

Impact of structured clinical workflows and retrieval-augmented generation on Large Language Model recommendations for colorectal liver metastases: The METATRON Study

Supplementary materials

Summary of contents

Supplementary methods 2

 SM1 – Generation prompts 2

 SM2 - WF-LLM Worked Example 10

 SM3 – Evaluation prompts 13

Supplementary results..... 17

Supplementary methods

SM1 – Generation prompts

The following templates report the exact prompt structures used across the evaluated strategies, with curly-brace placeholders ({ }) indicating where dynamic inputs, such as the relevant JSON schema portion, the clinical case, or retrieved passages, are inserted at inference time.

1. Structured case analysis

System prompt

You are a hepatobiliary surgeon.

You are required to examine clinical cases involving patients diagnosed with a primary colorectal tumor and liver metastases, with the possibility of additional extrahepatic metastases.

User prompt (item node)

Analyze the clinical case and extract the required information.

Follow these rules:

- If the information is present, extract it exactly.
- If the information is not mentioned, write "Missing".
- If an item name contains the word "Presence", "Performed" or "Feasible", the value must be a boolean:
 - use True if the information is reported as present,
 - use False if it is reported as absent.
- Do not invent or infer information.
- Do not use null, None, or empty strings.
- At the end of the task, output only a valid JSON object.

Clinical case: {clinical_case}

Information to be extracted:

{json_input}

Here is an example of the expected output format:

```
{{
  "Tumor_Characteristic": {{
    "Items": [
      {{"Information 1": "Example"}},
      {{"Information 2": "Example"}},
      {{"Information 3": "Example"}}
    ]
  }}
}}
```

Now produce the JSON output based on the actual clinical case.

User prompt (condition node)

Analyze the clinical case and determine whether the condition stated in the question is true based on the case details.

Respond with:

- "Yes" if the condition is confirmed,
- "No" if it is explicitly denied,
- "Missing" if the information is not mentioned.

Do not infer, assume, or interpret beyond what is explicitly stated in the clinical case. Only use the information provided.

Clinical case: {clinical_case}

Question: {question}

Answer: "Yes" or "No" or "Missing".

2. Chunk retrieval

Based on this clinical case, determine whether the patient is resectable, potentially resectable, or unresectable according to hepatobiliary oncologic criteria, and identify the most appropriate oncological strategy or sequence of strategies, including surgery.

CLINICAL CASE:
{clinical_case}

3. Recommendation generation

System prompt (for LLM-baseline and RAG-LLM strategies)

You are an expert hepatobiliary surgeon participating in a Tumor Board (Multidisciplinary Team discussion - MDT).

You are expected to contribute your expertise to the evaluation and management of complex oncological cases, involving patients diagnosed with a primary colorectal tumor and liver metastases, with the possibility of additional extrahepatic metastases.

In your answer, do not use any Markdown formatting: no bold, no italics, no underscores, no headings, and no tables.

Use lists and bullet points to structure your answer.

System prompt (for WF-LLM and WF-RAG-LLM strategies)

You are an expert hepatobiliary surgeon, expected to contribute to the evaluation and management of complex oncological cases, involving patients diagnosed with a primary colorectal tumor and liver metastases, with the possibility of additional extrahepatic metastases.

You are collaborating with a colleague, who performs a preliminary, structured analysis of each case. The junior colleague presents the information in a step-by-step format, identifying and summarizing the key elements relevant for clinical decision-making.

The case evaluation is organized into four main sections:

1. Primary colorectal tumor

2. Liver metastases
3. Extrahepatic metastases
4. Anatomical feasibility for surgery.

Within each section, a predefined list of items is provided, reflecting the elements that should be assessed according to standard oncological and surgical reasoning.

When reading the structured analysis, treat the two labels as follows:

- "Missing": the information was expected but is not available. Do not assume a value. Flag it only if it is truly essential to the recommendation.
- "Assumed False" or "False": the condition is confirmed absent. Use this actively in your reasoning: a confirmed absence can justify a clinical decision just as a confirmed presence does (e.g., a confirmed absence of a specific condition can exclude competing clinical concerns and sharpen the prioritization of those that are present).

In your answer, do not use any Markdown formatting: no bold, no italics, no underscores, no headings, and no tables.

Use lists and bullet points to structure your answer.

User prompt (LLM-baseline)

Analyze the provided clinical case and formulate a recommendation regarding the next steps in patient management, including the patient's resectability assessment and surgical strategy.

RECOMMENDATION STRUCTURE

The recommendation must be structured in three steps:

Step 1 - Oncological strategy

Recommend the most appropriate oncological strategy or sequence of strategies, including surgery. Provide a detailed, step-by-step approach tailored to the specific clinical presentation.

Step 2 - Missing data

Determine whether any missing information prevents a reliable treatment recommendation.

Classify missing information into three categories:

- A. Patient Assessment: Information related to the patient's fitness for major liver surgery.
- B. Primary Tumor Status: Information related to the control and oncological status of the primary colorectal cancer (symptomatic or not).
- C. Liver Metastatic Disease Assessment: Information related to the extent, distribution, and technical resectability of liver metastases.

