MammoClaw: Towards Skill-Evolving Agent Harness for Breast Cancer Mammography Analysis

Supplementary Material

Krishna Kanth Nakka

Munich, Bavaria, Germany
krishkanth.92@gmail.com

A Limitations and Future Work

`MammoClaw` is an initial exploratory study of skill-evolving agent frameworks for mammography. Our evaluation is limited to two assessment tasks, BI-RADS prediction and breast density estimation, using a single frozen MLLM backbone and one benchmark dataset per task. In addition, Qwen3.5-35B-A3B shows modest performance in the no-tools setting, which may limit the extent to which the benefits of the agent harness can be assessed. Future evaluations with stronger mammography-capable backbones, additional datasets, and broader tasks, such as abnormality classification, are needed to better characterize the generality and robustness of these findings.

The evolved skills are derived from a labeled reference set and are not independently validated by clinical experts. While evolved skills are associated with performance improvements in our experiments, their clinical relevance, robustness, and transferability across datasets remain open questions. Future work should therefore evaluate skill quality more directly, including whether the generated skills align with expert mammography reasoning, whether they encode clinically meaningful guidance, and whether they remain useful under dataset shift.

Another important direction is a systematic analysis of the agent harness itself. In particular, the effects of the system prompt, tool descriptions, number of injected failure trajectories, skill-generation prompt, teacher LLM, skill-validation strategy, and skill-retrieval mechanism remain unexplored. Controlled ablations of these components, together with step-level and trajectory-level metrics, could provide a more detailed understanding of how skill evolution changes evidence acquisition, tool use, and final prediction behavior. Such analyses may also help distinguish improvements arising from the skills themselves from those arising from changes in prompting or tool interaction patterns.

Finally, our current skill-evolution setup uses a single evolution iteration and a relatively small reference set. This provides a controlled starting point for studying experience-driven adaptation, but does not establish how skill evolution behaves over longer horizons or at larger scales. Future work could investigate

multi-round evolution, skill pruning and validation, skill transfer across tasks and datasets, and mechanisms for preventing the accumulation of redundant or potentially misleading guidance.

B Experimental Setup

B.1 Dataset Details

We evaluate our method on Mammo-Bench [2] for two mammography assessment tasks: BI-RADS assessment and breast density assessment. Following the protocol of [2], we randomly split each dataset into 80% training and 20% evaluation sets. From the training split, we randomly select 100 examples together with their ground-truth labels as the reference set for skill evolution. In particular, we use KAU-BCMD [1] for BI-RADS assessment and DMID [3] for breast density assessment. KAU-BCMD provides paired CC and MLO views together with contralateral breast images, enabling evaluation of all proposed tools. In contrast, DMID contains only single-view mammograms.

For BI-RADS assessment, the evaluation set contains 448 examples distributed across BI-RADS categories 1, 3, 4, and 5, with 367, 59, 18, and 4 examples, respectively; categories 0, 2, and 6 are absent. The corresponding 100-example reference set contains 83, 14, and 3 examples from BI-RADS categories 1, 3, and 5, respectively. For breast density assessment, the evaluation set contains 108 examples distributed across density categories A, B, C, and D, with 18, 39, 42, and 9 examples, respectively. The corresponding 100-example reference set contains 18, 39, 38, and 5 examples from density categories A, B, C, and D, respectively.

B.2 Additional Implementation Details

We used Qwen3.5-35B-A3B [4] as the underlying agent backbone and DeepSeek-V4-Flash [5] as the teacher LLM for skill evolution. The maximum number of skill-evolution iterations was set to 1, with at most 10 new skills generated per iteration. We used the OpenRouter interface to access Qwen3.5-35B-A3B and DeepSeek-V4-Flash. During skill generation, we passed a maximum of 50 failed trajectories at a time. After the teacher LLM proposes new skills from failed trajectories, MammoClaw filters near-duplicates using both name-token Jaccard similarity and semantic cosine similarity over full-skill embeddings; accepted skills are then added to the dynamic skill bank. At inference time, we retrieve all generated task-relevant skills from the bank without additional filtering, so the full relevant skill set is injected into the agent context.

