Implementing a context-augmented large language model to guide precision cancer medicine

Authors: Hyeji Jun, Yutaro Tanaka, Shreya Johri, Filipe LF Carvalho, Alexander C. Jordan, Chris Labaki, Matthew Nagy, Tess A. O'Meara, Theodora Pappa, Erica Maria Pimenta, Eddy Saad, David D Yang, Riaz Gillani, Alok K. Tewari, Brendan Reardon, Eliezer Van Allen

Description

This study evaluates a context-augmented large language model (LLM) workflow for recommending FDA-approved, biomarker-driven cancer therapies. By optimizing prompt design and integrating a structured precision oncology knowledge base, the approach demonstrates strong performance across both synthetic queries and real-world oncologist scenarios.

Abstract

BACKGROUND

The rapid expansion of molecularly informed therapies in oncology, coupled with evolving regulatory FDA approvals, poses a challenge for oncologists seeking to integrate precision cancer medicine into patient care. Large Language Models (LLMs) have demonstrated potential for clinical applications, but their reliance on general knowledge limits their ability to provide up-to-date and niche treatment recommendations. To address this challenge, we developed a Retrieval-Augmented Generation (RAG)-LLM approach that integrates with a precision oncology knowledge resource, and evaluated whether this approach improved accuracy in biomarker-driven treatment recommendations relative to alternative frameworks.

METHODS

We developed a RAG-LLM workflow that integrates Molecular Oncology Almanac (MOAlmanac) and evaluated this approach relative to alternative frameworks (i.e. LLM-only) in making biomarker-driven treatment recommendations using both unstructured and structured data. We evaluated LLM performance by calculating exact and partial match accuracies across 234 therapy-biomarker relationships. Finally, we assessed real-world applicability of the workflow by testing it on actual queries from practicing oncologists.

RESULTS

While LLM-only achieved 54–69% accuracy in biomarker-driven treatment recommendations, RAG-LLM achieved 73–85% accuracy with an unstructured database and 91–99% accuracy with a structured database. In addition to accuracy, structured context augmentation substantially increased precision (49.0% to 79.9%) and F1-score (55.7% to 84.4%) compared to unstructured data augmentation. In queries recommended by practicing oncologists, RAG-LLM achieved 75–94% accuracy and high precision (91.67%).

CONCLUSIONS
These findings demonstrate the effectiveness of the RAG-LLM framework in recommending FDA-approved precision oncology therapies based on individualized clinical information, and highlight the importance of integrating a well-curated, structured knowledge base in this process. While our RAG-LLM approach significantly improved accuracy compared to standard LLMs, further efforts will enhance the generation of reliable responses for ambiguous or unsupported clinical scenarios.

Introduction

Identifying therapeutically targetable molecular alterations to guide treatment options is a key component of precision cancer medicine. 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]. 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, the advancements in Large Language Models (LLMs) for addressing similar challenges across clinical medicine have gained attention. Studies have demonstrated the capability of LLMs in 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 application to support clinical decision-making in precision oncology[8–10].

Despite these advancements, LLMs have limitations when handling niche and constantly evolving queries, particularly in the field of oncology[11]. These challenges are due to insufficient domain-specific training and reliance on potentially outdated data, which raises concerns about the accuracy and relevance of their output[12]. This issue is especially critical given these strategies guide oncologists in clinical decision-making. Furthermore, the overwhelming variety of ever-growing available LLMs complicates their adoption, as benchmarking them for specific use cases can be both time-consuming and computationally demanding.

Efforts to address these limitations are an active area of research[13]. Notably, Retrieval-Augmented Generation (RAG) has emerged as a promising approach, dynamically retrieving relevant information from external domain-specific databases to supplement an LLM's general knowledge without modifying its internal weights[14–16]. In precision oncology, expertly curated databases such as Molecular Oncology Almanac (MOAlmanac), OncoKB, CIVIC, and MyCancerGenome, are used to guide oncologists treatment options that may be available for diverse genomic and disease indications[17–20]. Rather than envisioning LLMs replacing these resources, we hypothesized that LLMs augmented with expertly curated databases could further enhance clinical decision support, particularly for practicing medical professionals in underserved settings[21,22].

In this study, we introduce a RAG-LLM approach that enables accurate and approval-derived therapy recommendations based on patients' genomic biomarkers, disease type, treatment history, and other clinically relevant information for treatment planning (Figure 1). 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[17], 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 additionally 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.

Methods

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, and 3) assessing the real-world applicability of the RAG-LLM by testing it with clinical queries from oncologists regarding biomarker-driven treatment recommendations.

Database: To incorporate the latest knowledge on the clinical actionability of genomic biomarkers, we used FDA-approved drug indications from the April 11th, 2024 release of the Molecular Oncology Almanac (MOAlmanac) database (https://github.com/vanallenlab/moalmanac-db/releases/tag/v.2024-04-11). MOAlmanac contains both unstructured, free-text precision oncology genomic knowledge and structured data fields, including biomarkers, cancer types, and therapies[17]. 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[23], 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 (Table 1).

