

Structured Workflows and Retrieval-Augmented Generation Improve LLM Recommendations for Colorectal Liver Metastases: The METATRON Study

Maria Conticchio

`maria_cont@hotmail.it`

Inserm

Francesca Pescol

University of Pavia

Giorgia Carra

Lausanne University Hospital and University of Lausanne

Martin Hubner

Lausanne University Hospital and University of Lausanne

Stylianos Tzedakis

Cochin hospital, Université Paris Cité

Alexandra Nassar

Cochin hospital, Université Paris Cité

Ugo Marchese

Cochin hospital, Université Paris Cité

Lucia Sacchi

University of Pavia

Charles Hellec

Chaire Innovation BOPA

Riccardo Bellazzi

University of Pavia

Eric Vibert

Hôpital Paul-Brousse AP-HP

David Fuks

Cochin hospital, Université Paris Cité

Jean Louis Raisaro

Lausanne University Hospital and University of Lausanne

Keywords:

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11112101/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Large language models (LLMs) show promise for clinical decision support, but their ability to reproduce multidisciplinary oncologic reasoning is unclear. We evaluated whether structured clinical workflows and retrieval-augmented generation (RAG) improve concordance between LLM-generated and expert reference recommendations across 110 synthetic colorectal liver metastasis (CRLM) vignettes spanning eight clinical scenarios. Four inference strategies built on the same LLM (gpt-oss-120B)—baseline prompting, RAG, workflow-guided prompting, and their combination (WF-RAG-LLM)—were scored by three independent LLM judges across five dimensions. WF-RAG-LLM achieved the highest aggregated score (0.626 ± 0.070), ahead of RAG-LLM (0.559 ± 0.111), WF-LLM (0.527 ± 0.168), and baseline (0.444 ± 0.162), with the largest gains in post-chemotherapy, recurrent, and anatomically challenging scenarios. Combining expert-derived workflow guidance with retrieval-based knowledge grounding achieved the highest overall concordance with expert reference recommendations, supporting clinically informed reasoning architecture as an important component of LLM performance in complex oncologic decision support.

INTRODUCTION

Liver metastases develop in up to half of patients with colorectal cancer during the course of their disease, and the liver is the most frequent site of metastatic spread.¹ The management of colorectal liver metastases (CRLM) is among the most complex decision-making processes in surgical oncology. Treatment strategies require the integration of tumor biology, response to systemic therapy², technical resectability³, future liver remnant assessment, and the presence of extrahepatic disease.⁴ Consequently, recommendations are typically established through multidisciplinary tumor board (MDT) discussions.⁵ Despite their central role in cancer care, MDT decisions remain subject to variability arising from differences in expertise, institutional practices, and guideline interpretation, potentially leading to heterogeneous recommendations for similar clinical scenarios.^{6,7}

Recent advances in large language models (LLMs) have generated considerable interest in their potential role as clinical decision-support tools.⁸ LLMs have demonstrated strong performance across multiple medical tasks, including clinical question answering, report summarization, and therapeutic recommendation generation.^{9,10} In oncology, several studies have reported encouraging agreement between LLM-generated recommendations and expert opinions.¹¹ However, systematic reviews have also highlighted important limitations that currently preclude autonomous clinical deployment, including incomplete guideline adherence, hallucinations, poor recognition of missing information, and inconsistent performance across clinical scenarios.^{12,13,14}

Emerging evidence suggests that these limitations may depend not only on the underlying foundation model but also on the clinical architecture surrounding it. Retrieval-augmented generation (RAG) approaches can improve factual grounding^{15,16,17}, while structured prompting strategies may enhance

reasoning and workflow adherence.^{18,19,20} Nevertheless, most published studies have focused on simplified clinical tasks or non-surgical oncology settings, and the impact of structured clinical workflows on complex multidisciplinary decision-making remains largely unexplored.

CRLM represents a particularly demanding testbed for evaluating LLM-based decision support because management requires the integration of oncologic, anatomical, technical, and patient-specific factors within a sequential reasoning process. Whether expert-derived clinical workflows and guideline-grounded knowledge can improve the quality and safety of LLM-generated recommendations in this setting remains unknown.

The aim of the present study was therefore to evaluate the impact of structured clinical workflows and knowledge grounding on LLM-assisted decision-making for CRLM and to assess their ability to generate recommendations aligned with expert multidisciplinary practice. To this end, four inference strategies based on the same LLM (gpt-oss-120B)²¹ were evaluated: a baseline prompting strategy (LLM-baseline), a RAG approach (RAG-LLM), a workflow-guided clinical reasoning strategy (WF-LLM), and a combination of workflow guidance and retrieval augmentation (WF-RAG-LLM). All strategies shared an identical base prompt designed to generate structured recommendations for CRLM management, while the latter three incrementally incorporated additional knowledge-grounding and/or workflow-guidance components. A key methodological contribution of this study is the introduction of a workflow-guided approach that formalizes expert clinical knowledge into a structured reasoning framework. Embedded into the prompt, this framework guides case analysis and supports the generation of clinically relevant, consistent recommendations. The workflow was derived from the key evaluation steps routinely performed by hepatobiliary surgeons during MDT discussions, keeping expert clinical reasoning at the core of the LLM's decision-making process. In this sense, the hepatobiliary surgeon's contribution was not confined to validating the model's outputs after generation: clinicians were involved as experts in the design of the system, not only as experts in the loop reviewing it, shaping the reasoning architecture itself according to how a hepatobiliary surgeon actually analyzes a case, rather than simply checking its conclusions.

RESULTS

Study cohort and evaluation framework

A total of 110 simulated CRLM clinical vignettes were evaluated using four LLM inference strategies (LLM-baseline, RAG-LLM, WF-LLM, and WF-RAG-LLM; Fig. 1). The 110 cases covered eight representative CRLM clinical scenarios, enabling both overall and scenario-level analyses.

Each recommendation was independently assessed by three external LLM judges (Claude Sonnet 4.6, GPT-4o, and Llama 3.3 70B) against expert reference recommendations across five predefined evaluation dimensions: completeness, step concordance, case tailoring concordance, missing-data concordance, and MDT decision concordance. This multi-judge configuration was adapted from established evaluation frameworks, drawing on similar evaluation strategies applied in clinical and

oncology domains^{10,22}. In parallel, a zero-shot NER pipeline assessed entity-level concordance with the expert references.

Overall performance by evaluation dimension

Overall performance differed substantially across the five evaluation dimensions (Table 1, Fig. 2). WF-RAG-LLM achieved the highest normalized mean score across judges in four of the five dimensions, including completeness (0.980 ± 0.022), step concordance (0.393 ± 0.164), case tailoring concordance (0.802 ± 0.116), and MDT decision concordance (0.536 ± 0.139). Missing-data concordance was the only dimension in which RAG-LLM ranked first (0.486 ± 0.087), ahead of WF-RAG-LLM (0.419 ± 0.098), WF-LLM (0.239 ± 0.061), and LLM-baseline (0.130 ± 0.133).

Table 1

Overall performance of inference strategies by evaluation dimension. Values indicate mean $\pm$ standard deviation of normalized scores assigned to each strategy for each evaluation dimension, aggregated across LLM judges.

Evaluation dimension	LLM-baseline	RAG-LLM	WF-LLM	WF-RAG-LLM
Completeness	0.878 ± 0.100	0.974 ± 0.026	0.917 ± 0.106	0.980 ± 0.022
Step concordance	0.232 ± 0.260	0.260 ± 0.215	0.301 ± 0.293	0.393 ± 0.164
Case tailoring	0.647 ± 0.299	0.718 ± 0.222	0.713 ± 0.251	0.802 ± 0.116
Missing-data concordance	0.130 ± 0.133	0.486 ± 0.087	0.239 ± 0.061	0.419 ± 0.098
MDT decision concordance	0.332 ± 0.253	0.357 ± 0.165	0.467 ± 0.248	0.536 ± 0.139

Completeness showed consistently high scores across all inference strategies, whereas substantially larger differences were observed for reasoning-related dimensions, particularly step concordance, MDT decision concordance, and missing-data concordance. Workflow-guided strategies generally achieved higher normalized scores than the baseline model, and the combined WF-RAG strategy provided the highest aggregated score across evaluation dimensions.

