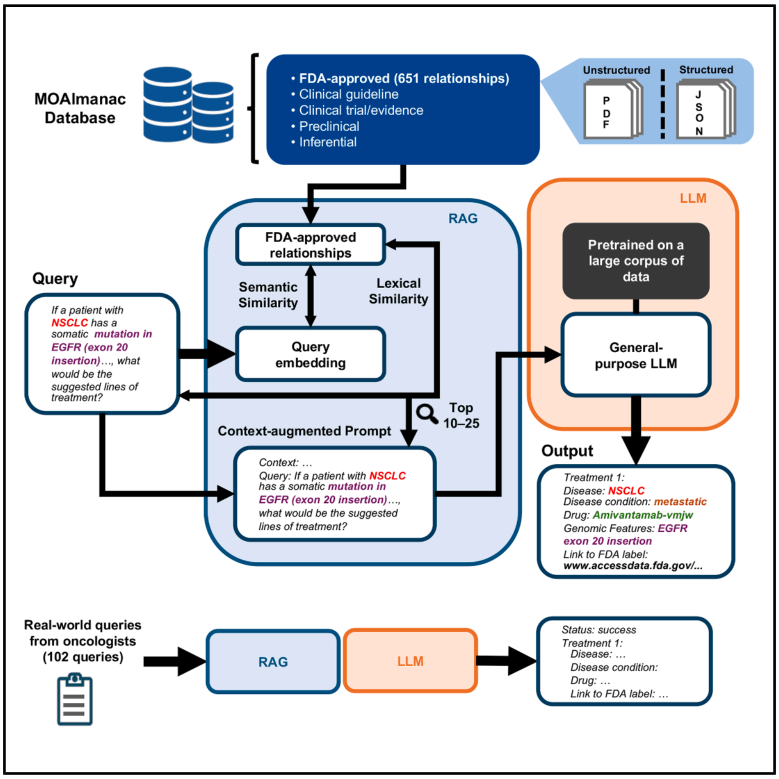
Cancer Cell

Report

A context-augmented large language model for accurate precision oncology medicine recommendations

Graphical abstract


MOAlmanac Database
• FDA-approved (651 relationships)
• Clinical guideline
• Clinical trial/evidence
• Preclinical
• Inferential
Unstructured
PDF
Structured
JSON
RAG
FDA-approved relationships
Semantic Similarity
Lexical Similarity
Query embedding
Query
If a patient with NSCLC has a somatic mutation in EGFR (exon 20 insertion)..., what would be the suggested lines of treatment?
Top 10–25
Context-augmented Prompt
Context: ...
Query: If a patient with NSCLC has a somatic mutation in EGFR (exon 20 insertion)..., what would be the suggested lines of treatment?
LLM
Pretrained on a large corpus of data
General-purpose LLM
Output
Treatment 1:
Disease: NSCLC
Disease condition: metastatic
Drug: Amivantamab-vmjw
Genomic Features: EGFR exon 20 insertion
Link to FDA label: www.accessdata.fda.gov/...
Real-world queries from oncologists (102 queries)
RAG
LLM
Status: success
Treatment 1:
Disease: ...
Disease condition:
Drug: ...
Link to FDA label: ...

Highlights

- Context-augmented LLM framework for precision oncology recommendations
- Framework validated on 102 oncologist-derived real-world queries
- Achieved 95% accuracy on synthetic and 93% accuracy on real-world queries
- Benchmarked prompting and retrieval strategies for optimal performance

Authors

Hyeji Jun, Yutaro Tanaka, Shreya Johri, ..., Alok K. Tewari, Brendan Reardon, Eliezer Van Allen

Correspondence

eliezerm_vanallen@dfci.harvard.edu

In brief

Precision oncology requires accurate biomarker-driven treatment guidance, yet LLMs often lack up-to-date clinicogenomic knowledge. Jun et al. develop a dynamically updated context-augmented LLM framework that improves biomarker-driven treatment recommendations compared to LLM-only approaches, achieving up to 93% accuracy on real-world queries and providing an adaptable framework for LLM deployment in oncology.

Jun et al., 2026, Cancer Cell *44*, 1–10
April 13, 2026 © 2025 The Author(s). Published by Elsevier Inc.
https://doi.org/10.1016/j.ccell.2025.12.017

CellPress

Cancer Cell

CellPress
OPEN ACCESS

Report

A context-augmented large language model for accurate precision oncology medicine recommendations

Hyeji Jun,[1,2] Yutaro Tanaka,[2,3] Shreya Johri,[1,2,4] Sabrina Y. Camp,[1,2] Erik L. Bao,[1,2] Filipe L.F. Carvalho,[7] Dan Y. Gui,[1,2,4] Alexander C. Jordan,[1,2] Chris Labaki,[1,9] Samantha D. Martin,[2,3,4,5] Matthew Nagy,[2,3,5] Tess A. O'Meara,[1,2] Theodora Pappa,[1,2,6] Erica Maria Pimenta,[1,2] Eddy Saad,[1,2] David D. Yang,[1,8] Riaz Gillani,[2,3,4,5] Alok K. Tewari,[1,2,4,6] Brendan Reardon,[1,2] and Eliezer Van Allen[1,2,4,10,*]

[1]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[2]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA
[3]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[4]Harvard Medical School, Boston, MA 02115, USA
[5]Boston Children's Hospital, Boston, MA 02115, USA
[6]Brigham and Women's Hospital, Boston, MA 02115, USA
[7]Department of Urology, Brigham and Women's Hospital, Boston, MA 02115, USA
[8]Department of Radiation Oncology, Brigham and Women's Hospital, Boston, MA 02115, USA
[9]Department of Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215, USA
[10]Lead contact
*Correspondence: eliezerm_vanallen@dfci.harvard.edu

https://doi.org/10.1016/j.ccell.2025.12.017

SUMMARY

The rapid expansion of molecularly informed therapies in oncology, coupled with evolving regulatory food and drug administration (FDA) approvals, poses a challenge for oncologists seeking to integrate precision oncology medicine into patient care. Large language models (LLMs) have clinical potential, but their reliance on general knowledge limits their ability to provide up-to-date and niche treatment recommendations. Here, we developed a retrieval-augmented generation (RAG)-LLM workflow using the molecular oncology almanac (MOAlmanac) and benchmarked it against an LLM-only approach for biomarker-driven treatment recommendations. Our RAG-LLM achieved up to 95% accuracy on synthetic queries and 93% on real-world queries collected from practicing oncologists. Finally, our study explored several prompting and retrieval strategies to enhance performance. Taken together, this approach may serve as valuable guidance for deploying LLMs to support cancer patients' treatment decisions in precision oncology clinical settings.

INTRODUCTION

Identifying therapeutically targetable molecular alterations to guide treatment options is a key component of precision oncology. However, the growing complexity and volume of regulatory approvals for such therapies make it increasingly challenging for clinicians to stay up to date with relevant clinicogenomic relationships.[1] Tracking these approvals often requires navigating multiple scattered sources, including clinical practice guidelines from professional organizations, published journal articles, and a regulatory agency's own website. Moreover, certain approvals occur with limited or no publicity, further delaying awareness. This knowledge gap may hinder the timely implementation of new approvals into clinical practice, especially for clinicians not well-versed in cancer genomics.[2]

In recent years, large language models (LLMs) have demonstrated the capability for tasks such as patient-to-clinical trial matching,[3–5] performing clinical summarization tasks,[6] and achieving physician-level performance on medical board examinations.[7] There is also emerging interest in their potential to support clinical decision-making in precision oncology.[8–10] Despite these advancements, LLMs have limitations when handling questions relating to niche and constantly evolving knowledge, particularly in the field of oncology.[11] These challenges are due to insufficient domain-specific training and reliance on frequently outdated data, which raise concerns about the accuracy and relevance of their output.[12] Overcoming them is especially important for potential applicability of these systems to aid oncologists with clinical decision-making, as it is critical that they utilize the most up to date knowledge.

