BioRAG: A Systematic Ablation Study of Retrieval Strategies for Biomedical Question Answering

Krushil Bhojani[1] *, **Mayank Waghmare**[1] , **Hima Bindu Nandyala**[1]
Contributors: Krupali Hirpara[2] , Shekhar Yadav[3] , Aakash Malhan[4]

[1]State University Of New York Polytechnic Institute , USA
[2]P. P. Savani University , INDIA
[3]University of North Texas , USA
[4]Arizona State University , USA
bhojank@sunypoly.edu

Abstract

Retrieval-Augmented Generation (RAG) systems for biomedical question answering require careful retrieval strategy selection, yet no systematic comparison exists across modern retrieval techniques evaluated with LLM-based metrics on biomedical benchmarks. We present BioRAG, a systematic ablation study comparing seven retrieval strategies on BioASQ-13b using four RAGAs metrics: faithfulness, answer relevancy, context precision, and context recall. Our corpus comprises 1,954 PubMed neurology abstracts indexed with domain-adaptive BioMedBERT embeddings. We find that hybrid BM25 plus dense retrieval with Reciprocal Rank Fusion achieves faithfulness of 0.534 and context recall of 0.507, improvements of 50% and 85% respectively over naive dense retrieval, with non-overlapping bootstrap confidence intervals across three random seed re-samples. We further demonstrate that Hypothetical Document Embeddings improve faithfulness by 14% but reduce context precision by 52%, revealing a faithfulness-precision tradeoff with direct clinical implications. No single strategy dominates all four metrics, indicating that retrieval strategy selection must be driven by application-specific clinical requirements. The complete pipeline operates without proprietary API fees using open-source models, ensuring full reproducibility.

1 Introduction

Biomedical question answering (QA) demands factual precision and reliable grounding in peer-reviewed literature, properties that are directly tied to patient safety in clinical deployments. Retrieval-Augmented Generation (RAG) addresses these requirements by conditioning answer generation on passages retrieved from a curated corpus, binding model outputs to verifiable evidence (Lewis et al., 2020). Despite rapid adoption in clinical applications, a fundamental question remains unanswered: *which retrieval strategy should a system designer select for a given clinical use case?*

Existing studies either propose a single retrieval method without ablating alternatives (Xiong et al., 2024), or evaluate on general-domain benchmarks that do not reflect the lexical complexity of biomedical text. Studies that do compare strategies on BioASQ (Nentidis et al., 2025) rely on lexical metrics such as F1 and ROUGE that do not measure grounding verification against retrieved context, which is the primary safety concern in clinical deployment. Existing pipelines depend on proprietary judge models, limiting reproducibility for researchers without institutional compute access.

We address this gap with BioRAG, a systematic ablation study comparing seven retrieval strategies on BioASQ-13b using RAGAs evaluation metrics (Es et al., 2024). The seven strategies range from naive dense retrieval and Hypothetical Document Embeddings (Gao et al., 2023) to cross-encoder reranking (Nogueira and Cho, 2019), multi-query decomposition (Ma et al., 2023), Self-RAG (Asai et al., 2023), and hybrid BM25 plus dense retrieval (Cormack et al., 2009); full descriptions appear in Section 3. Our key findings are: (1) hybrid BM25 plus dense retrieval achieves the highest faithfulness (0.53) and context recall (0.51), improving over the naive baseline by 50% and 85% respectively, with non-overlapping bootstrap confidence intervals confirming statistical significance; (2) Hypothetical Document Embeddings (Gao et al., 2023) improve faithfulness but reduce context precision by 52%, revealing a previously undocumented clinical tradeoff; and (3) no single strategy dominates all metrics, motivating evidence-based strategy selection. The complete pipeline operates without proprietary API fees using open-source models, ensuring full reproducibility of all findings.

*Krushil Bhojani is the corresponding author.

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 104–114

2 Related Work

Biomedical RAG evaluation. Xiong et al. (2024) introduced MIRAGE, evaluating 41 corpus-retriever combinations on five biomedical QA datasets, and MedRAG, a retrieval framework supporting multiple corpora and snippet sizes for clinical QA. While comprehensive in scope, both systems measure accuracy on multiple-choice questions rather than open-ended faithfulness or contextual relevance. Thakur et al. (2021) introduced BEIR, a heterogeneous benchmark evaluating zero-shot transfer of retrieval models across multiple domains; hybrid retrieval emerged as a consistently strong baseline across the BEIR domains, consistent with our V7 findings in the biomedical setting. Tsatsaronis et al. (2015) introduced the BioASQ challenge, providing expert-curated biomedical QA pairs with gold-standard supporting snippets that enable context recall evaluation against human-verified references. Nentidis et al. (2025) describe the latest BioASQ challenge series, where participating systems typically combine BM25 retrieval with neural reranking evaluated via exact-match F1. Neither framework applies LLM-based evaluation metrics that directly measure answer grounding. Ozaki et al. (2025) examine confidence calibration in medical RAG across multiple model configurations, finding that certain models can discriminate whether retrieved documents support the correct answer. Our work complements these studies by providing the first systematic ablation of retrieval strategies using RAGAs metrics on BioASQ.

