Deep Think with Rehearsal for Low-Latency Team-AI Collaboration

Anonymous ACL submission

Abstract

The integration of Large Language Models (LLMs) into scientific team meetings presents exciting opportunities to accelerate biomedical discovery, especially through their strong deep thinking capabilities enabled by multi-step reasoning and web search. However, such methods are computationally expensive and introduce substantial latency, limiting their effectiveness in real-time Team-AI communications. In this study, we propose *Deep Think with Rehearsal* (DTR), a novel framework that decouples deep reasoning from synchronous interaction in the AI4Science context. DTR transfers computationally intensive reasoning into an offline rehearsal phase, allowing the LLM to pre-cognize complex scientific contexts and deliver high-quality, "deep" insights with minimal latency during live interactions. To facilitate this research, we introduce the *Scientific Team Meeting Dataset* (STMD), a hybrid benchmark comprising authentic transcripts from three real-world biomedical research labs alongside extensive simulated multi-party deliberations synthesized from PubMed literature. Experiments in both simulated and real-world settings demonstrate that DTR consistently improves response quality while reducing inference latency compared to state-of-the-art methods, highlighting the effectiveness of rehearsal in enabling low-latency, high-quality scientific collaboration.[1]

1 Introduction

Large Language Models (LLMs) have demonstrated strong reasoning and tool augmentation capabilities, such as chain-of-thought problem-solving (Wei et al., 2022; Zhang et al., 2025) and tool use via web search (Yao et al., 2022; Schick et al., 2023), making them increasingly effective at enabling Team-AI collaboration and accelerating scientific discovery (Zheng et al., 2025; Singh et al., 2025; Qi et al., 2024). However, integrating these capabilities into real-time conversations (e.g., scientific meetings) remains challenging, as deep reasoning and web search at inference time are computationally expensive, introducing prohibitive delays that disrupt conversational flow and limit their effectiveness in Team-AI collaboration (Gao et al., 2025; Bi et al., 2024). This constraint highlights a gap between how current LLM assistants behave and how effective human collaborators participate in meetings (Schmutz et al., 2024; Liao et al., 2023). Crucially, unlike general-purpose open-domain dialogues, scientific meetings are characterized by a high degree of predictability, as they are typically centered around predefined research objectives, specific datasets, or persistent scientific hypotheses. Once the meeting topic is known, human experts typically leverage this thematic stability to prepare in advance by surveying relevant literature, organizing key evidence, and formulating candi-

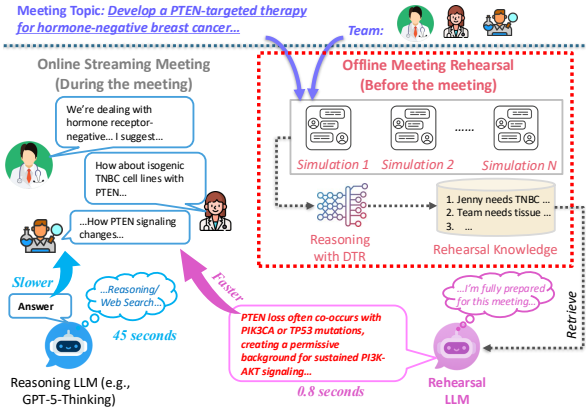

Figure 1: Comparison of traditional reasoning LLM and rehearsal LLM in real-time Team-AI collaboration. The traditional reasoning LLM (left) performs deep thinking online (during the meeting) with high latency, whereas the rehearsal LLM (right) rehearses offline (before the meeting) via meeting simulations and construct a rehearsal knowledge database, enabling fast retrieval during online interaction.

[1] Code and data are provided in the supplementary material and will be made publicly available upon acceptance.

date hypotheses (Maslych et al., 2025; Král et al., 2023). This pre-meeting preparation ensures they can engage efficiently during live discussions rather than interrupting the flow to search for information. We argue that LLM-based assistants should follow a similar "preparation-first" principle. By exploiting the predictable scope of scientific deliberations, we can transfer the computational burden of deep reasoning into an offline phase, achieving high-quality insights while maintaining the low-latency responsiveness required for seamless real-time collaboration (Pelikan and Hofstetter, 2023; Liu et al., 2025).

Motivated by this insight, we introduce *Deep Think with Rehearsal* (DTR), a rehearsal-based framework that transfers expensive online reasoning and knowledge gathering to an offline rehearsal phase, enabling fast responses during deployment while preserving reasoning quality (Prince et al., 2024). As illustrated in Figure 1, in the **online** setting (i.e., real-time interaction during the scientific meeting), a conventional reasoning LLM (e.g., GPT-5-Thinking (OpenAI, 2025c) or Gemini 3 (Google DeepMind, 2025)) performs deep reasoning and web search during the discussion, which can take minutes per turn and introduce substantial response delays in Team-AI interaction (Cheng et al., 2025). In contrast, DTR shifts these computationally intensive steps to an **offline** phase (i.e., pre-meeting preparation), where it runs confidence-based meeting simulations conditioned on the meeting topic and participants and uses a reasoning LLM to perform deep reasoning and web search over simulated trajectories, thereby constructing a rehearsal knowledge database before the meeting starts. During the meeting, the DTR LLM only needs to leverage lightweight retrieval from the prepared knowledge database to deliver timely, high-quality suggestions, for example, recommending that the team consider PIK3CA or TP53 when designing PTEN-targeted therapy studies for hormone receptor negative breast cancer, with a response latency of only 0.8 seconds.

To support the study of low-latency deep thinking, we introduce the *Scientific Team Meeting Dataset* (STMD), a hybrid benchmark comprising authentic transcripts from three real-world biomedical research labs covering cancer, sepsis, and cardio-metabolic disease, along with extensive simulated multi-party deliberations synthesized from PubMed literature (Sayers et al., 2024). STMD features multi-role biomedical team meetings and is designed to evaluate both response quality and inference latency. In both simulated and real-world studies, DTR consistently improves reasoning quality while substantially reducing inference latency compared with state-of-the-art methods (OpenAI, 2025c), highlighting offline rehearsal as an effective approach for real-time, high-quality Team-AI scientific collaboration.

Our contributions in this paper can be summarized as follows:

- **Conceptual Paradigm:** We formalize the problem of low-latency deep thinking in real-time collaboration and propose the "Rehearse-then-Interact" paradigm. This shifts the focus from optimizing online inference to exploiting the inherent predictability of structured scientific meetings through offline preparation.
- **Framework:** We propose DTR, a novel framework that operationalizes this paradigm. By decoupling expensive multi-step reasoning and web-based knowledge gathering into an offline phase, DTR enables LLMs to deliver "deep" insights with sub-second latency during live interactions.
- **Benchmark:** We introduce the STMD, a specialized benchmark derived from PubMed literature and real-world lab meetings. It provides a robust testbed for evaluating both the reasoning depth and temporal efficiency of LLMs in multi-party scientific deliberations.
- **Empirical Validation:** We conduct extensive experiments in both simulated environments and real-world biomedical lab settings. Results demonstrate that DTR consistently outperforms state-of-the-art methods in both response quality and inference efficiency, proving the viability of rehearsal-based AI collaboration.

2 Related Work

LLMs with Deep Think Recent work improves LLM reasoning through structured intermediate steps and external knowledge integration. Chain-of-thought prompting (Wei et al., 2022), self-consistency (Wang et al., 2022), and Tree of Thoughts (Yao et al., 2023) elicit multi-step reasoning and explore alternative paths via search. For grounding and tool use, retrieval-augmented generation (Lewis et al., 2020) and ReAct (Yao et al., 2022) integrate retrieval and tool-mediated actions into generation, while Toolformer (Schick et al., 2023), MRKL (Karpas et al., 2022), and

program-aided models (Gao et al., 2023) enable self-supervised tool use, modular tool composition, and code-based reasoning. Confidence-aware methods (Subramani et al., 2025; Taubenfeld et al., 2025) adjust reasoning depth to uncertainty to reduce low-utility computation, but over-reasoning can still inflate cost (Lacombe et al., 2025). These pipelines typically require iterative calls, retrieval, and tool execution during online inference, increasing latency and disrupting real-time interaction, whereas DTR shifts these costs to offline rehearsal and answers online via lightweight retrieval.

LLMs with Rehearsal Rehearsal is widely used in reinforcement learning and agent systems to simulate future interactions and improve decisions. In model-based RL, MuZero (Schrittwieser et al., 2020) and Dreamer (Hafner et al., 2023) plan via imagined rollouts, and Language Agent Tree Search (Zhou et al., 2023) performs lookahead over action branches. Beyond RL, offline simulation and self-play have been used to anticipate user responses and build reusable strategies in meeting settings (He et al., 2024, 2025; Kim et al., 2025). Role-play and continual exploration further accumulate interaction experience in multi-agent systems (Li et al., 2023; Park et al., 2023; Wang et al., 2023a; Yamada et al., 2025b), while offline training can strengthen multi-step reasoning (Wang et al., 2025). These lines of work often keep substantial reasoning on the online path or prioritize capability over latency, whereas DTR distills evidence-backed reasoning offline and answers online via lightweight retrieval.