Among efforts to address these limitations, retrieval-augmented generation (RAG) has emerged as a promising approach. RAG enables an LLM to dynamically retrieve relevant

Please cite this article in press as: Jun et al., A context-augmented large language model for accurate precision oncology medicine recommendations, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.12.017

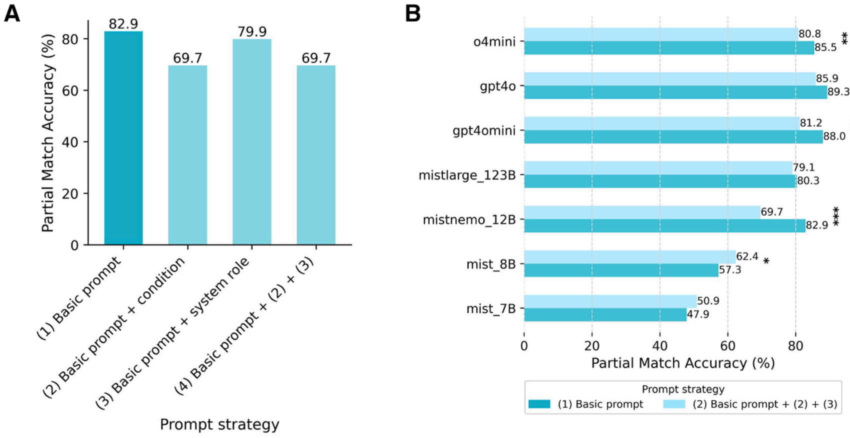


Figure 1. Performance of LLMs across various prompt engineering strategies

(A) Partial match accuracies of Mistral NeMo 12B using different prompt strategies.

(B) Partial match accuracies of LLMs with varying sizes across basic and combined prompt strategies. Accuracies between the basic and combined prompt strategies were compared using McNemar's test. p value significance is shown if the p value is less than 0.05 ($p < 0.001$: ***, $p < 0.01$: **, $p < 0.05$: *); otherwise, it is not displayed.

information from external domain-specific databases to supplement an LLM's general knowledge without modifying its internal weights.[13–17] Prior studies have focused on developing and optimizing RAG-based frameworks to answer open-ended medical questions.[15,18] More recently, Ferber et al. developed and evaluated a clinical decision-making AI agent that integrates multimodal precision oncology tools; however, the study was limited by a small real-world sample size.[16] RAG-based LLMs such as OpenEvidence have also emerged to assist with medical literature searches. While these can support physicians in making clinical decisions, they are usually trained on broad, general-purpose medical knowledge bases, and they are closed source or limited to specific users (e.g., verified healthcare professionals in the United States). In precision oncology, expertly curated databases like the molecular oncology almanac (MOAlmanac), OncoKB, CIViC, and MyCancerGenome provide structured evidence to guide treatment decisions across diverse genomic and disease contexts.[19–22] Kreimeyer et al. demonstrated the potential of an RAG-based pipeline in retrieving treatment relationships from OncoKB; however, the study lacked real-world clinical evaluation.[23] Similarly, Berman et al. highlighted the potential of integrating LLMs with RAG in molecular tumor board (MTB) workflows while also identifying key limitations, such as constrained scope of query inputs and generalizability.[24]

Here, we introduce an open-source RAG-LLM approach that enables accurate and approval-derived therapy recommendations based on patients' genomic biomarkers, cancer type, treatment history, and other clinically relevant information for treatment planning. Our RAG-LLM method leverages the MOAlmanac, an expert-curated clinicogenomic interpretation database that compiles the latest precision oncology knowledge regarding relationships between molecular features and clinical actionability,[19] to improve queries of retrieving appropriate FDA-approved biomarker-based oncology therapies. We evaluate this approach on a synthetic dataset, finding that it retrieves the approved therapeutic option based on the provided clinical information with high accuracy. We then benchmark this approach on real-world questions provided by practicing oncologists, finding that this approach accurately identifies approved therapies for different diagnoses, clinical histories, and known genomic alterations.

RESULTS

Prompt optimization and benchmarking of LLMs

Given the critical role of prompt optimization in enhancing LLM performance, we first evaluated whether designing a specific prompt structure would optimize the accuracy of a representative LLM (Mistral Nemo 12B) in accurately retrieving FDA-approved biomarker-based ("precision") oncology therapies.[25] We tested four distinct prompt strategies (see STAR MethodsTable 1). All evaluated queries were formulated using the structured data from MOAlmanac v2024-04-11 (Tables S1 and S2). To compare prompt effectiveness, we measured partial match accuracy, defined as the proportion of retrieved therapies that matched ground-truth FDA-approved therapies (see STAR Methods). Among the four strategies, the basic prompt demonstrated the highest accuracy in retrieving FDA-approved therapies (Figure 1A). The basic prompt achieved an accuracy of 82.9%, outperforming the second (69.7%), third (79.9%), and fourth (69.7%) strategies.

The superior performance of the basic prompting strategy was also consistently observed in other LLMs, particularly those of larger sizes ($n = 7$; Figure 1B). Across all models tested, GPT-4o achieved the highest accuracy for both the basic and combined prompt strategies of 89.3% and 85.9%, respectively. Thus, we selected the basic prompt and used GPT-4o for all subsequent evaluations.

RAG-integrated LLM outperforms LLM in therapy predictions

We next hypothesized that structured data augmentation would improve the accuracy and reliability of LLM for FDA-approved therapy predictions compared to unstructured data. We thus evaluated GPT-4o's performance using both unstructured and structured data formats. Each approach was benchmarked on synthetic prompts derived from single entities in MOAlmanac April 2024 ($n = 234$ entities) and October 2025 releases ($n = 651$ entities) to (1) incorporate the latest FDA approvals and (2) assess whether the context augmentation performance was agnostic to database version (Tables S3 and S4).

Without RAG-provided context, model accuracy ranged from 62% to 89% on synthetic queries from the 2024 release. However, augmenting the model with unstructured text data

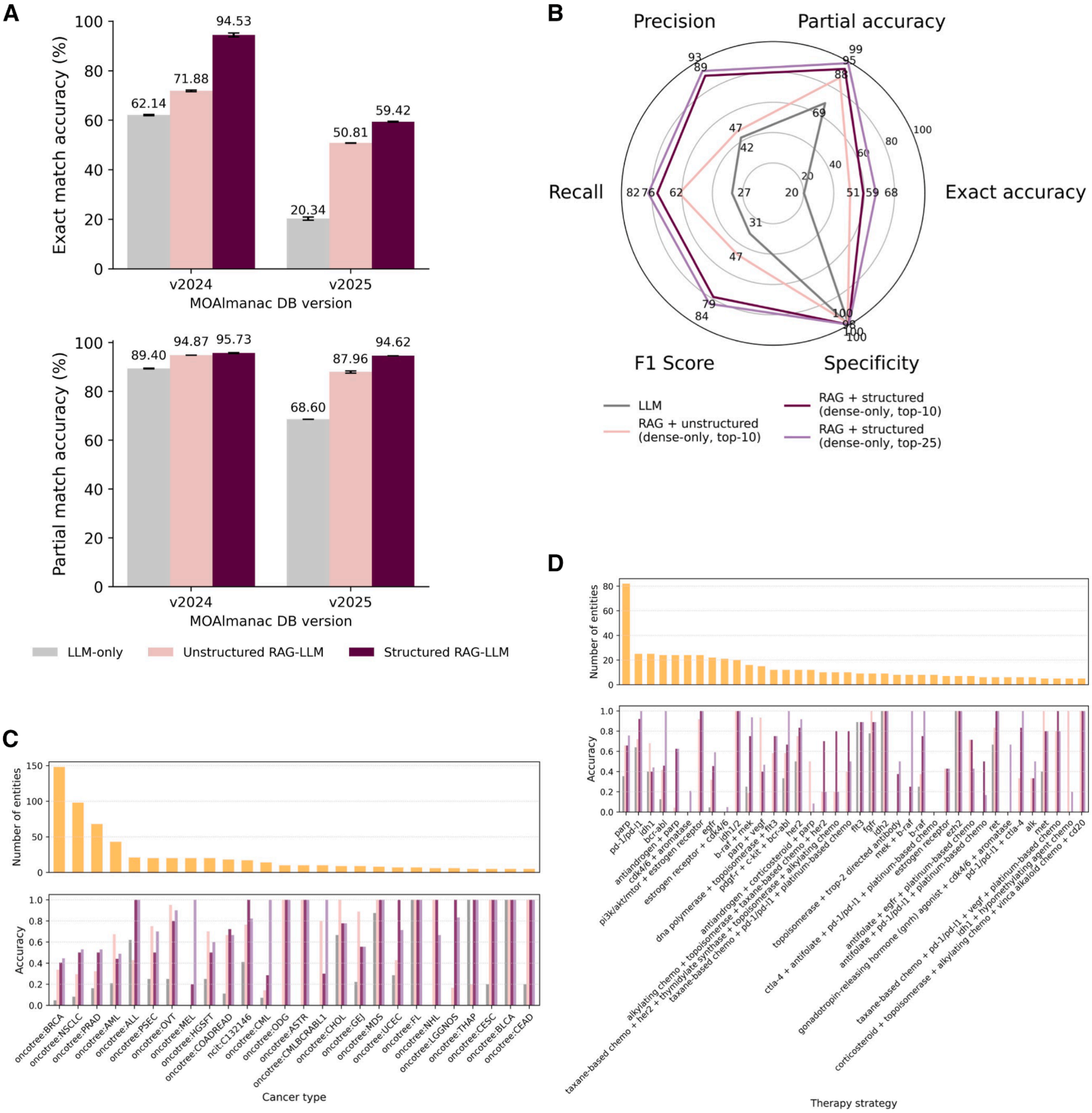


Figure 2. Enhancement through RAG using unstructured and structured datasets

(A) Exact match and partial match accuracies from LLM-only and RAG-LLM augmented with unstructured and structured datasets derived from MOAlmanac 2024 April and 2025 October releases.

