Automating Generation of Long-Form Queries

Shivani Upadhyay
sjupadhyay@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Daniel Campos
daniel.campos@zipf.ai
Zipf AI
New York, USA

Nandan Thakur
nandan.thakur@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Ronak Pradeep
rpradeep@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Nick Craswell
nickcr@microsoft.com
Microsoft
Seattle, USA

Jimmy Lin
jimmylin@uwaterloo.ca
University of Waterloo
Waterloo, Canada

Abstract

Traditional short keyword queries are increasingly being replaced by longer, more detailed queries that reflect complex and nuanced user information needs, especially in conversational assistants equipped with web search capabilities. In this work, we present a methodology for automatically generating such human-style long-form queries (narratives) by clustering raw short queries to form synthetic search sessions, designed to reflect a real user's search behavior. Based on our human interpretation study of a 50-narrative set (comprising both human-written and automated narratives), 44% of the automated narratives are misidentified as human-written, underscoring not only the realism and complexity of the generated content but also its indistinguishability from authentic human narratives. Furthermore, we share a collection of automated narratives as a testbed for evaluating LLMs on long-form question answering (QA), which was used in the TREC 2025 RAG track. Our code is available at https://github.com/castorini/narrative-generation.

CCS Concepts

• **Information systems** → **Query representation**.

Keywords

Large Language Model; Long-Form Query Generation; Evaluation.

ACM Reference Format:
Shivani Upadhyay, Daniel Campos, Nandan Thakur, Ronak Pradeep, Nick Craswell, and Jimmy Lin. 2026. Automating Generation of Long-Form Queries. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26), July 20–24, 2026, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3805712.3809917

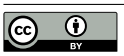
1 Introduction

Large language models (LLMs) have introduced a significant shift in how humans search for information. With their ability to process, reason, and think over extended contexts [12, 18], LLMs now support more complex information-seeking behavior than conventional keyword-based search systems. Instead of short search queries, users now formulate detailed questions that incorporate more context, intent details, and capture multiple aspects of their information need [19, 23]. This evolution reflects the growing expectation that search systems should not only retrieve relevant documents but also reason across them to deliver cohesive answers.

Current evaluation resources fail to capture the linguistic and semantic richness of these complex interactions. Existing benchmarks are largely developed around short, keyword-focused queries, which do not fully match the detailed, naturalistic style of modern search behavior [3, 11, 17]. Moreover, collecting these long and realistic search queries directly from humans is a tedious and resource-intensive process, making it difficult to develop such benchmarks effectively at scale. Without updated evaluation resources, system effectiveness may be overstated, overlooking real-world challenges in handling complex information needs.

In this paper, we introduce a methodology for automatically generating detailed, long-form queries using LLMs, which aims to mimic human-style query formulation. Throughout this paper, we adopt the TREC nomenclature and refer to these two to three sentence long-form queries as *narratives*. In general, long-form queries are multi-sentence, context-rich expressions of an information need that go beyond simple keyword searches. As such, a narrative is a detailed query that represents the search session that a user would have been part of in the pre-LLM era. Figure 1 presents a high level example of our automated generation framework. We cluster raw search queries into synthetic search sessions and use LLMs to generate narratives that reveal the underlying user intent. Using this process, we generated 105 narratives, which served as the test set for the TREC 2025 RAG Track [20].

Our analysis shows that the generated narratives of TREC 2025 RAG exhibit significantly higher linguistic and semantic complexity when compared with short search queries from the TREC 2024 RAG test set, with notable differences in syntactic richness and topical breadth [13]. Furthermore, our results indicate no significant differences in per-sentence structural attributes when comparing RAG 2025 narratives with human-written narratives from the TREC 2006 ciQA track [8]. Moreover, based on our retrieval evaluations, a comparison between short keyword queries and narratives shows that narratives are more difficult for retrieval. Additionally, after applying a style transformation step to rewrite the RAG 2025 narratives in the TREC 2006 ciQA style, we conduct a human interpretation test to assess whether participants could distinguish between automated and human-written narratives. From the interpretation analysis performed on a 50-narrative set, 44% of the transformed



