You need to MIMIC to get FAME: Solving Meeting Transcript Scarcity with Multi-Agent Conversations

Frederic Kirstein[1,2], Muneeb Khan[1], Jan Philip Wahle[1], Terry Ruas[1], Bela Gipp[1]
[1]University of Göttingen, Germany
[2]kirstein@gipplab.org

Abstract

Meeting summarization suffers from limited high-quality data, mainly due to privacy restrictions and expensive collection processes. We address this gap with FAME, a dataset of 500 meetings in English and 300 in German produced by MIMIC, our new multi-agent meeting synthesis framework that generates meeting transcripts on a given knowledge source by defining psychologically grounded participant profiles, outlining the conversation, and orchestrating a large language model (LLM) debate. A modular post-processing step refines these outputs, mitigating potential repetitiveness and overly formal tones, ensuring coherent, credible dialogues at scale. We also propose a psychologically grounded evaluation framework assessing naturalness, social behavior authenticity, and transcript difficulties. Human assessments show that FAME approximates real-meeting spontaneity (4.5/5 in naturalness), preserves speaker-centric challenges (3/5 in spoken language), and introduces richer information-oriented difficulty (4/5 in difficulty). These findings highlight that FAME is a good and scalable proxy for real-world meeting conditions. It enables new test scenarios for meeting summarization research and other conversation-centric applications in tasks requiring conversation data or simulating social scenarios under behavioral constraints[1].

1 Introduction

Meetings underlie collaboration and decision-making in corporations, academia, and government. Meeting summaries help record key discussion points, update absentees, and capture to-dos (Zhong et al., 2021; Hu et al., 2023; Laskar et al., 2023). While AI-based summaries are available on platforms such as Zoom[2], Microsoft Teams[3], they typically build on limited English data that does not represent the diversity of real meetings, e.g., multilingual sessions, specialized discussions (Kirstein et al., 2025a). Data scarcity for training and testing meeting summarization systems persists due to privacy and intellectual property concerns, along with expensive manual annotation (Ben Abacha et al., 2023). Existing corpora, such as AMI (Mccowan et al., 2005), ICSI (Janin et al., 2003), and MeetingBank (Hu et al., 2023), offer only a narrow range of scenarios, which primarily revolve around staged business, academic, or parliamentary meetings. Non-English resources like FREDSum (Rennard et al., 2023), containing manually transcribed and annotated political debates, remain sporadic and underscore the lack of linguistic diversity.

Researchers have explored synthetic transcripts to address this scarcity, but many methods are suboptimal in scalability and realism. Single-model, omniscient continuation (Qiu and Pan, 2024; Zhou et al., 2024a) can produce dialogues lacking actual knowledge and speaker interplay, while crowdsourced role-plays are expensive (Mccowan et al., 2005; Thulke et al., 2024). Automated heuristics (e.g., noising, swapping) often yield disjointed conversations (Chen and Yang, 2021; Park et al., 2022; Liu et al., 2022). These approaches struggle to balance large-scale generation with authentic group dynamics and credible topic evolution, even though they can process and generate thousands of tokens.

We introduce **MIMIC** (**M**ulti-agent **IMI**tation of **C**onversations, see Figure 1), a movie-production-inspired framework based on multi-agent debate (Liang et al., 2024; Du et al., 2024). MIMIC summarizes a knowledge source, expands the summary into an agenda, and orchestrates psychologically grounded agents debating turn-by-turn, allowing interruptions (e.g., phone calls). A modular refinement step mitigates repetitions or overly formal speech, ensuring plausible and coherent discourse at scale. With MIMIC, we generate **FAME** (**FA**ke

[1]Resources are available as per Appendix A on GitHub.
[2]https://www.zoom.com/en/ai-assistant
[3]https://copilot.cloud.microsoft

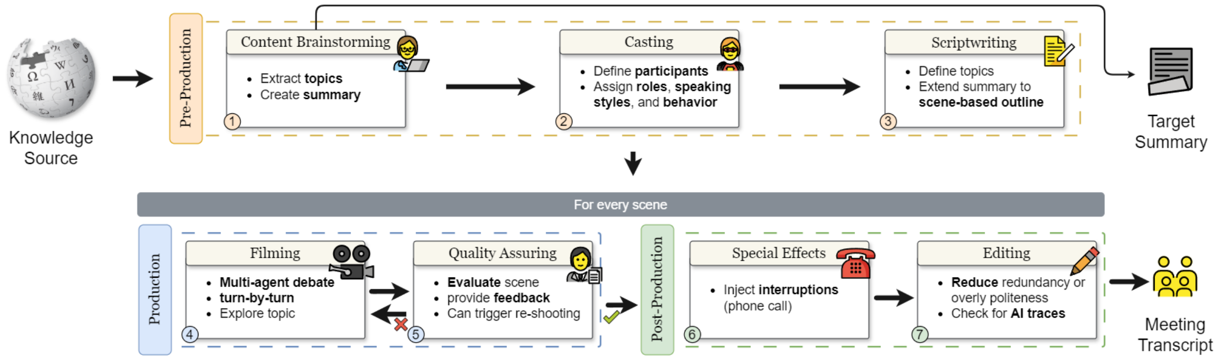

Figure 1: Architecture of our meeting generation framework MIMIC. Stages 1 to 3 are realized as a linear pipeline of LLM instances. Stage 4 is an agent orchestrating participants played by agents, who can decide on the length of the meeting through voting. Stage 5 checks and optionally provides feedback for refinement to stage 4 agent. The remaining stages, 6 and 7, polish the generated conversation.

MEetings), a corpus of 800 meetings (500 English, 300 German parallel in input) covering 14 meeting types (e.g., project updates, brainstorming) and 300 Wikipedia articles as a knowledge source.

We evaluate the synthetic meetings on quality criteria, e.g., naturalness, coherence, (Chen et al., 2023), and transcript difficulties, e.g., scattered information (Kirstein et al., 2024b), and propose a new psychology-based measure of behavioral authenticity (Choi et al., 2020). Human ratings confirm that FAME achieves near-real spontaneity (4.5/5 in naturalness), preserves speaker-related challenges (3/5 in spoken language), and intensifies low-information-density difficulties (4/5). Comparisons with real transcripts and 100 crowdsourced meeting experiences show a similar behavioral pattern. Evaluations of GPT-4o (OpenAI, 2024), Gemini 1.5 pro (Team et al., 2024), DeepSeek-R1 Distill Llama 70B (DeepSeek-AI and Zhang, 2025), Llama 3.3 70B[4] (Llama Team, 2024) reveal persistent context-handling issues. Ablation studies show that MIMIC reliably generates good transcripts using different inputs or LLMs.

Contributions.

- **MIMIC**: A multi-agent simulation method that captures realistic group dynamics.
- **FAME**: A corpus of 800 meetings in English (500) and German (300) on diverse topics and meeting formats, including quality annotations.
- A **psychology-based evaluation framework** for behavioral authenticity, addressing a gap in measuring agent interactions.

[4] We will refer to these as GPT, Gemini, DeepSeek, and Llama throughout the paper.

2 Related Works

Meeting summarization datasets. Most meeting summarization research relies on a few standard English-only corpora (Kirstein et al., 2025a), notably AMI (Mccowan et al., 2005), i.e., staged business, and ICSI (Janin et al., 2003), i.e., academic. While QMSum (Zhong et al., 2021) and MeetingBank (Hu et al., 2023) broaden coverage to parliamentary sessions and city councils, non-English data remains sparse (e.g., FREDSum (Rennard et al., 2023) for French and ELITR (Nedoluzhko et al., 2022) for Czech). We address this gaps by releasing FAME, a corpus of 800 synthetic English and German meetings spanning diverse topics, meeting formats, and speaking styles.

Synthetic dialogue generation. Existing synthetic meeting data generation approaches often rely on relatively simple text continuation by LLMs (Qiu and Pan, 2024) or heuristics, e.g., noising, swapping (Chen and Yang, 2021; Park et al., 2022; Liu et al., 2022), risking superficial turn-taking and unrealistic participant behavior (Kirstein et al., 2025a). Small-scale manual simulations (Thulke et al., 2024) offer greater realism but are costly and hard to scale. Recent tools like Google's NotebookLM (Google, 2024) and Nvidia's PDF-to-Podcast (Nvidia, 2025) transform documents into two-speaker podcasts but lack multi-participant group dynamics. In contrast, MIMIC simulates turn-by-turn interactions among psychology-based agents with their own memory, allowing spontaneous debates and evolving stances. A post-processing module addresses common LLM flaws (e.g., repetition, vocabulary), ensuring high-quality, naturally flowing conversations.

3 The MIMIC Methodology

We propose MIMIC, a multi-agent framework inspired by movie production to generate synthetic meeting transcripts from a knowledge source through multi-agent debate (see Figure 1). The basic idea is to summarize the knowledge source to distill its key highlights to be discussed in a multi-agent-LLM setup. This setup emulates a discussion among participants with distinct personas and their private memory of the knowledge source, including real-world dynamics such as turn-taking, disagreements, clarifications, and topic continuity. MIMIC operates in three phases, i.e., *pre-production*, *production*, and *post-production*, with seven stages overall (splitted 3/2/2 between each phase). We explain here the methodological background of MIMIC, the corresponding prompts, and implementation details are covered in Appendices C and H. In Appendix I, we present the result of each stage using an example knowledge source.

3.1 Pre-Production Phase

This phase establishes foundational elements, including the meeting's target summary, participant roles, and an agenda-like outline.

Stage 1: Content Brainstorming. Given a source text, we prompt an LLM to extract hierarchical topics and subtopics, following the approach outlined in Paoli (2023). An LLM composes an abstractive target summary following Gao et al. (2024), guided by five human-written QMSum summaries for consistent brevity, style, and structure. This summary covers the later discussion points and topic flow, acting as a basic outline.

Stage 2: Casting. We define participant profiles suited to the meeting context. Each profile contains a functional role (e.g., project manager, technical expert), background (e.g., experience, qualifications), domain expertise, and a distinct perspective (e.g., favoring practical solutions). An LLM iteratively creates these profiles, ensuring complementary viewpoints without redundancy. Next, each profile receives a speaking style, including tone (e.g., formal), language complexity (e.g., jargon), communication style (e.g., assertive), plus filler words (e.g., "um," "you know") and catchphrases to align the person's language with their role.

We distributed select knowledge-source paragraphs to each participant based on expertise (Li et al., 2025), introducing knowledge imbalances

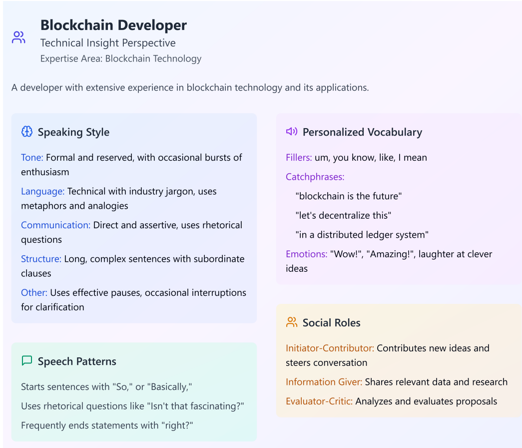

Figure 2: Example of a minimal participant definition.

that foster reliance on one another. Building on Serapio-García et al. (2023) research on LLM personality traits, each participant is assigned psychologically grounded behaviors, e.g., evaluator-critic, blocker (Benne and Sheats, 1948), drawn from a curated list (full list given in Table 6 in Appendix C). These behaviors can shift according to the meeting format or topic, allowing realistic group dynamics and participant evolution. An LLM also checks for contradictory traits (e.g., proactive yet blocking) to maintain role consistency.

Stage 3: Scripting. An LLM expands the target summary into a flexible outline of thematic scenes, each dedicated to a major subtopic of the knowledge source. A scene includes a title, an agenda description, and bullet points covering key content. Scenes also specify relevant aspects of the source material and invite participants to draw on personal experiences, reflecting each persona. Depending on the meeting type, extra scenes (e.g., a pre-meeting brainstorming phase) may be added for realism, even if they are not part of the target summary. This outline serves as a roadmap, allowing participants to briefly shift themes or discuss subtopics without straying from the broader structure.

3.2 Production Phase

This phase simulates turn-by-turn dialogue among multiple LLMs and validate each scene for quality.

Stage 4: Filming. We generate transcripts scene by scene, following the outline from Stage 3: Scripting. Each participant is an independent LLM instance, contributing one turn at a time. To emulate real-world scenarios (Zhou et al., 2024a), we implement a non-omniscient approach where each participant has private memory, seeing only their own profile, relevant source snippets, a summary of prior scenes, the last three turns of the previous scene and all turns so far in the current scene.

Scenes begin with the participant most relevant to the topic, determined by the exact LLM-based matching used during casting (Li et al., 2025).

At the end of each turn, the current speaker nominates the next speaker based on roles, current focus, and prior contributions (e.g., previously raised concerns or ideas introduced) (Nonomura and Mori, 2024). Reminding the current speaker about the other participants prevents them from being left out.

To conclude a turn, the speaker can propose a vote to end a scene (Wang et al., 2024). The vote can be called after the first turn if the speaker finds that they have stated all the details in their individual knowledge that fit the current discussion topic. If the majority (more than 50%) of the participants agree during voting, the scene ends. To avoid endless loops, a system reminds the group to finalize if no one initiates a vote after 50 turns. MIMIC then proceeds to the next scene or terminates if all scenes are complete.

Stage 5: Quality assuring. Inspired by 'Self-Refinement' (Madaan et al., 2023) where LLMs iteratively provide feedback and refine their output, we use a "director" model to review each scene along three dimensions: **topical alignment**, i.e., adherence to the target summary (Lin and Chen, 2023), **conversational naturalness**, i.e., turn-taking quality, dialogue flow (Liu et al., 2023), and **coherence/factual accuracy**, i.e., logical progression, consistency with knowledge source (Xie et al., 2024). The director LLM provides feedback to correct any critical issues identified (e.g., missing subtopics, overly formal language, and contradictions). We allow up to three "re-filming" cycles, considering this feedback. If a scene remains problematic after three attempts, MIMIC proceeds with the best version, deferring residual edits to the post-production phase. In practice, this fallback was never triggered for FAME.

3.3 Post-Production Phase

This phase injects disruptions for realism and polishes transcripts to ensure real-sounding meetings.

Stage 6: Special effects. We inject flow-breaking events (e.g., phone calls, technical glitches, side questions) with a 25% chance per scene[5], permitting multiple disruptions per meeting. These events lead to adding a few turns from the participants (e.g., acknowledging a ringing phone) before the main discussion resumes.

Stage 7: Editing. A two-step linguistic refinement removes repetitive phrases and overly formal speech and adds minor speech markers (e.g., hesitations). A subsequent detector–revision step targets any remaining synthetic cues (e.g., uniformly polite turns, unrealistic consensus). If necessary, minor disagreements or paraphrases are introduced to achieve more natural, human-like transcripts.

4 The FAME Dataset

In the following, we detail how FAME was generated, present its overall statistics, and analyze the authenticity of its simulated meetings and how challenging they are compared to real transcripts. All data and annotations performed here are available per Appendix A.