Table 1. Strategies tested for prompt design optimization

Strategy	Prompt

(1)
Basic prompt
	Please provide each treatment as a json format with the following JSON schema.
	{{
	"Treatment 1": {{
	"Disease Name": ,
	"Disease Phase or Condition": ,
	"Drug Name": ,
	"Prior Treatment or Resistance Status": ,
	"Genomic Features":
	"Link to FDA-approved Label":
	}}
	}}
	Query: {prompt}

(2)
Scope-limiting prompt
	Please only provide the therapies that are FDA-approved for the provided genomic biomarkers. Please provide each treatment as a json format with the following JSON schema.
	{{
	"Treatment 1": {{
	"Disease Name": ,
	"Disease Phase or Condition": ,
	"Drug Name": ,
	"Prior Treatment or Resistance Status": ,
	"Genomic Features":
	"Link to FDA-approved Label":
	}}
	}}
	Query: {prompt}

(3)
System role prompt
	You are a helpful chat bot specialized in suggesting FDA-approved drugs to treat cancer. Please provide each treatment as a json format with the following JSON schema.
	{{
	"Treatment 1": {{
	"Disease Name": ,
	"Disease Phase or Condition": ,
	"Drug Name": ,
	"Prior Treatment or Resistance Status": ,
	"Genomic Features":
	"Link to FDA-approved Label":
	}}
	}}
	Query: {prompt}

(4) Combination	You are a helpful chat bot specialized in suggesting FDA-approved drugs to treat cancer. Please only provide the therapies that are FDA-approved for the

prompt	provided genomic biomarkers. Please provide each treatment as a json format with the following JSON schema.
	{{
	"Treatment 1": {{
	"Disease Name": ,
	"Disease Phase or Condition": ,
	"Drug Name": ,
	"Prior Treatment or Resistance Status": ,
	"Genomic Features":
	"Link to FDA-approved Label":
	}}
	}}
	Query: {prompt}

Benchmark against other LLMs: In addition to the Mistral NeMo 12B model, we evaluated the performance of several other widely used LLMs. 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.

- Mistral 7B Instruct (mistral-7B-Instruct-v0.3)
- Mistral 8B (ministral-8b-2410)
- Mistral 123B (mistral-large-2407)
- OpenAI GPT-4o (gpt-4o-2024-05-13)
- OpenAI GPT-4o-mini (gpt-4o-mini-2024-07-18)

RAG-LLM approach: For the development and assessment of the RAG-LLM approach, we used 234 systematically generated prompts derived from each individual entity in the MOAlmanac database to mirror human-provided queries (Table S1–2).

To evaluate the RAG-LLM approach using an unstructured dataset, 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, which served as the context database for RAG integration (Table S3). A representative input-output example using unstructured data is available in Supplementary Note 1.

To evaluate the RAG-LLM approach using a structured dataset, we manually created an answer set for the synthetic prompts, incorporating it into the RAG-LLM workflow as a structured context database. Each answer chunk corresponded to each entity or therapy-biomarker relationship in MOAlamanc (Table S4). An example of an input prompt augmented with structured data and its output from the RAG-LLM workflow is provided in Supplementary Note 2.

For both approaches, the context database was embedded using the *mistral-embed* model (https://docs.mistral.ai/capabilities/embeddings/) and stored as a vector database via the *FAISS* library[24]. Each query embedding was then compared to the embedded context database using Euclidean distance, retrieving the top 10 most similar text chunks to supplement the prompt fed into the LLM for inference.

Real-world question survey and evaluation: To collect real-world clinical questions, we designed a survey and distributed it to collaborating physicians at the Dana-Farber Cancer Institute and Boston Children's Hospital. 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. 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. To account for potential variations, we conducted five iterations of the RAG-LLM workflow.

Evaluation: 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 relationships. 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. For exact match accuracy, predictions were considered correct if all the ground-truth therapies (Table S2) were present in the drug output:

$$Exact\ match\ accuracy = \frac{\#\ cases\ with\ predicted\ therapies\ exactly\ matching\ ALL\ the\ ground\ truth\ therapies}{Total\ number\ of\ relationships} \times 100$$

For partial match accuracy, predictions were considered correct if at least one ground truth therapy was present:

$$Partial\ match\ accuracy = \frac{\#\ cases\ with\ predicted\ therapies\ exactly\ matching\ one\ or\ more\ ground\ truth\ therapies}{Total\ number\ of\ relationships} \times 100$$

where the total number of relationships is 234.

Additionally, we calculated precision, recall, F1 score, and specificity to comprehensively assess model performance across all queries. For all evaluations, we used the general-purpose state-of-the-art Mistral NeMo model, known for its efficiency and small size[23]. To minimize the randomness in outputs, we set the temperature to 0.0 and initialized a fixed random seed.

Statistical test: We performed McNemar's test to compare accuracies between different approaches. When the number of discordant pairs was 25 or more, 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 account for multiple testing, Benjamini-Hochberg correction was performed to adjust the p-values obtained from multiple workflow iterations.

Results

Prompt engineering: Given the critical role of prompt optimization in enhancing LLM performance, we first evaluated whether a specific prompt design optimized 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: (1) a single instruction (referred to as the "basic prompt"), (2) a basic prompt with a condition statement restricting results to FDA-approved drugs only, (3) a basic prompt with a defined system role, and (4) a combination of the second and third additional statements (Table 1). All evaluated queries were formulated using the structured data from MOAlmanac (Table S1–2). To compare prompt effectiveness, we measured partial match accuracy, defined as the proportion of retrieved therapies that matched ground-truth FDA-approved therapies (Methods). Among the four strategies, the basic prompt demonstrated the highest accuracy in retrieving FDA-approved therapies (Figure 2A). The basic prompt achieved an accuracy of 69.23%, outperforming the second (53.42%), third (64.53%), and fourth (51.28%) strategies. The superior performance of the basic prompting strategy was also observed in other LLMs, particularly those of larger sizes (Figure 2B). Based on these results, we selected the basic prompt for all subsequent evaluations in the RAG-LLM workflow, where relevant information from an external database was integrated into the prompt alongside the user-provided query.