SIGIR '26, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-2599-9/2026/07
https://doi.org/10.1145/3805712.3809917

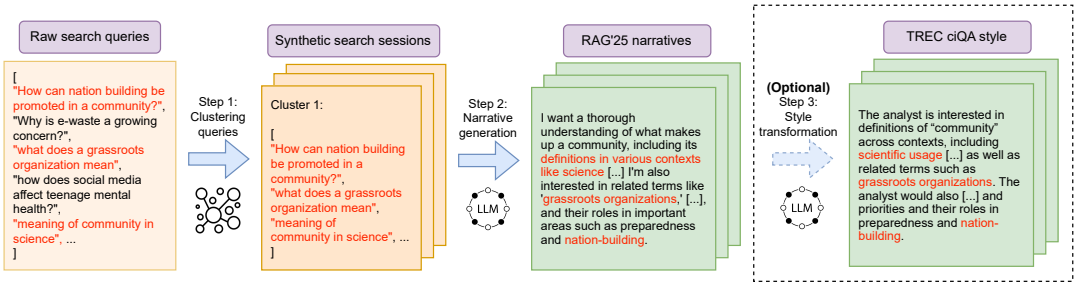


Figure 1: Overview of the automated narrative generation process.

automated narratives are misclassified as human-written, indicating that the generated narratives are often indistinguishable and can effectively mimic realistic search needs. To summarize, our contributions are as follows:

- We propose a novel framework for automatically generating human-style long-form queries using raw short search queries.
- We created a benchmark of 105 narratives for TREC 2025 RAG.
- Our human interpretation analysis shows that 44% of the transformed automated narratives are indistinguishable from human-written ones, showing their realism.

2 Related Work

Recent work increasingly uses LLMs to generate queries that better capture nuanced user intent. Jiang et al. [7] introduce DeepRetrieval, an RL-based framework for query generation and rewriting that improves retrieval effectiveness. Feng et al. [5] propose InteR, which synergistically combines retrieval models and LLMs for iterative query refinement. The InPars Toolkit [1] shows how LLMs can generate diverse, high-quality queries for training and evaluation, and Rahmani et al. [14] examine the feasibility of using LLMs to synthesize entire IR test collections, including relevance judgments. However, these focus primarily on query reformulation or synthesis rather than on generating realistic information needs.

BRIGHT [15] focuses on reasoning-intensive retrieval, drawing queries from human-generated community data. BrowseComp [21] uses an inverted methodology, starting from a verifiable fact and crafting questions hard to find but easy to verify; BrowseComp-Plus [2] extends this with a document corpus built by human verified supporting evidence. FreshStack builds nugget-level IR benchmarks from technical documentation [16]. MoNaCo [22] elicits multi-hop questions from human crowd workers through persona-based prompting, with manually annotated reasoning chains. DeepResearch Bench [4] includes 100 PhD-level tasks authored by domain experts to assess agentic synthesis and citation accuracy. Our approach focuses on automating the generation of human-style narratives that emulate human search behavior, making this a scalable process for creating testbeds for evaluating LLMs on realistic, complex and user-aligned retrieval tasks.

3 Methodology

We begin by collecting a selection of previous years' TREC datasets, including DL23, MSMARCO v2 (dev and dev2) used with DL tracks, and RAG 2024 (raggy-dev and researchy-dev), to ensure diversity and coverage across different query styles and domains [3, 13].