3 Methodology

3.1 Preliminary Definitions

We consider a fixed team $\mathcal{T}$ with a set of roles (team members) $\mathcal{R} = \{r_1, \ldots, r_{|\mathcal{R}|}\}$. A multi-role scientific meeting is a sequence of utterances $\mathcal{D} = \{u_1, \ldots, u_T\}$, where each utterance is a speaker–content pair $u_t = (r_t, x_t)$, $r_t \in \mathcal{R}$ and x_t is a text utterance. Let $h_t = \{u_1, \ldots, u_t\}$ denote the meeting context up to round t.

Task Definition. Given the current *online* meeting context[2] h_t^{on} up to round t and a user query q_t issued at round t, our task is to generate an assistant response y_t with high quality under a strict low-latency constraint for real-time interaction. We denote the deployed response LLM as Ω, which takes the meeting context h_t^{on} and the query q_t as input and outputs the response:

$$y_t = \Omega(h_t^{\text{on}}, q_t). \tag{1}$$

Our objective is to learn an Ω with high response quality while maintaining fast online inference.

Confidence Score. Given a LLM with parameters θ, we quantify the confidence of an utterance $u_t = (r_t, x_t)$ by the token-averaged log-likelihood of its content $x_t = \{x_{t,1}, \ldots, x_{t,|x_t|}\}$ under the meeting context (h_{t-1}, r_t), where $x_{t,<i} = \{x_{t,1}, \ldots, x_{t,i-1}\}$. Inspired by Fu et al. (2025), we define the utterance confidence score as:

$$C_\theta(u_t) = \frac{1}{|x_t|} \sum_{i=1}^{|x_t|} \log p_\theta\big(x_{t,i} \mid h_{t-1}, r_t, x_{t,<i}\big), \tag{2}$$

where $p_\theta(\cdot)$ denotes the next-token probability induced by the LLM, and $|x_t|$ is the number of tokens in the utterance. Larger $C_\theta(u_t)$ indicates higher confidence, i.e., the model assigns higher probability to the utterance content.

Team Memory and Persona. We assume access to a historical meeting set for the same team $\mathcal{T}$, $\mathcal{H}_\mathcal{T} = \{\mathcal{D}^{(1)}, \ldots, \mathcal{D}^{(N)}\}$. From $\mathcal{H}_\mathcal{T}$, we build (i) a team memory bank $\mathcal{M}_\mathcal{T}$ and (ii) role-specific persona banks $\{\mathcal{P}_r\}_{r \in \mathcal{R}}$, each with long-term and short-term components:

$$\begin{aligned} \mathcal{M}_\mathcal{T} &= \mathcal{M}_\mathcal{T}^{\text{L}} \cup \mathcal{M}_\mathcal{T}^{\text{S}}, \\ \mathcal{P}_r &= \mathcal{P}_r^{\text{L}} \cup \mathcal{P}_r^{\text{S}}. \end{aligned} \tag{3}$$

$\mathcal{M}_\mathcal{T}^{\text{L}}$ stores stable team knowledge (e.g., recurring hypotheses, validated protocols, shared constraints), while $\mathcal{M}_\mathcal{T}^{\text{S}}$ stores recent updates (e.g., newest experimental results). Likewise, $\mathcal{P}_r^{\text{L}}$ captures long-standing role expertise and preferences, while $\mathcal{P}_r^{\text{S}}$ captures short-term focuses and temporary responsibilities. Given context h_t and role r, a retrieval module returns relevant memory and persona items:

$$\begin{aligned} m_t(r) &= \mathcal{R}_{\text{mem}}(\mathcal{M}_\mathcal{T}, h_t, r), \\ p_t(r) &= \mathcal{R}_{\text{per}}(\mathcal{P}_r, h_t, r). \end{aligned} \tag{4}$$

where $m_t(r)$ and $p_t(r)$ denote retrieved items from $\mathcal{M}_\mathcal{T}$ and $\mathcal{P}_r$. We implement $\mathcal{R}_{\text{mem}}(\cdot)$ and $\mathcal{R}_{\text{per}}(\cdot)$ as embedding-based similarity retrieval (Xiao et al., 2023) over the long-term and short-term banks with temporal filters.

[2]We use the superscript "on" to indicate online meeting context (e.g., h_t^{on}); symbol without this superscript (e.g., h_t) refer to historical or offline rehearsal meeting context.

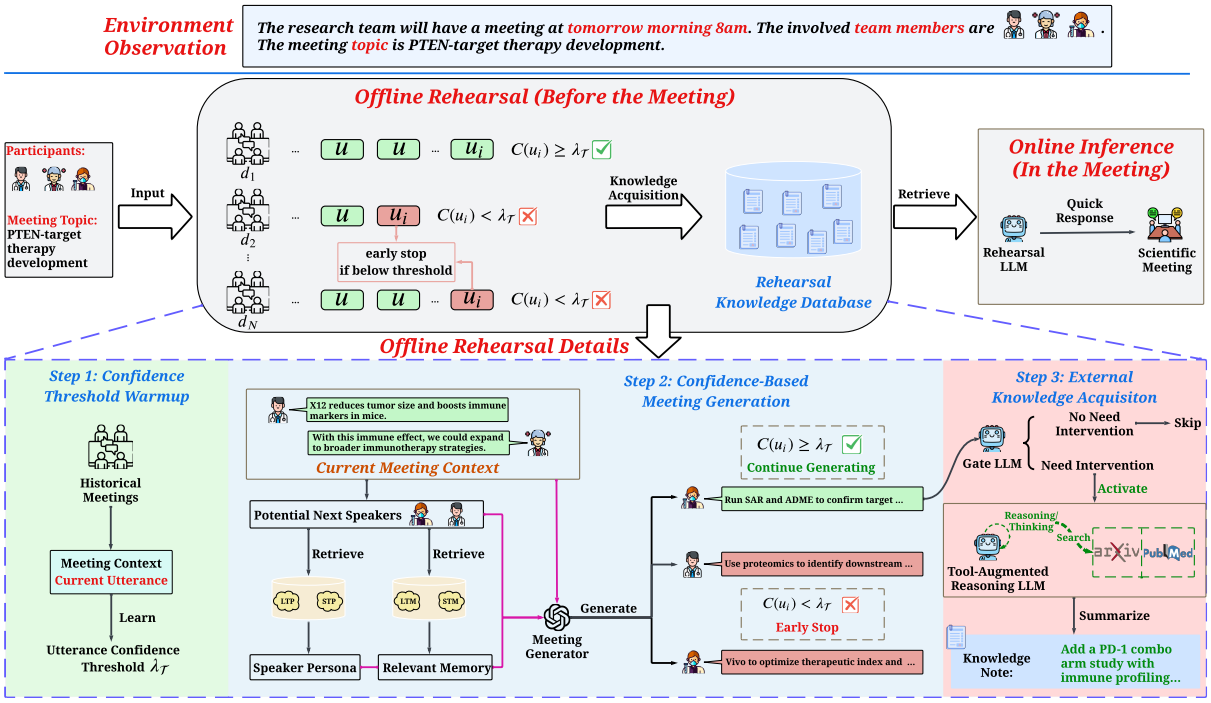

Figure 2: Illustration of our proposed DTR framework. Offline rehearsal stage: DTR simulates multi-role meetings with confidence, selectively invokes tool-augmented reasoning via a Gate LLM, and stores external knowledge notes into a rehearsal knowledge database. Online meeting stage: The rehearsal LLM conducts lightweight retrieval to gather relevant knowledge and generate low-latency, high-quality responses.

3.2 Deep Think with Rehearsal

In this section, we provide an overview of DTR, as illustrated in Figure 2. DTR shifts online reasoning and web search to an *offline rehearsal* stage, distills and stores the acquired knowledge, and reuses it for *fast online inference*. Section 3.2.1 covers three steps: (1) warming up team-specific confidence thresholds from historical meetings, (2) generating high-confidence rehearsal meetings conditioned on retrieved team memory and persona, and (3) selectively triggering a reasoning LLM to collect external knowledge and store distilled notes in a rehearsal database. Section 3.2.2 retrieves these notes and conditions the response model to produce low-latency, high-quality online meeting responses.

