Beyond the Answer: A Narrative-Driven RAG Framework for Biomedical Knowledge Synthesis

Anonymous ACL submission

Abstract

The synthesis of complex biomedical information is a critical challenge, with traditional search and question-answering systems often providing fragmented, decontextualized results. To address the limitations of conventional *answer-driven* Retrieval-Augmented Generation (RAG), we propose a paradigm shift to a *narrative-driven* approach. Our framework decomposes a user's query into thematic aspects, synthesizes retrieved evidence into a coherent narrative, and presents the output as a visually structured poster. On established datasets and our new Biomedical Narrative Poster Benchmark (BNPB), our method demonstrates superior performance. It enhances information coverage by up to 51% in ROUGE-1 Recall and, crucially, more than doubles the scores for narrative quality compared to a traditional RAG baseline. This work introduces the first query-driven poster generation framework, offering a more holistic and effective method for understanding complex biomedical knowledge.

1 Introduction

Synthesizing the ever-expanding volume of biomedical literature into a coherent understanding is a significant challenge for practitioners and researchers alike (Upadhyay and Viviani). For instance, a clinical question such as, *"What is the effect of TRH on myocardial contractility?"* requires integrating information across its background, mechanisms, applications, and challenges. Traditional search engines often provide fragmented, decontextualized answers, creating a barrier to effective decision-making and knowledge acquisition.

Retrieval-Augmented Generation (RAG) has emerged to mitigate the limitations of Large Language Models (LLMs), such as factual inaccuracies and hallucinations, by grounding them in external knowledge sources like medical textbooks and scientific literature (Li et al., 2024b; Wang et al., 2025; Li et al., 2024c). However, the prevailing RAG paradigm is predominantly *answer-driven*, generating concise but unstructured answers that fall short when a holistic understanding of a complex topic is required. Recognizing this limitation, recent research has trended towards more structured information synthesis through multi-step retrieval or knowledge graph (KG) integration (Li et al., 2025; Wu et al.; Zhao et al.).

In this paper, we argue for a fundamental shift from an answer-driven to a *narrative-driven* approach. Our framework constructs a comprehensive narrative by decomposing a high-level query into four fundamental aspects: *Introduction*, *Mechanism*, *Application*, and *Challenges*. By systematically retrieving and synthesizing evidence for each aspect, our method moves beyond simple knowledge recall to promote a deeper level of understanding.

To present this structured narrative in an accessible and digestible format, we propose the first framework for query-driven, automated poster generation. While prior work has focused on generating posters from full research papers (Sun et al., 2025; Pang et al., 2025), our query-driven approach offers greater flexibility for on-demand information synthesis. The final output is rendered as a structured PDF, aligning with the cognitive workflows of medical professionals who need to quickly grasp the key facets of a topic.

Our main contributions can be summarized as follows:

- To the best of our knowledge, we propose the first **query-driven poster generation**

framework, offering a more practical and on-demand alternative to existing paper-driven approaches.

- We introduce a ***narrative-driven RAG*** framework, which departs from traditional answer-driven methods by structuring information into a coherent narrative to enrich content and foster comprehensive understanding.
- We construct and release the **Biomedical Narrative Poster Benchmark (BNPB)**, a new dataset specifically designed to evaluate and advance research on this novel task.

2 Related Work

2.1 Retrieval-Augmented Generation (RAG)

The RAG framework enhances LLMs by retrieving relevant information from external knowledge bases to augment the context, thereby generating more accurate and reliable responses. The value of RAG is particularly pronounced in high-stakes fields like biomedicine, where accuracy and timeliness are critical, though it also faces unique challenges (Thapa et al., 2025).

A core challenge is the quality and selection of the knowledge source. In the biomedical domain, researchers have explored various high-quality sources to ensure authority, such as authoritative medical textbooks (Wang et al., 2025), structured Knowledge Graphs (KGs) for their reasoning capabilities (Li et al., 2024a; Wu et al.; Zhao et al.), and scientific literature via search engines for up-to-date information (Shi et al.; Pronesti et al., 2025). Our work adopts such a multi-source strategy, integrating knowledge from PubMed, web searches, and textbooks.