Goal: Synthesize a sequence of user queries from a single, focused session into one comprehensive, first-person query that represents the user's overall informational need, suitable for submission to an LLM like Gemini with search capabilities.
Context: The following list contains related queries issued by a user within a single search session. Assume the user has a coherent goal or topic they are exploring, even if the queries evolve or contain repetitions. Try to make the overall intent diverse such that it can lead to different paths. Maybe include some additional topic outside the questions and relate the overall intent.
Task:
1. **Analyze Intent:** Carefully examine the sequence of main queries and try to connect with sub-queries. Infer the user's underlying goal, the specific information they are seeking, and the overall topic. Notice how the queries might relate (e.g., refining a topic, asking for definitions, comparing items, seeking examples, correcting typos).
2. **Identify Core Need:** Determine the essential information the user is trying to gather by the end of the session. Disregard minor variations or repetitions if they don't change the core intent.
3. **Synthesize:** Combine the inferred intent and core informational needs into a *single, concise, and natural-sounding query*.
4. **Adopt Persona:** Phrase the synthesized query in the first person, as if you *are* the user asking for help (e.g., "I'm trying to understand...", "Can you provide information on...", "I need help finding...").
5. **Ensure Completeness:** The final query should be comprehensive enough to address the main informational goal demonstrated across the *entire* sequence of original queries.
6. **Be Efficient:** Avoid simply listing all original query terms. If a broader concept or question captures the essence, use that. For example, instead of "what is currency in France? what is currency in Germany?", prefer "What currencies are used in France and Germany?".
7: **Do not make Complicated Sentences: Don't over-complicate the sentences by adding too many things. Keep sentences simple so that they are easily understandable. Don't add too many connectors and make too long sentences.
8. **Keep it Concise:** Avoid over-simplifying the intent by adding every available information. Try to keep it concise.

Input Data Format:
–Begin User Queries – Queries: {queries} –End User Queries–

Example:
Main queries: [college football scores, auburn football score, lsu baseball schedule, alabama football schedule, football colleges in georgia]
Synthesized first-person query: "I'm looking for information on college sports, specifically the latest football scores for Auburn, the football schedule for Alabama, the baseball schedule for LSU, and a list of colleges in Georgia that have football teams."

Required Output:
Generate *only* the synthesized first-person query *with natural* flow containing (2-3 short sentences at max) based on the instructions above. Do not include any other explanatory text.

Figure 2: Prompt used to understand the high-level essence of the query list and generate a first-person narrative. The highlighted parts are the styling details.

Step 1: Clustering Search Queries. `BERTopic` [6], a topic modeling approach, is employed to combine transformer-based embeddings with clustering of raw search queries. Each query is first converted

Re-write the following given user narrative. Keep all information present in the given narrative. Just make it more natural and comprehensive. Don't remove any information.

narrative: {previous_response}

Generate *only* the rewritten first-person narrative *with natural* flow containing (2-3 short sentences at max). Do not include any other explanatory text.

Figure 3: Prompt used for rewriting the previously generated narrative. The highlighted parts are the styling details.

into a dense vector representation using the sentence transformer `all-MiniLM-L6-v2`. Then, a customized `HDBSCAN` [9] model with a minimum cluster size of 4 is used to ensure that only sufficiently large groups of queries are retained. To make the clustering more inclusive, the minimum number of samples is set to 1, which lowers the threshold for identifying core points. Stable clusters are then extracted with the Excess of Mass method, and for each cluster, `BERTopic` applies class-based TF-IDF to highlight the most distinctive terms, resulting in the assignment of queries to coherent topics. Finally, clusters are constructed, each containing semantically similar or related short search queries.

Step 2: Narrative Generation. Once the clusters (query lists) are obtained, each cluster or list is treated as a synthetic user session and used to generate corresponding first-person narratives. Figure 2 illustrates the prompt used to produce these narratives from the query lists. To ensure diversity in expression, both GPT-4.1 and Gemini 2.5 Flash are used to produce two distinct versions of the narrative. To mitigate potential biases from any single model, a second-stage rephrasing step is introduced in our framework (for prompt refer to Figure 3). Specifically, narratives initially generated by GPT-4.1 are rephrased by Gemini 2.5 Flash, and vice versa. Usage of the second model in rewriting the original narrative avoids any sort of intrinsic language pattern or behavior that might not be detected during manual inspection. Hence, the rewriting phase results in four variations of each narrative. Table 1 shows these variations for the "community" example. Finally, the best narrative is selected manually based on criteria such as coherence, interpretability, and clarity of the generated narrative. From the "community" example variations shown in Table 1, entry (3) (bold, shaded) represents

Convert the given narrative into TREC narrative style (1-2 sentence long).