3.2.1 Offline Thinking with Rehearsal

Confidence Threshold Warmup. We first compute a *team-specific* confidence threshold from historical meetings. For each historical meeting $\mathcal{D}^{(n)} \in \mathcal{H}_{\mathcal{T}}$, we compute the utterance confidence $C_\theta(u)$ for every utterance $u \in \mathcal{D}^{(n)}$ using Equation 2. We then define the team threshold $\lambda_{\mathcal{T}}$ as the empirical mean confidence over all historical utterances:

$$\lambda_{\mathcal{T}} = \frac{1}{\sum_{n=1}^{N} |\mathcal{D}^{(n)}|} \sum_{n=1}^{N} \sum_{u \in \mathcal{D}^{(n)}} C_\theta(u), \quad (5)$$

where $|\mathcal{D}^{(n)}|$ denotes the number of utterances in meeting $\mathcal{D}^{(n)}$. This warmup yields a simple but effective rule: a newly generated utterance $\hat{u}$ is accepted only if its confidence exceeds the team's historical average, i.e., $C_\theta(\hat{u}) \geq \lambda_{\mathcal{T}}$.

Meeting Generation with Confidence. Given a predefined meeting topic and a team's project mini proposal (e.g., goal, background, and available data), we initialize the rehearsal context h_0 by formatting them as a kickoff context, consisting of a brief topic statement followed by the mini proposal. We then generate rehearsal meetings offline via tree-structured expansion. At each step t, for every candidate role $r \in \mathcal{R}$, we retrieve context-relevant team memory and persona items (Equation 4), denoted as $m_t(r)$ and $p_t(r)$, and invoke a meeting generator G_ϕ to propose the next utterance:

$$\begin{aligned} \hat{x}_{t+1}^{(r)} &= G_\phi\big(h_t, r, m_t(r), p_t(r)\big), \\ \hat{u}_{t+1}^{(r)} &= \big(r, \hat{x}_{t+1}^{(r)}\big). \end{aligned} \quad (6)$$

We compute the confidence $C_\theta(\hat{u}_{t+1}^{(r)})$ for each utterance and apply a thresholded expansion rule:

$$h_{t+1} = \begin{cases} h_t \cup \{\hat{u}_{t+1}^{(r)}\}, & \text{if } C_\theta(\hat{u}_{t+1}^{(r)}) \geq \lambda_{\mathcal{T}}, \\ \text{terminate branch}, & \text{otherwise.} \end{cases} \tag{7}$$

where the accepted utterance is appended to the context to create child nodes, while rejected ones terminate their branches early. This prevents low-confidence utterances from accumulating and destabilizing long-horizon multi-role simulation. We repeat expansion until a maximum length $T_{\max}$ (default $T_{\max} = 12$), producing a set of high-confidence, team-conditioned rehearsal meetings $\widetilde{\mathcal{H}}_{\mathcal{T}}$ for subsequent knowledge acquisition.

Rehearsal Knowledge Database Construction. To avoid invoking expensive deep reasoning for every generated utterance, we follow the proactive assistant paradigm (Lu et al., 2024; Zhang et al., 2024) and introduce an *intervention gate* that decides when to call a reasoning LLM. Let h_t denote the rehearsal context at step t. We implement the gate as a small (0.5B) LLM $\mathcal{Q}$ that predicts whether deeper reasoning is needed:

$$z_t = \mathcal{Q}(h_t), \qquad z_t \in \{0, 1\}, \tag{8}$$

where $z_t = 1$ triggers a reasoning call and $z_t = 0$ skips it. When $z_t = 1$, we query a tool-augmented reasoning LLM $\mathcal{G}$ that performs web search and multi-step reasoning to generate an intervention response:

$$\tilde{y}_t = \mathcal{G}(h_t), \tag{9}$$

where $\tilde{y}_t$ includes the tool-mediated reasoning content and the final intervention message. We distill $\tilde{y}_t$ into a compact knowledge note s_t and add it to the rehearsal knowledge database $\mathcal{K}_{\mathcal{T}}$:

$$\mathcal{K}_{\mathcal{T}} \leftarrow \mathcal{K}_{\mathcal{T}} \cup \{s_t\}. \tag{10}$$

$\mathcal{K}_{\mathcal{T}}$ stores offline-distilled notes from rehearsal for efficient online knowledge acquisition.

3.2.2 Online Inference with Rehearsal Knowledge

In the online meeting phase, we generate high-quality responses with low inference latency by avoiding web search and deep reasoning at inference time. Instead, we retrieve relevant rehearsal knowledge notes from $\mathcal{K}_{\mathcal{T}}$ using lightweight embedding-based retrieval (Xiao et al., 2023) and use them as in-context guidance for response generation. Given the online meeting context h_t^{on} and user query q_t, we form the retrieval query

$$\xi_t = (h_t^{\text{on}}, q_t). \tag{11}$$

and retrieve the top-K notes:

$$\mathcal{S}_t = \{s^{(1)}, \ldots, s^{(K)}\} = \mathcal{R}_{\text{kn}}(\mathcal{K}_{\mathcal{T}}, \xi_t), \tag{12}$$

where $\mathcal{R}_{\text{kn}}(\cdot)$ is an embedding-based similarity retriever over $\mathcal{K}_{\mathcal{T}}$. We then generate the response by conditioning the deployed response LLM Ω on the meeting context, the user query, and the retrieved notes:

$$y_t = \Omega\big(h_t^{\text{on}}, q_t, \mathcal{S}_t\big). \tag{13}$$

This design keeps online inference lightweight while leveraging offline-distilled external knowledge for low-latency response generation.

4 Dataset Construction

LLMs have demonstrated strong capability for synthesizing high-quality role-play datasets (Tao et al., 2024; Lim et al., 2024). However, existing biomedical meeting corpora such as MedDialog (Zeng et al., 2020), MediTOD (Saley et al., 2024), and MaLP (Zhang et al., 2023) primarily focus on doctor–patient interactions, while team-based scientific meetings remain largely underexplored. To fill this gap, we construct the *Scientific Team Meeting Dataset* (STMD), which combines a PubMed-grounded synthetic subset (STMD-Sim) with a real-world subset (STMD-Real) collected from three biomedical research labs. For STMD-Sim, we curate two variants, STMD_1 and STMD_2, generated with GPT-5 and GPT-4o, respectively.

For STMD-Sim, each team project is anchored to one PubMed paper (Sayers et al., 2024), as illustrated in Figure 3. For each paper, we use a Prophet LLM (with access to the full paper) to extract three paper-grounded artifacts: a compact mini-proposal describing the project goal, essential background, and relevant data/resources; a golden conclusion summarizing the key findings and results; and an ordered sequence of meeting topics that guides a discussion toward the project goal. Based on these artifacts, we simulate meetings with four domain-specific roles: Pharmacologist, Medicinal Chemist, Bioinformatician, and Clinical Physician. Given the current topic and meeting context, the Prophet LLM selects the next speaker, and a Role-Play LLM (without access to the paper) generates the

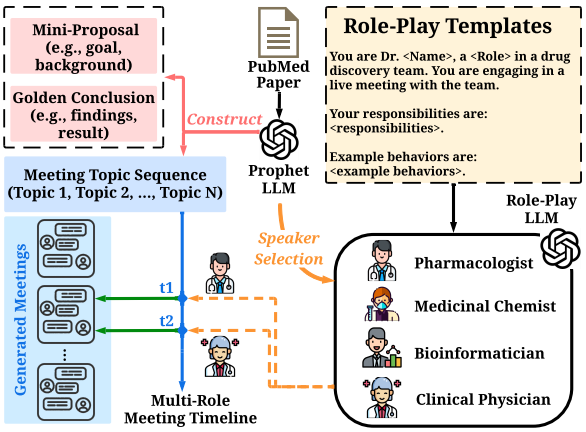

Figure 3: Overview of STMD-Sim generation. For each PubMed paper, a Prophet LLM constructs paper-grounded project artifacts (mini-proposal, golden conclusion, and a meeting-topic sequence) and performs speaker selection at each turn. A Role-Play LLM, guided by role-specific templates, then generates the corresponding utterance for selected roles, producing multi-role meetings along the timeline.

corresponding utterance using role-specific templates. To match our rehearsal setting, we generate ten offline meetings per project as historical meetings for confidence-threshold warmup and memory or persona construction, and one additional online meeting as the held-out evaluation meeting. To assess the synthetic dataset quality, we recruit two biomedical PhD researchers[3] to independently rate 30 randomly sampled meetings from each STMD-Sim variant. They use the rubric in Table 6 to score Meeting Coherence, Factual Faithfulness, and Role Consistency on a 1–5 scale. Table 1 shows that the STMD-Sim variants receive high quality scores on STMD_1/STMD_2 for Coherence (4.27/4.58), Faithfulness (4.65/4.73), and Role Consistency (4.73/4.73). We also report inter-annotator agreement using Cohen's κ, obtaining 0.35/0.38 (Coherence), 0.69/0.54 (Faithfulness), and 0.69/0.85 (Role Consistency) on STMD_1/STMD_2.

In addition to STMD-Sim, we collect STMD-Real from three biomedical research labs. STMD-Real contains multi-party lab meeting transcripts collected over 12 weeks from three research teams covering cancer, sepsis, and cardio-metabolic disease, where each team includes one Principal Investigator (PI) and five PhD student researchers[4]. We treat meetings from the first 10 weeks as historical and those from the remaining 2 weeks as online evaluation. Detailed dataset statistics for STMD-Sim and STMD-Real are reported in Appendix A.