Another key challenge is the optimization of the retrieval process. Advanced strategies include Adaptive Retrieval to determine when to retrieve (Baek et al., 2025), Query Decomposition to break down complex questions (Li et al., 2025), and Post-retrieval Re-ranking to filter and prioritize the most relevant context using the LLM's own semantic understanding (Yu et al., 2024). Our framework is built upon these ideas, using query decomposition to structure the narrative and a re-ranker to optimize retrieval quality.

2.2 Automated Poster Generation

Academic posters are a vital medium for scholarly communication, but their manual creation is time-consuming. Consequently, automated academic poster generation has become a valuable research direction. Most existing works follow a "paper-to-poster" paradigm, where a complete academic paper is input into a multi-agent system or pipeline that automatically extracts key elements (e.g., abstract, figures, conclusions) and organizes them into a poster (Sun et al., 2025; Pang et al., 2025).

Our work shares the same ultimate goal but differs in its core paradigm. We propose a "query-driven" framework rather than a "paper-driven" one. This means our system is designed to generate a structured, narrative poster that answers a specific user question by retrieving multi-source information. We believe this model offers greater utility and specificity, particularly in helping researchers quickly understand scientific topics or assisting clinicians with decision-making.

3 Methodology

In this section, we present our novel Narrative-Driven RAG framework, designed to generate structured biomedical posters from a single user query, as illustrated in Figure 1.

3.1 Query Processing

The Query Processing stage transforms a user's query into a structured plan for narrative synthesis through two key steps.

Narrative Decomposition and Entity Normalization. An LLM deconstructs the initial query into four thematic subqueries corresponding to our poster's sections: *Introduction*, *Mechanism*, *Application*, and *Challenges*. This forms a narrative scaffold to guide information gathering. Concurrently, the LLM identifies key biomedical entities in the query and normalizes them to the Medical Subject Headings (MeSH) thesaurus, a crucial step for improving retrieval precision from scientific databases.

Source-Specific Query Formulation. The subqueries and normalized entities are used to generate tailored search terms. For literature retrieval from PubMed and web search via Serper, the LLM formulates precise search strings, one for each narrative aspect. For our Textbook corpus, the LLM instead maps each subquery to the most relevant thematic category (e.g., "Pharmacology") to direct retrieval to appropriate textbooks.

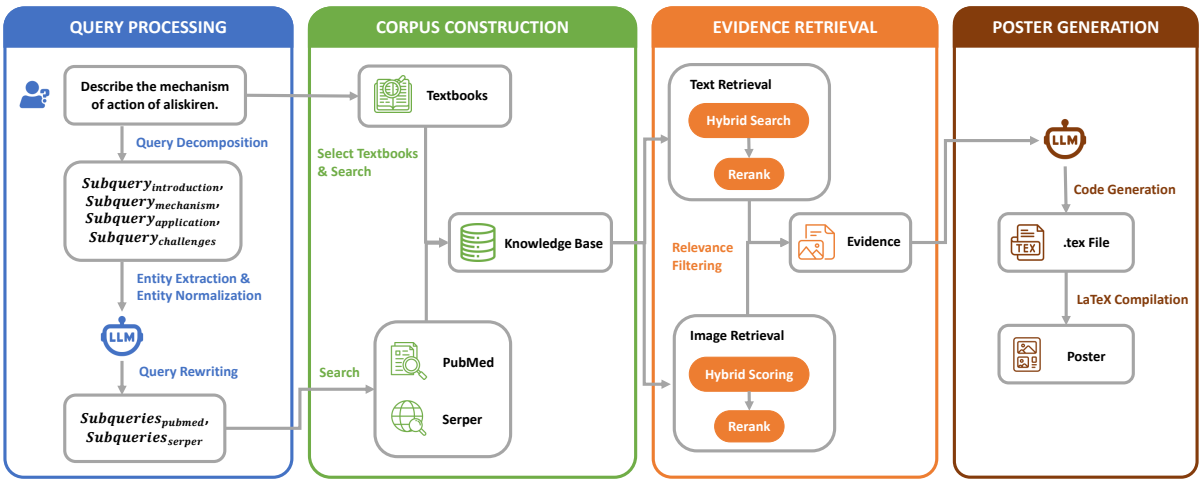

Figure 1: An overview of our narrative-driven RAG framework. The process transforms a user query into a structured poster through four stages: Query Processing, Corpus Construction, Evidence Retrieval, and Poster Generation.