TREC narrative example 1: "The analyst is interested in South African arms support to Pakistan and the effect such support or sales has on relations of both countries with India. Additionally, the analyst would like to know what nuclear arms involvement, if any, exists between South Africa and Pakistan."

TREC narrative example 2: "The analyst is interested in evidence of transport of goods from Syria to Iraq under the food-for-oil program."
Instructions: Try to make 1 sentence long narratives but don't over-complicate each sentence by adding too much information in a single sentence. If too much information then form 2 sentence long narrative. Make simpler narratives. If needed drop some info.

Given narrative: {narrative}

Follow the above given instructions properly.

Figure 4: Prompt used for transforming the previously generated TREC narrative to TREC 2006 ciQA track style narrative.

(1) I'm trying to understand what defines a community and why engaging communities in preparedness and nation-building is important. Can you also explain what a grassroots organization is and how the concept of community is viewed in science?
(2) I'm looking for a comprehensive understanding of what constitutes a community. This includes its definition across different contexts, like in science or what defines a 'civilized community,' and the meaning of related terms such as 'grassroots organizations'. I also want to understand how communities function internally, such as in setting boundaries and priorities, and their role in significant efforts like preparedness and nation-building.
(3) I want a thorough understanding of what makes up a community, including its definitions in various contexts like science and what it means to be a 'civilized community.' I'm also interested in related terms like 'grassroots organizations,' how communities set boundaries and priorities, and their roles in important areas such as preparedness and nation-building.
(4) I'm interested in understanding what defines a community and the importance of engaging them in preparedness and nation-building. Could you also explain what a grassroots organization is, and how the concept of community is viewed scientifically?

Table 1: Four variations of the narrative generation process: (1) GPT-4.1, (2) Gemini 2.5 Flash, (3) GPT-4.1 rephrased by Gemini 2.5 Flash, (4) Gemini 2.5 Flash rephrased by GPT-4.1. The bold shaded entry (3) is chosen during selection.

RAG 2025 Narrative: I want a thorough understanding of what makes up a community, including its definitions in various contexts like science and what it means to be a 'civilized community.' I'm also interested in related terms like 'grassroots organizations,' how communities set boundaries and priorities, and their roles in important areas such as preparedness and nation-building.
ciQA Style Narrative: The analyst is interested in definitions of "community" across contexts, including scientific usage and the notion of a "civilized community," as well as related terms such as grassroots organizations. The analyst would also like to know how communities set boundaries and priorities and their roles in preparedness and nation-building.
Human-Written Narrative: The analyst would like to know of efforts made to discourage narco traffickers from using Bonaire as a transit point for drugs to the United States. Specifically, the analyst would like to know of any efforts by local authorities as well as the international community.

Table 2: Examples of narratives written by humans and generated using our approach and its style-transformed version.

the option chosen during the human selection. Section 4 discusses details on delegating this manual selection process to LLMs.

Step 3: (Optional) Style Transformation. An optional step can also be performed to transform the generated narratives into a specific style. Figure 4 presents an example of a style transformation prompt that is used for converting to the TREC 2006 ciQA track [8] using GPT-5. Human narrative examples are included in the prompt to illustrate the desired style. Table 2 showcases examples of narratives: a human-written, an automated narrative in first-person, and its style-transformed counterpart. Section 4.2 presents comparison between two approaches for style transformed narrative generation: applying a post-hoc style transformation after narrative generation,

