Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Grounded ontogenic retrieval improves accuracy and citation fidelity of a medical oncology clinical reasoning AI assistant

Gary Takahashi

gtakahashi@aprilio.ai

Aprilio

Eyoha Mengistu

Aprilio

Ebrahim Tarshizi

Aprilio

Andrew Van Benschoten

Aprilio

Article

Keywords:

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11053038/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Introduction:

Large language models (LLMs) have demonstrated considerable utility in generating clinically informed text to assist physicians, yet uncertainty about the accuracy of their outputs has limited their broader adaptation. General purpose frontier models challenge the necessity of domain-optimized proprietary applications; however, they remain prone to attribution errors because their assertions are drawn from opaque model weights rather than explicitly verifiable references. Tethering model-generated outputs to auditable and verifiable ground truths is essential for establishing clinical confidence and minimizing the risk of medical error. Here, we introduce Grounded Ontogenic Retrieval (GOR), an architecture that couples effective prompt structuring with a harness-driven “ontology-on-the-fly” framework that creates a powerful neuro-symbolic disease management platform based on publicly available curated oncology management guidelines.

Methods:

The system employs a three-step pipeline—intent extraction, targeted source retrieval, and precise localization of relevant facts within retrieved material—implemented through a tailored architectural harness that mediates all input to and output from the core language model. Retrieval is performed via the Brave Search API, with capture restricted to abstracts and query-pertinent paragraphs. ASCO Rapid Recommendations and ESMO Living Guidelines are additionally checked for late-breaking updates. Each retrieved source is scored on four factors: recency, authority (domain tier plus a publication-type bonus), indication match (line-of-therapy alignment), and topical relevance. Topical relevance is assessed primarily by a separate LLM grader that accounts for medical synonymy, with a lexical word-overlap score serving as a fallback. A larger-context model (GLM-5.2) holds the full assembled corpus in-prompt, reinforcing the design principle that inference is driven by grounded retrieval rather than parametric weights. The pipeline also handles intent classification, agent extraction, and approval adjudication for limited-flag and limited-options clinical calls. The performance of this model was compared with that of several frontier and advanced open-source models from a mix of vendors, using an LLM-as-judge methodology. Models were provided online search facility, but received no additional prompting than a presentation of a clinical vignette, mirroring a typical physician use case.

Results:

Model performance was evaluated on a curated set of 100 generated clinical questions designed to drawn from research published within the preceding year. By restricting evaluation to recent literature, the design precludes answers sourced from parametric knowledge acquired during pretraining, thereby isolating and measuring the model's capacity to reason over current, retrieved evidence. The Aprilio model, based on grounded ontogenic retrieval (GOR), scored 100% accuracy on the test panel, with most

other models scoring between 97% and 99%; the open-weight Qwen3 Max scoring 89%. As these models were not prompted specifically to perform an online search, all but one model (GPT-5.6 Terra) reliably did so, although four of the citations it provided were non-existent or erroneous. Perplexity Sonar Pro and Qwen3 Max searched 100% of the time, but left only a citation factor in one and twelve citations, respectively. One model cited a patent-petition document to support an answer. Other models utilized search facility to answer queries for between 35% and 81% of the problems but left almost no locatable citation, making reference to a supportive research study only in prose. Most importantly, the Aprilio model provided locatable links in all of the vignettes (without hallucination or dead links), while all other models, save GPT-5.6 Terra, exhibited substantially lower performance. The Aprilio model also supplied a citation evaluable for claim support in 97 of 100 responses, against 20 for GPT-5.6 Terra and 5 or fewer for the rest. Notably, 55% of responses generated by the Aprilio model carried citations which were confirmed to directly substantiate its assertions, whereas only 15% of citations from the next-best-performing model substantiated its claims. Potential harm could have resulted from a few of the answers, such as by Perplexity Sonar where the recommended agent did not account for critical baseline patient factors, while Qwen3 Max recommended a drug that has been withdrawn from the market due to unacceptable toxicity. In addition to the lack of auditable citations, we found instances of evidentiary misrepresentation in 1 of 14 instances in which claims were made regarding the preference rung of a treatment option according to reference guidelines. While questions were answered correctly, the reasoning behind the recommended selection was faulty. A correct answer can therefore exist beside a misstated authority claim, which is why an MCQ score cannot be read as a statement about open-ended reliability.

Conclusion:

The proposed grounding architecture, encompassing real-time information retrieval, coupled with query processing to extract intent and impose semantic structure, binds each generated clinical statement to a verifiable, auditable source, shifting the selection criterion for assistant performance away from the model's internal factual knowledge and toward its fidelity to curated, up-to-date evidence. A grounded, ontogenic retrieval (GOR) framework offers a more consistent and less error-prone foundation for deploying LLM-based decision support in oncology, directly addressing the reliability concerns that have limited physician adoption of LLM assistants in clinical practice.

INTRODUCTION

AI assistant applications have proliferated in recent years, offering clinicians an efficient means of accessing burgeoning advances in the medical literature. Medical content providers and startups with proprietary models compete alongside general-purpose frontier models to provide clinical decision support.[1] The effectiveness of such assistance, however, remains an open question, as attitudes vary widely regarding their place in the clinical workflow and whether their input meaningfully improves decision-making performance.[2,3]

Recognition of LLM-only limitations has similarly motivated a shift toward grounding model output in retrieved evidence. Retrieval-augmented clinical assistants that return source-attributed answers, benchmarks that systematically evaluate medical RAG, and methods that couple adaptive retrieval with self-reflection have each advanced the field [4–6] and in oncology, retrieval-grounded systems have been applied to guideline-concordant decision support and clinical-trial matching [7,8]. Grounding, however, remains imperfectly realized. Automated auditing demonstrates that a substantial fraction of the statements produced by LLMs are not fully supported by the sources they cite[9], and frontier models cannot reliably reproduce verbatim guideline text on demand, frequently citing hallucinated or outdated information.[10] Knowledge-graph and formal-ontology approaches reduce hallucination but may suffer quality degradation from uncontrolled outsourcing[11], and depend upon curated graphs whose continual maintenance constrains their scope and currency[12]. What remains underexplored is a method that binds each generated statement to a precise, auditable location within authoritative, continually current public sources (including regulatory approval status) without recourse to domain fine-tuning or a hand-maintained knowledge graph.

A more consistent and less error-prone methodology grounds the assistant's output in precise, auditable locations within an authoritative medical source. This entails three steps: extracting the intent of the query, retrieving source material appropriate to that intent, and identifying the exact location of the relevant facts within the retrieved material. By binding each generated statement to a verifiable source, this approach constrains the model to reason over curated evidence rather than its unaudited internal representations.

This grounding methodology is realized architecturally as modern assistant systems employ a tailored harness structure that mediates the input to and output from the core large language model. Consequently, the selection criterion shifts away from the factual knowledge stored in the model's weights and toward capabilities better suited to grounded generation: a sufficient context window, and the capacity to faithfully synthesize text retrieved from disparate sources.