Overall performance by judge

When scores were aggregated across the five evaluation dimensions and the three independent judges, WF-RAG-LLM achieved the highest overall composite score (0.626 ± 0.070), followed by RAG-LLM (0.559 ± 0.111), WF-LLM (0.527 ± 0.168), and LLM-baseline (0.444 ± 0.162) (Table 2, Fig. 2). Judge-specific analyses nevertheless showed substantial differences in score calibration. Claude Sonnet 4.6 assigned the lowest scores overall, ranging from 0.231 ± 0.264 for LLM-baseline to 0.530 ± 0.264 for WF-RAG-LLM. GPT-4o showed intermediate values while preserving the same ranking, with WF-RAG-LLM again achieving the highest score (0.652 ± 0.264). Llama 3.3 70B consistently produced higher normalized scores than the other judges and ranked WF-LLM first (0.735 ± 0.275), slightly ahead of RAG-LLM (0.700

± 0.246) and WF-RAG-LLM (0.696 ± 0.211). Despite these differences in score calibration, all three judges confirmed an advantage of workflow-guided and/or retrieval-augmented inference strategies over baseline, although the identity of the top-ranked strategy varied between evaluators.

Table 2

Overall performance of inference strategies by LLM judge. Values indicate mean $\pm$ standard deviation of normalized scores assigned to each strategy by each judge, aggregated across evaluation dimensions.

Inference strategy	Claude Sonnet 4.6	GPT-4o	Llama 3.3 70B	Overall
LLM-baseline	0.231 $\pm$ 0.264	0.479 $\pm$ 0.335	0.622 $\pm$ 0.347	0.444 $\pm$ 0.162
RAG-LLM	0.429 $\pm$ 0.302	0.548 $\pm$ 0.300	0.700 $\pm$ 0.246	0.559 $\pm$ 0.111
WF-LLM	0.324 $\pm$ 0.239	0.523 $\pm$ 0.336	0.735 $\pm$ 0.275	0.527 $\pm$ 0.168
WF-RAG-LLM	0.530 $\pm$ 0.264	0.652 $\pm$ 0.264	0.696 $\pm$ 0.211	0.626 $\pm$ 0.070

Performance by clinical scenario

Performance differed across the eight CRLM clinical scenarios (Fig. 3). Completeness remained consistently high across all categories, whereas larger variations were observed for step concordance, missing-data concordance, and MDT decision concordance.

Metachronous liver metastases showed the highest overall performance across several evaluation dimensions, with WF-RAG-LLM reaching a completeness score of 1.000 and the highest MDT decision concordance (0.900). Conversely, post-chemotherapy metastases represented the most challenging scenario, with the lowest step concordance (0.13–0.45, depending on inference strategy) and the lowest MDT decision concordance (0.07–0.52). For both evaluation dimensions, this scenario showed a clear increasing trend in alignment scores from LLM-baseline to the most comprehensive strategy, WF-RAG-LLM. Anatomically challenging cases also remained difficult for missing-data concordance (0.21–0.58, with RAG-LLM achieving the best score), particularly for the baseline strategy. Overall, workflow-guided and hybrid strategies demonstrated the greatest advantage in clinically complex scenarios (post-chemotherapy metastases and synchronous potentially resectable disease after systemic therapy), especially for the step-concordance and MDT concordance dimensions. Differences were less pronounced in less clinically challenging scenarios.

Inference strategy-specific gains

Retrieval-augmented strategies improved missing-data recognition. Although missing-data concordance varied considerably across clinical scenarios and inference strategy, RAG-LLM achieved the highest score in most scenarios, while WF-RAG-LLM ranked first in the remaining categories (Fig. 3). The largest performance differences between strategies were observed for anatomically challenging cases, liver recurrences after surgery, and metachronous liver metastases. In contrast, differences were modest

for post-chemotherapy liver metastases, where all four strategies achieved relatively similar performance. Overall, across nearly all clinical scenarios, retrieval-augmented strategies (RAG-LLM and WF-RAG-LLM) achieved higher missing-data concordance scores than LLM-baseline, supporting the contribution of retrieval to identifying clinically relevant missing information.

Workflow-guided strategies improved anatomical resectability classification. Anatomical resectability classification, extracted from Step 3 of each recommendation, was further evaluated against the expert reference classification using confusion matrices, F1-score, precision and recall, computed separately for each class (anatomically compliant and non-compliant; Table 3, Fig. 4). Based on expert classification, non-compliant cases were considerably less frequent (15/110 cases) than compliant ones. Focusing on the non-compliant classification, LLM-baseline and RAG-LLM showed high specificity but identified only one true positive case, resulting in poor sensitivity. In contrast, workflow-guided strategies increased true-positive detection, with WF-LLM identifying five non-compliant cases and WF-RAG-LLM identifying six, corresponding to a sensitivity of 40%, which remains insufficient for autonomous clinical use despite the improvement over baseline. Consistent with these findings, classification performance improved from a macro F1-score of 0.495 for LLM-baseline and 0.510 for RAG-LLM to 0.632 for WF-LLM and 0.643 for WF-RAG-LLM. Similar improvements were observed for macro precision and macro recall, indicating that integrating a structured clinical workflow markedly enhanced, but did not fully resolve, anatomical resectability assessment.

Table 3
Anatomical resectability classification performance.
Values indicate macro-averaged F1-score, precision, and recall for each inference strategy, computed by comparing the anatomical resectability classification extracted from Step 3 of each recommendation against the expert reference classification, across the two classes (anatomically compliant, non-compliant).

Inference strategy	Precision	Recall	F1-score
LLM-baseline	0.503	0.502	0.495
RAG-LLM	0.559	0.518	0.510
WF-LLM	0.641	0.625	0.632
WF-RAG-LLM	0.640	0.647	0.643

NER-based entity-level concordance

Entity-level concordance with expert references, measured as the case-averaged mean of token set ratio, exact match, and Levenshtein similarity across five oncology-specific entity labels (biomarker, cancer surgery, imaging test, staging, therapy), was consistently higher at Step 3 than at Step 2 of recommendations, across all inference strategies (Fig. 5, Supplementary Figs. 2 and 3). Cancer surgery remained the most challenging label at both steps, with LLM-baseline showing a slight edge over workflow-based approaches. The improvement associated with workflow-based approaches was most

evident for biomarker and staging, and to a lesser extent for therapy; notably, most of this gap was already established at the workflow stage, prior to the complete WF-RAG-LLM approach. Overall, across most labels, workflow guidance appeared to increase lexical and syntactic alignment with the expert reference.

Statistical analysis

Statistical significance was assessed using the Friedman test on ordinal scores (1–5 Likert scale), performed separately for each judge and evaluation dimension. Sensitivity to inference strategy differences varied considerably across judges (Supplementary Table 1): Claude Sonnet 4.6 showed the highest sensitivity, with statistically significant differences observed across all evaluated criteria; GPT-4o confirmed significance on three criteria: step concordance, missing-data concordance, and MDT decision concordance; Llama 3.3 70B showed the lowest sensitivity, with significance confirmed only for missing-data concordance.

At the pairwise level (Supplementary Table 2), none of the comparisons reached statistical significance for Llama. For Claude, WF-RAG-LLM scored significantly higher than every other strategy across all criteria except missing-data concordance, with a further exception for completeness when compared to RAG-LLM, which likewise failed to reach significance. GPT, in turn, showed WF-RAG-LLM scoring significantly higher only against LLM-baseline, and only for the step concordance and MDT decision criteria. P values < 0.05 were considered statistically significant.

METHODS

Study design

This study was conducted as a methodological validation within the METATRON project, an expert-developed framework for LLM-assisted decision support in CRLM. The primary objective was to determine whether a structured clinical reasoning architecture, alone or combined with RAG, improves the alignment of LLM-generated recommendations with expert hepatobiliary MDT decision-making.

Four inference strategies built upon the same LLM (gpt-oss-120B, deployed locally) were evaluated (LLM-baseline, RAG-LLM, WF-LLM, and WF-RAG-LLM; Fig. 1) to generate recommendations for the same standardized set of CRLM clinical vignettes. Model outputs were assessed through a two-step evaluation approach. First, three independent LLM judges (Claude Sonnet 4.6, GPT-4o, and Llama 3.3 70B) rated concordance with expert references across multiple CRLM-specific clinical dimensions on a 1–5 Likert scale, adapting the multi-judge configuration proposed by established evaluation frameworks^{10,22,30,31}. Second, a zero-shot named-entity recognition (NER) pipeline assessed entity-level concordance with expert references.