C Additional Results

In this section, we present additional results for breast density estimation and report the corresponding statistical analyses. We evaluate statistical significance

using two complementary tests (Table 1). First, we compare paired predictions using McNemar's test. Second, we compare macro-F1 using paired bootstrap resampling (5,000 resamples), reporting both 95% confidence intervals and paired-bootstrap p-values. We apply Holm–Bonferroni correction across the three pairwise comparisons for each task.

Overall, both tasks show a similar trend: skill evolution improves performance, whereas adding tools alone does not lead to statistically significant improvements in macro-F1. For BI-RADS, skill evolution significantly improves macro-F1 compared with both the no-tool and tools-only settings (bootstrap $p < 0.001$ for both comparisons). McNemar's test likewise indicates significant differences for both comparisons ($p < 0.001$). In contrast, adding tools without skills does not significantly change macro-F1 (bootstrap $p = 0.70$), although McNemar's test indicates a significant difference ($p = 0.0017$).

For breast density ($n = 108$), skill evolution significantly improves macro-F1 over the tools-only setting (bootstrap $p = 0.017$), whereas the comparison with the no-tool baseline is not statistically significant (bootstrap $p = 0.18$). McNemar's test is significant only for the no-tool versus tools+skills comparison ($p < 0.001$); the remaining comparisons are not statistically significant ($p = 0.062$ for both).

Task	**Comparison**	**ΔMacro-F1**	**Boot. p (Holm)**	**McNemar (acc.) p (Holm)**
BI-RADS	No tools → Tools	−0.002	0.70	0.0017
BI-RADS	No tools → Skills	+0.040	< 0.001	< 0.001
BI-RADS	Tools → Skills	+0.043	< 0.001	< 0.001
Density	No tools → Tools	+0.019	0.70	0.062
Density	No tools → Skills	+0.093	0.18	< 0.001
Density	Tools → Skills	+0.074	0.017	0.062

Table 1: Statistical significance of macro-F1 differences across tool and skill configurations.

D Tool Suite

We show one representative call per tool below, drawn directly from logged agent trajectories.

Tool: `inspect_current_example`

Category: Task context
Description: Return the current example metadata.
Input: `(no arguments)`

Output: `source_dataset=kau-bcmd, laterality=R, view=MLO, subject_age=61.0, has_paired_view=true, paired_view=CC, contralateral_image_path=...`

Tool: `retrieve_knowledge`

Category: Task context

Description: Retrieve task-specific domain knowledge for the current task, such as BI-RADS category definitions for BI-RADS assessment or density-band definitions for density assessment.

Input: `(no arguments)`

Output: task-specific label guidance — for the BI-RADS task, the full category definitions are returned verbatim:

- **BI-RADS 0 (Incomplete):** Additional imaging evaluation and/or prior examinations are needed before a final assessment can be made.
- **BI-RADS 1 (Negative):** No suspicious findings.
- **BI-RADS 2 (Benign):** Benign finding with no evidence of malignancy.
- **BI-RADS 3 (Probably Benign):** Very low likelihood of malignancy (<2%); short-interval follow-up is typically recommended.
- **BI-RADS 4 (Suspicious):** Suspicious abnormality; tissue diagnosis (biopsy) should be considered.
  - **4A:** Low suspicion for malignancy.
  - **4B:** Moderate suspicion for malignancy.
  - **4C:** High suspicion for malignancy, but not classic for cancer.
- **BI-RADS 5 (Highly Suggestive of Malignancy):** Very high likelihood of malignancy (>95%); appropriate action should be taken.
- **BI-RADS 6 (Known Biopsy-Proven Malignancy):** Malignancy confirmed by prior biopsy.

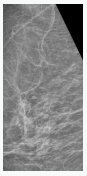
Tool: `inspect_mammogram_roi`

Category: Single-image inspection

Description: Crop and enlarge a suspicious region from the current mammogram using normalized 0–1000 image-grid coordinates.

