Into the *Unknown Unknowns*: Engaged Human Learning through Participation in Language Model Agent Conversations

Yucheng Jiang[1*] **Yijia Shao**[1*] **Dekun Ma**[2] **Sina J. Semnani**[1] **Monica S. Lam**[1]
[1]Stanford University [2]Yale University
{yuchengj, shaoyj}@cs.stanford.edu
dekun.ma@yale.edu, {sinaj, lam}@cs.stanford.edu

Abstract

While language model (LM)-powered chatbots and generative search engines excel at answering concrete queries, discovering information in the terrain of *unknown unknowns* remains challenging for users. To emulate the common educational scenario where children/students learn by listening to and participating in conversations with their parents/teachers, we create **Collaborative STORM** (**Co-STORM**). [1] Unlike QA systems that require users to ask all the questions, Co-STORM lets users observe and occasionally steer the discourse among several LM agents. The agents ask questions on the user's behalf, allowing the user to discover *unknown unknowns* serendipitously. To facilitate user interaction, Co-STORM assists users in tracking the discourse by organizing the uncovered information into a dynamic mind map, ultimately generating a comprehensive report as takeaways. For automatic evaluation, we construct the WildSeek dataset by collecting real information-seeking records with user goals. Co-STORM outperforms baseline methods on both discourse trace and report quality. In a further human evaluation [2], 70% of participants prefer Co-STORM over a search engine, and 78% favor it over a RAG (Retrieval Augmented Generation) chatbot.

1 Introduction

Recent advancements in language models (LMs) (Bai et al., 2022; OpenAI, 2023; Gemini Team, 2024) and retrieval-augmented generation (RAG) (Lewis et al., 2021) have led to more capable chatbots and emerging generative search engines (Liu et al., 2023a). Compared to traditional search engines and information retrieval (IR) models (Robertson, 1977), these systems fulfill user queries by generating direct responses, effectively addressing *known unknowns*, where users are aware of their information needs.

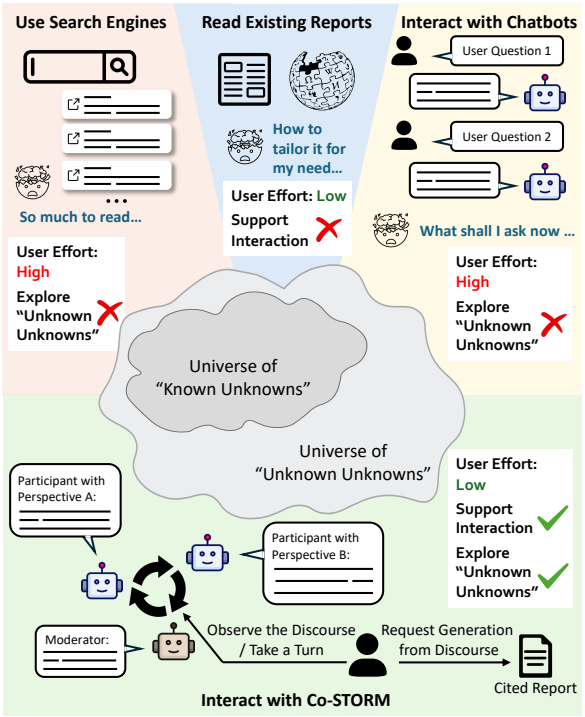

Figure 1: **Comparison of different paradigms for learning and information seeking.** Co-STORM enables humans to observe and participate in a collaborative discourse among LM agents with different roles. Users can request the system to generate a full-length cited report based on the discourse history and the information collected.

However, a gap remains in using these systems for complex information-seeking scenarios, such as academic research, market analysis, and decision-making, where the system should expose users to their *unknown unknowns* to facilitate knowledge discovery. While the concept of *unknown unknowns* originally referred to unexpected risks in the military, it is linked to the serendipitous discovery of information in the information research context (Foster and Ford, 2003; Agarwal, 2015). Specifically, Kirzner (1997) directly contrasts such

[*]Equal Contribution
[1]Our resources and code are released at https://github.com/stanford-oval/storm.
[2]A live research preview is now available publicly at https://storm.genie.stanford.edu

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 9917–9955
November 12-16, 2024

discovery ("the realization that one had overlooked something in fact readily available") with successful search ("the deliberate production of information which one knew one had lacked").

Prior work on automated expository writing attempts to help readers navigate the terrain of *unknown unknowns* by curating information from various sources into unified articles with substantial breadth and depth (Shen et al., 2023). In particular, the STORM writing system demonstrates that LMs paired with search engines can automatically generate a high-quality draft of Wikipedia-like articles on arbitrary topics (Shao et al., 2024). However, producing the static report as the final outcome, STORM does not support any user interaction which is crucial in complex information seeking where there is no single, gold query, but queries evolve dynamically towards a goal (Bates, 1989). This dynamic and exploratory nature makes designing assistance systems challenging. Traditional search engines and RAG chatbots passively react to users' search queries or conversational questions, often inducing echo chamber effects (Sharma et al., 2024) or high cognition load as users with limited prior knowledge may even struggle to formulate questions (Kuhlthau, 1991; Belkin et al., 1982).

To surface *unknown unknowns* and better support user interaction, we propose **Collaborative STORM** (**Co-STORM**), an information-seeking assistance system that supports collaborative discourse among users and multiple LM agents. Unlike the one-question-one-answer mode of interaction, Co-STORM allows users to learn by observing and participating occasionally in the discourse, emulating a common educational scenario (Nussbaum, 2008). To facilitate a thought-provoking discourse and serendipitous discovery, Co-STORM simulates two agent types grounded in the search engine: *experts* who participate by asking or answering questions with different perspectives and a *moderator*, a non-expert who knows enough to ask good questions and steers the discourse. The user can jump in at any time to steer the discourse and inject questions and opinions according to their interest. Co-STORM maintains a dynamic, hierarchical *mind map* to ensure they can easily follow and engage (Buzan, 1974). Upon the conclusion of the discourse, users can request the system to generate a cited report based on the mind map.

For evaluation, we introduce WildSeek, a dataset of topics and user goals from real users engaged in complex information-seeking across multiple domains. We propose automatic metrics to assess both discourse trace and final report quality. Our results show that Co-STORM significantly outperforms RAG chatbots in surfacing in-depth and serendipitous information while providing a more engaging learning experience.

We further conduct a human evaluation by inviting 20 users with diverse backgrounds to compare Co-STORM with a search engine and a RAG chatbot. 70% preferred Co-STORM over the search engine, and 78% preferred it over the RAG chatbot for the overall information-seeking experience. Participants find that Co-STORM facilitates serendipitous discovery and requires less mental effort.

Our main contributions include:

- We propose Co-STORM, a novel system that combines collaborative discourse emulation, human interaction, and information organization to assist learning and complex information seeking.
- We construct the WildSeek dataset from real-world human information-seeking records to evaluate information-seeking assistance tools.
- Results from both automatic and human evaluation show that Co-STORM can help humans discover *unknown unknowns* with less mental effort required.

2 Complex Information Seeking

2.1 Problem Formulation

Pirolli (2009) defines complex information-seeking as part of the broader sensemaking process, involving collecting, sifting, understanding, and organizing information from large collections to generate a knowledge product. Prevalent in domains such as investigative journalism, scientific research, and market analysis, this task has the following properties: (1) it requires seeking information from *multiple sources* to address various facets of a topic rather than retrieving a document that best matches a query; (2) it involves *ongoing user interaction* rather than processing a single query; (3) it produces *report-like curated information product* rather than a single short-form answer. As shown in Table 1, none of the existing information-seeking assistance systems (e.g., Robertson, 1977; Chen et al., 2017; Reddy et al., 2019; Shao et al., 2024) can fully support this task.

Given a user with an initial topic of interest t and an initial goal g, and a repository of information R,

	Multiple Sources	Ongoing Interact	Curated Report
Information Retrieval	✗	✗	✗
Single-Turn QA	✓	✗	✗
Conversational QA	✓	✓	✗
Report Generation	✓	✗	✓
Co-STORM	✓	✓	✓

Table 1: Comparison of different information-seeking assistance systems.

Domain: Economics
Topic: Development of a Shared Trading Currency to Facilitate International Trade
Goal: Investigate how a new shared currency could eliminate transaction costs and boost GDP among member countries.

Table 2: A sample data point in the WildSeek dataset for studying complex information-seeking tasks; the topic and goal are provided by users on the publicly available STORM website, the domain is assigned manually.

the task is to *interact* with the user and write a custom long-form report tailored to the users' interest; the report is a sequence of sentences $\mathcal{S} = s_1 s_2 ... s_n$, with each sentence citing a set of retrieved information $\mathcal{I} \subset \mathcal{R}$, for the sake of verifiability.

2.2 WildSeek: An In-the-Wild Information Seeking Dataset

To study users' interests in complex information-seeking tasks in the wild, we utilized data collected from the open-source STORM web application[3], which generates comprehensive long-form reports based on users' specified topics of interest and goals for using the site. Each data point is a pair comprising a topic and the user's goal. To improve the quality of the dataset, we retain only those data points that are well motivated by applying rule-based filtering followed by binary classification using an LM (`gpt-4o-2024-05-13`). Next, we use the same LM to predict the taxonomy class of each topic, followed by manual review and refinement. Finally, we downsample the data to create a dataset with 100 data points across 24 different domains. Table 2 shows a sample data point from the dataset; further details about the dataset are in Appendix A.

3 Method

"Tell me and I forget. Teach me and I remember. Involve me and I learn." — Benjamin Franklin

[3]Available at https://storm.genie.stanford.edu

3.1 Collaborative Discourse Protocol

Nussbaum (2008) emphasizes the importance of collaborative discourse in fostering deep understanding and critical thinking in human learning. Since it is difficult to assemble a group of human experts for collaborative discourse on any topic at any time, we propose Co-STORM (Figure 2) to emulate this process with multiple LM agents to assist human information seeking and learning. Formally, the collaborative discourse, $\mathcal{D} = \{u_1, u_2, ..., u_n\}$, consists of turn-based textual utterances u_i from one of three roles: the *user* (§3.3), *experts* with diverse perspectives (§3.4), and a *moderator* guiding the discourse and injecting questions (§3.5). The discourse begins with N experts, $\mathcal{P} = \{p_1, ..., p_N\}$, discussing the topic t for one turn per expert to warm up the conversation. Co-STORM dynamically maintains a mind map (§3.2) to track the discourse and construct shared knowledge between the user and the system.

Utterance Intent Inspired by the utterance intent taxonomy for information-seeking conversations proposed by Qu et al. (2019), we associate each agent utterance u_i with an agent intent type a_i, where a_i can be one of the following:

- ORIGINAL QUESTION initiates a new question
- INFORMATION REQUEST seeks additional information from the prior utterance
- POTENTIAL ANSWER offers a possible answer to a previously posed question
- FURTHER DETAILS provides supplementary information to a previous answer

We group ORIGINAL QUESTION and INFORMATION REQUEST as *question-asking* and the other two intents as *question-answering*.

Initiative Management Traum (2003) underscores the necessity of discourse management in multiparty dialogues. While existing systems support either just user initiatives (*e.g.*, QA systems) or just agent initiatives (*e.g.*, STORM), Co-STORM adopts a mixed-initiative approach. When the user actively engages, the system continues the discourse based on the user's question or argument, allowing for a more targeted discussion. Otherwise, the system automatically generates the next turn. The user controls who takes the initiative, as Co-STORM allows the user to take a turn anytime.

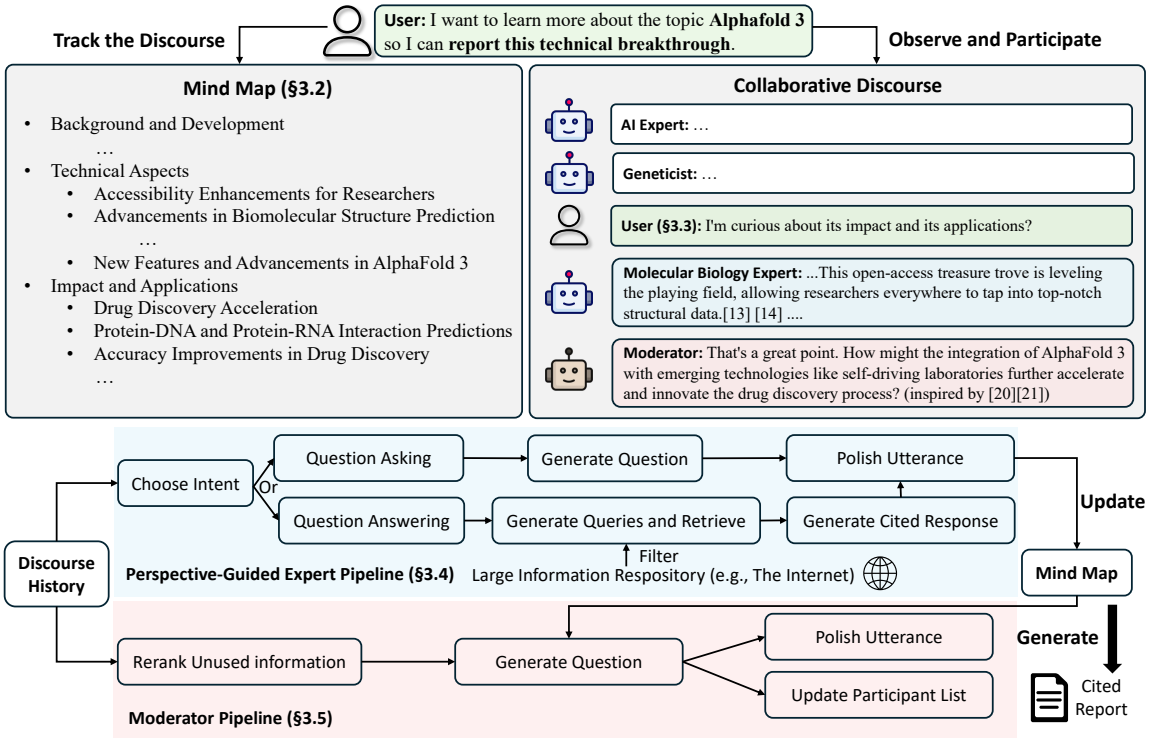

Figure 2: **Overview of Co-STORM.** Co-STORM emulates a collaborative discourse among the user, simulated perspective-guided experts, and a simulated moderator. It maintains a dynamically updated mind map (§3.2) to help user track and engage in the discourse (§3.3) . The simulated expert is prompted to determine the utterance intent based on discourse history and generate a question or an answer grounded in the Internet (§3.4). The simulated moderator is prompted with unused information and the mind map to generate a new question to automatically steer the discourse (§3.5). The mind map can be used to generate a full-length cited report as takeaways. Complete discourse transcript and the associated report are detailed in Appendix §G and §H.

Turn Management If the user does not take the turn at timestamp i, Co-STORM needs to determine which LM agent should generate the next utterance u_i. Its protocol is to let different experts, $p_1, ..., p_N$, take turns in sequence. To prevent them from only expanding on the same point, upon observing L consecutive turns of expert responses with intents being either POTENTIAL ANSWER or FURTHER DETAILS, the system asks the moderator to intervene. In §5.3, we analyze the benefit of this protocol design.

3.2 Tracking the Discourse with a Mind Map

Having shared knowledge or a shared conceptual space is critical for collaboration (Roschelle and Teasley, 1995). To help users track the discourse and reduce their cognition load, Co-STORM uses a tree-structured *mind map* $\mathcal{M}$ to dynamically organize collected information in the discourse $\mathcal{D}$. Specifically, $\mathcal{M} = (\mathcal{C}, \mathcal{E})$ is a hierarchical organization of concepts $\mathcal{C}$, where its directed edges $\mathcal{E}$ characterize latent parent-child relationships among topics (*e.g.*, in Figure 2, "Drug Discovery Acceleration" is a subtopic of "Impact and Applications"). Each concept $c \in \mathcal{C}$ is associated with a subset of retrieved information $I^c \subset \mathcal{I}$. To ensure $\mathcal{M}$ is an intent-driven organization of information, each piece of information is also associated with the question that leads to its retrieval.

Co-STORM dynamically updates the mind map through two operations, `insert` and `reorganize`. `Insert` places a piece of information under the most appropriate concept by first deriving a set of candidate concepts using semantic similarity between its associated question and each concept in $\mathcal{C}$, then prompting the LM to choose the final placement. When a concept c has more than K pieces of information, $\mathcal{M}$ triggers the `reorganize` operation. `Reorganize` prompts the LM to generate a list of new subtopic names under c, and applies `insert` to place each piece of information associated with c in the subtree rooted at c. After expansion, Co-STORM adopts a bottom-up cleaning process to iteratively delete concepts with no supporting information and collapse concepts with only one subtopic. More details on `insert` opera-

tion are included in Appendix B.

