From Retrieval to Verdict: A Hybrid LLM Pipeline for Evaluating Medical and Economic Claims

Acceptance code

A Livieratos[1], M Kudela[1,2], Y Zhao[1,2], A Chen[1,2], J Lin[1,3], D Zhang[1,4], X Luo[1,2], PA Ramos[1,2], S Dharmarajan[1,5], C Su[2], M Gamalo[1,2]

[1]SPAIML Scientific Working Group, USA; [2]Pfizer, USA; [3]Takeda Pharmaceuticals, USA; [4]Teva Pharmaceuticals, USA; [5]Sarepta Therapeutics, USA

INTRODUCTION

Health researchers face a time-consuming burden when manually verifying clinical and economic claims in literature. Systematic reviews and health technology assessments demand meticulous multi-reviewer scrutiny of large text corpora.[1] Early experiments have explored using GPT-4 to automate parts of this process. For example, GPT-4 matched human reviewers in screening abstracts during a PRISMA systematic review, suggesting it could replace one human screener.[1-2] However, naïve LLM evaluators remain imperfect: in one study GPT-4's agreement with human quality appraisal of case studies was only moderate, indicating current models aren't yet rigorous enough for full autonomy.[1-2]

Recent research is moving beyond one-shot GPT-4 evaluations toward more sophisticated "LLM-as-judge" models.[3] These approaches combine retrieval augmentation and iterative self-critiquing to improve factual accuracy. For instance, AlignRAG introduced a Critic LLM that iteratively refines answers via evidence-based critiques, actively aligning the reasoning with retrieved evidence.[3] Likewise, the TextGrad framework treats the LLM's feedback as a "gradient" – the model evaluates its output, criticizes flaws, and updates its response in a loop.[4] Such pipelines hold promise for expert-level claim verification in medical and health economics research, where answers must be not only correct but also well-supported by cited literature.

OBJECTIVE

This study aimed to design and evaluate a hybrid AI judge model for structured evaluation of medical and economic claims. Our primary objectives were:

- **Automate Evidence-Based Claim Review**
  Replace or augment manual screening and appraisal by integrating retrieval, reasoning, and critique steps in a single pipeline.
- **Improve Judgment Accuracy via Iterative Critique**
  Apply TextGrad-refinement to generate structured verdicts (TRUE / PARTLY TRUE / FALSE) that align more closely with PubMed evidence.
- **Test Cross-Domain Robustness**
  Validate the method across diverse medical claims (e.g., efficacy, safety, cost-effectiveness) to assess generalizability beyond a single disease.
- **Enforce Citation Transparency**
  Require every output to justify its verdict using explicitly cited PubMed IDs, increasing the accountability and verifiability of model reasoning.

METHODS

We developed a hybrid Retrieval-Augmented Generation (RAG) + TextGrad pipeline to serve as an automated judge model (Figure 1) for scientific claims. Key components of the method include:

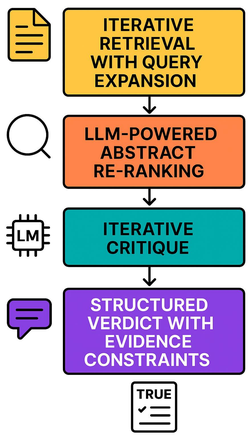
•**Iterative Retrieval with Query Expansion:** Instead of relying on a static query, it dynamically expands the search space by encoding the original claim and selecting the most semantically similar medical concept from a predefined list of candidate terms (e.g., drug classes). In each iteration the system uses a LLM (DeepSeek-R1) to generate new, focused search queries based on the most relevant previously retrieved abstracts, allowing the pipeline to uncover pivotal RCTs, head-to-head studies, and real-world, observational studies supporting the claim.[3,5]

•**LLM-Powered Abstract Re-ranking (DeepSeek-R1):** To improve evidence quality, the pipeline re-ranks retrieved PubMed abstracts using DeepSeek-R1 based on scientific relevance to the original claim. This LLM-based analysis emulates expert judgment and ensures that highly pertinent references—such as high-quality randomized trials or cost-effectiveness studies—are prioritized over less relevant articles like general reviews or observational reports.