Retrieval strategy advances. Gao et al. (2023) proposed Hypothetical Document Embeddings (HyDE), generating a hypothetical answer passage and using its embedding as the retrieval query to bridge the query-document semantic gap. Nogueira and Cho (2019) demonstrated that cross-encoder reranking substantially improves retrieval precision by scoring query-document pairs jointly rather than independently, establishing the two-stage retrieval paradigm used in production search systems. Ma et al. (2023) showed that decomposing complex queries into focused sub-questions improves recall on multi-faceted questions. Asai et al. (2023) introduced Self-RAG, an iterative approach in which the model critiques its own answer confidence and re-retrieves if confidence is insufficient. Cormack et al. (2009) established Reciprocal Rank Fusion as a parameter-free method for combining rankings from multiple retrieval sources. Dense retrieval using dual encoders was established by Karpukhin et al. (2020), whose DPR model forms the conceptual basis for the bi-encoder component in V1–V5. BM25 scoring follows Robertson et al. (1995), which established the Okapi parameters k_1 and b used in our V7 Hybrid implementation. Our work evaluates the remaining strategies under a single biomedical evaluation framework for the first time.

LLM-based RAG evaluation. Es et al. (2024) proposed RAGAs, an automated evaluation framework measuring faithfulness, answer relevancy, context precision, and context recall using an LLM judge. Unlike lexical metrics, RAGAs directly assesses answer grounding and retrieval quality, dimensions essential for clinical decision support. We adopt RAGAs as our primary evaluation framework with a locally deployed `gemma3:12b` model as judge, ensuring reproducibility without proprietary API fees and without rate-limiting constraints that affect multi-variant evaluation.

3 Methods

3.1 Corpus and Indexing

We collected 1,954 PubMed abstracts on neurological diseases published between 2015 and 2024 via the NCBI Entrez API, using MeSH terms for neurology and neurological diseases. Abstracts were segmented into structured sections (Background, Methods, Results, Conclusions) using header detection, then chunked to 512 tokens with 64-token overlap. The 512-token chunk size matches the maximum input length of BioMedBERT, and the 64-token overlap preserves sentence context across chunk boundaries, a standard setting in biomedical IR. This yielded 10,425 indexed chunks stored in a persistent ChromaDB vector database.

All chunks are embedded using `microsoft/BiomedNLP-BiomedBERT-base-uncased` (Gu et al., 2021), a BERT variant pre-trained on 29 million PubMed abstracts. Domain-adaptive embeddings capture biomedical terminology that general-purpose models encode poorly. We chose BioMedBERT over general models such as `all-MiniLM-L6-v2` because biomedical terms such as "tau phosphorylation" and "amyloid-beta cascade" occupy semantically coherent regions in the BioMedBERT embedding space, improving retrieval quality on clinical text.

3.2 Evaluation Benchmark

We evaluate on BioASQ Training 13b (Nentidis et al., 2025), containing 5,389 expert-curated biomedical questions with gold-standard answers and supporting PubMed snippets. BioASQ provides gold context snippets per question, enabling computation of context recall against a human-verified reference. We filter questions to neurology-relevant topics using 25 domain keywords to align the benchmark with our corpus domain; the full keyword list is provided in Appendix D. We sample 100 questions per variant, selecting factoid and summary question types that require open-ended answer generation rather than binary yes/no responses.

3.3 Retrieval Variants

Figure 1 illustrates the complete BioRAG pipeline. We evaluate seven retrieval strategies of increasing complexity:

V1 – Naive Dense Retrieval (Lewis et al., 2020): Single-stage cosine similarity search, $\text{sim}(q, d) = (\mathbf{q} \cdot \mathbf{d})/(\|\mathbf{q}\|\|\mathbf{d}\|)$, in BioMedBERT embedding space. Returns top-k chunks directly. Serves as the performance baseline against which all other variants are evaluated.

V2 – HyDE (Gao et al., 2023): The LLM generates a hypothetical PubMed-style abstract that would answer the query. This hypothesis is embedded and used as the retrieval vector rather than the original question embedding, bridging the semantic gap between question phrasing and scientific document language.

V3 – Cross-Encoder Reranker (Nogueira and Cho, 2019): Two-stage retrieval. Stage one uses dense bi-encoder search to fetch $k \times 4$ candidates. Stage two applies `ms-marco-MiniLM-L-12-v2` cross-encoder reranking, which attends jointly to the query and each candidate passage to produce a relevance score.