3.3 User Participation

When the user injects an utterance u, Co-STORM uses u as the query to retrieve information to prompt the LM to obtain an updated list of experts, $\mathcal{P}'$. Following this update, the system switches back to the auto-steering mode where the expert or the moderator takes turns according to the turn management protocol introduced in §3.1. Once the user is satisfied with the discourse, Co-STORM generates the final report $\mathcal{S}$ as the curated information product of the collaborative discourse. This report is generated using the mind map $\mathcal{M}$ as the outline and the retrieved information I^c associated with each concept c to generate the report section by section.

3.4 Simulating the Roundtable Participant

Following STORM (Shao et al., 2024), which uses perspective-guided question asking to improve the question diversity and quality, Co-STORM personalizes simulated experts with different expertise to represent different perspectives. Co-STORM retrieves the background of topic t with a search query and gives it to an LM to generate the expert list $\mathcal{P} = \{p_1, ..., p_N\}$. For example, for the topic "AlphaFold3" in Figure 2, the LM suggests an "AI Expert", a "Geneticist", and a "Molecular Biology Expert" to participate in the discourse.

If there is no interruption by the user or the moderator, each expert p_j sequentially takes turns with the following procedure: To generate an utterance from an expert p in turn i, (1) The LM is prompted to choose the intent a_i based on the discourse history $\{u_1, ..., u_{i-1}\}$ and the expert's perspective p. (2) If the intent a_i is POTENTIAL ANSWER or FURTHER DETAILS, we prompt the LM to generate a search query q, retrieve information with a search engine [4], and generate a response with citations; otherwise, we prompt the LM to directly generate a question based on the discourse history. (3) We use the LM to polish the utterance to make it more chatty and engaging.

3.5 Simulating the Moderator

If all the simulated participants are experts, we discover that the discourse tends to consist mostly of utterances with intent FURTHER DETAILS, leading to repetition and niche discussions. The moderator plays an important role of injecting new directions into the discourse. To generate the moderator's utterance, Co-STORM instructs the LM to generate informed questions based on the uncited sources retrieved since the last moderator turn. To choose among the many uncited sources, the moderator reranks each piece of information i based on the similarity to the topic t and the dissimilarity to its associated question q. Formally, the reranking score is

$$\cos(\mathbf{i}, \mathbf{t})^{\alpha}(1 - \cos(\mathbf{i}, \mathbf{q}))^{1-\alpha},$$

where $\mathbf{i}, \mathbf{t}, \mathbf{q}$ are corresponding text embeddings and α is a hyperparameter. This reranking function prioritizes information that does not directly answer the original question but is relevant to the topic t. Co-STORM concatenates these reranked sources along with concept names in $\mathcal{C}$ to avoid repetitive concepts. This combined context is used to prompt the LM to generate the question for the moderator turn and an updated list of experts, $\mathcal{P}'$.

4 Co-STORM Implementation

The LM component of Co-STORM is implemented using zero-shot prompting via the DSPy framework (Khattab et al., 2023) and `gpt-4o-2024-05-13` (see full prompts in Appendix D). We ground Co-STORM on the Internet using the You.com search API[5] although the system is compatible with other search engines or IR systems. Hyperparameters N, K, L, α are set to 3, 10, 2, and 0.5, respectively. The text embeddings in Eq. 1 are obtained from `text-embedding-3-small`. We set the LM temperature as 1.0 and top_p as 0.9 for all experiments. For human evaluation, we develop a web application (Figure 12) for users to interact with Co-STORM in real time.

5 Automatic Evaluation

Automatic evaluation enables scalable testing and allows for consistent simulation of user behavior. We compare Co-STORM with the following baselines: (1) **RAG Chatbot**, a baseline that retrieves information from the search engine and interacts with the user through a one-question-one-answer paradigm. (2) **STORM + QA**, a baseline that uses

[4] Following Shao et al. (2024), when retrieving information with a search engine, we apply rule-based filtering according to the Wikipedia guideline https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources.

[5] https://documentation.you.com/api-reference/search

	Report Quality					Question-Answering Turn Quality		
	Relevance	Breadth	Depth	Novelty	Info Diversity	Consistency	Engagement	# Unique URLs
RAG Chatbot	3.57	3.50	3.26	2.44	0.595	4.37	4.13	2.94
STORM + QA	3.61	3.61	3.43	2.50	0.592	4.34	4.11	2.89
Co-STORM	**3.78**	**3.79**	**3.77**†	**3.05**†	**0.602**	**4.40**†	**4.33**†	**6.04**†
w/o Multi-Expert	3.73	3.75	**3.77**	2.93	0.589	**4.40**	4.32	5.91
w/o Moderator	3.56	3.69	3.41	2.89	0.577	4.39	4.28	5.67

Table 3: Automatic evaluation results for report quality and the quality of question-answering turns in the discourse with simulated users. Ablations are included as follows: "w/o Multi-Expert" denotes 1 expert and 1 moderator, and "w/o Moderator" denotes N experts and 0 moderator.† denotes significant differences ($p < 0.05$) from a paired t-test between Co-STORM and both baselines. The rubric grading uses a 1-5 scale. All scores reported are the mean values.

the STORM framework (Shao et al., 2024) to generate a report for a given topic to provide general information. It then allows the user to ask follow-up questions and provides corresponding answers retrieved with the search engine.

5.1 Evaluation Setup

We use the WildSeek dataset where each data point consists of an initial topic t and goal g. We simulate the user with an LM (`gpt-4o-2024-05-13`) prompted with t, g, the discourse history $\mathcal{D}$, and the instruction for question generation. To ensure a fair comparison, we terminate the information-seeking session once it reaches 30 search queries for Co-STORM and both baselines. For all methods, the final report is generated using the two-stage approach of outline generation followed by section-by-section article generation, as proposed by STORM (Shao et al., 2024), based on the interaction history. We evaluate the system quality by assessing the final report and the interaction history (*i.e.*, discourse) with the automatic metrics defined in §5.2.

5.2 Automatic Metrics

Report Quality We evaluate the final report on four aspects, *Relevance*, *Broad Coverage (Breadth)*, *Depth*, and *Novelty*, as indicators of the quality of the whole information-seeking process.[6] We employ Prometheus 2 (Kim et al., 2024), a 7B evaluator LM, to score the report based on a 5-point rubric. To further quantify the diversity of the collected information, we also report the *Information Diversity* as the average pairwise dissimilarity of $\mathcal{I}$,

$$1 - \frac{\sum_{i,j \in \mathcal{I}, i \neq j} \cos(\mathbf{i}, \mathbf{j})}{|\mathcal{I}|(|\mathcal{I}| - 1)},$$

where $\mathbf{i}, \mathbf{j}$ are corresponding text embeddings obtained from OpenAI's `text-embedding-3-small`.

Discourse Quality Since the discourse itself is valuable for human learning, we also evaluate the discourse trace using a 5-point rubric to grade each turn. This grading assesses *Novelty*, *Intent Alignment*, and *No Repetition* for question-asking utterances (*i.e.*, utterances with the intent ORIGINAL QUESTION or REQUEST INFORMATION). For question-answering utterances that provide information, we assess *Consistency* and *Engagement*. We also report the number of unique cited URLs in these utterances to indicate information diversity at the turn level. Both the rubrics for report evaluation and utterance evaluation are included in Appendix D.

5.3 Automatic Evaluation Results

Table 3 presents the evaluation results for report quality and the quality of question-answering turns in the discourse. The question-answering turns and the final report are the primary sources for human learning when they interact with Co-STORM. STORM + QA considers multiple perspectives in researching the given topic, indeed leading to improved performance across all four grading dimensions of the report quality compared to the RAG Chatbot. However, Co-STORM outperforms it, particularly in the *Depth* and *Novelty* aspects, by simulating collaborative discourse with multiple agentic roles, akin to a thought-provoking round table discussion. For discourse quality, the question-answering turns in Co-STORM significantly outperform both baselines in terms of *Consistency* and *Engagement*. This improvement is attributed to collaborative discourse setup, where the LM is prompted to generate the answer only when the retrieved information matches the current question according to the discourse history (see Listing 2). The utterance polishing step (see Figure 2) also

[6]The same four aspects are used in human evaluation.

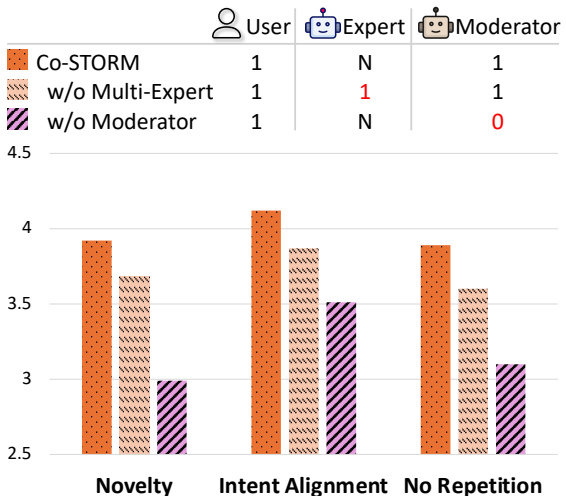

Figure 3: Rubric grading results for question-asking turn quality in automatic evaluation with simulated users.

helps as it serves as a self-improving mechanism.

5.4 Ablation Studies

As discussed in §3, a major innovation of Co-STORM is the orchestration of two types of LM agents. To assess the benefit, we compare Co-STORM with two ablations: (1) without multiple experts with different perspectives ("w/o Multi-Expert"), *i.e.*, only a single expert and a moderator, and (2) multiple experts but no moderator steering the discourse ("w/o Moderator"). As shown in Table 3, the ablated systems perform worse than the full system across all metrics in both report and question-answering turn quality. Notably, removing the moderator has a greater negative impact than reducing the number of experts.

A key feature of Co-STORM is that LM agents can ask questions on the user's behalf. As shown in Figure 3, the advantage of Co-STORM's multi-agent design becomes clearer when inspecting the question-asking turns. *Having just one expert and one moderator can already provide most of the benefits.* Importantly, the moderator role in Co-STORM raises questions based on unused information about the topic—such a role represents somebody with a much larger *known unknowns*, effectively steering the discourse to help users discover more in the space of their *unknown unknowns*.

Another key innovation of Co-STORM is the dynamic mind map. We include controlled experiments on mind map quality in Appendix B.

6 Human Evaluation

Human evaluation is essential for assessing systems designed for collaborative discourse, as it captures the complexities of human interaction, reflects familiar real-world interactions, and provides critical insights into the system's effectiveness. We compare Co-STORM with two baselines: (1) **RAG Chatbot** as detailed in section §5. (2) Traditional **Search Engine**.

6.1 Evaluation Setup

We conduct an IRB-approved human evaluation to compare Co-STORM with RAG Chatbot and Search Engine by recruiting 20 volunteers on the Internet. Participants are randomly split into two groups: one compared Co-STORM with Google Search, while the other group compared it with the RAG Chatbot. Participants are asked to seek information on two topics, one on each system, from the same domain and sharing the same goal. Note that we mitigate topic familiarity bias by using two different topics within the same domain. We have prepared five different domains, with each assigned to 2 users in each group. To counterbalance, one user starts with Co-STORM and switches to the baseline for the other topic, and vice versa.

After seeking information for each topic, participants are instructed to rate their experience based on four grading aspects defined in §5.2 (*Relevance*, *Breadth*, *Depth*, *Novelty/Serendipity*), using a 5-point Likert scale. After completing both tasks, participants are asked to provide pairwise preferences regarding the *required effort*, *user engagement*, *addressing echo chamber issues*, and *overall experience*. We also collect open-ended feedback and allow participants to optionally leave comments on each discourse turn and the mind map snapshots when interacting with Co-STORM. More details on the human evaluation are included in Appendix E.

6.2 Human Evaluation Results

Table 4 shows the human rating results and Figure 4 shows the pairwise comparison results. **Co-STORM helps users find broader and deeper information relevant to their goals.** Participants found that Co-STORM uncovers information with greater breadth and depth compared to the search engine and the RAG Chatbot. Specifically, Co-STORM is rated strictly higher in *Breadth* by 50% of the participants and strictly higher in *Depth* by 60% of the participants than the search engine. Compared to the RAG Chatbot, Co-STORM re-

[7]One participant in the Co-STORM v.s. RAG Chatbot group submitted the rating but did not leave a usage record, so we excluded this data point from the aggregated results.

	Co-STORM v.s. Search Engine				Co-STORM v.s. RAG Chatbot			
	Search Engine	Co-STORM	Win % (Lose %)	p-value	RAG Chatbot	Co-STORM	Win % (Lose %)	p-value
Relevance	3.90	**4.00**	30% (30%)	0.758	3.89	**4.22**	33% (0%)	0.081
Breadth	3.60	**4.10**	50% (10%)	0.096	3.11	**4.22**	67% (0%)	0.013
Depth	3.10	**4.00**	60% (10%)	0.081	3.11	**4.00**	56% (33%)	0.069
Serendipity	2.70	**3.90**	70% (10%)	0.030	2.78	**3.78**	67% (0%)	0.009

Table 4: Human ratings on different aspects of the information-seeking experience with Co-STORM and Search Engine (n=10) and with Co-STORM and RAG Chatbot (n=9)[7]. The ratings are given on a scale from 1 to 5 with 3 as "Average". We report the win rate of Co-STORM in pairwise comparison and the p-value in a paired t-test.

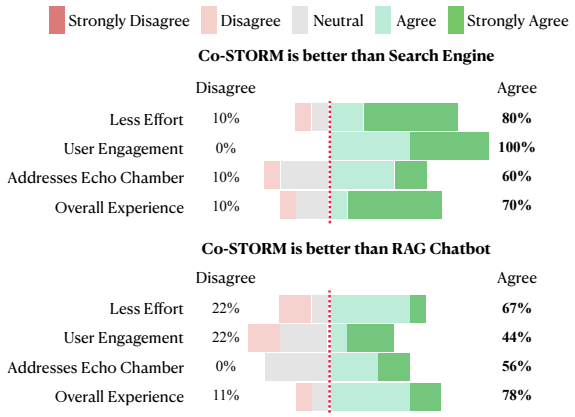

Figure 4: Survey results of the pairwise comparison (*i.e.*, agreement on whether Co-STORM is better than Search Engine/RAG Chatbot) in human evaluation.

ceives strictly higher scores in *Breadth* from 67% of participants and in *Depth* from 56% of participants. This finding aligns with the automatic evaluation results shown in Table 3. While helping users discover more information, Co-STORM remains aligned with their goals, as participants also rated Co-STORM higher in *Relevance* compared.

Co-STORM provides more serendipitous information with less mental effort required. Participants found that Co-STORM requires less effort, better mitigates the echo chamber issue, and provides a better overall experience. In a more fine-grained evaluation, participants evaluated 32% of Co-STORM 's total utterances, rating 89% of them as effectively "steering the discourse towards a new and interesting direction". One participant noted, "Co-STORM allows for almost full automation and much better understanding as it brings up topics that the user may not even think of". Moreover, participants found the mind map helpful. In total, they evaluated 80 snapshots of the dynamic mind map, finding it accurately tracked the discourse 71% of the time. One participant remarked, "Co-STORM is so much less mentally taxing for me to use".

Co-STORM should support more customization. Among the 19 participants, 4 noted that the RAG Chatbot better follows instructions that have a clear target and mentioned they expect Co-STORM to generate more concise utterances and provide less information in such cases. We view dynamically adapting Co-STORM to users' evolving mental states and personalizing their preferences as a meaningful direction for future work.

7 Related Works

Information-Seeking Support in NLP NLP research supporting human information-seeking has mainly focused on building question-answering (QA) systems (Chen et al., 2017; Lee et al., 2019; Dasigi et al., 2021; Levy et al., 2021; Yuan et al., 2020). These works often assume that the answer can be found within a single document (Clark et al., 2020) or that users can formulate complex queries (Yang et al., 2018; Chen et al., 2021; Ahmadvand et al., 2023), assumptions that do not hold true in complex information seeking (Butler, 2000; Booth et al., 2009; Byström and Järvelin, 1995).

Some more recent works have proposed long-form QA systems (Xu et al., 2023, 2024) and automatic expository writing systems (Balepur et al., 2023; Shen et al., 2023; Shao et al., 2024) to synthesize information from multiple sources. Some other studies have explored conversational search (Kumar and Callan, 2020; Nakamura et al., 2022). However, these works typically ignore human interaction or only passively answer user questions. We construct a multi-agent system with a human-in-the-loop protocol to support effective user interaction for complex and evolving information needs.