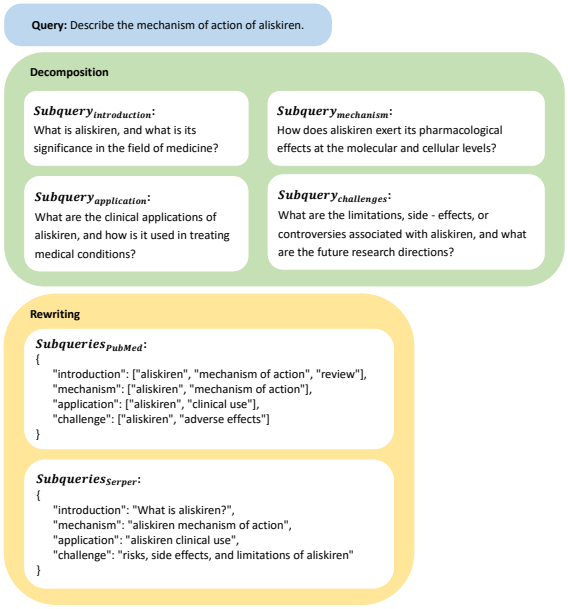

Figure 2: An example of the Query Processing stage.

3.2 Corpus Construction

To provide comprehensive evidence for our narrative framework, we construct a query-specific, ad-hoc corpus by amalgamating information from three diverse sources. This multi-source approach ensures that our generated narrative is grounded in a rich and balanced knowledge base, encompassing cutting-edge research, broader context, and foundational concepts.

- **Academic Literature (PubMed):** To source the most recent scientific evidence, we leverage the normalized MeSH entities generated during query processing. We query the PubMed database for relevant, open-access publications from the last five years. The full-text PDFs of the retrieved articles are then downloaded and parsed using MinerU [1], a tool designed for structured text extraction from scholarly documents.

- **Web Documents (Serper):** To capture broader context and information not yet published in academic literature, we perform a general web search using the Serper API. For this source, we intentionally use the original, non-normalized query entities to better align with common language used online. The system incorporates special handling for queries containing negations to improve retrieval relevance. Relevant PDFs found through this search are also parsed via MinerU.

- **Textbooks (Foundational Knowledge):** To ground the narrative in established medical knowledge, we incorporate a corpus of medical textbooks. This component is sourced directly from the MedQA dataset (Jin et al., 2021), which provides pre-processed plain text files. As this data is already in a clean format, no further downloading or parsing is necessary.

The textual content extracted from these three sources is unified to form a comprehensive, query-relevant corpus. This ad-hoc knowledge base

[1] https://github.com/opendatalab/MinerU

serves as the foundation for the subsequent retrieval stage described in the next section.

3.3 Evidence Retrieval

With the ad-hoc corpus constructed, the retrieval stage aims to identify the most salient textual and visual evidence for each of the four narrative subqueries. To ensure high relevance and precision, we execute a targeted, multi-stage retrieval process independently for each narrative aspect (*Introduction*, *Mechanism*, etc.). This process involves two parallel pipelines for textual and visual evidence respectively.

Textual Evidence Retrieval. Our text retrieval pipeline is a three-stage funnel designed to progressively refine the set of relevant passages.

1. **Hybrid Search:** For each subquery, we first perform a hybrid search on the corpus. This approach combines the lexical matching strength of the sparse retriever BM25 with the semantic understanding of a dense passage retriever (DPR), maximizing the recall of potentially relevant text chunks.