4.1 Setup to generate FAME

Knowledge source. Unlike synthetic datasets built on minimal or artificial context, FAME builds on grounded text from Wikipedia. We select 300 articles from 28 broad domains (e.g., Global Issues, Technological Innovations, Cultural Diversity, Philosophy, Environment & Ecology), each meeting three criteria: (1) have at least five subsections, (2) no negative flags regarding article quality, and (3) no reference deficiencies (to avoid contradictory claims). For the German subset, we choose 150 German-language articles from the same pool, ensuring a high content overlap with their English counterparts via BERTScore (Zhang et al., 2020) and an empirically chosen threshold of 0.7. For each article, we randomly assign meeting types, participant roles, and the number of participants, yielding 500 English and 300 German meetings.

Backbone model. We use GPT (OpenAI, 2024) for all stages of MIMIC, taking advantage of its 128k-token context window and robust role-playing capabilities (Kirstein et al., 2024a). In Section 6, we show that models with fewer parameters (Gemini, DeepSeek, Llama) can generate high-quality transcripts with minor drops in naturalness.

[5]Chosen empirically to balance realism without overwhelming the meeting.

Dataset	# Meetings	# Speaker	# Unique Spea.	# Turns	# Words	Vocab.	Token Overlap	Sum. Len.	Interruptions	Language
AMI	137	$4.0_{0.00}$	4	$513.5_{266.2}$	$4937.5_{1999.3}$	9388	-	$109.9_{27.1}$	no	informal
ICSI	44	$6.2_{1.3}$	35	$757.5_{374.8}$	$9889.4_{3794.9}$	9164	-	$93.3_{22.2}$	no	formal
WCPC	51	$16.8_{18.7}$	316	$337.3_{277.3}$	$11427.8_{4574.0}$	13780	-	$122.3_{39.2}$	no	informal
QMSum	232	$7.2_{10.1}$	330	$521.0_{320.4}$	$7303.4_{4232.2}$	20505	-	$109.5_{30.7}$	no	both
EN	500	$5.1_{2.8}$	3200	$405.0_{330.3}$	$6223.4_{4084.4}$	10347	0.081	$207.7_{22.7}$	yes ($\sim$ 0.5)	both
GER	300	$5.0_{2.8}$	1000	$393.3_{323.2}$	$6272.4_{3793.2}$	9589	0.096	$170.3_{29.0}$	yes ($\sim$ 0.5)	both

Table 1: Statistics on FAME for English (EN) and German (GER) and established corpora. Values are Mean_{Std}.

Baseline. We compare FAME to QMSum (Zhong et al., 2021), an established dataset combining academic (ICSI (Janin et al., 2003)), product (AMI (Mccowan et al., 2005)), and parliament (Welsh/Canadian, WPCP) meetings. While other corpora exist (e.g., MeetingBank (Hu et al., 2023), ELITR (Nedoluzhko et al., 2022)), they closely resemble QMSum's formal institutional settings. We also contrast our non-omniscient multi-LLM pipeline with a single omniscient GPT approach (Zhou et al., 2024a) that generates entire meetings in one shot given the same article, target summary, speaker count, and meeting type used for FAME. To match in meeting length, we prompt GPT multiple times to circumvent its 16k-token output limit and produce transcripts of ~20k tokens.

4.2 General Statistics

Subsection key finding. FAME's diversity enables large-scale benchmarking of summarizers, mimicking the varied conditions seen in real meetings.

Comparison. Table 1 summarizes statistics for FAME. Our average meeting length (~6,200 words), turn count (~400 turns), and participant count (~4) closely mirrors real corpora (e.g., AMI), though FAME has higher variance (e.g., larger standard deviation for turn counts).

Unique features. In addition to standard business, academic, and parliamentary contexts, FAME spans 14 new meeting types and 28 Wikipedia-based domains. FAME's dialogues primarily rephrase rather than copy source passages (average token overlap of English: 0.081, German: 0.096). Over 3,000 participanth adopt unique speech patterns and undergo up to four behavioral shifts, introducing dynamic role changes beyond the fixed-character setups typical of existing corpora. One-third of the meetings have interruptions, with 20% of those featuring three or more disruptions. Existing corpora typically lack these unpredictabilities.

4.3 Authenticity Evaluation

Subsection key finding. FAME meetings exhibit near-real conversation flow, and participants behave in ways that closely match human expectations.

Approach. We evaluate meeting authenticity along two axes: overall authenticity, i.e., coherence, consistency, interestingness, naturalness (Chen et al., 2023), and participant behavior authenticity, e.g., conflicts and power dynamics, defined in consultation with psychology and sociology experts. Therefore, eight overarching behavior categories, i.e., knowledge, power, conflict, status, trust, support, similarity, and fun (Bales et al., 2009; Choi et al., 2020), were divided into 18 items complete list given in Table 8 in Appendix D).

Six annotators (students to PhD candidates in Psychology, Computer Science, Communication; aged 23–28, at least C1 English/German) rate 30 English and 30 German meetings from FAME, as well as 30 meetings from QMSum, using a 1–5 Likert scales Krippendorff's $\alpha = 0.83$[6]. Because existing real-meeting corpora may exhibit a behavior bias due to the formal or staged nature, we also surveyed 100 crowdworkers (age 24–63, balanced by gender, diverse professional backgrounds such as law, engineering, and management) who frequently attend meetings to collect experiences with real meetings. Additional details on annotators, crowdworkers, and reliability are in Appendix E.

Quantitative analysis. Table 2 compares overall authenticity scores for FAME and QMSum. The English/German FAME subsets match QMSum in coherence, while both subsets score higher in conciseness and interestingness (4.5/5 in FAME vs. 4/5 in QMSum). ICSI meetings receive the lowest interestingness rating (3/5) due to slower topic evolution, whereas WPCP leads in naturalness (5/5).

[6]To support broader community use, we extend annotations across the entire FAME, we used GPT with the three-step approach of Kirstein et al. (2025b), aligned automatic scores through the self-training concept to match human ratings until the average discrepancy was below 0.8 points.

	EN	GER	AMI	ICSI	WPCP	QMSum	OMNI
COH	$\mathbf{4.5}_{0.00}$	$\mathbf{4.5}_{0.18}$	$\mathbf{4.5}_{0.36}$	$3.5_{0.36}$	$\mathbf{4.5}_{0.00}$	$\mathbf{4.5}_{0.73}$	$3.5_{0.00}$
CON	$\mathbf{4.5}_{0.07}$	$\mathbf{4.5}_{0.09}$	$\mathbf{4.5}_{0.68}$	$3.5_{0.87}$	$\mathbf{4.5}_{0.38}$	$4_{0.59}$	$3.5_{0.09}$
INT	$\mathbf{4.5}_{0.13}$	$\mathbf{4.5}_{0.23}$	$4_{0.68}$	$3_{0.87}$	$4_{0.38}$	$4_{0.77}$	$4_{3.08}$
NAT	$4.5_{0.12}$	$4_{1.37}$	$4.5_{0.55}$	$4.5_{0.82}$	$\mathbf{5}_{0.00}$	$4.5_{0.90}$	$3_{0.74}$

Table 2: Evaluation of synthetic meetings on COH = coherence, CON = consistency, INT = interestingness, and NAT = naturalness. Values are Median$_{Std}$. EN, GER are FAME English and German. OMNI is a single LLM. **Higher** is better.

The German FAME subset lags the English one by 0.5 points in naturalness (English: 4.5/5), which is similar to most QMSum data. Notably, transcripts generated by the single omniscient model rank lowest overall due to shallow, repetitive content.

As shown in Figure 3, social behavior patterns in FAME align closely with both QMSum (within 0.2 points) and the crowdsourced baseline (within 1.0 points). Modest differences likely reflect gaps between self-assessment and observed performance (Kruger and Dunning, 1999), often linked to social desirability or recency biases. The synthetic meetings' mild overperformance in some categories suggests a slight dramatization of behaviors. We show the behavior pattern in single-LLM meetings in Appendix H, which just minimal overlaps with the FAME patterns. In sum, MIMIC speakers reliably replicate real behaviors.

Qualitative insights. Annotators highlight the realistic interplay of participant knowledge, consistent role portrayal, and dynamic sub-topic exploration. Occasionally, overuse of transitional phrases (e.g., "This sounds great, but...") or overly cordial tones reveal the synthetic origin (noted in 57 out of 800 meetings). Despite these artifacts, FAME remains a near-realistic meeting proxy, offering a robust environment to develop and benchmark summarization methods. In contrast, the omniscient approach often yields shallow dialogues, word/passage repetition, or even recitations from well-known sources (e.g., the Bible).

4.4 Transcript Challenge Assessment

Subsection key finding. FAME preserves speaker-related complexities and extends information-centric challenges closer to real meetings.

Analysis. We adopt the seven challenges from Kirstein et al. (2024b), i.e., spoken language, speaker dynamics, coreference, discourse structure, contextual turn-taking, implicit context, and low information density, to gauge the difficulty of summarizing FAME transcripts (Table 3). We compare FAME's median 1–5 Likert ratings against human-annotated QMSum data (Kirstein et al., 2024b). Overall, FAME matches QMSum in spoken language (3/5) and coreference (2/5, German subset increases to 3/5), though it exhibits calmer speaker dynamics (2/5 vs. ∼3/5 in QMSum). Both FAME subsets surpass QMSum in implicit context (FAME: 4/5 vs. QMSum: 0/5) and low information density (FAME: 4/5 vs. QMSum: 2.5/5), indicating that participants rely on prior exchanges instead of explicitly reiterating every detail. By contrast, many existing corpora emphasize exhaustive information sharing (ICSI, WPCP) or feature staged, side-discussion-free scenarios (AMI).

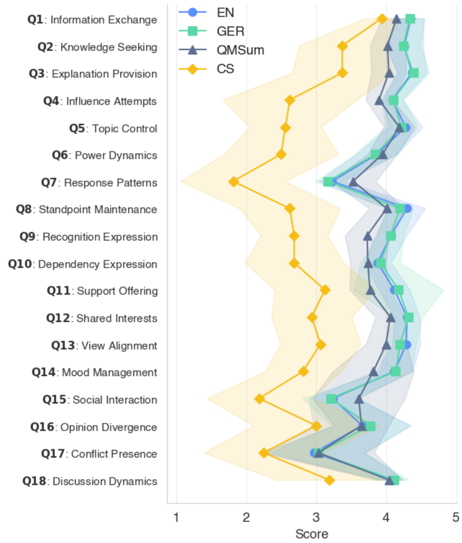

Figure 3: Evaluation of the social behavior in the meetings. All questions are detailed in Appendix H. EN and GER are the English and German subsets of FAME, and CS indicates the crowdsourced experiences.

Challenge	EN	GER	AMI	ICSI	WCPC	QMSum
spoken language	$3_{0.49}$	$3_{0.54}$	$\mathbf{4}_{0.22}$	$3_{0.70}$	$0.5_{0.55}$	$\mathbf{3}_{1.35}$
speaker dynamics	$2_{0.62}$	$2_{0.73}$	$\mathbf{4}_{0.66}$	$3_{0.70}$	$2_{0.63}$	$3_{0.91}$
coreference	$2_{0.74}$	$\mathbf{3}_{0.80}$	$2_{0.76}$	$2_{1.07}$	$1.5_{0.55}$	$2_{0.90}$
discourse structure	$\mathbf{3}_{1.00}$	$2_{1.20}$	$\mathbf{3}_{0.93}$	$2_{0.84}$	$\mathbf{3}_{1.17}$	$\mathbf{3}_{0.96}$
turn-taking	$\mathbf{4}_{0.56}$	$3.5_{0.51}$	$3_{0.93}$	$3_{0.67}$	$2_{0.84}$	$3_{1.04}$
implicit context	$\mathbf{4}_{0.16}$	$\mathbf{4}_{0.18}$	$0_{0.00}$	$2_{0.97}$	$0_{0.82}$	$0_{0.85}$
information density	$\mathbf{4}_{0.27}$	$\mathbf{4}_{0.00}$	$3_{0.55}$	$2_{0.57}$	$2_{0.84}$	$2.5_{0.88}$

Table 3: Scores on challenges of meeting transcripts. EN, GER are FAME English and German. Values are Median$_{Std}$. **Higher** is more difficult.

	QMSum				FAME [EN]				FAME [GER]			
	GPT	Gemini	Deepseek	Llama	GPT	Gemini	Deepseek	Llama	GPT	Gemini	Deepseek	Llama
Meeting Summary Focused Evaluation Metric (lower is better)												
Coreference	$0_{1.22}$	$3_{1.58}$	$0_{1.42}$	$1_{1.67}$	$0_{1.45}$	$3_{1.57}$	$1_{1.54}$	$2_{1.60}$	$0_{1.38}$	$3_{1.57}$	$1.5_{1.68}$	$2_{1.52}$
Hallucination	$3_{1.22}$	$4_{2.04}$	$2_{1.88}$	$3_{2.08}$	$4_{0.98}$	$4_{1.40}$	$4_{1.81}$	$4_{1.02}$	$4_{1.57}$	$4_{1.65}$	$3_{1.81}$	$4_{1.61}$
Incoherence	$4_{1.50}$	$4_{1.09}$	$4_{1.85}$	$4_{1.88}$	$4_{0.94}$	$4_{0.72}$	$4_{0.94}$	$4_{1.43}$	$4_{1.18}$	$4_{1.39}$	$3_{1.20}$	$3_{1.57}$
Irrelevance	$2_{1.70}$	$3_{1.32}$	$2_{1.48}$	$2_{1.52}$	$3_{1.14}$	$3_{1.07}$	$3_{1.11}$	$2_{1.42}$	$2_{1.16}$	$2_{1.23}$	$2_{1.29}$	$2_{1.27}$
Language	$1_{1.30}$	$2_{1.44}$	$2_{1.44}$	$1_{1.50}$	$1_{1.17}$	$1_{1.20}$	$1_{1.22}$	$1_{1.04}$	$0_{1.31}$	$0_{0.97}$	$0_{1.04}$	$1_{1.31}$
Omission	$3_{0.40}$	$3_{0.38}$	$3_{0.36}$	$3_{0.47}$	$4_{0.16}$	$4_{0.31}$	$4_{0.48}$	$4_{0.31}$	$4_{0.18}$	$4_{0.00}$	$4_{0.18}$	$4_{0.18}$
Repetition	$4_{1.05}$	$3_{0.98}$	$3_{1.01}$	$4_{1.19}$	$4_{0.74}$	$4_{0.44}$	$4_{0.39}$	$4_{0.51}$	$4_{0.52}$	$4_{0.79}$	$4_{0.34}$	$4_{0.47}$
Structure	$4_{0.90}$	$3_{1.70}$	$3_{1.65}$	$4_{1.69}$	$3_{1.57}$	$3_{1.53}$	$3_{1.52}$	$3_{1.36}$	$3_{1.66}$	$3_{1.42}$	$3_{1.50}$	$3_{1.54}$
General Evaluation Metrics (higher is better)												
R-1	$37.73_{5.85}$	$39.61_{7.21}$	$32.37_{5.18}$	$40.91_{5.21}$	$39.68_{5.73}$	$38.82_{5.79}$	$33.18_{7.38}$	$40.35_{5.46}$	$33.74_{4.01}$	$32.03_{5.14}$	$33.90_{5.37}$	$32.54_{4.50}$
R-2	$7.95_{4.18}$	$11.10_{4.86}$	$5.80_{2.94}$	$9.79_{4.08}$	$8.43_{3.45}$	$8.96_{3.50}$	$8.07_{3.33}$	$8.62_{3.18}$	$7.45_{2.98}$	$6.98_{3.72}$	$7.93_{3.67}$	$7.24_{3.17}$
R-L	$21.39_{4.05}$	$27.55_{6.36}$	$18.46_{3.44}$	$7.95_{4.18}$	$29.98_{4.88}$	$27.81_{4.18}$	$23.16_{5.98}$	$29.45_{3.67}$	$25.20_{3.68}$	$23.59_{4.32}$	$24.60_{3.29}$	$24.10_{3.11}$
BS (F1)	$61.61_{2.87}$	$60.64_{3.66}$	$59.44_{2.78}$	$62.28_{2.76}$	$63.80_{3.11}$	$63.66_{2.49}$	$63.16_{3.37}$	$63.53_{2.72}$	$64.75_{1.20}$	$64.58_{1.74}$	$64.88_{1.77}$	$64.61_{1.33}$

Table 4: Combined Evaluation Results: Human Evaluation top and Automatic Evaluation bottom. Values are Median_{Std}. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100.