(B) Average precision, recall, F1-score, specificity, and accuracies from RAG-LLM with unstructured and structured data augmentation (data from the October 2025 release).

(C) Top: Number of entities in the MOAlmanac database (2025) across cancer types. Bottom: Exact match accuracies from unstructured and structured RAG-LLM approaches across cancer types.

(D) Top: Number of entities in the MOAlmanac database (2025) across therapy strategies. Bottom: Exact match accuracies from unstructured and structured RAG-LLM approaches across therapy strategies. Only the cancer types and therapy strategies with greater than or equal to 5 relationships are displayed in (C) and (D), respectively. Abbreviations in (C) and (D) are defined in Table S12.

significantly improved performance, increasing accuracy to 72%–95% (exact match: $\chi^2(1) = 48.74, p = 5.85 \times 10^{-12}$, Cohen's g = 0.44; partial match: $\chi^2(1) = 21.57, p = 4.55 \times 10^{-6}$, Cohen's g = 0.35; McNemar's test) (Figure 2A). Integrating structured data further enhanced performance compared to unstructured data augmentation, yielding an accuracy of 95%–96% (exact match: $\chi^2(1) = 190.95, p = 7.90 \times 10^{-43}$, Cohen's g = 0.73; McNemar's test). A similar pattern was observed using synthetic queries derived from the 2025 release, with the structured RAG-LLM consistently outperforming all other modes. However, the exact match accuracy from the 2025 release (59%) was lower than that from the 2024 release (95%), likely due to the nearly 3-fold expansion of relationships of FDA approvals in the latest version.[26]

In addition to accuracy, structured data augmentation markedly improved other key performance metrics on queries derived from the 2025 release (Figure 2B). Specifically, precision and F1-score significantly improved by approximately 90% and 69%,

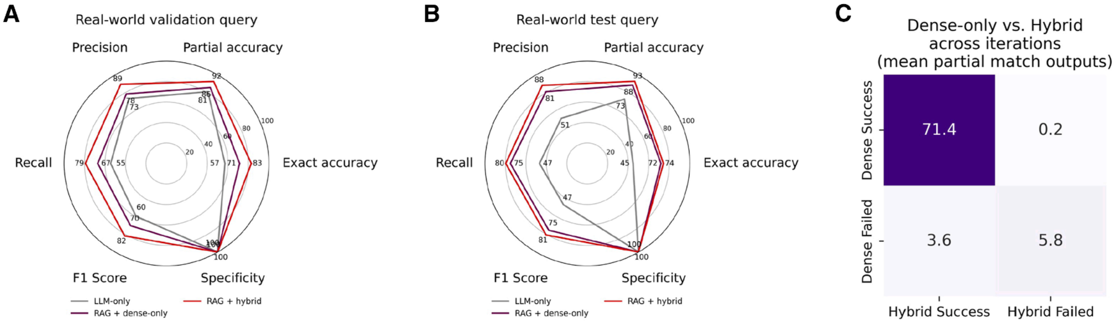


Figure 3. Enhancement through hybrid retrieval and real-world evaluation

(A) Dense-only vs. hybrid search RAG-LLM performance on real-world validation queries (n = 21).

(B) Dense-only vs. hybrid search RAG-LLM performance on real-world test queries (n = 81).

(C) Dense-only vs. hybrid search average partial match counts tbl1fnbacross all five iterations on real-world test queries.

respectively, with structured data augmentation compared to the unstructured data augmented model (mean precision: 47%–89%, $W = 15.0$, $p = 3.13 \times 10^{-2}$; mean F1-score: 47%–79%, $W = 15.0$, $p = 3.13 \times 10^{-2}$; one-sided Wilcoxon signed-rank test). RAG augments queries to an LLM by including relevant records from the supporting structured database with the query, and we observed that including more records (from 10 to 25) further improved model performance (Figure 2B). Together, these results demonstrate the role of the RAG approach in enhancing the model's ability to provide more precise and reliable FDA-approved therapy predictions, with structured data augmentation and a higher top-k further optimizing performance.

Structured data augmentation also consistently outperformed the unstructured approach across various cancer types and therapy categories (Figures 2C and 2D). For example, an increase in accuracy of at least 30% was observed when transitioning from unstructured to structured augmentation for several cancer types. Similar improvements were observed for therapy strategies. The model's performance was suboptimal for certain breast cancer combination therapies (aromatase inhibitor + CDK4/6 inhibitor and selective estrogen receptor modulator + CDK4/6 inhibitor), primarily due to incomplete retrieval of approved therapies as a combination strategy.

Of note, the structured data augmented model performed well in some cases where structured clinicogenomic relationships were densely represented. For example, antiandrogen + PARP inhibition is approved for prostate cancer patients with mutations in any of 14 homologous recombination repair (HRR) genes. For this context, the RAG-integrated model achieved a high exact match accuracy of 63%, relative to 4% from the unstructured RAG-LLM and 0% with the LLM alone. This highlights the advantage of structured augmentation in effectively recalling genomic knowledge statements for this approval. Together, these results demonstrate that structured data enables the model to better capture complex relationships between therapies and their approved indications, thereby improving performance across diverse clinical contexts.

RAG-LLM accurately predicts therapies in real-world scenarios

To assess the real-world applicability of the RAG-LLM approach, we collected 102 clinical queries from 15 oncologists spanning multiple specialties across four institutions (see STAR Methods). These queries primarily focused on precision oncology therapies given a specific cancer type and biomarker(s), encompassing 64 genomic biomarkers and 35 cancer types across diverse organ systems (Figures S1A–S1B). Of these, 21 queries were used for validation, and 81 queries were used as the final test set (Table S5). We used the structured context database constructed based on the October 2025 MOAlmanac release for RAG.

For real-world queries where no on-label FDA-approved therapies currently exist, we evaluated whether RAG-LLM incorrectly retrieved off-label treatment options. We found that, compared to the baseline retrieval model ("strategy R1"; see STAR Methods), a model augmented with an out-of-scope JSON schema ("strategy R2") exhibited improved performance on real-world validation queries (exact match: $\chi^2(1) = 12.8$, $p = 3.47 \times 10^{-4}$, Cohen's g = 0.56; partial match: $\chi^2(1) = 12.8$, $p = 3.47 \times 10^{-4}$, Cohen's g = 0.56; McNemar's test; Figure S2).

We also evaluated whether incorporation of a hybrid lexical and semantic search-based retrieval approach ("Hybrid"; see STAR Methods)[27] would enhance performance on real-world queries. On an initial set of real-world validation queries (n = 21), all major evaluation metrics (e.g., accuracy, precision, recall, and F1) improved by 8%–18% compared to the baseline "Dense-only" approach (Figure 3A). Similarly, on our final held-out real-world test set (n = 81), RAG-LLM with the hybrid search achieved the best accuracy (93%) and evaluation metrics increased by 3%–9% compared to the dense-only retrieval method, with partial match accuracy increasing significantly (partial match: $\chi^2(1) = 1.0$, $p = 1.53 \times 10^{-4}$, Cohen's g = 0.90; McNemar's test; Figure 3B). Across all iterations of the pipeline, the hybrid search consistently produced a higher number of partially correct predictions—on average, about four more than the dense-only approach on the real-world test queries (Figure 3C).

Representative interpretation challenges in real-world cases

These context-augmented LLM frameworks occasionally misinterpreted the retrieved contexts, especially for queries asking about first-line treatments. For instance, in response to the question "*73 yo patient with a new diagnosis of metastatic urothelial carcinoma. Molecular testing reveals fgfr3 g370c mutations and* *her2 ihc 3+ status. What options would you recommend for systemic treatment?*'', both dense-only and hybrid search modes predicted erdafitinib, even though the context retrieved explicitly stated that it was indicated in tumors after progression on prior systemic therapy. Similarly, in response to the question ''*I have a patient with newly diagnosed aml that has mutations in tp53 and idh1. Are there any therapies that I should be thinking about in this setting?*'', the model did not return any drugs under both approaches, even when the context stated that ivosidenib was indicated. While the LLM occasionally misinterpreted the retrieved contexts, the hybrid search demonstrated an advantage in scenarios where dense-only retrieval prioritized clinically irrelevant yet semantically similar contexts (Table 1).