The imposition of structure, across LLM input, user and system prompt generation, custom rule definition within the semantic layer, strict rule enforcement in data retrieval and analysis, and structured output generation, has been shown to enhance accuracy and mitigate hallucination. We describe this staged, cumulative imposition of ontologic structure as its ontogenesis, which we demonstrate can be performed over publicly available data to yield a useful clinical reasoning and decision-support tool.[13,14]

METHODS

AI Assistant Structure

The query submitted by the user should ask a question relevant to hematology or oncology, with the expectation that the answer is deterministic and not speculative, as the AI assistant is designed for

retrieval. An autocorrect spellcheck is available to ensure correct tokenization; the user is able to accept or modify the suggested corrections before submitting.

The processing occurs in two waves. In Wave 1, the query is then analyzed by an upfront LLM stage that performs three checks. First, it looks for clinical "red flags." These red flags include signs where there is discordance between tumor markers and disease burden, suggestive of misdiagnosis or latent disease evolution. Other suggestions of disease transition include evidence of rapid progression, an atypical metastatic pattern suggestive of a second or alternate primary neoplasm, new disease which has manifested in unusual sites, or missing molecular or histologic information that would change management (such as IDH1 status). The presence of a red flag prioritizes confirmation of the diagnosis prior to treatment discussion. Second, the query is analyzed to determine whether the setting is an edge case in which therapy options are limited by clinical factors such as frailty; when detected, this both relaxes the line-of-therapy penalty during source scoring and instructs the synthesizer to give weight to alternative options that would not otherwise be considered. Third, the query intent is classified in order to select appropriate resources for answer retrieval (e.g. supportive care, diagnostic confirmation, or toxicity management).

Llama 3.3 structures a query suitable for searching PubMed via the NCBI API. A separate worker filters the FDA oncology approval feed against this query, producing a setting-level list of candidate approvals. A fan-out retrieval planner makes a planning call that enumerates search targets across four axes (backbone regimens, biomarker-directed agents, alternative regimens, and guideline queries) rather than performing the retrieval itself. A thread pool executor coordinates the parallel dispatch of search terms to the Brave Search engine via API access, and the NCI PDQ database, queried based on the identified disease entity. Publicly available online guidelines are also consulted for treatment recommendations.

Online sources identified as suitable for the selected clinical categories are enumerated as text listings in an external YAML file. The resources consulted include abstracted listings in PubMed, NCI PDQ, and FDA New Approvals, as well as a general Internet search using the Brave search API. Capture is restricted to abstracts and the paragraphs pertinent to the query rather than full text. ASCO Rapid Recommendations and ESMO Living Guidelines are checked for late-breaking updates to their respective guidelines.

Each retrieved source is scored on four factors: recency, authority (domain tier plus a publication-type bonus), indication match (line-of-therapy alignment), and topical relevance. Topical relevance is judged primarily by a separate LLM grader that accounts for medical synonymy (e.g., recognizing "NSCLC" and "non–small cell lung cancer" as the same disease, while preserving distinctions such as IDH1 versus IDH2); a lexical word-overlap score is used as a fallback only if the grader fails. When the limited-options flag is set, the indication factor is relaxed so that adjacent-setting evidence is not penalized. The factors are combined as (authority + pub-type bonus) × recency × indication × topical relevance, and the composite score assigns each source to a PRIMARY, SECONDARY, TERTIARY, or BACKGROUND tier.

Wave 2 performs per-agent FDA/label adjudication after a candidate set is extracted from three channels — candidate agents named in the query text (only potential candidates, not those failed prior), agents in the candidate treatment-option set (if presented), and agents named in the Wave 1 retrieved text — unionized and de-duplicated. It is at this point that labels are retrieved from the NLM DailyMed database, and each label is reconciled with the approval announcements surfaced by the Wave 1 FDA oncology-approval filter, and the adjudicator judges each agent's approval scope: whether the label indicates approval for this indication, approval for a different setting, or whether the retrieved label was unclear or absent. Agents lacking an FDA label may still be considered on the basis of strong supporting data. This formatted constraint block is then injected into the synthesizer LLM prompt, and a formatted output is displayed (Fig. 1).

The response is formatted according to strict rules set in the system prompt, and the direct answer is required to be specially highlighted to remove ambiguity. Treatment recommendations or clinical answers deemed primary are rendered in red boldface and underlined. Treatment recommendations that are still considered standard of care although a modern alternative exists receive red boldface only. Throughout, the system is constrained to be retrieval-grounded rather than speculative: the model is instructed not to draw on its training knowledge, and if it cannot retrieve relevant documents, it is required to say so rather than answer from memory.

It should be noted that Llama3.3, a non-reasoning model, is also tasked with building the Brave search keyword string, the red-flag and limited-options calls, intent classification, agent extraction, and approval adjudication. A larger model, GLM-5.2, was selected for its larger context window to be able to hold in-prompt the full retrieved corpus of assembled information. This reinforces the principle that model inference is driven by grounded retrieval, rather than reliance on an advanced model’s parametric weights.

Model performance evaluation

To evaluate the performance of our model, we constructed a set of 100 clinical questions testing knowledge drawn from research published within the preceding year. By restricting questions to recent literature, we sought to preclude answers derived from parametric knowledge acquired during pretraining, thereby isolating the model’s ability to reason over current information.

The initial question set was generated by Z.AI’s GLM-5.2 model. To further decrease reliance on the training parameters of a particular model, Thinking Machines’ Inkling model was tasked to modify each question so as to alter its answer. This step was intended to mitigate the possibility that responses to the generated questions would derive from memorized answer options rather than from synthesis grounded in retrieved evidentiary data from available resources. The resulting questions were reviewed by a human board-certified oncologist (one of the authors) to confirm that each question was validly formulated, and that the designated answer was correct, using the NCCN guidelines as the reference standard. Questions were discarded if answer options erroneously contained duplicates (e.g. casting

trastuzumab deruxtecan and fam-trastuzumab deruxtecan-nxki as distinct choices) or if no supporting citation could be identified for the reference recommendation. The Aprilio model was given a single chance to answer 100 generated questions with no prompting other than the presentation of the vignette questions. Responses were compared with those of seven frontier models: Claude 5 Opus, Gemini 3.1 Pro Preview, and GPT-5.6 Terra, xAI Grok 4.5, DeepSeek V4 Pro, Perplexity Sonar, and Qwen3 Max. All models were provided with Internet access.

Answers generated by the frontier models, as well as the Aprilio model, were scored using the LLM-as-judge MCQ benchmark. The question-answer corpus was first validated across three sources (questions / answers / cross-check) before any model was called: vignette agreement, gold-letter agreement, option completeness, and answer-leak detection. Three frontier models were selected to be judges: GPT-5.5, Claude 5 Sonnet and Gemini 5.6 Flash. Every response was submitted to the three judges twice, yielding six structured extractions per response. The responses were collapsed and resolved by panel majority with a single tie-breaking judge. The extractor was validated against 26 known-truth fixtures at 100% letter accuracy before deployment for production. Extraction difficulty was not dependent on which system wrote the text ($p = 0.51$). Answer selection was scored by Answer Credit and tested with exact McNemar and Holm correction over 7 comparisons. Contraindicated endorsement uses a key built from the gold reasoning only, where a flag is admitted only with a verbatim quote verified by string containment, for 1,072 pair judgements. Citation reliability is source-neutral with per-response denominators of exactly 100, split between registry identifiers and bare URLs so that publisher bot-blocks are not counted as invented sources.