For each category:

- List only data that are truly required to make a treatment recommendation.
- Do not list information that would merely increase confidence.
- Assume that MDT decisions are routinely made despite incomplete datasets.

Procedure to identify only essential missing data (i.e., without which a reliable decision cannot be made): before listing missing information, answer the following question: "Can a reasonable MDT recommendation already be formulated based on the available information?" Only if the answer is NO, specify which items prevent a reliable recommendation.

Step 3 - Anatomic resectability

Determine whether any anatomical barrier to surgical resection can be identified from the available information and classify the case into one of two categories:

- Compliant: No anatomical barrier is identified, or any concern is solely due to missing data rather than a confirmed barrier.
- Non-compliant: A definite anatomical barrier precludes safe resection.

If the case is classified as Compliant but critical anatomical data are missing, respond with: "Compliant, provided that: [list the missing data required to confirm feasibility]".

Do not classify a case as Non-compliant solely on the basis of missing information.

Important rules

- Do not assume anatomical non-compliance solely because specific anatomical data are unavailable.
- If no anatomical barrier is described, assume anatomical compliance unless available information strongly suggests otherwise.
- Missing volumetry, vascular mapping, biliary mapping, or liver function assessment should not automatically result in anatomical non-compliance.

GENERAL INDICATIONS:

- Focus on the hepatobiliary surgical perspective.
- Use the labelled structure above exactly (Step 1, Step 2, Step 3). Clearly separate each step by following the structure provided, using the same step headers.
- Provide a clear statement or sequence of statements for each step. Be complete but non-redundant.

CLINICAL CASE:

{clinical_case}

YOUR RECOMMENDATION:

Step 1 -

Step 2 – Essential missing data for each category:

- A) Patient Assessment: [list of essential missing information];
- B) Primary Tumor Status: [list of essential missing information];
- C) Liver Metastatic Disease Assessment: [list of essential missing information]

Step 3 - [Anatomically compliant / Anatomically non-compliant / Anatomically compliant provided that <list the missing data required to confirm feasibility>]: brief rationale

User prompt (RAG-LLM)

Analyze the provided clinical case and formulate a recommendation regarding the next steps in patient management, including the patient's resectability assessment and surgical strategy.

RECOMMENDATION STRUCTURE

The recommendation must be structured in three steps:

Step 1 - Oncological strategy

Recommend the most appropriate oncological strategy or sequence of strategies, including surgery. Provide a detailed, step-by-step approach tailored to the specific clinical presentation.

Step 2 - Missing data

Determine whether any missing information prevents a reliable treatment recommendation.

Classify missing information into three categories:

- A. Patient Assessment: Information related to the patient's fitness for major liver surgery.
- B. Primary Tumor Status: Information related to the control and oncological status of the primary colorectal cancer (symptomatic or not).
- C. Liver Metastatic Disease Assessment: Information related to the extent, distribution, and technical resectability of liver metastases.

For each category:

- List only data that are truly required to make a treatment recommendation.
- Do not list information that would merely increase confidence.
- Assume that MDT decisions are routinely made despite incomplete datasets.

Procedure to identify only essential missing data (i.e., without which a reliable decision cannot be made): before listing missing information, answer the following question: "Can a reasonable MDT recommendation already be formulated based on the available information?" Only if the answer is NO, specify which items prevent a reliable recommendation.

Step 3 - Anatomic resectability

Determine whether any anatomical barrier to surgical resection can be identified from the available information and classify the case into one of two categories:

- Compliant: No anatomical barrier is identified, or any concern is solely due to missing data rather than a confirmed barrier.
- Non-compliant: A definite anatomical barrier precludes safe resection.

If the case is classified as Compliant but critical anatomical data are missing, respond with: "Compliant, provided that: [list the missing data required to confirm feasibility]".

Do not classify a case as Non-compliant solely on the basis of missing information.

Important rules

- Do not assume anatomical non-compliance solely because specific anatomical data are unavailable.
- If no anatomical barrier is described, assume anatomical compliance unless available information strongly suggests otherwise.
- Missing volumetry, vascular mapping, biliary mapping, or liver function assessment should not automatically result in anatomical non-compliance.

GENERAL INDICATIONS:

- Focus on the hepatobiliary surgical perspective.
- Use the labelled structure above exactly (Step 1, Step 2, Step 3). Clearly separate each step by following the structure provided, using the same step headers.
- Provide a clear statement or sequence of statements for each step. Be complete but non-redundant.

CLINICAL CASE:

{clinical_case}

SUPPORTING EVIDENCE

The following literature excerpts have been retrieved to support your recommendation. Where directly relevant, integrate them to justify your clinical reasoning.