2. **Re-ranking:** The initial set of retrieved candidates is then passed to a dedicated re-ranking model to perform a more fine-grained relevance assessment, sorting the candidates and retaining only the top-32 most promising passages.

3. **LLM-based Filtering:** As a final step, we leverage an LLM to perform a definitive relevance check. The LLM is prompted to analyze the top passages in the context of the subquery's specific intent, filtering out any content that is only tangentially related and ensuring that the final evidence is highly pertinent.

Visual Evidence Retrieval. To enrich the poster with meaningful illustrations, we retrieve relevant images (including figures and tables) from the parsed documents. This is achieved through a two-stage process.

1. **Hybrid Scoring:** We first rank all available images using a hybrid score, calculated as the average of textual similarity (between the subquery and the image caption) and cross-modal similarity (between the subquery and the image content, computed via CLIP (Radford et al., 2021)). The top-10 images based on this score are selected.

2. **VLM-based Filtering:** These 10 candidates are then evaluated by a Vision-Language Model (VLM). The VLM performs a final, context-aware assessment to verify that the image content directly and clearly illustrates a concept relevant to the subquery, ensuring the final selection is both pertinent and high-quality.

The output of this stage is a structured evidence package, where each subquery is populated with a curated list of text passages and images, ready for the final synthesis in the Poster Generation module.

3.4 Poster Generation

The final stage of our framework synthesizes the curated evidence into a structured, coherent poster. This process is orchestrated by an LLM, which acts as a domain-aware author to transform the raw evidence into a flowing narrative.

The generation is performed on a section-by-section basis. For each narrative aspect (*Introduction*, *Mechanism*, etc.), the LLM receives the corresponding evidence package. It is prompted to generate a descriptive and logically structured paragraph that summarizes the key information contained within the provided evidence. After generating the content for all four foundational sections, they are programmatically assembled into the final poster layout.

3.5 Generating Future Directions

Beyond summarizing existing knowledge, we introduce a novel module to generate a speculative *Future Directions* section. This component leverages the creative capacity of LLMs to explore potential research gaps in a controlled manner.

Our approach is a two-stage generate-then-evaluate pipeline. First, we prompt an LLM with an increased sampling temperature to generate a diverse set of initial hypotheses on unexplored topics. Second, these hypotheses are fed into a evaluative prompt, which instructs the LLM to act as a cautious scientist. In this stage, it critically assesses the feasibility and relevance of the initial ideas, filtering for those grounded in scientific plausibility. This self-correction process allows us to turn potential model hallucination into controlled, valuable innovation.

4 Experimental Setup

4.1 Datasets

To comprehensively evaluate our framework, we use three distinct datasets: BioASQ-QA (Krithara et al., 2023) for expert-level queries, MedQuAD (Ben Abacha and Demner-Fushman, 2019) for public health questions, and our self-constructed Biomedical Narrative Poster Benchmark (BNPB) to assess performance across diverse biomedical topics.

BioASQ-QA We utilize the summary-type questions from the BioASQ-QA track (1130 QA pairs). These questions emulate complex information needs from biomedical experts, necessitating multi-document synthesis and structured narrative generation.

MedQuAD To assess performance on public health inquiries, we use the MedlinePlus Health Topics subset of MedQuAD, selecting the first 200 of its 981 QA pairs. Its question structure, centered on topics like 'symptoms', 'causes', and 'treatments', reflects common user concerns.

Biomedical Narrative Poster Benchmark (BNPB) We introduce the BNPB, an LLM-generated question set designed to test our framework's generalization. The benchmark was generated via a structured pipeline:

1. **Category Definition:** We defined three broad categories: *Disease/Condition-Centric*, *Technology/Methodology-Centric*, and *Biological Process/System-Centric*.

2. **Topic Identification & Question Generation:** An LLM identified 10 topics per category and generated 5 questions per topic (150 total).

3. **Automated Filtering:** An LLM filtered questions for correctness, clarity, scope, and retrievability.

This process retained 149 high-quality questions, which constitute the final BNPB dataset.