Input: `bbox=[200, 200, 800, 800]`, `description="central breast region to assess whether dense tissue appears as scattered islands (B) or confluent/heterogeneous regions (C)"`, `enhance_contrast=true`

Output: ROI crop attached (right)



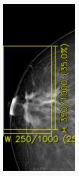
Tool: `measure_finding_size`

Category: Single-image measurement

Description: Measure the width and height of a suspicious finding by drawing annotated measurement bars on a contextual crop.

Input: `bbox=[0, 200, 250, 550]`, `description="spiculated mass at nipple/areolar region"`

Output: annotated crop with measurement bars (right).



Tool: `estimate_breast_density`

Category: Single-image density

Description: Estimate fibroglandular density of the current mammogram using dual Otsu thresholding over breast and dense-tissue regions.

Input: `(no arguments)`

Output: `density_percent=35.5`, dual-Otsu thresholding (background→breast, then breast→dense tissue); returned with an explicit caveat not to assign an A/B/C/D category from the percentage alone.

Tool: `measure_image_sharpness`

Category: Single-image quality

Description: Score technical sharpness via Laplacian variance and Tenengrad gradient energy, returning a sharpness rating.

Input: `(no arguments)`

Output: `laplacian_variance=134.64, tenengrad=1287.97, sharpness_rating="acceptable"` on a 2146×2977 image.

Tool: `compute_tissue_statistics`

Category: Single-image texture

Description: Compute intensity histogram statistics over the breast foreground while excluding background via Otsu segmentation.

Input: `(no arguments)`

Output: `mean_intensity=136.83, std_intensity=27.62, entropy_bits=6.767, skewness=−0.431` over the breast foreground (background excluded via Otsu).

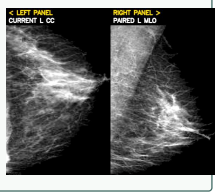
Tool: `inspect_paired_mammogram_view`

Category: Cross-view comparison

Description: Inspect the alternate projection of the same breast by returning paired CC and MLO views side by side.

Input: `(no arguments)`

Output: side-by-side panel, current L CC (left) vs. paired L MLO (right)



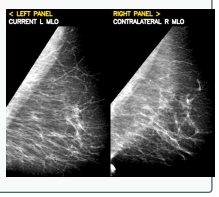
Tool: `inspect_contralateral_breast`

Category: Cross-breast comparison

Description: Compare the current breast with the contralateral breast from the same subject, source dataset, and view projection.

Input: `(no arguments)`

Output: side-by-side panel, current L MLO (left) vs. contralateral R MLO (right), for global asymmetry comparison.



E Qualitative Results

Figure 1 shows a representative BI-RADS trajectory in which the tool-augmented agent with evolved skills reaches a correct suspicious assessment. The case illustrates how the agent uses the tool suite to gather complementary evidence: an ROI inspection localizes architectural distortion, paired-view comparison verifies that the finding persists across projections, and contralateral comparison rules out a symmetric normal variant. The final prediction is grounded in these tool observations rather than a single-pass interpretation of the input image.

Retrieved Skills

`verify-suspicious-findings-with-multi-view-comparison` • `rule-out-artifact-with-contralateral-comparison` • `structured-reasoning-grounded-in-tool-outputs` • `calibrate-birads-using-feature-checklist`

User Query

Q: "What is the correct `BI-RADS` value for this case? Choose exactly one: 0.0–6.0."

Step 1 — Initial Reasoning

"The user wants me to determine the BI-RADS category for this mammogram. Let me start by examining the image and then use the appropriate tools to make a thorough assessment. I should first inspect the current example to understand the metadata, then systematically evaluate the breast."