Ethics

Institutional review board approval was not required because the study used exclusively synthetic clinical vignettes and involved no identifiable patient data or medical records. The system was evaluated exclusively for methodological research and was not used to support real-time clinical decision-making.

Clinical vignettes

A total of 110 synthetic CRLM clinical vignettes were developed by experienced hepatobiliary surgeons. Cases were not derived from identifiable patient records but were specifically designed to reproduce realistic multidisciplinary tumor board scenarios encountered in routine clinical practice. The dataset covered eight representative clinical scenarios: synchronous resectable liver metastases (13.64%), synchronous potentially resectable disease after systemic therapy (9.09%), metachronous metastases (13.64%), post-chemotherapy reassessment (9.09%), liver recurrence after previous hepatectomy (11.82%), liver metastases associated with extrahepatic disease (10.91%), molecularly complex situations (22.73%), and anatomically challenging cases (9.09%).

All vignettes were written in English using a semi-standardized structure including demographic, oncological, molecular, radiological and surgical information when appropriate. Selected cases deliberately omitted decision-critical variables to evaluate the ability of each inference strategy to recognize insufficient information before issuing a therapeutic recommendation.

Expert reference standard

For each vignette, an expert reference recommendation was independently generated by a single senior hepatobiliary surgeon experienced in CRLM multidisciplinary decision-making. Reference recommendations included the proposed oncological strategy, assessment of resectability, identification of missing decision-critical information, and anatomical feasibility whenever applicable. These expert recommendations were finalized before generation or review of any LLM output and served as the reference standard throughout the study.

Expert-derived clinical workflow implementation

To guide and formalize the analysis of CRLM clinical cases, we implemented a structured, expert-derived workflow reproducing the key evaluation steps routinely performed by hepatobiliary surgeons in tumor board discussions.

The workflow resulted from a multidisciplinary collaboration between clinicians and engineers. Hepatobiliary surgeons defined the main evaluation steps and associated clinical questions underlying CRLM decision-making, grounding them in both expert MDT practice and the major decision domains described in international CRLM management guidelines, including those of the European Society for Medical Oncology (ESMO)²³, the National Comprehensive Cancer Network (NCCN)²⁴, and the Americas

Hepato-Pancreato-Biliary Association (AHPBA)²⁵. Rather than reproducing individual recommendations, the workflow captured the sequence of clinical questions that expert surgeons systematically address before reaching a therapeutic decision.

This expert-defined framework was formalized as a tree-structured evaluation procedure, encoded in a JSON schema. Instead of operating as a linear checklist, it was organized as a clinical hierarchical reasoning architecture composed of independent decision domains, each expanding into predefined clinical items and conditional decision nodes that reproduce the adaptive nature of expert reasoning.

Four complementary reasoning domains (*steps*) constitute the backbone of this architecture, referred to hereafter as workflow domains A–D (Supplementary Fig. 1):

- Domain A – Primary tumor assessment, including characteristics of the primary tumor, molecular profile, treatment of the primary lesion and biological risk factors.
- Domain B – Hepatic disease assessment, including the number, size and distribution of liver metastases, vascular and biliary involvement, future liver remnant (FLR), technical resectability and response to systemic therapy.
- Domain C – Extrahepatic disease assessment, including the presence, anatomical distribution, biological behavior, disease control and potential resectability of extrahepatic metastases.
- Domain D – Anatomical feasibility for surgery, including future liver remnant (FLR) adequacy, vascular involvement, biliary drainage, and the achievability of R0 resection or, when oncologically and technically appropriate, R1 vascular resection.

Within each domain, major decision nodes (item nodes) expand hierarchically into clinically relevant sub-items (leaf nodes) and conditional branches (condition nodes). Some evaluations are activated only when clinically appropriate. For example, assessment of FLR adequacy and consideration of portal vein embolization are required only when a major hepatectomy is anticipated, whereas evaluation of extrahepatic disease control determines whether a curative-intent strategy remains appropriate. This conditional organization enables the architecture to reproduce the adaptive reasoning process of expert MDT discussions. The complete JSON schema implementing the workflow is provided in the project GitHub repository (https://github.com/BDSC-CDS/METATRON_project/blob/master/generation/workflow_based_case_analysis/workflow/).

LLM inference strategies

RAG-LLM. The retrieval component of the RAG-LLM inference strategy was implemented using a vector database as the underlying storage. The literature corpus consisted of 292 documents derived from international clinical guidelines, primarily the ESMO, NCCN and AHPBA recommendations, supplemented with peer-reviewed articles to incorporate up-to-date research findings (full list is provided in the project GitHub repository, <https://github.com/BDSC->

CDS/METATRON_project/blob/master/generation/data/studies_information.csv). A hybrid chunking strategy was adopted to optimize document representation, combining paragraph-based segmentation for structured sources with fixed-length chunking, resulting in 5,379 indexed text segments stored in the vector database. Each chunk was encoded into dense vector representations using Qwen3-0.6B embedding model²⁶. At inference time, retrieval was performed using case-specific clinical queries, encoded as embeddings and matched against indexed representations via cosine similarity. For each case, the top 30 most relevant corpus passages were selected and included into the prompt to support context-grounded generation.

WF-LLM. The workflow-guided strategy was designed to reproduce the cognitive reasoning process of expert hepatobiliary surgeons during MDT discussions. The objective was not to increase the amount of information produced by the LLM, but to constrain its reasoning through a reproducible clinical decision framework that explicitly reflects how CRLM cases are analyzed in routine multidisciplinary practice. The model is therefore required to explicitly evaluate every major component of CRLM management before synthesizing a final recommendation. In this study, the expert-derived workflow was used to guide recommendation generation through a two-phase process, consisting of separate inference calls to the same LLM. In both phases, the model was prompted to act as a hepatobiliary surgeon.

Phase 1 (structured case analysis)

guided by the JSON schema-based workflow, the LLM produced a structured analysis of the clinical case. This phase was implemented through an iterative inference approach that mirrored the hierarchical structure of the workflow. Specifically, the model was queried stepwise to extract structured information from the case description using portions of the JSON schema, which were progressively incorporated into a prompt template at each iteration. Across the workflow steps, for each item node, the corresponding portion of the JSON schema was injected into the prompt template, and the model was required to return a structured output with answers to the associated clinical evaluation questions (either a specific extracted value or "Missing" when the required information was unavailable). When an item node included a conditional element, the corresponding information was extracted only after verifying the specified condition, enabling adaptive navigation of the workflow. This approach allowed simulation of an interactive clinical tool, in which the model could explicitly identify missing information and structure intermediate assessments at the level of each evaluation step.

Phase 2 (recommendation generation)

the structured analysis produced in Phase 1 was incorporated as input into the base prompt shared across the four strategies, accompanied by dedicated instructions to contextualize it and guide its use in generating the final recommendation. In this step, the model was prompted to generate the recommendation following the analytical flow established by the structured analysis, which enforced a consistent reasoning order, from primary tumor assessment to anatomical resectability, and surfaced information that might otherwise remain implicit, including relevant missing data identified during Phase 1.

By separating structured analysis from final generation, this two-stage design promoted more consistent, interpretable, and clinically grounded outputs. The workflow constrained the clinical decision space by requiring explicit evaluation of each reasoning domain, systematic consideration of all decision-critical variables, and transparent reporting of missing information before formulating the final recommendation, ensuring that every output followed the same reproducible analytical framework.

The structured analyses generated in Phase 1 for all 110 cases included in this study are available in the project GitHub repository (https://github.com/BDSC-CDS/METATRON_project/tree/master/generation/structured_information_files), while a representative example is provided in the Supplementary Materials (SM2).

WF-RAG-LLM. The last tested strategy integrated both previous contributions, jointly incorporating the retrieved context and the structured analysis within the shared base prompt. In this way, we evaluated their combined impact on recommendation generation.

Prompt design

Three base prompt templates were used across the evaluated strategies: one for structured case analysis according to the workflow (Phase 1 of the WF-LLM strategy), comprising two template versions, one for item nodes and one for condition nodes; one for retrieving relevant chunks from the literature corpus; and one, shared across all four strategies, for generating the final recommendation. The complete templates are provided in the Supplementary Materials (SM1).

