Pride-Boiler at MedGenVidQA 2026: LLM-Augmented BM25 Retrieval with Corrective Self-Verification for Biomedical Evidence Retrieval

Basil T. Ebinesar, Keyuan Jiang, Charansai Maddineni, Ashok Vardhan Raja
Purdue University Northwest
{bebinesa, kjiang, cmaddine, raja22}@pnw.edu

Abstract

This paper describes the Pride-Boiler system submitted to MedGenVidQA 2026 Shared Task A, which asks for retrieving relevant PubMed articles and medical instructional videos in response to consumer health queries. Our approach pairs Pyserini BM25 retrieval with LLM-driven query rewriting and a corrective self-verification loop inspired by the Corrective Retrieval-Augmented Generation (CRAG) paradigm. Given a consumer query, the pipeline first asks Google Gemini to generate clinically optimized search text, one targeting PubMed abstracts with MeSH terms and clinical synonyms, and another targeting video subtitles with procedural action language. BM25 retrieves a broad candidate pool, and Gemini then scores each candidate against the original query, blending its relevance judgment with the normalized lexical signal. A quality grader assesses the top results: if they are judged insufficient, the pipeline triggers a corrective cycle with reformulated terminology and retries up to three attempts. The entire workflow is orchestrated as a LangGraph state machine. In the official shared task evaluation, Pride-Boiler ranked first among all participating systems on PubMed article retrieval, achieving an nDCG of 0.6532 and MAP of 0.5550, both exceeding the organizer-provided Text-RR baseline. Our performance on video (text) retrieval achieves 0.5304 in MAP and 0.5927 in nDCG, outperforming other systems but falling below that of baseline, indicating the structural limitations of lexical matching over noisy subtitle text. We release the pipeline code to support reproducibility on GitHub at https://github.com/basilll007/BioNLP.

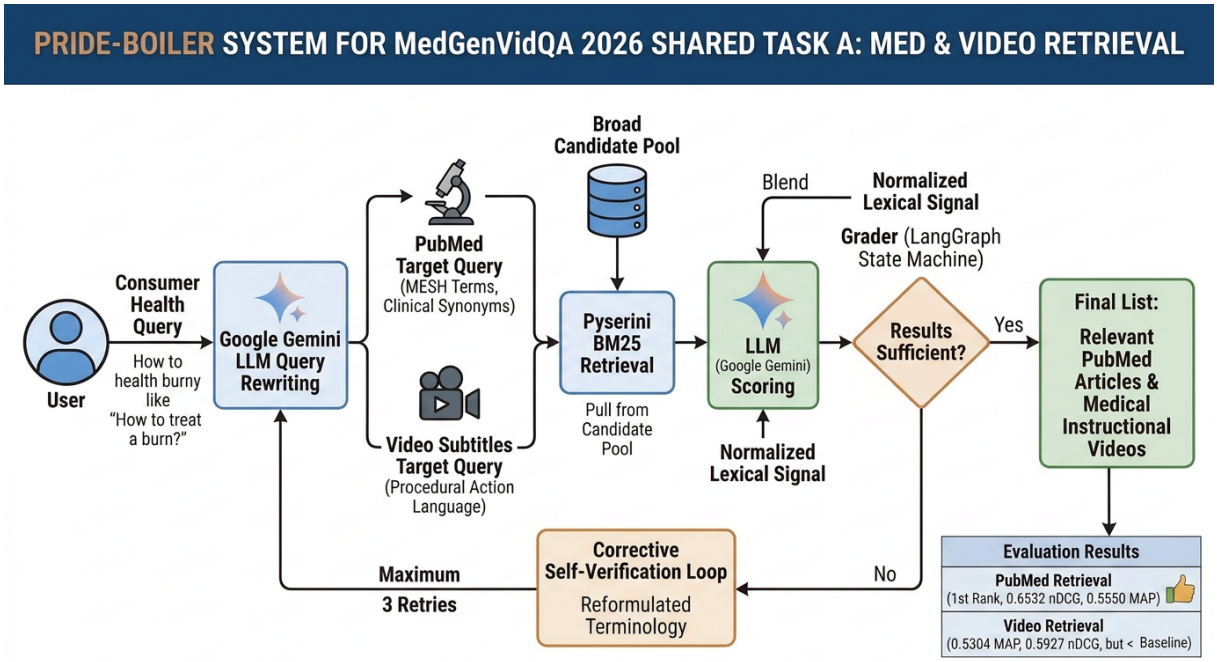

Figure 1: Workflow of the Pride-Boiler's system. Generated with Google Gimini.

