Tahmid Rahman Laskar, Elena Khasanova, Xue-Yong Fu, Cheng Chen, and Shashi Bhushan Tn. 2024. Query-OPT: Optimizing inference of large language models via multi-query instructions in meeting summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1140–1151. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Microsoft Support. 2026. Recap in Microsoft Teams. https://support.microsoft.com/en-us/teams/meetings/recap-in-microsoft-teams. Accessed: 2026-05-26.

OpenAI. 2022. Introducing ChatGPT. https://openai.com/index/chatgpt/. Accessed: 2026-05-26.

Otter.ai Help Center. 2025. Otter Notetaker overview. https://help.otter.ai/hc/en-us/articles/4425393298327-Otter-Notetaker-Overview. Accessed: 2026-05-26.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Virgile Rennard, Guokan Shang, Julie Hunter, and Michalis Vazirgiannis. 2023. Abstractive meeting summarization: A survey. *Preprint*, arXiv:2308.01691.

Fabian Retkowski, Maike Züfle, Andreas Sudmann, Dinah Pfau, Shinji Watanabe, Jan Niehues, and Alexander Waibel. 2025. Summarizing speech: A comprehensive survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Josef Steinberger, Karel Jezek, and 1 others. 2004. Using latent semantic analysis in text summarization and summary evaluation. *Proc. Isim*, 4(93-100):8.

Melanie Subbiah, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. 2025. Is the top still spinning? Evaluating subjectivity in narrative understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Ruotong Wang, Lin Qiu, Justin Cranshaw, and Amy X. Zhang. 2024. Meeting bridges: Designing information artifacts that bridge from synchronous meetings to asynchronous collaboration.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations*.

Zoom Support. 2026. Using meeting summary with AI companion. `https://support.zoom.com/hc/en/article?id=zm_kb&sysparm_article=KB0058013`. Accessed: 2026-05-26.