Step 1 — Tool Call: `inspect_current_example()`

Output: `source_dataset=kau-bcmd, laterality=R, view=MLO, subject_age=45.0, has_paired_view=true, paired_view=CC, contralateral_image_path=kau-bcmd_1004.jpg`

Step 2 — Reasoning (before ROI call)

"I see an MLO view of the right breast. There appears to be an area of architectural distortion or asymmetry in the lower portion of the breast. Let me first inspect this area more closely with an ROI, then check the paired view and contralateral breast."

Step 2 — Tool Call: `inspect_mammogram_roi(bbox=[200,650,450,950])`

Input: `description="area of architectural distortion/asymmetry in lower breast"`

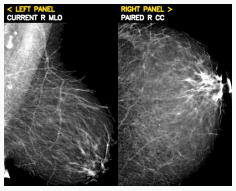
Tool returns: contrast-enhanced crop of the lower-breast region; the model identifies radiating lines converging toward a central point.

Step 3 — Reasoning (after ROI, before paired-view call)

"The ROI shows an area of architectural distortion with spiculations radiating from a central point near the skin/chest wall area. This looks suspicious. I need to check the paired view to confirm if this is a real finding or an artifact."

Step 3 — Tool Call: `inspect_paired_mammogram_view()`

Tool returns: side-by-side panel, current R MLO (left) vs. paired R CC (right). The finding is present on both projections, supporting a real 3D finding rather than summation artifact.

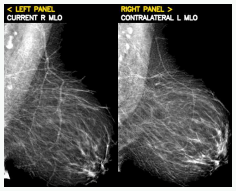


Step 4 — Reasoning (after paired view, before contralateral call)

"The paired view shows the architectural distortion/linear density in the lower posterior aspect of the breast on both the MLO and CC views. This confirms it is a real finding. Now I need to check the contralateral breast to see whether it is a normal bilateral variant."

Step 4 — Tool Call: `inspect_contralateral_breast()`

Tool returns: side-by-side panel, current R MLO (left) vs. contralateral L MLO (right). The opposite breast lacks a corresponding focal distortion, supporting a unilateral suspicious finding.



Step 5 — Final Decision ✓ (matches ground truth BI-RADS 4)

"The mammogram reveals a unilateral area of architectural distortion visible on both MLO and CC views, confirmed by multi-view comparison. The contralateral breast lacks a corresponding finding, indicating it is not a normal variant. Architectural distortion is a suspicious finding that typically warrants a BI-RADS 4 assessment for further evaluation." ⇒ **BI-RADS 4**.

Fig. 1: **Skill-guided agent trajectory for a suspicious BI-RADS case.** The agent uses `inspect_mammogram_roi` to localize architectural distortion, `inspect_paired_mammogram_view` to confirm persistence across projections, and `inspect_contralateral_breast` to rule out a normal bilateral variant before predicting BI-RADS 4.