RAG-LLM's performance based on unstructured and structured data: 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 Mistral NeMo 12B's performance using both unstructured and structured data formats. The unstructured dataset was constructed by extracting the 'Indication and Usage' section from drug label text, with a median length of 735 tokens (Interquartile range [IQR]: 552–977; Figure S1A, Table S3). In contrast, the structured dataset consisted of manually curated therapy-biomarker relationships, with a median length of 734 tokens (IQR: 649–827; Figure S1B, Table S4). Both approaches were tested on 234 synthetic prompts derived from single entities in MOAlmanac, and accuracy was evaluated against ground-truth therapies (Methods).

Without RAG-provided context, model accuracy ranged from 54–69%. However, augmenting the model with unstructured text data significantly improved performance, increasing accuracy to 73–85% (exact match: $\chi^2(1) = 23.90$, $p = 1.01 \times 10^{-6}$, McNemar's test; partial match: $\chi^2(1) = 18.25$, $p = 1.93 \times 10^{-5}$, McNemar's test) (Figure 3A). Integrating structured data further improved performance compared to the approach with unstructured data augmentation, yielding an accuracy of 91–99% (exact match: $\chi^2(1) = 25.40$, $p = 4.67 \times 10^{-7}$, McNemar's test; partial match: $\chi^2(1) = 28.26$, $p = 1.06 \times 10^{-7}$, McNemar's test) (Figure 3A). In addition to accuracy, structured data augmentation markedly improved other key performance metrics (Figure 3B). Specifically, precision, recall, F1-score, and specificity all increased, with precision and F1-score improved by approximately 50–60% compared to the unstructured data augmented model (precision: 49.0% to 79.9%; F1-score: 55.7% to 84.4%). These observations demonstrate the role of structured data in enhancing the model's ability to provide more precise and reliable FDA-approved therapy predictions with minimal errors.

Additionally, structured data augmentation consistently outperformed the unstructured approach across various cancer types and therapy categories (Figure 3C–D). For example, an increase in accuracy of at least 30% was observed when transitioning from unstructured to structured augmentation for melanoma (MEL), chronic myelogenous leukemia (CML), peritoneal serous carcinoma (PSEC), and non-small cell lung cancer (NSCLC). Similarly, among therapy strategies, prompts related to RAF inhibition, antiandrogen + PARP inhibition, RAF + MEK inhibition, RET inhibition, BCR-ABL inhibition, and MET inhibition also showed at least a 30% improvement in accuracy with structured augmentation. Together, these results suggest that structured data helped the model better capture relationships between therapies and their approved indications, making it more robust across different contexts.

However, the model's performance was suboptimal for certain complex therapy strategies, particularly antiandrogen + PARP inhibition and ER signaling inhibition. Most antiandrogen + PARP inhibition therapies are approved for Homologous Recombination Repair (HRR)-gene-mutated metastatic castration-resistant prostate cancer (mCRPC), which involves 16 HRR genes linked to drug approvals. The additional reasoning required to infer the connection from "HRR genes" to specific gene mutations, which are not explicitly stated within the 'Indications and Usage' section of FDA drug package inserts, challenged the retrieval algorithm's ability to retrieve all relevant entries from the unstructured dataset within its top 10 results.

RAG-LLM on real-world questions: To assess the real-world applicability of the RAG-LLM approach, we collected 21 clinical queries from 12 oncologists affiliated with Dana-Farber Cancer Institute. These queries focused on precision oncology therapies given a specific cancer type and biomarker(s). Of these, 10 queries had answers linked to on-label FDA-approved therapy indications listed in the MOAlmanac database (Table S5). We used the structured dataset as the context database for RAG and evaluated the model's performance using exact match and partial match accuracies (Methods).

For queries where no on-label FDA-approved therapies currently exist, we evaluated whether RAG-LLM incorrectly retrieved off-label treatment options. We found that RAG-LLM exhibited lower performance on clinical queries (n=21) compared to synthetic queries (n=234), particularly exhibiting high rates of misattributions in cases without any appropriate FDA-approved drugs available (mean exact match accuracy: 27.62%; mean partial match accuracy: 46.67%; Figure 3A, 4A–B). For example, the LLM misattributed crizotinib as an on-label treatment option when queried for on-label regimens for a patient with rhabdomyosarcoma and a *TFCP2* fusion (Supplementary Note 3).

However, when explicitly instructed to return no results when no approved therapy was available (Methods), the model correctly returned no drugs in most cases without FDA-approved drugs available (Figure 4C), achieving a mean exact match accuracy of 75.24% and a mean partial match accuracy of 94.29%, significantly outperforming the version without the out-of-scope instruction (adjusted p-values: all $p < 0.05$, McNemar's test; Figure 4A). The full output and predicted therapies from the workflow with the out-of-scope instruction are provided in Table S6. These results demonstrate the challenges in extrapolation of carefully designed queries to real-world queries, and potential strategies to counter common issues. They also demonstrate the effectiveness of RAG-LLM in retrieving relevant therapies for actual oncologist

queries and highlight the importance of accurately handling cases with no valid ground-truth answers.

For the 10 questions with FDA-approved therapies within the MOAlamanc database, the RAG-LLM model predicted at least one correct therapy for each case. Notably, in six cases, the predicted therapies perfectly matched all the ground-truth therapies. The model demonstrated high precision (91.67%), indicating strong reliability when making therapy predictions. However, recall (84.97%) was slightly lower, suggesting that the model occasionally failed to retrieve all relevant therapies and subsequently missed ground-truth therapies. Together, these findings underscore the RAG-LLM approach's effectiveness in retrieving clinically relevant therapies, while also suggesting the need for additional strategies to address recall limitations.

