Automated Knowledge Synthesis and Collaborative Discovery: A Unified Study of STORM and Co-STORM Frameworks for Grounded Long-Form Report Generation

Rony Preetam[1], Amit Patil[2], G. Banidhar[3], Ketan Bemalkhedkar[4], Pranav Patil[5]
[1,2,3,4,5] *Department of Artificial Intelligence and Machine Learning, Guru Nanak Dev Engineering College, Bidar*
doi.org/10.64643/IJIRTV12I12-200357-459

***Abstract*—Producing structured, factually grounded long-form articles entirely from scratch continues to pose significant difficulties for contemporary large language model (LLM) pipelines, chiefly because the cognitively demanding pre-writing stage—source identification, perspective mapping, question formulation, and evidence-based outline construction—remains poorly automated. This paper offers a unified examination of two interrelated systems that tackle complementary aspects of the knowledge-synthesis problem. STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking) mechanises the pre-writing phase by orchestrating simulated multi-perspective dialogues between persona-driven writers and a retrieval-grounded expert, subsequently assembling a hierarchical outline before article drafting begins. Co-STORM extends this paradigm into a collaborative, human-steerable discourse model: users observe and intermittently redirect conversations among several LM agents while a dynamic mind map tracks the evolving knowledge landscape in real time. Both frameworks leverage the DSPy prompting toolkit and retrieval-augmented generation (RAG). Evaluation on the FreshWiki benchmark and the newly curated WildSeek dataset confirms that STORM exceeds outline-driven RAG baselines in breadth and organisational coherence, and that Co-STORM surpasses conventional search engines and RAG chatbots in depth, novelty, and user-reported satisfaction. Expert Wikipedia editors and lay user cohorts alike express strong preference for the outputs of both systems, substantiating the practical viability of LLM-driven knowledge-synthesis pipelines for scholarly and encyclopaedic writing tasks.**

***Index Terms*—Large Language Models, Retrieval-Augmented Generation, Long-Form Text Generation, Collaborative Information Seeking, Pre-Writing Automation, Wikipedia-Like Article Generation, Multi-Agent Systems, Knowledge Discovery.**

I. INTRODUCTION

Large language models have demonstrated a remarkable capacity for generating fluent, coherent prose across a wide spectrum of domains. Translating that fluency into trustworthy, exhaustively researched documents on par with high-quality Wikipedia entries, however, demands considerably more than raw generation prowess. Writers—whether human or machine—must first conduct a rigorous pre-writing process: surveying pertinent sources, charting diverse viewpoints, posing incisive questions, and weaving the retrieved evidence into an organised outline before composing even a single sentence of the final document [1].

The challenge intensifies when the subject falls outside the model's training distribution. This phenomenon, sometimes dubbed the "long-tail knowledge problem" [19], means that parametric memory alone frequently yields incomplete or outright hallucinated content. Retrieval-augmented generation (RAG) [3] partially mitigates the issue by grounding responses in external corpora, yet a naïve single-query retrieval typically returns only a narrow slice of the available evidence, leaving considerable topical terrain unexplored.

A second, equally pressing challenge resides in the human dimension of information seeking. Users venturing into unfamiliar domains often operate within the territory of unknown unknowns—they do not know what they do not know [14]. Traditional search engines and question-answering chatbots are inherently reactive: they await a user query before retrieving. Consequently, users who lack the vocabulary or background knowledge to pose

precise questions end up with partial answers that reinforce existing mental models rather than expanding them [17].

This work presents a thorough analysis of two systems engineered to address these twin challenges. The first, STORM (Synthesis of Topic Outlines through Retrieval and Multi-perspective Question Asking), automates the pre-writing stage by orchestrating simulated conversations between perspective-driven Wikipedia writers and a retrieval-grounded expert [1]. The second, Co-STORM (Collaborative STORM), introduces a human-steerable, multi-agent discourse in which users observe—and occasionally redirect—ongoing dialogue among simulated domain experts and a moderator agent, while a dynamic mind map tracks accumulated knowledge in real time [2].

The principal contributions of this study are as follows: (i) a rigorous exposition of the STORM pre-writing framework together with its FreshWiki evaluation benchmark; (ii) a detailed description of the Co-STORM collaborative discourse architecture and the WildSeek evaluation dataset; (iii) comparative analyses of both systems against strong RAG and conversational baselines; and (iv) a synthesis of lessons drawn from expert and user evaluations that highlights open research problems in grounded, human-centred document creation.