Feature	Human Narratives	RAG 2024	RAG 2025	Without Rephrasing GPT-4.1	Without Rephrasing G 2.5 Flash
# Samples	25	25	25	25	25
# Tokens	22.24 ± 6.29	8.60 ± 2.47	20.60 ± 3.79	24.56 ± 3.53	16.97 ± 3.25
# Sentences	1.52 ± 0.51	1.00 ± 0.00	2.24 ± 0.43	2.00 ± 0.00	2.72 ± 0.46
# Clauses	0.58 ± 0.70	0.28 ± 0.61	0.63 ± 0.55	0.94 ± 0.58	0.28 ± 0.27
# Concept	7.72 ± 2.88	3.24 ± 1.51	6.83 ± 1.47	8.48 ± 1.81	5.42 ± 1.16
# Connector	2.22 ± 1.25	0.68 ± 0.48	4.69 ± 1.50	5.74 ± 1.43	3.35 ± 0.99
Readability grade	13.29 ± 3.68	8.67 ± 3.91	14.56 ± 2.30	14.75 ± 1.83	13.45 ± 2.78

Table 3: Summary statistics (mean ± std) for query features.

Feature	U	U_p	Cohen's d
# Tokens	270	0.41	−0.32
# Sentences	501	0.00	1.52
# Clauses	342	0.55	0.07
# Concept	251	0.24	−0.39
# Connector	552	0.00	1.79
Readability grade	397	0.10	0.41

Table 4: Mann-Whitney U test for comparing human-written and RAG 2025 narrative features.

and incorporating style constraints directly into the narrative generation prompts for the *analyst* role. The styling details for the later approach are applied during generation by adjusting the prompt styles (highlighted parts) shown in Figures 2 and 3.

4 Experiments

To evaluate narrative complexity, we analyze a set of linguistic and semantic features summarized in Table 3. These include token count (alphabetic tokens), sentence count (from sentence segmentation), clause count (based on dependency labels such as 'ccomp', 'advcl', and 'relcl'), concept count (named entities and noun chunks), and connector count (coordinating/subordinating conjunctions and connective punctuation). Readability is measured using the Flesch-Kincaid grade level via textstat.[1] Except for readability and sentence count, all features are computed at the sentence level to control for length. All features are extracted using spaCy and applied consistently throughout the paper unless otherwise noted.

Table 3 (columns 3 and 4) compares the summary statistics for randomly sampled researchy queries from the RAG 2024 test set [13] and our automated narratives (RAG 2025 test set). A consistent pattern of higher complexity is shown in automated narratives when compared to researchy queries (RAG 2024). Intuitively, automated narratives are longer and span more sentences. They contain many more syntactic elements in each sentence, including clauses (0.63 vs. 0.28), overall concepts (6.83 vs. 3.24), and connectors (4.69 vs. 0.68). Automated narratives' higher readability grade reflects more sophisticated syntactic organization. These findings confirm that automated narratives are not simple extensions of RAG 2024 queries but represent a more complex form of query articulation.

Upon comparing automated narratives with human-written narratives from the TREC 2006 ciQA track [8] (Table 3 columns 2 and 4), features display close structural similarity aside from sentences.[2] Furthermore, the Mann-Whitney U test comparing human-written ciQA narratives with 25 randomly sampled RAG 2025 automated narratives (Table 4) shows no significant differences in most linguistic features, including tokens, clauses, concepts, and readability (all $p > 0.05$). This suggests that RAG 2025 sentences closely match human text in density and structure. The only significant differences are in sentence count and connector usage ($p < 0.05$). The higher sentence count reflects stylistic choices in narrative structure.

Automatic Narrative Selection. The only step in our pipeline that involves human input is narrative selection. This choice is primarily driven by the large-scale evaluation setting of RAG 2025, and manual selection is applied to avoid introducing additional model-based bias or confounding effects into the final evaluation. To assess whether this remaining human step could be automated, we conduct a preliminary experiment in which GPT-5 selects the best narrative from a pool of candidates under low reasoning and verbosity settings, reflecting the lightweight role played by human selectors. We observe approximately 65% agreement between GPT-5 and human selections, despite the inherent subjectivity of narrative preference and the absence of task-specific tuning. These results indicate that manual selection is not a fundamental requirement of the approach, but a conservative, evaluation-driven choice specific to the use of these narratives as a RAG 2025 test, and that selection is amenable to automation in an end-to-end system.