{refs}

YOUR RECOMMENDATION:

Step 1 -

Step 2 – Essential missing data for each category:

- A) Patient Assessment: [list of essential missing information];
- B) Primary Tumor Status: [list of essential missing information];
- C) Liver Metastatic Disease Assessment: [list of essential missing information]

Step 3 - [Anatomically compliant / Anatomically non-compliant / Anatomically compliant provided that <list the missing data required to confirm feasibility>]: brief rationale

User prompt (WF-LLM)

Analyze the provided clinical case and formulate a recommendation regarding the next steps in patient management, including the patient's resectability assessment and surgical strategy.

RECOMMENDATION STRUCTURE

The recommendation must be structured in three steps:

Step 1 - Oncological strategy

Recommend the most appropriate oncological strategy or sequence of strategies, including surgery. Provide a detailed, step-by-step approach tailored to the specific clinical presentation.

Step 2 - Missing data

Determine whether any missing information prevents a reliable treatment recommendation.

Classify missing information into three categories:

- A. Patient Assessment: Information related to the patient's fitness for major liver surgery.
- B. Primary Tumor Status: Information related to the control and oncological status of the primary colorectal cancer (symptomatic or not).
- C. Liver Metastatic Disease Assessment: Information related to the extent, distribution, and technical resectability of liver metastases.

For each category:

- List only data that are truly required to make a treatment recommendation.
- Do not list information that would merely increase confidence.
- Assume that MDT decisions are routinely made despite incomplete datasets.

Procedure to identify only essential missing data (i.e., without which a reliable decision cannot be made): before listing missing information, answer the following question: "Can a reasonable MDT recommendation already be formulated based on the available information?" Only if the answer is NO, specify which items prevent a reliable recommendation.

Step 3 - Anatomic resectability

Determine whether any anatomical barrier to surgical resection can be identified from the available information and classify the case into one of two categories:

- Compliant: No anatomical barrier is identified, or any concern is solely due to missing data rather than a confirmed barrier.
 - Non-compliant: A definite anatomical barrier precludes safe resection.
- If the case is classified as Compliant but critical anatomical data are missing, respond with: "Compliant, provided that: [list the missing data required to confirm feasibility]".
- Do not classify a case as Non-compliant solely on the basis of missing information.

Important rules

- Do not assume anatomical non-compliance solely because specific anatomical data are unavailable.
- If no anatomical barrier is described, assume anatomical compliance unless available information strongly suggests otherwise.
- Missing volumetry, vascular mapping, biliary mapping, or liver function assessment should not automatically result in anatomical non-compliance.

GENERAL INDICATIONS:

- Focus on the hepatobiliary surgical perspective.
- Use the labelled structure above exactly (Step 1, Step 2, Step 3). Clearly separate each step by following the structure provided, using the same step headers.
- Provide a clear statement or sequence of statements for each step. Be complete but non-redundant.

CLINICAL CASE:

{clinical_case}

STRUCTURED ANALYSIS OF THE CLINICAL CASE

The following structured analysis was produced by a colleague. Follow the sequential order of its four sections (Primary Tumor → Liver Metastases → Extrahepatic Disease → Anatomical Feasibility) when building your reasoning across the three steps.

For each step of your recommendation, identify which features from the analysis are decision-relevant and explain how they drive your clinical reasoning.

{str_analysis}

Notes about the structured analysis:

When the structured analysis explicitly indicates that FLR assessment is not required (FLR trigger condition marked as False), do not list missing volumetry or liver function data as conditions for anatomical compliance. In such cases, anatomical compliance is determined solely by vascular involvement, biliary drainage, and surgical margin achievability.

YOUR RECOMMENDATION:

Step 1 -

Step 2 – Essential missing data for each category:

- A) Patient Assessment: [list of essential missing information];
- B) Primary Tumor Status: [list of essential missing information];
- C) Liver Metastatic Disease Assessment: [list of essential missing information]

Step 3 - [Anatomically compliant / Anatomically non-compliant / Anatomically compliant provided that <list the missing data required to confirm feasibility>]: brief rationale

User prompt (WF-RAG-LLM)

Analyze the provided clinical case and formulate a recommendation regarding the next steps in patient management, including the patient's resectability assessment and surgical strategy.

RECOMMENDATION STRUCTURE

The recommendation must be structured in three steps:

Step 1 - Oncological strategy

Recommend the most appropriate oncological strategy or sequence of strategies, including surgery. Provide a detailed, step-by-step approach tailored to the specific clinical presentation.

Step 2 - Missing data

Determine whether any missing information prevents a reliable treatment recommendation.

Classify missing information into three categories:

- A. Patient Assessment: Information related to the patient's fitness for major liver surgery.
- B. Primary Tumor Status: Information related to the control and oncological status of the primary colorectal cancer (symptomatic or not).
- C. Liver Metastatic Disease Assessment: Information related to the extent, distribution, and technical resectability of liver metastases.
- D. Extrahepatic disease: Information related to the extent, distribution, and technical resectability of extrahepatic metastases.

For each category:

- List only data that are truly required to make a treatment recommendation.
- Do not list information that would merely increase confidence.
- Assume that MDT decisions are routinely made despite incomplete datasets.