5 Baseline Models and Results

We evaluate current LLMs on abstractive meeting summarization for real and synthetic transcripts, sampling 30 meetings from each FAME (English, German) and QMSum (90 total).

5.1 Experimental Setup

Summarization approaches. We benchmark two closed-weight models (GPT, Gemini) and two open-weight models (Llama, DeepSeek), excluding refinement-based methods (Kirstein et al., 2025c), as these would produce self-refined GPT summaries. We use a simple zero-shot prompt requesting an abstract summary of up to 250 tokens, reflecting standard practices in meeting summarization (Kirstein et al., 2025c). Full prompt details are provided in Appendix H.

Evaluation metrics. We compare system outputs against reference summaries using the established ROUGE (R-1/R-2/R-L) (Lin, 2004) and BERTScore (rescaled F1) (Zhang et al., 2020) metrics along MESA (Kirstein et al., 2025b), an LLM-based metric for the error types of meeting summarization (e.g., structure, irrelevance). These metrics enable direct comparisons with prior work and provide detailed insights into model weaknesses.

5.2 Results

Table 4 contains the evaluation scores.

Reasoning boosts performance. Although Llama tops the ROUGE/BERTScore metrics on both QMSum and FAME, MESA shows that DeepSeek consistently matches or improves over other models by minimizing common error types (∼1 point lower per category), especially on QMSum. DeepSeek also outperforms on the German FAME subset, while Gemini generally trails behind. All models perform comparably on the English subset, with the least language and coreference issues (∼ 1/5).

FAME reveals LLM struggles. MESA scores for categories like incoherence, structure, and repetition remain similar to those on QMSum (∼ 4/5) but with lower deviation for the FAME subsets. We conclude that the varied topics and meeting formats in FAME add to the overall difficulty and negatively influence summary quality.

Contextualization deficits persist. The omission and irrelevance rows show that all models struggle with FAME's more difficult cross-turn information and information scarcity observed in Section 4.4. Omission rises from 3/5 on QMSum to 4/5 on FAME, and irrelevance from 2/5 to 3/5 for the English subset (German remains 2/5). We derive that reliable content understanding (Kirstein et al., 2024a, 2025c) can be tested with our FAME and that current LLMs struggle with this.

6 Ablations on MIMIC

Our ablations analyze MIMIC under different settings and assess its role consistency. Below, we briefly summarize the experiments; full details, tables, and figures appear in Appendix G. All evaluations use the same model settings, metrics, and annotation guidelines from Sections 4 and 5.

Knowledge source shapes discussion depth and structure. Beyond the semi-structured Wikipedia articles used for FAME, we test MIMIC on 10 research papers from arXiv (Cohan et al., 2018) (clearly sectioned, e.g., abstract, methods, results), 10 human-written stories from Writing-Prompts (Fan et al., 2018) (unstructured) and 10 research papers from PubMed (Xiong et al., 2024) (clearly structured, niche domain knowledge). Using GPT as the backbone, we generate one synthetic meeting per source (30 total) and apply our evaluation framework (Section 4). As shown in Table 10 (overall authenticity) and Figure 5 (behavior authenticity) in Appendix G.3, quantitative metrics remain stable, suggesting that meeting naturalness is not strongly tied to input structure. Research papers lead to deeper discussions, while short stories produce briefer, shallow meetings.

Other models yield shorter meetings. To see if other LLMs can drive MIMIC, we replace GPT with DeepSeek and Llama, generating 25 meetings per model using the FAME Wikipedia article pool. We assess quality (overall authenticity, behavior authenticity, challenges) and compare outputs using identical knowledge sources. The results are given in Table 11 (overall authenticity) and Figure 6 (behavior authenticity) in Appendix G.4. Although these models produce transcripts with about 50 fewer turns than FAME, their naturalness remains high (4/5), and they replicate participant roles and behaviors almost as consistently as GPT, close to human expectations. Occasionally, transcripts feature fewer back-and-forth exchanges but remain high-quality in qualitative reviews, thereby broadening MIMIC's real-world applicability.

Editing is model-dependent but rarely critical. Stage 7 (editing) addresses issues such as formal phrasing or repeated filler words. To measure its influence, we review 75 transcripts (25 each from GPT, DeepSeek, and Llama), examining chain-of-thought logs of stage 7 and final transcripts. Only *1 in 25* GPT transcripts require major edits to mask synthetic traits, rising to *2 in 25* for the Llama-based models[7]. Minor refinements correct model-specific wording or repeated transitions, indicating that MIMIC already yields coherent discussions while stage 7 polishes for full realism.

[7]FAME contains a flag if refinement is due to a pipeline issue or a model-specific artifact along the feedback from the editorial stage.

Roles and behaviors are reliably enacted. Drawing on Serapio-García et al. (2023), we evaluate whether participants preserve assigned roles and behaviors. We select 30 participants from the newly generated meetings (10 from GPT, Gemini, DeepSeek, and Llama) and ask our six annotators to judge each turn's alignment with predefined behavior (e.g., blocker). We cover 100 scenes and 400 simulated participants in total. More than 95% of GPT turns match their roles, dropping slightly to 92% for the other two[8]. To cross-check, we sample 50 participant profiles and prompt GPT enacting these profiles to answer the TREO questionnaire (Mathieu et al., 2015) containing 48 1–5 Likert-scored questions to determine someone's role in a group. We find 48 of the profiles consistent with the assigned behaviors[9]. We conclude that MIMIC enforces coherent persona dynamics for social simulation (Zhou et al., 2024b).

Fine-tuning on FAME improves QMSum performance. Fine-tuning LLAMA-3.2-3B on 100 synthetic FAME meetings and evaluating on the 35-meeting QMSUM test split produces clear transfer gains. Compared with the non-fine-tuned model, fine-tuning on FAME lowers MESA errors, and, against an equally sized QMSum fine-tuned checkpoint, further reduces *coreference*, *incoherence*, and *language* errors while trailing slightly in *irrelevance* and *structure*. ROUGE and BERTScore differ by no more than two points between the two fine-tuned variants, and manual inspection confirms that fine-tuning on FAME yields the most coherent, information-dense summaries. These results position FAME as a practical proxy when only limited labelled meeting data are available.

7 Final Considerations

We introduced MIMIC, a seven-stage multi-agent framework that uses psychologically grounded, non-omniscient LLMs to generate source-grounded meeting transcripts. MIMIC generated FAME, a multilingual corpus of 500 English and 300 German meetings on 28 domains and 14 meeting types. Human assessments showed that FAME closely mirrors real meetings (4.5/5 in naturalness) and amplified low-information density (4/5). Comparisons with real meetings suggest that FAME captures authentic group dynamics, while our ablation studies highlighted how varying knowledge sources

[8]Annotation will be extended and made available later.

[9]All responses are provided as per Appendix A.

and backbone models shape transcript quality and demonstrate the GPT's reliable role enactment.

FAME and its detailed annotations open new directions for meeting summarization, from multilingual model development to re-introducing fine-tuning for LLMs and reinforcement learning to address persistent shortcomings (Section 5). By releasing MIMIC as open-source, we provide a powerful toolkit that researchers can adapt to low-resource languages, diverse domains, and a range of conversational styles. The framework's psychology-based behavior definitions and evaluation methodology bring a higher level of realism into synthetic conversations, enabling deeper investigations of social dynamics. Our work bridges a data gap through human-like meeting simulations that foster advances in summarization, conversational AI, social simulation, and beyond.

Limitations

The quality of generated meetings partly depends on the underlying LLM's capabilities and biases. Models with smaller context windows or different linguistic styles may produce less coherent dialogues or more frequent artifacts. Nevertheless, our ablation study shows that even mid-scale LLMs (e.g., Llama 3.3) can produce high-quality transcripts, aided by our feedback loops and refinement stages, to address major flaws. Although FAME covers seven broad Wikipedia domains and features a German subset, it does not encompass all real-world meeting types or cultural nuances (e.g., corporate etiquette, cross-cultural communication). While specialized domains (e.g., medical conferences) remain outside our scope, our framework can be easily adapted to additional knowledge sources, as evidenced by our tests with short stories and research papers (see Section 6).

A small portion of transcripts shows recurring phrases (e.g., " That's an excellent point") or overly polite tones that may hint at synthetic origins. Our multi-stage post-production phase detects and revises these repeated patterns, minimizing mechanical politeness and introducing more diverse expressions. In practice, only 1 out of 30 transcripts required major edits, suggesting that the remaining artifacts do not substantially undermine FAME's overall realism.

Acknowledgements

This work was supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation.

References

Robert F Bales, K Krippendorff, and MA Bock. 2009. Interaction process analysis. *The Content Analysis Reader*, pages 75–83.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302. Association for Computational Linguistics.

Kenneth D. Benne and Paul Sheats. 1948. Functional roles of group members. *Journal of Social Issues*, 4(2):41–49.

Jiaao Chen and Diyi Yang. 2021. Simple conversational data augmentation for semi-supervised abstractive dialogue summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6605–6616. Association for Computational Linguistics.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. PLACES: Prompting Language Models for Social Conversation Synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868, Dubrovnik, Croatia. Association for Computational Linguistics.

Minje Choi, Luca Maria Aiello, Krisztián Zsolt Varga, and Daniele Quercia. 2020. Ten Social Dimensions of Conversations and Relationships. In *Proceedings of The Web Conference 2020*, WWW '20, pages 1514–1525, New York, NY, USA. Association for Computing Machinery.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621. Association for Computational Linguistics.

DeepSeek-AI and Zhongyu Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *Preprint*, arXiv:2501.12948.

Peter F Drucker. 1967. The effective executive.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *ICML'24*, pages 11733–11763, Vienna, Austria. JMLR.org.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Tianwei She, Yuang Jiang, and Irene Li. 2024. Evaluating Large Language Models on Wikipedia-Style Survey Generation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5405–5418, Bangkok, Thailand. Association for Computational Linguistics.

Google. 2024. Notebook LM. https://notebooklm.google/.

Yebowen Hu, Timothy Ganter, Hanieh Deilamsalehy, Franck Dernoncourt, Hassan Foroosh, and Fei Liu. 2023. MeetingBank: A Benchmark Dataset for Meeting Summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16409–16423, Toronto, Canada. Association for Computational Linguistics.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters. 2003. The ICSI Meeting Corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, volume 1, pages I–I.

Frederic Kirstein, Terry Ruas, Robert Kratel, and Bela Gipp. 2024a. Tell me what I need to know: Exploring LLM-based (Personalized) Abstractive Multi-Source Meeting Summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 920–939, Miami, Florida, US. Association for Computational Linguistics.

Frederic Kirstein, Jan Philip Wahle, Bela Gipp, and Terry Ruas. 2025a. CADS: A Systematic Literature Review on the Challenges of Abstractive Dialogue Summarization. *J. Artif. Int. Res.*, 82.

Frederic Kirstein, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2024b. What‘s under the hood: Investigating Automatic Metrics on Meeting Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6709–6723, Miami, Florida, USA. Association for Computational Linguistics.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025b. Is my Meeting Summary Good? Estimating Quality with a Multi-LLM Evaluator. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 561–574, Abu Dhabi, UAE. Association for Computational Linguistics.

Frederic Thomas Kirstein, Terry Lima Ruas, and Bela Gipp. 2025c. What‘s Wrong? Refining Meeting Summaries with LLM Feedback. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2100–2120, Abu Dhabi, UAE. Association for Computational Linguistics.

Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, 77(6):1121.

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, Singapore. Association for Computational Linguistics.

Zitao Li, Fei Wei, Yuexiang Xie, Dawei Gao, Weirui Kuang, Zhijian Ma, Bingchen Qian, Yaliang Li, and Bolin Ding. 2025. KIMAs: A Configurable Knowledge Integrated Multi-Agent System. *Preprint*, arXiv:2502.09596.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. 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.

Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, pages 47–58, Toronto, Canada. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Yongtai Liu, Joshua Maynez, Gonçalo Simões, and Shashi Narayan. 2022. Data augmentation for low-resource dialogue summarization. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 703–710. Association for Computational Linguistics.

AI @ Meta Llama Team. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. SELF-REFINE: Iterative refinement with self-feedback. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 46534–46594, Red Hook, NY, USA. Curran Associates Inc.

John E. Mathieu, Scott I. Tannenbaum, Michael R. Kukenberger, Jamie S. Donsbach, and George M. Alliger. 2015. Team Role Experience and Orientation: A Measure and Tests of Construct Validity. *Group & Organization Management*, 40(1):6–34.

Iain Mccowan, J Carletta, Wessel Kraaij, Simone Ashby, S Bourban, M Flynn, M Guillemot, Thomas Hain, J Kadlec, V Karaiskos, M Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska Masson, Wilfried Post, Dennis Reidsma, and P Wellner. 2005. The AMI meeting corpus. *Int'l. Conf. on Methods and Techniques in Behavioral Research*.

Anna Nedoluzhko, Muskaan Singh, Marie Hledíková, Tirthankar Ghosal, and Ondřej Bojar. 2022. ELITR Minuting Corpus: A Novel Dataset for Automatic Minuting from Multi-Party Meetings in English and Czech. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3174–3182, Marseille, France. European Language Resources Association.

Ryota Nonomura and Hiroki Mori. 2024. Who Speaks Next? Multi-party AI Discussion Leveraging the Systematics of Turn-taking in Murder Mystery Games. *Preprint*, arXiv:2412.04937.

Nvidia. 2025. PDF-to-Podcast. https://build.nvidia.com/nvidia/pdf-to-podcast.

OpenAI. 2024. GPT-4 Technical Report. *Preprint*, arXiv:2303.08774.

Alex F. Osborn. 1953. *Applied Imagination*. Applied Imagination. Scribner'S, Oxford, England.

Stefano De Paoli. 2023. Can Large Language Models emulate an inductive Thematic Analysis of semi-structured interviews? An exploration and provocation on the limits of the approach and the model. *Preprint*, arXiv:2305.13014.

Seongmin Park, Dongchan Shin, and Jihwa Lee. 2022. Leveraging non-dialogue summaries for dialogue summarization. In *Proceedings of the First Workshop on Transcript Understanding*, pages 1–7. International Conference on Computational Linguistics.

Ganbin Qiu and Yiteng Pan. 2024. Enhancing Meeting Summarization Models via Large Language Model Data Generation and Weighted Sampling. In *2024 5th International Conference on Computer Engineering and Intelligent Control (ICCEIC)*, pages 102–106.

Virgile Rennard, Guokan Shang, Damien Grari, Julie Hunter, and Michalis Vazirgiannis. 2023. FREDSum: A Dialogue Summarization Corpus for French Political Debates. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4241–4253, Singapore. Association for Computational Linguistics.