Although the Aprilio model contained its own internal system prompts, clinical vignettes from the hematology-oncology dataset were provided to each model to emulate the style of a physician, without auxiliary prompt engineering or retrieval instructions. Each frontier model relied solely on native, provider-configured default system prompts, allowing the use of trained weights or to access online information at its discretion.

Responses were scored against ground truth answers derived from the latest clinical practice guidelines from the National Comprehensive Cancer Network (NCCN), the American Society of Hematology (ASH), and the American Society of Clinical Oncology (ASCO). When a model generated multiple options, credit was awarded only if the guideline-concordant answer was identified as the preferred recommendation.

RESULTS

Model response accuracy

The Aprilio model was the only system to score 100% in answering the clinical vignette dataset correctly. Claude 5 Opus scored 99%, GPT-5.6 Terra, Gemini 3.1 Pro Preview and xAIGrok 4.5 each scored 98%. DeepSeep V4 and Perplexity Sonar Pro scored 97%, and Qwen3 Max scored 89% (Fig. 2). Seven of the eight models correctly answered at least 97 of the 100 questions correctly, while Qwen3 incorrectly

answered 11 questions. As a stylistic feature, the Aprilio model most often presented its output as a ranked set of responses amongst candidate options rather than stating a single answer as a preferred treatment. Cochran's Q, asking whether the eight systems differ as a set, is significant ($\chi 2 = 32.9$, df = 7, p = 2.7 x 10 – 5), driven by Qwen3 Max rather than by any separation amongst the top performers. The 24 incorrect responses were concentrated in 15 items, and 8 of those 15 were missed by Qwen3 Max alone.

Source validation

Citation validity is essential in clinical medicine, as therapy recommendations strive to be evidence-based. Clinical action based on hallucinated citations, especially resulting in an unfavorable outcome, would be indefensible. It is critical for an disease management AI assistant to be able to provide the basis for its recommendations, such as with citations that the physician can verify manually. This also enables auditability which becomes essential following an adverse event.

Prompt content and structure influence quality in AI inference.[15] However, when using chatbot apps, most practicing clinicians users input natural language queries without prefatory prompt scaffolding or procedural directives. To mirror real-world use, we supplied each model with the direct vignette queries, and deferred inference strategy to the internal system prompts intrinsic to each frontier model. In each case, LLM judges attempted to check each citation against a public registry (e.g. PMID, DOI, NCT, DailyMed SETID) and scored whether a registry ID could be retrieved. The judging platform also tracked each model's usage of online search tool capability.

When not explicitly instructed to source the answer to each clinical query, the frontier and open-source models differed in the degree of citation support provided (Fig. 3). By design, the Aprilio model conducted an online search to retrieve the necessary information to provide a grounded response in 100% of vignettes, each associated with a registry or a working URL. GPT-5.6 grounded its response in 97% of vignettes, and utilized the online search tool with every query, but not all citations were associated with a working locator. Lack of a working locators could signify a deprecated reference, or a fabricated citation. In contrast, the other models (Qwen3, Sonar Reasoning Pro, Grok 4.5, DeepSeek, Gemini 3.1, and Claude 5) left a citation locator in only 0 to 12% of the vignettes. Although tracking revealed engagement of the online search tool, these models did not provide source material citation. Furthermore, the latter four models answered many questions bypassing online retrieval entirely, relying solely on pre-trained knowledge. Gemini 3.1 Pro provided no machine-verfiable identifiers across any of the 100 responses. Using Fisher's 2-sided exact test, the difference in providing ≥ 1 working locator compared with the Aprilio model was significant for all models ($p < 10^{-43}$) except GPT-5.6 Terra ($p = 0.246$). Wilson and Newcombe 95% confidence intervals confirm that the Aprilio's leaving at least one working locator is significantly different from all models except GPT-5.6 Terra. The use of search in the Aprilio model is also significantly done more often compared to Grok 4.5, DeepSeek V4 Pro, Gemini 3.1 Pro and Claude 5 Opus (Fig. 4).

The record of what each search retrieved shows why election alone is a weak guarantee. Across the 520 responses in which a search was attested, providers returned 5,222 URLs spanning 727 distinct domains, 61% of which appear in only a single response. Breadth varied by an order of magnitude: Grok 4.5's median response drew 18 URLs across 10 domains and its widest drew 95, whereas GPT-5.6 Terra's median drew 2 URLs from a single domain. Primary sources and registries like PubMed, PubMed Central, ClinicalTrials.gov, FDA and DailyMed do account for 28.5% of domain visits, and for answering these vignettes such sources are sufficient, as the accuracy results show. The distribution matters for a narrower reason: a claim about an NCCN evidence category or preference tier can be settled against one source only, yet NCCN's domains account for 2.6% of visits and just 5% of searched responses reached nccn.org, much of the panel content sitting behind a login. Most traffic went instead to secondary literature like review articles, conference coverage and aggregator sites. Tier language was therefore most often reproduced from commentary about a guideline rather than from the guideline itself, which is the explanation behind the misattributions described below.

Evidentiary misrepresentation

Model inference dependent on trained weights can lead to **evidentiary misrepresentation** – a form of grounding fallacy in which the model cites a genuine, verifiable source, yet the underlying evidence is weaker, more conditional, or otherwise discordant with the certainty conveyed in the model's response. We preface the following discussion noting that of the 700 comparator model responses, search was elected in only 520 of the vignette cases, and of those, 398 responses left no citation locator that would permit verification. GPT-5.6 grounded 97% of its responses, while others grounded between 0% and 12%. Gemini 3.1 Pro left no machine-verifiable identifiers anywhere in its 100 responses.

Our investigation has exposed instances where a confident guideline claim with no locator attached resulted in an incorrect guideline rung. In question concerning a case of refractory intrahepatic cholangiocarcinoma with an IDH1 R132C mutation, Claude 5 Opus correctly answered ivosidenib but indicated that this drug was NCCN category 1, when it is actually listed as “useful in certain circumstances”. In another question regarding metastatic GIST harboring a PDGFRA exon 18 D842V mutation, Grok 4.5 correctly answered avapritinib as the correct therapy without searching, also claiming category 1 NCCN endorsement, when it is category 2A, as its approval rests on a single 43-patient phase I trial. DeepSeek V4 Pro answered a vignette correctly, quoted NCCN accurately, then compressed its own correct analysis into a summary line that moved a category 2B label onto a category 1 drug. In a case concerning a 68 y/o patient with castrate-resistant prostate cancer facing progression after enzalutamide who has never received chemotherapy, GPT-5.6 Terra, in a sentence concerning sipuleucel-T, cited a supporting reference which resolved to a Cancer Care Ontario monograph on abiraterone. Finally, in case concerning maintenance therapy for multiple myeloma after autologous transplantation, GPT-5.6 provided a guidelines link that could not be resolved for verification.

Approximately one rung claim in fourteen was inaccurate. Of the 82 claims that were scored as correct, 42 did not link to a locator that could be used for verification, and 34 claims were defective that were

only confirmable though a third-party restatement, a locator pointing to a document other than a guideline, or other indirect means. Restricted to responses that were scored as correct, 6 of 84 claims were wrong. While the vignette did not specifically request NCCN category and preference tier, these were offered by the models, unprompted, in the commentary space of the response (Fig. 5). In another case concerning a patient with metastatic ampullary carcinoma, GPT-5.6 Terra provided a reference in support of FOLFIRINOX (the correct answer) which linked to a USPTO patent-petition docket.