Proceedings of the BioNLP 2026 (Shared Tasks), pages 248–256

1 Introduction

The volume of biomedical knowledge published each year far exceeds what any individual can absorb. PubMed alone indexes over 36 million abstracts (Lu, 2011), and the number of medical instructional videos hosted on open platforms continues to grow rapidly (Gupta et al., 2023). For clinicians seeking the latest evidence on a treatment protocol, or for patients trying to understand a procedure they have been recommended, the challenge is no longer the access to information, but finding the most relevant pieces of information amid an overwhelming surplus.

The MedGenVidQA 2026 Shared Task A (Gupta et al., 2026) formalizes these challenges into a concrete benchmark. Given a natural-language medical question, participating systems are asked to return the ten most relevant PubMed articles and the ten most relevant medical instructional videos from 2 corpora: one containing over 28 million abstracts and another with 57,166 video entries (Gupta et al., 2025). The task is designed to approximate the real-world scenario in which a user searches for both textual evidence and visual demonstrations of a clinical procedure.

In this paper, we present our Pride-Boiler system submitted to Task A. Rather than training a domain-specific neural retriever which would demand substantial labeled data, embedding infrastructure, and compute, we take a deliberately lightweight approach. We retain Best Match 25 (BM25) as the retrieval backbone, served through Pyserini (Lin et al., 2021), and augment it with a large language model at two critical junctures: before retrieval, where the LLM rewrites each consumer query into clinically precise BM25 search strings with rich MeSH terms and procedural terminology; and after retrieval, where the LLM scores each candidate document against the original query, promoting articles it judges to be genuinely relevant over those that are merely topically adjacent.

The pipeline is further equipped with a corrective self-verification loop inspired by the Corrective Retrieval-Augmented Generation (CRAG) paradigm (Yan et al., 2024). After scoring, a separate LLM call evaluates whether the top results are relevant to the query. If the verdict is insufficient, the system loops back with the grader's explicit failure reason which is not a generic retry, but a targeted steer toward different terminology and re-executes retrieval from the scratch. The entire workflow is implemented as a four-node LangGraph state machine: query transformation, BM25 retrieval, LLM reranking, and quality grading. The corrective edge between the grader and the transformer closes the loop, capped at three attempts to bound latency.

The central question behind this design is whether a capable LLM can compensate for the well-known semantic limitations of BM25 bridging the lay-to-clinical vocabulary gap on the query side, and separating clinically specific evidence from merely related material on the scoring side without requiring dense retrieval, fine-tuned encoders, or domain-specific training data.

2 Related Work

Biomedical IR has long relied on lexical retrieval. BM25 (Robertson and Zaragoza, 2009) remains competitive in domain-specific settings like BioASQ (Tsatsaronis et al., 2015), where clinical terminology creates a vocabulary that dense models must be explicitly trained to handle. Dense passage retrieval (Karpukhin et al., 2020) has narrowed the gap on general benchmarks, but requires substantial labeled data that is scarce in the biomedical setting. Pyserini (Lin et al., 2021) has made BM25 the de facto reproducible baseline in the community. Our system keeps BM25 as the retrieval backbone precisely because it does not require training, and invests the complexity budget elsewhere.

Query reformulation addresses the vocabulary mismatch between consumer questions and clinical literature. Hypothetical Document Embedding (Gao et al., 2023) encodes an LLM-generated answer as a dense query vector; we take a simpler route, prompting the LLM to produce a clinical keyword string directly for BM25, avoiding the embedding infrastructure HyDE requires. Prior neural query rewriting work operates on formal queries; our setting is harder because the input is lay consumer language, not a clinician's search string.