Greg Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2023. Personality Traits in Large Language Models. *Preprint*, arXiv:2307.00184.

Herbert A Simon. 2013. *Administrative Behavior*. Simon and Schuster.

Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, and Timothy Lillicrap. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

David Thulke, Yingbo Gao, Rricha Jalota, Christian Dugast, and Hermann Ney. 2024. Prompting and Fine-Tuning of Small LLMs for Length-Controllable Telephone Call Summarization. *Preprint*, arXiv:2410.18624.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the Bounds of LLM Reasoning: Are Multi-Agent Discussions the Key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Yiqing Xie, Wenxuan Zhou, Pradyot Prakash, Di Jin, Yuning Mao, Quintin Fettes, Arya Talebzadeh, Sinong Wang, Han Fang, Carolyn Rose, Daniel Fried, and Hejia Zhang. 2024. Improving Model Factuality with Fine-grained Critique-based Evaluator. *Preprint*, arXiv:2410.18359.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. *arXiv preprint arXiv:2402.13178*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. *Preprint*, arXiv:1904.09675.

Ming Zhong, Da Yin, Tao Yu, Ahmad Zaidi, Mutethia Mutuma, Rahul Jha, Ahmed Hassan Awadallah, Asli Celikyilmaz, Yang Liu, Xipeng Qiu, and Dragomir Radev. 2021. QMSum: A New Benchmark for Query-based Multi-domain Meeting Summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5905–5921, Online. Association for Computational Linguistics.

Xuhui Zhou, Zhe Su, Tiwalayo Eisape, Hyunwoo Kim, and Maarten Sap. 2024a. Is this the real life? Is this just fantasy? The Misleading Success of Simulating Social Interactions With LLMs. *Preprint*, arXiv:2403.05020.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024b. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. *Preprint*, arXiv:2310.11667.

A Data Availability

The FAME dataset proposed in this paper, along with all (human and LLM-generated) annotations (including overall authenticity, behavior authenticity, and challenges), as well as the MIMIC framework and the psychology-grounded evaluation framework are available on GitHub under a CC BY-SA 4.0.

B Cost and Time

The generation of one meeting using MIMIC with a GPT backbone costs on average \$7.63 and takes ∼30 minutes. One meeting generated with a single, omniscient GPT costs \$0.22 and takes ∼3 minutes.

C Details on MIMIC Settings

In this appendix, we list the available options in MIMIC for the meeting types (Table 5) (Osborn, 1953; Drucker, 1967; Simon, 2013), psychological roles (Table 6) (Benne and Sheats, 1948), and special effects (Appendix C.1).

C.1 Possible Special Effects

- **Polite interruptions** to add a point or seek clarification.
- **Participants speaking over each other** briefly.
- **Side comments or asides** related to the main topic.
- **Brief off-topic remarks** or questions.
- **Moments of confusion** requiring clarification.
- **Laughter or reactions** to a humorous comment.
- **Time-checks or agenda reminders**.
- **Casual side comments or friendly banter**.
- **Rapid-fire idea contributions**.
- **Instructional interruptions** to provide examples.
- **Light-hearted jokes or humorous reactions**.
- **Strategic questions** about project goals.
- **Feedback requests** on presented material.
- **Technical difficulties**: Problems with audio, video, or presentation equipment (e.g., *"You're on mute."*).
- **Checking the time** or mentioning scheduling constraints.
- **Misunderstandings** that are quickly resolved.
- **External disruptions** such as phone calls, notifications, etc.

D Evaluation Framework Details

This appendix contains the definitions of the three quality measures applied in Section 4, i.e., overall authenticity (Table 7) from Chen et al. (2023), our eighteen questions on behavior authenticity (Table 8) defined from the knowledge, power, conflict, status, trust, support, similarity, and fun (Choi et al., 2020; Bales et al., 2009), and the challenges in meeting transcripts defined by Kirstein et al. (2024b).

E Human Annotations

E.1 Complete annotation process

Annotator selection: Our annotation team consisted of six graduate students, officially employed as interns or doctoral candidates through standardized contracts. We selected them from a pool of volunteers based on their availability to complete the task without time pressure and their proficiency in English and German (native speakers or C1-C2 certified). By that, we ensured they could comprehend meeting transcripts and summaries. We aimed for gender balance (three males, three females) and diverse backgrounds, resulting in a team of two computer science students, three psychology students, and one communication science student aged 23-28. The annotators consent that their annotations will be used anonymously in this work. The annotation process has been internally reviewed by an ethics committee.

Preparation: We prepared a comprehensive handbook for our annotators, detailing the project context and defining the criteria (a short version as

Meeting Type	Objectives	Expected Outcomes
Brainstorming Session	Generate a wide range of ideas Encourage creative thinking Foster a collaborative environment	List of potential ideas Prioritized concepts for further exploration
Decision-Making Meeting	Evaluate options Weigh pros and cons Reach a consensus or make a decision	Finalized decision Action items with assigned responsibilities
Problem-Solving Meeting	Identify the root cause of a problem Analyze potential solutions Implement actionable solutions	Clear understanding of the problem Viable solutions identified Action plan for implementation
Training and Workshop Session	Educate participants on new skills or knowledge Provide hands-on training Assess participant understanding	Enhanced participant skills Increased knowledge in specific areas Preparedness to apply new skills
Strategic Planning Meeting	Define long-term organizational goals Develop strategies Allocate resources effectively	Comprehensive strategic plan Defined organizational objectives Resource allocation roadmap
Committee or Board Meeting	Review and discuss policies Make governance decisions Oversee organizational performance	Approved or revised policies Governance decisions made Reviewed organizational performance
Innovation Forum	Encourage innovative thinking Explore new opportunities Foster a culture of innovation	Generated innovative ideas Identified new opportunities Enhanced culture of innovation
Agile/Scrum Meeting	Facilitate daily progress updates Plan and prioritize sprint tasks Review sprint outcomes	Daily progress transparency Well-defined sprint plans Identified process improvements
Remote or Virtual Meeting	Facilitate collaboration among remote participants Share information and updates Coordinate tasks virtually	Effective virtual collaboration Shared information and updates Coordinated tasks and projects
Project Kick-Off Meeting	Introduce project goals and objectives Define team roles and responsibilities Establish project timelines	Clear project roadmap Assigned roles and responsibilities Initial project timeline established
Stakeholder Meeting	Update stakeholders on project progress Gather stakeholder feedback Ensure alignment with expectations	Informed stakeholders Collected valuable feedback Aligned project goals with expectations
Casual Catch-Up	Build team rapport Share updates Discuss non-work-related topics	Strengthened team relationships Shared personal and professional insights
Cross-Functional Meeting	Facilitate collaboration across departments Align on shared project objectives Resolve interdepartmental issues	Aligned project objectives Resolved cross-departmental issues Enhanced interdepartmental collaboration
Retrospective Meeting	Reflect on past performance Identify successes and areas for improvement Implement process enhancements	Documented lessons learned Actionable improvement plans Enhanced future project processes

Table 5: Meeting Types and Their Descriptions (Multi-row Format)

presented in Appendix D and an extended version with more details). Each definition included two examples: one with minimal impact on quality and one with high impact. The handbook explained the 1–5 Likert rating for the individual questionnaires. The handbook did not specify an order for processing the items. We provided the handbook in English and the annotators' native languages, using professional translations.

We further elaborated a three-week timeline for the annotation process, preceded by a one-week onboarding period. The first week featured twice-weekly check-ins with annotators, which were reduced to weekly meetings for the following two weeks. Separate quality checks without the annotators were scheduled weekly. (Note: week refers to a regular working week)

Onboarding: The onboarding week was dedicated to getting to know the project and familiarization with the definitions and data. We began with a kick-off meeting to introduce the project and explain the handbook, particularly focusing on each definition. We noted initial questions to potentially revise the handbook. Annotators received ten transcripts and summaries generated by GPT (OpenAI, 2024) using MIMIC. After the first five samples, we held individual meetings to clarify any confusion and updated the guidelines accordingly.

Term	Definition
Group behavior	
Initiator-Contributor	Contributes new ideas and approaches and helps to start the conversation or steer it in a productive direction.
Information Giver	Shares relevant information, data, or research that the group needs to make informed decisions.
Information Seeker	Asks questions to gain clarity and obtain information from others.
Opinion Giver	Shares their views and beliefs on topics under discussion.
Opinion Seeker	Encourages others to share their opinions and beliefs in order to understand different perspectives.
Coordinator	Connects different ideas and suggestions of the group to ensure that all relevant aspects are integrated.
Evaluator-Critic	Analyzes and critically evaluates proposals or solutions to ensure their quality and feasibility.
Implementer	Puts plans and decisions of the group into action and ensures practical implementation.
Recorder	Documents the group's decisions, ideas, and actions in order to have a reference for future discussions.
Encourager	Provides positive feedback and praise to boost the morale and motivation of group members.
Harmonizer	Mediates conflicts and ensures that tensions in the group are reduced to promote a harmonious working environment.
Compromiser	Helps the group find a middle ground when there are differences of opinion and encourages compromise in order to move forward.
Gatekeeper	Ensures that all group members have the opportunity to express their opinions and encourages participation.
Standard Setter	Emphasizes the importance of adhering to certain norms and standards within the group to ensure quality and efficiency.
Group Observer	Monitors the dynamics of the group and provides feedback on how the group is functioning as a whole and what improvements can be made.
Follower	Supports the group by following the ideas and decisions of others without actively driving the discussions.
Individual behavior	
Aggressor	Exhibits hostile behavior, criticizes others, or attempts to undermine the contributions of others.
Blocker	Frequently opposes ideas and suggestions without offering constructive alternatives and delays the group's progress.
Recognition Seeker	Tries to draw attention to themselves by emphasizing their own successes or focusing on their own importance.
Dominator	Tries to control the group by dominating the flow of conversation and imposing their own views.
Help Seeker	Seeks sympathy or support by presenting as insecure or helpless, often to avoid responsibility.
Special Interest Pleader	Brings their own interests or concerns to the discussion that do not align with the goals of the group.

Table 6: Overview of usable behaviors defined by Benne and Sheats (1948).

Item	Description
Naturalness	How natural the conversation flows, like native English speakers (1-5)
Coherence	How well the conversation maintains logical flow and connection (1-5)
Interesting	How engaging and content-rich the conversation is (1-5)
Consistency	How consistent each speaker's contributions are (1-5)

Table 7: Overall authenticity evaluation following Section 4.

The remaining five samples were then annotated using these updated guidelines. A second group meeting this week addressed any new issues with definitions. After the group meeting, we met individually with the annotators to review their work, ensuring their quality and understanding of the task and samples. All six annotators demonstrated reliable performance and good comprehension of the task and definitions, judging from the reasoning they provided for each decision and annotation. We computed an inter-annotator agreement score using Krippendorff's alpha, achieving 0.86, indicating sufficiently high overlap.

Annotation Process: Each week, we distribute all samples generated by one model/source (on average 27 samples) to one of the annotators. Consequently, one annotator worked through all samples of one model/source in one week. On average, one annotator processes summaries from three models/sources (depending on other commitments, some annotators could only annotate two datasets, and others four or more). Three annotators annotate each sample. The annotators were unaware of the origin of the meeting transcript and summary. They were given a week to complete their set at their own pace and with their own break times. Quiet working rooms were provided if needed for concentration. To mitigate position bias, the sample order was randomized for each annotator. Annotators could choose their annotation order for each sample and were allowed to revisit previous samples.

Regular meetings were held to address any emerging issues or questions on definitions. During the quality checks performed by the authors, we looked for incomplete annotations, missing explanations, and signs of misunderstanding judging from the provided reasoning. If the authors had found such a lack of quality, the respective annotator would have been notified to re-do the annotation. After three weeks, we computed inter-annotator agreement scores on the error types (overall Krippendorff's alpha = 0.79). If we observed a significant difference among annotators, we planned a dedicated meeting with all annotators and a senior annotator to discuss such cases. On average, an-

Short Version	Category	Description
Q1: Information Exchange	Knowledge	Participants exchange information or knowledge.
Q2: Knowledge Seeking	Knowledge	Participants request or seek knowledge.
Q3: Explanation Provision	Knowledge	Participants clarify previous statements upon request.
Q4: Influence Attempts	Power	Participants attempt to influence another participant's behavior or decisions.
Q5: Topic Control	Power	Participants take control of a topic or subtopic to steer outcomes in their favor.
Q6: Power Dynamics	Power	A power dynamic exists among participants.
Q7: Response Patterns	Conflict	Participants fail to engage with others' suggestions.
Q8: Standpoint Maintenance	Power	Participants insist on their own perspective.
Q9: Recognition Expression	Status	Participants express recognition, gratitude, or admiration toward others.
Q10: Dependency Expression	Trust	Participants indicate reliance on another participant's actions or judgments.
Q11: Support Offering	Support	Participants offer help or support to others.
Q12: Shared Interests	Similarity	Participants discuss shared interests or motivations.
Q13: View Alignment	Similarity	Participants exhibit aligned views or opinions.
Q14: Mood Management	Fun	Participants attempt to lighten the atmosphere.
Q15: Social Interaction	Fun	Participants discuss leisure activities or enjoyable moments.
Q16: Opinion Divergence	Conflict	Participants express divergent opinions.
Q17: Conflict Presence	Conflict	Conflicts or tensions emerge among participants.
Q18: Discussion Dynamics	Conflict	Participants engage in discussions about disagreements, topics, or decisions.

Table 8: Psychology-grounded Framework to evaluate participant behavior.

notators spent 44 minutes per sample, completing about six samples daily.

Handling of unexpected cases: Given that our annotators had other commitments, we anticipated potential scheduling conflicts. We allowed flexibility for annotators to complete their samples beyond the week limit if needed, reserving a fourth week as a buffer. Despite these provisions, all annotators completed their assigned samples within the original weekly timeframes. We further allowed faster annotators to continue with an additional sample set. This additional work was voluntary.

E.2 Crowdworkers

The crowdworkers comprised 100 employees (officially employed through standardized contracts) with diverse backgrounds, including psychology, law, business administration, physics, and design. We selected them from a pool of volunteers and ensured that the crowdworkers did not overlap with the annotators. We aimed for gender balance (57 male, 43 female), covering ages 24 to 63. The crowdworkers consent that their answers (Likert scores) will be used anonymously in this work. The crowdsourcing has been internally reviewed by an ethics committee. The crowdworkers were given the behavior questionnaire in Table 8 (Appendix D) in their native language with the task of answering the statements according to their general experience with meetings. The 1–5 Likert scale rating scheme was initially explained to them with an example. In case of unclarities, these were directly resolved. The items were answered on a Laptop with an average time of 12 minutes.

F Omniscient Model Behavior

In Figure 4, we show the participant behavior pattern of the omniscient single model compared to FAME simulated participant behavior patterns. We observe that the patterns rarely match, and the single model struggles to reflect topic control, standpoint maintenance, or support offering. We conclude that the single model omniscient approach cannot simulate group behaviors directly.