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, while also showing recall limitations. 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[26]. We observed better performance with a basic prompt over a combined prompt, suggesting that enforcing JSON-format output may benefit from brevity. However, performance varied across models, especially smaller ones with fewer parameters, indicating no universal prompt design guarantees improvement across all LLMs in this use case.

RAG implementation significantly outperformed the LLM-only approach in predicting FDA-approved therapies, though performance varied with the format of the external dataset used. Our results demonstrate the model's ability to extract relevant information from unstructured free text, while underscoring that highly curated structured data can significantly enhance its reliability, making the RAG-LLM built with structured data better suited for integration into clinical decision-making workflows.

We demonstrated the effectiveness of the RAG-LLM approach using a structured dataset in the real-world scenarios we evaluated, but additionally identified a number of challenges this strategy faces. When no FDA-approved therapy was available for a given case, the model often misattributed treatments approved for other cancer types or biomarkers—although such treatments may still be used in practice based on their clinical relevance, this is beyond the scope of the current study. To address this issue, we implemented a predefined JSON schema to instruct the model to return "no matches" when appropriate. This enhancement substantially improved accuracy. For queries with on-label FDA-approved drugs, the model achieved high precision and slightly lower recall, indicating the model is less likely to predict false positives, which is critical in treatment recommendation tasks.

Notably, further improvements are needed for the deployment of the RAG-LLM framework at the point of care. One key challenge lies in the retrieval mechanism. RAG selects

information based on text similarity, which can result in irrelevant therapies when no perfect match exists. This process can mislead the model and hinder its ability to differentiate between cancer types and biomarkers. Additionally, in some instances, the model retrieved drugs that were not present in the context retrieved via RAG, indicating the model's occasional reliance on pretrained knowledge over the retrieved context provided with the query.

Another challenge stems from the nature of general-purpose LLMs, which are optimized to generate responses and often prioritize providing an answer over acknowledging uncertainty[27]. We found that this issue is amplified when the model is directed to generate structured JSON outputs; for example, if the schema requires a link to a drug label, the model may fabricate one to meet schema requirements. Consequently, enforcing a rigid schema may inadvertently increase the risk of hallucination. While we have demonstrated that this risk can be alleviated by explicitly defining out-of-scope cases (e.g. instructing the model to return no results when no match is found), the model remains prone to fabricating responses, pulling from either the RAG-provided context or its pretrained general knowledge.

Ultimately, deploying LLM-based tools in clinical workflows demands strategies to mitigate the inherent risks arising from their widespread adoption. While the primary purpose of our RAG-LLM framework is to guide oncologists in recommending biomarker-driven therapies, its deployment will inevitably increase the risk of misuse or vulnerability to adversarial prompts and attacks—risks shared by all general-purpose LLMs[28]. Furthermore, integrating this framework into clinical settings will present unique challenges, including patient data privacy, potential harm from errors and biases, and protection of intellectual property and proprietary data[29]. Ensuring regulatory compliance with data privacy laws is also critical to maintaining patient confidentiality[30]. Thus, successful implementation of this framework is dependent on a comprehensive strategy that not only tackles the technical complexities of LLM deployment but also prospectively addresses the ethical and regulatory challenges in the high-stakes realm of oncology patient care.

In total, our RAG-LLM framework designed to retrieve biomarker-based FDA-approved drugs significantly outperformed an LLM-only implementation. Structured data augmentation further boosted its performance, and it demonstrated strong reliability on real-world oncologist questions. Given that it requires fewer computational resources than finetuning-based approaches, this framework may also promote greater equity in precision cancer medicine, particularly by supporting non-academic oncologists with limited resources. However, achieving the necessary accuracy and reliability for real-world deployment of this approach requires further refinements. Key areas for improvement include enhancing retrieval mechanisms to prioritize relevant therapies over purely text-matched results, expanding the context database by integrating clinical guidelines and clinical trial data to provide treatment options often available to physicians beyond FDA-approved drugs, and lastly, ensuring that treatment recommendations remain current by incorporating a dynamically updated knowledge base. Addressing these enhancements will be essential for developing a clinically reliable, real-world applicable, and LLM-driven decision-support tool. As LLMs continue to rapidly evolve, ensuring that they are grounded in structured, high-quality knowledge is imperative to their successful implementation in precision oncology.

Figure Legend

Figure 1. Schematic representation of the RAG-LLM precision oncology workflow for biomarker-driven treatment recommendations.
A. Overview of the RAG-LLM precision oncology workflow. The RAG framework retrieves the top 10 most similar matches from the context database using Euclidean distance between query and context embeddings. To evaluate the impact of dataset structure on biomarker-driven treatment recommendations, we compared the performance of RAG-LLM with unstructured data augmentation versus that of RAG-LLM using structured data. The resulting context-augmented query was then fed into a general-purpose LLM for inference. **B.** To test real-world applicability of the RAG-LLM workflow, actual clinical queries collected from practicing oncologists were input into the RAG-LLM workflow.

Figure 2. Performance of Mistral Nemo 12B and other models 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.