II. RELATED WORK

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation was introduced as a principled mechanism for injecting dynamically retrieved evidence into generative pipelines [3]. Subsequent research extended this basic idea to active and adaptive retrieval strategies, multi-hop reasoning chains, and citation-grounded text generation [5]. STORM adopts a query-decomposition variant of RAG wherein the LLM first decomposes a complex research question into targeted search sub-queries before issuing retrieval calls, thereby improving source coverage relative to monolithic query approaches.

2.2 Automatic Expository and Wikipedia Generation

Early attempts at automatic Wikipedia article creation either assumed access to pre-existing outlines or confined themselves to single-paragraph generation within narrow domains [10], [12]. The WikiSum framework [11] reframed Wikipedia article generation as multi-document summarisation conditioned on retrieved references, sidestepping the outline-construction problem altogether. More recent scholarship recognises the critical role of the author's sensemaking process and outline planning in producing coherent expository text [23], thereby motivating the staged approach that STORM adopts.

2.3 Information-Seeking Support and Collaborative Discourse

The information-science literature draws a useful distinction between known unknowns—questions a user is aware of—and unknown unknowns, serendipitous knowledge encountered unexpectedly during exploratory search [14]. Conversational search systems have begun to address the former category but rarely the latter. Research on collaborative discourse in educational settings demonstrates that structured group discussion deepens understanding, raises engagement, and surfaces information that no single participant would independently discover [13], [16]. Co-STORM operationalises these insights within an LLM-agent framework, assigning distinct epistemic roles to each participant.

2.4 Multi-Agent LLM Systems

A growing body of literature investigates the use of LLMs in multi-agent configurations, demonstrating that disagreement and diverse role-play improve factuality and reasoning beyond single-model performance. Most prior systems, however, treat multi-agent interaction as an autonomous pipeline with no provision for human participation. Co-STORM bridges this gap by treating the human user as a first-class discourse participant who may interject at any turn, shifting initiative between human and machine as the conversation evolves [2].

III. PROBLEM STATEMENT AND NOTATION

Let t denote a topic of interest and g denote a user-specified goal or intent (g may be null in the non-interactive STORM setting). The objective is to produce a grounded long-form report $S = s_1 s_2 \ldots s_n$ by simultaneously discovering a reference set R from an external information repository and constructing an organisational outline O. Each sentence s_i cites a subset of documents drawn from R, thereby ensuring verifiability.

Existing systems satisfy at most two of three desiderata: (1) multi-source synthesis, (2) ongoing user interaction, and (3) curated report output. STORM satisfies desiderata (1) and (3); Co-STORM satisfies all three [2]. Formalising this tripartite requirement motivates the architectural choices described in subsequent sections.

IV. STORM: PRE-WRITING VIA MULTI-PERSPECTIVE QUESTION ASKING

STORM decomposes the article-creation workflow into two sequential stages: pre-writing, during which references and an outline are assembled; and writing, during which the outline is expanded into a full-length article. This mirrors the compositional process described in classical writing pedagogy, where outline planning is recognised as the cognitive engine that enables well-structured drafting [9].

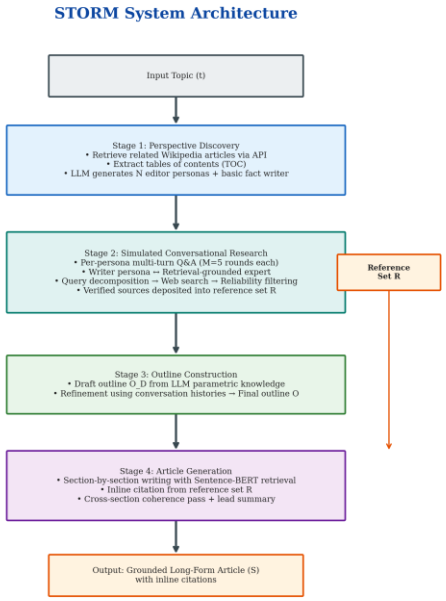

Fig. 1. End-to-end architecture of the STORM system showing the four-stage pipeline from topic input through perspective discovery, simulated conversational research, outline construction, and article generation.