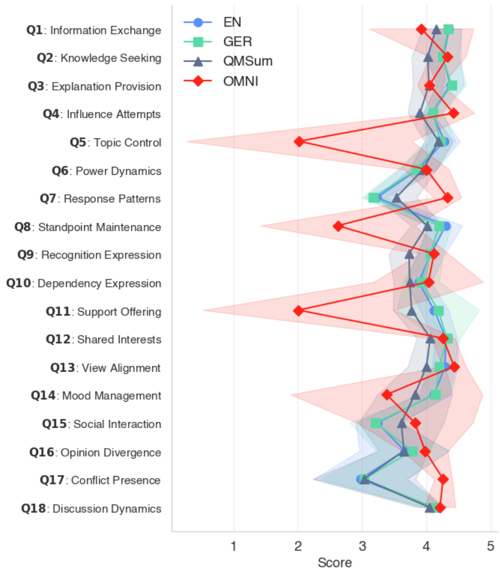

Figure 4: Behavior pattern of FAME's English (EN) and German (GER) subsets, QMSum, and a single omniscient model (OMNI).

G Ablation Experiments

This appendix provides additional information on our ablation experiments (Section 6), elaborating on the models, evaluation procedures, annotator

Category	Definition	Instructions
Spoken language	The extent to which the transcript exhibits spoken-language features—such as colloquialisms, jargon, false starts, or filler words—that make it harder to parse or summarize.	1. Are there noticeable filler words, false starts, or informal expressions? 2. Does domain-specific jargon disrupt straightforward summarization? 3. How challenging are these elements for generating a coherent summary?
Speaker dynamics	The challenge of correctly identifying and distinguishing between multiple speakers, tracking who said what, and maintaining awareness of speaker roles if relevant.	1. Is it difficult to keep track of speaker identities or roles? 2. How significantly do these dynamics affect clarity for summarization?
Coreference	The difficulty in resolving references (e.g., who or what a pronoun refers to) or clarifying references to previous actions or decisions, so the summary remains coherent.	1. Are references (e.g., pronouns like "he" or "that") ambiguous? 2. Do unclear references to earlier points or events appear? 3. How difficult is it to track these references for summary generation?
Discourse structure	The complexity of following the meeting's high-level flow—especially if it changes topics or has multiple phases (planning, debate, decision).	1. Does the transcript jump between topics or phases without clear transitions? 2. Are meeting phases or topical shifts difficult to delineate? 3. How challenging is it to maintain an overview for the summary?
Contextual turn-taking	The challenge of interpreting local context as speakers take turns, including interruptions, redundancies, and how each turn depends on previous utterances.	1. Do abrupt speaker turns or interjections complicate continuity? 2. Are important points lost or repeated inconsistently? 3. How difficult is it to integrate these nuances into a coherent summary?
Implicit context	The reliance on unspoken or assumed knowledge, such as organizational history, known issues, or prior decisions, only vaguely referenced in the meeting.	1. Do participants refer to hidden topics or internal knowledge without explaining? 2. Is there essential background or context missing? 3. How much does summarization rely on understanding this hidden context?
Low information density	Segments where salient information is sparse, repeated, or only occasionally surfaced—making it hard to find and isolate key points in a sea of less relevant content.	1. Are there long stretches with minimal new information? 2. Is meaningful content buried under trivial or repetitive remarks? 3. How challenging is it to isolate crucial points for the summary?

Table 9: Summary challenges from Kirstein et al. (2024b) and their evaluation instructions.

settings, and preliminary results. Unless otherwise stated, all evaluations follow the methodologies described in Section 4 and Section 5.

G.1 Experimental Setup and Models

Framework. We use the same seven-stage MIMIC approach (pre-production, production, post-production) outlined in the main paper to generate synthetic meetings.

Models. We primarily employ the GPT-4o model as a reference. In addition, we evaluate three LLMs, i.e., Gemini 1.5 pro, DeepSeek-R1 Llama Distill 70B, Llama 3.3 70B. Each model receives identical prompts at every MIMIC stage. Hyperparameters (e.g., temperature, top-p) remain at the defaults recommended in each model's documentation to maintain comparability.

Data Sources. In addition to the semi-structured Wikipedia articles that power FAME, we also experiment with:

- **Research Papers:** 10 arXiv papers (Cohan et al., 2018) with conventional sections (abstract, methods, results, conclusion).
- **Stories:** 10 human-written short stories from WritingPrompts, each at least 500 tokens and lacking formal structure (Fan et al., 2018).

G.2 Evaluation Approach

We use the same metrics and annotation protocols described in Section 4 and Section 5, summarized below:

- **Overall Authenticity:** Following Chen et al. (2023), we measure coherence, consistency, interestingness, and naturalness (1–5 Likert scale).
- **Behavior Authenticity:** Building on psychology/sociology literature (Choi et al., 2020; Bales et al., 2009) and Section 4, we consider eight overarching categories (e.g., knowledge sharing, conflict, power) subdivided into 18 items.
- **Challenge Scores:** Adopted from Kirstein et al. (2024b), focusing on complexities like spoken language, speaker dynamics, coreference, implicit context, and low information density.
- **Manual Qualitative Checks:** For each experiment, we sample transcripts to review dialogue flow, persona behavior, chain-of-thought traces from Stages 5 (Quality Assuring) and 7 (Editing), plus any repetitive patterns or artifacts that might reveal synthetic origins.

All annotation tasks were performed by six raters (backgrounds in psychology, computer science, communication). They used the labeling guidelines introduced in Section 5, holding weekly calibration sessions to maintain high inter-annotator agreement (Krippendorff's $\alpha > 0.80$). Additional annotator details are in Appendix E.

G.3 Knowledge Source Shapes Discussion Depth and Structure

Goal. We aim to discover whether the level of structure in source texts (structured vs. unstructured) and the level of niche (known concepts vs. highly specialized) significantly impacts the generated meetings' authenticity and behavioral patterns.

Setup. Using GPT as the backbone, we generated 30 synthetic meetings, one for each knowledge source (10 research papers from arXiv, 10 research papers from PubMed, 10 short stories). To ensure diversity, each transcript was configured to one of three meeting types (e.g., decision-making, brainstorming, innovation forum). We then applied our evaluation framework (Section 4), measuring overall authenticity, behavior authenticity, and challenge scores, supplemented by a qualitative analysis of discussion depth.

	FAME EN	arXiv	Stories	PubMed
COH	**4.5**$_{0.00}$	4.0$_{0.38}$	4.0$_{0.46}$	**4.5**$_{0.56}$
CON	4.5$_{0.07}$	4.5$_{0.05}$	4.5$_{0.08}$	4.5$_{0.05}$
INT	**4.5**$_{0.13}$	**4.5**$_{0.54}$	4$_{0.86}$	**4.5**$_{0.27}$
NAT	**4.5**$_{0.12}$	4$_{0.97}$	**4.5**$_{0.73}$	**4.5**$_{0.11}$

Table 10: Average authenticity scores for structured (research papers) vs. unstructured (short stories). FAME values taken from Section 4. Values are Median$_{Std}$. **Higher** is better.

Results. Table 10 and Figure 5 show that meetings derived from research papers often feature deeper discussions and smoother scene transitions, attributed to the documents' well-defined sections. In contrast, short-story inputs yield shorter and more tangentially structured meetings. Quantitative metrics (e.g., overall authenticity) remain close to those for Wikipedia-based transcripts, implying that meeting *naturalness* does not heavily depend on the input's structural clarity.

G.4 Mid-Size Backbone Models for MIMIC

Goal. We investigate whether MIMIC maintains comparable transcript quality when using other LLMs rather than GPT.

Setup. We replaced GPT with DeepSeek and Llama and randomly picked 25 Wikipedia articles from the FAME pool. Each article was used to generate one synthetic meeting per model (75 total). We then ran the same evaluation process, collecting both quantitative metrics (authenticity, behavior, challenge scores) and qualitative feedback on dialogue flow and realism.

Results. Table 11 shows that other models produce an average of 50 fewer turns per transcript compared to GPT, yet maintain near-equal naturalness (4/5 vs. 4.5/5). Annotators report that roles and psychological behaviors (Figure 6) are faithfully preserved, though about one-third of the transcripts have fewer back-and-forth exchanges. Despite this, reviewers found the outputs coherent and realistic.

G.5 Analysis of Editing Stage (Stage 7) Influence

Goal. We evaluate the necessity of Stage 7 (editing/refinement) in removing synthetic cues or repetitive language to achieve realistic transcripts.

Setup. We manually audited 75 transcripts generated during the mid-size model experiment: 25

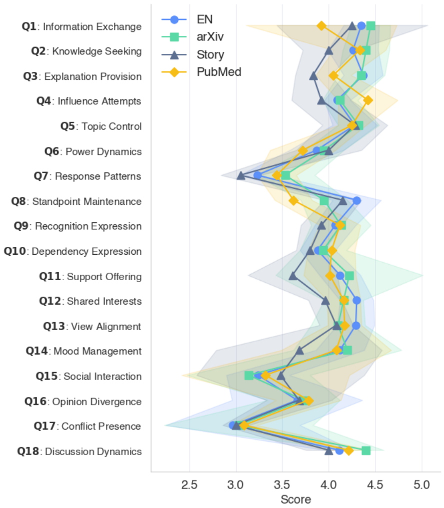

Figure 5: Behavior authenticity comparison for different knowledge sources.

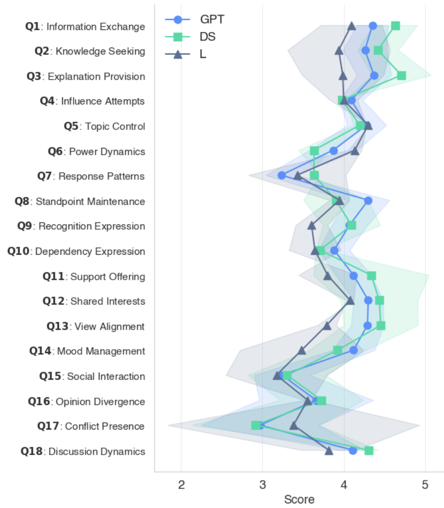

Figure 6: Behavior authenticity for GPT, DeepSeek (DS) and Llama (L) backbone models.

	GPT	DeepSeek	Llama
COH	$4.5_{0.00}$	$4.5_{0.15}$	$4.5_{0.18}$
CON	$4.5_{0.07}$	$4.5_{0.04}$	$4.5_{0.03}$
INT	$\mathbf{4.5}_{0.13}$	$\mathbf{4.5}_{0.94}$	$4_{0.37}$
NAT	$\mathbf{4.5}_{0.12}$	$\mathbf{4.5}_{0.84}$	$4_{0.71}$

Table 11: Average authenticity scores for DeepSeek and Llama vs. GPT-4o. GPT-4o values taken from Section 4. Values are Median$_{Std}$. **Higher** is better.

each from GPT, DeepSeek, and Llama. Reviewers compared chain-of-thought logs at Stage 7 with final transcripts to identify whether refinements corrected pipeline-wide issues (e.g., missing subtopics) or model-specific artifacts (e.g., repeated filler phrases).

Results. We observed that only *1 in 25* GPT transcripts needed major edits to hide synthetic traits, rising to *2 in 25* for mid-scale models. Most refinements were minor vocabulary adjustments or removal of repetitive transition phrases. Thus, MIMIC inherently produces coherent multi-agent dialogues, with Stage 7 serving primarily as a final polish to further mask synthetic cues.

G.6 Roles and Behaviors Are Reliably Enacted

Goal. We test whether participants adhere to their assigned roles (e.g., project manager) and psychological behaviors (e.g., conflict aversion, leadership) throughout a meeting.

Setup. We sampled 30 newly generated meetings, 10 each from GPT, Gemini, DeepSeek, and Llama, and had six annotators evaluate each turn for alignment with assigned persona traits (e.g., status consciousness, creative thinking). This encompassed 100 scenes and 400 participants. Inspired by Serapio-García et al. (2023), we further deployed the TREO questionnaire (Mathieu et al., 2015), containing 48 Likert-scored questions, to see whether a prompted model persona would self-report consistently with its designated role.

Results. Over 90% of GPT's turns (and 87% from DeepSeek/Llama) aligned with the assigned persona. In 48 of 50 TREO questionnaires,[10] the questionnaire results matched the participant's predefined behavior. One notable exception is the "Blocker" role, which GPT partially avoided due to its tendency toward supportive language, causing a small mismatch in self-reported behaviors. Overall, these findings confirm that MIMIC enforces coherent persona dynamics even when smaller LLMs serve as the backbone.

G.7 Fine-tuning on FAME transfers to QMSUM

Goal. We investigate whether a model fine-tuned to synthetic FAME meetings improves QMSUM summarization.

Setup. Starting from the LLAMA–3.2–3B backbone, we apply LoRA fine-tuning on 100 randomly sampled FAME dialogues for five epochs (learning rate 2×10^{-4}, $\alpha = 128$, rank $r = 256$, dropout 0.05), yielding checkpoint `FT-FAME`. As

[10] All participant responses are provided as per Appendix A.

	No Finetuning	Finetuning FAME	Finetuning QMSum
MESA (lower is better)			
Coreference	$2.5_{1.53}$	$0_{0.99}$	$1.5_{1.53}$
Hallucination	$5_{2.26}$	$5_{1.53}$	$5_{1.98}$
Incoherence	$4_{1.27}$	$3_{1.76}$	$3.5_{1.79}$
Irrelevance	$2_{1.48}$	$2_{1.83}$	$1.5_{1.70}$
Language	$2_{1.39}$	$1_{1.64}$	$2_{1.42}$
Omission	$4_{0.22}$	$4_{0.22}$	$4_{0.06}$
Repetition	$3_{0.64}$	$3_{0.86}$	$3_{1.19}$
Structure	$3_{1.70}$	$3_{1.63}$	$2.5_{1.41}$
General Evaluation Metrics (higher is better)			
R-1	$34.10_{6.31}$	$39.92_{5.74}$	$40.75_{4.71}$
R-2	$8.54_{4.18}$	$14.96_{2.22}$	$16.25_{2.70}$
R-L	$20.55_{3.67}$	$25.93_{2.58}$	$26.42_{4.67}$
BS (F1)	$58.86_{3.07}$	$67.54_{2.67}$	$68.77_{1.64}$

Table 12: Evaluation Results of Llama-3.2 3b with different fine-tuning. Values are Median$_{Std}$. MESA scores are 1–5 Likert ratings, ROUGE (R-1/R-2/R-L) and BERTScore (BS) are 0–100.

baselines, we use the not fine-tuned backbone (`NoFT`) and an identically configured LoRA fine-tuning on 100 samples from the QMSum training set (`FT-QMSum`). We evaluate the three models on the 35-meeting QMSum test split using MESA, ROUGE, and BERTScore, following the setup detailed in Section 5.

Results. Both fine-tuned checkpoints outperform the non-fine-tuned model on every automatic metric (Table 12). Relative to `FT-QMSum`, `FT-FAME` obtains lower (better) MESA scores for *coreference* (0 vs. 1.5), *incoherence* (3 vs. 3.5), and *language* (1 vs. 2), while lagging slightly in *irrelevance* (2 vs. 1.5) and *structure* (3 vs. 2.5). ROUGE and BERTScore of the fine-tuned checkpoints differ by at most two points, indicating that the two data sources confer complementary benefits. These results suggest that FAME can serve as a practical proxy for fine-tuning.

G.8 Summary of Ablation Findings

In summary, the ablation experiments demonstrate that:

- **Knowledge Source:** Meeting *naturalness* remains stable across structured (research papers) and unstructured (stories) inputs, though structured texts yield deeper, more cohesive scene discussions.

- **Backbone Models:** Mid-scale LLMs (Gemini, DeepSeek, Llama) produce slightly shorter transcripts but maintain strong authenticity and consistent role behaviors, indicating MIMIC's flexibility across model scales.

- **Editing Stage:** Stage 7 is rarely essential for correctness but provides a valuable final polish, mitigating repeated language and formal tones that might expose synthetic origins.