The recommendation prompt was iteratively refined through multiple rounds of review with hepatobiliary experts. It requires the model to structure the final recommendation into three sequential steps, detailed below.

Step 1 – Overall oncological strategy. The model first formulates an initial multidisciplinary treatment strategy using the information available. This step intentionally requires the model to commit to a management plan despite potential missing information, thereby reproducing the initial reasoning process performed during real MDT discussions.

Step 2 – Identification of missing decision-critical information. After formulating the initial strategy, the model is required to perform a structured verification of the clinical information supporting its recommendation. All variables considered necessary to validate, modify or refine the proposed strategy have to be explicitly identified. Positioning this step after the initial recommendation mirrors expert clinical reasoning: clinicians first develop a provisional therapeutic strategy and subsequently assess which additional information is required before confirming the final decision. This sequential verification transforms missing-data identification into an explicit assessment of recommendation robustness rather than a simple checklist.

Step 3 – Hepatobiliary surgical assessment. The final module provides the dedicated perspective of an expert hepatobiliary surgeon. Independently of the global oncological strategy, the model evaluates anatomical resectability according to predefined surgical criteria, including FLR adequacy, vascular inflow and outflow preservation, biliary anatomy, anticipated surgical margins and technical feasibility. This dedicated surgical assessment reflects the specific contribution of hepatobiliary surgeons during CRLM MDT discussions and ensures that technical operability is analyzed independently from oncological strategy.

Supplementary Materials (SM1) report the final version of the recommendation prompt, while earlier iterations developed during the expert refinement process are available in the project GitHub repository (https://github.com/BDSC-CDS/METATRON_project/tree/master/generation/recommendation_strategies/prompt).

LLM-as-a-jury evaluation framework

Building on the growing adoption of LLM-based automated evaluation approaches in clinical settings, each recommendation was independently evaluated using an LLM-as-a-jury framework^{10,22}. Three external judges, Claude Sonnet 4.6³², GPT-4o²⁷ and Llama 3.3 70B²⁸, received the clinical vignette, the expert reference recommendation and the candidate recommendation generated by a given inference strategy. Claude and Llama models were accessed via Amazon Bedrock, while GPT-4o was accessed using an API key. Recommendations were evaluated on factual alignment with the expert reference recommendation, across five predefined dimensions using five-point Likert scales (see full prompt in the Supplementary Materials, SM3):

1. *Completeness*: assesses whether the candidate recommendation covers the same major clinical domains as the expert reference.
2. *Step concordance*: assesses whether the candidate recommendation reproduces the same sequence of major clinical decisions as the expert reference.
3. *Case tailoring concordance*: assesses whether the candidate recommendation incorporates the same case-specific clinical factors and patient-specific reasoning as the expert reference.
4. *Missing-data concordance*: assesses whether the candidate recommendation identifies the same decision-critical missing information as the expert reference.
5. *MDT decision concordance*: assesses whether the candidate recommendation would lead to the same final MDT treatment decision as the expert reference, independent of individual supporting details.

This evaluation was performed independently for every vignette, inference strategy, and judge. The final score for each strategy along each evaluation criterion was computed as the mean rating across all cases, calculated separately for each judge.

Because the three LLM judges differed in their scoring calibration, criterion-level scores were then normalized within each judge before aggregation. These normalized scores were summarized in two complementary ways:

- mean value across judges for each evaluation dimension, to compare inference strategies and localize the specific dimensions along which they differ;
- mean value across evaluation dimensions within each judge, to assess whether judges rated strategies consistently or showed evidence of systematic rating discrepancies.

A final overall score was obtained by averaging normalized scores across both judges and evaluation dimensions.

NER-based pipeline

To complement the holistic scoring provided by the LLM-jury, we performed a granular, lexical-level comparison between candidate and expert reference recommendations. This analysis was designed to capture concordance at the level of individual clinical entities, rather than relying solely on the overall qualitative judgments produced by the jury.

Entity extraction was performed using the John Snow Labs (JSL) zero-shot named entity recognition model (zeroshot_ner_oncology_large), designed to detect and label oncology-related entities without requiring task-specific training^{29,33}. The label set includes, among others, imaging test, metastasis, and cancer surgery. The model was run using sparknlp_jsl version 6.3.0.

The NER pipeline was applied to both the candidate recommendations generated by each strategy and the expert reference recommendations, and the resulting entity sets were compared using similarity metrics (token set ratio, exact match, and Levenshtein similarity, computed as $1 - \text{normalized Levenshtein distance}$)³⁴. To enable a finer-grained comparison, entity extraction was performed separately for Step 2 and Step 3 of each recommendation.

Study outcomes

This study was designed as a multidimensional evaluation of concordance between LLM-generated and expert reference recommendations, without a single prespecified primary endpoint subjected to a global hypothesis test. Predefined outcomes comprised the five evaluation dimensions assessed by the LLM-as-a-jury framework (completeness, step concordance, case tailoring concordance, missing-data concordance, and MDT decision concordance), each analyzed separately by evaluation criterion and judge. As a descriptive summary measure, an aggregated composite concordance score was also computed for each inference strategy and vignette, obtained by normalizing scores within each judge and averaging normalized scores across the three independent LLM judges and the five evaluation

dimensions; this composite score is reported descriptively (Table 2, Fig. 2) and was not itself subjected to formal statistical hypothesis testing.

Additional predefined analyses included a scenario-level analysis: for each of the eight case scenarios, normalized mean scores assigned by LLM judges were aggregated for every evaluation dimension and inference strategy to identify scenario-specific strengths and limitations that may not be apparent from global performance metrics alone. A further analysis compared the anatomical resectability classification identified from Step 3 of the evaluated and expert recommendations, using confusion matrices together with macro-average precision, recall and F1-score. As an exploratory analysis, entity-level alignment with the expert reference was assessed via the NER-based pipeline and summarized by averaging similarity metrics across cases.

Statistical analysis

Statistical analyses were performed on the evaluation results obtained through the LLM-as-a-jury assessment. The different strategies were compared using the Friedman test for repeated measures, with cases treated as blocks, applied to the ordinal scores assigned on the Likert scale. A separate test was conducted for each evaluation criterion and each judge. Post hoc pairwise comparisons between strategies were performed using the Wilcoxon signed-rank test, with Holm correction applied for multiple comparisons.

DISCUSSION

LLMs are increasingly being investigated as decision-support tools in oncology. However, most studies have focused on comparing foundation models or prompting strategies, while relatively little attention has been paid to the architecture through which clinical reasoning is organized. In this methodological evaluation of AI-assisted multidisciplinary decision support for colorectal liver metastases (CRLM), we demonstrate that recommendation quality depended on the inference strategy guiding generation, shaped by both structured clinical reasoning and retrieval-augmented context. Across five complementary evaluation dimensions, the workflow-guided retrieval-augmented strategy (WF-RAG-LLM) achieved the best overall performance, whereas retrieval augmentation alone provided the greatest benefit for identifying missing decision-critical information. Together, these findings suggest that clinically structured reasoning and knowledge grounding represent complementary components of reliable AI-assisted multidisciplinary decision support rather than interchangeable prompting strategies.

A central contribution of this work is the introduction of an expert-derived clinical reasoning architecture rather than a conventional prompting strategy. The workflow was not designed to increase the amount of information generated by the model, but to reproduce the cognitive reasoning process routinely followed by hepatobiliary surgeons during multidisciplinary tumor board discussions. Instead of allowing the model to generate recommendations through unconstrained inference, the architecture explicitly constrains clinical reasoning into reproducible decision domains derived from international CRLM

guidelines and expert MDT practice. In this way, the workflow structures the clinical decision space by requiring systematic evaluation of primary tumor characteristics, hepatic disease, extrahepatic disease, missing decision-critical information and anatomical feasibility before a final recommendation is formulated. By making the reasoning pathway explicit, reproducible and auditable, the architecture provides a reproducible alternative to unconstrained prompting while remaining faithful to expert clinical practice.