Procedure to identify only essential missing data (i.e., without which a reliable decision cannot be made): before listing missing information, answer the following question: "Can a reasonable MDT recommendation already be formulated based on the available information?" Only if the answer is NO, specify which items prevent a reliable recommendation.

Step 3 - Anatomic resectability

Determine whether any anatomical barrier to surgical resection can be identified from the available information and classify the case into one of two categories:

- Compliant: No anatomical barrier is identified, or any concern is solely due to missing data rather than a confirmed barrier.
- Non-compliant: A definite anatomical barrier precludes safe resection.

If the case is classified as Compliant but critical anatomical data are missing, respond with: "Compliant, provided that: [list the missing data required to confirm feasibility]".

Do not classify a case as Non-compliant solely on the basis of missing information.

Important rules

- Do not assume anatomical non-compliance solely because specific anatomical data are unavailable.

- If no anatomical barrier is described, assume anatomical compliance unless available information strongly suggests otherwise.
- Missing volumetry, vascular mapping, biliary mapping, or liver function assessment should not automatically result in anatomical non-compliance.

GENERAL INDICATIONS:

- Focus on the hepatobiliary surgical perspective.
- Use the labelled structure above exactly (Step 1, Step 2, Step 3). Clearly separate each step by following the structure provided, using the same step headers.
- Provide a clear statement or sequence of statements for each step. Be complete but non-redundant.

CLINICAL CASE:

{clinical_case}

SUPPORTING EVIDENCE

The following literature excerpts have been retrieved to support your recommendation. Where directly relevant, integrate them to justify your clinical reasoning.

{refs}

STRUCTURED ANALYSIS OF THE CLINICAL CASE

The following structured analysis was produced by a colleague. Follow the sequential order of its four sections (Primary Tumor → Liver Metastases → Extrahepatic Disease → Anatomical Feasibility) when building your reasoning across the three steps.

For each step of your recommendation, identify which features from the analysis are decision-relevant and explain how they drive your clinical reasoning.

{str_analysis}

Notes about the structured analysis:

When the structured analysis explicitly indicates that FLR assessment is not required (FLR trigger condition marked as False), do not list missing volumetry or liver function data as conditions for anatomical compliance. In such cases, anatomical compliance is determined solely by vascular involvement, biliary drainage, and surgical margin achievability.

YOUR RECOMMENDATION:

Step 1 -

Step 2 – Essential missing data for each category:

- A) Patient Assessment: [list of essential missing information];
- B) Primary Tumor Status: [list of essential missing information];
- C) Liver Metastatic Disease Assessment: [list of essential missing information]

Step 3 - [Anatomically compliant / Anatomically non-compliant / Anatomically compliant provided that <list the missing data required to confirm feasibility>]: brief rationale

SM2 - WF-LLM Worked Example

Case description

Man, 64 years old, no significant medical history, presenting with a left colon tumor revealed by subocclusive syndrome. Thoraco-abdominal CT scan shows a 6 cm colonic mass and two liver metastases in segments II and V measuring 2.2 and 3 cm. CEA: 120 ng/mL. RAS wild-type, MSS. No extra-hepatic disease.

Structured analysis (generated in Phase 1 within WF-LLM inference strategy)

Information concerning Colorectal_Primary_Tumor_Characterization

- Location: left colon
- Grade_Of_Differentiation: Missing
- Molecular_Phenotype: RAS wild-type, MSS

Information concerning Colorectal_Primary_Tumor_Post_Resection_Characterization

- Has the colorectal primary tumor been already resected? False

Information concerning Colorectal_Primary_Tumor_Symptomatology

- Presence_Of_Symptoms_For_Obstruction: True
- Presence_Of_Symptoms_For_Perforation: False
- Presence_Of_Symptoms_For_Bleeding: False

Information concerning Colorectal_Primary_Tumor_CEA_Level

- CEA_Level: 120 ng/mL

Information concerning Liver_Metastases_Timing

Description: Liver metastases are defined as "Synchronous" if detected before or at the same time as the primary colorectal tumor; they are considered "Metachronous" if diagnosed at least one month after the colorectal primary tumor.

- Timing: Synchronous

Information concerning Liver_Metastases_Characterization

- Number_Of_Metastases: 2
- Maximum_Lesion_Diameter: 3 cm
- Minimum_Lesion_Diameter: 2.2 cm
- Hepatic_Segments_Involved: II, V
- Presence_Of_Vascular_Contact_Or_Invasion: Missing
- Presence_Of_Biliary_Structures_Involved: Missing
- Presence_Of_Necrosis: Missing
- Presence_Of_Fibrosis: Missing
- Presence_Of_Associated_Calcifications: Missing
- Presence_Of_Associated_Cystic_Changes: Missing
- Presence_Of_Associated_Portal_Thrombosis: Missing
- Arterial_Anatomy: Missing
- Presence_Of_Atypical_Features: Missing

Information concerning ImagingExams

- Thoraco_CT_Scan_Performed: True
- Abdomen_CT_Scan_Performed: True
- Has liver MRI been performed? Missing

Information concerning Liver_Metastases_Under_Systemic_Treatment

Description: Post chemotherapy characterization of the liver metastases.