4.2 Baselines

We evaluate our proposed **Narrative-Driven RAG (ND-RAG)** framework against two baselines designed to systematically ablate its core components:

- **LLM-Only:** Utilizes the large language model's parametric knowledge without any external retrieval. To ensure a fair structural comparison, it is prompted to generate the same four sections (*Introduction*, *Mechanism*, etc.) as our method. This baseline quantifies the performance gains attributable to the retrieval-augmentation component itself.

- **Answer-Driven RAG (AD-RAG):** A standard RAG framework that retrieves and synthesizes information based on the original, undecomposed user query. This is our primary baseline, as it directly isolates and measures the benefits of our core contribution: the narrative-driven query decomposition strategy.

4.3 Metrics

We evaluate the generated posters from three key perspectives: the factual correctness of the text, its narrative quality, and the overall aesthetic appeal of the visual presentation.

Correctness. To measure the factual overlap with ground-truth summaries, we employ ROUGE recall scores: ROUGE-1, ROUGE-2, and ROUGE-L (Lin). This set of metrics quantifies the coverage of essential n-grams from the reference text within our generated content, thereby assessing its information fidelity.

Narrative Quality. Assessing the high-level structure of a narrative requires moving beyond n-gram overlap. We, therefore, employ a LLM as an automated judge to evaluate the poster's textual narrative. The LLM provides scores from 1 to 5 on the following three dimensions:

- **Logical Coherence:** Evaluates the logical flow and consistency of the text, both within and between sections.

- **Structural Adherence:** Measures the topical alignment of content under its corresponding section heading (e.g., 'Mechanism').

- **Narrative Completeness:** Assesses whether the sections collectively form a comprehensive and holistic story that addresses the user's query.

Table 1: Evaluation of information coverage on the BioASQ and MedQuAD datasets. Our ND-RAG framework demonstrates significantly higher ROUGE Recall scores across all metrics compared to the AD-RAG baseline, indicating more comprehensive content generation.

Method	**ROUGE**	**BioASQ**			**MedQuAD**		
		Recall	Precision	F1-score	Recall	Precision	F1-score
AD-RAG	ROUGE-1	0.3447	0.1177	0.1579	0.3390	0.3443	0.3333
	ROUGE-2	0.1184	0.0351	0.0489	0.0733	0.0756	0.0727
	ROUGE-L	0.2500	0.0770	0.1064	0.1769	0.1780	0.1727
ND-RAG	ROUGE-1	0.4558	0.0828	0.1234	0.5119	0.1634	0.2419
	ROUGE-2	0.1556	0.0228	0.0355	0.1077	0.0333	0.0497
	ROUGE-L	0.3354	0.0539	0.0824	0.2725	0.0850	0.1265

Aesthetic Quality. Beyond the textual content, the visual presentation is a critical component of a poster. To assess this, we again utilize an LLM-as-a-judge to score the aesthetic quality of the final output on a scale of 1 to 5 across the following three dimensions:

- **Visual Attractiveness:** Assesses the overall visual appeal of the poster and its ability to capture and maintain the reader's interest.
- **Layout & Formatting:** Evaluates the professionalism, organization, and aesthetic coherence of the poster's layout and typography.
- **Informational Clarity:** Measures how effectively the visual design enhances the readability and comprehension of the information presented.

4.4 Implementation Details

Our framework utilizes the `doubao-1.5pro-32k` model for language tasks and `doubao-1.5vision-pro-32k` for vision tasks. All source texts were segmented into 256-token chunks with a 50-token overlap. The text retrieval pipeline employs a hybrid search combining BM25 with dense vectors from `BioLORD-2023`, followed by a `bge-reranker-v2-m3` model that selects the top 32 passages. For image retrieval, we rank candidates by a composite score of title similarity (using `BioLORD-2023`) and a CLIP score (from `clip-vit-large-patch14`), retaining the top 10 for final VLM-based filtering. The poster generation process is handled by the LLM with a default `temperature` of 0.3 and a `max_tokens` limit of 1000. To foster creativity in the "Future Directions" section, the `temperature` was increased to 0.7. ROUGE scores were calculated using the Python `rouge_score` library.