4.1 Perspective Discovery

A naïve approach to generating research questions is to prompt the LLM directly to enumerate queries about the topic. Such direct prompting reliably produces surface-level factual questions ("When was X founded?", "Where is Y located?") but seldom surfaces the nuanced angles specific to a given topic [1]. STORM replaces this with a two-step perspective-discovery protocol.

First, the system identifies a set of related Wikipedia topics by prompting the LLM with the target topic t. For each related topic whose Wikipedia article exists, the table of contents (TOC) is extracted via the Wikipedia API and concatenated to form a context document. Second, the LLM is prompted to analyse this aggregate context and generate N distinct editor personas $P = \{p_1, \ldots, p_n\}$, each characterised by a specific angle, expertise, or stakeholder role. A zeroth persona p_0 representing a "basic fact writer" is always prepended to ensure broad coverage of foundational information.

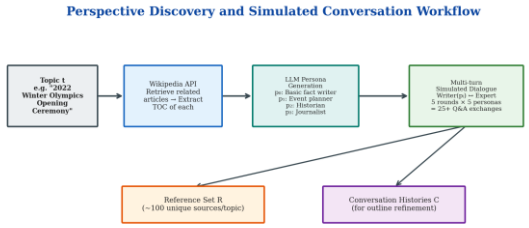

Fig. 2. Perspective discovery and simulated conversation workflow. The topic t yields related Wikipedia articles, from which editor personas are generated. Each persona drives a multi-turn dialogue producing reference set R and conversation histories C.

4.2 Simulated Conversational Research

For each persona p_i, STORM instantiates a simulated dialogue between an LLM-powered Wikipedia writer carrying perspective p_i and an LLM-powered domain expert whose answers are grounded in retrieved web sources. This design draws on question-asking theory [21], which posits that answering one question frequently gives rise to deeper follow-up questions. The conversational history h_i^{-1} conditions the writer's next question, enabling iterative depth of inquiry absent from one-shot question generation.

Each question q_i is decomposed into multiple search sub-queries by prompting the LLM. Retrieved web pages are subjected to a rule-based reliability filter derived from Wikipedia's verifiability guidelines, and surviving passages are synthesised by the expert LLM into a cited answer a_i. All verified sources are deposited into R for use during article generation. STORM limits conversations to M = 5 rounds per persona and instantiates up to N = 5 personas, yielding up to 30 question–answer exchanges per topic [1].

4.3 Outline Construction and Article Generation

Once all conversations are complete, STORM synthesises the outline in two passes. A draft outline O_D is first produced by prompting the LLM with only the topic t, leveraging the model's parametric knowledge to establish a high-level skeleton. This draft is then refined by a second prompt that incorporates both O_D and the concatenated conversation histories $C = \{C_0, C_1, \ldots, C_n\}$, yielding the final outline O enriched with topic-specific detail surfaced during the research stage.

The article is written section by section. For each section, semantically relevant documents are retrieved from R using Sentence-BERT embeddings [8], and the LLM generates the section text with inline citations. Sections are composed in parallel to reduce latency, then concatenated. A final coherence pass removes cross-section redundancy, and a lead summary section is prepended following Wikipedia's stylistic conventions.

V. CO-STORM: COLLABORATIVE DISCOVERY

While STORM treats article generation as a fully automated pipeline, Co-STORM reconceives the process as a human-steerable collaborative discourse. The underlying motivation is epistemic: users who do not know what questions to ask cannot benefit from reactive question-answering systems. By allowing users to observe—and occasionally redirect—an ongoing multi-agent conversation, Co-STORM surfaces unknown unknowns through a mechanism akin to eavesdropping on expert debate [2].

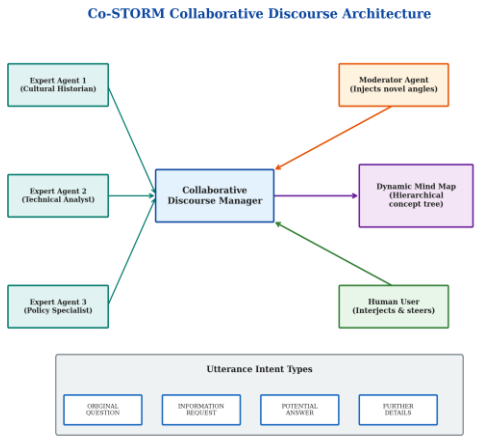

Fig. 3. Co-STORM collaborative discourse architecture showing the interaction between perspective-guided expert agents, a moderator agent, the human user, and the dynamic mind map. Utterance intent types are listed at bottom.