- Has a systemic treatment (chemotherapy) already been performed? Missing

Information concerning Liver_Metastases_Occurrence

- Does the current liver metastases represent a recurrence — meaning, were there previous liver metastases that had been treated surgically or with microwave or radiofrequency ablation?

Missing

Information concerning Extrahepatic_Pulmonary_Involvement

- Do imaging studies show evidence of metastatic extrahepatic pulmonary involvement without discussion of differential diagnosis? False

Information concerning Extrahepatic_Peritoneal_Involvement

- Do imaging studies show evidence of metastatic extrahepatic peritoneal involvement without discussion of differential diagnosis? False

Information concerning Extrahepatic_Lymph_Node_Involvement

- Do imaging studies show evidence of metastatic extrahepatic lymph node involvement without discussion of differential diagnosis? False

Information concerning Extrahepatic_Bone_Involvement

Description: Metastatic bone lesions description.

- Do imaging studies show evidence of metastatic extrahepatic bone involvement without discussion of differential diagnosis? False

Information concerning Extrahepatic_Brain_Involvement

Description: Metastatic brain lesions description.

- Do imaging studies show evidence of metastatic extrahepatic brain involvement without discussion of differential diagnosis? False

Information concerning Future_Liver_Remnant_FLR_Ratio

Description: FLR-Ratio (FLR/total liver volume) is compatible with surgery (liver metastases resection) when >25% in healthy livers; >30% in chemotherapy-damaged livers; >40% in cholestatic livers, and >50% in cirrhotic livers. FLR considerations become mandatory only when available information suggests one of the following conditions: major hepatectomy, extended hepatectomy, bilobar disease requiring staged procedures, vascular reconstruction, significant chemotherapy-associated liver injury, cirrhosis, cholestasis, or marginal liver reserve.

- Does the case require management under any of the following conditions: major hepatectomy, extended hepatectomy, bilobar disease requiring staged procedures, vascular reconstruction, significant chemotherapy-associated liver injury, cirrhosis, cholestasis, or marginal liver reserve? False

Information concerning Vascular_Involvement

Description: Surgical feasibility requires preserved portal inflow and hepatic venous outflow of future liver remnant (FLR), with vascular involvement amenable to unilateral resection/reconstruction. Surgery is contraindicated when resection of all three hepatic veins, both portal vein branches, and/or the retrohepatic inferior vena cava is required.

- Vascular_Involvement_Compatibility_Surgery: Missing

Information concerning Biliary_Drainage

Description: Surgical feasibility requires preserved biliary drainage of the future liver remnant (FLR). Impaired biliary drainage of the FLR contraindicates surgery.

- Biliary_Drainage_Compatibility_Surgery: Missing

Information concerning R0_Surgical_Margins

Description: Surgical feasibility requires the possibility of achieving R0 resection or R1 vascular resection margins. R1 parenchymal margins and R2 resection contraindicate surgery.

- Surgical_Margins_Compatibility_Surgery: True

SM3 – Evaluation prompts

The following template reports the exact prompt structure used for the LLM-as-a-jury evaluation, with curly-brace placeholders ({ }) indicating where dynamic inputs, such as the clinical vignette, the expert reference recommendation, and the candidate recommendation, are inserted at inference time.

System prompt

You are an expert hepatobiliary surgeon participating in a multidisciplinary tumor board (MDT) for colorectal liver metastases (CRLM).

You will be provided with:

1. A clinical case
2. Recommendation 1 (Expert Reference Recommendation)
3. Recommendation 2 (Candidate Recommendation)

Your task is to determine how closely Recommendation 2 reproduces the clinical reasoning, treatment strategy, and decision-making process of Recommendation 1.

Recommendation 1 should be considered the expert reference recommendation.

Recommendation 2 should be evaluated against Recommendation 1.

GENERAL PRINCIPLES

1. Focus on decisions, not wording. Ignore differences in: writing style, sentence structure, terminology, wording, formatting, recommendation length, level of detail. Evaluate only the underlying clinical reasoning and management decisions.

2. Clinical equivalence. Two recommendations should be considered concordant when they would lead to the same decisions, even if they differ in wording, sequencing, or level of detail. The goal is to evaluate decision concordance rather than textual similarity.

3. Additional information. Do not penalize Recommendation 2 for including additional clinically relevant information if:

- it is consistent with Recommendation 1
- it does not contradict Recommendation 1
- it does not alter the management strategy

Additional clinically relevant details should be considered neutral or beneficial.

4. Missing information. Do not penalize Recommendation 2 for omitting information that is not essential to clinical decision-making. Focus only on decision-critical information.

5. Clinical enrichment. When Recommendation 2 provides clinically appropriate additional reasoning or details that are absent from Recommendation 1, these additions should not decrease concordance scores. The objective is to evaluate agreement with the expert recommendation, not to penalize clinically valid enrichment.