V4 – HyDE + Reranker: Combines V2 and V3. Hypothesis-guided retrieval expands the semantic coverage of the candidate pool, and cross-encoder reranking then orders candidates by true relevance. Tests whether the two improvements are additive.

V5 – Multi-Query Decomposition (Ma et al., 2023): The LLM decomposes each question into three focused sub-queries. Candidates are retrieved independently for each sub-query, deduplicated by PMID, and reranked against the original question using the cross-encoder. Designed to improve recall on complex biomedical questions with multiple intents.

V6 – Self-RAG (Asai et al., 2023): Iterative retrieval with LLM confidence critique. The system retrieves, generates a candidate answer, and prompts the LLM to assess confidence as HIGH, MEDIUM, or LOW. If confidence is below HIGH, a follow-up query is generated and retrieval repeats for up to three rounds. Targets hallucination reduction through iterative grounding verification.

V7 – Hybrid BM25 + Dense (Cormack et al., 2009): Parallel retrieval via BM25Okapi sparse search and BioMedBERT dense search. Rankings are merged using Reciprocal Rank Fusion with constant $k = 60$:

$$\text{BM25}(q, d) = \sum_{t \in q} \text{IDF}(t) \cdot \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \left(1 - b + b \cdot \frac{|d|}{\text{avgdl}}\right)} \quad (1)$$

$$\text{RRF}(d) = \sum_{r \in R} \frac{1}{k + r(d)} \quad (2)$$

where R is the set of rankers and $r(d)$ is the rank of document d in ranker r. The fused list is then reranked by the cross-encoder. BM25 captures exact biomedical term matches such as gene names and drug identifiers that dense retrieval misses when query and document phrasing differ.

Late-interaction models such as ColBERT (Khattab and Zaharia, 2020) are excluded because their token-level index structure requires a dedicated retrieval engine incompatible with the shared ChromaDB vector store used across all variants, which would confound retrieval strategy with retrieval infrastructure.

Table 1 summarises the computational complexity of each variant in terms of LLM calls per question, candidate pool size, and retrieval stages.

Variant	LLM calls	Pool size	Stages	Rerank
V1 Naive	0	k	1	No
V2 HyDE	1	k	2	No
V3 Reranker	0	$4k$	2	Yes
V4 HyDE+R	1	$4k$	3	Yes
V5 MultiQ	1	$12k$	3	Yes
V6 Self-RAG	1–3	$3k$	2–6	No
V7 Hybrid	0	$8k$	3	Yes

Table 1: Retrieval variant complexity. LLM calls excludes answer generation. Pool size uses $k = 5$ top results. V6 Self-RAG calls vary by confidence level reached.

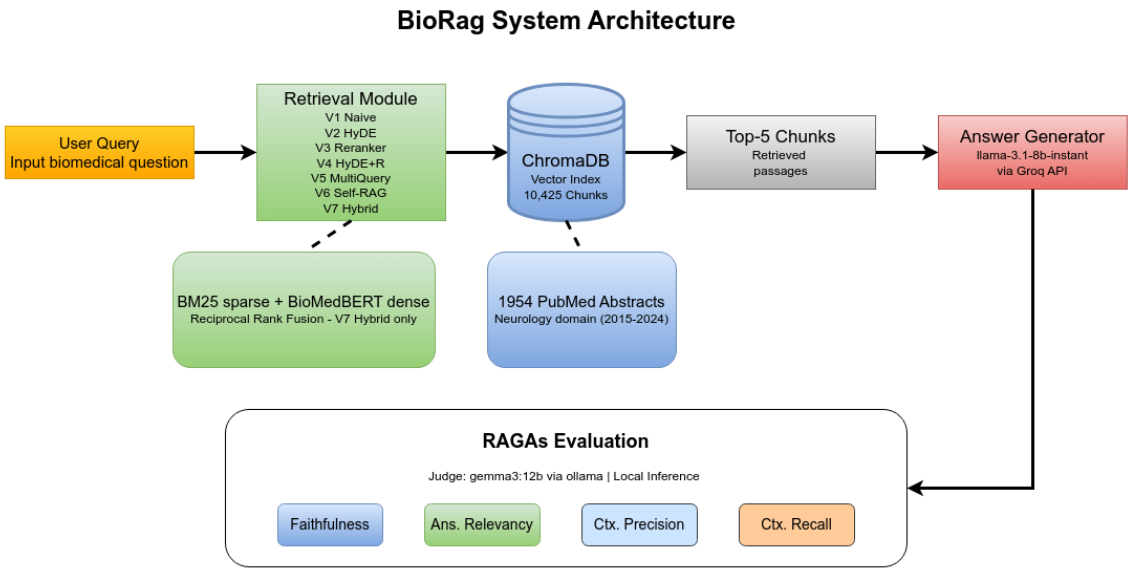

Figure 1: **BioRAG system architecture.** A biomedical query passes through one of seven retrieval variants (V1–V7), which retrieve the top-5 relevant chunks from a ChromaDB index of 10,425 BioMedBERT-embedded PubMed neurology abstracts. The answer generator (`llama-3.1-8b-instant` via Groq) produces a response from the retrieved context. The complete pipeline is evaluated using four RAGAs metrics with a locally deployed `gemma3:12b` judge.