Figure 3. Enhancement through RAG using unstructured and structured datasets.
A. Exact match and partial match accuracies from RAG-LLM augmented with unstructured and structured datasets. **B.** Precision, recall, F1-score, specificity and accuracies from RAG-LLM with unstructured and structured data augmentation. **C.** Top: Number of entities in the MOAlmanac database 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 across therapy strategies. Bottom: Exact match accuracies from unstructured and structured RAG-LLM approaches across therapy strategies. *Abbreviations in **C** and **D** are defined in Table S7.*

Figure 4. Performance of structured context-augmented RAG-LLM on real-world questions from practicing oncologists.
A. Exact match and partial match accuracies from structured data-based RAG-LLM with or without explicit out-of-scope instructions. **B.** Number of matching, missing, hallucinated, and no-drug cases predicted by the workflow without explicit out-of-scope instructions, based on first-iteration predictions. Hallucinated drugs include all off-label drugs that were incorrectly attributed as FDA-approved. **C.** Same as **B**, but with explicit out-of-scope instructions (i.e., explicitly directing the model to return no matches if no FDA-approved drugs are available).

References

1. Suehnholz SP, Nissan MH, Zhang H, et al. Quantifying the expanding landscape of clinical actionability for patients with cancer. Cancer Discov [Internet] 2024;14(1):49–65. Available from: http://dx.doi.org/10.1158/2159-8290.CD-23-0467

2. Chow-White P, Ha D, Laskin J. Knowledge, attitudes, and values among physicians working with clinical genomics: a survey of medical oncologists. Hum Resour Health [Internet] 2017;15(1):42. Available from: http://dx.doi.org/10.1186/s12960-017-0218-z

3. Jin Q, Wang Z, Floudas CS, et al. Matching patients to clinical trials with large language models. Nat Commun [Internet] 2024 [cited 2025 Apr 24];15(1):9074. Available from: https://www.nature.com/articles/s41467-024-53081-z

4. Cerami E, Trukhanov P, Paul MA, et al. MatchMiner-AI: An Open-Source Solution for Cancer Clinical Trial Matching [Internet]. arXiv [cs.AI]. 2024;Available from: http://arxiv.org/abs/2412.17228

5. Wong C, Zhang S, Gu Y, et al. Scaling clinical trial matching using large language models: A case study in oncology [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2308.02180

6. Van Veen D, Van Uden C, Blankemeier L, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med [Internet] 2024 [cited 2025 Apr 24];30(4):1134–42. Available from: https://www.nature.com/articles/s41591-024-02855-5

7. Katz U, Cohen E, Shachar E, et al. GPT versus resident physicians — A benchmark based on official board scores. NEJM AI [Internet] 2024;1(5). Available from: http://dx.doi.org/10.1056/aidbp2300192

8. Benary M, Wang XD, Schmidt M, et al. Leveraging large language models for decision support in personalized oncology. JAMA Netw Open [Internet] 2023;6(11):e2343689. Available from: http://dx.doi.org/10.1001/jamanetworkopen.2023.43689

9. Xu S, Most A, Chase A, et al. Large language models management of complex medication regimens: a case-based evaluation [Internet]. medRxiv. 2024;2024.07.03.24309889. Available from: http://medrxiv.org/content/early/2024/07/08/2024.07.03.24309889.abstract

10. Elemento O, Khozin S, Sternberg CN. The use of artificial intelligence for cancer therapeutic decision-making. NEJM AI [Internet] 2025;2(5). Available from: http://dx.doi.org/10.1056/aira2401164

11. Verlingue L, Boyer C, Olgiati L, Brutti Mairesse C, Morel D, Blay J-Y. Artificial intelligence in oncology: ensuring safe and effective integration of language models in clinical practice. Lancet Reg Health Eur [Internet] 2024;46:101064. Available from: http://dx.doi.org/10.1016/j.lanepe.2024.101064

12. Jeffrey C, Marc M, Orion W, Dawn L, Daniel K, Van Durme B. Dated data: Tracing knowledge cutoffs in Large Language Models [Internet]. arXiv [cs.CL]. 2024;Available from:

http://arxiv.org/abs/2403.12958

13. Towhidul Islam Tonmoy S, Mehedi Zaman SM, Vinija J, et al. A comprehensive survey of hallucination mitigation techniques in Large Language Models [Internet]. arXiv [cs.CL]. 2024;Available from: http://arxiv.org/abs/2401.01313

14. Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Neural Inf Process Syst [Internet] 2020 [cited 2025 Apr 24];abs/2005.11401. Available from: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

15. Shanghua G, Richard Z, Zhenglun K, et al. TxAgent: An AI agent for therapeutic reasoning across a universe of tools [Internet]. arXiv [cs.AI]. 2025;Available from: http://arxiv.org/abs/2503.10970

16. Zakka C, Shad R, Chaurasia A, et al. Almanac - retrieval-augmented language models for clinical medicine. NEJM AI [Internet] 2024;1(2). Available from: http://dx.doi.org/10.1056/aioa2300068

17. Reardon B, Moore ND, Moore NS, et al. Integrating molecular profiles into clinical frameworks through the Molecular Oncology Almanac to prospectively guide precision oncology. Nat Cancer [Internet] 2021 [cited 2025 Apr 24];2(10):1102–12. Available from: https://www.nature.com/articles/s43018-021-00243-3

18. Chakravarty D, Gao J, Phillips SM, et al. OncoKB: A precision oncology knowledge base. JCO Precis Oncol [Internet] 2017;2017. Available from: http://dx.doi.org/10.1200/PO.17.00011

19. Griffith M, Spies NC, Krysiak K, et al. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. Nat Genet [Internet] 2017 [cited 2025 Apr 24];49(2):170–4. Available from: https://www.nature.com/articles/ng.3774