•**Iterative Critique via TextGrad:** An initial verdict on the claim is generated by an LLM given the top retrieved evidence. We then refine this output through TextGrad-based iterative feedback. In each iteration, a critic model analyzes the answer while its citations detect any reasoning misalignments or missed evidence.[4] This process is repeated, guided by a relevance-weighted objective that penalizes unsupported statements and rewards alignment with the cited evidence. Over successive critique rounds, the claim evaluation becomes more refined and evidence-grounded.[4]

•**Structured Verdict with Evidence Constraints:** The final output is constrained to a structured format: a categorical judgment (TRUE, PARTLY TRUE, or FALSE) accompanied by supporting references (identified by PubMed ID). The model must justify its decision using the retrieved literature, citing specific PMIDs as evidence. This structured approach imposes discipline on the LLM's output – it mirrors frameworks in scientific fact-checking that label claims as supported, refuted or neutral based on literature evidence.[6] Crucially, the judge model is not free to generate open-ended text; it must produce a verdict and rationale that directly reference the retrieved documents, enforcing accountability and verifiability in its evaluation.

Figure 1: Structured Medical Claim Evaluation



RESULTS

The LLM-powered, relevance, distribution plot (Figure 2) showed that most retrieved studies were recent—published between 2023 and 2025—highlighting the system's focus on up-to-date evidence. However, a subset of the retrieved articles (approximately 8%) were non-relevant to the original claim, covering unrelated topics. This confirms the mixed-quality retrieval landscape common in IBD-related queries and reinforces the importance of re-ranking and query reformulation. This initial heterogeneity was substantially mitigated by the downstream TextGrad judge model (Table 1), which refined the noisy abstract set into structured, citation-backed verdicts specific to gastroenterology treatment comparisons.

For each claim we manually provided, the model iteratively refined its output, shifting from coarse initial assessments to nuanced final judgments. Importantly, each verdict was constrained to cite specific PubMed IDs, enabling transparent clinical reasoning. This layered architecture—combining augmented retrieval, LLM-based reranking, and TextGrad optimization—demonstrates how progressive stages can convert a semi-relevant document set into precise, evidence-aligned decision support for IBD.

The result is a reproducible framework for AI-led claim adjudication, grounded in trial data and real-world risk-benefit profiles in gastroenterology.

Figure 2: Retrieved Abstracts

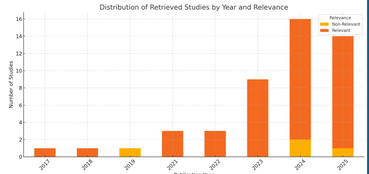


Table 1: Verdict Examples across Medical Claims

Claim	Final Verdict	Justification Bullet
In moderate-to-severe ulcerative colitis, JAK inhibitors should always be used before any TNF-alpha inhibitor to maximize mucosal healing rates.	PARTLY TRUE	JAK inhibitors demonstrate superior mucosal-healing rates vs some TNF inhibitors in specific populations (PMID 40102788, 39894450).
	PARTLY TRUE	The "always first" recommendation is unsupported because efficacy varies across agents and safety concerns exist (PMID 38659488, 38399538).
	PARTLY TRUE	Thrombosis risks with JAK inhibitors contraindicate universal first-line use per FDA warnings (PMID 40407729).
	PARTLY TRUE	Guidelines prioritize individualized therapy over fixed sequencing (PMID 40398832, 40386337).
IL-23 inhibitors provide significantly better sustained remission than IL-12/23 inhibitors and should replace them as preferred therapy in biologic-naïve patients.	PARTLY TRUE	IL-23 inhibitors show superior sustained-remission rates vs IL-12/23 inhibitors in pivotal trials (PMID 40102788, 38651771).
	PARTLY TRUE	Limited long-term comparative safety data exist (PMID 40398832, 40386337).
	PARTLY TRUE	Cost-effectiveness analyses favor IL-12/23 inhibitors in some healthcare systems (PMID 39894450).
	PARTLY TRUE	Insufficient evidence for universal replacement until more real-world outcomes is available (PMID 38399538).
S1P modulators are safer and more effective than JAK inhibitors and should be used as first-line oral therapy for ulcerative colitis.	PARTLY TRUE	S1P modulators show lower thromboembolic risk than JAK inhibitors (PMID 40102788, 39894450) but comparable infection risks in some studies (PMID 38823040).
	PARTLY TRUE	Efficacy is similar between classes for moderate UC (PMID 38659488, 38651771), though S1P shows advantages in specific subgroups (PMID 38399538).
	PARTLY TRUE	Guidelines position both classes as alternatives—not first-line standards—due to variable individual responses (PMID 37820832, 40407729).
	PARTLY TRUE	Long-term data support personalized selection over universal first-line use (PMID 40398832, 40386337).

CONCLUSIONS

This study presents a novel hybrid LLM pipeline that integrates iterative retrieval with query expansion, open-source LLM-based abstract re-ranking (DeepSeek-R1), and multi-hop critique via TextGrad to automate the evaluation of medical and economic claims in recent medical literature, with a specific focus on treatment comparisons in ulcerative colitis.

By combining retrieval, re-ranking, and structured verdict enforcement, the system transforms noisy literature sets into citation-backed judgments.