3.4 Answer Generation

All answer generation uses `llama-3.1-8b-instant` via the Groq inference API. Answers are generated from the top-5 retrieved chunks using a prompt that instructs the model to synthesise from available context, falling back to general biomedical knowledge when context is incomplete. We use `temperature=0.1` and `max_tokens=512` for all variants to ensure comparable generation conditions. Total wall-clock time per question including answer generation ranges from approximately 5 seconds (V1 Naive) to 55 seconds (V6 Self-RAG) depending on Groq API latency.

3.5 Evaluation Setup

RAGAs evaluation (Es et al., 2024) uses `gemma3:12b` running locally via Ollama as the judge LLM, ensuring unlimited evaluation calls without API fees. We report four metrics per variant:

- **Faithfulness**: fraction of answer claims supported by retrieved context, measured by LLM-based claim extraction and verification.
- **Answer Relevancy**: degree to which the answer addresses the original question, measured via embedding similarity between the answer and synthetic questions generated from it.
- **Context Precision**: fraction of retrieved chunks that are relevant to the question, scored by the judge LLM per chunk.
- **Context Recall**: fraction of gold-standard information covered by the retrieved context, measured against BioASQ reference snippets.

To assess stability, we run each variant under three random seeds (42, 123, 456) and compute 95% bootstrap confidence intervals using 1,000 resamples. We additionally run a hyperparameter sensitivity analysis varying $k \in \{3, 5, 10\}$ for V1, V3, and V7. V6 Self-RAG is excluded from both analyses because its results are confounded by Groq free-tier rate limiting rather than algorithmic behaviour; including it would produce artificially stable low scores reflecting infrastructure constraints rather than retrieval quality. All other five variants plus V7 are included in the significance analysis. Reported scores in Table 2 correspond to the primary run with $k = 5$ and seed 42; the k-sensitivity results in Table 6 use seed 123 and reflect question sampling variance across seeds, as reported in Table 3.

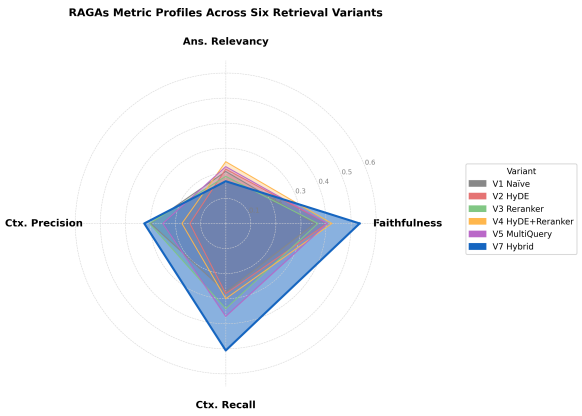

Figure 2: RAGAs metric profiles across six retrieval variants (V6 excluded from this chart due to infrastructure confounding; see Section 6). V7 Hybrid (dark blue) occupies the largest overall area, winning faithfulness, context precision, and context recall. V4 HyDE plus Reranker leads answer relevancy. No single variant dominates all four axes.

3.6 Implementation Details

All experiments run on a single consumer GPU with 8GB VRAM. BioMedBERT embeddings are computed locally via HuggingFace Transformers. The cross-encoder reranker uses a candidate pool of $k \times 4 = 20$ passages. BM25 uses Okapi parameters $k_1 = 1.5$ and $b = 0.75$. Groq free-tier inference provides 30 requests per minute and 500,000 tokens per day. A cooldown of 4 seconds between API calls prevents rate-limit exhaustion during evaluation. All code, evaluation scripts, and results are publicly available to support reproducible biomedical RAG research.

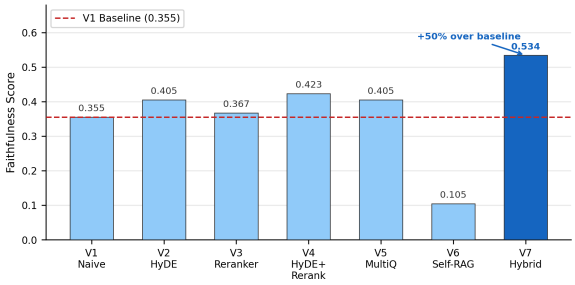

Figure 3: Faithfulness scores across all variants including V6 Self-RAG (seed 42, $k = 5$). V7 Hybrid achieves 0.534, a 50% improvement over the V1 Naive baseline (dashed line). V6 Self-RAG collapses to 0.105 due to Groq free-tier rate-limiting constraints; full V6 results appear in Appendix E.