- **Role Consistency:** Participants consistently adhere to their assigned personas and behaviors, as evidenced by both manual turn-based checks and the TREO questionnaire alignment.

- **Cross-Domain Transfer:** Fine-tuning on 100 synthetic meetings cuts coreference, incoherence, and language errors by up to 1.5 MESA points while matching in-domain training on all remaining metrics.

Collectively, these findings underscore MIMIC's robustness: it adapts to diverse inputs, model scales, and persona assignments while preserving conversational quality. Further details, including complete transcripts, chain-of-thought logs, and extended annotations, are available as per Appendix A.

H MIMIC Prompts

In this appendix, we present the central prompts of MIMIC in detail. These include:

- **Stage 1 (Content Brainstorming):** Target summary generation (Figure 7) and article tag generation (Figure 8).
- **Stage 2 (Casting):** Meeting participant generation (Figures 9 and 10), speaking style definition (Figures 11 and 12) and behavior assignment (Figures 13 and 14).
- **Stage 3 (Scripting):** Meeting planning (Figure 15).
- **Stage 4 (Filming):** Starting participant selection (Figure 16) and conversation (Figures 17 and 18).
- **Stage 5 (Quality assuring):** LLM-judge director (Figure 19).
- **Stage 6 (Special effects):** Special effects injection (Figure 20).
- **Stage 7 (Editing):** Editorial refinement (Figures 21 and 22), AI content detection (Figure 23), and humanizing (Figure 24).

I Examples for MIMIC Stages

This appendix shows intermediate results of MIMIC with a GPT backbone generating a meeting transcript from the "Pandemics" Wikipedia article.

Given this input to Stage 1: Content Brainstorming, the pipeline extracts the topics of the article and generates a target summary (Figure 25). Informed about the summary, the pipeline defines during Stage 2: Casting a set of participants such as in Figure 26 and further extends the summary into a meeting outline (Figure 27) during Stage 3: Scripting. The participants are then orchestrated to discuss the points on the outline turn-by-turn, producing one scene per topic such as Figure 28. This raw scene undergoes a refinement step during Stage 5: Quality assuring from the director LLM which provides a thorough feedback (Figure 29). If the director LLM approves a scene, the pipeline may inject, with a probability of 25%, a special effect into the scene (Figure 30) during Stage 6: Special effects. Finally, during Stage 7: Editing, the scene is assessed by an editorial LLM and a detector-revision LLM to polish AI content and generate the final scene (Figure 31). After all scenes undergo this procedure, the whole meeting is generated.

Target Summary Generation

You are a professional meeting summarizer, drawing inspiration from the QMSUM dataset's organized and concise style.
Your task is to summarize a Wikipedia article as if the various facts in the article were discussed in a meeting, now being summarized for participants or readers.
The summary should:
1. **Reflect a 'Meeting Summary' Style**: Adopt a systematic structure, clearly presenting main points, relevant decisions, and/or action items.
2. **Remain Concise Yet Sufficiently Detailed**: Aim for brevity but do not omit crucial details needed to understand the discussion.
3. **Stay True to the Article**: Ensure accuracy by covering the principal topics while preserving the meeting context.
4. **Match language-speaking Conventions**: Generate the summary in language, mirroring the phrasing and cultural norms typical of real meetings in that language.
Follow these rules:
Structural Requirements:
1. **Opening**: Start with the meeting's primary objective or central topic (e.g., 'The meeting focused on standardizing...').
2. **Flow**: Group related points into logical sequences (e.g., proposals $\rightarrow$ concerns $\rightarrow$ resolutions).
3. **Decisions/Actions**: Conclude each topic with clear outcomes (e.g., 'agreed to explore alternatives').
4. **Paragraphs**: Use 1-2 dense paragraphs without section headers, bullets, or lists.
Language Requirements:
- **Avoid**: Phrases like 'we discussed,' 'the meeting covered,' or 'participants mentioned.'
- **Use Direct Language**: Frame points as decisions or facts (e.g., 'The team proposed...' instead of 'They talked about...').
- **Tense**: Use past tense and passive voice where appropriate (e.g., 'It was agreed...').
- **Concision**: Omit filler words (e.g., 'then,' 'next').
Below are several example meeting summaries illustrating the level of clarity, organization, and balance between detail and concision:
Examples of QMSUM Style Summaries (Note Structure & Tone):

> Meeting participants wanted to agree upon a standard database to link up different components of the transcripts. The current idea was to use an XML script, but it quickly seemed that other options, ...
> The meeting discussed the progress of the transcription, the DARPA demos, tools to ensure meeting data quality, data standardization, backup tools, and collecting tangential meeting information. The ...

These examples demonstrate an orderly, concise approach. Summarize the Wikipedia article **strictly** as a QMSUM-style meeting summary — presenting the main topics, relevant decisions, key points of contention, and concluding remarks in cohesive paragraph(s) **without using bullet points**.
Generate an abstractive summary with **at most num_words words** in **language**. Ensure it is systematically organized and remains consistent with the meeting type: **meeting_type**.

Your Task:
Meeting Type: **meeting_type**, Article Title: **article_title**, Content: **content**.
Now generate an abstractive meeting summary in **language**.

Figure 7: Prompt template for generating meeting-style summaries based on Wikipedia articles, designed to align with QMSUM conventions.

Article Tags

You are a Wikipedia Editor tasked with assigning five highly relevant and specific tags to a given Wikipedia article.
The tags should accurately reflect the main topics, themes, and subjects covered in the article.

User Input:
Here is the Wikipedia article. Only return a Python list of strings including the five most relevant tags for the specified article, reflecting the main topics, themes, and subjects covered in it.

Wikipedia Article: `< article >`

Output Format:
`['tag1', 'tag2', 'tag3', 'tag4', 'tag5']`

Ensure that the list contains exactly five concise, meaningful tags without additional text or formatting.

Figure 8: Prompt template for extracting five relevant tags for a Wikipedia article.

Generate Meeting Participants - Part 1

When faced with a task, begin by identifying the participants who will contribute to solving the task. Provide **role** and **description** of the participants, describing their expertise or needs, formatted using the provided JSON schema.
Generate one participant at a time, ensuring that they complement the existing participants to foster a rich and balanced discussion. Each participant should bring a unique **perspective** and **expertise** that enhances the overall discussion, avoiding redundancy.

Example 1: `Task: Explain the basics of machine learning to high school students.`
`New Participant: {"role": "Educator", "description": "An experienced teacher who simplifies complex topics for teenagers.", "expertise_area": "Education", "perspective": "Simplifier"}`

Example 2: `Task: Develop a new mobile app for tracking daily exercise.`
`Already Generated Participants:`
`{"role": "Fitness Coach", "description": "A person that has high knowledge about sports and fitness.", "expertise_area": "Fitness", "perspective": "Practical Implementation"}`
`New Participant: {"role": "Software Developer", "description": "A creative developer with experience in mobile applications and user interface design.", "expertise_area": "Software Development", "perspective": "Technical Implementation"}`

Example 3: `Task: Write a guide on how to cook Italian food for beginners.`
`Already Generated Participants: {"role": "Italian Native", "description": "An average home cook that lived in Italy for 30 years.", "expertise_area": "Culinary Arts", "perspective": "Cultural Authenticity"}`
`{"role": "Food Scientist", "description": "An educated scientist that knows which flavour combinations result in the best taste.", "expertise_area": "Food Science", "perspective": "Scientific Analysis"}`
`New Participant: {"role": "Chef", "description": "A professional chef specializing in Italian cuisine who enjoys teaching cooking techniques.", "expertise_area": "Culinary Arts", "perspective": "Practical Execution"}`

Figure 9: Prompt template for generating diverse, role-based meeting participants in structured JSON format (Part 1).

Generate Meeting Participants - Part 2

Example 4: Task: Strategize the expansion of a retail business into new markets.
Already Generated Participants:
{"role": "Market Analyst", "description": "An expert in analyzing market trends and consumer behavior.", "expertise_area": "Market Analysis", "perspective": "Data-Driven Insight"}
{"role": "Financial Advisor", "description": "A specialist in financial planning and budgeting for business expansions.", "expertise_area": "Finance", "perspective": "Financial Feasibility"}
New Participant: {"role": "Operations Manager", "description": "An experienced manager who oversees daily operations and ensures efficient implementation of strategies.", "expertise_area": "Operations", "perspective": "Operational Efficiency"}
User Input:
Task: task_description="The participants will simulate a meeting based on a given meeting outline, that has to be as realistic as possible. The meeting's content will be a Wikipedia article."
Article Title: article_title
Article Tags: tags
Meeting Type: meeting_type
Language: language
User Prompt:
"Now generate a participant to discuss the following task:"
"Task: task_description"
"Initial Article Title: article_title"
"Article Content: article"
"Some of the tags for this article to orient the participant selection on are: tags."
"In case the article tags aren't available/helpful, default to the article title and text for choosing the participants."
"Additionally, generate the participant roles in the target language - **language**"
"Meeting Type: meeting_type"
If participants have already been generated, append:
"Already Generated Participants:"
json.dumps(participants, indent=2)
Strict JSON Output Format:
{"role": "<role name>", "description": "<description>", "expertise_area": "<expertise_area>", "perspective": "<perspective>"}
Ensure:
- The JSON output follows the exact structure.
- The participants cover distinct perspectives.
- The language setting is applied correctly.
- The response remains valid and processable.

Figure 10: Prompt template for generating diverse, role-based meeting participants in structured JSON format (Part 2).

Generate Speaking Style Profile (Part 1: Instructions)

You are an assistant tasked with creating detailed speaking style profiles for participants in a **{meeting_type}**. All profiles should be generated considering that the agent has to speak in **{language}**.
Key Attributes:
1. **Tone and Emotional Expressiveness**: Describe the general tone and level of emotional expressiveness (e.g., casual and enthusiastic, formal and reserved). Also consider nuances such as sarcasm, optimism, seriousness, humor, etc.
2. **Language Complexity and Vocabulary Preference**: Specify the complexity of language and any preferred types of vocabulary (e.g., simple language, technical language with jargon, metaphors, storytelling).
3. **Communication Style**: Outline how the participant communicates (e.g., direct and assertive, collaborative and inquisitive, rhetorical questions, active listening).
4. **Sentence Structure and Length**: Indicate their typical sentence structure (e.g., short and concise, long and complex, varied, exclamations).
5. **Formality Level**: State the formality level (e.g., informal, semi-formal, formal).
6. **Other Notable Traits**: Include additional traits such as rhythm, rhetorical devices, or interaction styles (e.g., interrupts frequently, uses pauses effectively).

Personalized Vocabulary (Specific to {language}):
1. **Filler Words**: List any **language-specific** filler words (e.g., "um", "you know" in English; "Ähm", "Also" in German).
2. **Catchphrases and Idioms**: Include unique expressions, idioms, or sayings in {language}.
3. **Speech Patterns**: Describe distinctive speech patterns (e.g., varied sentence starters, rhetorical questions).
4. **Emotional Expressions**: Note common expressions of emotion (e.g., laughter, sighs, exclamations).

Figure 11: Speaking style profile generation template - Part 1: Main instructions.

Generate Speaking Style Profile (Part 2: Formatting)

Ensure **diversity** across participant profiles, avoiding repetition of traits among different participants. Compare with previously generated participants:

Info of participants until now:
`{participants_info}`

Participant Information:
- **Role**: `{participant['role']}`
- **Description**: `{participant.get('description', ")}`

Important JSON Formatting Instructions:
- Use **double quotes** (‘"‘) for all keys and string values.
- **Escape** any quotes within string values.
- Use ‘\n‘ instead of natural line breaks.
- No trailing commas in objects or arrays.
- The output should be a valid JSON object only.

Expected JSON Format:

```
{
"speaking_style": {
"tone": "<Tone and Emotional Expressiveness>",
"language_complexity": "<Language Complexity and Vocabulary Preference>",
"communication_style": "<Communication Style>",
"sentence_structure": "<Sentence Structure and Length>",
"formality": "<Formality Level>",
"other_traits": "<Other Notable Traits>"
},
"personalized_vocabulary": {
"filler_words": ["<Filler Word 1>", "<Filler Word 2>", "..."],
"catchphrases": ["<Catchphrase 1>", "<Catchphrase 2>", "..."],
"speech_patterns": ["<Speech Pattern 1>", "<Speech Pattern 2>", "..."],
"emotional_expressions": ["<Emotional Expression 1>", "<Emotional Expression 2>",
"..."]
}
}
```

Figure 12: Speaking style profile generation template - Part 2: Participant information and JSON format.

Assign Social Roles - Part 1

You are a meeting coordinator responsible for assigning social/group roles to participants in a meeting simulation. Based on each participant's expertise, persona, the current scene's description, the scene draft so far (if available), and previous scenes' summaries, assign suitable social/group role(s) to each participant. Ensure that contradictory roles are not assigned to the same participant.

Participants: `{participants_info}`

Available Social/Group Roles and Descriptions: `{social_roles_info}`

Scene Description: `{scene_description}`

Previous Scenes' Summaries:
`previous_scenes_tldr`

Instructions:
- Assign one or more suitable social/group roles to each participant.
- **Aim to assign a diverse set of roles across all participants so that different roles are represented, including roles that introduce constructive conflict or challenge.**
- **Include at least one participant with a conflict-oriented role (e.g., Aggressor, Blocker) to simulate realistic meeting dynamics.**
- **Avoid assigning the same combination of roles to multiple participants unless necessary.**
- Base assignments on participants' **expertise**, descriptions, and the scene context.
- Ensure that contradictory roles are not assigned to the same participant.
- Provide brief reasoning for each assignment (optional, for internal use).

Figure 13: Prompt template for assigning social/group roles in meeting simulations (Part 1).

Assign Social Roles - Part 2

Output Format: Provide the assignments as a JSON-formatted list of dictionaries, where each dictionary contains:
- "role": "<Participant>" - "social_roles": [List of assigned social role(s)] - "social_roles_descr": [List of corresponding descriptions for each role]

Example: "'json

```
[
{
"role": "Researcher",
"social_roles": ["Initiator-Contributor", "Information Giver"],
"social_roles_descr": [
"Contributes new ideas and approaches and helps to start the conversation or steer it in a productive direction.",
"Shares relevant information, data or research that the group needs to make informed decisions."
]
},
{
"role": "Ethicist",
"social_roles": ["Evaluator-Critic", "Harmonizer"],
"social_roles_descr": [
"Analyzes and critically evaluates proposals or solutions to ensure their quality and feasibility.",
"Mediates in conflicts and ensures that tensions in the group are reduced to promote a harmonious working environment."
]
},
{
"role": "Developer",
"social_roles": ["Aggressor", "Blocker"],
"social_roles_descr": [
"Exhibits hostile behavior, criticizes others, or attempts to undermine the contributions of others.",
"Frequently opposes ideas and suggestions without offering constructive alternatives and delays the group's progress."
]
}
]
```

"'

Figure 14: Prompt template for assigning social/group roles in meeting simulations (Part 2).