One of the most striking observations is that all the evaluated inference strategies achieved similarly high completeness scores. Modern LLMs therefore appear capable of generating clinically comprehensive recommendations even without explicit reasoning guidance. In contrast, the greatest differences emerged in reasoning-dependent dimensions, including step concordance, case tailoring and MDT decision concordance. These findings suggest that the principal contribution of the workflow is not to generate more information but to organize existing information according to an expert clinical reasoning process. In CRLM management, recommendations are expected to integrate tumor biology, previous treatments, anatomical feasibility, future liver remnant assessment and therapeutic sequencing into a coherent multidisciplinary strategy. Explicitly embedding this reasoning sequence within the inference strategy improved agreement with the expert reference recommendations.

Retrieval augmentation and workflow guidance contributed differently to model performance, highlighting that they address complementary cognitive functions. Retrieval augmentation achieved the highest performance for missing-data concordance across most clinical scenarios, whereas workflow guidance primarily improved reasoning coherence and agreement with the expert reference recommendations. This distinction is clinically meaningful. These findings are consistent with the hypothesis that retrieval augmentation does not simply provide additional knowledge, but may also enhance the model's ability to recognize when the available information is insufficient to support a reliable recommendation, whereas workflow guidance may constrain how available information is interpreted and sequentially integrated into clinical decision-making. The superior overall performance of the hybrid WF-RAG strategy is consistent with a combination of these complementary mechanisms: retrieval improving awareness of informational uncertainty, and the workflow-guided strategy improving organization of clinical reasoning.

The deliberate separation of the recommendation into three sequential reasoning modules also deserves attention. The model is first required to formulate a global oncological strategy using the available information, then to explicitly identify missing decision-critical variables, and finally to provide an independent hepatobiliary surgical assessment of anatomical resectability. This sequence mirrors the cognitive process typically observed during expert MDT discussions. Clinicians frequently formulate a provisional therapeutic strategy before critically reassessing its robustness in light of missing information and finally incorporating the specific perspective of the hepatobiliary surgeon regarding technical feasibility. Positioning missing-data identification after the initial recommendation therefore transforms it from a simple checklist into an explicit assessment of recommendation robustness, while

the dedicated surgical module ensures that anatomical feasibility is evaluated independently from the global oncological strategy.

The scenario-level analyses further demonstrate that benchmarking AI systems using a single global performance score may obscure clinically important heterogeneity. Completeness remained consistently high across all CRLM scenarios, whereas substantially greater variability was observed for reasoning-related dimensions. In particular, post-chemotherapy cases consistently represented the most challenging scenarios, while metachronous liver metastases achieved the highest overall performance. These differences are consistent with variations in cognitive complexity rather than factual knowledge, although this interpretation was not directly tested and warrants further investigation. Post-chemotherapy decision-making requires simultaneous integration of treatment history, radiological response, tumor biology, evolving anatomical resectability and therapeutic sequencing, whereas other scenarios follow more constrained decision pathways. These findings suggest that future benchmarking studies in surgical oncology should complement global performance metrics with scenario-specific analyses capable of identifying clinical contexts in which AI support remains less reliable.

The analysis of missing-data concordance addresses a particularly important safety aspect of AI-assisted decision support. In multidisciplinary oncology, inappropriate recommendations frequently arise because critical information has not yet been obtained rather than because clinical reasoning itself is incorrect. Variables such as chemotherapy exposure, molecular profile, control of extrahepatic disease, liver function or future liver remnant characteristics may fundamentally alter management decisions. The consistently higher performance of retrieval-augmented strategies is consistent with the hypothesis that retrieval mechanisms may support not only factual grounding but also recognition of informational uncertainty. From a clinical perspective, explicitly acknowledging that available information is insufficient may be as valuable as producing the final recommendation itself, particularly in multidisciplinary tumor boards where incomplete datasets are common.

The dedicated evaluation of anatomical resectability provides an additional and complementary perspective. Unlike the global recommendation scores, anatomical assessment represents a specific surgical reasoning task based on predefined technical criteria. Baseline and retrieval-only strategies detected very few anatomically non-compliant cases, whereas workflow-guided strategies increased true-positive detection of this category and achieved the highest macro-average classification performance, although sensitivity remained insufficient for autonomous clinical use, underscoring the continued need for human oversight of anatomical feasibility assessment. Although the limited number of anatomically non-compliant cases precludes definitive conclusions, these findings suggest that explicit clinical reasoning architectures facilitate structured surgical assessment in addition to improving overall recommendation quality. More broadly, they indicate that global recommendation concordance and anatomical feasibility assessment represent complementary endpoints that should not be considered interchangeable when evaluating AI-assisted MDT systems.

Beyond CRLM, the present work supports a broader methodological concept. Rather than considering prompting as a simple implementation detail, our results suggest that clinical reasoning architecture represents an important component of AI performance. The way clinical information is structured, sequentially analyzed and integrated may therefore complement improvements achieved through increasingly capable foundation models. Although the present workflow was specifically developed for CRLM management, its underlying hierarchical reasoning principles are potentially transferable to other complex oncological settings through adaptation of disease-specific decision domains. This distinction shifts the focus from language generation towards explicit modelling of expert clinical reasoning as the central object of evaluation.

This study has several strengths. First, it evaluates AI-assisted decision support within a highly specialized hepatobiliary oncology setting rather than generic medical question answering. Second, the use of 110 expert-designed CRLM vignettes spanning eight clinically relevant scenarios enabled systematic evaluation across a broad spectrum of multidisciplinary decision problems. Third, unlike previous studies primarily comparing foundation models or prompting techniques, the present work evaluates an expert-derived clinical reasoning architecture specifically designed to reproduce multidisciplinary expert cognition. Finally, the use of three independent LLM judges provided a scalable evaluation framework: the overall advantage of workflow-guided and retrieval-augmented inference strategies was observed across judges, although the magnitude and, for one judge, the ranking of individual strategies varied between evaluators.

Several limitations should nevertheless be acknowledged. First, this was a simulation-based study using expert-designed clinical vignettes rather than prospective MDT cases. Although this approach enabled controlled comparisons, it cannot fully reproduce the uncertainty, interpersonal interactions and contextual complexity of real multidisciplinary discussions. Second, the reference recommendations were defined by a single senior hepatobiliary expert rather than by multidisciplinary consensus, and therefore do not constitute an absolute gold standard; involvement of additional independent reviewers, with formal adjudication of disagreements, could further strengthen the reference standard in future work. Third, performance was evaluated using an LLM-as-judge framework, and although judge-specific rankings confirmed the overall advantage of augmented inference strategies, we did not compute a formal inter-rater agreement statistic (such as an intraclass correlation coefficient) between the three LLM judges; future work should quantify this agreement and compare automated assessments with larger panels of hepatobiliary experts to better characterize agreement between human and AI-based evaluation. Fourth, the present study assessed recommendation quality rather than downstream clinical utility, workflow efficiency, user acceptance or patient outcomes. Fifth, all four inference strategies were built upon the same underlying LLM (gpt-oss-120B); future work should extend this evaluation to alternative models to assess the generalizability of these findings across different LLM backbones. Finally, although the proposed reasoning architecture was specifically developed for CRLM, disease-specific adaptation and prospective validation will be required before extending this methodology to other oncological indications.

These findings have important implications for the future development of AI-assisted multidisciplinary decision-support systems. Rather than focusing exclusively on increasingly capable foundation models, future efforts should emphasize clinically informed reasoning architectures capable of reproducing expert decision pathways, explicitly identifying missing information and integrating disease-specific expertise into transparent and auditable clinical recommendations. Evaluation frameworks should likewise move beyond single global performance metrics and incorporate scenario-level analyses together with task-specific endpoints that better reflect the complexity of multidisciplinary decision-making.

The next step is not immediate clinical deployment but prospective validation within real multidisciplinary tumor boards. Future studies should evaluate consecutive CRLM cases, compare AI recommendations with prospective MDT consensus and assess clinically meaningful outcomes, including correction of omitted variables, effects on discussion efficiency, inter-specialty agreement, user confidence and ultimately patient management. Independent benchmarking by expert hepatobiliary surgeons, particularly for anatomical resectability assessment, will remain essential before routine clinical implementation.

In conclusion, this study demonstrates that the performance of LLM-assisted decision support for CRLM depends on the inference strategy guiding generation, shaped by both structured clinical reasoning and retrieval-augmented context. By formalizing the cognitive process of expert hepatobiliary surgeons into a reproducible and auditable reasoning framework, workflow guidance improved reasoning coherence and agreement with the expert reference recommendations, whereas retrieval augmentation enhanced recognition of missing decision-critical information. Their combination achieved the highest aggregated overall performance. More broadly, these findings suggest that clinically informed reasoning architectures may represent an important component of performance complementary to foundation-model capability, capable of reproducing expert clinical reasoning in a transparent, reproducible and disease-specific manner.