Resolvable Citations that Support the Claim

The number of citations provided by AI assistants varied by more than two orders of magnitude from 1206 in the Aprilio model, to 181 provided by GPT-5.6 Terra, 12 by Qwen3 Max, and 5 or less in the other models.

The 1206 citations presented by the Aprilio model were resolvable in all cases, and none were hallucinated. Twelve of Aprilio-provided links could not be resolved by automated crawler, but all of these resolved on manual attempts. Two of the citations provided by GPT-5.6 Terra generated a 404 FDA server error, while the third was confirmed to be a dead ASCO link. Seven other citations provided by GPT-5.6 did resolve, however one was intended to support sipuleucel-T but linked to an abiraterone monograph. (Fig. 6).

It is clear that existence of a citation does not imply support of the underlying claim. To assess whether cited literature substantiated the responses to the clinical vignettes, each reference was systematically retrieved and examined. A minimum length threshold of 120 characters was established to ensure sufficient substantive text to evaluate reference integrity and contextual fidelity. A three-judge panel selected one citation per vignette (PMID over DOI over NCT over DailyMed over URL; preference for locators inside the recommendation span) to evaluate whether the answer was supported by the associated citation. Each judge voted SUPPORTS / PARTIALLY_SUPPORTS / UNRELATED_OR_CONTRADICTED / INSUFFICIENT_SOURCE_TEXT, resolved by majority of votes.

Although only 57% of the references cited by the Aprilio model were judged as supportive under the 120-character threshold, the model provided candidate citations in 97% of vignettes, a rate markedly higher than those of comparator models (Fig. 7). While this substantial disparity in evaluable citations precludes a direct cross-comparison of reference quality, clinicians are likely better served by an AI assistant that proactively supplies evidentiary support rather than shifting the burden of literature verification entirely onto the user.

Another key advantage of grounded retrieval, as contrasted with approaches reliant on trained LLM model weights or static vector databases, is its capacity to surface breaking information relevant to a medical query. General purpose frontier models may not be equipped to supply such information, and even domain-specific AI assistants may prove similar as unresponsive. For example, on 26 August 2026, the multi-selective RAS(ON) inhibitor, daraxonrasib, was approved by the FDA for adults with pancreatic

adenocarcinoma who had received at least one prior systemic chemotherapy, or who were not candidates for standard multiagent systemic therapy.[16] When queried, the Aprilio model correctly reported this new approval, as did Claude 5 Opus and GPT-5.6 Sol. Perplexity and Gemini 3.7 Flash, however, continued to recommend conventional regimens as second-line options, as did OpenEvidence when queried on the day of approval. A medical oncologist consultating such tools for therapy options might remain unaware of this important breakthrough option when advising a candidate patient.

It is important in disease management consultation that a recommendation not faciliate patient harm. Although AI assistants can rationalize therapeutic recommendations based on clinical history, they may fail to recognize and integrate critical patient factors. For example, Perplexity Sonar was able to confirm that loncastuximab was appropriate for relapsed/refractory DLBCL but failed to recognize that baseline edema and a pleural effusion would necessitate alternate therapy. Qwen3 Max failed to recognize that mobocertinib was withdrawn from the market in 2023 and still recommended it as a treatment option.

DISCUSSION

AI assistants powered by large language models (LLMs) have demonstrated considerable utility in generating text informed by extensive training data. As physicians have grown increasingly familiar with the integration of these tools into clinical practice, and as many have gained firsthand experience evaluating their outputs, interest has grown in exploring how such assistants might find a place in clinicians' workflows. But while the outputs of advanced LLMs may appear authoritative and convincingly factual, the accuracy of the generated content still cannot be guaranteed. For this reason, aside from clinical documentation applications, a significant proportion of physicians remain cautious about adopting LLM-based assistants in clinical practice.[17]

While these AI assistants can be tasked with constructing plausible clinical scenarios consistent with the clinical features and variables provided, it has become clear that designing and implementing metrics to rank model performance—whether comparing these tools to one another or to human clinicians—remains challenging. A more reliable and trustworthy application may instead lie in assisting healthcare providers with information management, a task for which demand continues to grow. Establishing success in this domain may serve as a productive initial step toward the broader goal of replicating a knowledgeable and trusted clinical consultant.

The harness component of AI assistant systems—the infrastructure managing the information pipeline within LLM deployment—is very often proprietary. While specialized models can be fine-tuned on a domain-specific corpus, this approach is constrained by training costs, the eventual obsolescence of training knowledge, and the degradation of performance on previously learned tasks (i.e., 'catastrophic forgetting'). Retrieval-augmented generation (RAG) addresses several of these limitations; however, concerns regarding the stochastic behavior of vector database search have motivated interest in strategies that pursue more grounded methods of sourcing verifiable evidence.

Our model also demonstrates that when supported by an appropriate harness, less powerful and earlier-generation open-weight LLMs, such as GLM-5.2 and the non-reasoning Llama 3.3, achieve parity in accuracy with leading foundation models, even outperforming them in citation integrity. While on-premises (local) AI deployments are increasingly favored to comply with data privacy regulations, manage operational expenditures, and address data governance, hardware limitations and intellectual property constraints preclude local hosting of leading frontier models. Our work highlights a path for achieving state-of-the-art oncology AI assistant performance using local infrastructure by augmenting smaller models with a structured ontological framework.

Deploying LLMs to automate retrieval from trusted disease-specific guidelines curated by recognized thought leaders, supplemented by point of query retrieval can surface critical information potentially missed by reliance on the training of more expensive large foundational models alone. Beyond offering convenience, this approach – as our use cases illustrate – provides access to current information that can substantially impact awareness of updated treatment options and enhance delivery of state-of-the-art oncologic care.

Imposing a standardized structure on queries submitted to an LLM can improve retrieval accuracy, as can reference to ontology-grounded knowledge graphs. [18,19] The provision of structure is likewise a defining feature of neurosymbolic processing, a machine learning paradigm in which model comprehension is scaffolded by an explicit framework for organizing factual knowledge.[20] To remain useful, however, knowledge graphs and formal ontologies require continual updating as underlying information networks evolve and expand, a maintenance burden that constrains their broader application.

LLMs can operationally skirt this limitation by generating a working symbolic struture derived from the model's interpretation of query components, linked to supplementary resource files (e.g. YAML) files that require substantially less frequent updating. This process provides the basis for what might be termed an "ontology-on-the-fly". We exploit the capacity of LLMs to extract intent from a query, thereby guiding a focused retrieval of information from relevant curated source material, rather than pursuing a more diffuse search strategy – one that tends to generate greater noise than signal. Queries pertaining to supportive care and symptom management, for instance, draw upon a different resource set than those addressing treatment selection for a medically unfit patient, or those seeking stage-specific or biomarker-directed therapeutic recommendations. Because these tasks are grounded extraction rather than clinical inference, the model used for query parsing for intent classification, keyword extraction, search string construction, agent extractions and approval adjudication did not require advanced models with reasoning capabilities.