[1] https://pypi.org/project/textstat

[2] human-written TREC 2006 ciQA narratives are usually 1–2 sentences, while RAG 2025's automated ones are designed to be 2–3, so the length difference is expected.

Feature	GPT-4.1			Gemini 2.5 Flash		
	U	U_p	Cohen's d	U	U_p	Cohen's d
# Tokens	417	0.04	0.45	139	0.00	−1.05
# Sentences	462	0.00	1.33	579	0.00	2.47
# Clauses	429	0.02	0.56	244	0.17	−0.56
# Concept	398	0.10	0.32	121	0.00	−1.04
# Connector	599	0.00	2.62	467	0.00	1.00
Readability grade	412	0.05	0.50	331	0.73	0.05

Table 5: Mann-Whitney U test comparing human-written narratives from TREC 2006 ciQA [8] with narratives generated by a single model (without rephrasing step) features.

4.1 Multi vs. Single Model Narratives

In Table 3, GPT-4.1 (Column 5) and G 2.5 Flash (Column 6; using Gemini 2.5 Flash) show the features of the single model outputs before rephrasing. Notably, the RAG 2025 narratives (column 4) strike a balanced middle ground between the single model generations. This suggests that narratives obtained after the rephrasing step effectively aggregate the strengths of both models while smoothing out extremes. These trends are further supported by the Mann-Whitney U tests with human narratives in Tables 4 and 5, which show that single model narratives tend to deviate more strongly and unevenly across structural features. In contrast, the RAG 2025 narratives, which are prepared with multiple models, exhibit more moderate differences and fewer pronounced effects, indicating that rephrasing reduces model-specific stylistic biases. Overall, this highlights clear qualitative improvements in the generated narratives when multiple models are combined, resulting in outputs that more closely resemble human narrative structure and style.

Feature	Style Transform			Style Integration		
	U	U_p	Cohen's d	U	U_p	Cohen's d
# Tokens	225	0.09	−0.34	270	0.42	−0.14
# Sentences	200	0.01	−0.84	187	0.00	−0.97
# Clauses	347	0.48	0.26	421	0.02	0.75
# Concept	296	0.76	0.06	305	0.90	−0.04
# Connector	40	0.00	−2.00	89	0.00	−1.24
Readability grade	171	0.01	−0.80	135	0.00	−1.09

Table 6: Mann-Whitney U test comparing human-written narratives from TREC 2006 ciQA [8] with style transformed and prompt-modified narrative features.

Technique	RAG 2024			RAG 2025		
	Recall@10	MAP	nDCG@10	Recall@10	MAP	nDCG@10
BM25	0.0323	0.1539	0.3361	0.0169	0.1235	0.3613
BM25 + Rocchio	0.0347	0.1450	0.3738	0.0168	0.0712	0.3603
Arctic-Embed-l	0.0591	0.2514	0.6028	0.0264	0.2227	0.5303

Table 7: Retrieval metrics: RAG 2025 vs. RAG 2024 topics.

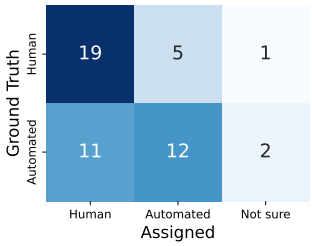


Figure 5: Confusion matrix comparing human annotator and ground-truth labels using the TREC 2006 ciQA track dataset.