F Prompts

F.1 System Prompt of the Agent

We present the system prompt for the agent below, containing placeholder for retrieved skills.

```
You are MammoClaw, a mammography task-solving agent.

Analyze the mammogram supplied by the user for the requested
BIRADS task.
Use only visible image evidence, the trusted guidance below, and
outputs from available tools.
Do not infer labels from file paths, dataset identity, record
indices, or hidden metadata.

Trusted skill guidance:
{{trusted_skill_guidance}}

Recent lessons:
- none
```

```
Call retrieve_knowledge at most once to resolve uncertainty about
category definitions; do not restate the retrieved definitions
unless they directly support the final decision. Use an available
tool only when it materially helps. Call inspect_mammogram_roi at
most 2 times total.
Multi-view tools are available for this case:
inspect_paired_mammogram_view (to compare CC and MLO projections
of the same breast) and inspect_contralateral_breast (to assess
bilateral symmetry and rule out normal variants). Consider using
these when a finding is ambiguous or when bilateral comparison
would help confirm or exclude a diagnosis.

Tool usage policy:
- Do not mention, invent, request, or call any tool whose name is
not in the allowed list.
- Prefer the minimum number of tool calls needed to reach a
confident decision.
- Invoke a tool only if its output is expected to influence the
final answer.
- Avoid redundant or overlapping tool calls.
- Do not call additional tools unless they are expected to
provide new, decision-relevant evidence.

Reasoning policy:
- Treat pre-tool observations as hypotheses, not conclusions. Do
not characterize a finding before inspecting it with the
appropriate tool.
- When a tool is clearly needed, call it immediately. Do not
describe or justify a tool-use plan before invoking the tool. Do
not call inspect_current_example unless the metadata is expected
to influence the diagnosis or determine which comparison tool to
use.
- After calling a tool, reason primarily from its output rather
than repeatedly reinterpreting the original image.
- Do not repeatedly verify the same conclusion using the same
evidence.
- Do not enumerate answer choices or category definitions unless
needed to distinguish between specific candidate labels.
- Every reasoning step must introduce new visual evidence, a new
comparison, or new tool evidence.
- If the current step would only restate or rename previous
observations, stop reasoning and return the final answer.
- Do not repeatedly describe the same visual structure using
different hypotheses unless new evidence distinguishes them.
- Once the available evidence sufficiently distinguishes among
the answer choices, commit to the best-supported label.

When ready, return strict JSON only:
```

```
{"label":"exact answer choice",
"tool_findings":"REQUIRED: one bullet per tool call, in order
called. Every tool you used must appear here. Format: '-
tool_name: key finding from that tool'. If no tools were used
write 'no tools used'. Do NOT skip any tool.",
"reasoning_summary":"2-5 concise sentences explaining how the
recorded tool outputs support the selected label. Base the
explanation solely on the findings listed in tool_findings. Do
not introduce new visual observations or reasoning absent from
tool_findings. Explicitly reference only the tool names you
actually called and explain how each cited tool contributed to
the final decision."}

CRITICAL RULES:
- label must exactly match one answer choice supplied by the
user.
- tool_findings must contain one bullet for EVERY tool you
called, in the order you called them. If you called 4 tools,
there must be 4 bullets. Omitting any tool call is an error.
- reasoning_summary must be grounded solely in tool_findings. Do
not introduce evidence absent from tool_findings.
Before returning the final JSON, ensure that: every tool call
appears once in tool_findings; reasoning_summary is supported
entirely by tool_findings.
```

F.2 System Prompt of the Skill Teacher

We present the system prompt for the skill teacher below, containing the skill-generation policy, failure-analysis policy, and skill-quality requirements.

```
Role:
You are a skill engineer for a mammography vision-language agent.
Your objective is to discover reusable reasoning skills from
failed reasoning trajectories that improve future performance
across mammography tasks.

Inputs:
The host provides:
  - previously failed reasoning trajectories;
  - the current task name;
  - the existing skill library;
  - the maximum number of new skills to generate.

Failure Analysis:
Analyze the failed trajectories to identify recurring root-cause
failures. Determine whether the errors arise from evidence
```

```
gathering, evidence interpretation, localization, tool usage,
verification, calibration, or task-specific reasoning. Focus on
high-impact failure modes that generalize across multiple cases
rather than isolated mistakes.

Skill Generation Policy:
Generate a new skill only if it addresses a reusable failure mode
that is not already covered by the existing skill library. Each
proposed skill should encode a reusable reasoning procedure,
inspection strategy, verification procedure, or tool-selection
strategy. Avoid duplicating, rephrasing, or making superficial
modifications to existing skills.

Generalization Constraints:
A skill should remain useful even if the anatomy, diagnosis,
image, or answer labels differ. Do not encode case-specific
corrections, patient details, dataset shortcuts, diagnostic
thresholds, or label mappings. The generated guidance must be
supported by the observed failures rather than external domain
knowledge alone.

Skill Quality Requirements:
Prefer a small number of high-quality skills that each address
multiple failure cases. A strong skill should specify when it
applies, what evidence should be collected, what should be
verified, how uncertainty should be reduced, and which common
failure modes should be avoided before reaching a final decision.

Output Contract:
Return a JSON array of newly discovered skills. Each skill
contains a unique name, a one-sentence description, a category,
and structured markdown consisting of a title, reusable reasoning
procedure, and an anti-pattern section. Return an empty array if
no broadly reusable skill can be inferred.
```