The effectiveness of the approach we describe is dependent on the availability of curated information accessible through online retrieval protocols. We have shown that semantic deconstruction of a raw query into symbolic entities permits information retrieval of relevant content that serves as the basis for semantic search. This process calls upon the ability of a language model to identify components of a

provided text. Note that the query-processing LLM generates its output through next-token prediction. The task is nonetheless designed to keep that generation tightly-scoped, via prompt instruction and deterministic decoding, with output validated against a fixed label set. While the encoder-only transformer models offer a strictly non-generative and deterministic framework, better able to support claims of being free of off-menu output, these models are less flexible, may require task-specific fine-tuning, and risk imprecision in parsing the complex structure of general medical queries, not all of which follow a specific clinical vignette format style. Similarly, the salient portions of the retrieval material, containing metadata, such as source provenance, URL, and publication date, undergo consolidative synthesis using an LLM, constrained to the assembled material. Rank weightings of authority, recency, publication type, relevance to the indication and topic are set by expert judgment, auditable per-factor, and tuned to produce the output.

We note that our model is designed to assist in disease management and is not intended to provide diagnostic support. Applications that encourage such cognitive offloading are vulnerable to a distinct set of reasoning failures: salience bias and association bias, which originate in the model's training and are modulated by prompt structure; and anchoring bias, which originates in the physician's cognitive state. AI assistance used for diagnosis generation is especially hazardous for the inexperienced clinician, in whom independent clinical reasoning may be insufficiently mature to recognize and challenge flawed output. AI assistants can empower physicians by streamlining access to timely and relevant reference data, providing educational value, while fostering active engagement in clinical reasoning.

Relentless advances in molecular biology and pathophysiology research, alongside corresponding advances in targeted pharmaceuticals and clinical trial evidence, have posed a formidable challenge to the medical oncologist tasked with selecting appropriate treatment from among a panoply of patient factors – biomarker profile, performance status, prior treatment history – and matching these against the clinical trial eligibility criteria underlying regulatory approval status. Both patients and insurance companies demand such precision from clinicians, and although AI-powered clinical tools promise to provide assistance, factors such as the absence of an accepted benchmark performance standard and the secrecy surrounding proprietary model architectures undermine transparency and complicate trust. Furthermore, continued advances in general purpose frontier models threaten to erode the purported advantage of proprietary domain-specific models, which depend on accurate inference from trained weights, vectorized stored resources, or even static knowledge graphs. Grounded, ontology-driven retrieval across a panel of trusted, curated resources yielding verifiable output may offer the optimal solution.

Declarations

Dataset availability:

The datasets used and/or analyzed during the current study are not publicly available to protect the dataset from crawlers which could allow frontier models to train on these evaluation cases, which would potentially undermine reproducibility, invalidating subsequent benchmarking efforts. The datasets used in this study are available from the corresponding author to qualified researchers for benchmarking and replication purposes, with the stipulation that they are not used for model training.

Funding statement:

The authors received no financial support for the research, authorship, and/or publication of this article.

Author Contribution

GT: cowrote manuscript text, generated med oncology dataset, designed LLM harness, provided medical input, prepared figures 1,3,4,7EM: cowrote manuscript text, designed and deployed LLM-as-judge judging platform, prepared figures 2,5,6AVB: cowrote manuscript text, provided input on harness design and implementationET: provided input on harness design and implementationAll authors reviewed the manuscript.

Data Availability

To develop the medical oncology clinical vignettes, we utilized Z.ai's GLM-5.2 model to generate a set of questions reflecting advances in medical oncology and hematology from the past year. To mitigate the risk of data contamination, we strictly avoided published datasets to which leading frontier models might have already been exposed during pretraining. Furthermore, we deliberately excluded the most ubiquitous frontier models from the initial question design phase. The resulting question-answer pairs were subsequently processed by Thinking Machine's Inkling model, which revised the vignettes to alter the correct responses while maintaining their focus on recent research developments. This secondary modification was implemented specifically to minimize both association bias and answer selection position bias. The datasets used and/or analyzed during the current study are not publicly available to protect the dataset from crawlers which could allow frontier models to train on these evaluation cases, which would potentially undermine reproducibility, invalidating subsequent benchmarking efforts. The datasets used in this study are available from the corresponding author to qualified researchers for benchmarking and replication purposes, with the stipulation that they are not used for model training.

References

1. Vishwanath, K. et al. General-purpose large language models outperform specialized clinical AI tools on medical benchmarks. *Nat. Med.* **32** (7), 2405–2409. 10.1038/s41591-026-04431-5 (2026).

2. Waldock, W. J., Guni, A., Darzi, A. & Ashrafian, H. Performance of predictive AI-based clinical decision support systems across clinical domains: A systematic review and meta-analysis. *PLOS Digit. Health*. **5** (3), e0001310. 10.1371/journal.pdig.0001310 (2026).
3. Jackson, N. J. et al. Factors influencing the effectiveness of artificial intelligence-assisted decision-making in medicine: a scoping review. *J. Am. Med. Inf. Assoc.* **33** (5), 1054–1064. 10.1093/jamia/ocag002 (2026).
4. Xiong, G., Jin, Q., Lu, Z. & Zhang, A. Benchmarking Retrieval-Augmented Generation for Medicine. In: *Findings of the Association for Computational Linguistics ACL 2024*. *Association Comput. Linguistics*. 6233–6251. 10.18653/v1/2024.findings-acl.372 (2024).
5. Zakka, C. et al. Almanac — Retrieval-Augmented Language Models for Clinical Medicine. *NEJM AI*. **1** (2). 10.1056/AIoa2300068 (2024).
6. Jeong, S., Baek, J., Cho, S., Hwang, S. J. & Park, J. C. Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity. *arXiv*. Preprint posted online March 28, 2024:arXiv:2403.14403. 10.48550/arXiv.2403.14403
7. Lammert, J. et al. Expert-Guided Large Language Models for Clinical Decision Support in Precision Oncology. *JCO Precis Oncol.* **8**, e2400478. 10.1200/PO-24-00478 (2024).
8. Hung, T. K. W. et al. Performance of Retrieval-Augmented Large Language Models to Recommend Head and Neck Cancer Clinical Trials. *J. Med. Internet Res.* **26**, e60695. 10.2196/60695 (2024).
9. Wu, K. et al. An automated framework for assessing how well LLMs cite relevant medical references. *Nat. Commun.* **16** (1), 3615. 10.1038/s41467-025-58551-6 (2025).
10. Belligoli, P., Bitterman, D. & Miller, T. Do They Really Know? Evaluating Large Language Models' Ability to Reference and Cite Oncology Guidelines. In: *Artificial Intelligence in Medicine: 23rd International Conference, AIME* Pavia, Italy, June 23–26, 2025, Proceedings, Part II. Springer-Verlag; 2025:30–34. (2025). 10.1007/978-3-031-95841-0_6
11. Schaftner, S. & Gaedke, M. ORKG Properties Ontology Consolidated: LLM-Driven Refinement of Crowdsourced Knowledge for Machine-Actionability. In: *Companion Proceedings of the ACM Web Conference 2026*. WWW Companion '26. Association for Computing Machinery; :1009–1016. (2026). 10.1145/3774905.3795080
12. Clinical Knowledge Graph Construction and Evaluation with Multi-LLMs via Retrieval-Augmented Generation. Accessed August 28. (2026). https://arxiv.org/html/2601.01844v1
13. Rumiantsau, M., Vertsel, A., Hrytsuk, I. & Ballah, I. Beyond Fine-Tuning: Effective Strategies for Mitigating Hallucinations in Large Language Models for Data Analytics.
14. Sundararajan, B., Sripada, S. & Reiter, E. Input Matters: Evaluating Input Structure's Impact on LLM Summaries of Sports Play-by-Play. *arXiv*. Preprint posted online October 27, 2025:arXiv:2510.21034. 10.48550/arXiv.2510.21034
15. Liu, J., Liu, F., Wang, C. & Liu, S. Prompt Engineering in Clinical Practice: Tutorial for Clinicians. *J. Med. Internet Res.* **27** (1), e72644. 10.2196/72644 (2025).