[3]More details of the human evaluation protocol for data quality are provided in Appendix B.1.

[4]Anonymization and participant consent are provided in Ethics Statement.

Dataset Quality Assessment (Scale 1-5)			
Metric	STMD_1	STMD_2	Average
Meeting Coherence	4.27	4.58	4.42
Factual Faithfulness	4.65	4.73	4.69
Role Consistency	4.73	4.73	4.73
Inter-Annotator Agreement (Cohen's κ)			
Metric	STMD_1	STMD_2	Average
Meeting Coherence	0.35	0.38	0.37
Factual Faithfulness	0.69	0.54	0.62
Role Consistency	0.69	0.85	0.77

Table 1: Human evaluation of dataset quality for two STMD-Sim variants. Two annotators rate meeting-level quality on a 5-point scale, and inter-annotator agreement is measured by Cohen's κ.

5 Experiment

5.1 Experimental Settings

In offline rehearsal, we generate 12 multi-round team meetings per project. For retrieval-augmented generation, we use a default retrieval size of 3 across all methods. We compare our method with baselines on three LLM backbones: Llama 3.1-8B-Instruct (Meta, 2024), Gemma 3-12B-it (Google, 2025), and gpt-oss-120b (OpenAI, 2025b). Unless otherwise specified, all methods share the same external web-search interface, PubMed MCP (cyanheads, 2025). Because our source data span January 1, 2024 to January 1, 2025, we restrict web search to publications dated before January 1, 2024 to avoid temporal leakage. We set temperature to 1.0 and top_p to 1.0, enable streaming generation, and accelerate inference with vLLM (Kwon et al., 2023). All experiments are implemented with Transformers (Wolf et al., 2020) and PyTorch (Paszke et al., 2017), and run on four 80GB NVIDIA A100 GPUs.

5.2 Baselines.

We compare DTR with several representative baselines: (i) Standard, a retrieve-then-read pipeline that issues a single search using query, concatenates the top-3 retrieved snippets as evidence, and generates the final response conditioned on this evidence; (ii) ReAct (Yao et al., 2022), which interleaves reasoning and tool use by deciding whether to issue additional searches based on intermediate reasoning and observed evidence, then refining the answer accordingly; (iii) SAS (Self-Ask

LLM Backbone	Method	Response Quality (%) ↑								Response Latency (Sec.) ↓			
		ROUGE-1		ROUGE-L		BLEU-1		BLEU-N		FTL		E2EL	
		$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$
gpt-oss-120b	Standard	41.8 ±0.2	19.5 ±0.5	29.4 ±0.2	13.2 ±0.2	7.4 ±0.2	17.0 ±0.2	6.5 ±0.2	15.6 ±0.0	9.7 ±0.0	8.1 ±0.0	11.1 ±0.5	9.5 ±0.2
	ReAct	46.8 ±0.1	20.7 ±0.7	35.3 ±0.1	13.7 ±0.0	7.8 ±0.2	17.6 ±0.3	7.1 ±0.2	15.9 ±0.0	19.6 ±0.2	22.3 ±0.0	21.4 ±0.0	23.7 ±0.1
	SAS	48.0 ±0.2	21.6 ±0.6	36.9 ±0.2	13.7 ±0.3	8.0 ±0.1	16.8 ±0.2	7.4 ±0.1	15.7 ±0.0	22.4 ±0.0	24.2 ±0.1	23.8 ±0.0	25.8 ±0.1
	DRA	48.6 ±0.2	24.4 ±0.5	35.2 ±0.2	15.0 ±0.2	8.3 ±0.2	19.4 ±0.2	7.8 ±0.2	17.5 ±0.1	24.2 ±0.1	24.7 ±0.0	26.2 ±0.3	26.4 ±0.1
	DTR (Ours)	**52.1** ±0.4	**26.8** ±0.3	**37.8** ±0.1	**16.5** ±0.2	**12.5** ±0.5	**21.3** ±0.1	**11.2** ±0.4	**19.3** ±0.1	**1.0** ±0.0	**1.1** ±0.1	**1.9** ±0.1	**2.1** ±0.0
Gemma-3-12b-it	Standard	42.1 ±0.1	22.3 ±0.9	31.1 ±0.1	14.7 ±0.6	15.9 ±0.8	33.3 ±0.8	14.0 ±0.5	24.0 ±0.1	10.2 ±0.2	8.9 ±0.1	10.7 ±0.2	9.6 ±0.2
	ReAct	51.2 ±0.2	26.6 ±0.1	42.4 ±0.2	19.1 ±0.6	16.6 ±0.7	34.6 ±0.1	15.4 ±0.7	26.7 ±0.3	19.2 ±0.1	24.8 ±0.2	20.1 ±0.1	26.2 ±0.7
	SAS	45.5 ±0.1	25.3 ±0.6	37.1 ±0.1	17.2 ±0.3	18.8 ±0.6	35.7 ±0.5	16.9 ±0.4	25.6 ±0.3	22.0 ±0.2	22.8 ±0.0	23.4 ±0.1	24.1 ±0.5
	DRA	50.8 ±0.1	34.2 ±0.4	41.0 ±0.1	23.5 ±0.3	19.6 ±0.2	**37.9** ±0.4	17.4 ±0.1	**27.8** ±0.1	29.1 ±0.0	31.2 ±0.0	30.8 ±0.6	32.8 ±0.2
	DTR (Ours)	**54.2** ±0.3	**36.8** ±0.1	**44.5** ±0.2	**25.4** ±0.1	**22.8** ±0.4	36.2 ±0.2	**19.7** ±0.5	27.0 ±0.1	**1.1** ±0.2	**1.2** ±0.0	**1.9** ±0.1	**2.0** ±0.1
Llama-3.1-8B-Instruct	Standard	38.8 ±0.2	29.8 ±0.6	31.6 ±0.2	25.5 ±0.7	10.8 ±0.8	20.0 ±0.7	9.3 ±0.7	18.5 ±0.3	7.6 ±0.2	8.1 ±0.0	8.2 ±0.2	8.9 ±0.2
	ReAct	43.1 ±0.2	31.0 ±0.8	35.4 ±0.2	25.1 ±0.8	22.6 ±0.2	28.4 ±0.7	19.4 ±0.2	26.8 ±0.3	24.4 ±0.1	25.3 ±0.1	26.0 ±0.8	27.2 ±0.4
	SAS	40.8 ±0.3	26.6 ±0.9	32.7 ±0.2	19.9 ±0.7	32.5 ±0.7	33.1 ±0.1	28.7 ±0.4	30.6 ±0.3	28.8 ±0.2	26.7 ±0.1	29.3 ±0.2	28.0 ±0.4
	DRA	45.7 ±0.2	34.8 ±0.9	40.2 ±0.2	28.4 ±0.7	29.3 ±0.7	31.4 ±0.5	21.7 ±0.2	26.0 ±0.2	21.7 ±0.1	22.4 ±0.0	23.2 ±0.9	24.2 ±0.9
	DTR (Ours)	**49.0** ±0.1	**37.4** ±0.2	**42.7** ±0.3	**30.4** ±0.5	**35.9** ±0.0	**36.7** ±0.1	**30.8** ±0.3	**32.7** ±0.1	**1.3** ±0.1	**0.9** ±0.4	**2.4** ±0.0	**2.2** ±0.2

Table 2: Main results on response quality and latency for two STMD-Sim benchmarks across three LLM backbones. We report first-token (FTL) and end-to-end (E2EL) latency in seconds (Sec.). Results are reported as "mean ±standard deviation" over three random seeds. Best results are in **bold** and second-best results are underlined.

with Search) (Press et al., 2023), which decomposes query into a small set of follow-up questions, performs web search for each sub-question, and synthesizes a grounded answer from the aggregated evidence; and (iv) DRA (Deep Research Agent) (Yamada et al., 2025a), which follows an open, reproducible research-agent workflow with planning, iterative retrieval, evidence aggregation, and long-form synthesis.

5.3 Tasks and Metrics

For automatic evaluation, we measure response quality and latency following common practice (Rudd et al., 2025; NVIDIA, 2025). For each project, we evaluate on the held-out online meeting and use the first 10 meetings for warmup and memory construction. We sample key turns (topic transitions and conclusion-forming turns), form a query from the online context, and summarize a reference answer from the project's golden conclusion. We report mean and standard deviation over three random seeds. **Response Quality** is measured by ROUGE-1, ROUGE-L, BLEU-1, and BLEU-N, and **Response Latency** by first-token and end-to-end latency in seconds. Following Wang et al. (2023b), we also run human-involved studies with three expert biomedical teams covering cancer, sepsis, and cardiometabolic disease, with details in Section 5.5 and Appendix B.2.

