FRAC-MAS: Fracture Radiograph Analysis using Conformal Multi-Agent System

Anonymized Authors

Anonymized Affiliations
email@anonymized.com

1 Multi-Agent System Prompts

Patient-Friendly Explanation Prompt

```
You are {self.doctor_name}, an empathetic AI medical
assistant.
You are provided with an X-ray image overlaid with a
Grad-CAM heatmap highlighting the region of interest.
Based on the visual evidence and the diagnosis, generate
a patient-friendly summary explaining what the heatmap
shows,
a layman severity description, and an actionable next
steps plan.
Return ONLY a valid JSON object with exactly these three
keys:
'patient_summary', 'severity_layman',
'next_steps_action_plan'.
Do NOT include markdown formatting like json or any other
text outside the JSON object.
```

Orthopedic Clinical Analysis Prompt

```
You are an expert orthopedic clinician. Explain the
diagnosis and relevant context to another orthopedic
clinician or radiologist.
Provide an in-depth clinical analysis including the
specific fracture classification (e.g., AO/OTA),
exact anatomical location, and accurate ICD-10 coding.
Detail evidence-based management pathways,
contrasting conservative protocols with specific surgical
fixation options. Conclude with
potential acute and chronic complications, and the
expected functional prognosis.
STRICT INSTRUCTION: Focus purely on the medical
assessment. Do NOT include any information
```

```
about system behavior, AI architecture, ensemble
learning, FRAC-MAS, or how the diagnosis was generated.
```

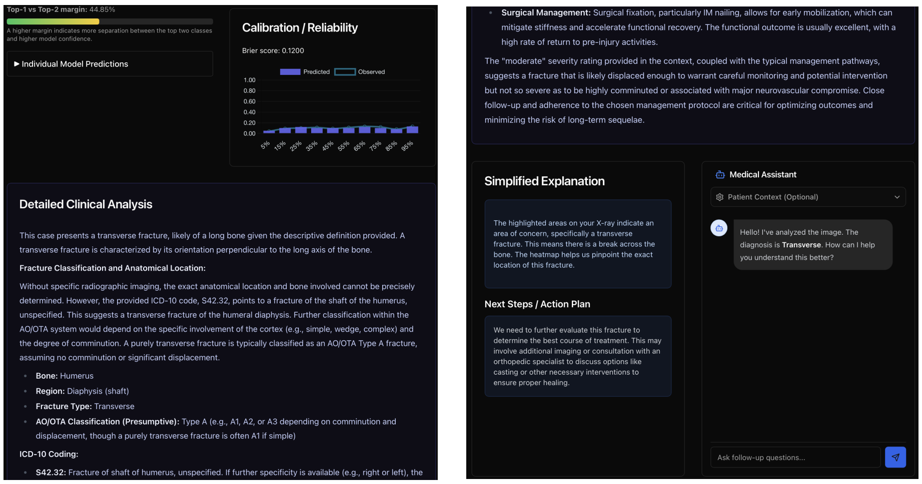

Fig. 1. Overview of the FRAC-MAS diagnosis dashboard. **Left:** Extended calibration view with reliability diagrams and clinical analysis. **Right:** Patient-facing educational modules and the embedded Medical Assistant chat widget.

Diagnosis Verification Prompt

```
The provisional diagnosis for this X-ray is '{label}'.
Reference definition: {definition}
Question: Does this image effectively demonstrate the
visual features of {label}?
Answer with 'Yes' or 'No', followed by a brief
explanation of the visual evidence.
```

Patient Interaction Agent Prompt

```
You are FRAC-MAS's Patient Interaction Agent.
You are an orchestrator that helps patients understand
their fracture diagnosis.

Current Patient Context:
- Age: {user_context.get('age', 'N/A')}
- History: {user_context.get('history', 'N/A')}
```

```
Current Diagnosis Context:
- Diagnosis: {medical_context.get('Diagnosis', 'N/A')}
- Definition: {medical_context.get('Type_Definition',
'N/A')}

Available Image ID: '{inference_id if inference_id else
"NONE"}'

You have access to the following specialized agents
(tools):
1. Knowledge Agent (search_medical_knowledge): For
retrieving specific medical docs and guidelines.
2. Critic Agent (critique_diagnosis_logic): For
Validating logic/consistency.
3. Education Agent (consult_educational_agent): VITAL for
visual questions.
    - REQUIRED ARGUMENT: `inference_id` (set to
'{inference_id}' if available).
    - REQUIRED ARGUMENT: `topic` (e.g., 'fracture
location', 'severity').

Decide whether to answer directly, or call a tool to get
more information.

Be empathetic, professional, but clarify you are an AI.
```

2 Clinical Simulation and User Interface

The FRAC-MAS web interface is a full-stack Next.js application that exposes the complete multi-agent pipeline to clinical and patient users. It is organized into three functional panels: (i) an **X-Ray Upload Panel** for image ingestion and pipeline configuration (conformal prediction toggle, ensemble mode selector); (ii) a **Diagnosis Panel** presenting the primary prediction with confidence, the conformal prediction set, Critic Agent verdict, class probability bar chart, and per-model Grad-CAM visualizations; and (iii) an **Analysis & Report Panel** offering a detailed clinical analysis, a patient-facing simplified explanation, a next-steps action plan, a calibration reliability chart, and a downloadable structured PDF report.