16. U.S. Food and Drug Administration. FDA approves daraxonrasib for metastatic pancreatic adenocarcinoma. *FDA*. Published online August 26, Accessed August 26, 2026. (2026). https://www.fda.gov/drugs/resources-information-approved-drugs/fda-approves-daraxonrasib-metastatic-pancreatic-adenocarcinoma
17. Poon, E. G., Lemak, C. H., Rojas, J. C., Guptill, J. & Classen, D. Adoption of artificial intelligence in healthcare: survey of health system priorities, successes, and challenges. *J. Am. Med. Inf. Assoc.* **32** (7), 1093–1100. 10.1093/jamia/ocaf065 (2025).
18. Ayoub, M. et al. Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. *Commun. Med.* **6** (1), 86. 10.1038/s43856-025-01348-x (2026).
19. Ali, M., Taha, Z. & Morsey, M. M. Ontology-grounded knowledge graphs for mitigating hallucinations in large language models for clinical question answering. *J. Biomed. Inf.* **175**, 104993. 10.1016/j.jbi.2026.104993 (2026).
20. Velik, R. & Bruckner, D. Neuro-symbolic networks: introduction to a new information processing principle. In: *6th IEEE International Conference on Industrial Informatics*. IEEE; 2008:1042–1047. IEEE; 2008:1042–1047. (2008). 10.1109/INDIN.2008.4618256

Figures

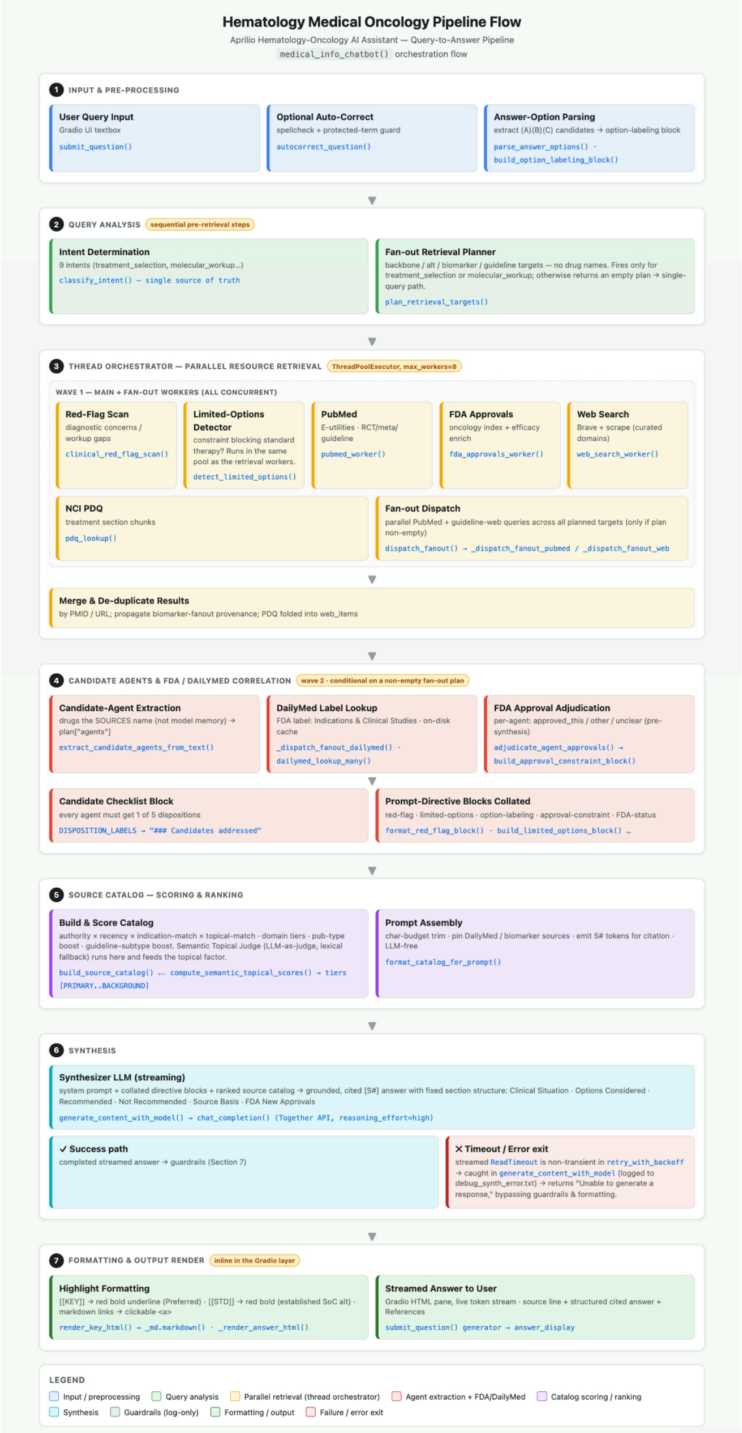


Figure 1

Flow diagram of the Aprilio model, illustrating the processing of the query input to response synthesis.

System	Answer Credit	AC = 1.0	Wrong-answer rate	Multi-option answers
Aprilio	1.00	100%	0%	24%
Claude Opus 5	0.99	99%	1%	0%
GPT-5.6 Terra	0.98	98%	2%	3%
Gemini 3.1 Pro	0.98	98%	2%	3%
Grok 4.5	0.98	98%	2%	0%
DeepSeek V4 Pro	0.97	97%	3%	0%
Sonar Reasoning Pro	0.97	97%	3%	1%
Qwen3 Max	0.89	89%	9%	2%

Figure 2

Primary endpoint: Answer credit measures the ability to correctly answer each clinical vignette query without refusal. Nearly all models were able to respond with the correct answer with a high degree of accuracy. For multi-option responses where several approaches were considered viable depending on clinical context, full credit required the model to rank the alternatives and clearly designate the preferred option.