20. Holt ME, Mittendorf KF, LeNoue-Newton M, et al. My Cancer Genome: Coevolution of precision oncology and a molecular oncology knowledgebase. JCO Clin Cancer Inform [Internet] 2021;5(5):995–1004. Available from: http://dx.doi.org/10.1200/CCI.21.00084

21. Malone ER, Oliva M, Sabatini PJB, Stockley TL, Siu LL. Molecular profiling for precision cancer therapies. Genome Med [Internet] 2020;12(1):8. Available from: http://dx.doi.org/10.1186/s13073-019-0703-1

22. Rahman B, McEwen A, Phillips JL, Tucker K, Goldstein D, Jacobs C. Genetic and genomic learning needs of oncologists and oncology nurses in the era of precision medicine: a scoping review. Per Med [Internet] 2022;19(2):139–53. Available from: http://dx.doi.org/10.2217/pme-2021-0096

23. mistralai/Mistral-Nemo-Base-2407 · Hugging Face [Internet]. [cited 2025 Apr 24];Available from: https://huggingface.co/mistralai/Mistral-Nemo-Base-2407

24. Matthijs D, Alexandr G, Chengqi D, et al. The Faiss library [Internet]. arXiv [cs.LG]. 2024;Available from: http://arxiv.org/abs/2401.08281

25. Reynolds L, McDonell K. Prompt programming for large language models: Beyond the few-shot paradigm [Internet]. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: ACM; 2021. Available from: http://dx.doi.org/10.1145/3411763.3451760

26. Pranab S, Singh AK, Sriparna S, Vinija J, Samrat M, Aman C. A systematic survey of prompt engineering in large language models: Techniques and applications [Internet]. arXiv [cs.AI]. 2024;Available from: http://arxiv.org/abs/2402.07927

27. Miao X, Zhiyuan H, Xinyang L, et al. Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2306.13063

28. Erfan S, Al Mamun MA, Yu F, Pedram Z, Yue D, Nael A-G. Survey of vulnerabilities in large Language Models revealed by adversarial attacks [Internet]. arXiv [cs.CL]. 2023;Available from: http://arxiv.org/abs/2310.10844

29. Umeton R, Kwok A, Maurya R, et al. GPT-4 in a cancer center — institute-wide deployment challenges and lessons learned. NEJM AI [Internet] 2024;1(4). Available from: http://dx.doi.org/10.1056/aics2300191

30. Ong JCL, Chang SY-H, William W, et al. Ethical and regulatory challenges of large language models in medicine. Lancet Digit Health [Internet] 2024;6(6):e428–32. Available from: http://dx.doi.org/10.1016/S2589-7500(24)00061-X

Figure 1. Schematic representation of the RAG-LLM precision oncology workflow for biomarker-driven treatment recommendations

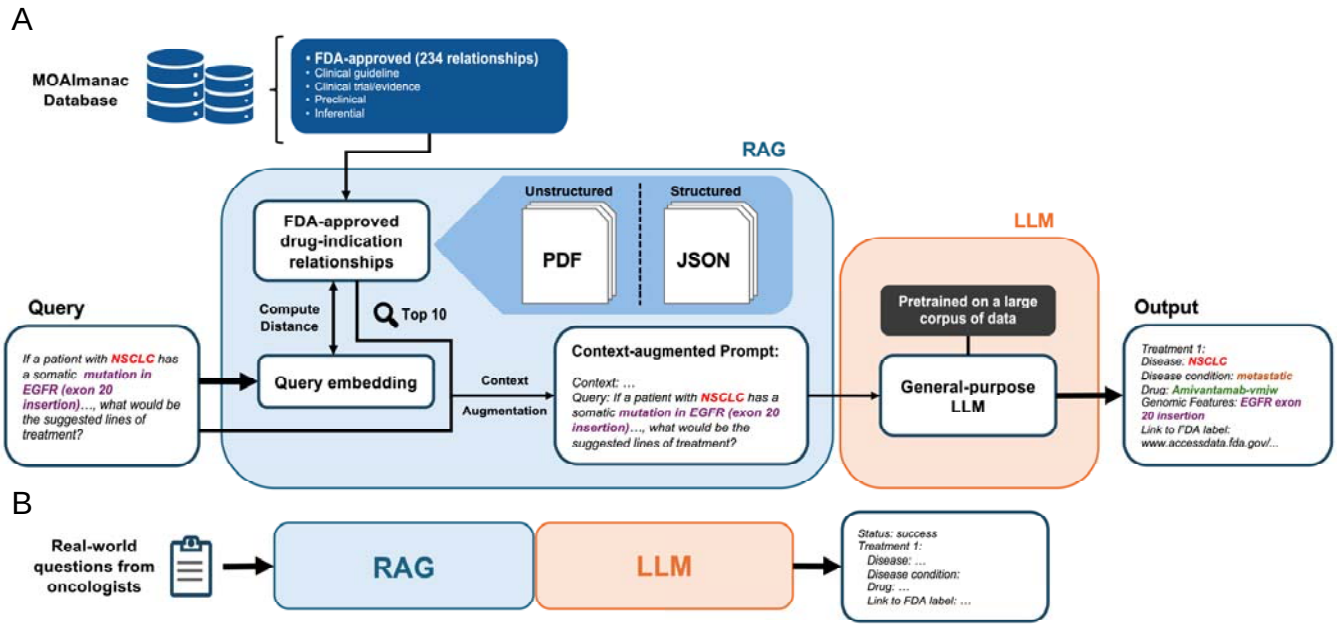

Figure 2. Performance of Mistral Nemo 12B and other types of models across various prompt engineering strategies