Retrieval-augmented generation (Lewis et al., 2020) grounds LLM outputs in external evidence. CRAG (Yan et al., 2024) introduced the corrective loop we adapt here: when retrieved results are judged insufficient, the system reformulates and retries. Self-RAG (Asai et al., 2024) pursues a similar self-evaluation idea but interleaves it with text generation. Video retrieval for medical content is a younger problem. MedVidQA (Gupta et al., 2023) introduced the instructional video setting that

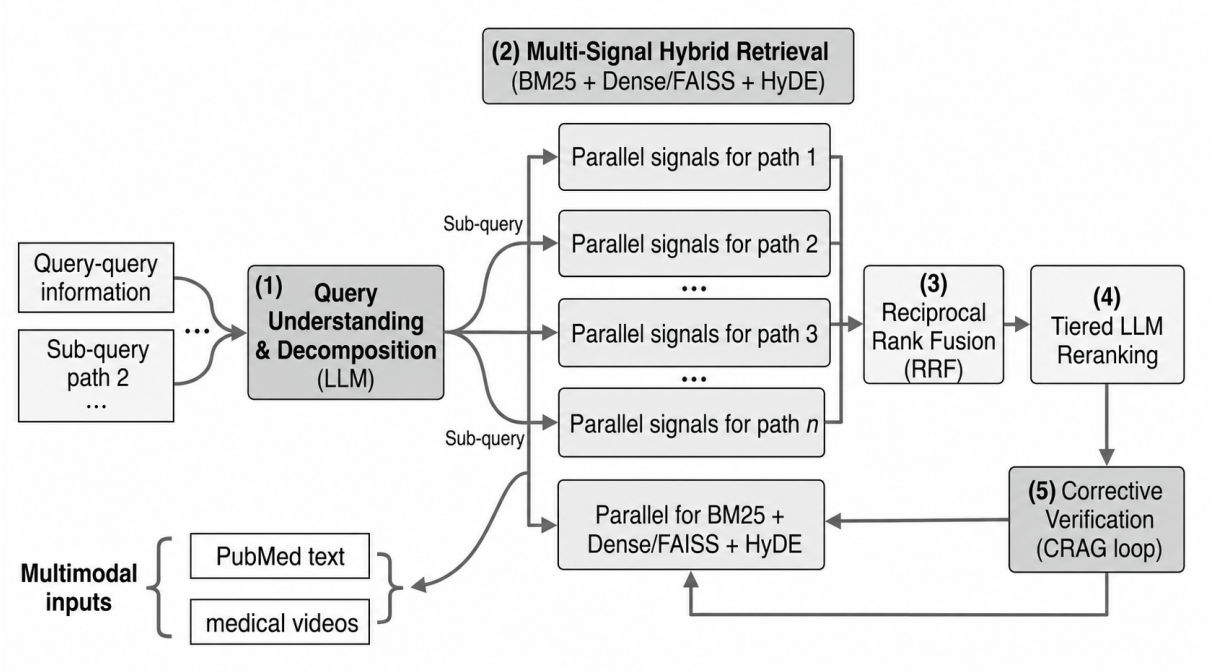

Figure 2: Architecture of the Pride-Boiler retrieval pipeline. A medical query enters the **Query Understanding & Decomposition** node (1), which uses an LLM to rewrite it into BM25-optimized and video-optimized search strings. These are issued to PubMed and video BM25 indexes via **Multi-Signal Retrieval** (2). Candidates are fused (3) and scored by the LLM in a **Tiered Reranking** step (4). A **Corrective Verification** node (5) evaluates result quality: if deemed insufficient, the pipeline loops back to node 1 with a reformulated query, up to three attempts.

the MedGenVidQA 2026 task (Gupta et al., 2026, 2025) extends to joint textual and video retrieval. No prior system has tackled this joint setting; existing work either addresses PubMed retrieval or video understanding in isolation. The subtitle-text challenge we encounter—spoken, fragmented, procedurally oriented—has no direct precedent in prior biomedical IR, which motivates the modality-specific query strategy we develop.

3 System Architecture

Figure 1 illustrates the workflow of the Pride-Boiler system with the functional units and data flow. Figure 2 presents an architectural overview of the Pride-Boiler pipeline, which is implemented as a LangGraph state machine with four processing nodes and a conditional retry edge.

3.1 Node 1: Query Transformation

The raw consumer query is passed to Google Gemini[1] with a prompt that requests two optimized search strings. The first, `BM25_QUERY`, targets PubMed abstracts and emphasizes MeSH terms, clinical synonyms, anatomical terms, and drug or procedure names, restricted to 8–12 keywords with no filler. The second, `VISUAL_QUERY`, targets medical video subtitles and emphasizes step-by-step procedural language and action-oriented phrases.