4.2 Style Transform vs. Style Integration

Table 6 summarizes the Mann-Whitney U test comparing human-written narratives with those produced via style transformation and prompt-based style integration. Narratives with style transform show smaller deviations from human-written narratives for most features. Token count differences are modest, and sentence count shows a statistically significant but moderate effect. Clause count differences are minimal, and readability grade also deviates moderately. In contrast, prompt style integration introduces larger deviations, particularly for sentence count, clause count, and readability grade. These indicate stronger structural and complexity shifts compared to human narratives. For concept count, both methods closely approximate human narratives, suggesting semantic content remains largely unaffected. However, connector usage differs substantially, and prompt style integration performs better.

Furthermore, incorporating a dedicated style transform step improves adherence to the specified style, leading to more accurate and consistently formatted outputs. This separation allows for more precise alignment with prescribed stylistic conventions, resulting in narratives that are both structurally coherent and stylistically consistent. Overall, the style transform approach not only better approximates human sentence-level complexity and readability but also improves stylistic and formatting accuracy.

4.3 Human Interpretation Analysis

To further examine how automated narratives are interpreted by humans, we compare them with human-written narratives from the TREC 2006 ciQA track [8]. We compile a set of 50 narratives, consisting of 25 written by humans and randomly sample 25 TREC ciQA style transformed from our automated narratives set.

Three human annotators were recruited and tasked with distinguishing between the two types of narratives. Annotators were instructed to classify each narrative as either *human* or *automated* to the best of their ability. If they were unable to make a clear determination, they were advised to select the *'Not Sure'* option. The labels assigned by three annotators were combined based on majority voting, and the results of this annotation process are summarized in the confusion matrix shown in Figure 5. These results show that, although assessors could often distinguish narrative origins, they frequently misclassified automated narratives as human (≈44%). Based on the human assessor's reasoning, many misclassified automated queries appeared human because of their coherent structure, purposeful phrasing, and organized flow. χ^2 test on the confusion matrix (Figure 5) yields $\chi^2 = 5.35$ with $p = 0.07$, slightly above the 0.05 threshold. This indicates no statistically significant difference in narrative distribution, though the marginal p-value indicates a weak trend rather than a strong effect.

4.4 Retrieval Comparison

Table 7 compares retrieval effectiveness for RAG 2025 and RAG 2024 test sets using the MS MARCO V2.1 corpus. Evaluation was performed using manual assessments created by NIST assessors. Scores decline across all three methods from 2024 to 2025, with consistent drops in Recall@10 and MAP. This suggests that the longer, more expressive narratives in RAG 2025 introduce greater semantic complexity, making it harder for retrieval models to capture all relevant aspects of the information need. While BM25 exhibits a slight increase in nDCG@10 for RAG 2025, this improvement occurs alongside a substantial drop in recall, indicating that although a small number of highly relevant documents may be ranked near the top, overall coverage of relevant documents is reduced. Snowflake's dense model Arctic-Embed-l [10] remains the strongest performer across both years, yet it also experiences notable declines in recall and MAP for RAG 2025, showing that even semantic embedding approaches are challenged by our multi-sentence narratives. Overall, these results support the interpretation that RAG 2024 topics are inherently easier to retrieve against than the more complex and nuanced narratives generated by our approach for RAG 2025.

5 Conclusion

This work introduces a methodology for curating long-form narratives by clustering raw search queries to form synthetic search sessions using LLMs. As part of this approach, we generated a set of narratives to exemplify the technique. Our analysis shows that these narratives exhibit considerable complexity, both linguistically and semantically. Furthermore, when evaluated against human-written narratives, the interpretations provided by human assessors indicate that the automated narratives are comparable in quality and structure to some extent with those written by humans. This suggests that such narratives have strong potential to enhance IR systems by simulating realistic search needs.

Acknowledgments

The authors thank Dhara Mehta, Joshua Green, and Shreyas Upadhyay for their valuable assistance with the human interpretation assessment. They helped out with the manual evaluation performed to compare human-written and automatically generated narratives.