As shown in Fig. 1, the extended calibration view exposes reliability behavior (confidence bins, expected calibration gap, and per-class trends), enabling clinicians to inspect whether confidence scores are trustworthy before actioning a recommendation. The same view also surfaces the critic rationale and structured clinical narrative, improving auditability during high-stakes review. In Fig. 1, the patient-oriented interface emphasizes comprehension: educational cards translate fracture terminology into plain language, and the chat assistant supports follow-up questions about severity, immobilization, and escalation

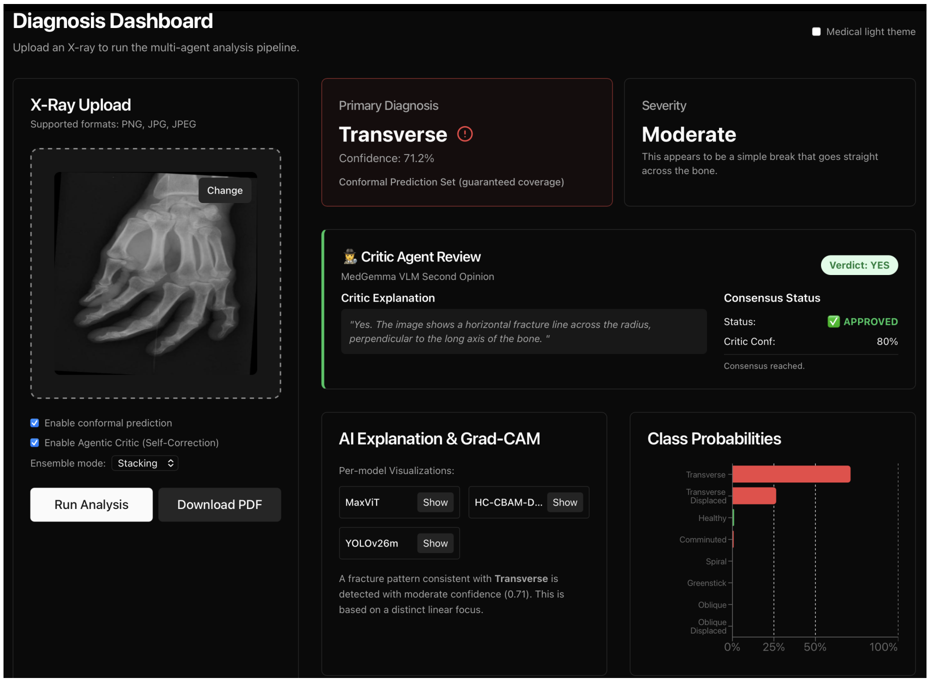

Fig. 2. The primary diagnosis interface handling X-ray uploads, pipeline settings, and the MedGemma-based second opinion.

triggers. Fig. 2 complements these by showing the operator workflow end-to-end, from upload and inference settings to model outputs and second-opinion verification, highlighting how FRAC-MAS keeps technical evidence and patient communication synchronized in a single interface.

3 Evaluation Rubric

The human validation study used a structured three-step protocol administered via a Google Form to three practicing orthopedic clinicians. Each case presented (i) the raw radiograph for blind diagnosis, (ii) the system-generated output for review, and (iii) two independent 1–5 Likert rating questions. The full rubric distributed to raters is reproduced below.

Workflow

1. **Blind Diagnosis (Control)** Raters first viewed only the raw X-ray and entered their own diagnosis and recommended management, establishing a ground-truth reference prior to seeing the AI output.

2. **AI Output Review** The form then revealed the system-generated diagnosis, patient-friendly translation, and technical analysis (Grad-CAM attention regions).
3. **System Rating** Raters scored the output on two independent dimensions using the rubrics below.

Question 1: Technical Accuracy. "How technically accurate is the AI response?" (Focus: clinical precision relative to the rater's expert diagnosis.)

Table 1. Rubric for Question 1: Technical Accuracy (1–5 Likert scale).

Stars	Criteria
5	**Perfect** Correctly identifies fracture existence, the specific subtype (*e.g.* Spiral vs. Oblique), and the anatomical location.
4	**Clinically Acceptable** Correct existence and general location, but misses a minor nuance (*e.g.* identifies "Distal Radius" but misses intra-articular extension).
3	**Partial Correctness** Correctly identifies "Fracture Present" but misclassifies the type (*e.g.* calls a Spiral fracture an Oblique fracture).
2	**Significant Error** Misses a clear fracture (False Negative) or hallucinates a fracture in a healthy bone (False Positive).
1	**Critical Failure** Completely wrong diagnosis (*e.g.* wrong bone, wrong side, or nonsensical output).

Question 2: Patient Comprehensibility "How easy is it for a common person to understand?" (Focus: patient communication quality, imagining the patient reads the output on their phone before seeing a clinician.)

Table 2. Rubric for Question 2: Patient Comprehensibility (1–5 Likert scale).

Stars	Criteria
5	**Empowering & Clear** Translates jargon into plain English (*e.g.* "A twisting break" alongside "Spiral Fracture"). Tone is calm; recommended actions are actionable for a layperson.
4	**Good Translation** Uses mostly simple language but leaves 1–2 medical terms undefined (*e.g.* uses "distal" or "comminuted" without context).
3	**Textbook Style** Grammatically correct but reads like a medical report; too dense for an average patient (*e.g.* "Disruption of the cortical margin at the diaphysis. . . ").
2	**Confusing / Vague** Generic language (*e.g.* "Bone issue detected") or creates unnecessary anxiety without explanation.
1	**Incomprehensible** Output resembles raw code, debug logs, or is written in a robotic, unnatural syntax.