On retry attempts triggered when the CRAG grader judges the previous results insufficient, the prompt additionally includes the grader's failure reason (e.g., *"results focused on pediatric indications rather than adult surgical approaches"*), and instructs the LLM to use completely different medical terminology. This explicit feedback mechanism ensures that successive attempts do not simply repeat the same search query. Appendix A provides a concrete example of the query transformer output and retrieved results for topic A1.

3.2 Node 2: Retrieval

Both optimized queries are issued to their respective indexes across two retrieval signals. For PubMed, we run a BM25 search over a Pyserini Lucene index of 28,372,706 abstracts with $k_1 = 0.9$ and $b = 0.4$, retrieving $k = 1000$ candidates per query.

For video retrieval, we use a dense FAISS index over subtitle embeddings encoded with SapBERT (Liu et al., 2021), a transformer pre-trained

[1]We used `gemini-3-flash-preview` throughout the pipeline.

on biomedical entity synonyms via self-alignment with UMLS concepts. Subtitle texts are encoded into 768-dimensional vectors and stored in a FAISS `IndexFlatIP` for cosine similarity search, retrieving $k = 1000$ candidates per query. We chose SapBERT over a purely lexical approach for video because subtitle text is short, spoken, and procedurally oriented conditions under which BM25 term overlap is unreliable.

In addition to BM25 and dense retrieval, we incorporate a Hypothetical Document Embedding (HyDE) signal (Gao et al., 2023) as a third parallel retrieval path. For each query, Gemini generates a short synthetic PubMed abstract that would constitute an ideal answer, using clinical vocabulary, MeSH terms, and procedure names. With the same SapBERT encoder used for the dense index, this hypothetical abstract is encoded as a dense vector against the FAISS index query. The intuition is that a synthetic clinical document sits closer to the relevant literature in the embedding space than the original consumer query, bridging the lay-to-clinical vocabulary gap at the vector level. The BM25, dense, and HyDE ranked lists are then fused using Reciprocal Rank Fusion (Cormack et al., 2009) before being passed to the reranker.

3.3 Node 3: LLM Reranking

This node is the primary quality filter. The top 100 BM25 candidates are batched into a single prompt with each document's ID and a 300-character content snippet. The LLM scores each item on a 0.0–1.0 scale against the original query, where 1.0 indicates a document that directly answers the query and 0.0 indicates a completely irrelevant one.

The final score for each candidate is a weighted blend:

$$s = 0.7 \times s_{\text{LLM}} + 0.3 \times \frac{s_{\text{BM25}}}{s_{\text{BM25}}^{\max} + \epsilon}, \qquad (1)$$

where $s_{\text{LLM}} \in [0, 1]$ is the relevance score assigned by the LLM, s_{BM25} is the raw BM25 score for the candidate, $s_{\text{BM25}}^{\max}$ is the highest BM25 score in the batch, and $\epsilon = 10^{-9}$ prevents division by zero when all BM25 scores are zero. The division by $s_{\text{BM25}}^{\max}$ normalises lexical scores into $[0, 1]$ so they are directly comparable to s_{LLM}.

The 0.7/0.3 weighting reflects a deliberate design choice: the LLM signal is the primary judge of clinical relevance, while BM25 acts as a tiebreaker among candidates the LLM scores similarly. A pure LLM score would discard the lexical signal entirely; a pure BM25 score would undo the semantic reasoning the LLM provides. The 0.7/0.3 split was set empirically on a small held-out sample of 10 queries drawn from the training set, where it produced the highest mean nDCG@10 among the splits $\{0.5/0.5, 0.6/0.4, 0.7/0.3, 0.8/0.2\}$ we tested. Candidates are sorted by s and the top 10 are retained per modality.

Metric	Docs	Videos
Mean top-1 score	0.926	0.823
Top-1 ≥ 0.90	53/60	43/60
Top-1 ≥ 0.80	58/60	49/60
Perfect top-1 score (1.00)	3/60	3/60
Score drop (rank 1 $\rightarrow$ rank 10)	0.118	0.404

Table 1: Relevance score statistics across 60 submitted topics. "Score drop (rank 1 $\rightarrow$ rank 10)" is the average difference between the relevance score of the top-ranked result and the tenth-ranked result, computed per topic and averaged over all 60 topics. A smaller value indicates that the system returns a uniformly confident shortlist rather than a single strong result followed by weaker candidates.