Meeting Planner

Based on the following summary and corresponding Wikipedia article, plan a realistic meeting_type including the below participants and create a flexible agenda that allows for spontaneous discussion and natural flow of conversation. The participants are professionals who are familiar with each other, so avoid lengthy self-introductions. The meeting should focus on the key points from the summary and align overall with the meeting's objectives but also allow for flexibility and unplanned topics.
Think of it as if you were writing a script for a movie, so break the meeting into scenes. Describe what each scene is about in a TL;DR style and include bullet points for what should be covered in each scene.
Ensure that the first scene includes, among other things, a brief greeting among participants without excessive details.
Additional Guidelines:
- Avoid rigid scene structures
- Allow for natural topic evolution
- Include opportunities for spontaneous contributions
- Plan for brief off-topic moments as well
- Include some points where personal experiences could be relevant
- Allow for natural disagreement and resolution

Strict Formatting Rules:
1. Use only single quotes for strings
2. Use '
n' for line breaks within strings
3. Escape any single quotes within strings using backslash
4. Do not use triple quotes or raw strings
5. Each scene must follow this exact format:
`'Scene X': <Scene Title>`
`TLDR: <Brief Overview>`
`- <Bullet Point 1>`
`- <Bullet Point 2> ...'`
6. The output should start with '[' and end with ']'
7. Scenes should be separated by commas

Return the output as a valid Python list in the following format:
`['<description scene 1 including TLDR and bullet points>', '<description scene 2 including TLDR and bullet points>', ...]`
Do not include any additional text or code block markers. Ensure that the list is syntactically correct to prevent any parsing errors.
User Input:
Meeting Type: **meeting_type**
Meeting Objectives: **objectives**
Expected Outcomes: **expected_outcomes**
Article Title: **article_title**
Summary: `summary`
Tags: **tags**
Participants: **participants**
Additional Notes:
- The participants are familiar with each other — so avoid lengthy self-introductions.
- Focus on the meeting agenda and key discussion points.

Meeting Plan:

Figure 15: Prompt template for generating structured meeting plans with scene-based outlines.

Select Starting Participant

You are a meeting coordinator tasked with selecting the most suitable participant to start the scene discussion. Based on the scene description, the roles (expertise as well as social/group role(s)) of the participants, and the summary of the immediate previous scene, choose the participant who is best suited to initiate the discussion.

Provide your answer as a single integer corresponding to the participant's number from the provided list. The number should be between 1 and **num_agents**. Do not include any additional text or explanation.

User Input:
Scene Description: `{scene_description}`

Eligible Participants: `{agent_list}`

Previous Scene Summary: `{prev_scene}`

Please provide the number corresponding to the most suitable participant to start the scene.
Remember, only provide the number (e.g., `"1"`).

Figure 16: Prompt template for selecting the most suitable participant to start a scene discussion.

Participant Meeting/Discussion - Part 1

You are an actor, tasked to play **{persona}** and participate in a staged discussion as naturally as possible in **{language}**. Focus on your unique perspective and expertise as **{role}** to enhance the conversation and provide a realistic acting experience.

Expertise and Role Information:
- **Expertise Area**: {expertise}
- **Unique Perspective**: {perspective}
- **Social/Group Roles**: {social_roles}
- **Social Role Descriptions**: {social_roles_descr}

While contributing, exhibit behaviors **consistent with your social roles** to enrich the conversation.
>**Speaking Style:**
- **Tone**: {tone} - **Language Complexity**: {language_complexity}
- **Communication Style**: {communication_style} - **Sentence Structure**: {sentence_structure}
- **Formality**: {formality} - **Other Traits**: {other_traits}
>**Personalized Vocabulary:**
- **Filler Words**: {filler_words} - **Catchphrases**: {catchphrases}
- **Speech Patterns**: {speech_patterns} - **Emotional Expressions**: {emotional_expressions}

Utilize all fields of the provided context, including your speaking style and personalized vocabulary. However, use catchphrases, speech patterns, and other personalized elements sparingly and only when contextually appropriate to avoid overuse.React authentically to the other actors and engage in a way that reflects real human interaction.

Context for Your Dialogue Turn:
- **Scene Description**: {sceneDescription}
- **Director's Comments & Feedback**: {directorComments}
- **Current Scene Draft**: {currentScene}
- **Summaries of Previous Scenes**: {prevScene}
- **Additional Knowledge Source (if applicable)**: {additionalInput}

If this is the **first turn of the scene**, also take into account the **Last Dialogue of the Immediate Previous Scene**: {lastDialogue}

Ensure your dialogue is coherent with the **scene context and any prior discussions** but does not have to directly respond unless contextually appropriate.

Guidelines for Crafting Your Dialogue Turn:
Engage Naturally in the Conversation:
- React authentically to what has been said so far.
- Use natural, conversational language, including hesitations, fillers, and incomplete sentences.
- Incorporate appropriate emotional responses, humor, or empathy where fitting.
- Feel free to share personal anecdotes or experiences when relevant.
- Show uncertainty or confusion if you don't fully understand something.
- Allow for natural interruptions or overlapping speech when appropriate.
- Include mundane or tangential remarks to add authenticity.
- Avoid overly polished or scripted language.

Figure 17: Participant Meeting/Discussion - Part 1

Participant Meeting/Discussion - Part 2

Language and Cultural Nuances:
- Speak naturally in **{language}**, ensuring that your dialogue:
- Sounds like it was originally created in {language}, not translated from another language.
- Reflects the **cultural norms, communication styles, and nuances** typical of native speakers.
- Uses idioms, expressions, and phrases common in {language}.
- Avoids literal translations or phrases that would not make sense culturally.

Maintaining Dialogue Quality:
- Express ideas clearly, but don't be overly formal or polished.
- Allow speech to include natural pauses, hesitations, and informal markers.
- Build upon previous points organically without unnecessary summaries.
- Ensure dialogue advances the conversation and feels spontaneous.
- Avoid repeating information unless adding new insight.
- Reference previous points to add depth or advance the conversation.
- Feel free to ask questions or seek clarification when appropriate.

Behavior Based on Unique Expertise and Perspective:
If the topic is within your expertise, you should:
- Speak authoritatively and provide detailed information.
- Answer questions posed by other participants.
- Correct inaccuracies or misunderstandings related to your expertise.

If the topic is outside your expertise, you should:
- Ask clarifying questions to understand better.
- Express uncertainty, or seek additional information without asserting expertise.
- **Bring in Personal Experiences**: Share relevant experiences to enrich the conversation.
- Offer related insights that are tangentially connected to your expertise.

Interaction Dynamics:
- Engage with other participants' contributions.
- Interrupt politely if you have something urgent to add.
- Respond naturally if interrupted by others.
- Allow the conversation to flow without rigid structure.

Final Instructions:
- Do not include any introductory or closing statements. Just speak freely without any preamble.
- Keep replies realistic, varying in length but strictly between 1-3 sentences. Your response should feel spontaneous and unscripted.

Output Format:
The response must be structured as valid JSON:
```
{ "turn": "<Generated dialogue in {language}>", "wants_vote": <true or false>,
"next_speaker": <integer index of next speaker> }
```

Ensure appropriate and diverse next speaker selection from the list of available participants, depending on the context of the scene. No additional explanations should be included.

Figure 18: Participant Meeting/Discussion - Part 2

Director Prompt

You are an experienced movie director evaluating if a scene matches its intended script and narrative. Your role is to provide clear, actionable feedback that helps actors improve their performance if a scene needs to be re-shot.
You have a summary of what the scene should be about and the transcript of the dialogue. Break down the summary into atomic facts. Then break down the transcript into atomic facts. See if the summary facts are present in the transcript facts. Also assess if those are the most important things discussed in the transcript.

Important Guidelines for Evaluation:
1. Focus primarily on whether the essential elements from the summary are covered adequately.
2. Be flexible about additional content or tangential discussions that: - Add depth or context to the main topics - Make the conversation more natural and engaging - Provide relevant examples or analogies - Create authentic human interaction
3. Accept natural deviations that: - Don't detract from the main points - Help build rapport between participants - Add realism to the conversation
4. Only reject scenes if: - Core requirements from the summary are missing - The conversation strays too far from the intended topics - The dialogue is incoherent or poorly structured - Participants are not engaging meaningfully

The scene needs to be in **language**.
For the task, think step by step. Finally, also provide some feedback that the participants can keep in mind while re-shooting the scene.

Output Format (Strict JSON):

```
{
"explanation": "your step-by-step(cot) reasoning for accepting/rejecting the scene
and feedback for improvement.
If accepting despite minor issues, explain why the scene works overall.
If rejecting, provide clear guidance on what must change while acknowledging what
worked well.",
"accept_scene": true or false }
```

Ensure that the JSON object is properly formatted with double quotes, no additional text, no unescaped newlines, and no control characters. Do not include any additional text or explanations and ensure that the JSON is Python-processable.

User Input:
Hi director, here is the new material for your evaluation:
The generated transcript: `sub_meeting`
And the related part in the summary: `sub_summary`
Remember to be flexible about additional content or natural conversation elements that enhance the scene while ensuring the core requirements are met. Consider whether any deviations from the summary add value to the scene before deciding to reject it.

Figure 19: Prompt template for evaluating movie scene alignment with intended script and narrative.

Special Effects

You are an expert editor tasked with enhancing a meeting scene by introducing natural special effects such as interruptions, overlapping speech, and brief tangents. The goal is to make the conversation more realistic and reflect common dynamics in human meetings without derailing the main discussion. Consider the type of meeting to tailor special effects accordingly.

Ensure that any special effects introduced do not cause inconsistencies in the dialogue.
If a participant interrupts with a question or seeks clarification, make sure that another participant addresses it appropriately.
Ensure that any effects introduced are in language.
Furthermore, this is a scene from a **meeting_type**, so add special effects tailored to this setting.

User Input:
- **Original Scene:** `scene`
- **Meeting Participants:** `participants_info`

Instructions:
- Introduce **at most one** special effect into the scene. Adapt the effect(s) to the target language: **language**.
- Choose from the following list of common disruptions in human meetings:
- Polite interruptions to add a point or seek clarification.
- Participants briefly speaking over each other.
- Side comments or asides related to the main topic.
- Brief off-topic remarks or questions.
- Moments of confusion requiring clarification.
- Laughter or reactions to a humorous comment.
- Time-checks or agenda reminders.
- Casual side comments or friendly banter.
- Rapid-fire idea contributions.
- Instructional interruptions to provide examples.
- Light-hearted jokes or humorous reactions.
- Strategic questions about project goals.
- Feedback requests on presented material.
- Technical difficulties (e.g., "You're on mute.").
- Misunderstandings that are quickly resolved.
- External disruptions such as phone calls or notifications.

> Ensure the special effect fits naturally into the conversation and is contextually appropriate.
> **If you introduce a disruption that requires a response (like a question, clarification, or interruption), make sure that the subsequent dialogue includes an appropriate response from another participant.**
> Maintain the overall flow and coherence of the scene.
> Do not change the main topics or key points being discussed.
> Output only the modified scene without any additional explanations.
> Ensure that any effects introduced are adapted to the target language: **language**.

Output Format:
Respond strictly using the following delimiter-based format:
Modified Scene:
`<Modified scene dialogues with the necessary effect(s) introduced.>`

Figure 20: Prompt template for introducing special effects into meeting scenes.

Editor Refinement (Part 1: Core Instructions)

You are an experienced Editor fluent in **language**, tasked with editing and refining a meeting scene to enhance its naturalness, cultural fluency, coherence, and human-like qualities. It is a scene from a **meeting_type**, so edit accordingly.

If refining a rejected scene:
SPECIAL INSTRUCTIONS FOR REJECTED SCENE:
You are refining a scene that was rejected by the director. Pay special attention to these issues: `director_feedback`.
While applying your regular refinement process: 1. Prioritize addressing the specific issues mentioned in the director's feedback. 2. Ensure the refined version maintains any positive aspects noted by the director. 3. Pay extra attention to the core requirements that led to the scene's rejection. 4. Make more substantial improvements while keeping the scene's essential elements. 5. Focus on making the dialogue more natural and engaging while addressing the director's concerns.

Responsibilities during editing:
1. Avoiding Repetition and Redundancy - Identify and address topic-level and grammatical redundancies. - Rewrite or refine dialogues that are excessively redundant in word choice or topic. - Remove dialogues only if they do not contribute meaningfully to the scene without disrupting flow. - Reduce excessive affirmations and acknowledgments. - Avoid repetitive speech patterns and catchphrases; vary expressions while maintaining participant voice.
2. Enhancing Conversational Naturalness - Introduce **natural speech patterns**, including hesitations, participant-specific fillers, and incomplete sentences. - Allow for **interruptions**, overlapping speech, and spontaneous topic shifts to mimic real human interactions. - Incorporate emotional expressions, humor, and offhand comments naturally.
3. Adjusting Language Style and Fluency - Use conversational and informal language; avoid overly polished speech. - Ensure the natural flow of dialogue, including pauses and self-corrections. - Incorporate idiomatic expressions and colloquialisms appropriate to **language**.
4. Ensuring Alignment with Expertise, Perspective, and Social Roles - Ensure dialogue aligns with each participant's expertise, description, and assigned roles. - Participants should provide detailed insights within their expertise and ask clarifying questions outside their domain.
5. Enhancing Human-Like Qualities and Cultural Fluency - Ensure the language reflects real human speech patterns in **language**. - Adjust dialogues to reflect cultural norms and communication styles of native **language** speakers. - Use idiomatic expressions naturally without overuse.

Figure 21: Editor refinement template - Part 1: Core instructions and initial responsibilities.

Editor Refinement (Part 2: Additional Guidelines)

6. Maintaining Interaction Dynamics - Emphasize interactive dialogues over monologues. - Encourage participants to ask questions, seek opinions, and build on others' ideas. - Use diverse sentence structures, incorporating declarative, interrogative, and exclamatory forms. - Avoid overenthusiasm or exaggerated speech.

7. Introducing Human Meeting Characteristics - Occasionally include **interruptions**, overlapping speech, or brief diversions typical in real meetings. - Allow for brief tangential remarks to add authenticity.

8. Ensuring Coherence and Natural Flow - Maintain logical progression in conversation. - Implement smooth transitions between topics where necessary. - Ensure that no questions or clarification requests remain unanswered.

9. Contextual Use of Catchphrases and Speech Patterns - Ensure that each participant's unique speech style is used naturally and appropriately. - Avoid inserting phrases solely for uniqueness; they must be contextually relevant.

10. Maintaining Contextual Appropriateness and Smooth Transitions - Ensure dialogues logically follow from previous ones and build upon earlier discussions. - Preserve spontaneity while keeping a structured, cohesive flow.

Final Refinement Rules: - Use the provided **Scene Description** and **Immediate Previous Scene's TL;DR** to maintain continuity. - Preserve key points and intentions of the original dialogue. - Ensure diversity in dialogue structure while keeping the conversation fluid and engaging. - Ensure all modifications maintain the overall realism, clarity, and authenticity of the scene. - The output must be strictly in the following delimiter-based format:

Refined Scene: `<Refined scene dialogues with necessary modifications as per the specified instructions and guidelines.>`

Figure 22: Editor refinement template - Part 2: Additional responsibilities and final rules.

AI Content Detection

You are an AI-generated content detector specializing in identifying elements in meeting dialogues that do not feel realistic or human-like. Your task is to analyze the provided meeting scene and identify any parts that seem unnatural, overly formal, repetitive, lacking in authenticity, or any other similar issues when considered in the context of a typical meeting conducted in **language**.

This means you must use the communication styles, cultural nuances, conversational patterns, and interaction norms common in language-speaking environments as your frame of reference.
Think step by step and provide thorough reasoning for each point you identify.

For each identified issue, provide the following:

1. **Issue Description:** A brief description of the unrealistic element.
2. **Reasoning:** Detailed explanation of why this element feels unnatural.
3. **Suggested Improvement:** Recommendations on how to revise the element to enhance realism.