5 Results and Analysis

5.1 Information Coverage

To assess the extent to which our generated posters cover the essential information present in gold-standard summaries, we employ the ROUGE metric, focusing on Recall. Table 1 summarizes the results on the BioASQ and MedQuAD datasets.

The results clearly indicate that our ND-RAG framework consistently outperforms the AD-RAG baseline in ROUGE-1, ROUGE-2, and ROUGE-L recall across both datasets. This demonstrates the effectiveness of our query decomposition strategy in retrieving a wider range of relevant information.

We argue that ROUGE-2 Recall is a particularly salient metric in the biomedical domain, where critical concepts are often represented by bigrams (e.g., "immune response," "spike protein"). The notable improvement in ROUGE-2 Recall (e.g., from 0.1184 to 0.1556 on BioASQ) suggests that our framework is more adept at capturing and reproducing key biomedical concepts and their relationships.

As anticipated, the Precision scores for ND-RAG are lower than for AD-RAG. This is an expected outcome and, in fact, reflects a core design choice of our framework. Our goal is to generate a comprehensive, structured narrative that is inherently much longer and more detailed than the concise gold-standard answers. Therefore, a lower precision is a direct consequence of this enriched information content. Similarly, the modest ROUGE-L scores are expected, as our framework intentionally restructures information into a new narrative

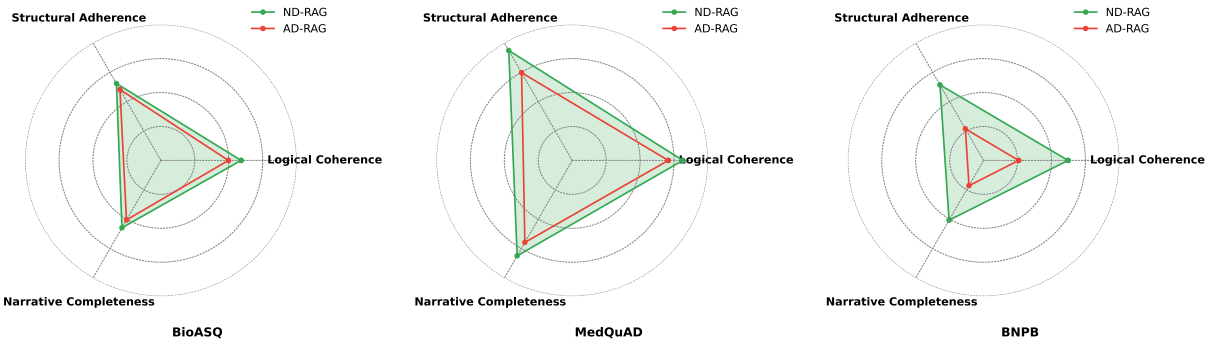

Figure 3: Evaluation of narrative quality using an LLM-as-a-judge. Our Narrative-Driven RAG (ND-RAG) consistently outperforms the Answer-Driven RAG (AD-RAG) baseline across all three metrics and datasets. The performance improvement is particularly pronounced on our custom BNPB benchmark (c).

Table 2: Aesthetic scores (1-5 scale) for generated posters on a random sample of 50 queries from the BioASQ dataset. Scores were assigned by an LLM-as-a-judge. Best results among RAG methods are shown in bold.

Method	Visual Attractiveness	Layout & Formatting	Informational Clarity
LLM-Only	1.7000	2.8000	2.8600
AD-RAG	1.1200	1.8800	1.9600
ND-RAG	**1.7200**	**2.8000**	**2.8800**

flow (Introduction, Mechanism, etc.) rather than preserving the sentence structure of the source documents.

5.2 Narrative Quality

Beyond factual coverage, we evaluated the core contribution of our work: the ability to generate high-quality narratives. Using an LLM-as-a-Judge approach, we scored the outputs on three dimensions: Logical Coherence, Structural Conformance, and Narrative Completeness. The results, presented in Figure 3, span all three datasets.