3.4 Node 4: Quality Grading (CRAG Loop)

Before finalizing the results, a separate Gemini call inspects the top three documents and top three videos against the original query. The grader assigns one of three labels:

RELEVANT: At least one top result directly addresses the query. The pipeline terminates and returns the results.

INCOMPLETE: Results are partially related but miss key aspects. If the attempt count is below three, control returns to Node 1 with the grader's reason.

IRRELEVANT: All results are off-topic. Same retry logic applies.

4 Experiments

4.1 Submission Output Quality

The shared task requires exact top ten ranked PubMed documents and top ten ranked videos per query, each accompanied by a relevance score in $[0, 1]$. Table 1 summarizes the score distributions across all 60 evaluation topics.

The document retrieval performance is strong: the system places a high-confidence article at the

Rank	Doc	Video
1	0.926	0.823
2	0.891	0.700
3	0.869	0.619
4	0.854	0.575
5	0.842	0.549
6	0.834	0.496
7	0.825	0.470
8	0.821	0.441
9	0.814	0.431
10	0.808	0.420

Table 2: Mean relevance scores by rank position, averaged over 60 topics. Document scores degrade gradually; video scores decline more noticeably after Rank 3.

top for 88.3% of queries, and the gentle score decline across ranks (mean score drop 0.118) suggests that the reranker is producing a coherent, well-ordered shortlist rather than a noisy grab-bag. Three topics achieved a perfect document relevance score of 1.0 at rank 1 (A34, A36, A37), all involving highly specific procedural queries where the BM25 query transformation precisely matched the target literature. A separate set of three topics achieved a perfect video relevance score of 1.0 at rank 1; notably, only topic A37 appears in both groups, suggesting that strong document retrieval and strong video retrieval are largely independent outcomes driven by different corpus characteristics.

4.2 Score Progression Across Ranks

Table 2 reports the mean relevance scores at each output position, averaged over all 60 topics.

Document scores remain above 0.80 even at rank 10, indicating that the entire returned list tends to be relevant, not just the top few items. Video scores tell a different story: they drop from 0.823 at rank 1 to 0.420 at rank 10, with a particularly steep cliff between ranks 3 and 6. This is consistent with the nature of the video corpus many medical topics simply do not have ten highly relevant instructional clips available, particularly for specialized procedural queries like *"How do you perform endoscopic transsphenoidal salvage surgery?"* (topic A37). In those cases, the system fills the lower ranks with the best available candidates, which are only tangentially related.

4.3 The CRAG Loop in Practice

The corrective loop fires when the quality grader judges the initial retrieval insufficient. Across our 60-topic run, the majority of queries achieved RELEVANT status within one or two attempts. The retry mechanism proved most valuable for queries with ambiguous lay terminology, where the first BM25 query missed the right clinical angle. For example, a query about *"crowded arteries"* (topic A20) initially retrieved literature on arterial crowding in dental contexts; the grader flagged this, and the reformulated query correctly targeted coronary artery disease and atherosclerotic stenosis.

5 Implementation Details

The entire pipeline runs on a single RunPod instance equipped with an NVIDIA RTX 6000 GPU (98 GB VRAM), 188 GB system RAM, and 16 vCPUs, running Ubuntu 24.04 with Python 3.10 (RunPod PyTorch image `1.0.2-cu128-torch280`). The PubMed BM25 index occupies roughly 15 GB on disk and the video index under 400 MB. All LLM calls use Gemini 3 Flash Preview via the Google GenAI API. BM25 search is handled by Pyserini 1.2.0 with OpenJDK 21. The pipeline is orchestrated with LangGraph, which manages the state machine transitions and the conditional retry edge.

Per-query wall-clock time is dominated by the LLM calls, and each query transformation takes roughly 1–2 seconds. BM25 retrieval responds in less than a second, and a batch reranking call finishes in 5–15 seconds depending on the candidate count. A single pass through the pipeline completes in approximately 10–20 seconds; queries that trigger the CRAG retry loop take proportionally longer, up to about a minute for the maximum three attempts.

6 Results and Discussion

The official evaluation results are summarized in Table 3.

On PubMed retrieval, our Pride-Boiler system ranked first among all participating systems, achieving an nDCG of 0.6532 and MAP of 0.5550, both above the Text-RR baseline. This is notable given that the less powerful machinery the system uses: no dense index, no fine-tuned encoder, no learned reranker. The result validates the core hypothesis behind our design that a capable LLM, applied at the query reformulation and candidate scoring stages, can extract substantial relevant documents from a classical BM25 backbone.