Multi-Agent Systems As LMs advance, a growing body of research explores their use in multi-agent applications (Wu et al., 2023; Nakajima, 2023; Liu et al., 2023b; Wang et al., 2024). Several studies show that multi-agent debate enhances factuality and reasoning compared to using a single LM (Du et al., 2023; Liang et al., 2023), and cooperative role-playing frameworks improve performance on coding or mathematical benchmarks (Li

et al., 2023; Hong et al., 2023). While these studies primarily focus on automating tasks, the potential applications extend further. For instance, Generative Agents (Park et al., 2023) instantiate an interactive environment with twenty-five LM agents to study emergent social behaviors, and Michael et al. (2023) show that multi-agent debates help humans supervise model outputs. Our work aligns with these broader applications by constructing a multi-agent system to facilitate human learning.

Collaborative Discourse for Human Learning Collaborative discourse has long been valued in classroom settings for its ability to deepen learners' understanding of concepts, enhance peer learning, and increase engagement (Nussbaum, 2008; Osborne, 2010; Kolodner, 2007; Chinn et al., 2000). Specifically, Nussbaum (2008) argues not all types of collaborative discourse are equally beneficial to students' learning, emphasizing the importance of critical discussion where participants assume different points of view. Furthermore, the facilitator role is important in collaborative discourse, with asking questions and providing complementary information as popular strategies (Onrubia et al., 2022).

8 Conclusion

We propose Co-STORM, an information-seeking assistance system that emulates collaborative discourse among users and multiple LM agents. By creating an interactive environment where users can both observe and participate, Co-STORM enhances learning and the complex information-seeking process. To facilitate automatic evaluation, we construct the WildSeek dataset, which captures the information-seeking needs and intents of real Internet users. Experimental results, including extensive human assessments, show that Co-STORM outperforms traditional search engines and RAG chatbots in surfacing *unknown unknowns* for human learning and reducing users' mental effort.

Limitations

We design Co-STORM to create an immersive human learning experience by enabling humans to participate in LM agent conversations. Despite the advantages demonstrated through both automatic and human evaluations, several limitations remain. First, the system could better tailor the collaborative discourse to the user's prior knowledge, skipping basic facts for knowledgeable users and introducing concepts progressively for novices. Second, while Co-STORM employs an effective discourse management mechanism, users sometimes desire more control over the discourse, including managing expert perspectives and customizing the utterance length. Third, extending Co-STORM to support multiple languages would significantly enhance its usefulness and impact. Although current LMs often possess multilingual capabilities, implementing a multilingual Co-STORM requires integrating search engines or retrieval models capable of accessing diverse language sources. Furthermore, managing content across different languages demands robust content moderation and the ability to identify conflicting information to ensure a reliable human learning experience. Finally, compared to the RAG Chatbot, Co-STORM has higher latency due to the need to decide the utterance intent and update the mind map. Although the current latency is acceptable for real-time interaction, as demonstrated in human evaluations, further improving the efficiency of the LM system would provide a smoother user experience.

Acknowledgements

We thank Omar Khattab, Eric Zelikman, Rose E. Wang, Yen-Jen Wang, and Qingyang Tang for their helpful feedback, and the ACL ARR reviewers for their valuable comments. We also appreciate the 20 participants in our user study who made this work possible. We are grateful to You.com for providing a discounted search API that supported our experiments. Yijia Shao is supported by a Stanford School of Engineering Fellowship. This work is supported in part by the Verdant Foundation and Microsoft Azure AI credits.

Ethics Statement

We build and evaluate our work to strictly adhere to ethical standards. The construction of the WildSeek dataset involves collecting data with users' explicit approval, and we carefully remove all personally identifiable information. In contrast to creative generation tasks, our tasks generate content that may impact how people perceive information and shape their opinions. We design our system to ground generated content on openly accessible external sources available on the general internet, with proper citations. Our experiments and evaluations ensure the accurate delivery of information and significantly reduce hallucinations. We avoid publishing or posting any generated content with-

out careful examination of information accuracy. We believe there are no data privacy issues as we ground our generated content from information accessible to the general public.

The primary risk of our work is the common bias issues originating from biases present on the general internet. Following Shao et al. (2024), we mitigate this problem by applying rule-based filtering according to the Wikipedia guideline[8] and incorporating multiple sources. Additionally, in our web application, we have implemented input topic moderation to reject topics that are sensitive, illegal, or potentially violate personal privacy. However, further information processing modules that serve as filters for internet sources and more robust modules to verify the accuracy of information can be implemented. Additionally, our current work only considers generating and retrieving information from English sources. Extending our system to be compatible with multilingual sources and generation will be beneficial.

References

Naresh Kumar Agarwal. 2015. Towards a definition of serendipity in information behaviour. *Information research: an international electronic journal*, 20(3):n3.

Ali Ahmadvand, Negar Arabzadeh, Julia Kiseleva, Patricio Figueroa Sanz, Xin Deng, Sujay Jauhar, Michael Gamon, Eugene Agichtein, Ned Friend, and Aniruddha. 2023. Making large language models interactive: A pioneer study on supporting complex information-seeking tasks with implicit constraints. *Preprint*, arXiv:2205.00584.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, and Jackson Kernion et al. 2022. Constitutional ai: Harmlessness from ai feedback. *Preprint*, arXiv:2212.08073.

Nishant Balepur, Jie Huang, and Kevin Chang. 2023. Expository text generation: Imitate, retrieve, paraphrase. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11896–11919, Singapore. Association for Computational Linguistics.

Marcia J. Bates. 1989. The design of browsing and berrypicking techniques for the online search interface. *Online Review*, 13(5):407–424.

Nicholas J Belkin, Robert N Oddy, and Helen M Brooks. 1982. Ask for information retrieval: Part i. background and theory. *Journal of documentation*, 38(2):61–71.

Wayne C Booth, Gregory G Colomb, and Joseph M Williams. 2009. *The craft of research*. University of Chicago press.

Ruth Butler. 2000. What learners want to know: The role of achievement goals in shaping information seeking, learning, and interest. In *Intrinsic and extrinsic motivation*, pages 161–194. Elsevier.

Tony Buzan. 1974. Noting. In *Using both sides of your brain*, chapter 4, pages 71–116. E. P. Dutton, New York.

Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information processing & management*, 31(2):191–213.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Yang Wang, and William W. Cohen. 2021. Open question answering over tables and text. In *International Conference on Learning Representations*.

Clark A Chinn, Angela M O'donnell, and Theresa S Jinks. 2000. The structure of discourse in collaborative learning. *The Journal of Experimental Education*, 69(1):77–97.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610, Online. Association for Computational Linguistics.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Allen Foster and Nigel Ford. 2003. Serendipity and information seeking: an empirical study. *Journal of documentation*, 59(3):321–340.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

[8] https://en.wikipedia.org/wiki/Wikipedia:Reliable_sources

Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. *Preprint*, arXiv:2308.00352.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. *Preprint*, arXiv:2405.01535.

Israel M Kirzner. 1997. Entrepreneurial discovery and the competitive market process: An austrian approach. *Journal of economic Literature*, 35(1):60–85.

Janet L Kolodner. 2007. The roles of scripts in promoting collaborative discourse in learning by design. In *Scripting Computer-Supported Collaborative Learning: Cognitive, Computational and Educational Perspectives*, pages 237–262. Springer.

Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science*, 42(5):361–371.

Vaibhav Kumar and Jamie Callan. 2020. Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3971–3980, Online. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Sharon Levy, Kevin Mo, Wenhan Xiong, and William Yang Wang. 2021. Open-Domain question-Answering for COVID-19 and other emergent domains. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 259–266, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

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

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.

Nelson Liu, Tianyi Zhang, and Percy Liang. 2023a. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7001–7025, Singapore. Association for Computational Linguistics.

Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023b. Dynamic llm-agent network: An llm-agent collaboration framework with agent team optimization. *Preprint*, arXiv:2310.02170.

Julian Michael, Salsabila Mahdi, David Rein, Jackson Petty, Julien Dirani, Vishakh Padmakumar, and Samuel R. Bowman. 2023. Debate helps supervise unreliable experts. *Preprint*, arXiv:2311.08702.

Yohei Nakajima. 2023. Babyagi. `https://github.com/yoheinakajima/babyagi`.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

E Michael Nussbaum. 2008. Collaborative discourse, argumentation, and learning: Preface and literature review. *Contemporary Educational Psychology*, 33(3):345–359.

Javier Onrubia, Begona Roca, and Marta Minguela. 2022. Assisting teacher collaborative discourse in professional development: An analysis of a facilitator's discourse strategies. *Teaching and Teacher Education*, 113:103667.

OpenAI. 2023. Gpt-4 technical report.

Jonathan Osborne. 2010. Arguing to learn in science: The role of collaborative, critical discourse. *science*, 328(5977):463–466.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA. Association for Computing Machinery.

Peter Pirolli. 2009. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42(3):33–40.

Chen Qu, Liu Yang, W. Bruce Croft, Yongfeng Zhang, Johanne R. Trippas, and Minghui Qiu. 2019. User intent prediction in information-seeking conversations. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19. ACM.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Stephen E. Robertson. 1977. Theories and models in information retrieval. *Journal of Documentation*, 33:126–148.

Jeremy Roschelle and Stephanie D Teasley. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer supported collaborative learning*, pages 69–97. Springer.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. Raptor: Recursive abstractive processing for tree-organized retrieval. In *International Conference on Learning Representations (ICLR)*.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *Preprint*, arXiv:2402.14207.

Nikhil Sharma, Q Vera Liao, and Ziang Xiao. 2024. Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–17.

Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond summarization: Designing ai support for real-world expository writing tasks. *Preprint*, arXiv:2304.02623.

David Traum. 2003. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*, pages 201–211. Springer.

Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the emergent cognitive synergy in large language models: A task-solving agent through multi-persona self-collaboration. *Preprint*, arXiv:2307.05300.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework.

Fangyuan Xu, Kyle Lo, Luca Soldaini, Bailey Kuehl, Eunsol Choi, and David Wadden. 2024. Kiwi: A dataset of knowledge-intensive writing instructions for answering research questions. *Preprint*, arXiv:2403.03866.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Xingdi Yuan, Jie Fu, Marc-Alexandre Côté, Yi Tay, Chris Pal, and Adam Trischler. 2020. Interactive machine comprehension with information seeking agents. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2325–2338, Online. Association for Computational Linguistics.

Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts. In *WSDM'23*, pages 429–437.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2024. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. *Preprint*, arXiv:2403.00165.

A Dataset Details

We constructed WildSeek using a web application[9] we built that hosts the open-sourced STORM project (Shao et al., 2024), as detailed in §2.2. User privacy was strictly maintained by explicitly obtaining consent each time users logged into our web application. No personally identifiable information was collected, and the entire dataset was manually reviewed to ensure compliance with this standard. We rejected topics that were illegal, harmful, violent, racist, sexual, non-English, based on personal experience, or contained personal information.

At the time of dataset construction, 8,777 users accessed our web application, resulting in the collection of 6,608 unique topic and purpose pairs. Participants were sourced from the general internet and all held valid Google accounts, in accordance with our IRB-approved policy. To ensure broad coverage, we conducted topic classification using `gpt-4o-2024-05-13` and human inspection, and then downsampled the collected data to 100 cases, covering 24 fine-grained categories in 6 domains: Science, Health and Fitness, Culture and Society, Lifestyle and Leisure, Social Science and Humanities, and Others. We applied rule-based filtering to exclude non-informative or trivial information-seeking purposes, and the 100 selected topic-purpose pairs were manually labeled by the authors. Table 5 includes example data points from each domain and Figure 5 shows the full taxonomy of the WildSeek dataset.

B Mind Map `Insert` Operation

As revealed in human evaluation results (see §6), the mind map is crucial for helping users track the discourse and the collected information. Co-STORM dynamically updates the mind map through `insert` and `reorganize` operations. In this section, we conduct controlled experiments on different implementations of `insert` and verify the quality of the mind map updates.

Dynamically organizing collected information into a mind map is challenging. Unlike classic document classification tasks (Zhang et al., 2024, 2023) and recursive summarization tasks (Sarthi et al., 2024; Gao et al., 2023), where either the hierarchical organization or the information to be organized is fixed, mind map insertion involves an evolving hierarchical organization of concepts and an incremental set of information. We compare the Co-STORM `insert` (§3.2) with two alternative approaches: (1) The **Embedding Only** baseline selects the placement with the highest semantic similarity using embedding cosine similarity. (2) The **Language Model Only** baseline directly prompts an LM to choose the best placement within the given hierarchical organization.

We construct an evaluation dataset for the controlled experiments by leveraging the FreshWiki dataset (Shao et al., 2024), which is a collection of recent, high-quality Wikipedia articles. We use the Wikipedia article outline as the concept hierarchy and require each candidate method to find the best placement for a given citation used in the article. The original placement of the citation in the article is deemed as the ground truth. We apply rule-based filtering to retain articles with up to three levels of hierarchy and English citation sources only. Inserting one cited source back into the outline is considered as one task. After downsampling, we derive a dataset consisting of 111 tasks: 33 from first-level sections, 64 from second-level sections, and 14 tasks from third-level sections.

We report the insertion accuracy in Table 6. For tasks where the ground truth placement is in the second or third level, we also consider a placement is partially correct if the information is inserted into one of the ancestors of the ground truth placement and report the partial accuracy. The experimental results show that solely relying on the LM performs poorly as the hierarchical organization can be wide and deep and the performance heavily depends on the quality of concept names. Co-STORM `insert` operation consistently outperforms both baseline approaches.

C Full Prompts in Co-STORM

In §3.1, we introduce Co-STORM 's collaborative discourse protocol which includes three key roles: the *user*, *experts*, and a *moderator*. We implement the perspective-guided expert and moderator pipeline using zero-shot prompting of `gpt-4o-2024-05-13`. Listing 2 and Listing 3 documents the full prompts for simulating the expert and the moderator respectively. Co-STORM uses a hierarchical mind map to track the discourse (§3.2) and the mind map `insert` operation is detailed in Appendix B. Prompts used for the mind map operations can be found in Listing 1.

[9] Our institution's IRB approved the web application

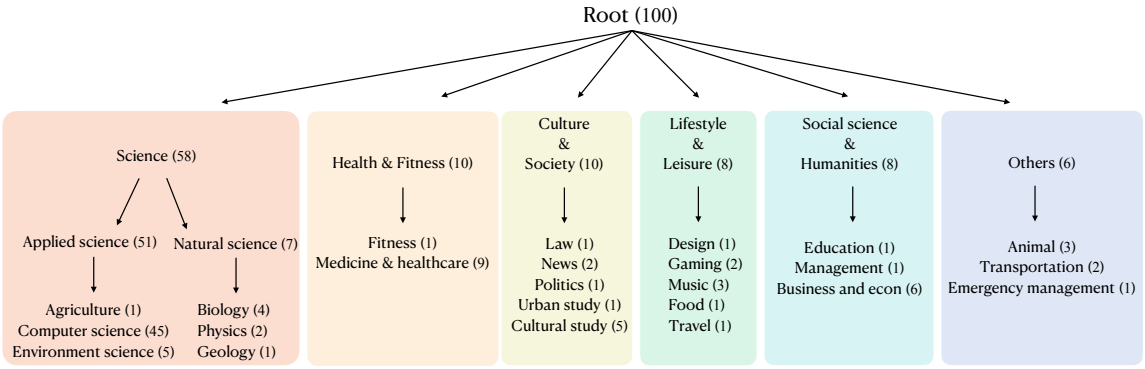

Figure 5: WildSeek taxonomy. The number in the parenthesis denotes the number of data points classified under the corresponding category or its descendants.

Domain	Example Task
Computer Science	**Topic:** Blockchain anomaly detection using large models **Intent:** To evaluate the effectiveness of large models in detecting anomalies in blockchain systems compared to existing models.
Healthcare	**Topic:** The effects of NMN supplements on human anti-aging **Intent:** To investigate the efficacy and mechanisms of NMN supplements in slowing down or reversing the aging process in humans.
Environmental Science	**Topic:** Utilization of Weather Forecasting for Wind and Solar Energy Assessment **Intent:** To explore advanced methodologies in integrating weather forecast data for optimizing wind and solar energy evaluations.
Law	**Topic:** Recent legal cases in the US involving hardware technology innovations **Intent:** To investigate the legal precedents and implications of hardware technology innovations in the US.
Economics	**Topic:** Development of a Shared Trading Currency to Facilitate International Trade **Intent:** Investigate how a new shared currency could eliminate transaction costs and boost GDP among member countries.

Table 5: Examples of complex information seeking tasks from the WildSeek dataset.

D Automatic Evaluation Details

Following Shao et al. (2024), we use the Prometheus model (Kim et al., 2024), an open-source rubric grading model for evaluating long-form text based on user-defined criteria. For our experiments, we use `prometheus-7b-v2.0` [10] with its default temperature as 1.0 and top_p as 0.9, the state-of-the-art version at the time of our experiments. As the model has a limited context window, for report evaluation, we omit references and trim the input text to under 2000 words to fit into the model's context window, following the practice in Shao et al. (2024); for discourse quality evaluation, we reduce the discourse history length by taking the last 2000 words as context. The report quality evaluation and discourse quality evaluation rubrics can be found in Table 9, Table 10, and Table 11.