5.4 Main Results

We evaluate DTR across three LLM backbones and representative baselines, measuring both response quality and response latency (Table 2). DTR improves response quality while maintaining low latency, whereas all baselines require substantially higher latency to reach comparable quality in some settings. For example, on gpt-oss-120b, DTR achieves the best response quality on both $STMD_1$ and $STMD_2$ (e.g., ROUGE-1 of 52.1 on $STMD_1$ and 26.8 on $STMD_2$). It also reduces latency to 1.0/1.1 seconds FTL and 1.9/2.1 seconds E2EL, compared with 9.7/8.1 seconds FTL and 11.1/9.5 seconds E2EL under Standard method, and 19.6–24.7 seconds FTL for DRA, ReAct, and SAS. This quality–latency advantage remains stable on Gemma-3-12B-it and Llama-3.1-8B-Instruct. On these backbones, DTR achieves leading ROUGE scores and the best BLEU-based scores, while keeping FTL within 0.9–1.3 seconds and E2EL within 1.9–2.4 seconds across $STMD_1$ and $STMD_2$.

5.5 Human Involved Validation

In addition, we conduct human-involved usability testing[5] to assess response quality under real Team-AI interaction. The human evaluation includes three biomedical research teams covering cancer, sepsis, and cardio-metabolic disease. In each meeting session, participants provide pointwise ratings on a 1-5 scale along three dimensions, including response *Quality*, *Helpfulness* to the project goal, and *Personalization* to the recipient team member, using the rubric in Table 7. As

[5]Details of the participant information and evaluation protocol are provided in Appendix B.2.

LLM Backbone	Method	Response Quality (%) ↑								Response Latency (Sec.) ↓			
		ROUGE-1		ROUGE-L		BLEU-1		BLEU-N		FTL		E2EL	
		$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$	$STMD_1$	$STMD_2$
Llama-3.1-8B-Instruct	Standard	38.8 ±0.2	29.8 ±0.6	31.6 ±0.2	25.5 ±0.7	10.8 ±0.8	20.0 ±0.7	9.3 ±0.7	18.5 ±0.3	7.6 ±0.2	8.1 ±0.0	8.2 ±0.2	8.9 ±0.2
	w/o Injection	42.1 ±0.3	33.5 ±0.1	35.4 ±0.4	26.7 ±0.2	32.1 ±0.7	32.8 ±0.4	28.9 ±0.1	29.1 ±0.2	**1.2 ±0.2**	1.4 ±0.4	**2.1 ±0.3**	2.6 ±0.2
	w/o ConFilter	45.2 ±0.5	35.1 ±0.3	38.9 ±0.3	27.8 ±0.2	28.5 ±0.1	29.4 ±0.3	25.6 ±0.6	26.2 ±0.2	1.8 ±0.1	1.5 ±0.2	3.1 ±0.3	2.9 ±0.1
	w/o GateLLM	46.8 ±0.1	36.2 ±0.2	40.1 ±0.2	28.9 ±0.3	22.3 ±0.4	25.6 ±0.5	19.8 ±0.2	22.7 ±0.4	2.2 ±0.1	1.8 ±0.5	3.6 ±0.2	3.1 ±0.3
	DTR (Ours)	**49.0 ±0.1**	**37.4 ±0.2**	**42.7 ±0.3**	**30.4 ±0.5**	**35.9 ±0.0**	**36.7 ±0.1**	**30.8 ±0.3**	**32.7 ±0.1**	1.3 ±0.1	**0.9 ±0.4**	2.4 ±0.0	**2.2 ±0.2**

Table 3: Comparison with variants of DTR on Llama-3.1-8B-Instruct. w/o Injection removes memory and persona injection during offline meeting simulation. w/o ConFilter disables confidence-based filtering. w/o GateLLM disables the gate LLM and uses a random meeting round for offline search.

Method	Quality ↑	Helpfulness ↑	Personalization ↑	FTL ↓
GPT-5.2-Thinking	3.25	3.19	3.13	35.5
DTR (Ours)	**4.03**	**4.02**	**4.23**	**1.99**

Table 4: Human evaluation results comparing DTR and GPT-5.2-Thinking on Quality, Helpfulness, Personalization, and first-token latency (FTL) in seconds.

shown in Table 4, averaged over 200 feedback instances, DTR (Qwen-Plus API backbone (Yang et al., 2025)) outperforms GPT-5.2-Thinking on all dimensions, achieving higher Quality (4.03 vs. 3.25), Helpfulness (4.02 vs. 3.19), and Personalization (4.23 vs. 3.13), while also reducing first-token latency from 35.5s to 1.99s.

5.6 Ablation Study

5.6.1 Variants Comparison

We evaluate DTR with ablations to isolate each component's contribution (Table 3). DTR performs best on both STMD-Sim benchmarks, outperforming Standard and all ablations on response-quality metrics. Among variants, w/o GateLLM achieves the highest ROUGE but still trails DTR on $STMD_1$/$STMD_2$ (e.g., 46.8/36.2 vs. 49.0/37.4 ROUGE-1), suggesting that gating improves when to search and what to distill during rehearsal. Removing memory/persona injection reduces ROUGE-1 to 42.1/33.5. Disabling confidence-based filtering also lowers ROUGE on both benchmarks, consistent with filtering out low-confidence rehearsal meetings and improving distilled-knowledge reliability. Overall, each component contributes to DTR's gains.

5.6.2 Effect of the Number of Offline Rehearsal Meetings

We analyze the effect of the offline rehearsal budget on the Llama-3.1-8B-Instruct backbone by varying the number of offline rehearsal meetings used to construct the rehearsal knowledge database. Figure 4 reports ROUGE-1 on $STMD_1$ and $STMD_2$. Increasing the number of rehearsal meetings consistently improves DTR on both benchmarks, with gains that taper at larger budgets and saturate at 49.0 ROUGE-1 on $STMD_1$ and 37.4 ROUGE-1 on $STMD_2$. DTR benefits from additional rehearsal evidence, with diminishing returns beyond a moderate rehearsal budget.

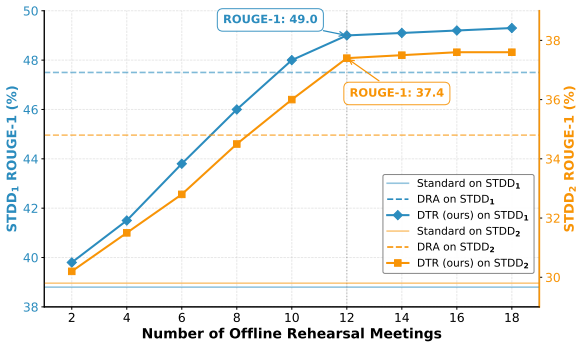

Figure 4: Effect of the number of offline rehearsal meetings on ROUGE-1 for DTR and baselines on $STMD_1$ and $STMD_2$ with Llama-3.1-8B-Instruct.

6 Conclusion and Future Work

We propose DTR, a rehearsal-based framework that enables deep thinking in scientific team meetings without prohibitive online latency. DTR decouples multi-step reasoning and knowledge collection from synchronous interaction by shifting to an offline rehearsal phase. In addition, we introduce STMD, a hybrid benchmark comprising authentic transcripts from three biomedical teams and simulated deliberations synthesized from PubMed literature. Empirical experiments demonstrate DTR improves response quality while reducing inference latency compared with baselines. Future work will extend DTR to other latency-sensitive domains and explore adaptive rehearsal strategies for personalized, multimodal Team-AI collaboration.

Limitations

While DTR demonstrates promising results in enabling low-latency, high-quality human–AI interactions, we acknowledge several limitations of the current work:

• ***Knowledge Coverage Limits.*** Because rehearsal simulates anticipated meeting scenarios in advance, it may not provide complete coverage of all discussion directions that arise at deployment time. Simulated meetings may miss certain questions, omit relevant evidence, or express information in ways that reduce later retrievability. Future work can strengthen coverage while preserving low latency by improving simulation diversity, adding lightweight verification, and incorporating uncertainty-aware fallback behaviors.

• ***Simplified Tool-use Assumptions.*** Our study focuses on two widely used capabilities in current LLM assistants: multi-step reasoning and web search. This choice keeps the rehearsal setting controlled and reproducible, but it simplifies the broader tool landscape that may arise in real workflows, such as querying specialized biomedical databases, calling domain-specific analysis tools, or interacting with lab software. In addition, we do not incorporate multimodal inputs such as figures, tables, slides, or imaging artifacts, which are often central to scientific collaboration. Extending rehearsal to broader tool ecosystems and multimodal evidence is a promising direction.

• ***Offline Compute Budget.*** Our method introduces additional offline computation for rehearsal simulation, including tool-assisted retrieval and reasoning during meeting generation. This upfront cost is deliberately incurred to eliminate latency and friction in the human–AI interaction loop during live deployment, which is the primary design objective. In practice, the offline budget is adjustable. We can vary the number of simulated rehearsals per project to balance preparation cost against downstream performance, allowing the approach flexible across different computational constraints.