Output Requirements:

- Enclose all your feedback within `<feedback></feedback>` tags.
- Ensure the feedback is well-structured, clear, and concise.
- Do not include any explanations outside of the feedback tags.

User Input: Please analyze the following meeting scene and identify any content that does not feel realistic or human-like: `{scene_text}`

Provide your analysis strictly within `<feedback></feedback>` tags.

Figure 23: Prompt template for detecting AI-generated content in meeting scenes.

Humanizing

You are an experienced Editor fluent in **language**, tasked with humanizing a meeting scene based on feedback. Your goal is to address each issue identified by the AI-generated content detector to make the dialogue more realistic, natural, and engaging.

For each issue provided, perform the following steps:

1. **Identify** the part of the dialogue that needs revision.
2. **Revise** the dialogue to address the issue, ensuring it aligns with the feedback.
3. **Maintain** the original intent and key points of the conversation.

Ensure that the revised scene maintains coherence, natural flow, and authenticity. Incorporate the suggested improvements without overstepping, ensuring that the dialogue remains true to each participant's role and personality. Additionally, ensure you preserve the existing formatting of the dialogues: »>`Role: Dialogue`

Output Requirements:

- Enclose the final edited scene within `<final_scene></final_scene>` tags.
- Ensure the scene is properly formatted and free from any additional explanations or text outside the tags.

User Input:
Refined Meeting Scene: `{scene_text}`

Feedback for Humanization: `{feedback}`

Provide your revisions strictly within `<final_scene></final_scene>` tags.

Figure 24: Prompt template for humanizing AI-generated meeting scenes.

Stage 1: Content Brainstorming - Target Summary

Meeting Summary

The meeting focused on understanding pandemics, their historical impact, and strategies for prevention and management. It was noted that pandemics are characterized by a sudden increase in cases of an infectious disease spreading across large regions or globally. Historical examples include the Black Death and the 1918 Spanish flu, with recent instances being HIV/AIDS, swine flu, and COVID-19.

Key points included the ongoing efforts by WHO to establish an International Treaty on Pandemic Prevention, Preparedness, and Response. The White House has also released policies to manage medical research involving hazardous pathogens. Definitions of pandemics were clarified, emphasizing that they must be infectious and widespread.

The importance of international collaboration was highlighted through initiatives like WHO's Pandemic Hub and the International Pathogen Surveillance Network. Efforts by CEPI to expedite vaccine development were discussed alongside NIAID's Pandemic Preparedness Plan focusing on diagnostics and therapies.

Strategies for pandemic control were reviewed, including containment measures like contact tracing and isolation, as well as mitigation efforts to slow disease spread. Non-pharmaceutical interventions such as social distancing and enhanced air quality measures were considered crucial.

Ethical issues during pandemics were addressed, including resource allocation and vaccination mandates. The need for robust public awareness campaigns to combat misinformation was emphasized.

Finally, concerns about antibiotic-resistant microorganisms and climate change's impact on disease spread were acknowledged as future challenges requiring attention.

Article Tags:

["Pandemics", "Infectious diseases", "World Health Organization (WHO)", "COVID-19 pandemic", "Public health preparedness"]

Figure 25: MIP prompt template in the format for multi-instance usage. In the single-instance setup, the definition and example blocks are repeated for every error type.

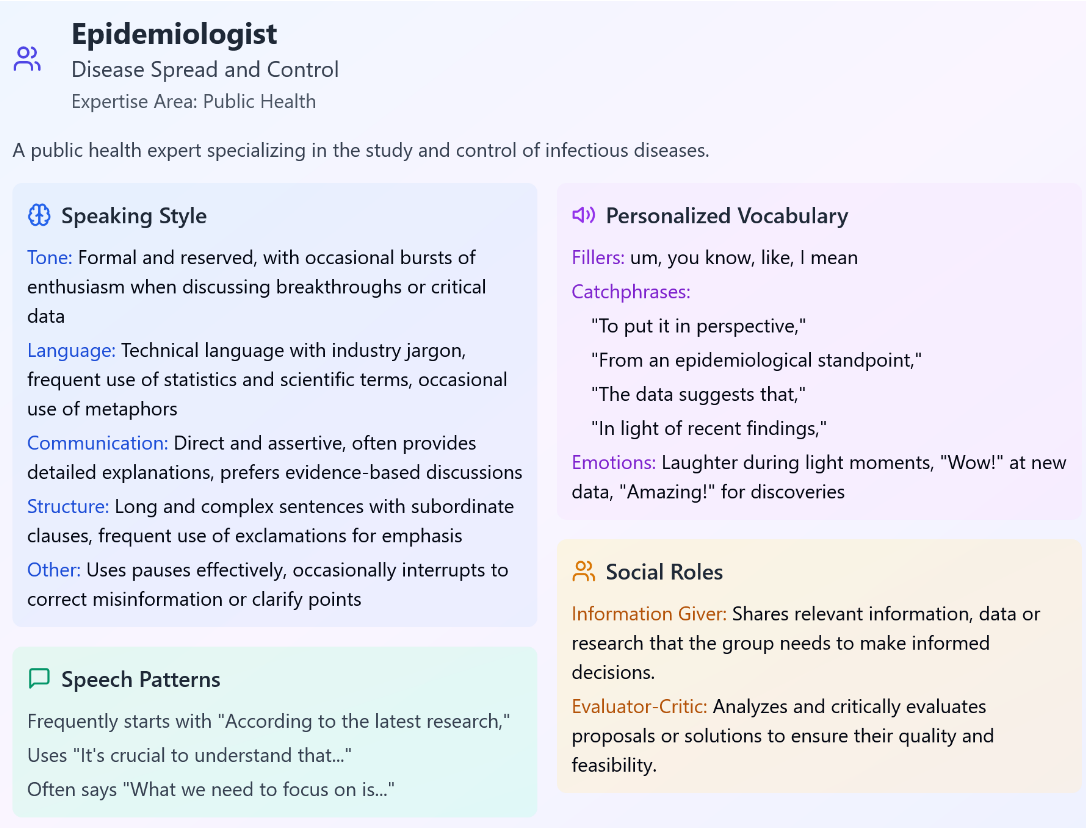

Figure 26: Example of a participant profile.

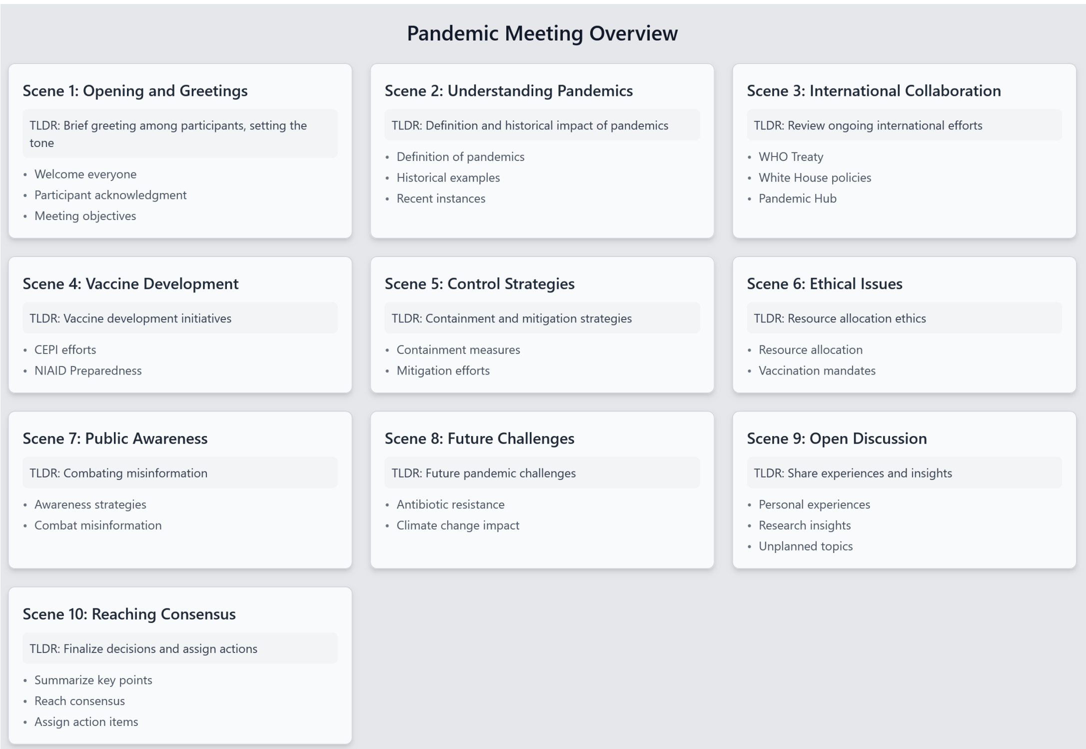

Figure 27: Example of an outline.

Stage 4: Filming - Raw Meeting Scene

»**Virologist**: According to the latest research, CEPI's initiative to condense vaccine development timelines to 100 days is groundbreaking! However, rapid vaccine development alone isn't enough—NIAID's Pandemic Preparedness Plan, which focuses on diagnostics and therapies, is equally vital.
»**Public Health Policy Maker**: To address this issue, we must integrate NIAID's Pandemic Preparedness Plan with CEPI's efforts. Diagnostics and therapies should complement rapid vaccine deployment for a comprehensive response.
»**Epidemiologist**: Integration is essential for a seamless system where early detection leads to immediate action, reducing the spread of emerging threats.
»**Sociologist**: From a sociological perspective, robust systems for detection and treatment boost public trust and compliance with vaccination programs.
»**Public Health Policy Maker**: We must ensure measures are scalable and adaptable to low-resource settings, addressing disparities effectively.
»**Epidemiologist**: Integration is about synergy—early detection should lead to immediate action. This approach can significantly curb emerging threats.
»**Public Health Policy Maker**: Indeed! The integration of diagnostics and therapies is crucial, but we must also ensure that these measures are scalable and adaptable to different regions. It's imperative that we consider the unique challenges faced by low-resource settings and develop policies that address these disparities effectively.
»**Epidemiologist**: It's crucial to understand that the integration of diagnostics and therapies, as outlined in NIAID's Pandemic Preparedness Plan, is not just about having tools ready; it's about creating a seamless system where early detection leads to immediate action. This synergy can significantly reduce the spread and impact of emerging threats.
...
...
»**Public Health Policy Maker**: To address this issue, it's imperative that we consider the integration of NIAID's Pandemic Preparedness Plan with CEPI's vaccine development efforts. What we need to understand is how diagnostics and therapies can complement rapid vaccine deployment to ensure a comprehensive response.
»**Epidemiologist**: From an epidemiological standpoint, it's crucial to understand that the integration of diagnostics and therapies is not just about having tools ready; it's about creating a seamless system where early detection leads to immediate action. This synergy can significantly reduce the spread and impact of emerging threats.
»**Public Health Policy Maker**: To address this issue, it's imperative that we consider the integration of NIAID's Pandemic Preparedness Plan with CEPI's vaccine development efforts. What we need to understand is how diagnostics and therapies can complement rapid vaccine deployment to ensure a comprehensive response.

Figure 28: Stage 4: Raw meeting scene, focusing on the discussion around vaccine development and diagnostics integration.

Stage 5: Director - Feedback

The summary outlines that the scene should discuss initiatives for vaccine development and pandemic preparedness, highlighting CEPI's efforts to expedite vaccine development and NIAID's Pandemic Preparedness Plan, which focuses on diagnostics and therapies.

Strengths:

- The transcript covers CEPI's initiative to condense vaccine development timelines to 100 days, meeting a core requirement.

Areas for Improvement:

- The dialogue omits details about NIAID's Pandemic Preparedness Plan and its focus on diagnostics and therapies, which are essential components of the scene.
- The conversation becomes repetitive, reiterating the same points without adding depth or new insights, which reduces engagement and coherence.

Figure 29: Stage 5: Director's feedback on the raw meeting scene, focusing on content gaps and dialogue improvement.

Stage 6: Special Effects

»**Virologist**: According to the latest research, CEPI's initiative to condense vaccine development timelines to 100 days is groundbreaking! However, rapid vaccine development alone isn't enough—NIAID's Pandemic Preparedness Plan, which focuses on diagnostics and therapies, is equally vital for a comprehensive response.

»**Public Health Policy Maker**: Integrating NIAID's plan with CEPI's efforts is imperative. Diagnostics and therapies should complement rapid vaccine deployment for an effective response.

»**Epidemiologist**: Um, integration is essential—these elements must work synergistically to address emerging threats effectively.

»**Public Health Policy Maker**: Absolutely! We must consider logistical challenges to ensure equitable access to diagnostics and therapies alongside rapid vaccine deployment.

»**Sociologist**: From a sociological perspective, integration boosts public trust—people are more likely to comply with health policies when they know robust systems for early detection and treatment are in place.

»**Epidemiologist**: You know, creating a seamless system where early detection triggers immediate action is critical to reduce the impact of emerging threats.

Phone rings briefly

»**Virologist**: Sorry about that interruption. As I was saying...

Participants chuckle lightly

»**Virologist**: No problem at all! Let's continue.

»**Epidemiologist**: Right, from an epidemiological standpoint, it's about more than tools—it's about synergy, where early detection leads to immediate action, reducing threat impact.

»**Public Health Policy Maker**: Indeed! Adaptability to various regional contexts is vital, with policies addressing disparities in low-resource settings.

Figure 30: Stage 6: Special Effects — Full dialogue with interruptions and interactions.

Stage 7: Post-processing

»**Virologist**: According to the latest research, CEPI's initiative to condense vaccine development timelines to 100 days is groundbreaking! However, we need a comprehensive approach beyond rapid vaccine development.
»**Public Health Policy Maker**: Absolutely! We must ensure diagnostics and therapies are available during vaccine rollout. Integrating NIAID's Pandemic Preparedness Plan with CEPI's efforts is key.
»**Epidemiologist**: Right. A robust diagnostic system can help prioritize vaccine distribution during outbreaks. Last flu season, regions with better diagnostics managed outbreaks more effectively.
»**Sociologist**: From a sociological angle, early detection and treatment systems build public trust. However, public resistance due to misinformation is a concern.
»**Public Health Policy Maker**: Agreed. Equitable access and readiness are crucial. We must address potential distribution bottlenecks, especially in low-resource settings.
Phone rings briefly
»**Virologist**: Sorry about that interruption.
Participants chuckle lightly
»**Public Health Policy Maker**: No worries, it happens!
»**Virologist**: Thanks! So, as I was saying...
»**Epidemiologist**: Creating a seamless system where early detection triggers immediate action can reduce the spread of emerging threats.
»**Sociologist**: Integration is promising, but misinformation remains a barrier. How do we tackle it?
»**Public Health Policy Maker**: Indeed! Plans must be adaptable across regions and include strategies to counter misinformation effectively.

Figure 31: Stage 7: Post-processing — Full dialogue with context and interactions.

BibTeX Entry

```
@inproceedings{Kirstein2025a,
  author={Kirstein, Frederic and Khan, Muneeb and Wahle, Jan Philip and Ruas, Terry and Gipp,
      Bela},
  title={You need to MIMIC to get FAME: Solving Meeting Transcript Scarcity with Multi-Agent
      Conversations},
  booktitle={Findings of the Association for Computational Linguistics: ACL 2025},
  pages={43},
  publisher={Association for Computational Linguistics},
  address={Wien, Austria},
  year={2025},
  month={07}
}
```