Exploring RAG-LLM's generalizability beyond FDA approvals

Lastly, to assess whether the RAG-LLM precision oncology framework could be generalizable beyond FDA contexts, we curated and integrated the CIViC database into the framework and compared its performance to the original pipeline augmented with MOAlmanac database. Like MOAlmanac, CIViC is continuously updated and contains knowledge statements originating from varied levels of evidence ranging from regulatory approvals to preclinical evidence. We additionally expanded the ground-truth therapies to include expert-curated ones that were initially outside the scope of our study (e.g., not an FDA-approved or biomarker-guided therapy). CIViC contained 1,552 biomarker-drug-cancer relationships, encompassing 170 cancer types and 660 genomic biomarkers. Nevertheless, the MOAlmanac-augmented model yielded higher overall accuracy (53–76%) than the CIViC-augmented model (37%–62%) across all real-world test queries, primarily due to differences in FDA-approval entries between the knowledge base (Figure S3A). In contrast, for real-world test queries involving off-label or out-of-scope options (n = 10), the CIViC-augmented model's performance surpassed the MOAlmanac-augmented model across all evaluation metrics (Figure S3B; Table S6). Thus, this demonstrates a tradeoff between breadth and precision, as performance can degrade with larger and less-targeted context database.[28,29] To maximize accessibility and exploration, the RAG-LLM implementations presented herein are available at llm.moalmanac.org.

DISCUSSION

Broadly, this study investigated the potential of RAG-LLM to guide precision oncology decision-making across diverse use cases. We found that prompt optimization, particularly using a simple prompt, markedly improved LLM accuracy in retrieving FDA-approved biomarker-driven therapies, with structured data augmentation further boosting performance to achieve near-perfect accuracy in therapy predictions. Real-world oncologist queries validated the model's ability to retrieve relevant therapies. Together, these findings highlight the transformative potential of RAG-LLM for precision oncology support.

Prompt design is essential for adapting general-purpose LLMs to specific applications.[30] We observed improved performance with a basic prompt over a combined prompt, likely due to factors, such as the impact of enforcing a JSON format on LLM's reasoning, the addition of personas for objective tasks, and inherent limitations in an LLM's ability to follow multiple requirements simultaneously.[31–33] While performance varied across models, the basic prompt consistently improved performance for larger models as per prior studies.[34] RAG implementation significantly outperformed the LLM-only approach in predicting FDA-approved therapies, underscoring the value of structured data in enhancing the reliability of the framework, making it better suited for integration into clinical decision-making workflows.

However, challenges to these approaches remained in real-world scenarios. When no FDA-approved therapy existed for a case, the model often misattributed treatments approved for other cancers or biomarkers. To address this, we implemented a predefined JSON schema instructing the model to return ''no matches'' when appropriate (strategy R2), which improved performance on real-world queries. Even so, the model under strategy R1 (no explicit schema) outperformed R2 in some queries with FDA-approved drugs. These findings align with prior studies in other domains demonstrating that imposing constraints on LLM outputs can reduce correctness for reasoning and classification tasks,[31,35] highlighting a tradeoff between conservatism and flexibility in retrieval and reasoning, driven by the model's adherence to output constraints. These behaviors warrant further systematic exploration, particularly in the context of precision oncology.

Also, the incorporation of a hybrid retrieval approach further improved performance on real-world queries, compared to the semantic-only approach. While the hybrid approach does not explicitly model structural relationships, it partially preserves these relationships by ensuring that the most relevant entities are accurately identified and prioritized in the retrieved context space. This allows the RAG-LLM to capture fine-grained determinants of precision oncology therapy recommendations. Structure-aware alternatives, including knowledge graph-based approach, should be explored further to enhance structural reasoning improve structural reasoning, although this usually requires extensive ontology engineering and often still relies on embeddings for semantic flexibility.[36–39]

There is ongoing collective interest in adapting retrieval concepts in medicine, and RAG-based LLMs like OpenEvidence have emerged as key tools for a variety of medical search queries.[40,41] While they can support physicians in making treatment decisions, the knowledge bases used for such general-purpose approaches are broad, which can lead to challenges like missing critical evidence (e.g., FDA drug labels) and retrieving less relevant sources.[42] In contrast, precision oncology decision-support AI tools may continue to benefit from augmentation with expertly curated, regularly updated, and domain-specific databases that are not typically part of general-purpose medical models, which can enhance the accuracy, relevance, and efficiency of the decision-making process tailored to these specific uses.

Ultimately, deploying LLM-based tools in clinical workflows demands strategies to mitigate risks arising from their widespread adoption. While our RAG-LLM framework is designed to guide oncologists in recommending biomarker-driven therapies, its deployment will inevitably increase the risk of misuse or vulnerability to adversarial prompts.[43] Furthermore, its integration into clinical settings will present unique challenges,

Table 1. Representative real-world test queries that failed across all iterations, stratified by failure types and retrieval strategies

Case	Real-world test query	Ground-truth[a]	Dense-LLM output therapy[b]	Hybrid-LLM output therapy[b]	Reason
Both failed	*73 yo patient with a new diagnosis of metastatic urothelial carcinoma. molecular testing reveals fgfr3 g370c mutations and her2 ihc 3+ status. what options would you recommend for systemic treatment?*	none	erdafitinib	erdafitinib	LLM misinterpretation (erdafitinib is not approved as first-line treatment)
	I have a patient with newly diagnosed aml that has mutations in tp53 and idh1. are there any therapies that i should be thinking about in this setting?	ivosidenib	none	none	LLM misinterpretation (LLM failed to retrieve ivosidenib)
	I have a patient with metastatic ascending colon adenocarcinoma, that is kras and braf wt, what first line targeted therapy therapy options are there in addition to folfox or folfiri	none	panitumumab	panitumumab	LLM misinterpretation (panitumumab is indicated as monotherapy following disease progression after prior treatment)
	My patient with prostate cancer had genomic testing on their tumor and her2 amplification was detected. can I give them trastuzumab deruxtecan or another her2-targeted therapy?	trastuzumab deruxtecan	none	none	Context retrieved, LLM misinterpretation and query ambiguity (trastuzumab deruxtecan is indicated for any unresectable or metastatic solid tumor that is HER2-positive)
Dense failed, Hybrid success	*I have a patient with metastatic hormone receptor-positive, her2-low (ihc 1 +) breast cancer who progressed on 1st-line treatment with letrozole and ribociclib. circulating tumor dna analysis reveals a pathogenic deletion in pten. what are approved treatment options in the second-line setting?*	capivasertib + fulvestrant, trastuzumab deruxtecan	none	capivasertib + fulvestrant	Dense-only: Context missing
	66 yo patient with previously resected clear-cell renal cell carcinoma, presents now with local unresectable recurrence. genetic testing is notable for vhl and pbrm1 mutations, as well as msi-h. what oprions are recommended for systemic treatment?	none	pembrolizumab	none	Dense-only: LLM misinterpretation (somatic alterations are not predictive of response to treatment in ccRCC)
	I have a male patient, never smoker, with de novo metastatic lung adenocarcinoma involving the cns. molecular testing reveals an alk rearrangement and pd-l1 tps 15%. what are the first-line treatment options?	alectinib, brigatinib, crizotinib, lorlatinib	none	alectinib, brigatinib, crizotinib, lorlatinib	Dense-only: Context missing
Dense success, Hybrid failed	No queries fall under this category				

[a] Ground-truth FDA-approved precision oncology therapies; “none” denotes no approved precision oncology indications for the corresponding query.
[b] Therapies generated by the RAG-LLM pipelines; “none” denotes no therapies returned.

including patient data privacy, potential harm from errors and biases, and protection of intellectual property and proprietary data.[44] Indeed, one potential solution is to deploy tools through a community-level, private, and HIPAA-compliant LLMs, which allows users to maintain the protection confidential information.[45] Since our RAG-LLM framework's retrieval module is external and does not rely on the model's internal weights, we postulate that it is largely model-agnostic, allowing integration with any hospital's HIPAA-compliant LLM workflow.

To further mitigate bias, particularly regional bias arising from variations in demographics and treatment efficacy, a practical strategy is to integrate regional databases into the framework as context sources, allowing users or the system itself to prioritize region-specific evidence when generating treatment recommendations. Our RAG-LLM can toggle between FDA and European Medicines Agency (EMA) approvals, which could be expanded to other regions for world-wide utility. Complementary approaches, such as bias auditing and embedding calibration, may also improve fairness and equity across diverse patient populations.[46,47] Thus, successful implementation of this framework is dependent on a comprehensive strategy that addresses privacy, regulatory, and bias-related challenges in the high-stakes realm of oncology patient care.