Video retrieval tells a different story. Our nDCG of 0.5927 falls below the Video-RR baseline's 0.6616, and the MAP gap is similarly unfavorable.

	MAP	R@5	R@10	P@5	P@10	nDCG
PubMed Article Retrieval						
Baseline	0.5404	0.5505	0.5863	0.5133	0.2700	0.6460
Ours	**0.5550**	**0.5571**	**0.5866**	**0.5333**	**0.2817**	**0.6532**
Instructional Video Retrieval						
Baseline	**0.5884**	**0.6067**	**0.6528**	**0.4100**	**0.2217**	**0.6616**
Ours	0.5304	0.5478	0.5833	0.3900	0.2100	0.5927

Table 3: Performance of our system in comparison with the Baseline results. Boldfaced numbers indicate the higher values within each modality.

This is the clearest evidence that neither lexical nor entity-centric dense retrieval fully addresses the subtitle matching problem. BM25 over subtitle text has a structural ceiling that query reformulation alone cannot raise: subtitles are short, fragmented, often auto-transcribed, and written in spoken register, so even a well-crafted procedural query can only match videos that happen to contain the exact terms. The dense index, encoded with SapBERT (Liu et al., 2021), does not resolve this. SapBERT was pre-trained on biomedical entity synonyms via UMLS concept alignment, making it strong at matching clinical terminology across surface forms for example, recognising that *myocardial infarction* and *heart attack* refer to the same concept. Subtitle text, however, is not a collection of biomedical entities; it is procedural narration (*"now we retract the flap and identify the vessel"*), and SapBERT has no training signal for this register. The result is that both retrieval signals fail for the same underlying reason: the query and the subtitle live in different linguistic worlds, and neither BM25 term overlap nor entity-level embedding alignment can bridge that gap without a model trained specifically on procedural spoken language. The pattern across modalities is instructive. PubMed abstracts are long, terminologically dense, and written in standardized clinical prose exactly the kind of text BM25 handles well once the right query terms are in place. Video subtitles violate nearly every assumption BM25 makes about document structure. The performance gap is not a failure of the pipeline design but a clean signal about where denser retrieval signals are needed.

The CRAG corrective loop proved most valuable on queries with ambiguous lay terminology, where the first reformulation landed in the wrong clinical neighborhood. The grader's failure reason gave the transformer a concrete steer on retry, not just "try again", but "the results discussed X when they should address Y." That said, the loop has a hard ceiling: when ground-truth documents never enter BM25's candidate pool, no amount of reformulation will surface them.

7 Error Analysis

Across 60 topics, the system places a highly relevant document (grade 2) at rank 1 for 38 queries and a partially relevant document for a further 15. The remaining 7 topics return an irrelevant result at rank 1, and 4 topics A15, A44, A49, and A56 return zero relevant documents in the entire top-10 (*FP2: missed top-ranked documents* (Barnett et al., 2024)). Inspection of these four queries reveals a consistent pattern: the query transformer confidently reformulated toward the broader condition using incorrect clinical terminology, steering BM25 toward adjacent literature rather than the target. Topic A44, for example, uses lay phrasing that maps to a narrow clinical subterm; the query transformer confidently reformulated toward the broader condition, and the relevant abstracts which use the specific subterm never entered the candidate pool. No amount of CRAG retry can recover from this: once the wrong clinical neighborhood is established in the first reformulation, subsequent attempts explore nearby but equally incorrect regions. A second failure mode affects 7 topics (A10, A16, A17, A19, A25, A51, A57), which retrieve some relevant documents but miss more than half the available relevant set. These are multi-faceted queries with both pharmacological and procedural dimensions. The LLM transformer consistently resolves the ambiguity by committing to one clinical angle usually the more salient one while the documents covering the secondary angle rank below the top-10 cutoff. Recall here is structurally capped by query specificity. The most pervasive issue cuts across all categories: on average, 6.77 of the 10 retrieved documents per topic are unjudged meaning they appear nowhere in the relevance pool assembled by the task organizers.

These are not necessarily irrelevant; they may be genuinely useful documents that assessors never evaluated. Under standard TREC scoring, however, unjudged documents count as non-relevant, which suppresses all precision metrics. This gap reflects a fundamental tension between a system that searches a 28-million-document corpus freely and an annotation pool necessarily bounded by human annotator effort.