To assess the quality of the automatic evaluation results in §5.3, we randomly sampled 50 data points from the automatic evaluation of discourse quality, with 10 data points for each rubric item, *i.e. Novelty*, *Intent Alignment*, *No Repetition* for question-asking utterances, and *Consistency* and *Engagement* for question-answering utterances, as defined in §5.2. Each data point represents the automatic grading of one utterance on one rubric item. Two independent evaluators provided human grading. We calculate the Pearson correlation between the automatic evaluation scores and the average human grading scores. Table 7 shows that the automatic rubric grading exhibits a positive correlation with human grading, with statistical significance observed for 4 out of the 5 rubric items. Additionally, the experimental results from the human evaluation with real users (Table Table 4 and Fig-

[10] https://huggingface.co/prometheus-eval/prometheus-7b-v2.0

	First-Level	Second-Level		Third-Level	
	Acc.	*Acc.*	Partial *Acc.*	*Acc.*	Partial *Acc.*
Embedding only	24.24	35.94	65.62	**35.71**	57.14
Language Model only	3.03	7.81	62.50	7.14	**71.43**
Co-STORM `insert`	**39.39**	**51.56**	**68.75**	**35.71**	**71.43**

Table 6: Controlled experiment results of different mind map insertion methods (%). A placement is deemed as partially correct if the information is inserted into one of the ancestors of the ground truth placement.

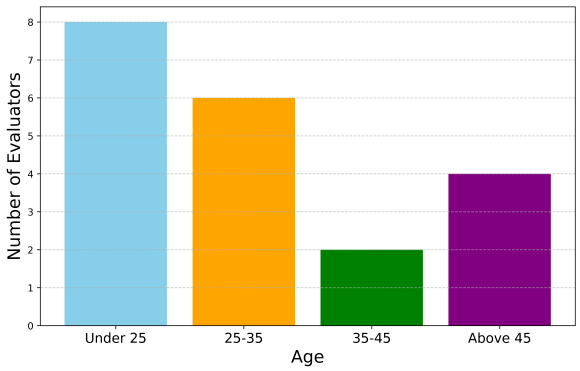

Figure 6: Age distribution of participants in the human evaluation.

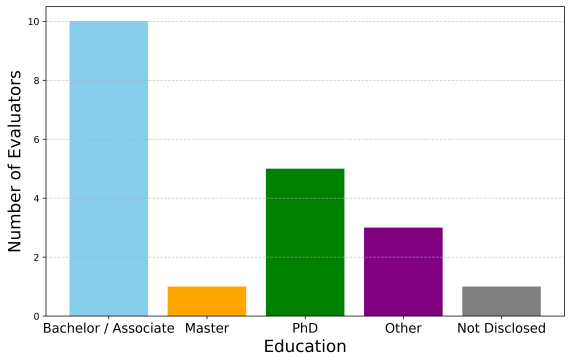

Figure 7: Education level distribution of participants in the human evaluation.

ure 4) also reveal similar findings to the automatic evaluation results, verifying our automatic evaluation setup.

E Human Evaluation Details

Human evaluation participants voluntarily provided demographic data, including their ages and highest education levels. As shown in Figure 6 and Figure 7), our human evaluation covers a diverse demographic. All participants gave consent to feedback data collection and we ensured no personal identifiable information was stored (see Figure 11). Feedback was collected via an online questionnaire

	Pearson Correlation (p-value)
Novelty	0.32 (< 4e-1)
Intent Alignment	0.55 (< 2e-2)
No Repetition	0.50 (< 7e-3)
Consistency	0.50 (< 2e-3)
Engagement	0.34 (< 2e-2)

Table 7: Pearson correlation between average human rubric grading scores and automatic rubric grading scores on discourse turn quality (n=50).

platform [11] and a web application we built.

The web application provides participants an interface to perform real-time interaction with Co-STORM. The web application has two modes, Co-STORM mode and RAG chatbot mode. Figure 12 shows a screenshot of the web application in Co-STORM mode. The RAG chatbot mode is similar to the common chatbot interface.

As discussed in §6, we crafted five pairs of complex information-seeking tasks for human evaluation (see Table 8). After completing each task, participants were instructed to rate the information-seeking assistance system they used (*i.e.*, Google Search, RAG Chatbot, or Co-STORM) from four grading aspects defined in §5.2 using 1 to 5 Likert scale (Likert question shown in Figure 8). After completing both two tasks, participants were asked to provide a pairwise preference by comparing Co-STORM with either Google Search (see Figure 9) or the RAG chatbot (see Figure 10) with the Likert questions.

F Case Study

We present two examples from different topics where the moderator effectively steers the conversation toward engaging directions. Example 13 shows an example of discourse on the topic "The effects of NMN supplements on human anti-aging" where the moderator effectively steers the ongoing discourse to the anti-aging benefits of personalized

[11] https://www.qualtrics.com

NMN and then further directs the discourse towards genetic profiling for personalized NMN supplementation plans. Example 14 highlights moderator effectively raises a new concept and shifts the discussion to the topic "The Emergence of Artificial Super Intelligence: Future Prospects and Impacts". The moderator steers the ongoing discourse from technology hurdles, the role of computation power, societal impact, risk, and mitigation toward a discussion on the quantum digital twin.

Additionally, we include a complete discourse transcript (Appendix §G) and the associated report (Appendix §H) on the topic of "AlphaFold 3," as referenced in Figure 2. In the discourse, the system initiates the discussion with steering by the moderator, focusing on the background and development of AlphaFold 3, as well as the technical advancements in biomolecular structure prediction, protein-DNA interactions, and its impact on genetic regulation. The user then directs the discourse towards its applications. Several participants provide insights into AlphaFold 3's applications in drug discovery, personalized medicine, and biotechnology. This is followed by a discussion on self-driving laboratories (SDLs), again steered by the moderator. Finally, the user shifts the discussion towards the economic impact and market implications of AlphaFold 3.

Topic	Goal
GPT-4 Omni AlphaFold 3	To investigate the latest technology breakthrough and discover a unique angle to report on it, ensuring more people know about the technology.
Gaza war protests in US colleges The conviction of Donald J. Trump in 2024	To investigate the latest news and provide comprehensive coverage, ensuring people receive diverse perspectives on the events.
Privacy Norm with Digital Technologies Copyright Issues with Language Models	To gain an in-depth understanding of the topic and prepare for a one-hour presentation in a college reading group.
Social Organism Social Statics and Social Dynamics	To conduct a literature review on a given topic in preparation for a class discussion in a sociology course.
China's dropping population in recent years The Humanitarian crisis in Gaza in recent years	To investigate the latest news and find an engaging angle to report it, incorporating background stories and connections to related events to enhance its appeal.

Table 8: Information-seeking tasks used in human evaluation.

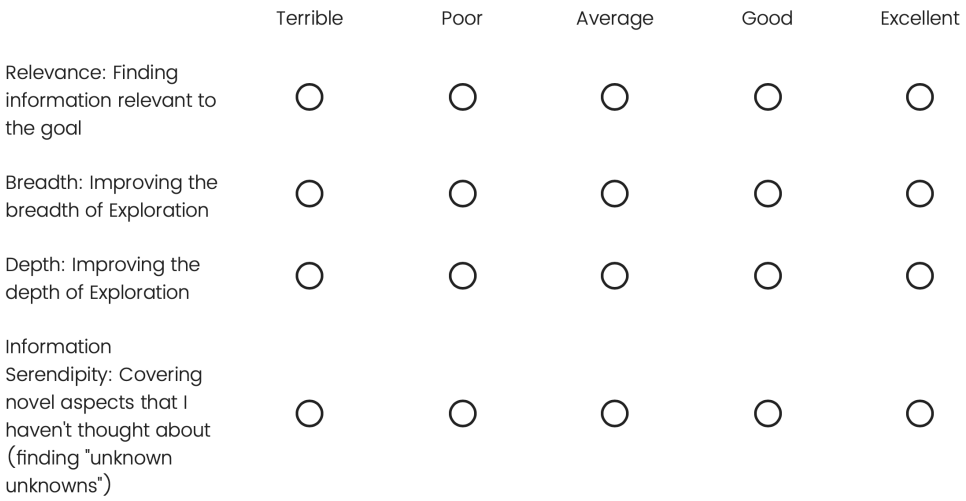

Figure 8: Human evaluation grading rubrics for each method (search engine, RAG Chatbot, and Co-STORM). Evaluation results are shown in Table 4

.

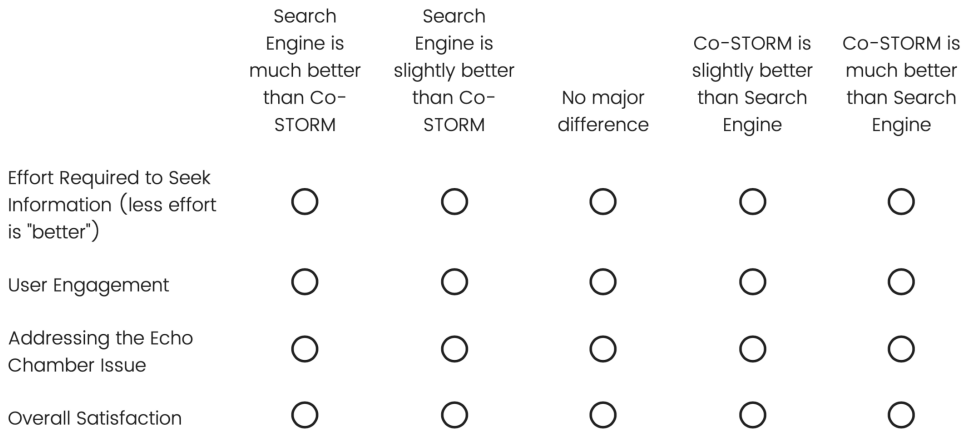

Figure 9: Likert question for comparing Co-STORM with Google Search in human evaluation. Evaluation results are shown in Figure 4.

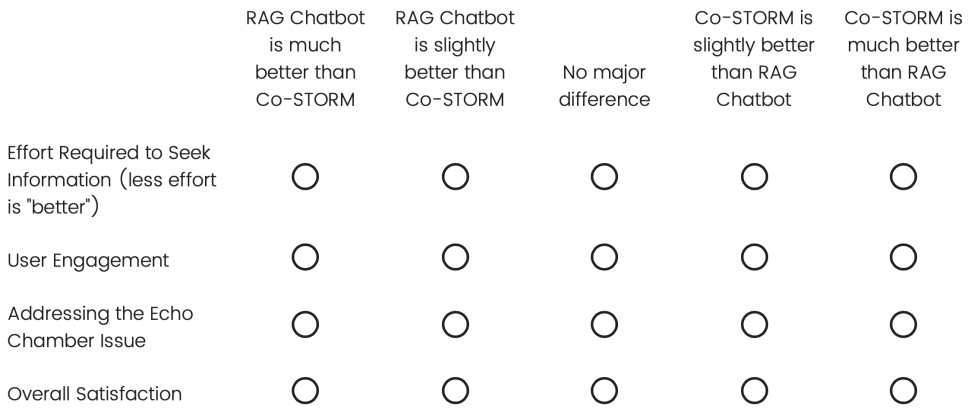

Figure 10: Likert question for comparing Co-STORM with RAG Chatbot in human evaluation. Evaluation results are shown in Figure 4.

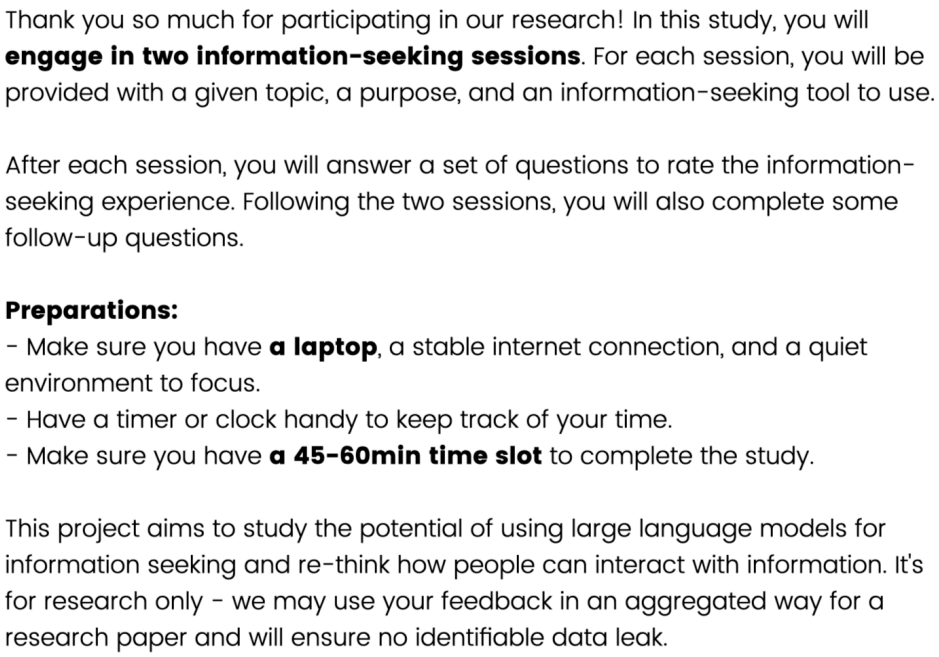

Figure 11: Screenshot to get consent from participants for gathering feedback data during human evaluations using Qualtrics.

```
class InsertInformation(dspy.Signature):
    """Your job is to insert the given information to the knowledge base. The knowledge base is a tree based data structure to
       organize the collection information. Each knowledge node contains information derived from themantically similar
      question or intent.
    To decide the best placement of the information, you will be navigated in this tree based data structure layer by layer.
    You will be presented with the question and query leads to ththeis information, and tree structure.

    Output should strictly follow one of options presetned below with no other information.
    - 'insert': to place the information under the current node.
    - 'step: [child node name]': to step into a specified child node.
    - 'create: [new child node name]': to create new child node and insert the info under it.

    Example outputs:
    - insert
    - step: node2
    - create: node3
    """
    intent = dspy.InputField(prefix="Question and query leads to this info: ", format=str)
    structure = dspy.InputField(prefix="Tree structure: \n", format=str)
    choice = dspy.OutputField(prefix="Choice:\n", format=str)

class InsertInformationCandidateChoice(dspy.Signature):
    """Your job is to insert the given information to the knowledge base. The knowledge base is a tree based data structure to
       organize the collection information. Each knowledge node contains information derived from themantically similar
      question or intent.
    You will be presented with the question and query leads to this information, and candidate choices of placement. In these
      choices, -> denotes parent-child relationship. Note that reasonable may not be in these choices.

    If there exists reasonable choice, output "Best placement: [choice index]"; otherwise, output "No reasonable choice".
    """
    intent = dspy.InputField(prefix="Question and query leads to this info: ", format=str)
    choices = dspy.InputField(prefix="Candidate placement:\n", format=str)
    decision = dspy.OutputField(prefix="Decision:\n", format=str)
```

Listing 1: Prompts used for dynamically updating the mind map in Co-STORM.

```
class QuestionToQuery(dspy.Signature):
    """You want to answer the question or support a claim using Google search.
        What do you type in the search box?
        The question is raised in a round table discussion on a topic. The question may or may not focus on the topic itself.
        Write the queries you will use in the following format:
        - query 1
        - query 2
        ...
        - query n"""

    topic = dspy.InputField(prefix='Topic context:', format=str)
    question = dspy.InputField(prefix='I want to collect information about: ', format=str)
    queries = dspy.OutputField(prefix="Queries: \n", format=str)


class AnswerQuestion(dspy.Signature):
    """ You are an expert who can use information effectively. You have gathered the related information and will now use the
      information to form a response.
    Make your response as informative as possible and make sure every sentence is supported by the gathered information.
    If [Gathered information] is not directly related to the [Topic] and [Question], start your response with "Based on the
      available information, I cannot fully address the question." Then, provide the most relevant answer you can based on the
       available information, and explain any limitations or gaps.
    Use [1], [2], ..., [n] in line (for example, "The capital of the United States is Washington, D.C.[1][3]."). 
    You DO NOT need to include a References or Sources section to list the sources at the end. The style of writing should be
      formal.
    """

    topic = dspy.InputField(prefix='Topic you are discussing about:', format=str)
    question = dspy.InputField(prefix='You want to provide insight on: ', format=str)
    info = dspy.InputField(
        prefix='Gathered information:\n', format=str)
    style = dspy.InputField(prefix="Style of your response should be:", format=str)
    answer = dspy.OutputField(
        prefix="Now give your response. (Try to use as many different sources as possible and do not hallucinate.)",
        format=str
    )

class ConvertUtteranceStyle(dspy.Signature):


    """
    You are an invited speaker in the round table conversation.
    Your task is to make the question or the response more conversational and engaging to facilitate the flow of conversation.
    Note that this is ongoing conversation so no need to have welcoming and concluding words. Previous speaker utterance is
      provided only for making the conversation more natural.
    Note that do not hallucinate and keep the citation index like [1] as it is. Also,
    """
    expert = dspy.InputField(prefix="You are inivited as: ", format=str)
    action = dspy.InputField(prefix="You want to contribute to conversation by: ", format=str)
    prev = dspy.InputField(prefix="Previous speaker said: ", format=str)
    content = dspy.InputField(prefix="Question or response you want to say: ", format=str)
    utterance = dspy.OutputField(prefix="Your utterance (keep the information as much as you can with citations, prefer
      shorter answers without loss of information): ", format=str)
```