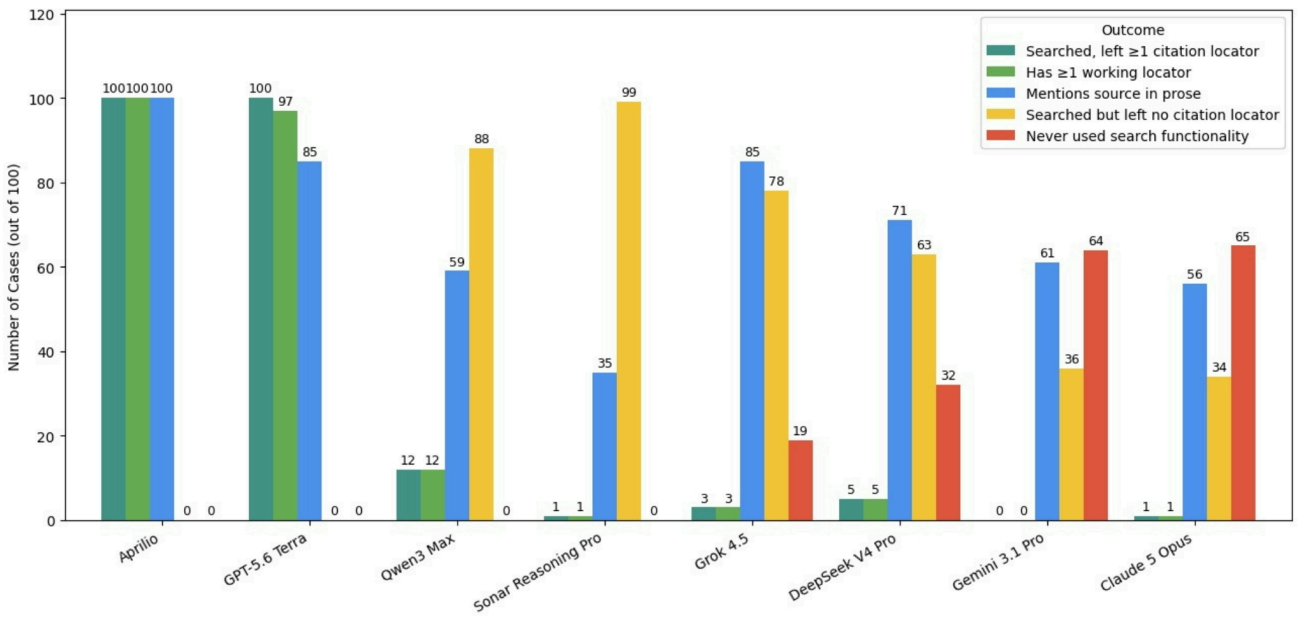
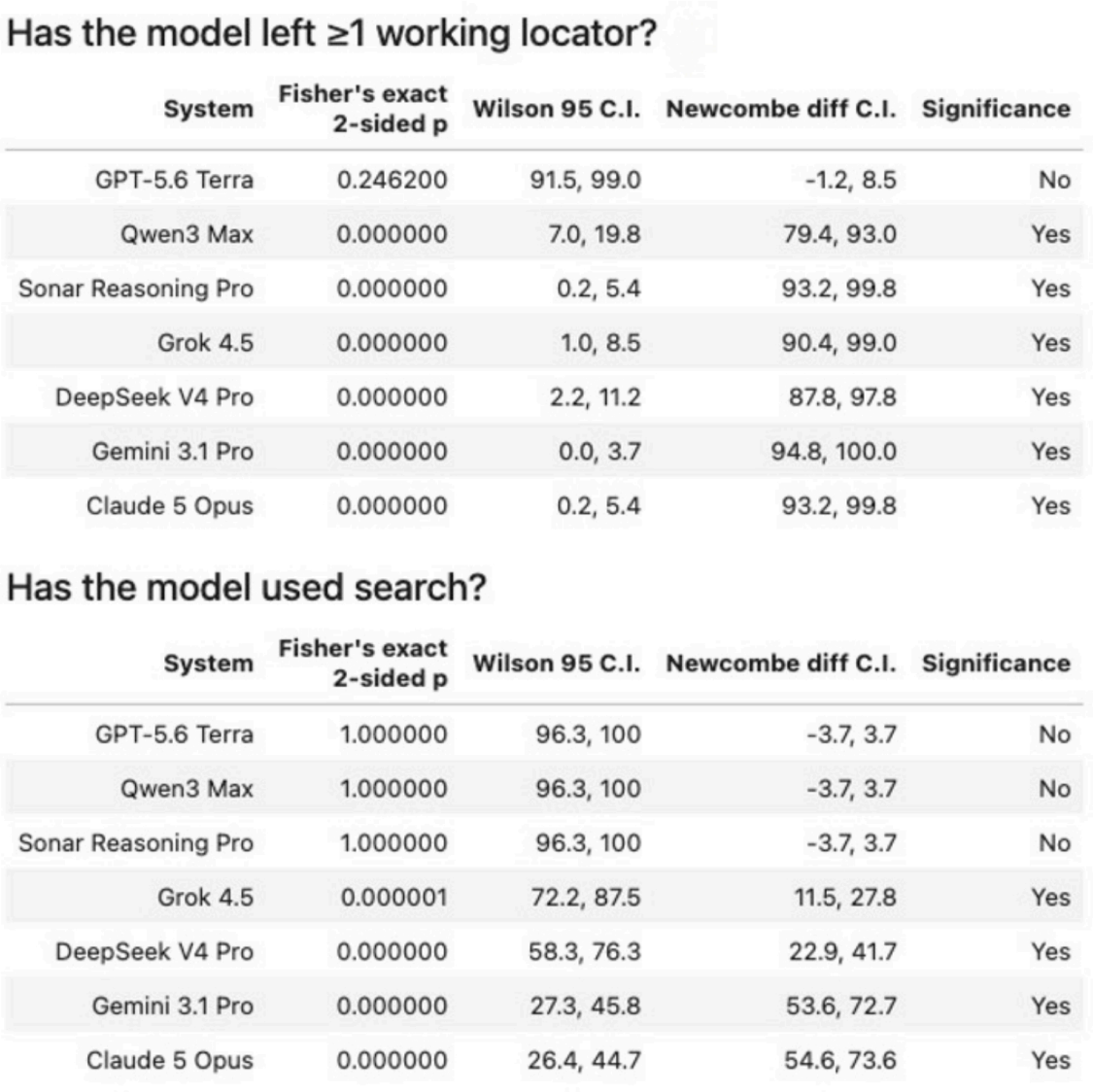


Figure 3

Citation support: Online search was inconsistent, and several modes never confirmed their answers with online sources. Even when online search was conducted, several modes did not cite sources, or only referred to supporting literature in prose but did not provide a location for the user to confirm directly.



Has the model left ≥1 working locator?

System	Fisher's exact 2-sided p	Wilson 95 C.I.	Newcombe diff C.I.	Significance
GPT-5.6 Terra	0.246200	91.5, 99.0	-1.2, 8.5	No
Qwen3 Max	0.000000	7.0, 19.8	79.4, 93.0	Yes
Sonar Reasoning Pro	0.000000	0.2, 5.4	93.2, 99.8	Yes
Grok 4.5	0.000000	1.0, 8.5	90.4, 99.0	Yes
DeepSeek V4 Pro	0.000000	2.2, 11.2	87.8, 97.8	Yes
Gemini 3.1 Pro	0.000000	0.0, 3.7	94.8, 100.0	Yes
Claude 5 Opus	0.000000	0.2, 5.4	93.2, 99.8	Yes

Has the model used search?