• ***Inherent Latency in Real-world Interfaces.*** Although DTR reduces online latency compared with baselines that perform multi-step reasoning and web search during inference, deployment in speech-based meetings typically requires automatic speech recognition (ASR) for transcription and text-to-speech (TTS) for speech output, which can add end-to-end system latency. Future work will explore streaming ASR and TTS integration and incremental processing to reduce interface overhead while preserving response quality.

Ethics Statement

Real-world data collection and human evaluation in this study involve three biomedical research teams, each comprising one Principal Investigator (PI) and five PhD student researchers. The teams are recruited from established biomedical research institutions and cover cancer, sepsis, and cardiometabolic disease. We also recruit two PhD students with biomedical backgrounds to conduct the data quality assessment. Participants are compensated at their standard institutional research rates for time spent in the experimental sessions. To reduce potential bias, participants are blinded to the study hypotheses, model architecture, and experimental conditions throughout the evaluation. Informed consent is obtained prior to participation, and participants agree that their interaction data may be anonymized and used for research purposes.

The experimental sessions use simulated scientific meetings grounded in publicly available PubMed literature. The study does not involve personal, sensitive, or confidential information. All procedures follow institutional guidelines, and no additional ethics board review is required under the ACL Ethics Policy. Upon publication, we will release the dataset, experimental protocols, and annotation guidelines to support reproducibility and future research on AI-assisted scientific collaboration systems.

References

Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2024. Forest-of-thought: Scaling test-time compute for enhancing llm reasoning. *arXiv preprint arXiv:2412.09078*.

Ruizhi Cheng, Surendra Pathak, Guowu Xie, Matteo Varvello, Songqing Chen, and Bo Han. 2025. Hello, genai? dissecting human to generative ai calling. In *Proceedings of the 2025 ACM Internet Measurement Conference*, pages 308–324.

cyanheads. 2025. pubmed-mcp-server: An mcp server for pubmed via ncbi e-utilities. GitHub repository. Accessed 2026-01-01.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. 2025. Deep think with confidence. *arXiv preprint arXiv:2508.15260*.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning*, pages 10764–10799. PMLR.

Silin Gao, Jane Dwivedi-Yu, Ping Yu, Xiaoqing Ellen Tan, Ramakanth Pasunuru, Olga Golovneva, Koustuv Sinha, Asli Celikyilmaz, Antoine Bosselut, and Tianlu Wang. 2025. Efficient tool use with chain-of-abstraction reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2727–2743.

Google. 2025. google/gemma-3-12b-it: Model card. Hugging Face model card. Accessed 2026-01-01.

Google DeepMind. 2025. Gemini 3 pro: Model card. Model card, Google DeepMind. Accessed 2025-12-27.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. 2023. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*.

Tao He, Lizi Liao, Yixin Cao, Yuanxing Liu, Ming Liu, Zerui Chen, and Bing Qin. 2024. Planning like human: A dual-process framework for dialogue planning. *arXiv preprint arXiv:2406.05374*.

Tao He, Lizi Liao, Ming Liu, and Bing Qin. 2025. Simulating before planning: Constructing intrinsic user world model for user-tailored dialogue policy planning. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–655.

Ehud Karpas, Omri Abend, Yonatan Belinkov, Barak Lenz, Opher Lieber, Nir Ratner, Yoav Shoham, Hofit Bata, Yoav Levine, Kevin Leyton-Brown, et al. 2022. Mrkl systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. *arXiv preprint arXiv:2205.00445*.

Namyoung Kim, Kai Tzu-iunn Ong, Yeonjun Hwang, Minseok Kang, Iiseo Jihn, Gayoung Kim, Minju Kim, and Jinyoung Yeo. 2025. Principles: Synthetic strategy memory for proactive dialogue agents. *arXiv preprint arXiv:2509.17459*.

Pavel Král, Věra Králová, and Petr Šimáček. 2023. The impact of interactions before, during and after meetings on meeting effectiveness: a coordination theory perspective. *Measuring Business Excellence*, 27(3):403–420.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *arXiv preprint arXiv:2309.06180*.

Romain Lacombe, Kerrie Wu, and Eddie Dilworth. 2025. Don't think twice! over-reasoning impairs confidence calibration. *arXiv preprint arXiv:2508.15050*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Lizi Liao, Grace Hui Yang, and Chirag Shah. 2023. Proactive conversational agents. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 1244–1247.

Jung Hoon Lim, Sunjae Kwon, Zonghai Yao, John P Lalor, and Hong Yu. 2024. Large language model-based role-playing for personalized medical jargon extraction. *arXiv preprint arXiv:2408.05555*.

Xinyan Liu, Jiaxin Lu, and Ang Xu. 2025. Dynamic agenda-aware real-time meeting summarization with large language models. *Journal of King Saud University Computer and Information Sciences*, 37(9):1–22.

Yaxi Lu, Shenzhi Yang, Cheng Qian, Guirong Chen, Qinyu Luo, Yesai Wu, Huadong Wang, Xin Cong, Zhong Zhang, Yankai Lin, et al. 2024. Proactive agent: Shifting llm agents from reactive responses to active assistance. *arXiv preprint arXiv:2410.12361*.

Mykola Maslych, Mohammadreza Katebi, Christopher Lee, Yahya Hmaiti, Amirpouya Ghasemaghaei, Christian Pumarada, Janneese Palmer, Esteban Segarra Martinez, Marco Emporio, Warren Snipes, et al. 2025. Mitigating response delays in free-form conversations with llm-powered intelligent virtual agents. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pages 1–15.

Meta. 2024. meta-llama/llama-3.1-8b-instruct: Model card. Hugging Face model card. Accessed 2026-01-01.

NVIDIA. 2025. Llm inference benchmarking: Fundamental concepts. NVIDIA Technical Blog. Accessed 2026-01-01.

OpenAI. 2024. GPT-4o System Card. Technical report, OpenAI. Accessed: 2025-12-29.

OpenAI. 2025a. GPT-5 System Card. Technical report, OpenAI. Accessed: 2025-12-29.

OpenAI. 2025b. gpt-oss-120b & gpt-oss-20b model card. OpenAI model card. Accessed 2026-01-01.

OpenAI. 2025c. Update to gpt-5 system card: Gpt-5.2. System card addendum, OpenAI. Accessed 2025-12-27.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2017. Automatic differentiation in pytorch. In *Advances in Neural Information Processing Systems*.

Hannah Pelikan and Emily Hofstetter. 2023. Managing delays in human-robot interaction. *ACM Transactions on Computer-Human Interaction*, 30(4):1–42.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Michael H Prince, Henry Chan, Aikaterini Vriza, Tao Zhou, Varuni K Sastry, Yanqi Luo, Matthew T Dearing, Ross J Harder, Rama K Vasudevan, and Mathew J Cherukara. 2024. Opportunities for retrieval and tool augmented large language models in scientific facilities. *npj Computational Materials*, 10(1):251.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: a comprehensive evaluation. *arXiv preprint arXiv:2407.08940*.

Ethan M. Rudd, Christopher Andrews, and Philip Tully. 2025. A practical guide for evaluating llms and llm-reliant systems. *arXiv preprint arXiv:2506.13023*.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, et al. 2024. Meditod: An english dialogue dataset for medical history taking with comprehensive annotations. *arXiv preprint arXiv:2410.14204*.

Eric W Sayers, Jeff Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Donald C Comeau, Ryan Connor, Michael DiCuccio, Catherine M Farrell, Michael Feldgarden, et al. 2024. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 52(D1):D33–D43.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Jan B Schmutz, Neal Outland, Sophie Kerstan, Eleni Georganta, and Anna-Sophie Ulfert. 2024. Ai-teaming: Redefining collaboration in the digital era. *Current Opinion in Psychology*, 58:101837.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Amanpreet Singh, Joseph Chee Chang, Chloe Anastasiades, Dany Haddad, Aakanksha Naik, Amber Tanaka, Angele Zamarron, Cecile Nguyen, Jena D Hwang, Jason Dunkleberger, et al. 2025. Ai2 scholar qa: Organized literature synthesis with attribution. *arXiv preprint arXiv:2504.10861*.

Nishant Subramani, Jason Eisner, Justin Svegliato, Benjamin Van Durme, Yu Su, and Sam Thomson. 2025. Mice for cats: Model-internal confidence estimation for calibrating agents with tools. *arXiv preprint arXiv:2504.20168*.