This framework has accountability and transparency of automated claim adjudication, important for AI work, particularly in complex disease areas such as ulcerative colitis.

These findings suggest that multi-agent LLM systems are not only viable replacements for manual reviewers but can also enhance the speed and reliability of evidence synthesis for HTA, regulatory, and market access decisions.

By analyzing data derived from the healthcare environment and streamlining operations in medical content generation, AI advances data-based prioritization and evaluation.[7]

Future research should address computational trade-offs and explore cross-disease generalizability. Extension of this pipeline to query trial registries and grey literature and apply bias-aware LLM critique—could also enable detection and adjustment for publication bias in future projects.

Table 2: Summary of novel AI methodology utilized for judging medical claims

Stage	Core Method
RAG + Query Expansion	**Dynamically reformulates search queries using embedding-guided expansion** • Leverages DistilBERT similarity to inject high-relevance terms (e.g., "JAK inhibitors") • Maximizes retrieval of relevant PubMed abstracts across multi-hop RAG iterations
DeepSeek-R1	• Re-ranks retrieved abstracts • Prioritizes evidence most likely to support or refute claims
TextGrad	• Applies iterative self-critique to improve claim judgments • Produces structured verdicts (TRUE / PARTLY TRUE / FALSE) with cited PubMed IDs

Key Messages

- ***Layered AI Judgment:*** *A hybrid pipeline combining embedding-based query expansion, LLM-powered abstract re-ranking (DeepSeek-R1), and iterative multi-hop critique (TextGrad) produces structured, citation-backed verdicts for medical and economic treatment claims in ulcerative colitis.*
- ***Query Optimization:*** *Dynamic query reformulation using semantic similarity and multi-hop expansion significantly improves retrieval quality, surfacing high-relevance PubMed studies—including pivotal RCTs and real-world comparative evidence.*
- ***Evidence-Backed Verdicts:*** *TextGrad refines initial outputs through iterative feedback and citation validation, delivering transparent, PubMed-anchored judgments (TRUE / PARTLY TRUE / FALSE).*
- ***Domain Scalability:*** *This modular framework is applicable to any therapeutic area—extending to domains like oncology, rare diseases, and health economics—where structured literature verification supports payer negotiations and regulatory filings.*

Limitations:

- **Computational complexity and scalability** – The multi-stage process, iterative retrieval with query expansion, LLM-powered re-ranking, and TextGrad critique incurs substantial compute and latency overhead.
- **Reliance on LLM accuracy** – Both DeepSeek-R1's re-ranking and TextGrad's self-critique depend on the underlying language models' reasoning and factual alignment.

REFERENCES

1. Landschaft A, Antweiler D, Mackay S, Kugler S, Rüping S, Wrobel S, Höres T, Allende-Cid H. Implementation and evaluation of an additional GPT-4-based reviewer in PRISMA-based medical systematic literature reviews. Int J Med Inform. 2024 Sep;189:105531. doi: 10.1016/j.ijmedinf.2024.105531. Epub 2024 Jun 26. PMID: 38943806.
2. Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-bench and chatbot arena. arXiv preprint, arXiv:2306.05685.
3. Wei, J., Zhou, H., Zhang, X., Zhang, D., Qiu, Z., Wei, W., Li, J., Ouyang, W., & Sun, S. (2025). AlignRAG: An adaptable framework for resolving misalignments in retrieval-aware reasoning of RAG. arXiv preprint, arXiv:2504.14858
4. Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. Automatic "Differentiation" via Text (TEXTGRAD). arXiv preprint arXiv:2406.0749
5. Tang, X., Gao, Q., Li, J., Du, N., Li, Q., Xie, S., & Li, J. (2025). MBA-RAG: A Bandit Approach for Adaptive Retrieval-Augmented Generation through Question Complexity. arXiv preprint, arXiv:2412.01572.
6. Liu H, Soroush A, Nestor JG, Park E, Idnay B, Fang Y, Pan J, Liao S, Bernard M, Peng Y, Weng C. Retrieval augmented scientific claim verification. JAMIA Open. 2024 Feb 21;7(1):ooae021. doi: 10.1093/jamiaopen/ooae021. PMID: 38455840; PMCID: PMC10919922.
7. Fröling E, Rajaeean N, Hinrichsmeyer KS, Domrös-Zoungrana D, Urban JN, Lenz C. Artificial Intelligence in Medical Affairs: A New Paradigm with Novel Opportunities. Pharmaceut Med. 2024 Sep;38(5):331-342. doi: 10.1007/s40290-024-00536-9. Epub 2024 Sep 11. PMID: 39259426; PMCID: PMC11473552.

Email: maria.kudela@pfizer.com