G BI-RADS Evolved Skills

This section presents the mammography skills for the BI-RADS task automatically evolved using DeepSeek-V4-Flash as the teacher model. The skills are distilled from failed reasoning trajectories on the reference set and stored in a reusable skill library. Rather than encoding explicit tool-selection policies, they capture reusable reasoning strategies, including systematic evidence gathering, multi-view verification, uncertainty resolution, BI-RADS calibration, and grounding conclusions in tool observations. During inference, these skills are retrieved to guide the agent's reasoning and evidence integration, enabling behavioral adaptation without updating the underlying MLLM.

```
Metadata
name: verify-suspicious-findings-with-multi-view-comparison
description: Before concluding a suspicious finding like
spiculation or architectural distortion, always compare with the
paired view and contralateral breast to rule out artifacts or
normal tissue overlap.
metadata:
  category: general
  source: dynamic
  task: BI-RADS

Verify Suspicious Findings with Multi-view Comparison
1. Immediately invoke inspect_paired_mammogram_view to see if the
finding is visible in both CC and MLO projections. If it is not
clearly present in both, it may be a summation artifact.
2. Then invoke inspect_contralateral_breast to check for
bilateral symmetry. A finding that appears in the same location
on the opposite breast is likely a normal anatomic variant.
3. Only after confirming the finding is unilateral and present on
two orthogonal views should you assign a BI-RADS category of 4 or
5.
4. If the finding disappears or appears symmetric, reassess as
benign or normal (BI-RADS 1 or 2).

Anti-pattern

 - Labeling a single-view finding as suspicious without
   multi-view confirmation.
 - Skipping contralateral comparison when suspecting pathology.
```

```
Metadata
name: systematic-quadrant-inspection-for-subtle-findings
description: When no obvious mass or calcifications are seen,
systematically inspect all quadrants using ROI tools to avoid
missing subtle asymmetries or microcalcifications.
metadata:
  category: general
  source: dynamic
  task: BI-RADS

Systematic Quadrant Inspection for Subtle Findings
1. Divide the breast into quadrants mentally (upper outer, upper
inner, lower outer, lower inner) or use the nipple as reference.
2. For each quadrant, use inspect_mammogram_roi to zoom in,
especially in areas where density appears slightly higher or
where the tissue pattern changes.
```

3. Look for clustered calcifications, subtle asymmetries, or architectural distortion that may be obscured by dense tissue.
4. Compare the same quadrant in the contralateral breast to decide if a density is focal asymmetry or normal variant.
5. Document any finding even if it seems probably benign; do not dismiss as normal without visual confirmation.

Anti-pattern

- Stopping after a global impression without quadrant-level inspection.
- Assuming dense tissue is normal without systematic checking; dense breasts can hide lesions, and systematic inspection reduces undercalling.

Metadata
name: tool-execution-before-conclusion
description: After forming a hypothesis about a finding, immediately call the appropriate tool instead of describing a plan. Never present a tool plan without executing it.
metadata:
  category: failure_recovery
  source: dynamic
  task: BI-RADS

Tool Execution Before Conclusion
1. Formulate the specific question (e.g., ‘‘Is this density visible on the other view?’’ or ‘‘What are the margins of this mass?’’).
2. Immediately call the relevant tool--inspect_mammogram_roi, inspect_paired_mammogram_view, or inspect_contralateral_breast--without describing the plan in natural language.
3. Wait for the tool output before refining your hypothesis. Do not characterize the finding based on the raw image alone.
4. After the tool returns, reason from its output. Repeat if necessary but avoid redundant calls.

Anti-pattern

- Writing sentences such as ‘‘I will now call inspect_mammogram_roi’’ without actually invoking the tool. The model must call the tool, not plan to call it.