Declarations

DATA AVAILABILITY

Data are available at the following link: https://github.com/BDSC-CDS/METATRON_project/blob/master/generation/data/.

CODE AVAILABILITY

Code and materials supporting the study are available at the following link: https://github.com/BDSC-CDS/METATRON_project.

COMPETING INTERESTS

The authors declare no competing interests.

Author Contribution

MC and FP: study conception and design. MC, FP, GC, ST, AN, UM, LS and CH: acquisition, curation and analysis of data. JLR and MH: statistical analysis. GC, ST, AN, UM, LS, CH, RB, EV and DF: interpretation of results and critical revision of the manuscript for important intellectual content. MC, EV, JLR and DF: supervision. All authors read and approved the final manuscript.

ACKNOWLEDGEMENTS

The authors received no specific funding for this work and have no individuals to acknowledge.

References

1. Engstrand J, Nilsson H, Strömberg C, Jonas E, Freedman J. Colorectal cancer liver metastases – a population-based study on incidence, management and survival. *BMC Cancer*. 2018;18:78. doi:10.1186/s12885-017-3925-x.
2. Loupakis F, et al. Initial therapy with FOLFOXIRI and bevacizumab for metastatic colorectal cancer. *N Engl J Med*. 2014;371(17):1609–1618. doi:10.1056/NEJMoa1403108.
3. Kanas GP, et al. Survival after liver resection in metastatic colorectal cancer: review and meta-analysis of prognostic factors. *Clin Epidemiol*. 2012;4(1):283–301. doi:10.2147/CLEP.S34285.
4. Chinese College of Surgeons, Section of Gastrointestinal Surgery, Branch of Surgery, Chinese Medical Association, Section of Colorectal Surgery, Branch of Surgery, Chinese Medical Association, Colorectal Cancer Professional Committee, Chinese Anti-Cancer Association, et al. Guideline for diagnosis and comprehensive treatment of colorectal liver metastases (Version 2025). *Chin J Colorectal Dis (Electron Ed)*. 2025;14(5):398–411. doi:10.3877/cma.j.issn.2095-3224.2025.05.004.
5. Williams GJ, Thompson JF. Management changes and survival outcomes for cancer patients after multidisciplinary team discussion: a systematic review and meta-analysis. *Cancer Treat Rev*. 2025;139:102997. doi:10.1016/j.ctrv.2025.102997.
6. Walraven JEW, et al. Factors influencing the quality and functioning of oncological multidisciplinary team meetings: results of a systematic review. *BMC Health Serv Res*. 2022;22(1):829.
7. Grössmann-Waniek N. Tumour boards and their quality of structures, processes, and team performance in multidisciplinary cancer care: a systematic review. *BMC Health Serv Res*. 2026;26:676. doi:10.1186/s12913-026-14447-9.
8. Yum JEI, Naqvi SAA, Zhou B, Riaz IB. Reimagining cancer care with generative artificial intelligence: the promise of large language models. *JCO Clin Cancer Inform*. 2025;9:e2500134.

9. Singhal K, et al. Large language models encode clinical knowledge. *Nature*. 2023;620(7972):172–180.
10. Bedi S, et al. MedHELM: holistic evaluation of large language models for medical tasks. *arXiv*. Preprint posted online June 2, 2025:arXiv:2505.23802.
11. Bertolo R, et al. Role of large language models in the multidisciplinary decision-making process for patients with renal cell carcinoma: a pilot experience. *NPJ Precis Oncol*. 2025;9:257. doi:10.1038/s41698-025-01014-4.
12. Chen D, et al. Large language models in oncology: a review. *BMJ Oncol*. 2025;4(1):e000759. doi:10.1136/bmjonc-2025-000759.
13. Zhong R, et al. Large language models in lung cancer: systematic review. *J Med Internet Res*. 2025;27:e74177.
14. Ji Z, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):Article 248. doi:10.1145/3571730.
15. Kotal A, Garza L, Grasso MA, Umucu E. Retrieval-augmented framework for LLM-based clinical decision support. *arXiv*. Preprint posted online 2025:arXiv:2510.01363.
16. Testagrose CT, et al. Automated prediction of radiological protocols using retrieval augmented generation. *Res Sq*. Preprint posted online 2025. doi:10.21203/rs.3.rs-7623430/v1.
17. Berk B. MicroRAG: development of a novel artificial intelligence retrieval-augmented generation model for microsurgery clinical decision support. *Microsurgery*. 2025. doi:10.1002/micr.70138.
18. Ayoub M, et al. Structured clinical approach to enable large language models to be used for improved clinical diagnosis and explainable reasoning. *Commun Med*. 2026;6:86. doi:10.1038/s43856-025-01348-x.
19. Balachandran A, Durgapraveen B, Sudhagar GS, J S VV, Rajkumar S. EXL Health AI Lab at MEDIQA-OE 2025: evaluating prompting strategies with MedGemma for medical order extraction. *arXiv*. Preprint posted online 2025:arXiv:2511.10583.
20. Wei J, et al. Chain-of-thought prompting elicits reasoning in large language models. *arXiv*. Preprint posted online January 2022:arXiv:2201.11903.
21. Agarwal S, et al. gpt-oss-120b & gpt-oss-20b model card. *arXiv*. Preprint posted online August 2025:arXiv:2508.10925.
22. Zheng L, et al. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *arXiv*. Preprint posted online June 9, 2023:arXiv:2306.05685.
23. Cervantes A, et al. Metastatic colorectal cancer: ESMO Clinical Practice Guideline for diagnosis, treatment and follow-up. *Ann Oncol*. 2023;34(1):10–32. doi:10.1016/j.annonc.2022.10.003.
24. Benson AB, et al. Colon cancer, version 3.2024, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw*. 2024;22(2D):e240029. doi:10.6004/jnccn.2024.0029.
25. Yohanathan L, et al. Assessment and treatment considerations for patients with colorectal liver metastases: AHPBA consensus guideline and update for surgeons. *HPB (Oxford)*. 2025;27(3):263–

278. doi:10.1016/j.hpb.2024.12.006.
26. Zhang Y, et al. Qwen3 embedding: advancing text embedding and reranking through foundation models. arXiv. Preprint posted online June 2025:arXiv:2506.05176.
27. Hurst A, et al. GPT-4o system card. arXiv. Preprint posted online October 2024:arXiv:2410.21276.
28. Grattafiori A, et al. The Llama 3 herd of models. arXiv. Preprint posted online July 2024:arXiv:2407.21783.
29. Kocaman V, Talby D. Biomedical named entity recognition at scale. In: Pattern Recognition. ICPR International Workshops and Challenges. Lecture Notes in Computer Science, vol 12661. Springer; 2021. doi:10.1007/978-3-030-68763-2_48.
30. Croxford E, et al. Evaluating clinical AI summaries with large language models as judges. NPJ Digit Med. 2025;8(1):640. doi:10.1038/s41746-025-02005-2.
31. Li D, Liang J, Li W, Wang X, Cao L, Yu K. CliCARE: grounding large language models in clinical guidelines for decision support over longitudinal cancer electronic health records. arXiv. Preprint posted online July 2025:arXiv:2507.22533.
32. Anthropic. Claude Sonnet 4.6 system card. Published online February 2026. Available from: <https://www.anthropic.com/claude-sonnet-4-6-system-card>.
33. Kocaman V, Talby D. Spark NLP: natural language understanding at scale. Softw Impacts. 2021;8:100058. doi:10.1016/j.simpa.2021.100058.
34. Zhang XYC, Ferreira CR, Rossignol F, Ng RT, Wasserman W, Zhu J. CaseReportBench: an LLM benchmark dataset for dense information extraction in clinical case reports. Proc Mach Learn Res. 2025;287:527–542.