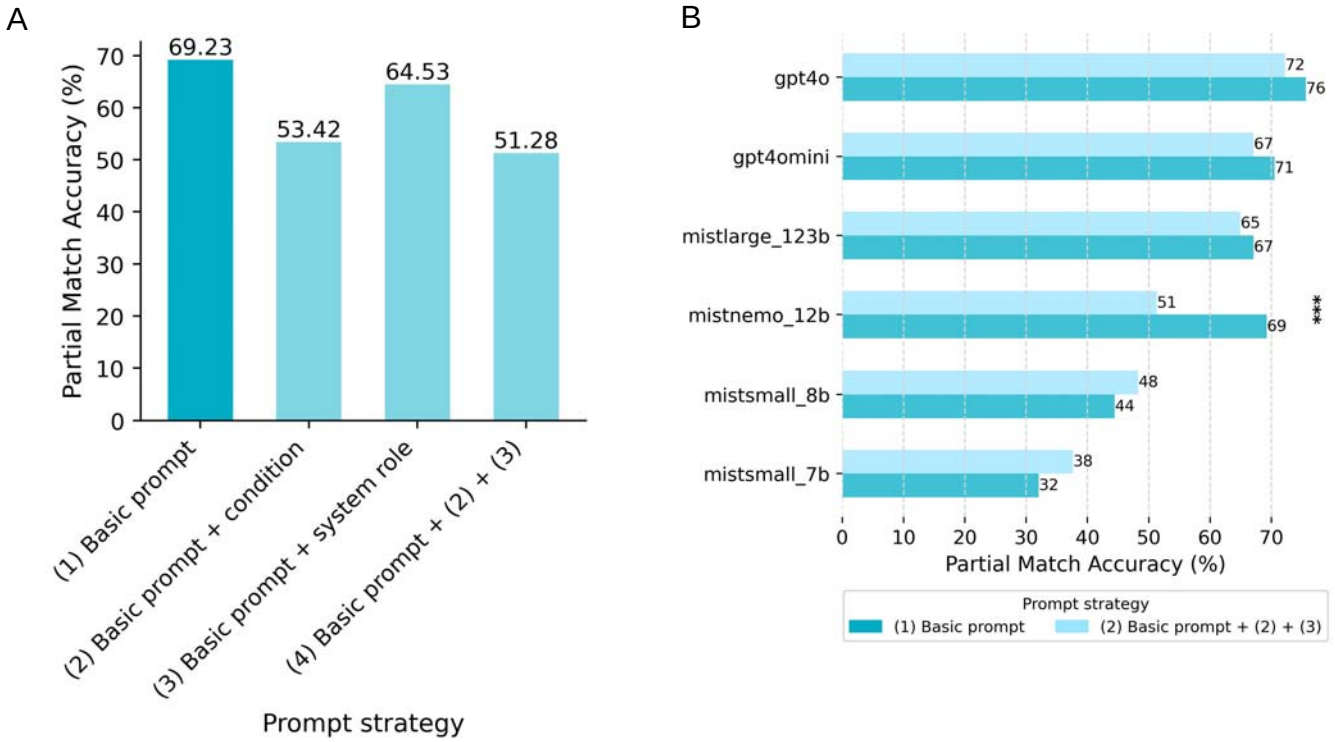

Figure 3. Enhancement through RAG using unstructured and structured datasets

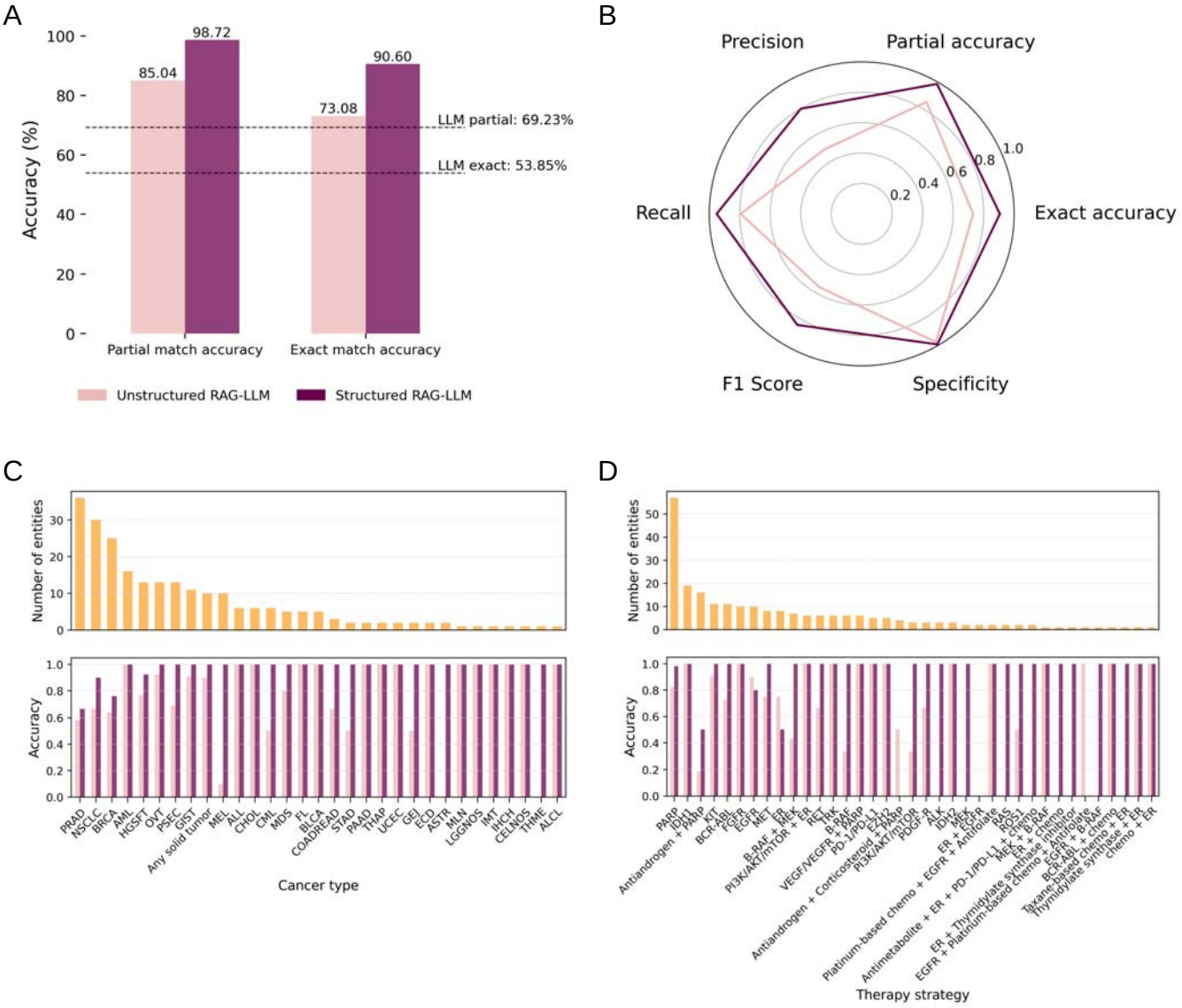