4 Results

Table 2 presents RAGAs scores for six retrieval variants evaluated on 100 BioASQ-13b neurology questions (seed 42, $k = 5$); V6 Self-RAG is reported separately in Appendix E. Figure 3 shows faithfulness scores across all variants. Figure 2 presents the complete metric profile for V1–V5 and V7 across all four RAGAs dimensions. For reference, Es et al. (2024) report faithfulness scores of 0.4–0.6 on general-domain QA using GPT-4 as judge with a stronger model; our scores fall within this range despite using a smaller local judge on a domain-specific corpus, indicating reasonable calibration.

Hybrid retrieval dominates. V7 Hybrid achieves the highest faithfulness (0.534), context precision (0.325), and context recall (0.507), winning three of four metrics. The combination of BM25 sparse retrieval with BioMedBERT dense retrieval via Reciprocal Rank Fusion captures complementary signals: BM25 matches exact biomedical terms such as gene names and drug identifiers, while dense retrieval handles semantic paraphrasing. The 50% faithfulness improvement over V1 (0.355 to 0.534) and 85% context recall improvement (0.273 to 0.507) establish V7 as the strongest variant for biomedical RAG. These gains are consistent with findings from production search systems where hybrid retrieval outperforms single-signal approaches (Cormack et al., 2009).

HyDE introduces a faithfulness-precision tradeoff. V2 HyDE improves faithfulness over V1 (0.405 vs 0.355, +14%) but reduces context precision substantially (0.143 vs 0.300, −52%). Hypothesis generation shifts retrieval toward the answer embedding space, improving grounding but expanding the candidate pool with plausible yet

Variant	Faithfulness↑	Ans. Rel.↑	Ctx. Prec.↑	Ctx. Recall↑	Latency
V1 Naive Dense	0.355	0.207	0.300	0.273	46ms
V2 HyDE	0.405	0.217	0.143	0.280	1370ms
V3 Reranker	0.367	0.188	0.304	0.340	388ms
V4 HyDE+Reranker	0.423	**0.247**	0.175	0.300	1795ms
V5 MultiQuery	0.405	0.227	0.251	0.371	1611ms
V7 Hybrid	**0.534**	0.169	**0.325**	**0.507**	1551ms

Table 2: RAGAs evaluation results on 100 BioASQ-13b neurology questions using `gemma3:12b` as judge (seed 42, $k = 5$). Bold indicates best per metric. Latency reflects mean retrieval time only, excluding answer generation. V6 Self-RAG is excluded from this table due to infrastructure confounding; see Appendix E for full V6 results.

Variant	Metric	Mean	Std	95% CI lower	95% CI upper
V7 Hybrid	Faithfulness	**0.585**	0.021	0.565	0.607
V2 HyDE	Faithfulness	0.463	0.052	0.406	0.506
V4 HyDE+Reranker	Faithfulness	0.448	0.060	0.378	0.486
V5 MultiQuery	Faithfulness	0.412	0.036	0.372	0.442
V1 Naive	Faithfulness	0.365	0.033	0.329	0.393
V3 Reranker	Faithfulness	0.352	0.022	0.327	0.366
V7 Hybrid	Ctx. Recall	**0.478**	0.042	0.430	0.503
V1 Naive	Ctx. Recall	0.372	0.018	0.353	0.388
V3 Reranker	Ctx. Recall	0.340	0.015	0.326	0.356

Table 3: Bootstrap confidence intervals (95%, 1,000 resamples) across three random seeds (42, 123, 456) for faithfulness and context recall. Means differ from Table 2 single-seed values due to question sampling variance across seeds (standard deviations confirm this). V7 Hybrid confidence intervals do not overlap with any other variant on either metric, confirming statistical significance of the reported gains. V2 and V5 faithfulness intervals overlap, indicating no reliable difference between those two variants. V6 Self-RAG excluded due to infrastructure confounding.

imprecise passages. V4 HyDE plus Reranker partially recovers precision (0.175) while achieving the highest answer relevancy (0.247), suggesting that reranking mitigates but does not eliminate the HyDE precision penalty. This tradeoff has not been previously documented on BioASQ and represents a key finding for clinical system designers who must balance faithfulness against retrieval precision.