The performance gap is particularly pronounced on our custom BNPB dataset, which was specifically designed to evaluate the generation of structured biomedical narratives. On BNPB, ND-RAG more than doubles the scores of AD-RAG in all categories. This strongly validates that our narrative-driven approach is not merely retrieving more information, but is critically more effective at synthesizing that information into a coherent, well-structured, and comprehensive story that aligns with the intended narrative framework.

5.3 Aesthetic Quality

We evaluated the quality of the final visual artifact—the poster. We randomly sampled 50 outputs from the BioASQ dataset and used an LLM judge to score them on Visual Attractiveness, Layout & Formatting, and Informational Clarity.

As shown in Table 2, the posters generated by ND-RAG are rated significantly higher than those from AD-RAG. Notably, our ND-RAG framework achieves scores comparable to the LLM-Only baseline. This suggests that the superior internal structure of the narrative-driven content allows the generation module to produce a more professional, organized, and visually appealing layout. The AD-RAG's less coherent content likely leads to poorer layout choices, resulting in lower scores for both aesthetics and clarity. This finding highlights an important synergy: a well-structured narrative not only improves the content's quality but also enhances its final visual presentation.

5.4 Qualitative Analysis

To provide a concrete illustration of our method's advantages, Figure 4 presents a side-by-side qualitative comparison for several biomedical queries.

The LLM-Only model generates text without grounding or visual aids, which are critical for understanding complex topics like the molecular pathways of LSD1. The Answer-Driven RAG (AD-RAG) improves upon this by grounding its output in literature, but its "answer-focused" objective leads to dense, text-heavy responses. While it may

Figure 4: A side-by-side comparison illustrating the advantages of our Narrative-Driven approach. While the LLM-Only model produces text-only outputs and AD-RAG yields text-heavy responses with minimal visuals, our ND-RAG framework synthesizes a rich visual narrative by seamlessly integrating text with multiple, contextually relevant diagrams and tables.

include an image, the visual is often isolated and not integrated into a broader narrative.

In contrast, our Narrative-Driven RAG (ND-RAG) is explicitly designed to build a comprehensive narrative. By decomposing the query into sub-topics (e.g., Introduction, Mechanism), it creates a logical flow and identifies opportunities for meaningful visual integration. For the "U1 small nuclear RNA" query, our method seamlessly incorporates multiple diagrams illustrating cellular functions alongside explanatory text. This synthesis of structured text and relevant visuals results in a far more informative and accessible "visual abstract," substantiating our claim that a narrative-driven approach yields more useful outputs for scientific understanding.

6 Conclusion

In this paper, we challenged the conventional *answer-driven* RAG paradigm by introducing a novel *narrative-driven* framework that synthesizes complex biomedical queries into structured posters. Our experiments conclusively show that this approach not only enhances information coverage, evidenced by higher ROUGE-Recall scores, but more importantly, delivers substantial gains in narrative coherence, structural integrity, and final aesthetic quality. By contributing the first query-driven poster generation system and the new BNPB benchmark, our work paves the way for more holistic and effective tools for biomedical knowledge synthesis. Future efforts should focus on human-centric validation to further refine the framework's real-world applicability.

7 Limitations

Our work is subject to several limitations. First, our Biomedical Narrative Poster Benchmark (BNPB) is LLM-generated and lacks validation from human domain experts. Second, our reliance on an LLM-as-a-judge for evaluating narrative and aesthetic quality, while scalable, may introduce biases and is not a substitute for human assessment. Third, our query decomposition approach can lead to "query drift," where the generated narrative, though comprehensive, may not provide a sufficiently direct answer to the original question. Finally, the speculative "Future Directions" module, despite safeguards, inherently risks generating scientifically implausible hypotheses, blurring the line between controlled innovation and hallucination.

References

Ingeol Baek, Hwan Chang, Byeongjeong Kim, Jimin Lee, and Hwanhee Lee. 2025. Probing-RAG: Self-Probing to Guide Language Models in Selective Document Retrieval. *arXiv preprint*. ArXiv:2410.13339 [cs].