5.1 Collaborative Discourse Protocol

Co-STORM maintains a discourse $D = \{u_1, u_2, \ldots, u_n\}$ consisting of turn-based textual utterances attributed to one of three roles: (1) perspective-guided expert agents, each personalised with a distinct background derived from retrieval on the topic t; (2) a moderator agent, responsible for injecting novel discussion angles based on unused retrieved information; and (3) the human user, who may interject at any turn to redirect or deepen the discourse [2].

Each expert utterance is associated with one of four intent types: ORIGINAL QUESTION (initiating a new line of inquiry), INFORMATION REQUEST (seeking elaboration on the prior utterance), POTENTIAL ANSWER (providing a retrieved and synthesised response), or FURTHER DETAILS (supplying supplementary information). This intent taxonomy, inspired by conversational information-seeking research, allows the system to reason about discourse coherence and progress. Turn management follows a deterministic rotation among experts, with the moderator intervening whenever $L = 2$ consecutive turns consist solely of answers or elaborations, signalling that the discourse has stagnated within a narrow sub-topic.

5.2 Dynamic Mind Map

Tracking multi-party discourse is cognitively demanding for human observers. Co-STORM addresses this by maintaining a hierarchical mind map $M = (C, E)$, a tree in which nodes represent discovered concepts and directed edges encode parent–child topical relationships [18]. Every retrieved information chunk is inserted into M via two operations: insert, which places new content under the most semantically appropriate concept node; and reorganize, which automatically subdivides over-populated nodes into sub-topics.

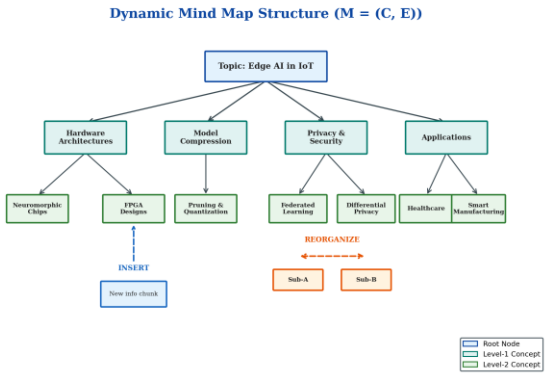

Fig. 4. Dynamic mind map structure showing hierarchical concept organisation with INSERT and REORGANIZE operations. The tree expands as new information chunks are retrieved during the collaborative discourse.

Controlled experiments on FreshWiki confirm that Co-STORM's hybrid insert strategy—first computing embedding similarity to identify candidate nodes, then invoking an LLM for final placement—outperforms both pure embedding matching (24.24% first-level accuracy) and pure LLM placement (3.03% first-level accuracy), achieving 39.39% first-level accuracy. The mind map is rendered in real time in the Co-STORM web interface, allowing users to click any node to highlight associated discourse turns and verify provenance [2].

5.3 Moderator Design

A key design challenge lies in ensuring that the moderator introduces genuinely novel angles rather than simply restating material already covered. The moderator reranks candidate unused information snippets using the scoring function cos(i, t)^α · (1 − cos(i, q))^(1−α), where i, t, and q are embedding vectors of the snippet, the topic, and the question that originally retrieved the snippet, respectively, and α = 0.5. This scoring mechanism rewards topical relevance while penalising redundancy with the retrieval query, biasing the moderator toward information that broadens rather than deepens the current thread [2].

VI. EXPERIMENTAL SETUP AND EVALUATION

6.1 FreshWiki Benchmark

Existing Wikipedia-generation benchmarks suffer from data leakage: training corpora for contemporary LLMs inevitably contain the Wikipedia articles used as evaluation targets. To mitigate this, the STORM authors curate FreshWiki, a collection of recent, high-quality English Wikipedia articles created or heavily revised after the training cutoffs of evaluated models. Selection criteria require ORES-predicted B-class quality or above, exclude list articles and stub pages, and sample from the top-100 most-edited pages for each month from February 2022 to September 2023 [1].