6. Clinical case context. Use the clinical case when evaluating the recommendations. The clinical case should be used to determine:

- whether proposed management strategies are addressing the same clinical problem
- whether identified missing data are truly decision-critical
- whether additional information is clinically relevant.

User prompt

You are given two recommendations. Your task is to evaluate their similarity by scoring each of the following criteria: completeness, step-by-step concordance, case-tailoring concordance, and missing data concordance. To perform your evaluation, follow the sequence of actions indicated below.

Action 1 - PRE-SCORING DECISION EXTRACTION

Before assigning any score, identify the key clinical decisions contained in both recommendations. For each recommendation, summarize the following elements:

- Overall oncological strategy
- Systemic therapy strategy
- Surgical strategy
- Timing of surgery
- Key staging/restaging investigations
- Anatomical compliance assessment
- Critical missing data
- Final MDT decision

Important: The purpose of this step is to identify the underlying clinical decisions. Thus,

- Do not evaluate wording, style, length, formatting, or level of detail.
- Only extract the clinical meaning of each recommendation.

The subsequent scoring should be based primarily on the extracted decisions rather than on the original text.

Action 2 - EVALUATION CRITERIA

Criterion 1 — Clinical Coverage

Evaluate whether Recommendation 2 addresses the same major clinical domains as Recommendation 1.

Examples include:

- resectability assessment
- systemic therapy
- imaging and staging
- surgical management
- postoperative management
- surveillance strategy

Scoring:

- 5 = Same clinically relevant domains covered
- 4 = Minor omissions with no impact on management
- 3 = Some relevant domains missing
- 2 = Major clinical domains missing
- 1 = Multiple critical domains absent

Criterion 2 — Step-By-Step Concordance

Evaluate whether Recommendation 2 reproduces the same major clinical decisions and management sequence as Recommendation 1.

Focus on:

Step 1

- overall treatment strategy
- timing of surgery
- systemic therapy decisions
- staging/restaging decisions
- treatment sequencing

Step 3: anatomical compliance assessment

Scoring:

- 5 = Same management strategy and same resectability conclusions
- 4 = Same overall strategy with minor differences
- 3 = Partially overlapping strategy
- 2 = Important strategic disagreements
- 1 = Opposite management strategy or opposite resectability assessment

Minor differences in wording, sequencing, or level of detail should not reduce the score.

Criterion 3 — Case-Tailoring Concordance

Evaluate whether Recommendation 2 incorporates the same case-specific clinical factors used by Recommendation 1. Focus on whether similar patient-specific reasoning is applied.

Examples include:

- molecular profile
- tumor biology
- chemotherapy response
- lesion distribution
- vascular involvement
- extrahepatic disease
- patient-related factors
- timing of metastatic presentation

Scoring:

- 5 = Same case-specific reasoning
- 4 = Minor differences
- 3 = Partial overlap
- 2 = Important patient-specific factors ignored
- 1 = Major case-specific reasoning absent

Criterion 4 — Missing Data Concordance

Determine whether Recommendation 2 identifies the same decision-critical missing information as Recommendation 1.

Instructions:

- Evaluate Step 2 only.
- Assess the overlap in missing data identified across the different categories (A, B, C, D).
- Focus exclusively on information that could realistically alter treatment decisions.
- Do not penalize Recommendation 2 for omitting desirable but non-essential information.

Scoring:

5 = Same critical missing data identified

4 = Minor differences

3 = Partial overlap

2 = Important discrepancies

1 = Completely different missing-data assessment

Criterion 5 — MDT DECISION CONCORDANCE

Evaluate whether Recommendation 2 would ultimately lead to the same MDT decision as Recommendation 1.

Focus on the final treatment recommendation rather than individual details.

Examples of high concordance:

- upfront surgery vs upfront surgery
- conversion chemotherapy vs conversion chemotherapy
- two-stage hepatectomy vs two-stage hepatectomy
- surgery contraindicated vs surgery contraindicated

Examples of low concordance:

- upfront surgery vs induction chemotherapy
- resectable vs unresectable
- surgery vs no surgery

Consider the final treatment pathway that would be proposed to the patient.

Scoring:

5 = Same MDT decision

- Same overall treatment pathway
- Same resectability conclusion
- Differences limited to details or supporting arguments

4 = Highly similar MDT decision

- Minor differences in sequencing
- Final treatment recommendation remains essentially unchanged

3 = Partially similar MDT decision

- Similar therapeutic goals
- Important implementation differences

2 = Different MDT decision

- Significant management differences
- Different treatment pathway likely

1 = Opposite MDT decision

- Opposite conclusions regarding surgery, resectability, or treatment intent
- Fundamentally different patient management

Action 3 - EVALUATION PROCESS

Using the clinical decisions extracted in Action 1, compare the two recommendations against each criterion, applying the definitions provided for each score value in Action 2. Support your scores with a brief reasoning for each criterion, based on the clinical decisions identified in Action 1 and the content of the two recommendations.

Important: Scores should reflect agreement in clinical reasoning and management decisions rather than textual similarity.