Listing 2: Prompts used for simulating perspective-guided experts in Co-STORM.

```
class KnowledgeBaseSummmary(dspy.Signature):
    """Your job is to give brief summary of what's been discussed in a roundtable conversation. Contents are themantically
      organized into hierarchical sections.
    You will be presented with these sections where "#" denotes level of section.
    """
    topic = dspy.InputField(prefix="topic: ", format=str)
    structure = dspy.InputField(prefix="Tree structure: \n", format=str)
    output = dspy.OutputField(prefix="Now give brief summary:\n", format=str)

class GroundedQuestionGeneration(dspy.Signature):
    """Your job is to find next discussion focus in a roundtable conversation. You will be given previous conversation summary
       and some information that might assist you discover new discussion focus.
      Note that the new discussion focus should bring new angle and perspective to the discussion and avoid repetition. The
        new discussion focus should be grounded on the available information and push the boundaries of the current
        discussion for broader exploration.
      The new discussion focus should have natural flow from last utterance in the conversation.
      Use [1][2] in line to ground your question.
    """
    topic = dspy.InputField(prefix="topic: ", format=str)
    summary = dspy.InputField(prefix="Discussion history: \n", format=str)
    information = dspy.InputField(prefix="Available information: \n", format=str)
    last_utterance = dspy.InputField(prefix="Last utterance in the conversation: \n", format=str)
    output = dspy.OutputField(prefix="Now give next discussion focus in the format of one sentence question:\n", format=str)

class GenerateExpertWithFocus(dspy.Signature):
    """
    You need to select a group of speakers who will be suitable to have roundtable discussion on the [topic] of specific [
      focus].
    You may consider inviting speakers having opposite stands on the topic; speakers representing different interest parties;
      Ensure that the selected speakers are directly connected to the specific context and scenario provided.
    For example, if the discussion focus is about a recent event at a specific university, consider inviting students, faculty
       members, journalists covering the event, university officials, and local community members.
    Use the background information provided about the topic for inspiration. For each speaker, add a description of their
      interests and what they will focus on during the discussion.
    No need to include speakers name in the output.
    Strictly follow format below:
    1. [speaker 1 role]: [speaker 1 short description]
    2. [speaker 2 role]: [speaker 2 short description]
    """

    topic = dspy.InputField(prefix='Topic of interest:', format=str)
    background_info = dspy.InputField(prefix='Background information:\n', format=str)
    focus = dspy.InputField(prefix="Discussion focus: ", format=str)
    topN = dspy.InputField(prefix="Number of speakers needed: ", format=str)
    experts = dspy.OutputField(format=str)
```

Listing 3: Prompts used for simulating the moderator in Co-STORM

Criteria Description	**Broad Coverage**: Does the article provide an in-depth exploration of the topic and have good coverage?
Score 1 Description	Severely lacking; offers little to no coverage of the topic's primary aspects, resulting in a very narrow perspective.
Score 2 Description	Partial coverage; includes some of the topic's main aspects but misses others, resulting in an incomplete portrayal.
Score 3 Description	Acceptable breadth; covers most main aspects, though it may stray into minor unnecessary details or overlook some relevant points.
Score 4 Description	Good coverage; achieves broad coverage of the topic, hitting on all major points with minimal extraneous information.
Score 5 Description	Exemplary in breadth; delivers outstanding coverage, thoroughly detailing all crucial aspects of the topic without including irrelevant information.
Criteria Description	**Novelty**: Does the report cover novel aspects that relate to the user's initial intent but are not directly derived from it?
Score 1 Description	Lacks novelty; the report strictly follows the user's initial intent with no additional insights.
Score 2 Description	Minimal novelty; includes few new aspects but they are not significantly related to the initial intent.
Score 3 Description	Moderate novelty; introduces some new aspects that are somewhat related to the initial intent.
Score 4 Description	Good novelty; covers several new aspects that enhance the understanding of the initial intent.
Score 5 Description	Excellent novelty; introduces numerous new aspects that are highly relevant and significantly enrich the initial intent.
Criteria Description	**Relevance and Focus**: How effectively does the report maintain relevance and focus, given the dynamic nature of the discourse?
Score 1 Description	Very poor focus; discourse diverges significantly from the initial topic and intent with many irrelevant detours.
Score 2 Description	Poor focus; some relevant information, but many sections diverge from the initial topic.
Score 3 Description	Moderate focus; mostly stays on topic with occasional digressions that still provide useful information.
Score 4 Description	Good focus; maintains relevance and focus throughout the discourse with minor divergences that add value.
Score 5 Description	Excellent focus; consistently relevant and focused discourse, even when exploring divergent but highly pertinent aspects.
Criteria Description	**Depth of Exploration**: How thoroughly does the report explore the initial topic and its related areas, reflecting the dynamic discourse?
Score 1 Description	Very superficial; provides only a basic overview with significant gaps in exploration.
Score 2 Description	Superficial; offers some detail but leaves many important aspects unexplored.
Score 3 Description	Moderate depth; covers key aspects but may lack detailed exploration in some areas.
Score 4 Description	Good depth; explores most aspects in detail with minor gaps.
Score 5 Description	Excellent depth; thoroughly explores all relevant aspects with comprehensive detail, reflecting a deep and dynamic discourse.

Table 9: Report scoring rubrics on a 1-5 scale for the Prometheus model.

Criteria Description	Evaluates the extent to which the conversation turn introduces new and unexpected information that is relevant to the topic at hand. **Novelty**: High novelty indicates the conversation is providing fresh insights or perspectives that the user might not have considered, thereby enriching the dialogue and enhancing the user's understanding of the subject.
Score 1 Description	The turn fails to introduce any new or unexpected information, repeating known facts or irrelevant content.
Score 2 Description	The turn introduces some new information, but it is mostly predictable or only slightly relevant.
Score 3 Description	The turn provides moderately novel information that is relevant and somewhat unexpected.
Score 4 Description	The turn introduces new and relevant information that is largely unexpected, sparking interest.
Score 5 Description	The turn consistently introduces highly novel and relevant information that is completely unexpected, significantly enhancing the conversation.
Criteria Description	**Engaging**: Measures how interesting and captivating the conversation turn is. An engaging turn holds the user's attention and encourages them to continue interacting. It often includes elements that are thought-provoking, entertaining, or particularly relevant to the user's interests.
Score 1 Description	The turn is dull and uninteresting, likely causing the user to lose interest.
Score 2 Description	The turn has limited engagement, with occasional interesting points but generally fails to captivate the user.
Score 3 Description	The turn is moderately engaging, holding the user's interest but lacking captivating elements.
Score 4 Description	The turn is engaging and interesting, encouraging further interaction with minor lapses.
Score 5 Description	The turn is highly engaging, consistently holding the user's interest and encouraging further interaction.
Criteria Description	Assesses whether the conversation turn contradicts previous statements or established facts. Minimizing contradictionsis essential **Consistency**: for maintaining trust and coherence in the conversation. A high score indicates that the turn is free from inconsistencies and logically fits with the preceding dialogue.
Score 1 Description	The turn frequently contradicts previous statements or established facts, causing confusion.
Score 2 Description	The turn occasionally contradicts itself, with some inconsistencies present.
Score 3 Description	The turn is mostly free of contradictions, with only minor inconsistencies that do not significantly impact coherence.
Score 4 Description	The turn is nearly free of contradictions, with only very rare and minor inconsistencies.
Score 5 Description	The turn is entirely free of contradictions, maintaining perfect coherence and logical consistency.

Table 10: Question-answering turn scoring rubrics on a 1-5 scale for the Prometheus model.

Criteria Description	Assesses how well the conversation turn aligns with the user's latent intent or goals. It measures the relevance and appropriateness of the **Intent Alignment**: response in contributing towards the user's overall objectives. High intent alignment ensures that the conversation stays focused on the user's needs and drives towards meaningful outcomes.
Score 1 Description	The turn does not align with the user's latent intent or goals, and may confuse the conversation's purpose.
Score 2 Description	The turn slightly aligns with the user's latent intent, but does not significantly contribute to the overall goals.
Score 3 Description	The turn moderately aligns with the user's latent intent, contributing to the overall goals in a limited way.
Score 4 Description	The turn aligns well with the user's latent intent, contributing meaningfully to the overall goals.
Score 5 Description	The turn perfectly aligns with the user's latent intent, significantly driving the conversation towards the overall goals.
Criteria Description	Looks at the degree to which the conversation turn repeats information that has already been provided. Lower scores indicate higher repetition, **Repetition**: which can detract from the value of the conversation by failing to introduce new content. Ideally, each turn should add new information or perspectives to the dialogue.
Score 1 Description	The turn repeats information already provided without adding any new value.
Score 2 Description	The turn has noticeable repetition, with limited new information added.
Score 3 Description	The turn includes some repetition, but provides enough new information to be moderately valuable.
Score 4 Description	The turn has minimal repetition, mostly introducing new and relevant information.
Score 5 Description	The turn does not repeat any information, consistently providing new and valuable content.

Table 11: Question-asking turn scoring rubrics on a 1-5 scale for the Prometheus model.

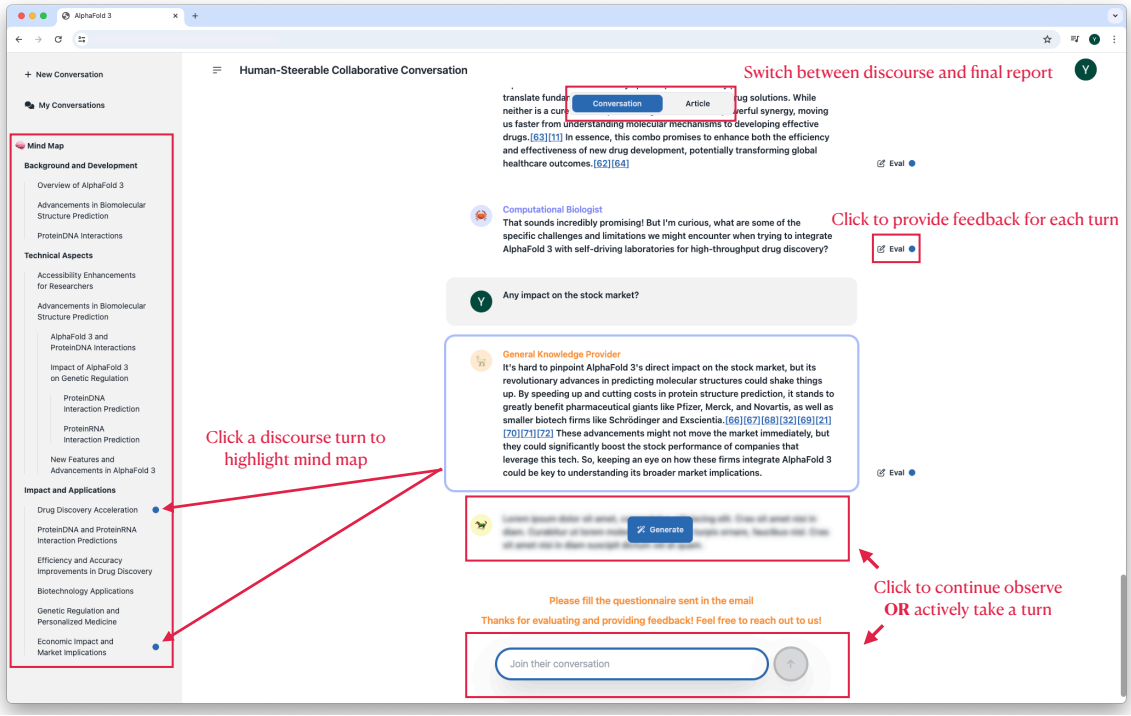

Figure 12: Screenshot of the interactive web application for human evaluation.

Topic: The effects of NMN supplements on human anti-aging

{Previous discourse focuses on the definition of NMN, Cellular Mechanisms of NMN, Side Effects and Genetic Factors, Genetic Variations and NMN Effectiveness, Methodologies of Recent Clinical Trials on NMN, and Safety and Side Effects.}

Moderator:
That's a fascinating angle! Considering the potential influence of genetic variations, how do you think personalized NMN supplementation-tailored to individual genetics, lifestyle, and health factors could optimize its anti-aging benefits? (inspired by [89][90])

{1 turn omitted}

Geneticist specializing in personalized medicine:
NMN supplementation by considering our unique genetic make-up and lifestyle. Since NMN boosts NAD+ levels, which drop with age and are crucial for improving insulin sensitivity, mitochondrial function, and lifespan [107][110] {More content omitted}

{1 turn omitted}

Moderator:
How can we create a thorough genetic profiling method to tailor NMN supplementation plans for individuals, ensuring we maximize those anti-aging benefits?

{rest of discourse omitted}

Figure 13: Example discourse highlight on the topic "The effects of NMN supplements on human anti-aging".

Topic: The Emergence of Artificial Super Intelligence: Future Prospects and Impacts

{Previous discourse focuses on Technological Hurdles, Role of Computational Power, Societal Impact, Risks and Mitigation.}

Moderator:
Building on that, how do you think the combination of digital twins with AI and quantum computing could speed up the development and practical use of artificial super intelligence (ASI) across different industries? (inspired by [77])

General Knowledge Provider:
Integrating digital twins with AI and quantum computing is a game-changer for various industries by offering incredibly precise simulations and optimizations{More content omitted}

In healthcare, QDTs can revolutionize our understanding and delivery of health services through advanced simulations of biological systems[81]{More content omitted}

QDTs allow for enhanced data management from IoT devices, leading to better asset performance and reduced failure risks [82][85].

{rest of discourse omitted}

Figure 14: Example discourse highlight on the topic "The Emergence of Artificial Super Intelligence: Future Prospects and Impacts".

G Full Collaborative Discourse Example

Topic: AlphaFold 3

Default Background Researcher:
Based on the available information, I cannot fully address the question in a conversational style. However, I can provide a detailed and accurate explanation in an informative manner.

AlphaFold 3 is the latest advancement in the field of molecular structure prediction, introduced in 2024 by Google DeepMind and Isomorphic Labs, both subsidiaries of Alphabet[2]. This groundbreaking AI model represents a substantial leap beyond its predecessor, AlphaFold 2, by expanding its capabilities to predict the structures and interactions of a broader spectrum of biomolecules[1][2].

Unlike AlphaFold 2, which primarily focused on single-chain proteins, AlphaFold 3 is designed to predict the structures of complex biomolecular systems. These systems include not only proteins but also DNA, RNA, post-translational modifications, and small molecules such as drugs and ligands[2][6][9]. The model operates by allowing scientists to input descriptions of these biomolecular complexes, which it then processes using a sophisticated neural network architecture. This architecture employs a custom Transformer with triangular attention and a diffusion process to generate the 3D coordinates of every atom in the specified system[6].

One of the most notable advancements in AlphaFold 3 is its potential application in drug design. By accurately predicting the structure and interactions of protein-ligand complexes, the model can aid in identifying and designing new molecules that could serve as effective drugs[3][4][10]. This has significant implications for the pharmaceutical industry, where Isomorphic Labs is already collaborating with companies to tackle real-world drug design challenges[3][4].

Researchers and scientists can access the majority of AlphaFold 3's capabilities for free through the newly launched AlphaFold Server, which is an easy-to-use research tool designed to make these advanced predictions accessible to a broader scientific community[3][4][7]. The introduction of AlphaFold 3 is expected to dramatically accelerate biological research by providing detailed atomic-level views of biomolecular interactions, thus enhancing our understanding of the biological world and facilitating the development of new treatments[9][11].

Overall, AlphaFold 3's ability to predict the structures of a wide variety of biomolecules more broadly and accurately than previous models makes it a revolutionary tool in the fields of structural biology and drug discovery[5][8].

Bioinformatics Specialist:
Based on the available information, I cannot fully address the question. However, I can provide some insights into how AlphaFold 3 has improved accessibility and usability for researchers aiming to integrate its predictions into their biological studies.