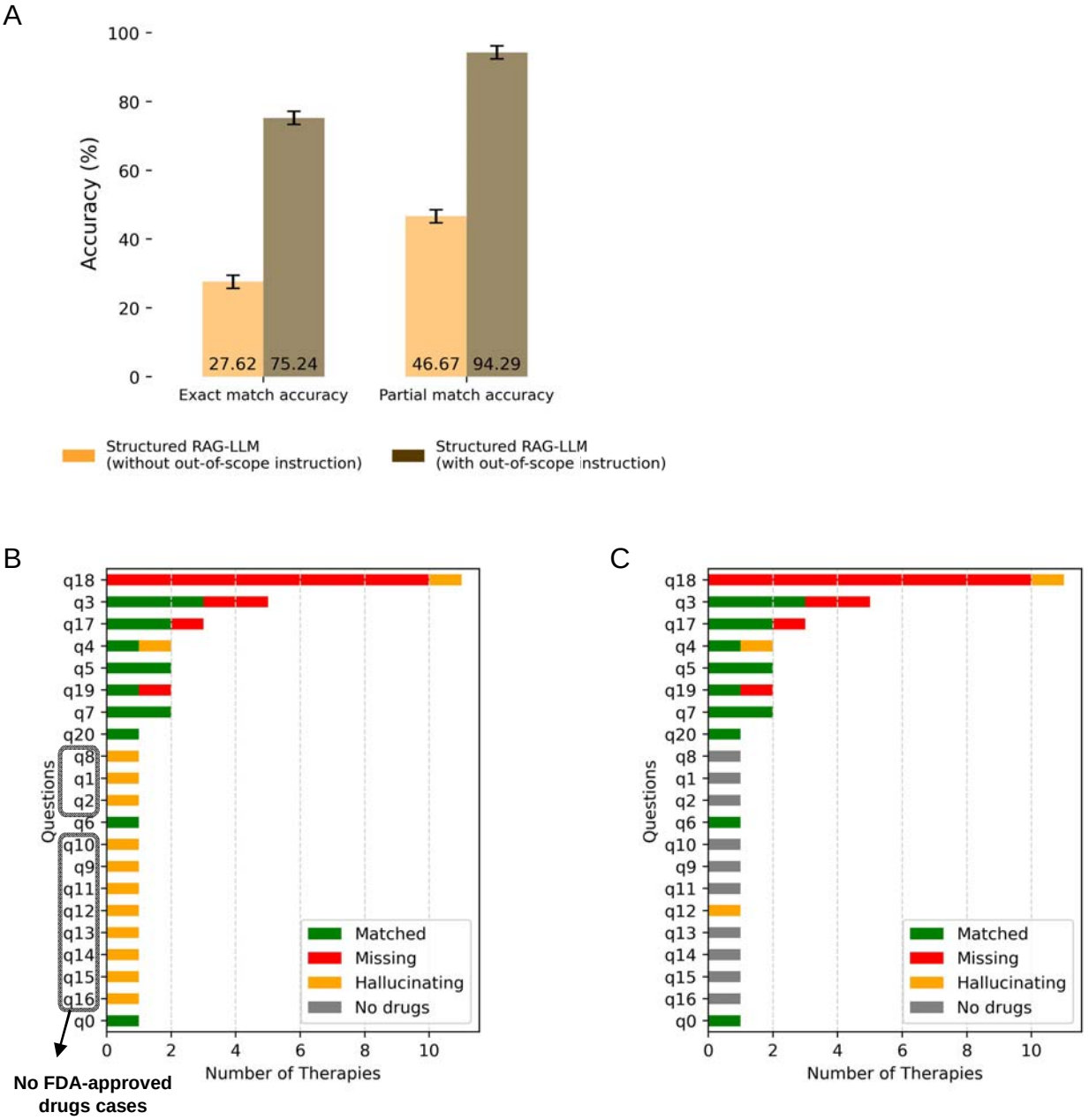
Figure 4. Performance of structured context-augmented RAG-LLM on real-world questions from practicing oncologists