OUTPUT FORMAT

```
{{
  "completeness": <integer from 1 to 5>,
  "step_concordance": <integer from 1 to 5>,
  "case_tailoring": <integer from 1 to 5>,
  "missing_data_concordance": <integer from 1 to 5>,
  "mdt_decision_concordance": <integer from 1 to 5>,
  "reasoning":""
}}
```

CLINICAL CASE

{case}

RECOMMENDATION 1 (EXPERT REFERENCE)

{reference}

RECOMMENDATION 2 (CANDIDATE)

{candidate}

Supplementary results

Supplementary Table 1. Friedman test results. For each judge and evaluation dimension, p-values indicate whether ordinal scores differed significantly across the four inference strategies.

Judge	Evaluation dimension	Friedman test p-value
Claude Sonnet 4.6	Completeness	<<0.05
Claude Sonnet 4.6	Step concordance	0.0018
Claude Sonnet 4.6	Case tailoring concordance	<<0.05
Claude Sonnet 4.6	Missing data concordance	0.0005
Claude Sonnet 4.6	MDT decision concordance	0.0004
GPT-4o	Completeness	0.2252
GPT-4o	Step concordance	0.0109
GPT-4o	Case tailoring concordance	0.0872
GPT-4o	Missing data concordance	0.0227
GPT-4o	MDT decision concordance	0.0166
Llama 3.3 70B	Completeness	0.9058
Llama 3.3 70B	Step concordance	0.6153
Llama 3.3 70B	Case tailoring concordance	0.8997
Llama 3.3 70B	Missing data concordance	0.0383
Llama 3.3 70B	MDT decision concordance	0.4409

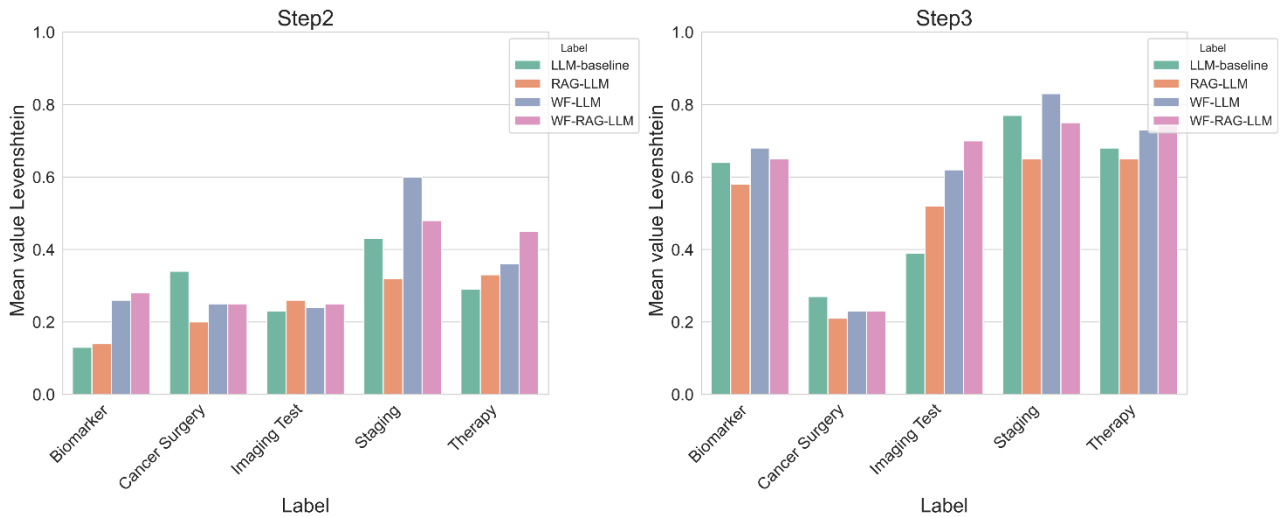
Supplementary Table 2. Statistically significant pairwise comparisons between inference strategies. For each judge and evaluation dimension, only pairwise comparisons reaching statistical significance (Wilcoxon signed-rank test, Holm-corrected $p < 0.05$) are reported, together with the corresponding strategy pair, their mean ordinal scores, the score difference, the original and corrected p-value.

Judge	Evaluation dimension	Comparison (strategy 1 vs 2)	Mean score strategy 1	Mean score strategy 2	Diff. among mean values (1 – 2)	p-value (raw)	p-value (after Holm correction)
Claude Sonnet 4.6	Completeness	LLM-baseline vs RAG-LLM	4	5	-0.291	1.924e-05	7.698e-05
Claude Sonnet 4.6	Completeness	LLM-baseline vs WF-RAG-LLM	4	5	-0.309	9.366e-07	5.619e-06
Claude Sonnet 4.6	Completeness	RAG-LLM vs WF-LLM	5	4	0.255	0.0001	0.0004
Claude Sonnet 4.6	Completeness	WF-LLM vs WF-RAG-LLM	4	5	-0.273	1.490e-05	7.451e-05
Claude Sonnet 4.6	Step concordance	LLM-baseline vs WF-RAG-LLM	4	4	-0.291	6.356e-05	0.0004
Claude Sonnet 4.6	Step concordance	WF-LLM vs WF-RAG-LLM	4	4	-0.218	0.0090	0.0450
Claude Sonnet 4.6	Case tailoring	LLM-baseline vs RAG-LLM	4	4	-0.255	0.0011	0.0042
Claude Sonnet 4.6	Case tailoring	LLM-baseline vs WF-RAG-LLM	4	4	-0.473	1.668e-09	1.001e-08
Claude Sonnet 4.6	Case tailoring	RAG-LLM vs WF-RAG-LLM	4	4	-0.218	0.0042	0.0125