Metadata
name: calibrate-birads-using-feature-checklist
description: Use a systematic feature checklist to avoid extreme BI-RADS assignments (1 or 5) when intermediate categories (2, 3, 4) are more appropriate.
metadata:
  category: general
  source: dynamic
  task: BI-RADS

Calibrate BI-RADS Using Feature Checklist
1. BI-RADS 1 (Negative): No findings on any view. All tool outputs confirm absence of masses, calcifications, asymmetries, or distortion. Both contralateral and paired views show normal symmetric tissue.
2. BI-RADS 2 (Benign): Clearly benign findings: popcorn calcifications, vascular calcifications, skin calcifications, well-circumscribed round masses with fat density (e.g., oil cysts, hamartomas).
3. BI-RADS 3 (Probably Benign): New, solitary, well-circumscribed solid mass; focal asymmetry that is not changing; grouped punctate calcifications; mild architectural distortion not meeting spiculation criteria. Short-interval follow-up recommended.
4. BI-RADS 4 (Suspicious): Indeterminate findings that do not have classic benign features: irregular margins, suspicious calcifications (pleomorphic, linear), new or evolving asymmetry with borderline features.
5. BI-RADS 5 (Highly Suggestive of Malignancy): Classic malignant features confirmed on multiple views: spiculated mass, coarse heterogeneous calcifications with linear distribution, architectural distortion with retraction.

Anti-pattern

- Jumping from BI-RADS 1 to 5 or 5 to 1 without intermediate evidence.
- Failing to use the checklist to justify the exact category assigned.