System	Fisher's exact 2-sided p	Wilson 95 C.I.	Newcombe diff C.I.	Significance
GPT-5.6 Terra	1.000000	96.3, 100	-3.7, 3.7	No
Qwen3 Max	1.000000	96.3, 100	-3.7, 3.7	No
Sonar Reasoning Pro	1.000000	96.3, 100	-3.7, 3.7	No
Grok 4.5	0.000001	72.2, 87.5	11.5, 27.8	Yes
DeepSeek V4 Pro	0.000000	58.3, 76.3	22.9, 41.7	Yes
Gemini 3.1 Pro	0.000000	27.3, 45.8	53.6, 72.7	Yes
Claude 5 Opus	0.000000	26.4, 44.7	54.6, 73.6	Yes

Figure 4

Pairwise comparison between the Aprilio model and comparators regarding differences in the provision of at least one working locator to verify evidentiary claims, and whether the model has accessed online searching to confirm claim recency and validity.

Claude 5 Opus, verbatim	What NCCN prints
"FDA-approved (2021) for previously treated, locally advanced or metastatic IDH1-mutated cholangiocarcinoma; **NCCN category 1 preferred subsequent-line option**"	"Ivosidenib is a **category 1 useful-in-certain-circumstances** subsequent-line systemic therapy option for unresectable or metastatic progressive CCA with IDH1 mutations." — NCCN Biliary Tract Guidelines BIL-C 1.2026 (2 of 15)
"a matched biomarker-directed agent **takes precedence over cytotoxic chemotherapy in NCCN guidelines**"	No such precedence rule appears. NCCN lists ivosidenib on and FOLFOX in different Principle of System Therapy pages, and the regimen it labels “preferred” for subsequent-line chemotherapy if disease progression is FOLFOX – NCCN Biliary Tract Guidelines BIL-C 1.2026 (2 of 15 and 3 of 15)

Aprilio, verbatim	What NCCN prints
"**Preferred:** ivosidenib (Option C) is the preferred targeted therapy **for this patient**. It is an oral, small-molecule inhibitor of mutant IDH1 that is FDA-approved for adult patients with previously treated, locally advanced or metastatic cholangiocarcinoma harboring an IDH1 mutation [S65]."	The ivosidenib DailyMed label and ClarIDHy (PMID 32416072), both linked. "Preferred" ranks the four options in front of it; it is not offered as an NCCN tier.

Figure 5

An example illustrating output from Claude 5 Opus correctly selecting ivosidenib as the appropriate treatment for a patient with progressive cholangiocarcinoma, but providing an incorrect guidelines rung.

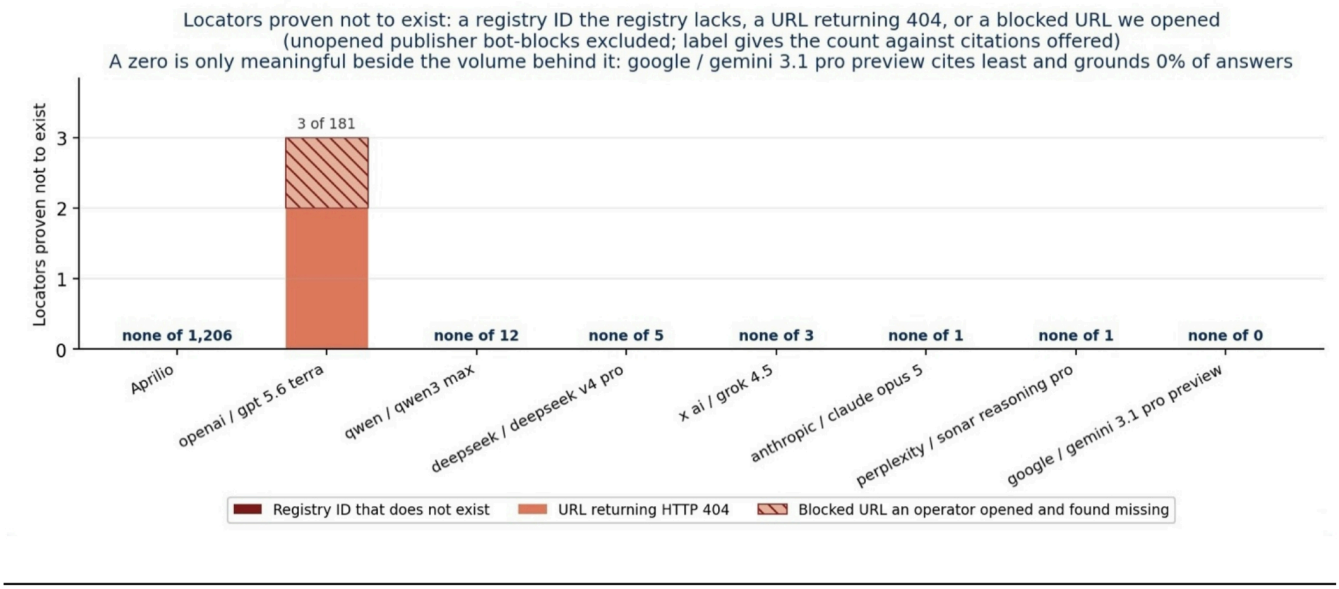


Figure 6

Every citation generated by the Aprilio model resolved without failure. GPT-5.6 provided 181 citations, with all but three resolving via automated crawler. For the remaining comparator models, the apparent absence of resolution errors is attributable to their near-total lack of citation generation rather than higher retrieval integrity.

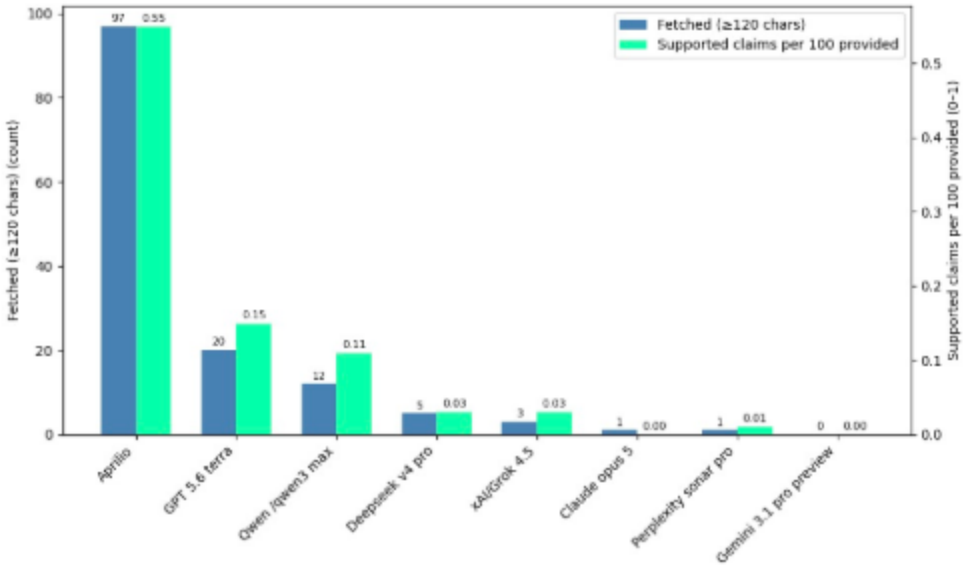


Figure 7

The Aprilio model provided candidate references suitable for judging support integrity more frequently than comparator models, which provided evidentiary support in a minority of cases.