Finally, given that our RAG-LLM approach requires fewer computational resources and greater adaptability than fine tuning-based approaches, it may also facilitate greater access to precision oncology, particularly in supporting non-academic oncologists with limited availability of clinical expertise in a dynamically changing domain of medical oncology. In addition, this flexible framework allows users to control the model's conservatism depending on their needs. For rare cancers, users may select a database comprising both approved and under investigation drugs or specify a less stringent prompt to return a larger number of potentially relevant results beyond exact matches. Furthermore, as the current process of staying up-to-date with regulatory approvals is highly fragmented, demanding toggling between multiple sources of information, our framework could also serve as a unified and reliable query layer across an otherwise fractured system. Lastly, given that our study has experimented with various prompting and retrieval strategies in the context of precision oncology medicine, it may also serve as guidance for future work on deploying context-augmented LLM pipelines to support cancer patients' treatment decisions in the clinical setting.

Limitations of the study

The primary focus of our study was to develop a reliable LLM-based framework that can serve as a decision-support tool for physicians in oncology. However, achieving the necessary accuracy and reliability for clinical deployment of this approach requires further refinements. Potential improvements include experimenting with different embedding models, chain-of-thought prompting, and uncertainty or confidence quantification.[48–51] Future work will also involve expanding the scope of the study and the context database by integrating clinical guidelines and clinical trial data to provide treatment options available to physicians beyond FDA-approved drugs, potentially with direct integration into electronic health records. To that end, further experimentation with strategies to mitigate the tradeoff between breadth and precision will be necessary, such as refining the retrieval step by performing a broad or individual searches across FDA-approved and under investigation drugs followed by evidence-based re-ranking. These enhancements will be essential for developing a clinically reliable, real-world applicable, and robust LLM-driven decision-support tool.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Eliezer Van Allen (eliezerm_vanallen@dfci.harvard.edu).

Materials availability

This study did not generate any new unique reagents.

Data and code availability

All the scripts and datasets for running the LLM-only and RAG-LLM pipelines, along with the corresponding outputs generated in this study, are publicly available at https://github.com/hjjshine/rag-llm-cancer-paper. Pipeline usage instructions are provided in the README file.

ACKNOWLEDGMENTS

This work was supported by P50CA272390, DOD W81XWH-21-PCRP-DSA, DOD HT94252410415, and the Mark Foundation Emerging Leader award.

AUTHOR CONTRIBUTIONS

Conceptualization, H.J., Y.T., S.J., B.R., and E.V.A.; data curation, H.J., Y.T., B.R., E.L.B., F.L.F.C., D.Y.G., A.C.J., C.L., S.D.M., M.N., T.A.O., T.P., E.M.P., E.S., D.D.Y., R.G., A.K.T., and E.V.A.; methodology, H.J. and Y.T.; formal analysis and visualization, H.J.; software: H.J., Y.T., and S.Y.C.; supervision, R.G., A.K.T., and E.V.A.; writing – original draft, H.J.; writing – review/editing, H.J., Y.T., S.J., B.R., and E.V.A.

DECLARATION OF INTERESTS

R.G. has equity in Google, Microsoft, Amazon, Apple, Moderna, Pfizer, and Vertex Pharmaceuticals; his spouse is employed by Carrum Health. E.S. receives research funding from Genentech/imCORE and Oncohost. C.L. receives research funding from Genentech/imCORE. E.M.V.A. holds consulting roles with Enara Bio, Manifold Bio, Monte Rosa, Novartis Institute for Biomedical Research, Serinus Bio, and TracerBio; he previously held consulting roles with Tango Therapeutics, Invitae, Syapse, Janssen, Genome Medical, Genomic Life, and Riva Therapeutics; he receives research support from Novartis, Bristol-Myers Squibb, Sanofi, and NextPoint; he has equity in Tango Therapeutics, Genome Medical, Genomic Life, Enara Bio, Manifold Bio, Microsoft, Monte Rosa, Riva Therapeutics, Serinus Bio, Syapse, and TracerDx; he received travel reimbursement from Roche and Genentech; he has filed institutional patents on chromatin mutations and immunotherapy response, and methods for clinical interpretation, and provides intermittent legal consulting on patents for Foaley & Hoag.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- METHOD DETAILS
  - ○ Database
  - ○ Prompt engineering

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.ccell.2025.12.017.

Received: July 15, 2025
Revised: November 3, 2025
Accepted: December 16, 2025

REFERENCES

1. Suehnholz, S.P., Nissan, M.H., Zhang, H., Kundra, R., Nandakumar, S., Lu, C., Carrero, S., Dhaneshwar, A., Fernandez, N., Xu, B.W., et al. (2024). Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discov. *14*, 49–65. https://doi.org/10.1158/2159-8290.CD-23-0467.
2. Chow-White, P., Ha, D., and Laskin, J. (2017). Knowledge, attitudes, and values among physicians working with clinical genomics: a survey of medical oncologists. Hum. Resour. Health *15*, 42. https://doi.org/10.1186/s12960-017-0218-z.
3. Jin, Q., Wang, Z., Floudas, C.S., Chen, F., Gong, C., Bracken-Clarke, D., Xue, E., Yang, Y., Sun, J., and Lu, Z. (2024). Matching patients to clinical trials with large language models. Nat. Commun. *15*, 9074. https://doi.org/10.1038/s41467-024-53081-z.
4. Cerami, E., Trukhanov, P., Paul, M.A., Hassett, M.J., Riaz, I.B., Lindsay, J., Mallaber, E., Klein, H., Gungor, G., Galvin, M., et al. (2024). MatchMiner-AI: An Open-Source Solution for Cancer Clinical Trial Matching. Preprint at arXiv. https://doi.org/10.48550/arXiv.2412.17228.
5. Wong, C., Zhang, S., Gu, Y., Moung, C., Abel, J., Usuyama, N., Weerasinghe, R., Piening, B., Naumann, T., Bifulco, C., et al. (2023). Scaling clinical trial matching using large language models: A case study in oncology. Preprint at arXiv. https://doi.org/10.48550/arXiv.2308.02180.
6. Van Veen, D., Van Uden, C., Blankemeier, L., Delbrouck, J.-B., Aali, A., Bluethgen, C., Pareek, A., Polacin, M., Reis, E.P., Seehofnerová, A., et al. (2024). Adapted large language models can outperform medical experts in clinical text summarization. Nat. Med. *30*, 1134–1142. https://doi.org/10.1038/s41591-024-02855-5.
7. Katz, U., Cohen, E., Shachar, E., Somer, J., Fink, A., Morse, E., Shreiber, B., and Wolf, I. (2024). GPT versus resident physicians — A benchmark based on official board scores. NEJM AI *1*, 101–109. https://doi.org/10.1056/aidbp2300192.
8. Benary, M., Wang, X.D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., et al. (2023). Leveraging large language models for decision support in personalized oncology. JAMA Netw. Open *6*, e2343689. https://doi.org/10.1001/jamanetworkopen.2023.43689.
9. Xu, S., Most, A., Chase, A., Hedrick, T., Murray, B., Keats, K., Smith, S., Barreto, E., Liu, T., and Sikora, A. (2024). Large language models management of complex medication regimens: a case-based evaluation. Preprint at medRxiv. https://doi.org/10.1101/2024.07.03.24309889.
10. Elemento, O., Khozin, S., and Sternberg, C.N. (2025). The use of artificial intelligence for cancer therapeutic decision-making. NEJM AI *2*. https://doi.org/10.1056/aira2401164.
11. Verlingue, L., Boyer, C., Olgiati, L., Brutti Mairesse, C., Morel, D., and Blay, J.-Y. (2024). Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. Lancet Reg. Health. Eur. *46*, 101064. https://doi.org/10.1016/j.lanepe.2024.101064.
12. Jeffrey, C., Marc, M., Orion, W., Dawn, L., Daniel, K., and Van Durme, B. (2024). Dated data: Tracing knowledge cutoffs in Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2403.12958.
13. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Kuttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Neural Inf Process Syst *abs/2005*. Preprint at arXiv. https://doi.org/10.48550/arXiv.2005.11401.
14. Shanghua, G., Richard, Z., Zhenglun, K., Ayush, N., Xiaorui, S., Curtis, G., Theodoros, T., and Marinka, Z. (2025). TxAgent: An AI agent for therapeutic reasoning across a universe of tools. Preprint at arXiv. https://doi.org/10.48550/arXiv.2503.10970.
15. Zakka, C., Shad, R., Chaurasia, A., Dalal, A.R., Kim, J.L., Moor, M., Fong, R., Phillips, C., Alexander, K., Ashley, E., et al. (2024). Almanac - retrieval-augmented language models for clinical medicine. NEJM AI *1*, 104435. https://doi.org/10.1056/aioa2300068.
16. Ferber, D., El Nahhas, O.S.M., Wölflein, G., Wiest, I.C., Clusmann, J., Leßmann, M.-E., Foersch, S., Lammert, J., Tschochohei, M., Jäger, D., et al. (2025). Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. Nat. Cancer *6*, 1337–1349. https://doi.org/10.1038/s43018-025-00991-6.
17. Towhidul Islam Tonmoy, S., Mehedi Zaman, S.M., Vinija, J., Anku, R., Vipula, R., Aman, C., and Amitava, D. (2024). A comprehensive survey of hallucination mitigation techniques in Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.01313.
18. Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., and Shoham, Y. (2023). In-Context retrieval-augmented language models. Trans. Assoc. Comput. Linguist. *11*, 1316–1331. https://doi.org/10.1162/tacl_a_00605.
19. Reardon, B., Moore, N.D., Moore, N.S., Kofman, E., AlDubayan, S.H., Cheung, A.T.M., Conway, J., Elmarakeby, H., Imamovic, A., Kamran, S.C., et al. (2021). Integrating molecular profiles into clinical frameworks through the Molecular Oncology Almanac to prospectively guide precision oncology. Nat. Cancer *2*, 1102–1112. https://doi.org/10.1038/s43018-021-00243-3.
20. Griffith, M., Spies, N.C., Krysiak, K., McMichael, J.F., Coffman, A.C., Danos, A.M., Ainscough, B.J., Ramirez, C.A., Rieke, D.T., Kujan, L., et al. (2017). CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat. Genet. *49*, 170–174. https://doi.org/10.1038/ng.3774.
21. Holt, M.E., Mittendorf, K.F., LeNoue-Newton, M., Jain, N.M., Anderson, I., Lovly, C.M., Osterman, T., Micheel, C., and Levy, M. (2021). My Cancer Genome: Coevolution of precision oncology and a molecular oncology knowledgebase. JCO Clin. Cancer Inform. *5*, 995–1004. https://doi.org/10.1200/CCI.21.00084.
22. Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A precision oncology knowledge base. JCO Precis. Oncol. https://doi.org/10.1200/PO.17.00011.
23. Kreimeyer, K., Canzoniero, J.V., Fatteh, M., Anagnostou, V., and Botsis, T. (2024). Using retrieval-augmented generation to capture molecularly-driven treatment relationships for precision oncology. Stud. Health Technol. Inform. *316*, 983–987. https://doi.org/10.3233/SHTI240575.
24. Berman, E., Sundberg Malek, H., Bitzer, M., Malek, N., and Eickhoff, C. (2025). Retrieval augmented therapy suggestion for molecular tumor boards: Algorithmic development and validation study. J. Med. Internet Res. *27*, e64364. https://doi.org/10.2196/64364.
25. Reynolds, L., and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (ACM)). https://doi.org/10.1145/3411763.3451760.
26. Moalmanac-db (Github). github.com/vanallenlab/moalmanac-db/releases/tag/v.2025-10-03.