TABLE I: FreshWiki Dataset Statistics

Statistic	Value
Avg. top-level sections	8.4
Avg. total headings	15.8
Avg. references per article	90.1
Quality threshold	ORES B-class+
Time range	Feb 2022 – Sep 2023

6.2 WildSeek Dataset

WildSeek is constructed from topic–goal pairs submitted by real users of the publicly accessible STORM web application. Raw submissions are filtered using rule-based heuristics (removing personally identifiable information, trivial queries, and non-English inputs) followed by binary classification with GPT-4o to retain well-motivated goals. The final dataset comprises 100 instances spanning 24 fine-grained categories across six domains: Science, Health and Fitness, Culture and Society, Lifestyle and Leisure, Social Science and Humanities, and Others [2].

6.3 Baselines

Three LLM-based baselines are evaluated for STORM: (a) Direct Generation (DG), which prompts the LLM to produce outline and article without external retrieval; (b) RAG, which augments the topic query with web search before outline construction; and (c) Outline-driven RAG (oRAG), which extends RAG by issuing per-section search queries to collect additional evidence prior to writing, yielding the strongest baseline. For Co-STORM, two baselines are considered: a RAG Chatbot operating in one-question-one-answer mode, and STORM+QA, which combines STORM's report generation with a retrieval-backed QA module for follow-up queries. All methods use the same underlying LLMs for fairness [1], [2].

VII. RESULTS AND ANALYSIS

7.1 STORM Outline Quality

STORM achieves heading soft recall of 86.26% with GPT-3.5 and 92.73% with GPT-4, compared with 80.23% and 87.66% for Direct Generation, respectively. Heading entity recall improves from 32.39% (DG, GPT-3.5) to 40.52% (STORM, GPT-3.5), a gain of approximately 8 percentage points. Ablation studies confirm that both components of STORM's question-asking mechanism—persona-guided prompting and conversational follow-up—are necessary: removing either degrades recall, with the removal of conversational structure causing the steeper drop [1].

TABLE II: Outline Quality Comparison Across Methods on FreshWiki

Method	LLM	Heading Soft Recall (%)	Heading Entity Recall (%)
Direct Gen.	GPT-3.5	80.23	32.39
Direct Gen.	GPT-4	87.66	39.81
RAG	GPT-3.5	81.42	34.71
oRAG	GPT-3.5	82.50	36.18
oRAG	GPT-4	89.55	43.53
STORM	GPT-3.5	86.26	40.52
STORM	GPT-4	92.73	45.91

A comparison of unique reference counts further illustrates the value of multi-perspective conversational research: the full STORM pipeline collects an average of 99.83 unique sources per topic, versus 54.36 without persona guidance and 39.56 without conversational follow-up. Greater source diversity directly correlates with richer outline coverage [1].

7.2 STORM Article Quality and Human Evaluation

TABLE III: Full-Article Quality Metrics: STORM vs. oRAG Baseline (GPT-4)

Metric	oRAG	STORM	Improvement
Entity Recall	12.57	14.10	+1.53
Interest Level	3.90	3.99	+0.09
Relevance & Focus	4.09	4.45	+0.36
Broad Coverage	4.70	4.88	+0.18
ROUGE-1	44.26	45.82	+1.56
ROUGE-L	16.51	16.70	+0.19

Ten experienced Wikipedia editors (500+ edits, at least one year of tenure) evaluated 20 article pairs. STORM articles were judged more organised by a margin of 25 percentage points and superior in coverage by 10 percentage points relative to oRAG outputs. Editors unanimously agreed that STORM aids pre-writing; 80% found it useful for editing articles on new topics; and 70% endorsed its broader utility to the Wikipedia community. Qualitative feedback, however, identified two persistent weaknesses: (1) emotional or promotional tone inherited from biased web sources, and (2) over-association of topically adjacent but logically unrelated facts—an issue termed the "red herring fallacy" [1].

7.3 Co-STORM Evaluation Results

TABLE IV: Co-STORM vs. Baselines on WildSeek (Automatic Metrics)

Metric	RAG Chatbot	STORM +QA	Co-STORM
Depth	3.43	3.43	3.77
Novelty	2.30	2.50	3.05
Unique cited URLs/turn	2.01	2.89	6.04
Info. diversity	0.521	0.547	0.602
Consistency	4.24	4.34	4.40

Engagement	3.87	4.11	4.33

Co-STORM significantly outperforms both baselines on depth (3.77 vs. 3.43 for STORM+QA), novelty (3.05 vs. 2.50), and unique cited URLs per turn (6.04 vs. 2.89). Information diversity is also highest for Co-STORM (0.602). The utterance polishing step—which converts raw question or answer text into natural conversational prose—contributes to this improvement by serving as a self-refinement mechanism [2].