References

[1] Hugo Abonizio, Luiz Bonifacio, Vitor Jeronymo, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. InPars Toolkit: A Unified and Reproducible Synthetic Data Generation Pipeline for Neural Information Retrieval. *arXiv:2307.04601* (2023).

[2] Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. 2025. BrowseComp-Plus: A More Fair and Transparent Evaluation Benchmark of Deep-Research Agent. *arXiv:2508.06600* (2025).

[3] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2024. Overview of the TREC 2023 Deep Learning Track. In *Text REtrieval Conference (TREC)*. NIST, TREC.

[4] Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv:2506.11763* (2025).

[5] Jiazhan Feng, Chongyang Tao, Xiubo Geng, Tao Shen, Can Xu, Guodong Long, Dongyan Zhao, and Daxin Jiang. 2024. Synergistic Interplay between Search and Large Language Models for Information Retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9571–9583.

[6] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794* (2022).

[7] Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025. DeepRetrieval: Hacking Real Search Engines and Retrievers with Large Language Models via Reinforcement Learning. *arXiv:2503.00223* (2025).

[8] Diane Kelly and Jimmy Lin. 2007. Overview of the TREC 2006 ciQA Task. *SIGIR Forum* 41, 1 (June 2007), 107–116.

[9] Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical Density Based Clustering. *Journal of Open Source Software* 2, 11 (2017), 205.

[10] Luke Merrick, Danmei Xu, Gaurav Nuti, and Daniel Campos. 2024. Arctic-Embed: Scalable, Efficient, and Accurate Text Embedding Models. *arXiv:2405.05374* (2024).

[11] Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, Andreas Vlachos and Isabelle Augenstein (Eds.). Dubrovnik, Croatia, 2014–2037.

[12] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, et al. 2024. GPT-4 Technical Report. *arXiv:2303.08774* (2024).

[13] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented Generation Track. *arXiv:2406.16828* (2024).

[14] Hossein A. Rahmani, Nick Craswell, Emine Yilmaz, Bhaskar Mitra, and Daniel Campos. 2024. Synthetic Test Collections for Retrieval Evaluation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Washington DC, USA) *(SIGIR '24)*. Association for Computing Machinery, New York, NY, USA, 2647–2651.

[15] Hongjin Su, Howard Yen, Mengzhou Xia, Weijia Shi, Niklas Muennighoff, Han-yu Wang, Haisu Liu, Quan Shi, Zachary S. Siegel, Michael Tang, Ruoxi Sun, Jinsung Yoon, Sercan O. Arik, Danqi Chen, and Tao Yu. 2025. BRIGHT: A Realistic and Challenging Benchmark for Reasoning-Intensive Retrieval. *arXiv:2407.12883* (2025).

[16] Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. 2025. FreshStack: Building Realistic Benchmarks for Evaluating Retrieval on Technical Documents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[17] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[18] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv:2307.09288* (2023).

[19] Johanne R. Trippas, Damiano Spina, Lawrence Cavedon, and Mark Sanderson. 2017. How Do People Interact in Conversational Speech-Only Search Tasks: A Preliminary Analysis. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval* (Oslo, Norway) *(CHIIR '17)*. Association for Computing Machinery, New York, NY, USA, 325–328.

[20] Shivani Upadhyay, Nandan Thakur, Ronak Pradeep, Nick Craswell, Daniel Campos, and Jimmy Lin. 2026. Overview of the TREC 2025 Retrieval Augmented Generation (RAG) Track. *arXiv:2603.09891* (2026).

[21] Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. *arXiv:2504.12516* (2025).

[22] Tomer Wolfson, Harsh Trivedi, Mor Geva, Yoav Goldberg, Dan Roth, Tushar Khot, Ashish Sabharwal, and Reut Tsarfaty. 2025. MoNaCo: More Natural and Complex Questions for Reasoning Across Dozens of Documents. *arXiv:2508.11133* (2025).

[23] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2023. Conversational Information Seeking. *arXiv:2201.08808* (2023).