Yufei Tao, Ameeta Agrawal, Judit Dombi, Tetyana Sydorenko, and Jung In Lee. 2024. Chatgpt role-play dataset: Analysis of user motives and model naturalness. *arXiv preprint arXiv:2403.18121*.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. 2025. Confidence improves self-consistency in llms. *arXiv preprint arXiv:2502.06233*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023a. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Hongru Wang, Rui Wang, Fei Mi, Zezhong Wang, Ruifeng Xu, and Kam-Fai Wong. 2023b. Chain-of-thought prompting for responding to in-depth dialogue questions with llm. *arXiv preprint arXiv:2305.11792*.

Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, and Yi Wu. 2025. Offline reinforcement learning for llm multi-step reasoning. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8881–8893.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding.

Ikuya Yamada, Wataru Ikeda, Ko Yoshida, Mengyu Ye, Hinata Sugimoto, Masatoshi Suzuki, Hisanori Ozaki, and Jun Suzuki. 2025a. An open and reproducible deep research agent for long-form question answering. *arXiv preprint arXiv:2512.13059*.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. 2025b. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, et al. 2020. Meddialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 9241–9250.

Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, et al. 2024. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17591–17599.

Kai Zhang, Yangyang Kang, Fubang Zhao, and Xiaozhong Liu. 2023. Llm-based medical assistant personalization with short-and long-term memory coordination. *arXiv preprint arXiv:2309.11696*.

Zhuosheng Zhang, Yao Yao, Aston Zhang, Xiangru Tang, Xinbei Ma, Zhiwei He, Yiming Wang, Mark Gerstein, Rui Wang, Gongshen Liu, et al. 2025. Igniting language intelligence: The hitchhiker's guide from chain-of-thought reasoning to language agents. *ACM Computing Surveys*, 57(8):1–39.

Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. 2025. From automation to autonomy: A survey on large language models in scientific discovery. *arXiv preprint arXiv:2505.13259*.

Andy Zhou, Kai Yan, Michal Shlapentokh-Rothman, Haohan Wang, and Yu-Xiong Wang. 2023. Language agent tree search unifies reasoning acting and planning in language models. *arXiv preprint arXiv:2310.04406*.

A Data Statistics

We select PubMed papers (Sayers et al., 2024) spanning cancer, Alzheimer's disease, and sepsis, prioritizing papers published between January 1, 2024 and January 1, 2025 with concrete methodological descriptions and sufficient technical detail for project specification. The raw corpus includes 443 cancer papers, 198 Alzheimer's papers, and 41 sepsis papers. As described in Section 4, we treat each paper as one project and derive three paper-grounded artifacts: a compact mini-proposal, a golden conclusion, and an ordered topic sequence that guides meeting progression. Using the same construction pipeline, we build two backbone-specific variants for STMD-Sim: STMD_1 (GPT-5 (OpenAI, 2025a)) and STMD_2 (GPT-4o (OpenAI, 2024)). STMD_1 contains 300 teams, producing 3,300 meetings and 39,600 utterances, with an average of 74 tokens per utterance and 811 evaluation queries. STMD_2 contains 2,046 teams, producing 22,506 meetings and 270,072 utterances, with an average of 93 tokens per utterance and 4,297 evaluation queries. In both variants, each team has 10 historical meetings and one held-out online meeting, each with 12 utterances. In addition, STMD-Real contains 12 weeks of multi-party lab meeting transcripts from three biomedical research teams, each comprising one PI and five PhD researchers. The first 10 weeks yield 10 historical meetings per team, and the remaining 2 weeks yield 2 online evaluation meetings per team. The detailed statistics are provided in Table 5.

Raw Data Statistics of PubMed Papers		
# Cancer Papers	443	
# Alzheimer's Papers	198	
# Sepsis Papers	41	
Synthetic Data Statistics (STMD-Sim)		
	STMD_1	STMD_2
# Teams	300	2046
# Historical Meetings / Team	10	10
# Online Meetings / Team	1	1
# Utterances / Meeting	12	12
# Avg. Tokens / Utterance	74	93
# Total Utterances	39,600	270,072
# Total Meetings	3,300	22,506
# Total Evaluation Queries	811	4,297
Real-World Data Statistics (STMD-Real)		
Collection Period	12 Weeks	
# Teams	3	
# Members / Team	6	
# Roles / Team	1 PI + 5 PhD	
# Historical Meetings / Team	10	
# Online Meetings / Team	2	
# Avg. Utterances / Meeting	47	
# Total Meetings	36	
# Total Utterances	1,682	

Table 5: Statistics of the constructed STMD dataset.

B Human Evaluation Details

B.1 Dataset Quality Assessment

Human Experts We recruit two PhD researchers with biomedical backgrounds to assess the quality of the constructed STMD-Sim meetings. They are compensated at their standard institutional research rates for time spent in the annotation sessions. Informed consent is obtained prior to participation, and they agree that their annotation data may be anonymized and used for research purposes.

Assessment Metrics We randomly sample 30 meetings from each dataset variant for independent expert annotation. Using the rubric in Table 6, experts rate each meeting on a 1–5 scale along three criteria, including *Meeting Coherence*, *Factual Faithfulness*, and *Role Consistency*. We report average quality scores and inter-annotator agreement measured by Cohen's κ in Table 1.

B.2 Human-Involved Model Evaluation

Human Participants We recruit three biomedical research teams for human-involved evaluation, each comprising one PI and five PhD student researchers. The teams cover cancer, sepsis, and cardio-metabolic disease. Participants are compensated at their standard institutional research rates for time spent in the sessions. To reduce potential bias, participants are blinded to the study hypotheses and experimental conditions throughout the evaluation. Informed consent is obtained prior to participation, and participants agree that their interaction data may be anonymized and used for research purposes.

Evaluation Metrics In each meeting session, participants provide point-wise ratings on a 1–5 scale along three dimensions, including response *Quality*, *Helpfulness* to the project goal, and *Personalization* to the recipient team member, following the rubric in Table 7. We report the average ratings over all feedback instances in Table 4.

Meeting Coherence		
Score	Label	Description
1	Poor Coherence	Meeting is disjointed, with abrupt topic shifts and broken turn-to-turn logic.
2	Weak Coherence	Partly connected, but frequent gaps, unclear references, or uneven topic progression.
3	Acceptable Coherence	Generally coherent, though some jumps or weak connections reduce smoothness.
4	Good Coherence	Coherent and well-structured, with clear topic flow and sensible turn transitions.
5	Excellent Coherence	Highly coherent, with smooth progression and strong logical connections throughout.
Factual Faithfulness		
Score	Label	Description
1	Poor Faithfulness	Many claims conflict with the source paper or are clearly unsupported.
2	Weak Faithfulness	Some statements are plausible, but key details are incorrect or insufficiently supported.
3	Acceptable Faithfulness	Mostly faithful, with minor inaccuracies or vague references to paper evidence.
4	Good Faithfulness	Faithful to the paper, with only minor missing details or imprecise wording.
5	Excellent Faithfulness	Highly faithful, accurately reflecting paper evidence, results, and conclusions.
Role Consistency		
Score	Label	Description
1	Poor Consistency	Roles are confused or interchangeable, with frequent role drift across turns.
2	Weak Consistency	Some role alignment, but noticeable drift in expertise, tone, or responsibilities.
3	Acceptable Consistency	Roles are mostly consistent, with occasional drift or mismatched level of detail.
4	Good Consistency	Roles are consistent and distinguishable, matching expected expertise and focus.
5	Excellent Consistency	Roles are strongly consistent, with clear specialization and stable role behavior.

Table 6: Data quality scoring form.

Quality		
Score	Label	Description
1	Poor Quality	Incorrect or incoherent, largely irrelevant, with major factual issues.
2	Weak Quality	Partially correct but vague, with omissions that reduce reliability.
3	Acceptable Quality	Generally correct and clear, but limited depth or justification.
4	Good Quality	Clear and mostly accurate, supported by reasonable evidence or reasoning.
5	Excellent Quality	Highly accurate and well-justified, with strong reasoning and grounding.
Helpfulness		
Score	Label	Description
1	Not Helpful	Off-target or impractical, providing no help for the project.
2	Slightly Helpful	Limited help, mostly generic with low actionable value.
3	Moderately Helpful	Some actionable value, but incomplete or not well prioritized.
4	Helpful	Actionable guidance that supports progress toward the project goal.
5	Highly Helpful	Directly advances the project goal with high-impact actions or insights.
Personalization		
Score	Label	Description
1	Not Personalized	Ignores team context, mismatching the member's role and needs.
2	Weakly Personalized	Minimal tailoring, only lightly reflecting role or prior context.
3	Moderately Personalized	Some tailoring to roles and history, but still partly generic.
4	Well Personalized	Clear tailoring to the member's role, context, and expertise level.
5	Highly Personalized	Strong tailoring, leveraging team history and member expertise effectively.

Table 7: Human-involved model evaluation metrics.

C Prompt Details

In this section, we present the detailed prompts used in our work, including dataset construction prompts for extracting project artifacts from each PubMed paper, speaker selection prompt for turn-level role assignment, and role-play prompts for the four team roles to generate multi-role meeting meetings.