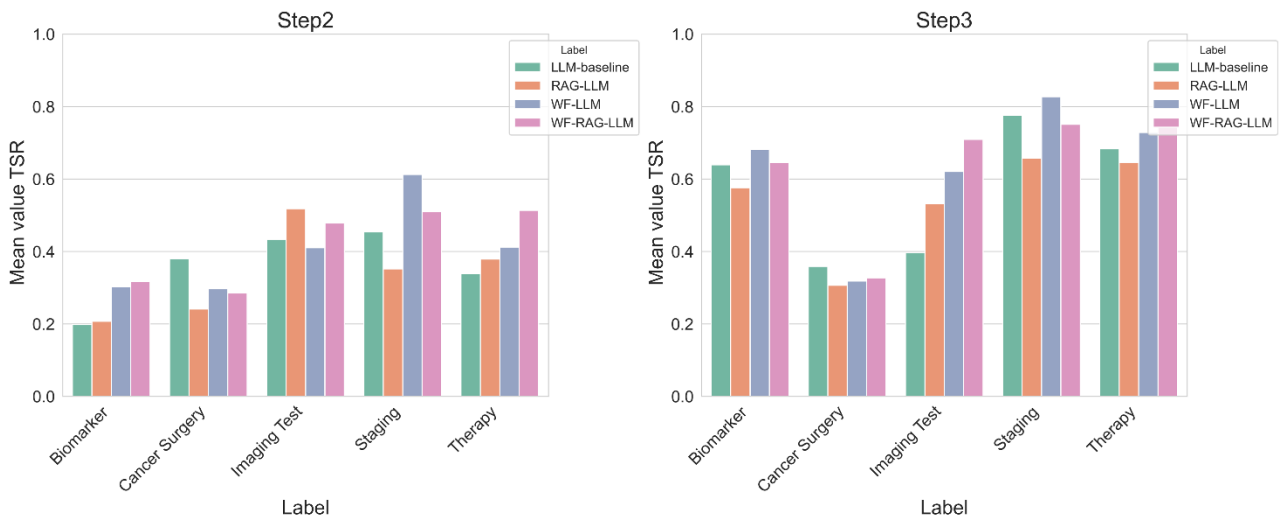
Claude Sonnet 4.6	Case tailoring	WF-LLM vs WF-RAG-LLM	4	4	-0.3	0.0004	0.0022
Claude Sonnet 4.6	Missing data concordance	LLM-baseline vs RAG-LLM	4	4	-0.355	0.0012	0.0070
Claude Sonnet 4.6	Missing data concordance	RAG-LLM vs WF-LLM	4	4	0.255	0.0018	0.0091
Claude Sonnet 4.6	MDT decision concordance	LLM-baseline vs WF-RAG-LLM	4	4	-0.4	1.095e-05	6.575e-05
Claude Sonnet 4.6	MDT decision concordance	RAG-LLM vs WF-RAG-LLM	4	4	-0.255	0.0015	0.0074
Claude Sonnet 4.6	MDT decision concordance	WF-LLM vs WF-RAG-LLM	4	4	-0.236	0.0085	0.0338
GPT-4o	Step concordance	LLM-baseline vs WF-RAG-LLM	4	4	-0.191	0.0073	0.043
GPT-4o	Missing data concordance	RAG-LLM vs WF-LLM	5	4	0.245	0.0071	0.0427
GPT-4o	MDT decision concordance	LLM-baseline vs WF-RAG-LLM	4	5	-0.236	0.0034	0.0205
Llama 3.3 70B	Missing data concordance	LLM-baseline vs RAG-LLM	4	5	-0.273	0.0062	0.0369



Supplementary Figure 1. Schematic overview of the expert-derived clinical workflow. The workflow is organized into four main decision domains (Steps). Within each domain, major decision nodes (item nodes) expand hierarchically into clinically relevant sub-items (leaf nodes) and conditional branches (condition nodes).



Supplementary Figure 2. Levenshtein similarity by entity label and inference strategy. Bars indicate mean Levenshtein similarities, 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 Levenshtein similarity 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 Figure 3. Token set ratio by entity label and inference strategy. Bars indicate mean token set ratios (TSR), 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 token set ratios reflect agreement with entities extracted from the corresponding expert reference recommendations. Separate panels show results for Step 2 and Step 3 of the recommendations.