Video retrieval fails more severely: 13 of 60 topics return zero relevant videos, and the top-ranked result is irrelevant in 18 topics nearly three times the document rate. The root cause is a linguistic register mismatch that query reformulation cannot bridge. BM25 is built on term overlap. PubMed abstracts and the clinical queries we generate share vocabulary by design. Subtitle text does not. A reformulated query like *uvulopalatopharyngoplasty obstructive sleep apnea surgical outcomes"* finds nothing useful in a subtitle corpus where the same procedure is described as *removing tissue from the back of the throat to widen the airway."* The query transformer generates clinical prose; subtitles speak in lay procedural language. The gap is not a vocabulary mismatch in the abstract sense it is a mismatch in the very register of the text. A secondary cause is corpus sparsity. For specialized procedural queries robotic surgery variants, rare metabolic conditions, niche anatomical procedures fewer than ten relevant instructional videos may exist in the entire 57,166-entry corpus. The system correctly identifies the available relevant videos at top ranks for these queries, but must fill lower positions with tangentially related content. This is not a retrieval failure; it is a coverage limitation of the corpus that no retrieval algorithm can overcome. Finally, the LLM reranker operates on 300-character subtitle snippets. For PubMed abstracts, 300 characters reliably captures the core claim of a structured abstract. For subtitles, 300 characters may span a single procedural instruction *"now we retract the flap and identify the vessel"* with no indication of the broader procedure being demonstrated. The reranker has insufficient context to judge video relevance accurately, and in 18 topics it ranks an irrelevant video first as a result.

8 Conclusion

We presented Pride-Boiler, a retrieval system for MedGenVidQA 2026 Shared Task A that demonstrates how the performance of a classical BM25 backbone can improve when augmented with LLM-driven query transformation and corrective self-verification. The system is deliberately simple, with four nodes in a LangGraph state machine, no fine-tuned encoders, no learned reranker, and dense retrieval limited to video subtitle embeddings, and yet it ranked first among all participating teams on textual retrieval. The architecture is modular by design: each node can be replaced or extended independently, and the corrective loop generalizes beyond the specific retrieval backend. We view this submission less as a finished system and more as a baseline that establishes what lightweight LLM augmentation can achieve on its own, so that future work on denser retrieval components can be measured against a clear reference point.

Looking ahead, the most immediate priority is stronger retrieval for the video modality, where BM25 over subtitle text has a structural ceiling that query reformulation alone cannot raise. We plan a comparative evaluation of three biomedical encoders MedCPT (Jin et al., 2023), BMRetriever (Xu et al., 2024), and BiCA (Sinha et al., 2026) fused with BM25 via Reciprocal Rank Fusion (Cormack et al., 2009) to recover ground-truth documents that lexical retrieval misses entirely. Three further extensions follow naturally from the error analysis: replacing the 300-character snippet reranker with a cross-encoder over full text, decomposing compound queries into modality-specific sub-queries, and replacing the CRAG grader's three-way categorical label with a continuous confidence score for more calibrated retry decisions.

Limitations

Three limitations bear directly on the results reported here. First, For PubMed retrieval, the system relies entirely on BM25 for its initial candidate pool. When ground-truth documents use vocabulary that diverges from even a well-crafted clinical query—as in the four complete-miss topics no reranking or CRAG retry can surface them. Dense retrieval over document embeddings would mitigate this by matching semantic intent rather than surface terms, but we deliberately excluded it to test how far a lexical system could go when paired with a capable LLM.

Second, the single-query strategy applies the same reformulated clinical string to both PubMed and video subtitle retrieval. PubMed abstracts reward clinical precision. Subtitle text rewards

procedural, action-oriented lay language. Using one query for both modalities guarantees that one of them is systematically mismatched. A modality-specific query generator producing a separate action-oriented search text for video would directly address the 13 complete-miss topics in the video track.

Third, the LLM reranker scores candidates from a 300-character snippet. For PubMed abstracts this is usually sufficient; the opening sentences of a structured abstract convey the central claim. For subtitle chunks it is rarely sufficient: 300 characters of spoken procedural narration lack the context to determine what intervention the video is actually demonstrating. A cross-encoder operating over full transcript text, or over title and metadata fields, would give the reranker the context it currently lacks.