TABLE V: Human Evaluation Results (5-point Likert Scale, 20 Participants)

Metric	Google Search	RAG Chatbot	Co-STORM
Relevance	3.60	3.44	3.89
Breadth	3.40	3.11	4.22
Depth	3.20	3.22	3.67
Serendipity	2.70	2.78	3.90
Overall pref. vs Co-STORM	30%	22%	—

Twenty volunteers with diverse backgrounds conducted information-seeking sessions on two topics each. Against the search engine, Co-STORM achieves statistically significant advantages in serendipity (3.90 vs. 2.70, $p < 0.05$). Against the RAG Chatbot, Co-STORM outperforms on breadth (4.22 vs. 3.11, $p < 0.05$) and serendipity (3.78 vs. 2.78, $p < 0.05$). Pairwise preference questions reveal that 70% of participants preferred Co-STORM over the search engine and 78% over the RAG Chatbot [2].

VIII. DISCUSSION

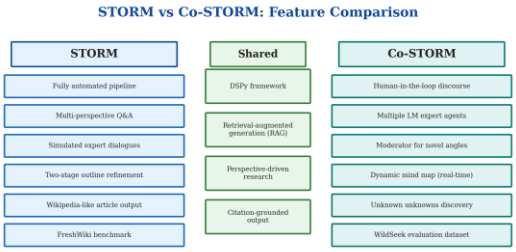

Fig. 5. Feature comparison of STORM and Co-STORM systems, highlighting shared foundations (DSPy, RAG, perspective-driven research) and distinct capabilities.

8.1 Source Bias and Neutrality

Both systems inherit the biases embedded in their retrieval sources. Internet content disproportionately represents commercially motivated, English-language, and Western perspectives. Expert evaluators observed that STORM-generated articles frequently employed superlatives, emotive adjectives, and promotional phrasing characteristic of marketing copy. Mitigating source bias requires complementary advances in source diversity (prioritising academic, governmental, and non-profit sources), content sifting (detecting and down-weighting promotional signals), and neutral paraphrasing (rewriting retrieved content in encyclopaedic voice) [1]. These remain important open problems.

8.2 Verifiability Beyond Hallucination

Citation-quality analysis of STORM-generated articles reveals that approximately 15% of sentences are unsupported by their cited references. The dominant error category is improper inferential linking: the LLM combines two independently retrieved facts into an assertion not supported by either source in isolation. This form of unfaithfulness differs fundamentally from hallucination (fabricating non-existent content) and demands distinct remediation strategies, such as entailment checking at the inference level rather than the fact level [20], [24]. Addressing inferential overreach is a critical step toward building genuinely trustworthy knowledge-synthesis systems.

8.3 Latency and Scalability

Co-STORM incurs higher end-to-end latency than STORM owing to the sequential nature of collaborative discourse, intent classification, and mind-map updates. In human evaluations, this latency was deemed acceptable, but real-world deployment at scale would benefit from parallelised retrieval pipelines, cached embedding computations,

and hierarchical discourse summarisation to avoid reprocessing the full conversation history at each turn [2].

8.4 Multilingual and Multimodal Extensions

Both systems presently generate English-language articles grounded exclusively in textual sources. Extending to multilingual generation requires search engines capable of retrieving diverse-language content and generation models with verified multilingual fidelity. Incorporating structured data, tables, and images—prevalent in high-quality Wikipedia articles—would necessitate multimodal generation capabilities not yet integrated into either pipeline.

IX. CONCLUSION

This paper has presented a comparative and integrative analysis of STORM and Co-STORM, two LLM-powered systems that advance the state of the art in automated knowledge synthesis and grounded long-form document generation. STORM introduces perspective-guided conversational research and two-stage outline refinement, enabling the production of Wikipedia-like articles with significantly improved structural coverage relative to RAG baselines. Co-STORM extends this paradigm to interactive collaborative discourse, leveraging a dynamic mind map and a moderator agent to surface unknown unknowns for users engaged in complex information-seeking tasks.