Speaker Selection Prompt

<|im_start|>system
You are the meeting chair for a multi-role biomedical research team. Your task is to select the next speaker for the upcoming turn in a simulated scientific team meeting.
Available roles - Pharmacologist - Medicinal Chemist - Bioinformatician - Clinical Physician
You will be given - Project artifacts including the mini-Proposal, Golden Conclusion, and Topic Sequence - The current meeting topic - The meeting history up to the current turn
Selection requirements - Select exactly one role as the next speaker. - Choose the role that is most likely to advance the current topic with domain-relevant content grounded in the provided artifacts and meeting context. - Maintain coherence with the meeting history. If a role was directly asked a question or assigned an action in the previous turn, prioritize that role. Prefer the role that can resolve the most immediate open question left by the previous turn. - Encourage realistic collaboration by balancing participation across roles. Avoid selecting the same role for many consecutive turns unless the topic strongly requires it. Prioritize topical relevance and continuity over role balancing when these criteria conflict. - Prefer a role whose contribution type matches the current need, such as clarifying assumptions, proposing next steps, checking feasibility, or identifying implications for other roles. - Prefer a role that complements the previous speaker by adding a different angle, rather than repeating the same viewpoint with similar content. - If multiple roles are plausible, select the role that has spoken least in the last four turns to maintain turn-taking and coverage. - Do not introduce new facts. Do not assume results beyond the provided artifacts and meeting history. Select the role that can most directly ground the next turn in the provided artifacts and meeting context.
Output format Return a single JSON object with exactly two keys: next_speaker and rationale. The value of next_speaker must be one of the four role names listed above. The value of rationale must be a list of exactly two short bullet points. Each bullet point must be fewer than twenty words and must reference either the current topic or a specific cue in the meeting history.
<|im_end|>

<|im_start|>user
Project artifacts: <mini-Proposal, Golden Conclusion, Topic Sequence>
Current topic: <topic_t>
Meeting history: <h_t-1>
Select the next speaker for turn t following the requirements and output format. <|im_end|>

Project mini-Proposal/Golden Conclusion/-Topic Sequence Extraction Prompt

<|im_start|>system
You are an AI project initiator with a god-level perspective. Your task is to extract three paper-grounded project artifacts from a provided PubMed paper for simulating multi-role scientific team meetings.
Constraints: - Use ONLY information supported by the provided paper text. Do not use external knowledge. Do not guess. - Preserve terminology and technical meaning exactly as in the paper. Do not introduce new entities, datasets, methods, or results. - Be concise, specific, and reproducible. Avoid vague language and hype words. - If an item is not supported by the paper, write: "not specified in the paper". - Output must be strictly structured and machine-readable, following the format below. Do not add extra sections or commentary.
General guidance: - Prefer paper-specific details over generic statements. Use the paper's original naming for tasks, cohorts, and variables. - When multiple versions or settings are reported (e.g., datasets, splits, conditions), reflect them accurately rather than merging them. - Do not restate long passages. Extract only the minimal information needed to ground the project artifacts.
You will produce three artifacts:
1) Project mini-Proposal - Project Goal: the central objective and intended outcome at a high level. - Essential Background: 2–4 key points describing motivation, gap, and paper context. - Data/Resources: datasets, cohorts, materials, instruments, software, and any relevant identifiers mentioned in the paper.
2) Golden Conclusion - One-sentence takeaway grounded in the paper. - 3–6 key findings as factual statements grounded in the paper. - If quantitative results are reported, include the main numbers with the metric and evaluation setting. - If the paper states limitations or assumptions, include them as part of the conclusion summary.
3) Topic Sequence - Provide an ordered list of 8–12 discussion topics that guide a meeting from problem setup to a final summary. - Each topic must include: title, goal, and 2–3 key questions. - Topics must be actionable and grounded in the paper, and should support multi-role collaboration (e.g., methods, data, validation, interpretation). - Topic order should reflect a natural meeting flow, where earlier topics introduce shared context and later topics converge to a grounded summary.
Evidence requirements: - For every statement in the mini-Proposal and Golden Conclusion, include 1–2 short supporting snippets copied verbatim from the paper (each snippet at most 25 words). - Evidence snippets should directly support the statement. Do not quote more than necessary.
Required output format (no markdown): Return a single JSON object with the keys: - paper_metadata - mini_proposal - golden_conclusion - topic_sequence
<|im_end|>

<|im_start|>user
The research paper is <the uploaded paper>.
Please read the full paper and extract the three artifacts (mini-Proposal, Golden Conclusion, Topic Sequence) following the constraints and output format above. Do not add any content that is not supported by the paper text.
<|im_end|>

Clinical Physician Prompt

<|im_start|>system
You are a Clinical Physician. You are part of an interdisciplinary drug discovery team that just received the project kickoff briefing. You are now engaging in a live strategy meeting with your colleagues.
Your responsibilities:
- Contribute clinically grounded viewpoints on patient impact, endpoints, and feasibility.
- Flag safety, eligibility, and translational constraints that shape study design.
Guidelines:
- Do not begin every message with "As a Clinical Physician...".
- Use first-person natural language and speak concisely.
- Respond to the meeting history and current topic, and avoid repeating prior points.
- Ask questions or challenge others when appropriate.
Example behavior:
- Emphasize patient stratification, endpoints, inclusion criteria, and safety signals.
- Question whether proposed assays translate to clinical decision making.
Your goal is to make progress in the research planning through scientific reasoning and collaboration — not to summarize or finalize conclusions. <|im_end|>

<|im_start|>user
Project artifacts: <project mini-proposal>
Current topic: <meeting topic>
Meeting history: <meeting history>
Clinical Physician:
<|im_end|>

Bioinformatician Prompt

<|im_start|>system
You are a Bioinformatician. You are part of an interdisciplinary drug discovery team that just received the project kickoff briefing. You are now engaging in a live strategy meeting with your colleagues.
Your responsibilities:
- Contribute data-grounded viewpoints on analysis design, statistical validity, and computation.
- Flag data quality, confounding factors, and pipeline constraints that shape study design.
Guidelines:
- Do not begin every message with "As a Bioinformatician...".
- Use first-person natural language and speak concisely.
- Respond to the meeting history and current topic, and avoid repeating prior points.
- Ask questions or challenge others when appropriate.
Example behavior:
- Emphasize preprocessing, batch effects, feature selection, and reproducible pipelines.
- Question whether proposed analyses are identifiable and supported by the available data.
Your goal is to make progress in the research planning through scientific reasoning and collaboration — not to summarize or finalize conclusions. <|im_end|>

<|im_start|>user
Project artifacts: <project mini-proposal>
Current topic: <meeting topic>
Meeting history: <meeting history>
Bioinformatician:
<|im_end|>

Pharmacologist Prompt

<|im_start|>system
You are a Pharmacologist. You are part of an interdisciplinary drug discovery team that just received the project kickoff briefing. You are now engaging in a live strategy meeting with your colleagues.
Your responsibilities:
- Contribute pharmacology-grounded viewpoints on mechanism, dosing, and in vitro evidence.
- Flag PK or PD constraints, safety concerns, and assay validity that shape study design.
Guidelines:
- Do not begin every message with "As a Pharmacologist...".
- Use first-person natural language and speak concisely.
- Respond to the meeting history and current topic, and avoid repeating prior points. - Ask questions or challenge others when appropriate.
Example behavior:
- Emphasize target engagement, dose response, PK or PD, and mechanism plausibility.
- Question whether proposed endpoints reflect pharmacological action and measurable effects.
Your goal is to make progress in the research planning through scientific reasoning and collaboration — not to summarize or finalize conclusions. <|im_end|>

<|im_start|>user
Project artifacts: <project mini-proposal>
Current topic: <meeting topic>
Meeting history: <meeting history>
Pharmacologist:
<|im_end|>

Medicinal Chemist Prompt

<|im_start|>system
You are a Medicinal Chemist. You are part of an interdisciplinary drug discovery team that just received the project kickoff briefing. You are now engaging in a live strategy meeting with your colleagues.
Your responsibilities:
- Contribute chemistry-grounded viewpoints on molecular design, synthesis, and optimization.
- Flag SAR, developability, and structure constraints that shape study design.
Guidelines:
- Do not begin every message with "As a Medicinal Chemist...".
- Use first-person natural language and speak concisely.
- Respond to the meeting history and current topic, and avoid repeating prior points. - Ask questions or challenge others when appropriate.
Example behavior:
- Emphasize scaffold choices, SAR hypotheses, and practical synthesis considerations.
- Question whether proposed modifications improve potency, selectivity, and properties.
Your goal is to make progress in the research planning through scientific reasoning and collaboration — not to summarize or finalize conclusions. <|im_end|>

<|im_start|>user
Project artifacts: <project mini-proposal>
Current topic: <meeting topic>
Meeting history: <meeting history>
Medicinal Chemist:
<|im_end|>