27. Sawarkar, K., Mangal, A., and Solanki, S.R. (2024). Blended RAG: Improving RAG (retriever-Augmented Generation) accuracy with semantic search and hybrid query-based Retrievers. Preprint at arXiv. https://doi.org/10.48550/arXiv.2404.07220.

28. Liu, Y., Huang, L., Li, S., Chen, S., Zhou, H., Meng, F., Zhou, J., and Sun, X. (2023). RECALL: A benchmark for LLMs robustness against external counterfactual knowledge. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.08147.

29. Xu, R., Zhuang, Y., Yu, Y., Wang, H., Shi, W., and Yang, C. (2025). RAG in the Wild: On the (in)effectiveness of LLMs with mixture-of-knowledge retrieval augmentation. Preprint at arXiv. https://doi.org/10.48550/arXiv.2507.20059.

30. Pranab, S., Singh, A.K., Sriparna, S., Vinija, J., Samrat, M., and Aman, C. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. Preprint at arXiv. https://doi.org/10.48550/arXiv.2402.07927.

31. Tam, Z.R., Wu, C.-K., Tsai, Y.-L., Lin, C.-Y., Lee, H.-Y., and Chen, Y.-N. (2024). Let me speak freely? A study on the impact of format restrictions on performance of large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2408.02442.

32. Zheng, M., Pei, J., Logeswaran, L., Lee, M., and Jurgens, D. (2023). When ''A helpful assistant'' is not really helpful: Personas in system prompts do not improve performances of Large Language Models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2311.10054.

33. Chenyang, Y., Yike, S., Qianou, M., Liu, M.X., Christian, K., and Tongshuang, W. (2025). What prompts don't say: Understanding and managing underspecification in LLM prompts. Preprint at arXiv. https://doi.org/10.48550/arXiv.2505.13360.

34. He, J., Rungta, M., Koleczek, D., Sekhon, A., Wang, F.X., and Hasan, S. (2024). Does prompt formatting have any impact on LLM performance?. Preprint at arXiv. https://doi.org/10.48550/arXiv.2411.10541.

35. Long, D.X., Ngoc, H.N., Sim, T., Dao, H., Joty, S., Kawaguchi, K., Chen, N.F., and Kan, M.-Y. (2024). LLMs are biased towards output formats! Systematically evaluating and mitigating output format bias of LLMs. Preprint at arXiv. https://doi.org/10.48550/arXiv.2408.08656.

36. Kuhn, M., Letunic, I., Jensen, L.J., and Bork, P. (2016). The SIDER database of drugs and side effects. Nucleic Acids Res. *44*, D1075–D1079. https://doi.org/10.1093/nar/gkv1075.

37. Liang, K., Meng, L., Liu, M., Liu, Y., Tu, W., Wang, S., Zhou, S., Liu, X., Sun, F., and He, K. (2024). A survey of knowledge graph reasoning on graph types: Static, dynamic, and multi-modal. IEEE Trans. Pattern Anal. Mach. Intell. *46*, 9456–9478. https://doi.org/10.1109/TPAMI.2024.3417451.

38. Chandak, P., Huang, K., and Zitnik, M. (2023). Building a knowledge graph to enable precision medicine. Sci. Data *10*, 67. https://doi.org/10.1038/s41597-023-01960-3.

39. Lin, Y., He, J., Chen, J., Zhu, X., Zheng, J., and Bo, T. (2025). BioGraphFusion: graph knowledge embedding for biological completion and reasoning. Bioinformatics *41*, btaf408. https://doi.org/10.1093/bioinformatics/btaf408.

40. Hurt, R.T., Stephenson, C.R., Gilman, E.A., Aakre, C.A., Croghan, I.T., Mundi, M.S., Ghosh, K., and Edakkanambeth Varayil, J. (2025). The use of an artificial intelligence platform OpenEvidence to augment clinical decision-making for primary care physicians. J. Prim. Care Community Health *16*, 21501319251332215. https://doi.org/10.1177/21501319251332215.

41. Akpınar, H. (2025). Comparison of responses from different artificial intelligence-powered chatbots regarding the All-on-four dental implant concept. BMC Oral Health *25*, 922. https://doi.org/10.1186/s12903-025-06294-7.

42. Shengming, Z., Yuchen, S., Yuheng, H., Jiayang, S., Zhijie, W., Chengcheng, W., and Lei, M. (2024). Understanding the design decisions of retrieval-Augmented Generation systems. Preprint at arXiv. https://doi.org/10.48550/arXiv.2411.19463.

43. Erfan, S., Al Mamun, M.A., Yu, F., Pedram, Z., Yue, D., and Nael, A.-G. (2023). Survey of vulnerabilities in large Language Models revealed by adversarial attacks. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.10844.

44. Ong, J.C.L., Chang, S.Y.-H., William, W., Butte, A.J., Shah, N.H., Chew, L.S.T., Liu, N., Doshi-Velez, F., Lu, W., Savulescu, J., and Ting, D.S.W. (2024). Ethical and regulatory challenges of large language models in medicine. Lancet Digit. Health *6*, e428–e432. https://doi.org/10.1016/S2589-7500(24)00061-X.

45. Umeton, R., Kwok, A., Maurya, R., Leco, D., Lenane, N., Willcox, J., Abel, G.A., Tolikas, M., and Johnson, J.M. (2024). GPT-4 in a cancer center — institute-wide deployment challenges and lessons learned. NEJM AI *1*. https://doi.org/10.1056/aics2300191.

46. Wu, X., Li, S., Wu, H.-T., Tao, Z., and Fang, Y. (2024). Does RAG introduce unfairness in LLMs? Evaluating fairness in retrieval-Augmented Generation systems. Int. Conf. Comput. Linguistics *19804*, 10021–10036. https://doi.org/10.48550/arXiv.2409.19804.