Empirical evidence drawn from both automatic benchmarks and extensive human evaluations affirms that each system outperforms competitive alternatives on its primary evaluation criteria. Remaining challenges—source bias, inferential unfaithfulness, computational latency, and monolingual scope—constitute fertile directions for future work. Taken together, these systems demonstrate that LLMs, when scaffolded with structured retrieval, multi-perspective reasoning, and human-in-the-loop interaction, can meaningfully augment the knowledge work of writers, researchers, and learners alike.

REFERENCES

[1] Y. Shao, Y. Jiang, T. A. Kanell, P. Xu, O. Khattab, and M. S. Lam, "Assisting in writing Wikipedia-like articles from scratch with large language models," arXiv preprint arXiv:2402.14207, 2024.

[2] Y. Jiang, Y. Shao, D. Ma, S. J. Semnani, and M. S. Lam, "Into the unknown unknowns: Engaged human learning through participation in language model agent conversations," arXiv preprint arXiv:2408.15232, 2024.

[3] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems, vol. 33, pp. 9459–9474, 2020.

[4] O. Khattab et al., "DSPy: Compiling declarative language model calls into self-improving pipelines," arXiv preprint arXiv:2310.03714, 2023.

[5] T. Gao, H. Yen, J. Yu, and D. Chen, "Enabling large language models to generate text with citations," in Proc. Conf. Empirical Methods in Natural Language Processing, pp. 6465–6488, 2023.

[6] S. Kim et al., "Prometheus: Inducing fine-grained evaluation capability in language models," arXiv preprint arXiv:2310.08491, 2023.

[7] A. Q. Jiang et al., "Mistral 7B," arXiv preprint arXiv:2310.06825, 2023.

[8] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in Proc. Conf. Empirical Methods in NLP, pp. 3982–3992, 2019.

[9] D. G. Rohman, "Pre-writing the stage of discovery in the writing process," College Composition and Communication, vol. 16, no. 2, pp. 106–112, 1965.

[10] N. Balepur, J. Huang, and K. Chang, "Expository text generation: Imitate, retrieve, paraphrase," in Proc. Conf. Empirical Methods in Natural Language Processing, pp. 11896–11919, 2023.

[11] P. J. Liu et al., "Generating Wikipedia by summarizing long sequences," in Proc. Int. Conf. Learning Representations, 2018.

[12] A. Fan and C. Gardent, "Generating biographies on Wikipedia: The impact of gender bias on the retrieval-based generation of women biographies," in Proc. 60th Annual Meeting ACL, vol. 1, pp. 8561–8576, 2022.

[13] E. M. Nussbaum, "Collaborative discourse, argumentation, and learning: Preface and literature review," Contemporary Educational Psychology, vol. 33, no. 3, pp. 345–359, 2008.

[14] A. Foster and N. Ford, "Serendipity and information seeking: An empirical study," Journal of Documentation, vol. 59, no. 3, pp. 321–340, 2003.
[15] P. Pirolli, "Powers of 10: Modeling complex information-seeking systems at multiple scales," Computer, vol. 42, no. 3, pp. 33–40, 2009.
[16] J. Roschelle and S. D. Teasley, "The construction of shared knowledge in collaborative problem solving," in Computer Supported Collaborative Learning, pp. 69–97, Springer, 1995.
[17] C. C. Kuhlthau, "Inside the search process: Information seeking from the user's perspective," J. Amer. Soc. Information Science, vol. 42, no. 5, pp. 361–371, 1991.
[18] T. Buzan, Using Both Sides of Your Brain, ch. 4, pp. 71–116. E. P. Dutton, New York, 1974.
[19] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," in Int. Conf. Machine Learning, pp. 15696–15707, 2023.
[20] S. Semnani, V. Yao, H. Zhang, and M. Lam, "WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia," in Findings of ACL: EMNLP, pp. 2387–2413, 2023.
[21] A. Ram, "A theory of questions and question asking," J. Learning Sciences, vol. 1, no. 3–4, pp. 273–318, 1991.
[22] S. Kim et al., "Prometheus 2: An open source language model specialized in evaluating other language models," arXiv preprint arXiv:2405.01535, 2024.
[23] Z. Shen et al., "Beyond summarization: Designing AI support for real-world expository writing tasks," arXiv preprint arXiv:2304.02623, 2023.
[24] L. Huang et al., "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," arXiv preprint, 2023.