Ethics Statement

This system retrieves published biomedical literature and educational instructional videos for informational purposes. It is not intended as a clinical decision-support tool and should not be used to make medical decisions without professional guidance. All corpora are publicly available. LLM-generated content (query rewrites, relevance scores, quality judgments) is used exclusively to guide retrieval; no LLM output is presented to users as clinical evidence.

Acknowledgments

We thank Dr. Deepak Gupta and Dr. Dina Demner-Fushman from the National Library of Medicine for organizing the MedGenVidQA 2026 Shared Task and for providing the evaluation infrastructure and datasets. We are especially grateful to Dr. Gupta for his prompt and generous responses to our questions throughout the manuscript preparation process. We also thank the developers of Pyserini and LangGraph for the open-source tooling that underpins this system.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the International Conference on Learning Representations*.

Scott Barnett, Stefanus Kurniawan, Srikanth Thudumu, Zach Brannelly, and Mohamed Abdelrazek. 2024. Seven failure points when engineering a retrieval augmented generation system. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering – Software Engineering for AI (CAIN 2024)*, Lisbon, Portugal.

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 on Research and Development in Information Retrieval*, pages 758–759.

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.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. 2023. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1):158.

Deepak Gupta, Davis Bartels, and Dina Demner-Fushman. 2025. A dataset of medical questions paired with automatically generated answers and evidence-supported references. *Scientific Data*, 12(1):1035.

Deepak Gupta, Collin Scott Campbell, Pedram Golnari, and Dina Demner-Fushman. 2026. Overview of the MedGenVidQA 2026 shared task on medical generative video question answering. In *Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026)*, San Diego, USA. Association for Computational Linguistics.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. 2023. MedCPT: Contrastive pre-trained transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16649–16664. Association for Computational Linguistics.

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.

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.

Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python toolkit for reproducible information retrieval research with sparse and dense representations. *SIGIR*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4228–4238.

Zhiyong Lu. 2011. PubMed and beyond: A survey of web tools for searching biomedical literature. *Database*, 2011:baq036.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Aarush Sinha, Pavan Kumar S, Roshan Balaji, and Nirav Pravinbhai Bhatt. 2026. BiCA: Effective biomedical dense retrieval with citation-aware hard negatives. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33010–33018.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, and 1 others. 2015. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16(1):138.

Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Yanqiao Zhu, May Dongmei Wang, Joyce C. Ho, Chao Zhang, and Carl Yang. 2024. BMRetriever: Tuning large language models as better biomedical text retrievers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22234–22254, Miami, Florida, USA. Association for Computational Linguistics.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. 2024. Corrective retrieval augmented generation. *Computing Research Repository*, arXiv:2401.15884.

A An Example Query Decomposition

For topic A1 *"What surgical approaches are effective for reducing symptoms in adult patients with obstructive sleep apnea? How are they performed?"* the query transformer produced the following optimized search strings:

```
{
  "original_query": "What surgical approaches
     are effective
    for reducing symptoms in adult patients with
      obstructive
    sleep apnea? How are they performed?",
  "BM25_QUERY": "uvulopalatopharyngoplasty
     maxillomandibular
    advancement hypoglossal nerve stimulation
     surgical
    treatment obstructive sleep apnea adults
     outcomes",
  "VISUAL_QUERY": "surgical procedure sleep
     apnea palate
    tissue removal airway widening step-by-step
     operating
    room technique demonstration"
}
```

The `BM25_QUERY` correctly introduced clinical procedure names (uvulopalatopharyngoplasty, maxillomandibular advancement, hypoglossal nerve stimulation) absent from the original consumer question this reformulation is what enables BM25 to bridge the vocabulary gap. The pipeline returned the following top-3 results per modality for this topic:

```
{
  "query_id": "A1",
  "pubmed_results": [
    {"id": "33418179", "score": 0.9307},
    {"id": "20451028", "score": 0.9255},
    {"id": "19784401", "score": 0.9199}
  ],
  "video_results": [
    {"id": "PMC3197034_jove-52-2652.mp4", "score
     ": 0.9170},
    {"id": "PMC5755377_jove-129-56369.mp4", "
     score": 0.8566},
    {"id": "PMC2841570_jove-35-1437.mp4",  "
     score": 0.8339}
  ]
}
```

The sharp score drop after rank 3 in the video results is consistent with the corpus sparsity discussed in the error analysis: highly relevant surgical demonstration videos for obstructive sleep apnea are sparse in the 57,166-entry corpus.