AlphaFold 3 has introduced several significant enhancements over its predecessors, specifically designed to improve both accessibility and usability for researchers. One of the most notable advancements is the launch of the AlphaFold Server, which is a free, user-friendly platform that allows biologists to generate molecular complexes with just a few clicks[13][14]. This means that researchers, regardless of their computational resources or expertise in machine learning, can now easily access high-accuracy protein structure predictions.

Additionally, AlphaFold 3 extends its predictive capabilities beyond just protein structures to include a variety of biomolecular systems, such as nucleic acids, small molecules, ions, and modified residues[12][17]. This broader scope of prediction is particularly beneficial for researchers involved in complex biological studies and drug discovery, as it allows them to model interactions across different types of molecules more accurately and comprehensively.

The new diffusion-based architecture of AlphaFold 3 enhances its prediction accuracy and reliability[12][16]. For example, the model shows a significant improvement in predicting protein-protein complexes and antibody-protein interfaces[18]. Such improvements are crucial for researchers focusing on immunology and therapeutic antibody development.

Despite these advancements, it's important to note that the downloadable code of AlphaFold 3 has not been released by DeepMind, unlike its previous models[14][15]. This may limit the flexibility for researchers who wish to modify or integrate the model deeply into their own systems. Nevertheless, the open accessibility of the AlphaFold Server provides a substantial step towards democratizing access to cutting-edge protein structure prediction technology.

In summary, AlphaFold 3 enhances accessibility through its user-friendly AlphaFold Server, allowing researchers to easily generate and use high-accuracy predictions. Its improved predictive capabilities across a range of biomolecular systems further support its integration into various biological studies. However, the absence of a downloadable code version may be a limitation for some users.

Pharmaceutical Industry Expert:
Based on the available information, I cannot fully address the question. However, I can provide some relevant insights into the impact of AlphaFold 3 on drug discovery projects.

AlphaFold 3, developed by Google DeepMind and Isomorphic Labs, represents a significant advancement in the field of protein structure prediction and has transformative potential for drug discovery[21]. Compared to previous computational methods, AlphaFold 3 has improved the efficiency and accuracy of these projects in several ways.

Firstly, AlphaFold 3 has enhanced the prediction accuracy for protein structures, including complex interactions like antibody-antigen bindings, which are crucial for immunology and therapeutic antibody development[24]. This increased accuracy allows researchers to gain a better understanding of molecular interactions, which is essential for designing effective drugs.

Moreover, AlphaFold 3's ability to predict structures for nearly all molecules in the Protein Data Bank significantly broadens the scope of potential drug targets[20]. This capability is particularly valuable for tackling neglected diseases, which

often suffer from a lack of research funding[25].

A notable example of AlphaFold 3's impact can be seen in a project conducted by Insilico and the University of Toronto. They applied AlphaFold to an end-to-end AI-powered drug discovery platform and successfully identified a new drug for a novel target to treat hepatocellular carcinoma (HCC)[23]. This project demonstrates how AlphaFold 3 can streamline the drug discovery process by integrating structure prediction with other AI models to identify effective therapeutic compounds efficiently.

Another example involves the discovery of a potent hit molecule, ISM042-2-048, using AlphaFold in combination with other AI tools. This molecule showed significant inhibitory activity and selective anti-proliferation effects, showcasing the practical application of AlphaFold in the hit identification phase of drug discovery[22].

In summary, AlphaFold 3 has made significant strides in improving the efficiency and accuracy of drug discovery projects. By providing highly accurate protein structure predictions and enabling the identification of new drug targets, it has the potential to revolutionize the development of new therapies. However, while the implications are promising, the full transformative impact on drug discovery may still be incremental at this stage[19].

Structural Biologist:
Based on the available information, I cannot fully address the question. However, I can provide the most relevant answer based on the gathered information, and explain any limitations or gaps.

AlphaFold 3 introduces several specific advancements in predicting complex biomolecular structures, which significantly enhance our understanding of interactions involving proteins, DNA, RNA, and ligands at the atomic level. One of the key innovations in AlphaFold 3 is the updated diffusion-based architecture. This architecture enables joint structure prediction for complexes that include not only proteins but also nucleic acids (DNA and RNA), small molecules, ions, and modified residues[8]. This comprehensive capability allows for more accurate and detailed modeling of biomolecular interactions, surpassing many previous specialized tools[8][16].

Furthermore, AlphaFold 3 achieves unprecedented accuracy in predicting protein-ligand interactions, which are crucial for drug discovery. It is reported to be 50\% more accurate than the best traditional methods on the PoseBusters benchmark, and it does so without the need for any structural information input[26]. This makes AlphaFold 3 the first AI system to outperform physics-based tools in this domain[26].

In addition to protein-ligand interactions, AlphaFold 3 is capable of modeling the interactions between proteins and nucleic acids such as DNA and RNA[27]. This capability is particularly important for understanding the molecular mechanisms underlying genetic regulation and expression, as well as the role of RNA in various cellular processes.

AlphaFold 3's advancements also extend to the modeling of protein interactions with antibodies, which is another critical area for drug development[28]. The ability to predict these interactions with high accuracy can aid in the design of new therapeutic molecules and improve the efficiency of drug discovery pipelines[28].

Overall, AlphaFold 3 represents a significant leap forward in biomolecular structure prediction. Its ability to model a wide range of molecular interactions with high accuracy enhances our understanding of the atomic-level details of these interactions, which is essential for both basic biological research and practical applications in drug discovery and development[9][11].

While the gathered information highlights many advancements of AlphaFold 3, it does not provide exhaustive details on all potential applications and specific improvements over its predecessors. Nonetheless, the advancements mentioned above underline the transformative impact of AlphaFold 3 in the field of structural biology.

moderator:
That's fascinating! Speaking of advancements, how do you think AlphaFold 3's improved accuracy in predicting protein-DNA and protein-RNA interactions might deepen our insights into genetic regulation and expression?

General Knowledge Provider:
Absolutely, AlphaFold 3 (AF3) is a game-changer for understanding genetic regulation and expression. Its diffusion-based architecture for predicting raw atom coordinates marks a significant leap, allowing it to model protein-DNA and protein-RNA interactions with much higher accuracy[29][8]. Unlike its predecessor, AlphaFold 2, which focused on individual protein structures, AF3 can now tackle these complex interactions, making it possible to gain deeper insights into how genetic processes are regulated.

What's really impressive is the 50\% improvement in prediction accuracy over existing methods, with some categories even doubling in accuracy[30][33]. This means we can now more reliably model these interactions, shedding light on the mechanisms of gene regulation and expression at a molecular level.

Additionally, AF3's ability to predict interactions across all of life's molecules, including nucleic acids, supports a wide range of scientific applications-from drug design to genomics research[31][32]. For instance, understanding protein-RNA interactions is crucial for RNA-based therapeutics, opening new avenues in medical research.

So, in a nutshell, AF3 is significantly enhancing our understanding of genetic regulation and expression by providing highly accurate models of these essential interactions.

Biochemist/Structural Biologist:
Absolutely, the strides AlphaFold 3 (AF3) has made in predicting protein-DNA and protein-RNA interactions are nothing short of revolutionary for genetic research and personalized medicine. With its diffusion-based architecture, AF3 goes beyond what AlphaFold 2 could achieve by predicting raw atom coordinates, allowing it to model a wide range of biomolecular interactions with high accuracy, including those involving nucleic acids, small molecules, ions, and modified residues [29][8].

One of the standout features of AF3 is its accuracy; it's shown to improve prediction accuracy by at least 50\%, and in some cases, even double it [30]. This is a game-changer for understanding gene regulation and expression, providing atomic-level insights into how proteins interact with DNA and RNA [11][5]. Imagine being able to visualize these interactions in such detail-this could lead to breakthroughs in developmental biology, oncology, and neurogenetics [31].

Moreover, AF3's capabilities have profound implications for personalized medicine. By accurately modeling protein-RNA interactions, it helps us understand how genetic variations affect individual responses to treatments, paving the way for more personalized and effective therapies [36][37]. It's not just a tool for basic research but a potential catalyst for transforming medical practices and treatment strategies.

Geneticist:
Building on AlphaFold 3's (AF3) advancements, its predictive capabilities can really push the boundaries of gene regulation research and genetic therapies. AF3 doesn't just improve on AlphaFold 2's accuracy; it broadens the scope to predict complexes involving proteins, nucleic acids, small molecules, and ions[8][41]. This expanded understanding of protein interactions in their full biological context, such as with DNA, RNA, and ligands, is crucial for effective gene regulation and therapy [38][42].

AF3's enhanced ability to model protein-ligand interactions with high precision can significantly aid in designing targeted genetic therapies and drugs[8]. For instance, its predecessor has already contributed to drug discovery, including cancer immunotherapy and COVID-19 vaccine development[43][40]. With AF3, we can expect even more precise modeling of these interactions, leading to the identification of novel therapeutic targets.

Moreover, DeepMind's commitment to open science, sharing over 200 million protein structures, accelerates collaborative efforts in genetic therapies by providing valuable data to researchers globally[39]. This open-access resource is a goldmine for identifying new protein targets and understanding their roles in genetic regulation, paving the way for breakthroughs in gene therapy and personalized medicine.

In essence, AF3's advanced capabilities offer unprecedented opportunities in both scientific research and therapeutic development, enhancing our ability to identify novel protein targets for genetic therapies[38][8][41][42].

User:
What are AlphaFold3 impacts and what are its applications?

General Knowledge Provider:
Sure, let's dive into what makes AlphaFold 3 such a game-changer and where it's making waves.

AlphaFold 3, developed by DeepMind, has taken our ability to predict protein structures to a whole new level, hitting almost 99\% accuracy for human proteins[44][45]. This isn't just a cool tech feat; it's a leap in how we understand biology, so much so that some think it could snag a Nobel Prize down the line[45].

For researchers, AlphaFold 3 is like a supercharged toolkit. It complements traditional methods like X-ray crystallography, cryo-EM, and NMR spectroscopy, speeding up the pace of discovery and offering new insights into the molecular basis of diseases[49].

One of the biggest wins is in drug discovery. By predicting protein interactions with other molecules, AlphaFold 3 can help identify new drug candidates much faster, tackling one of the biggest hurdles in the field[13][1].

Then there's the AlphaFold Database, now with over 200 million protein structure predictions. This open-access treasure trove is leveling the playing field, allowing researchers everywhere to tap into top-notch structural data[13][48]. It's fueling advancements in enzyme design, cancer research, and even vaccine development[46][47].

Beyond proteins, AlphaFold 3 is branching out to predict interactions involving DNA, RNA, and small molecules, opening doors in genomics, biotechnology, and synthetic biology[8][1].

Of course, it's not a silver bullet. Turning these insights into real-world applications will need lots of teamwork between scientists and industry experts[45].

All in all, AlphaFold 3 is revolutionizing structural biology and paving the way for breakthroughs in drug discovery, biotechnology, and beyond. Its impact is vast, touching everything from basic research to practical medical and industrial innovations.

Biochemist:
Absolutely, AlphaFold 3 (AF3) represents a huge leap in protein structure prediction, especially with its diffusion-based architecture. Unlike AlphaFold 2's Evoformer model, AF3 can accurately model interactions between proteins and various biomolecules like DNA, RNA, and small molecules with remarkable precision[52][29]. This is groundbreaking for genetic regulation research, as it allows us to predict how proteins interact with DNA and uncover the mechanisms behind gene activation and repression[50][8][53].

This enhanced capability is crucial for developing targeted gene therapies. With AF3, we can pinpoint key regulatory elements and design interventions to modify gene expression precisely. This has the potential to lead to innovative treatments for genetic disorders, where adjusting gene function is essential[9][31][34].

However, it's worth mentioning that access to AlphaFold 3 is currently limited by DeepMind, which could slow down its widespread application[51]. Despite this hurdle, the advancements made by AF3 offer a bright future for genetic research and therapeutic development.

In short, AlphaFold 3's ability to predict protein-DNA interactions with high accuracy could revolutionize our understanding of genetic regulation and pave the way for targeted gene therapies, even if access constraints may delay its broader use.

Pharmaceutical Industry Representative:
Absolutely, the potential of AlphaFold 3 in drug discovery is immense. Given its advanced ability to predict 3D structures of complex biomolecular interactions-including proteins, DNA, RNA, and small molecules-it offers a significant advantage in understanding how drugs bind to their targets [55][7].

One of the standout features of AlphaFold 3 is its capability to predict the effects of post-translational modifications and ions on molecular systems [55]. This can be a game-changer for identifying new drug targets that were previously considered too challenging to tackle [55][56].

By improving the accuracy of drug metabolism and toxicity predictions, AlphaFold 3 can substantially cut down the number of necessary wet lab experiments, saving both time and money [58]. Moreover, its near-complete prediction of human protein structures opens up unprecedented opportunities for fundamental research and therapeutic development [44].

While these advancements are promising, some experts still see the current impact on drug discovery as incremental, requiring further validation [54]. Nonetheless, the potential for AlphaFold 3 to streamline drug discovery, especially for complex diseases, is incredibly exciting [21][20].

moderator:
That's a great point. How do you think integrating AlphaFold 3 with emerging technologies like self-driving laboratories could further accelerate and innovate the drug discovery process? (inspired by [60][61])

General Knowledge Provider:
Absolutely, integrating AlphaFold 3 with self-driving laboratories (SDLs) can truly revolutionize drug discovery. AlphaFold 3's prowess in predicting 3D structures of complex biomolecular systems offers deep insights into molecular interactions crucial for drug design[55][9]. This means we can now target proteins that were previously considered untouchable with remarkable precision[55][44].

On the other hand, SDLs bring in the power of machine learning, lab automation, and robotics to conduct high-throughput experiments autonomously[64][65]. By merging AlphaFold 3's predictive capabilities with SDLs, we can streamline the process of identifying and optimizing drug candidates. SDLs can swiftly test numerous hypotheses and chemical compounds, guided by the insights from AlphaFold 3, making the identification of promising therapeutic candidates more efficient than ever[64].

Together, AlphaFold 3 and SDLs not only speed up the discovery process but also translate fundamental scientific insights into practical drug solutions. While neither is a cure-all alone, their integration creates a powerful synergy, moving us faster from understanding molecular mechanisms to developing effective drugs[63][11].

In essence, this combo promises to enhance both the efficiency and effectiveness of new drug development, potentially transforming global healthcare outcomes[62][64].

Computational Biologist:
That sounds incredibly promising! But I'm curious, what are some of the specific challenges and limitations we might encounter when trying to integrate AlphaFold 3 with self-driving laboratories for high-throughput drug discovery?

User:
Any impact on the stock market?

General Knowledge Provider:
It's hard to pinpoint AlphaFold 3's direct impact on the stock market, but its revolutionary advances in predicting molecular structures could shake things up. By speeding up and cutting costs in protein structure prediction, it stands to greatly benefit pharmaceutical giants like Pfizer, Merck, and Novartis, as well as smaller biotech firms like Schrodinger and Exscientia[66][67][68][32][69][21][70][71][72]. These advancements might not move the market immediately, but they could significantly boost the stock performance of companies that leverage this tech. So, keeping an eye on how these firms integrate AlphaFold 3 could be key to understanding its broader market implications.

Financial Analyst:
Absolutely, let's dive into how AlphaFold 3 is being integrated into the workflows of both pharmaceutical giants and smaller biotech firms, and the financial implications of this integration.

AlphaFold 3, developed by Google DeepMind, goes beyond predicting individual protein structures to modeling interactions between multiple proteins, DNA, RNA, and small molecules[73][74]. This makes it a game-changer in drug discovery and development.

Big pharma companies, especially those partnering with Isomorphic Labs, are using AlphaFold 3 to tackle complex drug design challenges[80][81]. By integrating this tech, they can speed up the drug discovery process, cut costs, and improve prediction accuracy, which enhances their development pipelines[78][79].

Smaller biotech firms are also leveraging AlphaFold 3 through platforms like Benchling, which incorporates AlphaFold beta for predicting 3D structures of novel proteins[75][77]. This helps them innovate more efficiently without needing extensive computational resources or deep machine learning expertise[78].

However, scaling AlphaFold 3 poses challenges like building the necessary IT infrastructure and ensuring data integrity and interoperability[76].

In essence, both large and small firms are using AlphaFold 3 to boost efficiency and innovation in drug discovery, with potentially significant long-term financial benefits.

H Final Report Example

AlphaFold 3

 # Background and Development

Overview of AlphaFold 3

AlphaFold 3, co-developed by Google DeepMind and Isomorphic Labs, was announced on May 8, 2024, and represents a significant advancement in the prediction of biomolecular structures and interactions[2][7]. Unlike its predecessor, AlphaFold 2, AlphaFold 3 extends its capabilities beyond single-chain proteins to a wide array of biomolecules including DNA, RNA, post-translational modifications, and small molecules such as ligands, which are critical in drug discovery[1][2].