47. Kim, T., Springer, J.M., Raghunathan, A., and Sap, M. (2025). Mitigating bias in RAG: Controlling the embedder. In Findings of the Association for Computational Linguistics: ACL 2025, W. Che, J. Nabende, E. Shutova, and M.T. Pilehvar, eds. (Association for Computational Linguistics), pp. 18999–19024. https://doi.org/10.18653/v1/2025.findings-acl.974.

48. Myers, S., Miller, T.A., Gao, Y., Churpek, M.M., Mayampurath, A., Dligach, D., and Afshar, M. (2025). Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. J. Am. Med. Inform. Assoc. *32*, 357–364. https://doi.org/10.1093/jamia/ocae308.

49. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Preprint at arXiv. https://doi.org/10.48550/arXiv.2201.11903.

50. Gao, Y., Myers, S., Chen, S., Dligach, D., Miller, T., Bitterman, D.S., Chen, G., Mayampurath, A., Churpek, M.M., and Afshar, M. (2025). Uncertainty estimation in diagnosis generation from large language models: next-word probability is not pre-test probability. JAMIA Open *8*, ooae154. https://doi.org/10.1093/jamiaopen/ooae154.

51. Chen, J., and Mueller, J. (2024). Quantifying Uncertainty in Answers from any Language Model and Enhancing their Trustworthiness. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (Association for Computational Linguistics), pp. 5186–5200. https://doi.org/10.18653/v1/2024.acl-long.283.

52. Matthijs, D., Alexandr, G., Chengqi, D., Jeff, J., Gergely, S., Pierre-Emmanuel, M., Maria, L., Lucas, H., and Hervé, J. (2024). The Faiss library. Preprint at arXiv. https://doi.org/10.48550/arXiv.2401.08281.

53. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7B. Preprint at arXiv. https://doi.org/10.48550/arXiv.2310.06825.

54. Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., and Altman, S. (2023). GPT-4 Technical Report. Preprint at arXiv. https://doi.org/10.48550/arXiv.2303.08774.

55. Brown, D., and Jain, S.; nlp4whp (2021). dorianbrown/rank_bm25 0.2.1 (Zenodo). https://doi.org/10.5281/ZENODO.4520057.

56. Bachmann, M., layday, thomasryde, Kokkinou, G., Schreiner, H., Fihl-Pearson, J., dheeraj, V., pekkarr, and Górny, M. (2025). Rapidfuzz/RapidFuzz: Release 3.13.0 (Zenodo). https://doi.org/10.5281/ZENODO.15133267.

57. Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., et al. (2024). PyTorch 2: Faster machine learning through dynamic python bytecode transformation and graph compilation. In Proceedings of the 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, *2*, pp. 929–947. https://doi.org/10.1145/3620665.3640366.

58. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. (2019). HuggingFace's

transformers: State-of-the-art natural language processing. Preprint at arXiv. https://doi.org/10.48550/arXiv.1910.03771.

59. mistralai/Mistral-Nemo-Base-2407 · Hugging Face. https://huggingface.co/mistralai/Mistral-Nemo-Base-2407.

60. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234–1240. https://doi.org/10.1093/bioinformatics/btz682.

61. judithrosell/BioBERT_BioNLP13CG_NER_new · Hugging Face. https://huggingface.co/judithrosell/BioBERT_BioNLP13CG_NER_new.

62. New embedding models and API updates. https://openai.com/index/new-embedding-models-and-api-updates.

Please cite this article in press as: Jun et al., A context-augmented large language model for accurate precision oncology medicine recommendations, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.12.017

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Synthetic and real-world precision oncology queries	This study	https://github.com/hjjshine/rag-llm-cancer-paper
Molecular Oncology Almanac (MOAlmanac)	Reardon et al.[19]	https://github.com/vanallenlab/moalmanac-db/releases
Clinical Interpretation of Variants in Cancer (CIViC)	Griffith et al.[20]	https://civicdb.org/downloads/01-Oct-2025/01-Oct-2025-ClinicalEvidenceSummaries.tsv
Software and algorithms		
RAG-LLM pipeline and analysis scripts	This study	https://github.com/hjjshine/rag-llm-cancer-paper
Python (version 3.10.14)	Python Software Foundation	https://www.python.org/
faiss-cpu (version 1.8.0.post1)	Douze et al.[52]	https://github.com/facebookresearch/faiss
mistralai (version 1.9.11)	Jiang et al.[53]	https://github.com/mistralai
openai (version 2.7.1)	OpenAI et al.[54]	https://github.com/OPENAI
rank-bm25 (version 0.2.2)	Brown et al.[55]	https://github.com/dorianbrown/rank_bm25
rapidfuzz (version 3.14.3)	Bachmann et al.[56]	https://github.com/rapidfuzz/RapidFuzz
torch (version 2.2.0)	Ansel et al.[57]	https://github.com/pytorch/pytorch
transformers (version 4.42.0)	Wolf et al.[58]	https://github.com/huggingface/transformers

METHOD DETAILS

The development and evaluation of this RAG-LLM precision oncology strategy involved three steps: 1) optimizing prompt design using a standard LLM, 2) evaluating the impact of both unstructured and structured context databases on the RAG-LLM's ability to recommend biomarker-driven treatments, 3) enhancing the retrieval mechanism through hybrid lexical and semantic retrieval and 4) assessing the real-world applicability of the RAG-LLM by testing it with clinical queries from oncologists regarding biomarker-driven treatment recommendations.

Database

For the initial development of the RAG-LLM pipeline, we used the FDA-approved drug relationships from the April 2024 release of the Molecular Oncology Almanac (MOAlmanac) database (https://github.com/vanallenlab/moalmanac-db/releases/tag/v.2024-04-11). To incorporate the latest knowledge on the clinical actionability of genomic biomarkers, we used FDA-approved drug indications from the October 2025 release of the MOAlmanac database (https://github.com/vanallenlab/moalmanac-db/releases/tag/v.2025-10-03). MOAlmanac contains both unstructured, free-text precision oncology genomic knowledge and structured data fields, including biomarkers, cancer types, and therapies.[19] Since the study's primary use case is to assist medical professionals and molecular tumor boards in making treatment decisions, we focused exclusively on FDA-approved drugs to be conservative.

Prompt engineering

To optimize prompt design for optimal model performance, we conducted a preliminary prompt engineering phase using the Mistral NeMo 12B model,[59] released in July 2024. Several prompt engineering strategies were evaluated, including:

1. Basic prompt: "Please provide each line of treatment as a json format with the following JSON schema … Query: {prompt}"
2. Scope-limiting prompt: *"Please only provide the therapies that are FDA-approved for the provided genomic biomarkers … Query: {prompt}"*
3. System role prompt: *"You are a helpful chatbot specialized in suggesting FDA-approved drugs to treat cancer … Query: {prompt}"*
4. Combination prompt: Merging strategies 2 and 3 with the basic prompt.

Additionally, the prompt included an example JSON schema to help the model generate structured output, which facilitated accurate and consistent evaluation (Tables S7 and S8).

Benchmark against other LLMs

In addition to the Mistral NeMo 12B model, we evaluated the performance of several other widely used LLMs for prompt engineering (Table S9). All models were accessed via their respective APIs, except Mistral 7B Instruct, an older model nearing deprecation, which was loaded and run locally on a Google Cloud virtual machine with 4 NVIDIA T4 GPUs. We used the top-ranked GPT-4o model for all subsequent analyses.

Creation of the synthetic queries

For the development and assessment of the RAG-LLM approach, we systematically generated queries derived from each individual entity in the MOAlmanac database to mirror human-provided queries (Tables S2 and S4). A total of 234 synthetic queries were derived from the MOAlmanac April 2024 release, and a total of 651 queries were created from the MOAlmanac October 2025 release.

Construction of the context database

To evaluate the RAG-LLM approach using an unstructured context database, we generated a dataset of FDA drug label text in PDF format for each biomarker-based therapy approval and extracted each label's 'indication and usage' section. We selected drug labels for FDA-approved oncology therapies involving a biomarker for at least one approved indication, as curated by the MOAlmanac database. The resulting unstructured dataset consisted of 56 original 'Indications and Usage' sections from the April 2024 version and 106 sections from the October 2025 release, with a median length of 162 tokens from the 2024 release (Interquartile range [IQR]: 127–240 tokens) and a median length of 132 tokens from the 2025 release (IQR: 95–231 tokens; Figures S4A–S4B; Table S10). A representative input-output example using unstructured data is available in Note S1.

To evaluate the RAG-LLM approach using a structured context database, we manually created an answer set for the synthetic prompts, with a median length of 181 tokens from the 2024 release (IQR: 165–204 tokens) and a median length of 156 tokens from the 2025 release (IQR: 136–174 tokens; Figures S4C–S4D; Table S11). A total of 234 and 651 structured contexts were derived from the 2024 and 2025 versions, respectively. Each answer chunk corresponded to each entity or therapy-biomarker relationship in MOAlmanac. An example of an input prompt augmented with structured data and its output from the RAG-LLM workflow is provided in Note S2.