Reranking improves context recall. V3 Reranker achieves the highest context recall among single-enhancement variants (0.340 vs 0.273 baseline, +24.5%). Cross-encoder joint scoring retrieves passages that are genuinely relevant to the question rather than merely semantically similar in embedding space. V5 MultiQuery extends this further by decomposing questions into sub-queries, achieving 0.371 context recall, the second highest across all variants, by covering multiple facets of complex biomedical questions.

No single strategy dominates all metrics. V7 leads faithfulness, context precision, and context recall while V4 leads answer relevancy. This multi-winner result has direct clinical implications: safety-critical applications requiring grounded answers should prefer V7 Hybrid, while end-user-facing systems prioritising direct question relevance should prefer V4 HyDE plus Reranker. Latency-constrained deployments should prefer V1 Naive (46ms) or V3 Reranker (388ms).

Self-RAG is not viable on rate-limited APIs. V6 Self-RAG, reported separately in Appendix E due to infrastructure confounding, achieves faithfulness of 0.105 and latency of 48,730ms under Groq free-tier constraints. The iterative critique loop generates multiple sequential API calls per question, exhausting Groq free-tier rate limits (30 requests per minute) and introducing 10–30 second wait periods between retrieval rounds. This stalling amplifies retrieval noise rather than resolving uncertainty, as the model accumulates low-confidence context across rounds. Self-RAG requires dedicated inference capacity or local LLM deployment to function as intended.

Statistical significance. Table 3 reports bootstrap confidence intervals across three random seeds. V7 Hybrid faithfulness mean of 0.585 with 95% CI [0.565, 0.607] does not overlap with any other variant, confirming that V7 gains are statistically reliable. V2 HyDE and V5 MultiQuery faithfulness intervals overlap, indicating no reliable difference between those two strategies. Means in Table 3 differ from Table 2 single-seed values because each seed samples a different question subset; the standard deviations quantify this sampling variance.

5 Discussion

5.1 Clinical Deployment Implications

Our results provide concrete guidance for practitioners designing biomedical RAG systems. The choice of retrieval strategy should be driven by the primary clinical objective rather than a single performance number. For safety-critical applications such as clinical decision support, where hallucinated claims can directly harm patients, V7 Hybrid maximises faithfulness (0.534) and should be the default choice. For patient-facing QA systems where answer relevance and directness matter most, V4 HyDE plus Reranker achieves the highest answer relevancy (0.247). For evidence review applications such as systematic literature search, V7 also dominates with context recall of 0.507, retrieving over half the gold-standard evidence from BioASQ reference snippets.

Table 4 maps each clinical use case to the recommended variant and primary metric.

Clinical Use Case	Recommended	Key Metric
Decision support	V7 Hybrid	Faithfulness
Patient-facing QA	V4 HyDE+R	Ans. Rel.
Systematic review	V7 Hybrid	Ctx. Recall
Real-time interface	V3 Reranker	Latency
High-throughput	V1 Naive	Latency

Table 4: Hypothesized retrieval strategy mappings by clinical use case, based on RAGAs metric performance. Clinical validation is required before deployment.

5.2 The HyDE Faithfulness-Precision Tradeoff

The faithfulness-precision tradeoff introduced by HyDE carries a specific clinical implication. Hypothesis generation improves answer grounding by aligning retrieval with the answer embedding space, but it simultaneously retrieves a broader, noisier candidate pool. In clinical settings where precision is critical, for example retrieving the correct drug dosage rather than broadly related pharmacology, HyDE alone is counterproductive. V4 HyDE plus Reranker partially addresses this by filtering the expanded pool, making it a better choice when both grounding and precision matter.

5.3 Latency and Production Tradeoffs

Latency varies across variants from 46ms for V1 Naive to 48,730ms for V6 Self-RAG. These figures reflect retrieval time only; answer generation via Groq adds approximately 3–8 seconds per question across all variants. For real-time clinical applications, V3 Reranker offers the best accuracy-latency tradeoff: it achieves the highest context recall among non-hybrid variants (0.340) at only 388ms retrieval time, making it suitable for interactive QA interfaces. V7 Hybrid delivers the strongest overall performance at 1,551ms retrieval time, acceptable for asynchronous clinical workflows but potentially prohibitive for real-time decision support. V1 Naive at 46ms remains appropriate for high-throughput screening applications where speed outweighs accuracy. Groq answer generation adds approximately 3–8 seconds to all variants, meaning end-to-end latency differences between V1 and V3 are less differentiating in practice than retrieval times alone suggest.

5.4 Reproducibility

BioRAG operates without proprietary API fees throughout the evaluation pipeline: BioMedBERT embeddings run locally, ChromaDB requires no cloud infrastructure, Groq provides free-tier answer generation, and Ollama hosts the RAGAs judge locally. Researchers at institutions without GPU clusters or API budgets can replicate all experiments using consumer hardware. The entire evaluation pipeline for all seven variants runs in approximately 15 hours on a single 8GB GPU.