The AI model employs a neural network architecture that utilizes a custom Transformer with triangular attention and a diffusion process to generate 3D coordinates of every atom within the specified biomolecular system[6]. This allows researchers to input descriptions of complex biomolecular systems and receive highly accurate predictions of their three-dimensional structures[6].

One of the standout features of AlphaFold 3 is its ability to predict the structures of protein complexes with enhanced accuracy, particularly in protein-DNA and protein-RNA interactions. This improvement is expected to facilitate groundbreaking discoveries in gene regulation and expression, potentially revolutionizing genetic research and personalized medicine[5].

To facilitate broader scientific research, the capabilities of AlphaFold 3 are accessible for free through the AlphaFold Server, an easy-to-use research tool[3][4]. This accessibility is aimed at accelerating biological research and drug discovery processes by enabling more precise identification of drug targets and reducing the time and costs associated with developing new medications, especially for complex diseases[7][10]. Isomorphic Labs is also collaborating with pharmaceutical companies to leverage AlphaFold 3's potential in real-world drug design challenges, with the ultimate goal of developing new life-changing treatments for patients[3][4].

Advancements in Biomolecular Structure Prediction

AlphaFold 3 represents a significant leap in the field of biomolecular structure prediction, building upon the successes of its predecessor, AlphaFold 2. The new model boasts a substantially updated diffusion-based architecture, which enables joint structure prediction of not only proteins but also nucleic acids, small molecules, ions, and modified residues[8][38]. This expansion beyond proteins marks a pivotal advancement, allowing for a more comprehensive understanding of complex biomolecular interactions and enhancing the accuracy of these predictions.

One of the most notable achievements of AlphaFold 3 is its improved accuracy, which far surpasses many specialized tools previously used for protein-ligand interactions[8]. This heightened precision has broad implications, particularly in identifying novel protein targets for genetic therapies, thereby offering unprecedented opportunities in gene regulation research[38][39]. Furthermore, the model's capabilities extend to predicting the structures of complexes involving DNA and RNA, which could significantly advance our understanding of genetic regulation and aid in the development of targeted gene therapies[8][42].

DeepMind's commitment to open science has also played a crucial role in the impact of AlphaFold 3. By freely publishing the predicted structures of over 200 million proteins, DeepMind fosters collaboration and knowledge sharing, accelerating scientific discoveries and pharmaceutical development worldwide[39]. This initiative ensures that the benefits of AlphaFold 3 are widely accessible, contributing to collective scientific progress.

The applications of AlphaFold 3 extend beyond academic research, impacting practical fields such as vaccine development. For instance, the structure-guided design of COVID-19 vaccines by companies like Pfizer, Moderna, and Johnson & Johnson benefited from advancements in protein structure prediction, highlighting the model's potential in addressing global health challenges[40].

Despite its significant advancements, some researchers have noted limitations in AlphaFold 3's accuracy for a subset of its predictions, and the model does not fully reveal the underlying mechanisms of protein folding[42]. Nevertheless, the broader understanding of biomolecular contexts provided by AlphaFold 3, including the interactions of drug targets with protein binding partners, DNA, RNA, and ligand cofactors, is expected to lead to more effective therapeutic interventions[43]. This richer contextual insight underscores the potential for rational, structure-based drug design, as demonstrated in the examination of TIM-3, a potential target for cancer immunotherapy[43].

ProteinDNA Interactions

AlphaFold 3 represents a significant advancement in the field of computational biology, specifically in the prediction of the structure of biomolecular systems. Building upon the foundational work of AlphaFold 2, which accurately predicted the structure of individual proteins, AlphaFold 3 extends these capabilities to include complex interactions involving multiple proteins, DNA, RNA, and small molecule ligands[9][11]. This includes an accurate atomic-level view of how these biomolecules come together and interact, providing critical insights into the structural impact of post-translational modifications and ions[11].

By providing detailed predictions of protein-DNA interactions, AlphaFold 3 enhances our understanding of gene regulation and the molecular basis of various diseases. This capability is crucial for advancing drug discovery, as it allows researchers to identify potential therapeutic targets more accurately and to understand how drugs can modify these interactions to produce desired effects[9][11]. The integration of AlphaFold 3 with emerging technologies, such as self-driving laboratories, promises to further accelerate and innovate the drug discovery process by automating the synthesis, testing, and optimization of new drug candidates based on precise structural data[9].

 # Technical Aspects

Accessibility Enhancements for Researchers

AlphaFold 3 significantly enhances accessibility and usability for researchers looking to integrate its predictions into their biological studies. One of the major strides in accessibility is the launch of the AlphaFold Server, a free and user-friendly research tool powered by AlphaFold 3. This server is touted as the most accurate tool globally for predicting protein interactions with other molecules within the cell. Researchers, irrespective of their computational resource availability or machine learning expertise, can generate molecular complexes with just a few clicks on a single platform[13].

Google DeepMind has made AlphaFold Server openly accessible to facilitate the global research community's use of AlphaFold 3, driving advancements in fields such as drug discovery, biotechnology, genomics, and our foundational understanding of biological systems. However, it is noteworthy that, unlike its predecessors, DeepMind has not released the downloadable code for AlphaFold 3[14]. Despite this, the open access to the AlphaFold Database, which houses over 200 million protein structure predictions, continues to accelerate scientific research[13].

Additionally, the structural predictions provided by AlphaFold 3 extend beyond proteins to include nucleic acids, small molecules, ions, and modified residues. This comprehensive predictive ability marks a substantial improvement over the specialized models of AlphaFold 2, which were more limited in scope[12][16][17]. The enhancements in prediction accuracy, particularly for protein-protein complexes and antibody-protein interfaces, offer researchers more reliable data to advance their studies[18]. These advancements collectively contribute to a more accessible and powerful tool for the scientific community.

Advancements in Biomolecular Structure Prediction

AlphaFold 3 and ProteinDNA Interactions

AlphaFold 3 has revolutionized the modeling of protein-DNA interactions, an essential component in understanding genetic regulation. The updated diffusion-based architecture of AlphaFold 3 enables the joint structure prediction of complexes, including not just proteins, but also nucleic acids such as DNA, small molecules, ions, and modified residues[8]. This comprehensive approach allows for significantly improved accuracy over many previous specialized tools, especially in predicting protein-ligand interactions[8]. By accurately modeling these interactions, AlphaFold 3 provides deeper insights into the mechanisms of genetic regulation and opens new avenues for developing targeted gene therapies[8].

Impact of AlphaFold 3 on Genetic Regulation

ProteinDNA Interaction Prediction

AlphaFold 3 marks a significant advancement in the prediction of protein-DNA interactions, offering enhanced capabilities compared to its predecessor. Unlike AlphaFold 2, which was optimized for predicting the structure of individual proteins, AlphaFold 3 employs a diffusion-based model that predicts raw atom coordinates, allowing it to accurately model an array of biomolecular interactions including those between proteins and nucleic acids like DNA and RNA[29].

The shift to a diffusion-based architecture enables AlphaFold 3 to achieve a remarkable improvement in prediction accuracy. Specifically, the model shows at least a 50% improvement in predicting the interactions of proteins with other molecule types, and in certain crucial categories, the accuracy has doubled compared to existing methods[30]. This enhanced prediction capability can lead to groundbreaking discoveries in gene regulation mechanisms and revolutionize our approach to genetic research and personalized medicine[30].

Introduced in collaboration with Isomorphic Labs, AlphaFold 3 goes beyond proteins to encompass a broad spectrum of biomolecules, including DNA, RNA, and small molecules known as ligands. This comprehensive approach opens new avenues for transformative science, from developing biorenewable materials and more resilient crops to accelerating drug design and genomics research[31]. By accurately predicting the interactions of proteins with DNA, AlphaFold 3 holds the potential to significantly advance our understanding of genetic regulation and assist in the development of targeted gene therapies[29][31].

ProteinRNA Interaction Prediction

AlphaFold 3 has marked a significant leap forward in the field of structural biology by enhancing its prediction accuracy for protein-DNA and protein-RNA interactions. Building upon the foundational work of AlphaFold 2, the latest iteration of AlphaFold developed by Google’s DeepMind and Isomorphic Labs in London can now predict the structure and interactions of a wide array of biomolecular systems with unprecedented precision[5][11][37]. This includes a dramatic improvement, with at least a 50% enhancement in accuracy for interactions between proteins and other molecule types compared to existing methods, and in certain crucial categories, the prediction accuracy has doubled[33].

These advancements hold transformative potential for understanding genetic regulation and expression, as the more accurate predictions can provide deeper insights into the mechanisms of gene regulation[36][37]. Such detailed atomic-level views of molecular interactions are expected to revolutionize approaches in genetic research and personalized medicine, paving the way for groundbreaking discoveries in how genes are regulated and expressed within biological systems[5][11]. This progress also means that the model is not limited to proteins but extends to DNA, RNA, and other small molecules, enabling a more comprehensive understanding of biomolecular dynamics[11].

New Features and Advancements in AlphaFold 3

AlphaFold 3 introduces several groundbreaking features and advancements in the field of biomolecular structure prediction. One of the most significant improvements is the ability to predict the structure of a wide variety of biomolecular systems more broadly and accurately than its predecessor, AlphaFold 2. This has been achieved through the use of diffusion techniques to enhance the underlying architectural model, allowing for more general predictions[16].

Notably, AlphaFold 3 has set a new benchmark in accuracy for predicting drug-like interactions, including the binding of proteins with ligands and antibodies with their target proteins. It is 50% more accurate than the best traditional methods on the PoseBusters benchmark, and it achieves this without requiring any input of structural information. This makes AlphaFold 3 the first AI system to outperform physics-based tools for biomolecular structure prediction[26].

Another significant advancement is AlphaFold 3's ability to model proteins interacting not only with other proteins but also with other biomolecules, such as DNA and RNA strands[27]. This capability is particularly valuable for understanding complex biological processes and interactions at the atomic level. Additionally, AlphaFold 3 excels in modeling protein-ligand interactions, a feature crucial for drug discovery efforts[27][28]. Accurate predictions of protein-ligand structures facilitate the identification and design of new molecules, which could potentially be developed into therapeutic drugs[28].

Early analyses have shown that AlphaFold 3 greatly outperforms AlphaFold 2.3 in certain protein structure prediction problems relevant to drug discovery, such as antibody binding[28]. This underscores the system's potential to significantly impact the pharmaceutical industry by improving the efficiency and accuracy of drug discovery processes[28].

 # Impact and Applications

AlphaFold 3, developed by Google DeepMind in collaboration with Isomorphic Labs, has made significant strides in biotechnology by accurately predicting the structure and interactions of a wide range of biological molecules, including proteins, DNA, RNA, and small molecules such as drugs[1][7][9]. This advancement has substantial implications for several fields, most notably drug discovery and genetic research.

One of the key impacts of AlphaFold 3 is its potential to dramatically accelerate the drug discovery process. By enabling precise identification of drug targets, it reduces both the time and costs associated with developing new medications, particularly for complex diseases[7][19][20][21]. The model's ability to predict how proteins interact with other molecules offers invaluable insights into the mechanisms of diseases and the development of targeted therapies[7][11]. Additionally, the integration of AlphaFold 3 with emerging technologies like self-driving laboratories could further innovate the drug discovery process, enhancing efficiency and accuracy[9][11].

In genetic research, AlphaFold 3's capability to predict protein-DNA interactions could significantly advance our understanding of genetic regulation, thereby aiding in the development of targeted gene therapies[8]. By providing an atomic-level view of biomolecular systems, including the structural impact of post-translational modifications and ions, AlphaFold 3 deepens our understanding of the biological world[11].

The introduction of AlphaFold Server, a free and accessible research tool powered by AlphaFold 3, has further democratized access to this groundbreaking technology. Researchers can now generate molecular complexes with minimal computational resources or expertise in machine learning, accelerating scientific research across the globe[13]. The server and the AlphaFold database provide open access to over 200 million protein structure predictions, fostering an environment of collaborative scientific discovery[13][20].

Drug Discovery Acceleration

AlphaFold 3 represents a significant advancement in drug discovery, offering the potential to revolutionize the field by enabling more precise identification of drug targets and reducing the time and costs associated with developing new medications, particularly for complex diseases[56][57]. Developed by Google DeepMind and Isomorphic Labs, AlphaFold 3 builds upon the success of its predecessor, AlphaFold 2, by providing accurate atomic-level views of the structure of biomolecular systems. This includes not only proteins but also DNA, RNA, and small molecule ligands, along with their interactions and structural impacts due to post-translational modifications and ions[11][55].

The AI model's ability to predict complex protein interactions and structures with high accuracy offers a new set of drug target candidates to explore, potentially leading to groundbreaking therapeutic developments[56][57]. Furthermore, the application of AlphaFold 3 in predicting the structural impact of various molecular systems opens up exciting possibilities for rational drug development against targets that were previously difficult to modulate[55].

Although the initial impact of AlphaFold and similar models like RoseTTAFold on drug discovery has been incremental, the potential commercial and scientific value of AlphaFold 3 is vast, with its transformative potential already being acknowledged as "Nobel Prize-worthy"[19][21]. By accurately predicting the three-dimensional shapes of proteins and other biomolecules, AlphaFold 3 helps streamline the process of identifying compounds that will successfully bind to these targets, producing beneficial health outcomes[57].

Moreover, integrating AlphaFold 3 with emerging technologies such as self-driving laboratories could further accelerate and innovate the drug discovery process. The combination of AlphaFold 3's structural predictions with automated, high-throughput experimentation could dramatically speed up the validation and optimization of new drug candidates, transforming our understanding and approach to drug R&D[9][44][55].

ProteinDNA and ProteinRNA Interaction Predictions

AlphaFold 3 represents a significant advancement in the prediction of biomolecular interactions, specifically those involving proteins, DNA, and RNA. Unlike its predecessor, AlphaFold 2, which primarily focused on predicting the structure of individual proteins, AlphaFold 3 employs a diffusion-based architecture to predict raw atom coordinates. This allows it to model a variety of biomolecular interactions with high accuracy, including those between proteins and nucleic acids such as DNA and RNA[29].

Introduced in 2024 by Google DeepMind and Isomorphic Labs, AlphaFold 3 expands its predictive capabilities beyond proteins to encompass all of life's molecules. This includes small molecules known as ligands, which are significant in the context of drug discovery[31]. The ability to predict interactions between proteins and DNA holds particular promise for advancing genetic regulation understanding and developing targeted gene therapies[29][31].

The predictive power of AlphaFold 3 extends to complex biomolecular interactions, including those involving protein complexes with DNA, RNA, and various ligands and ions. This enhanced capability allows for a more comprehensive understanding of biological processes and has the potential to unlock transformative scientific developments, from biorenewable materials to more resilient crops and accelerated genomics research[34]. Additionally, AlphaFold 3's success rate of approximately 70% in accurately predicting protein-protein interactions underscores its effectiveness[34].

Perhaps one of the most exciting aspects of AlphaFold 3 is its ability to model interactions between proteins and a wide range of biological molecules, including DNA and RNA. This advancement is critical for understanding the fundamental mechanisms of life and for identifying potential drug candidates, reflecting the extensive training set that includes a broad spectrum of molecules[53]. By accurately predicting these complex interactions, AlphaFold 3 has the potential to revolutionize various fields within biological research and biotechnology.

Efficiency and Accuracy Improvements in Drug Discovery

AlphaFold 3 has significantly improved the efficiency and accuracy of drug discovery processes, enabling more precise identification of drug targets and reducing the time and costs associated with developing new medications. This advancement is particularly impactful in the context of complex diseases, where traditional methods have struggled to provide swift and accurate results[7][20][59].

One notable example is the discovery of a more potent hit molecule, ISM042-2-048, through AI-powered compound generation. This compound demonstrated good inhibitory activity against CDK20, a crucial protein in hepatocellular carcinoma (HCC), with an IC50 value of 33.4 ± 22.6 nM. It also showed selective anti-proliferation activity in an HCC cell line, marking the first instance of AlphaFold being applied to hit identification in drug discovery[22]. Furthermore, scientists at Insilico and the University of Toronto have integrated AlphaFold into an end-to-end AI-powered drug discovery platform, leading to the identification of a new drug for a novel target aimed at treating HCC[23].

AlphaFold 3 has also enhanced prediction accuracy for antibody-antigen interactions, a critical area for immunology and therapeutic antibody development. By blending bioinformatics and physics, AlphaFold offers a more precise understanding of the exact binding between antibodies and antigens, surpassing the capabilities of previous computational methods[24].

Moreover, AlphaFold opens new avenues for exploring drug targets, especially in neglected diseases. These are conditions that receive little research funding due to affecting small or low-income populations, making them less attractive to commercial markets. The expanded scope of AlphaFold 3 to include a diverse range of biomolecules further paves the way for transformational science, including bio-renewable materials and more resilient crops, alongside accelerating drug discovery and genomics research[25][59].