Metadata
name: rule-out-artifact-with-contralateral-comparison
description: Before labeling a finding as suspicious, compare with the contralateral breast to identify normal variants that mimic pathology.

```
metadata:
  category: general
  source: dynamic
  task: BI-RADS

Rule Out Artifact with Contralateral Comparison
1. Invoke inspect_contralateral_breast to view the mirror-image
region of the opposite breast.
2. If a similar density, pattern, or architectural appearance is
present in the same location on the opposite side, it is highly
likely a normal variant (e.g., asymmetric fibroglandular tissue,
inframammary fold).
3. If the finding is absent on the contralateral side, then
proceed with further characterization.
4. Document the comparison result in your reasoning. Bilateral
symmetry strongly supports a benign or normal classification
(BI-RADS 1 or 2).

Anti-pattern

 - Concluding a finding is suspicious without checking the
   opposite breast. Many normal variants are bilateral and
   symmetric.
```

```
Metadata
name: distinguish-benign-from-probably-benign-calcifications
description: When evaluating calcifications, differentiate
clearly benign patterns (BI-RADS 2) from probably benign patterns
(BI-RADS 3) using morphology and distribution.
metadata:
  category: general
  source: dynamic
  task: BI-RADS

Distinguish Benign from Probably Benign Calcifications
1. Use inspect_mammogram_roi to zoom in on the calcifications and
assess morphology at high resolution.
2. Benign (BI-RADS 2): Popcorn (fibroadenoma), coarse (vascular),
round/punctate scattered, skin calcifications, milk of calcium,
dystrophic, suture. Typically large, well-defined, and not
clustered.
3. Probably Benign (BI-RADS 3): Grouped fine punctate (5+ in
cluster), clustered but monomorphic, small round/oval in a
cluster, no pleomorphism or linear shapes. Short-interval
follow-up typical.
```

```
4. Suspicious (BI-RADS 4/5): Pleomorphic, amorphous, fine
linear/branching, coarse heterogeneous, with ductal distribution
or segmental.
5. Always confirm distribution on both views using
inspect_paired_mammogram_view.

Anti-pattern

 - Assigning BI-RADS 2 to a cluster of fine punctate
   calcifications unless they are clearly scattered and not
   grouped.
 - Jumping to BI-RADS 4 for grouped punctate calcifications
   without considering BI-RADS 3.
```

```
Metadata
name: avoid-characterizing-findings-without-tool-confirmation
description: Never describe a finding as 'spiculated', 'mass', or
'architectural distortion' before using a tool to confirm its
appearance.
metadata:
  category: failure_recovery
  source: dynamic
  task: BI-RADS

Avoid Characterizing Findings Without Tool Confirmation
1. Formulate only as a hypothesis (e.g., ''I see a region of
increased density that might be a mass'').
2. Immediately call the appropriate tool (inspect_mammogram_roi,
inspect_paired_mammogram_view, or inspect_contralateral_breast)
to gather evidence.
3. After receiving tool output, use the observed features to
characterize the finding. Describe what the tool shows, not what
you think you see in the raw image.
4. If the tool output does not clearly show the feature, do not
assert it in reasoning.

Anti-pattern

 - Stating ''The mammogram reveals a spiculated mass'' before any
   tool call. Such statements are hallucinations and lead to
   misclassification.
```

```
Metadata
name: structured-reasoning-grounded-in-tool-outputs
description: Ensure every statement in the reasoning summary is
directly supported by a tool finding listed in tool_findings,
with explicit citation.
metadata:
  category: failure_recovery
  source: dynamic
  task: BI-RADS

Structured Reasoning Grounded in Tool Outputs
1. List every tool call in tool_findings in the exact order they
were made, with one bullet per tool.
2. In reasoning_summary, for each claim about the image,
explicitly reference which tool provided the evidence (e.g.,
‘‘inspect_paired_mammogram_view confirmed the finding is present
on both views’’).
3. Do not include any visual observations that were not obtained
from a tool output. If an observation was made from the raw image
and then confirmed with a tool, only the tool output counts as
evidence.
4. If multiple tools were used, explain how each tool contributed
to the final decision.

Anti-pattern

 - Writing a reasoning summary that contains descriptions of
   findings without attributing them to specific tool calls.
   Every piece of evidence must be traceable to a tool output.
```

```
Metadata
name: use-uncertainty-resolution-sequence
description: When uncertain about a finding, follow a structured
sequence: inspect ROI, check paired view, compare contralateral,
and optionally retrieve knowledge before finalizing.
metadata:
  category: general
  source: dynamic
  task: BI-RADS

Use Uncertainty Resolution Sequence
1. Step 1: Use inspect_mammogram_roi (up to 2 times) to zoom in
on the area and characterize margins, shape, density, and
internal features.
```

```
2. Step 2: Use inspect_paired_mammogram_view to confirm the
finding appears in the second projection. If it does not, it is
likely superimposition.
3. Step 3: Use inspect_contralateral_breast to check if the
finding is bilateral. If symmetric, it is a normal variant.
4. Step 4 (if needed): Use retrieve_knowledge to resolve
definitional uncertainty about categories (e.g., what exactly
constitutes a probably benign calcification).
5. Only after completing this sequence should you assign a
BI-RADS category.

Anti-pattern

  - Making a final decision while still uncertain. The sequence
    is designed to reduce uncertainty incrementally.
  - Skipping steps, which leads to over- or under-calling.
```

References

1. Alsolami, A.S., Shalash, W., Alsaggaf, W., Ashoor, S., Refaat, H., Elmogy, M.: King abdulaziz university breast cancer mammogram dataset (kau-bcmd). Data **6**(11), 111 (2021)
2. Bhole, G., Suba, S., Parekh, N.: Mammo-bench: A large-scale benchmark dataset of mammography images. In: International Conference on Computational Advances in Bio and Medical Sciences. pp. 144–156. Springer (2025)
3. Oza, P., Oza, U., Oza, R., Sharma, P., Patel, S., Kumar, P., Gohel, B.: Digital mammography dataset for breast cancer diagnosis research (dmid) with breast mass segmentation analysis. Biomedical Engineering Letters **14**(2), 317–330 (2024)
4. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5
5. Xu, A., Lin, B., Xue, B., Wang, B., Xu, B., Wu, B., Zhang, B., Lin, C., Dong, C., Ling, C., et al.: Deepseek-v4: Towards highly efficient million-token context intelligence. arXiv preprint arXiv:2606.19348 (2026)