Entity extraction and tokenization

User queries and database contexts were processed to extract and tokenize entities of interest, specifically cancer types and biomarkers. Key entities including cancer type and genomic biomarkers were extracted using BioBERT model fine-tuned for named entity recognition (NER) tasks (judithrosell/BioBERT_BioNLP13CG_NER_new).[60,61] To reduce noise and avoid bias in lexical scoring, generic terms such as "cancer" or "tumor" were excluded. Additionally, free-text context was also tokenized to support lexical scoring.

Dense and hybrid retrieval

User queries and database contexts were embedded using the *text-embedding-3-small* model[62] and stored vector representations via the *FAISS* library.[52] For dense (semantic) retrieval, each query embedding Q was compared against every i-th database context embedding D_i using cosine similarity s_i:

$$sim(Q, D_i) = s_i$$

The top 10 or 25 most semantically similar context chunks were retrieved and used to supplement the prompt fed into the LLM.

For hybrid retrieval, the top 50 semantically similar contexts were first selected, then re-ranked based on lexical similarity using the on the Okapi BM25 method implemented in the *rank_bm25* package. All scores were min-max normalized. The final BM25 score was computed as the weighted sum of all scores derived from cancer type, biomarker, and free-text context fields. If the tokenized user query entities are $q_{\text{cancer}}, q_{\text{biomarker}}, q_{\text{all}}$, and for each i-th database context, the tokenized database entities are $d_i^{\text{cancer}}, d_i^{\text{biomarker}}, d_i^{\text{context}}$, then BM25 score for the i-th database context was:

$$\text{score\_BM25}\left(q_{\text{cancer}}, d_i^{\text{cancer}}\right) = b_i^{\text{cancer}}, \text{MinMax}\left(b_i^{\text{cancer}}\right) = b_i^{\text{cancer,norm}}$$

$$\text{score\_BM25}\left(q_{\text{biomarker}}, d_i^{\text{biomarker}}\right) = b_i^{\text{biomarker}}, \text{MinMax}\left(b_i^{\text{biomarker}}\right) = b_i^{\text{biomarker,norm}}$$

$$\text{score\_BM25}\left(q_{\text{all}}, d_i^{\text{context}}\right) = b_i^{\text{context}}, \text{MinMax}\left(b_i^{\text{context}}\right) = b_i^{\text{context,norm}}$$

The combined BM25 score for context i was:

$$score\_finalBM25_i = w_c\, b_i^{\text{cancer,norm}} + w_b\, b_i^{\text{biomarker,norm}} + w_x\, b_i^{\text{context,norm}}$$

To adaptively balance lexical and semantic relevance, we defined a similarity gap, G, which measures how much the semantic similarity differs from lexical similarity by median difference between the two distributions. Low lexical similarity may arise from missing key entities (e.g., cancer type) or poor NER performance. We introduced an adaptive coefficient alpha, which linearly decays with the similarity gap G and is clipped between 0 and 1.

$$G = \mathrm{median}(score\_finalBM25) - \mathrm{median}(s^{\mathrm{norm}})$$

If $G > G_{\mathrm{thres}}$, $\alpha = 0$; otherwise, $\alpha = \max\left(0, \min\left(1, 1 - \frac{G}{G_{\mathrm{thres}}}\right)\right)$

The final hybrid retrieval score was then computed as:

$$score\_hybrid_i = \alpha\, score\_finalBM25_i + (1 - \alpha)\, s_i^{\mathrm{norm}},$$

where s_i^{norm} denotes scaled cosine similarities to the $[0,1]$ range

The top 25 re-ranked context chunks based on contexts based on the hybrid score were provided to the LLM as contextual input alongside the user query for inference.

While our approach does not explicitly model multi-hop or logical relationships, the hybrid retrieval step partially preserves these relationships by ensuring that the most relevant entities are accurately identified and prioritized in the retrieved context space, searching for fine-grained determinants of precision oncology therapy recommendations, and self-adjusting the weighting between semantic and lexical signals.

Real-world question survey and evaluation

To collect real-world clinical questions and corresponding ground-truth drugs, we designed a survey and distributed it to collaborating physicians at the Dana-Farber Cancer Institute, Boston Children's Hospital, Brigham and Women's Hospital, and Beth Israel Deaconess Medical Center. The survey introduced the MOAlmanac database, focusing on FDA-approved drug indications and clinical genomic biomarkers. Physicians were asked to submit questions related to their clinical practice regarding drug actionability, treatment regimens, or biomarker associations, without including identifiable patient data. These responses were then used as prompts to evaluate our RAG-LLM approach, assessing its ability to generate relevant and accurate answers in real-world clinical settings. Only the FDA-approved drugs were included in the final ground-truth drug set for evaluation. The prompt design was refined to handle cases where no FDA-approved therapies exist. To account for variability across model runs, we conducted five iterations of the RAG-LLM workflow.

Exploratory analysis of external database CIViC integration

While the primary focus of this study is on FDA-approved therapies, we performed an exploratory evaluation using a broader underlying database that contains drugs beyond FDA-approved therapies. To do this, we used the CIViC database, a community-driven, expert moderated database of oncology related genomic variants that includes investigational therapies.[20] For this, we downloaded the clinical evidence summaries from the October 1st, 2025, data release (available at https://civicdb.org/downloads/01-Oct-2025/01-Oct-2025-ClinicalEvidenceSummaries.tsv). To ensure interoperability with the RAG-LLM pipeline, we processed and extracted fields including cancer type, biomarker, and evidence statement, mirroring the structure of the MOAlmanac-based context database.

To accommodate the broader scope of CIViC, a more flexible prompt strategy was used (see Note S3) and tested on the real-world test queries and further expanded the ground-truth therapies to include expert-curated ground-truth drugs that are not FDA-approved (i.e., off-label drugs) or biomarker-associated (i.e., no biomarkers are provided in the FDA indication). Additionally, to compare results with the pipeline's performance on detecting FDA-approved therapies, we tested the pipeline augmented with the CIViC database under prompt strategy R2 ("scope-aware" strategy). This analysis was added in response to our reviewer feedback regarding the study's limited scope. Its purpose is to provide an illustrative example of pipeline generalizability across different databases and to lay the groundwork for future study on broader precision oncology queries.

Evaluation metrics

We evaluated LLM performance with and without RAG by calculating the proportion of exact and partial matches of correctly predicted therapy recommendations across all the queries. Each LLM output was generated in a JSON format and was parsed line by line for drug names following the 'Drug Name' entity to compute the accuracy. To account for the non-deterministic behavior of LLMs, we ran five iterations of the workflow and averaged the accuracy across iterations. For exact match accuracy, predictions were considered correct if all the ground-truth therapies were present in the drug output. For partial match accuracy, predictions were considered correct if at least one ground truth therapy was present:

$$Exact\ Match\ Accuracy = \frac{N_{exact\ match}}{N_{total}} \quad Partial\ Match\ Accuracy = \frac{N_{partial\ match}}{N_{total}}$$

Where:

$$N_{exact\ match} = cases\ with\ predicted\ drugs\ exactly\ matching\ all\ the\ ground\ truth\ drugs$$

$$N_{partial\ match} = cases\ with\ predicted\ drugs\ exactly\ matching\ one\ or\ more\ ground\ truth\ drugs$$

$$N_{total} = Total\ number\ of\ queries$$

Please cite this article in press as: Jun et al., A context-augmented large language model for accurate precision oncology medicine recommendations, Cancer Cell (2025), https://doi.org/10.1016/j.ccell.2025.12.017

Additionally, we calculated precision, recall, F1 score, and specificity to comprehensively assess model performance across all queries. To minimize the randomness in outputs, we set the temperature to 0.0 and initialized a fixed random seed.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis

We performed McNemar's test to compare accuracies between different approaches. When the total number of discordant pairs was 25 or greater, we used the chi-square approximation with a continuity correction to prevent overestimation of significance. For fewer than 25 discordant pairs, we applied the exact binomial test. To assess overall performance across iterations, we pooled the discordant pair counts from all runs and then performed the McNemar's test. When comparing evaluation metrics other than accuracy, we used a one-sided Wilcoxon signed-rank test. To account for multiple hypothesis testing, *p*-values were adjusted using the Benjamini-Hochberg correction. In addition to statistical significance, we quantified effect sizes using a standardized measure of discordant pairs from McNemar's test, calculated as the proportion of cases favoring the better-performing model (e.g., $\frac{|b-c|}{b+c}$), which is analogous to Cohen's g.