Biotechnology Applications

In 2024, together with Isomorphic Labs, we introduced AlphaFold 3, which predicts the structure and interactions of all of life's molecules[1]. AlphaFold 3 goes beyond proteins to a broad spectrum of biomolecules including DNA, RNA, and even small molecules, also known as ligands, which encompass many drugs[1].

Genetic Regulation and Personalized Medicine

AlphaFold 3 has heralded a significant advancement in our understanding of genetic regulation and the development of personalized medicine. By leveraging a diffusion-based architecture, AlphaFold 3 can predict the structure and interactions of various biomolecular systems with unprecedented accuracy, including proteins, nucleic acids, small molecules, ions, and modified residues[8][35]. This enhanced capability allows for joint structure prediction of complex biological systems, which is critical for understanding the intricate interactions within cells[35].

One of the most groundbreaking features of AlphaFold 3 is its ability to predict protein-DNA and protein-RNA interactions with far greater accuracy compared to previous models. The new AlphaFold model has shown a significant improvement—up to 50% or more—in predicting these interactions, which are essential for understanding gene regulation and expression[30]. Such precise predictions could lead to revolutionary discoveries in the mechanisms of gene regulation, potentially transforming genetic research and opening new avenues for personalized medicine[30][35].

Furthermore, AlphaFold 3's ability to model how DNA interacts with proteins offers profound insights into cellular processes and the regulation of genetic codes. This capability can significantly advance our understanding of genetic regulation and help in the development of targeted gene therapies[50][52]. The improved accuracy in predicting these molecular interactions means that scientists can now explore genetic pathways with a level of detail previously unattainable, facilitating the creation of more effective and personalized treatment plans for various genetic disorders[8][52].

Despite these advancements, it is important to note that access to AlphaFold 3 is currently restricted, which may limit the widespread application of its capabilities in the short term[51]. However, the potential implications for genetic research and personalized medicine remain vast and promising as the technology continues to evolve and become more accessible to the scientific community.

Economic Impact and Market Implications

AlphaFold 3, a groundbreaking artificial intelligence program developed by Google DeepMind and Isomorphic Labs, has been heralded for its transformative potential in drug discovery and development, which could have substantial economic implications[21]. The program predicts the structure and interactions of all of life's molecules with remarkable accuracy, a significant advancement in the field of genetics[32][67].

By dramatically reducing the cost and time associated with protein structure determination, AlphaFold 3 has the potential to expedite research and development processes in the pharmaceutical industry, leading to significant cost savings[66]. This reduction in costs can benefit multiple sectors, particularly companies focused on biotechnology and pharmaceuticals. Stocks of companies involved in these sectors, such as Alphabet Inc., ABBV, EXAI, IBM, MRK, MSFT, NVS, ORCL, PFE, SDGR, and SLP, could see a positive impact due to the advancements brought about by AlphaFold 3[68].

Moreover, the program's ability to predict protein structures has practical applications in addressing global health challenges. For instance, researchers at the University of Cambridge are utilizing AlphaFold to develop a more effective malaria vaccine, while teams at the University of Colorado are exploring solutions to antibiotic resistance, a major public health concern[70]. These efforts underscore the potential for AlphaFold 3 to contribute to significant medical breakthroughs, which could further bolster investor confidence in related sectors.

However, it remains uncertain how exactly AlphaFold 3 will catalyze drug discovery and development. Many drugs fail to reach the market due to unforeseen interactions between their components and various parts of the body[71]. Despite this uncertainty, the AI-driven advancements of AlphaFold 3 hold promise for overcoming some of these traditional hurdles, thereby accelerating the pipeline from research to market.

 # References

[1] AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/technologies/alphafold/

[2] AlphaFold - Wikipedia. Retrieved June 3, 2024, from https://en.wikipedia.org/wiki/AlphaFold

[3] AlphaFold 3 predicts the structure and interactions of all of life's molecules - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/alphafold-3-predicts-the-structure-and-interactions-of-all-of-lifes-molecules

[4] AlphaFold 3 predicts the structure and interactions of all of life's molecules. Retrieved June 3, 2024, from https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/

[5] AlphaFold 3 upgrade enables the prediction of other types of biomolecular systems. Retrieved June 3, 2024, from https://phys.org/news/2024-05-alphafold-enables-biomolecular.html

[6] Rational drug design with AlphaFold 3 - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/rational-drug-design-with-alphafold-3

[7] Google DeepMind's AlphaFold 3 Could Transform Drug Discovery. Retrieved June 3, 2024, from https://time.com/6975934/google-deepmind-alphafold-3-ai/

[8] Accurate structure prediction of biomolecular interactions with AlphaFold 3 | Nature. Retrieved June 3, 2024, from https://www.nature.com/articles/s41586-024-07487-w

[9] AlphaFold 3 predicts the structure and interactions of all of life's molecules. Retrieved June 3, 2024, from https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/

[10] A glimpse of the next generation of AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/

[11] Rational drug design with AlphaFold 3 - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/rational-drug-design-with-alphafold-3

[12] Accurate structure prediction of biomolecular interactions with AlphaFold 3 | Nature. Retrieved June 3, 2024, from https://www.nature.com/articles/s41586-024-07487-w

[13] AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/technologies/alphafold/

[14] AlphaFold 3: Google DeepMind's latest AI tech in drug discovery. Retrieved June 3, 2024, from https://www.prescouter.com/2024/05/google-deepmind-alphafold-3/

[15] AlphaFold - Wikipedia. Retrieved June 3, 2024, from https://en.wikipedia.org/wiki/AlphaFold

[16] AlphaFold 3 upgrade enables the prediction of other types of biomolecular systems. Retrieved June 3, 2024, from https://phys.org/news/2024-05-alphafold-enables-biomolecular.html

[17] AlphaFold 3: A Leap Forward in Biomolecular Structure Prediction — Opportunities and Limitations | by Freedom Preetham | Meta Multiomics | May, 2024 | Medium. Retrieved June 3, 2024, from https://medium.com/meta-multiomics/alphafold-3-a-leap-forward-in-biomolecular-structure-prediction-opportunities-and-limitations-924350af1181

[18] AlphaFold3 and its improvements in comparison to AlphaFold2 | by Falk Hoffmann | May, 2024 | Medium. Retrieved June 3, 2024, from https://medium.com/@falk_hoffmann/alphafold3-and-its-improvements-in-comparison-to-alphafold2-96815ffbb044

[19] What does AlphaFold mean for drug discovery?. Retrieved June 3, 2024, from https://www.nature.com/articles/d41573-021-00161-0

[20] DeepMind's latest AlphaFold model is more useful for drug discovery | TechCrunch. Retrieved June 3, 2024, from https://techcrunch.com/2023/10/31/deepminds-latest-alphafold-model-is-more-useful-for-drug-discovery/

[21] Why AlphaFold 3 is stirring up so much buzz in pharma | PharmaVoice. Retrieved June 3, 2024, from https://www.pharmavoice.com/news/google-alphafold-3-drug-discovery-pharma-buzz/716496/

[22] AlphaFold accelerates artificial intelligence powered drug discovery: efficient discovery of a novel CDK20 small molecule inhibitor - PMC. Retrieved June 3, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9906638/

[23] First Application of AlphaFold in Identifying Potential Liver Cancer Drug. Retrieved June 3, 2024, from https://www.genengnews.com/topics/drug-discovery/first-application-of-alphafold-in-identifying-potential-liver-cancer-drug/

[24] AlphaFold 3: Google DeepMind's latest AI tech in drug discovery. Retrieved June 3, 2024, from https://www.prescouter.com/2024/05/google-deepmind-alphafold-3/

[25] AlphaFold Is The Most Important Achievement In AI—Ever. Retrieved June 3, 2024, from https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/

[26] AlphaFold 3 predicts the structure and interactions of all of life's molecules. Retrieved June 3, 2024, from https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/

[27] AlphaFold 3 offers even more accurate protein structure prediction. Retrieved June 3, 2024, from https://www.drugdiscoverytrends.com/meet-alphafold-3-which-can-accurately-model-more-than-99-of-molecular-types-in-the-protein-data-bank/

[28] A glimpse of the next generation of AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/

[29] AlphaFold 3 offers even more accurate protein structure prediction. Retrieved June 3, 2024, from https://www.drugdiscoverytrends.com/meet-alphafold-3-which-can-accurately-model-more-than-99-of-molecular-types-in-the-protein-data-bank/

[30] AlphaFold 3 predicts the structure and interactions of all of life's molecules - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/alphafold-3-predicts-the-structure-and-interactions-of-all-of-lifes-molecules

[31] AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/technologies/alphafold/

[32] AlphaFold 3 predicts the structure and interactions of all of life's molecules - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/alphafold-3-predicts-the-structure-and-interactions-of-all-of-lifes-molecules

[33] AlphaFold 3 predicts the structure and interactions of all of life's molecules. Retrieved June 3, 2024, from https://blog.google/technology/ai/google-deepmind-isomorphic-alphafold-3-ai-model/

[34] AlphaFold - Wikipedia. Retrieved June 3, 2024, from https://en.wikipedia.org/wiki/AlphaFold

[35] AlphaFold 3 upgrade enables the prediction of other types of biomolecular systems. Retrieved June 3, 2024, from https://phys.org/news/2024-05-alphafold-enables-biomolecular.html

[36] AlphaFold 3 offers even more accurate protein structure prediction. Retrieved June 3, 2024, from https://www.drugdiscoverytrends.com/meet-alphafold-3-which-can-accurately-model-more-than-99-of-molecular-types-in-the-protein-data-bank/

[37] A glimpse of the next generation of AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/discover/blog/a-glimpse-of-the-next-generation-of-alphafold/

[38] AlphaFold3: A Game Changer in Protein Structure Prediction — Part 1 | by Chithra Srinivasan | May, 2024 | Medium. Retrieved June 3, 2024, from https://medium.com/@csn289/alphafold3-a-game-changer-in-protein-structure-prediction-part-1-b8d9c361bda2

[39] AlphaFold 3 offers even more accurate protein structure prediction. Retrieved June 3, 2024, from https://www.drugdiscoverytrends.com/meet-alphafold-3-which-can-accurately-model-more-than-99-of-molecular-types-in-the-protein-data-bank/

[40] Frontiers | Before and after AlphaFold2: An overview of protein structure prediction. Retrieved June 3, 2024, from https://www.frontiersin.org/articles/10.3389/fbinf.2023.1120370/full

[41] AlphaFold 3: A Leap Forward in Biomolecular Structure Prediction — Opportunities and Limitations | by Freedom Preetham | Meta Multiomics | May, 2024 | Medium. Retrieved June 3, 2024, from https://medium.com/meta-multiomics/alphafold-3-a-leap-forward-in-biomolecular-structure-prediction-opportunities-and-limitations-924350af1181

[42] AlphaFold - Wikipedia. Retrieved June 3, 2024, from https://en.wikipedia.org/wiki/AlphaFold

[43] Rational drug design with AlphaFold 3 - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/rational-drug-design-with-alphafold-3

[44] The Drug Discoverer - Reflecting on DeepMind's AlphaFold artificial intelligence success – what's the real significance for protein folding research and drug discovery? - The Institute of Cancer Research, London. Retrieved June 3, 2024, from https://www.icr.ac.uk/blogs/the-drug-discoverer/page-details/reflecting-on-deepmind-s-alphafold-artificial-intelligence-success-what-s-the-real-significance-for-protein-folding-research-and-drug-discovery

[45] AlphaFold Is The Most Important Achievement In AI—Ever. Retrieved June 3, 2024, from https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/

[46] AlphaFold 3 predicts the structure and interactions of all of life's molecules - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/alphafold-3-predicts-the-structure-and-interactions-of-all-of-lifes-molecules

[47] Great expectations – the potential impacts of AlphaFold DB | EMBL. Retrieved June 3, 2024, from https://www.embl.org/news/science/alphafold-potential-impacts/

[48] AlphaFold: Accelerating biological research - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/impact/meet-the-scientists-using-alphafold/

[49] How is AlphaFold2 used by scientists? | AlphaFold. Retrieved June 3, 2024, from https://www.ebi.ac.uk/training/online/courses/alphafold/validation-and-impact/how-is-alphafold-used-by-scientists/

[50] AlphaFold 3 Will Change the Biological World and Drug Discovery. Retrieved June 3, 2024, from https://www.analyticsvidhya.com/blog/2024/05/deepmind-ai-alphafold/

[51] Major AlphaFold upgrade offers boost for drug discovery. Retrieved June 3, 2024, from https://www.nature.com/articles/d41586-024-01383-z

[52] AlphaFold 3 upgrade enables the prediction of other types of biomolecular systems. Retrieved June 3, 2024, from https://phys.org/news/2024-05-alphafold-enables-biomolecular.html

[53] AlphaFold 3: Stepping into the future of structure prediction - Front Line Genomics. Retrieved June 3, 2024, from https://frontlinegenomics.com/alphafold-3-stepping-into-the-future-of-structure-prediction/

[54] Analyzing the potential of AlphaFold in drug discovery | MIT News | Massachusetts Institute of Technology. Retrieved June 3, 2024, from https://news.mit.edu/2022/alphafold-potential-protein-drug-0906

[55] Rational drug design with AlphaFold 3 - Isomorphic Labs. Retrieved June 3, 2024, from https://www.isomorphiclabs.com/articles/rational-drug-design-with-alphafold-3

[56] AlphaFold 3: Google DeepMind's latest AI tech in drug discovery. Retrieved June 3, 2024, from https://www.prescouter.com/2024/05/google-deepmind-alphafold-3/

[57] AlphaFold Is The Most Important Achievement In AI—Ever. Retrieved June 3, 2024, from https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/

[58] AlphaFold 3: Revolutionizing drug discovery and molecular biology. Retrieved June 3, 2024, from https://www.prescouter.com/2024/05/alphafold-3/

[59] Google DeepMind's AI model AlphaFold 3 can be a gamechanger in drug discovery. Retrieved June 3, 2024, from https://indiaai.gov.in/article/google-deepmind-s-ai-model-alphafold-3-can-be-a-gamechanger-in-drug-discovery

[60] AlphaFold Is The Most Important Achievement In AI—Ever. Retrieved June 3, 2024, from https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/

[61] New study uses AlphaFold and AI to accelerate design of novel drug for liver cancer. Retrieved June 3, 2024, from https://acceleration.utoronto.ca/news/new-study-uses-alphafold-and-ai-to-accelerate-design-of-novel-drug-for-liver-cancer

[62] AlphaFold 3: Google DeepMind's latest AI tech in drug discovery. Retrieved June 3, 2024, from https://www.prescouter.com/2024/05/google-deepmind-alphafold-3/

[63] AlphaFold Is The Most Important Achievement In AI—Ever. Retrieved June 3, 2024, from https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/

[64] The rise of self-driving labs in chemical and materials sciences | Nature Synthesis. Retrieved June 3, 2024, from https://www.nature.com/articles/s44160-022-00231-0

[65] Self-driving laboratories to autonomously navigate the protein fitness landscape | Nature Chemical Engineering. Retrieved June 3, 2024, from https://www.nature.com/articles/s44286-023-00002-4

[66] r/stocks on Reddit: Which stocks will benefit most from alphafold protein fold prediction advancement?. Retrieved June 3, 2024, from https://www.reddit.com/r/stocks/comments/k7z8hi/which_stocks_will_benefit_most_from_alphafold/

[67] Did Google's DeepMind Just Revolutionize Medicine? Retrieved June 3, 2024, from https://www.fool.com/investing/2020/12/05/did-googles-deepmind-just-revolutionize-medicine/

[68] Alphabet's AlphaFold Is A Gift To Humanity, Not Investors | Seeking Alpha. Retrieved June 3, 2024, from https://seekingalpha.com/article/4522954-alphabets-alphafold-is-a-gift-to-humanity-not-investors

[69] DeepMind's AlphaFold AI Origin Story, Influence on Biotech. Retrieved June 3, 2024, from https://www.businessinsider.com/deepmind-alphafold-ai-origin-story-impact-on-biotech-2023-2

[70] AlphaFold - Google DeepMind. Retrieved June 3, 2024, from https://deepmind.google/technologies/alphafold/

[71] Folds and faults: On AlphaFold 3 - The Hindu. Retrieved June 3, 2024, from https://www.thehindu.com/opinion/editorial/folds-and-faults-on-alphafold-3/article68161671.ece

[72] Software Engineers Lead Drug Breakthough Terry 11:07 AM - Markman on Tech. Retrieved June 3, 2024, from https://www.thestreet.com/tech/news/sdgrjdm121420

Table 12: Co-STORM's generated final report for "Alpha Fold 3". "#", "##" indicate the section title and subsection title respectively. Numbers in brackets indicate the cited references.