Figures

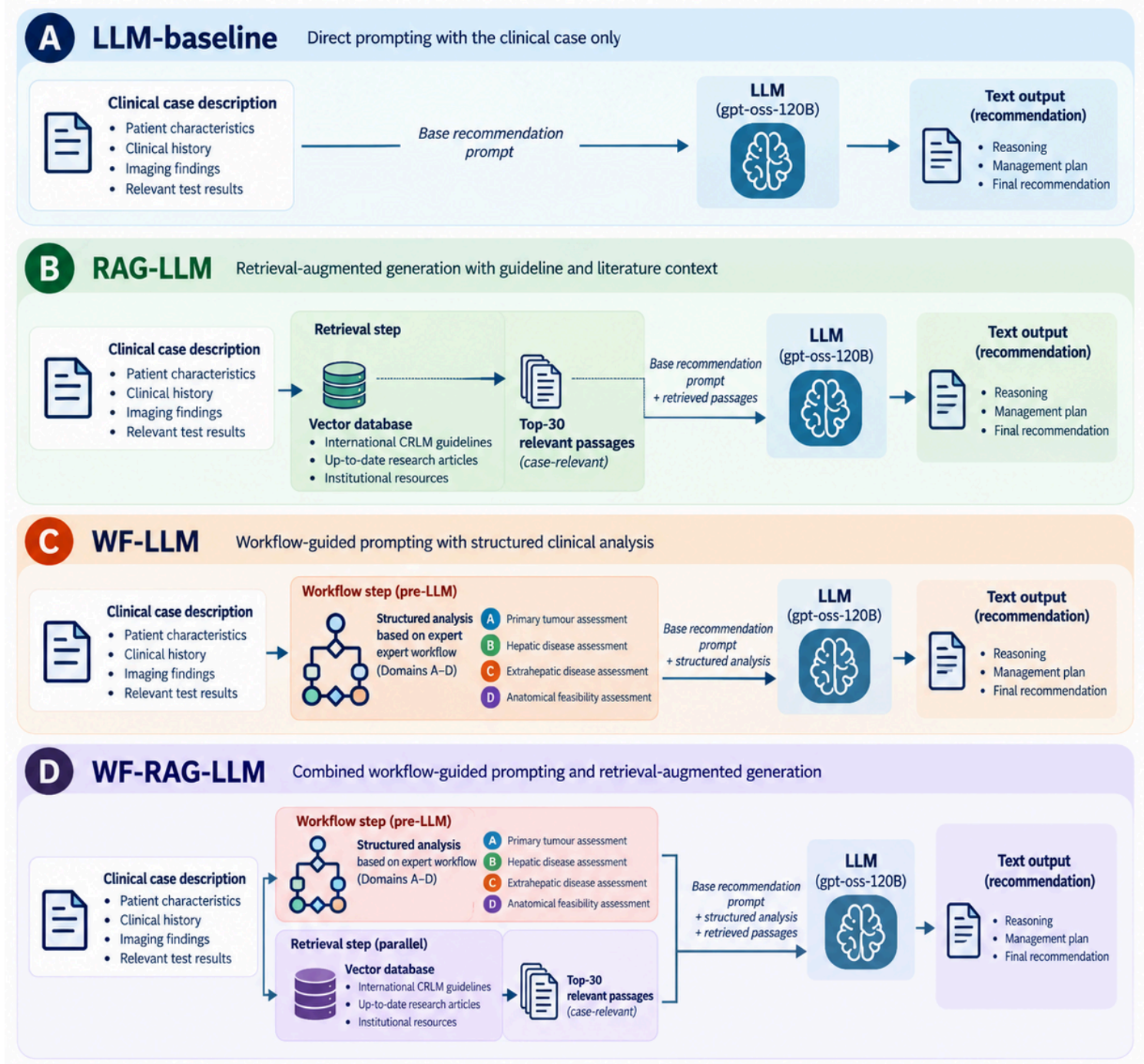


Figure 1

Overview of the four evaluated inference strategies. The figure highlights differences in the input provided to the base, shared prompt. LLM-baseline receives only the clinical case. RAG-LLM additionally receives case-relevant passages retrieved from a vector database indexing international CRLM guidelines and up-to-date research studies. WF-LLM additionally receives a structured analysis of the clinical case, derived from the expert-derived workflow through a separate inference step. WF-RAG-LLM receives both the retrieved passages and the workflow-derived structured analysis.

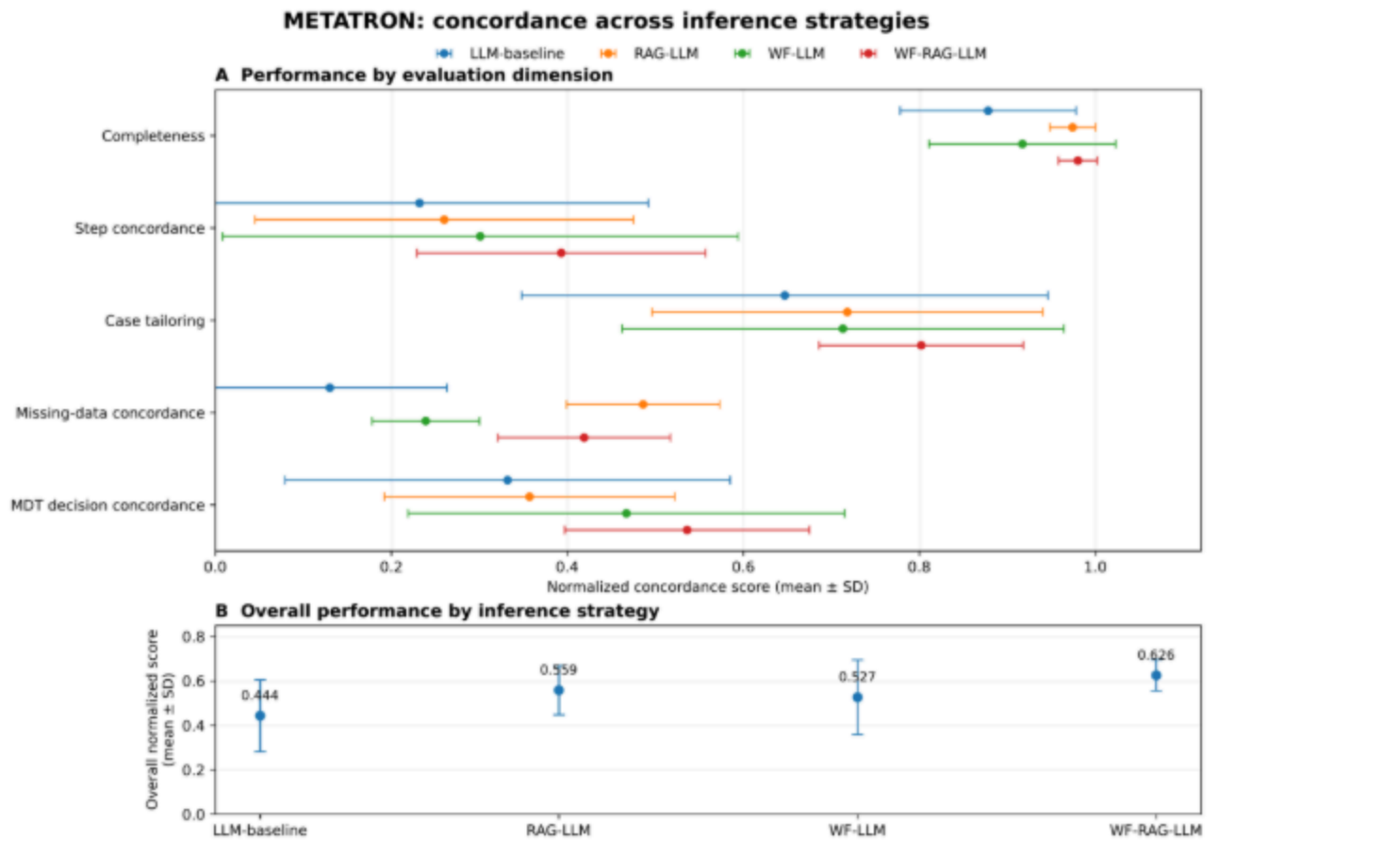


Figure 2

Descriptive overall composite score by inference strategy.(A) Normalized concordance scores (mean $\pm$ standard deviation) for each inference strategy across the five evaluation dimensions, aggregated across the three independent LLM judges. (B) Overall composite concordance score (mean $\pm$ standard deviation) for each inference strategy, aggregated across evaluation dimensions and judges.