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1):511.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences*, 11(14):6421.

Anastasia Krithara, Anastasios Nentidis, Konstantinos Bougiatiotis, and Georgios Paliouras. 2023. BioASQ-QA: A manually curated corpus for Biomedical Question Answering. *Scientific Data*, 10(1):170.

Dawei Li, Shu Yang, Zhen Tan, Jae Young Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, Huan Liu, Li Shen, and Tianlong Chen. 2024a. DALK: Dynamic Co-Augmentation of LLMs and KG to answer Alzheimer's Disease Questions with Scientific Literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2187–2205, Miami, Florida, USA. Association for Computational Linguistics.

Jiarui Li, Ye Yuan, and Zehua Zhang. 2024b. Enhancing LLM Factual Accuracy with RAG to Counter Hallucinations: A Case Study on Domain-Specific Queries in Private Knowledge-Bases. *arXiv preprint*. ArXiv:2403.10446 [cs].

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Pengjun Xie, Philip S Yu, Fei Huang, and Jingren Zhou. 2025. BENCHMARKING MULTIMODAL RETRIEVAL AUG- MENTED GENERATION WITH DYNAMIC VQA DATASET AND SELF-ADAPTIVE PLANNING AGENT.

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024c. Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach. *arXiv preprint*. ArXiv:2407.16833 [cs].

Chin-Yew Lin. ROUGE: A Package for Automatic Evaluation of Summaries.

Wei Pang, Kevin Qinghong Lin, Xiangru Jian, Xi He, and Philip Torr. 2025. Paper2Poster: Towards Multimodal Poster Automation from Scientific Papers. *arXiv preprint*. ArXiv:2505.21497 [cs].

Massimiliano Pronesti, Joao Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisin Redmond, Anya Belz, and Yufang Hou. 2025. Query-driven Document-level Scientific Evidence Extraction from Biomedical Studies. *arXiv preprint*. ArXiv:2505.06186 [cs].

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*. ArXiv:2103.00020 [cs].

Yucheng Shi, Tianze Yang, Canyu Chen, Quanzheng Li, Tianming Liu, Xiang Li, and Ninghao Liu. SearchRAG: Can Search Engines Be Helpful for LLM-based Medical Question Answering?

Tao Sun, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongliang Li, Wenhao Huang, Ge Zhang, Jian Yang, and Zhoujun Li. 2025. P2P: Automated Paper-to-Poster Generation and Fine-Grained Benchmark. *arXiv preprint*. ArXiv:2505.17104 [cs].

Rahul Thapa, Qingyang Wu, Kevin Wu, Harrison Zhang, Angela Zhang, Eric Wu, Haotian Ye, Suhana Bedi, Nevin Aresh, Joseph Boen, Shriya Reddy, Ben Athiwaratkun, Shuaiwen Leon Song, and James Zou. 2025. Disentangling Reasoning and Knowledge in Medical Large Language Models. *arXiv preprint*. ArXiv:2505.11462 [cs].

Rishabh Upadhyay and Marco Viviani. Enhancing Health Information Retrieval with RAG by Prioritizing Topical Relevance and Factual Accuracy.

Yubo Wang, Xueguang Ma, and Wenhu Chen. 2025. Augmenting Black-box LLMs with Medical Textbooks for Biomedical Question Answering. *arXiv preprint*. ArXiv:2309.02233 [cs].

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation.

Yue Yu, Wei Ping, Zihan Liu, Boxin Wang, Jiaxuan You, Chao Zhang, Mohammad Shoeybi, and Bryan Catanzaro. 2024. RankRAG: Unifying Context Ranking with Retrieval-Augmented Generation in LLMs. *arXiv preprint*. ArXiv:2407.02485 [cs].

Xuejiao Zhao, Siyan Liu, Su-Yin Yang, and Chunyan Miao. MedRAG: Enhancing Retrieval-augmented Generation with Knowledge Graph-Elicited Reasoning for Healthcare Copilot.