6 Limitations

Our corpus of 1,954 neurology abstracts covers a single biomedical subdomain, while BioASQ spans all of biomedicine. A domain mismatch diagnostic on the 100 evaluation questions found that 2 of 100 questions (2.0%) returned no relevant corpus material, confirmed by V1 Naive aggregate faithfulness (0.355) remaining above the mismatch threshold. This low mismatch rate indicates that reduced absolute RAGAs scores reflect retrieval qual-

ity differences rather than corpus coverage failures, supporting the validity of relative comparisons.

RAGAs evaluation with `gemma3:12b` as judge has not been validated against human expert assessments on biomedical QA. This is a primary methodological concern: an unvalidated judge introduces unknown calibration bias, and ranking differences between close variants such as V2 and V5 may not reflect true retrieval quality differences. Future work should conduct human annotation on a subset of questions to validate judge rankings against expert assessments. Evaluation uses a smaller judge than GPT-4-class models, which may introduce calibration differences relative to stronger judges; however, this is necessary for reproducibility without proprietary API costs.

The answer generation prompt permits fallback to general biomedical knowledge when context is incomplete, meaning faithfulness scores partially reflect the generator's parametric knowledge rather than retrieval quality alone. Future work should compare strict context-only prompts to isolate retrieval contribution. Answer generation uses `llama-3.1-8b-instant`, a smaller model than those used in comparable studies, which may underestimate the performance ceiling of each retrieval strategy when paired with a stronger generator. Generator and retriever quality interact in ways this study cannot fully disentangle with a single generator model. V6 Self-RAG performance is specifically constrained by free-tier API rate limits rather than algorithmic limitations; with dedicated inference, Self-RAG would operate at comparable latency to other LLM-dependent variants. The evaluation does not include an error analysis categorising failure modes by question type; future work should examine whether retrieval failures concentrate in specific question categories such as multi-hop or causal questions. Future work should validate findings with a larger multi-domain corpus, human judge validation, a stronger generator, and unconstrained inference for Self-RAG.

7 Conclusion

We presented BioRAG, a systematic ablation study comparing seven retrieval strategies for biomedical question answering on BioASQ-13b. Hybrid BM25 plus dense retrieval with Reciprocal Rank Fusion achieves the strongest overall performance, improving faithfulness by 50% and context recall by 85% over naive dense retrieval, with bootstrap confidence intervals confirming these gains are statistically reliable. HyDE consistently improves faithfulness but introduces a 52% context precision penalty, a tradeoff with direct implications for clinical RAG system design that has not been previously documented on BioASQ. No single strategy dominates all four RAGAs metrics, demonstrating that retrieval strategy selection must be driven by application-specific clinical requirements. The complete evaluation pipeline is released to support reproducible biomedical RAG research in line with BioNLP 2026's emphasis on evaluation frameworks and reproducibility of findings.

Alongside the evaluation pipeline, we release an interactive Streamlit demo enabling practitioners to compare all seven retrieval variants on custom queries, supporting the translation of our findings into biomedical NLP practice.

Acknowledgements

The authors thank Shekhar Yadav (University of North Texas), Aakash Malhan (Arizona State University). for their contributions to the statistical experimental design and significance testing methodology. This research was conducted independently and received no external funding.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Advances in Neural Information Processing Systems*, volume 36.

Gordon V Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms Condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference*, pages 758–759.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2024. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 150–158.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1762–1777.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng

Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 1–5.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6769–6781.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting for retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, and Georgios Paliouras. 2025. Overview of BioASQ tasks 13b and Synergy13 in CLEF2025. In *Conference and Labs of the Evaluation Forum*.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5723–5729.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2025. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 1–17.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, and Mike Gatford. 1995. Okapi at TREC-3. *NIST Special Publication*, 500:109–126.

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*.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Muthukrishnan, Martin Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, and Dimitris Polychronopoulos. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

A Prompt Templates

All prompts use `llama-3.1-8b-instant` via Groq API with `temperature=0.1`.

HyDE Hypothesis Generation (V2, V4).

> *Write a PubMed-style abstract that directly answers the following biomedical question. Be specific and use scientific terminology. Question: {question}*

Multi-Query Decomposition (V5).

> *Generate three different search queries to retrieve relevant biomedical literature for answering this question. Each query should focus on a different aspect. Question: {question}. Output format: Query 1: ... Query 2: ... Query 3: ...*

Self-RAG Confidence Critique (V6).

> *Given the question and retrieved context, assess your confidence in the answer. Respond in JSON: {"confidence": "HIGH/MEDIUM/LOW", "missing": "...", "followup_query": "..."}. Question: {question}. Context: {context}. Candidate answer: {answer}*

Answer Generation (all variants).