Figure 3

Inference strategy performance by case scenario and evaluation dimension. The heatmap shows normalized scores assigned to each inference strategy and evaluation dimension, stratified by case scenario. Values represent the mean normalized scores across the LLM judges. Case-scenario abbreviations: *Anatomical Resectability*, anatomically challenging cases; *LM with Extra-Hepatic Disease*, liver metastases associated with extrahepatic disease; *Liver Recurrences After Surgery*, liver recurrence after previous hepatectomy; *Metachronous LM*, metachronous liver metastases; *Metastases After Chemotherapy*, post-chemotherapy reassessment; *Molecular Cases and Complex Situations*, molecularly complex situations; *Sync. LM Post-Chemo Resectable*, synchronous potentially resectable disease after systemic therapy; *Sync. Resectable LM*, synchronous resectable liver metastases.

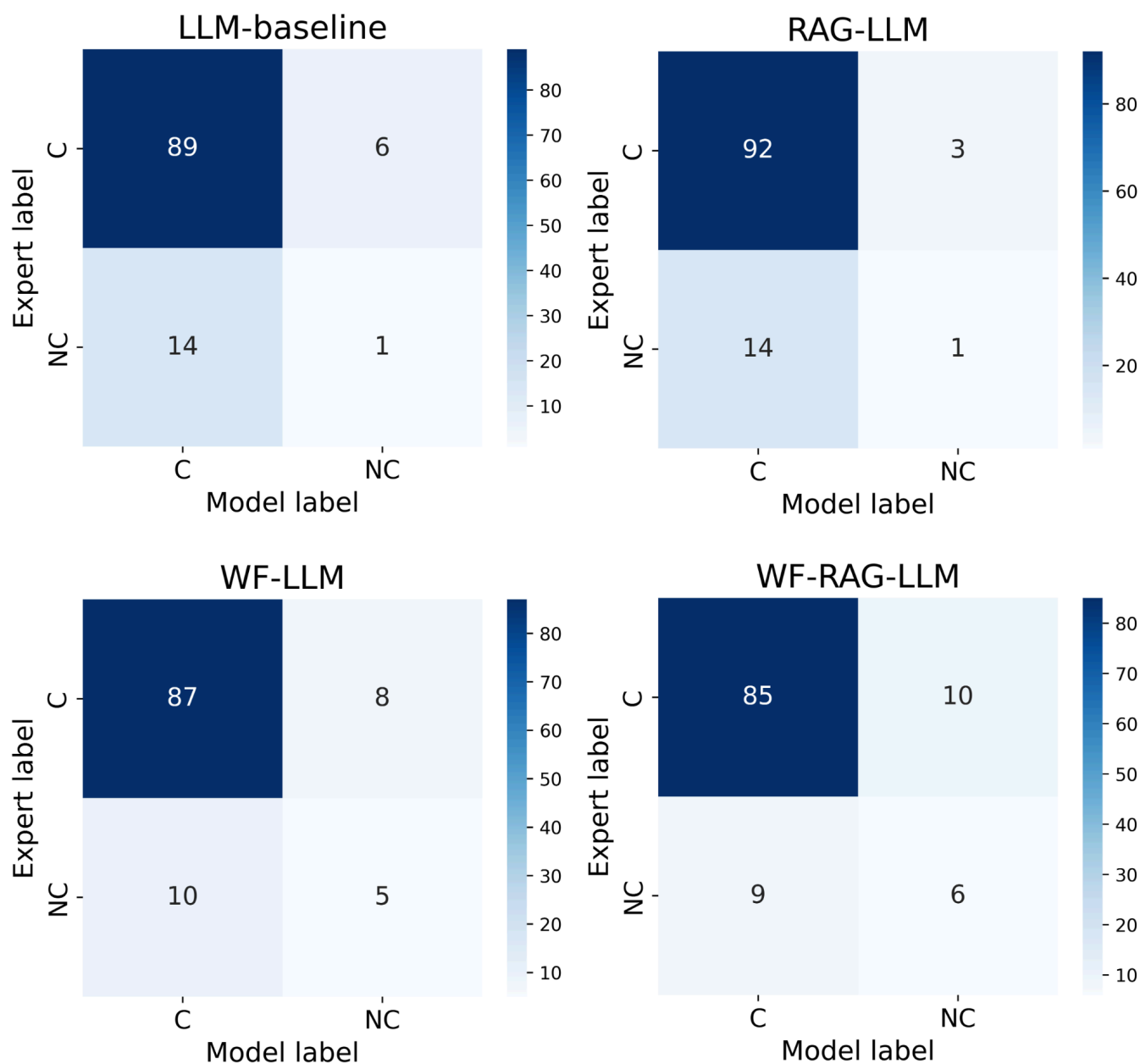


Figure 4

Confusion matrices for anatomical resectability classification. Each matrix shows the classification performance of one inference strategy, based on the anatomical resectability classification extracted from Step 3 of each recommendation, against the expert reference (C = anatomically compliant; NC = non-compliant).

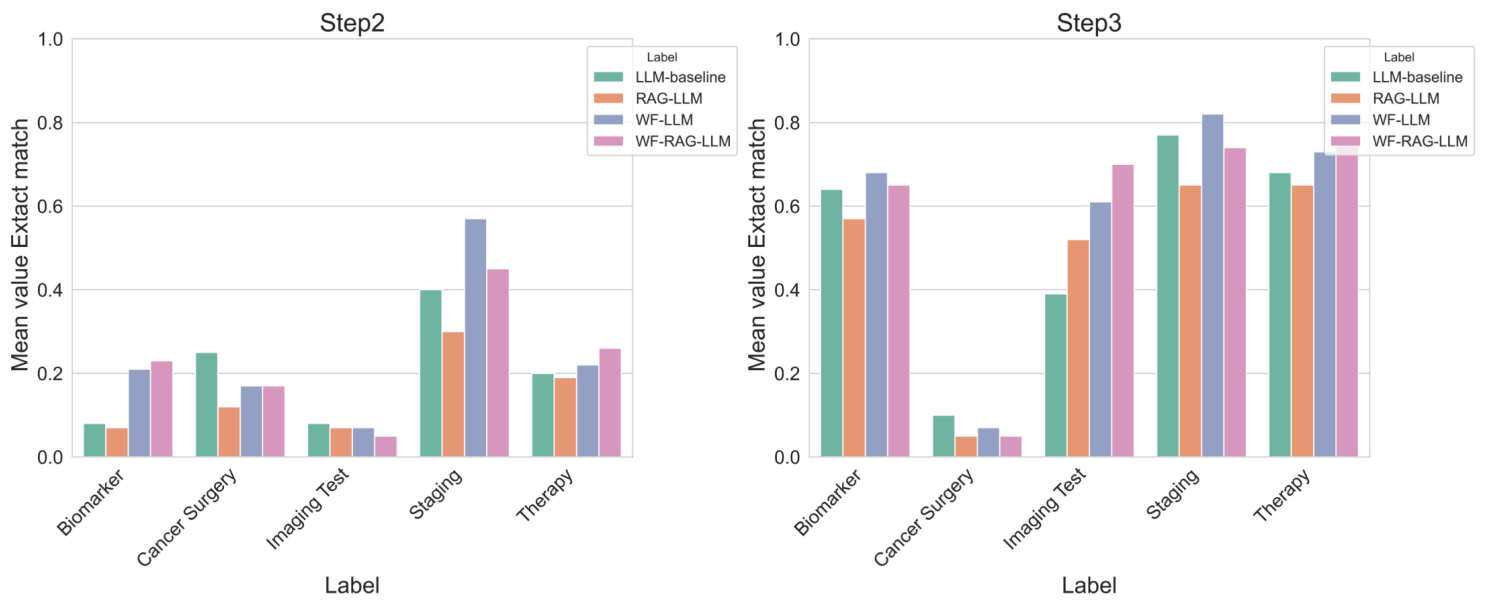


Figure 5

Exact match scores by entity label and inference strategy. Bars indicate mean exact match scores, averaged across cases, for each inference strategy, grouped by entity label (biomarker, cancer surgery, imaging test, staging, therapy). Entities were extracted using the NER-based pipeline, and exact match scores reflect agreement with entities extracted from the corresponding expert reference recommendations. Separate panels show results for Step 2 and Step 3 of the recommendations.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [METATRONSupplementarydef.pdf](#)