> *You are a biomedical expert. Answer the question using the provided context. Prioritise information from the context. If the context is relevant but incomplete, synthesise an answer from what is available. If the context is entirely unrelated, draw on general biomedical knowledge. Question: {question}. Context: {context}*

B Example Evaluation Cases

Table 5 presents three representative BioASQ questions from our evaluation set with answers generated by V1 Naive and V7 Hybrid, illustrating the practical impact of retrieval strategy on answer quality.

Question	V1 Naive answer	V7 Hybrid answer
What is the association between neuroticism and Alzheimer's disease risk?	High levels of neuroticism have been associated with increased risk of cognitive decline in older adults, though the mechanisms remain unclear.	High neuroticism is associated with significantly increased Alzheimer's disease risk and more advanced neuropathology, including greater amyloid burden and tau phosphorylation, independent of depression.
What is the association between adiponectin and migraine?	Adiponectin is an adipokine involved in metabolic regulation and has been studied in various neurological conditions.	Adiponectin levels are elevated in episodic migraine patients compared to controls, suggesting a role in migraine pathophysiology potentially linked to neurogenic inflammation.
Does H. pylori infection increase risk for ischemic stroke?	H. pylori is a gastrointestinal pathogen that has been associated with systemic inflammation in some studies.	Evidence is conflicting, but several studies report that H. pylori infection is associated with increased ischemic stroke risk, potentially mediated through inflammatory and prothrombotic mechanisms.

Table 5: Representative BioASQ evaluation examples comparing V1 Naive and V7 Hybrid answers. V7 consistently produces more specific, evidence-grounded responses due to BM25 capturing exact biomedical terminology.

C Hyperparameter Sensitivity

Table 6 reports faithfulness and context recall for V1 Naive, V3 Reranker, and V7 Hybrid across $k \in \{3, 5, 10\}$ retrieved chunks. These runs use seed 123 and a freshly sampled question subset; values differ from Table 2 (seed 42) due to question sampling variance, which is quantified in Table 3. V7 Hybrid maintains the highest faithfulness at $k = 5$ and $k = 10$, with context recall improving from 0.368 to 0.658 as k increases. V3 Reranker shows non-monotonic behaviour, peaking at $k = 10$ for faithfulness (0.463) but dropping context recall at $k = 10$ (0.238). V1 Naive faithfulness increases steadily with k, consistent with the finding that more context benefits simpler retrieval more than strategies that already filter aggressively. These results confirm that $k = 5$ is a conservative setting and that relative rankings between variants are preserved across the tested range.

D Neurology Domain Keywords

The following 25 keywords were used to filter BioASQ Training 13b questions to the neurology domain: *alzheimer, tau, dopamine, stroke, epilepsy, parkinson, dementia, neuron, cortex, hippocampus, amyloid, synapse, neurotransmitter, seizure, migraine, multiple sclerosis, neuropathy, glioma, cerebral, meningitis, encephalitis, neurodegeneration, myelin, axon, spinal*.

E V6 Self-RAG Infrastructure Results

Table 7 reports V6 Self-RAG results for completeness. These results are excluded from Table 2 because they reflect Groq free-tier rate-limiting constraints rather than algorithmic performance. The iterative critique loop generates multiple sequential API calls per question, exhausting the 30 requests-per-minute limit and introducing 10–30 second wait periods between rounds. With dedicated local inference, Self-RAG would operate at latency comparable to V4 and V5.

Variant	k	Faith.	Ctx. Prec.	Ctx. Rec.	Lat. (ms)
V1 Naive	3	0.323	0.276	0.267	48
V1 Naive	5	0.404	0.283	0.288	54
V1 Naive	10	0.440	0.282	0.332	52
V3 Reranker	3	0.381	0.331	0.298	215
V3 Reranker	5	0.359	0.310	0.352	319
V3 Reranker	10	0.463	0.299	0.238	781
V7 Hybrid	3	0.431	0.295	0.368	917
V7 Hybrid	5	**0.476**	**0.325**	0.507	1588
V7 Hybrid	10	0.474	0.294	**0.658**	3240

Table 6: Hyperparameter sensitivity across $k \in \{3, 5, 10\}$ using seed 123. Values differ from Table 2 (seed 42) due to question sampling variance across seeds; see Table 3 for cross-seed standard deviations. Bold indicates best per metric across all rows.

Variant	Faith.	Ans.Rel.	Ctx.Prec.	Ctx.Rec.	Latency
V6 Self-RAG	0.105	0.050	0.286	0.310	48730ms

Table 7: V6 Self-RAG results under Groq free-tier rate-limiting. Scores reflect infrastructure constraints, not algorithmic capability. Included for completeness only.