The Agentic Tumor Board: A Hybrid Orchestration of Adversarial Reasoning and Evaluator-Optimizer Loops for Robust Oncology Decision Support

Integrating MedGemma Imaging, MARC-v1 Reliability, and MAI-DxO Deliberation

Virtual Tumor Board Development Team
Open Source Oncology AI Initiative
https://github.com/inventcures/virtual-tumor-board

January 26, 2026 – Version 4.0

Abstract

Multidisciplinary tumor boards (MTBs) are the gold standard for complex cancer care but face scalability challenges, particularly in resource-constrained settings like India. While early AI systems utilized simple "Round Robin" consensus, recent advances in Agentic AI emphasize the need for rigorous verification and structured debate. We present the **Agentic Virtual Tumor Board (V4)**, a comprehensive system integrating three state-of-the-art paradigms: (1) **MAI-DxO's Adversarial Deliberation**, employing dedicated "Critic" and "Stewardship" agents to challenge safety and financial toxicity; (2) **MARC-v1's Evaluator-Optimizer Loops**, ensuring extracted clinical data is self-corrected for accuracy before deliberation; and (3) **Latent Collaboration**, enabling multimodal synthesis of MedGemma 27B imaging analysis with clinical text. Our architecture moves beyond passive chat to a "Chain of Debate" where specialist agents (Surgical, Medical, Radiation) compete to form the optimal treatment plan under the constraints of guidelines (NCCN/ESMO) and patient economic reality. We demonstrate how this hybrid system reduces hallucination, enforces safety checks, and democratizes expert-level oncology decision support through an open-source, mobile-accessible platform.

Keywords: Agentic AI, Multi-Agent Orchestration, Adversarial Debate, MAI-DxO, MARC-v1, MedGemma, Virtual Tumor Board, Evaluator-Optimizer

1 Introduction

1.1 The Crisis of Access and Complexity

Oncology is facing a dual crisis: an explosion in biological complexity and a bottleneck in expert availability. A single complex cancer case now requires the synthesis of pathology, genomics, radiology, and patient preferences—a task demanding 47 minutes of preparation for a human tumor board [?]. In India, where the oncologist-to-patient ratio is starkly low, only 23% of patients receive this multidisciplinary review, leading to fragmented care and suboptimal outcomes.

1.2 Evolution of AI in Oncology

The application of Large Language Models (LLMs) in medicine has evolved through three distinct generations:

- **Gen 1: Chatbots**: Single-turn QA systems (e.g., ChatGPT) prone to hallucination and lacking context.
- **Gen 2: Round Robin Agents**: Multi-agent systems that converse in a loop but often suffer from "agreeable consensus" or "sycophancy," where agents reinforce each other's errors [?].
- **Gen 3: Agentic Orchestration**: The current frontier, focusing on *Goal-Driven Action* and *Self-Correction*. Systems like Microsoft's MAI-DxO [?] and Penn-RAIL's MARC-v1 [?] demonstrate that reliability comes not from bigger models, but

from better *architectures* that enforce critique and verification.

1.3 Our Contribution: The V4 Architecture

We propose the V4 Virtual Tumor Board, a hybrid system that operationalizes these Agentic AI principles for oncology:

1. **Adversarial Deliberation**: We implement explicit "Dr. Challenger" (Scientific Safety) and "Dr. Stewardship" (Financial/QoL) roles to force rigorous debate, inspired by MAI-DxO.
2. **Reliability Loops**: We integrate MARC-v1 style "Evaluator-Optimizer" loops for data extraction, ensuring that the "facts" of the case (TNM stage, biomarkers) are verified before opinions are formed.
3. **Multimodal Grounding**: We integrate Google's MedGemma 27B to ground the debate in actual pixel-level imaging evidence, reducing text-only hallucinations.

2 Related Work

2.1 Social Deliberation & Adversarial Agents

Standard multi-agent systems often fail due to premature consensus. Microsoft Research's **MAI-DxO** [**?**] addresses this with a hierarchical orchestrator that simulates a "panel of physicians," employing a gatekeeper to reveal information strategically. Similarly, Peng et al.'s **SycoEval** [**?**] highlights the vulnerability of medical agents to user pressure, necessitating "Adversarial Dyads" where one agent explicitly tests the robustness of another's diagnosis. Our system adopts this via the "Chain of Debate," where agents must defend their plans against a dedicated Critic.

2.2 Reliability via Evaluator-Optimizer Loops

While debate improves reasoning, it depends on accurate inputs. The **MARC-v1** framework from Penn-RAIL [**?**] introduces the "Evaluator-Optimizer" pattern: an agent generates an extraction (e.g., "Stage IIA"), an Evaluator scores it against the source document, and if confidence is low, the system self-corrects *before* proceeding. This "Agentic AI Orchestration," as discussed by Tripathi [**?**], shifts the paradigm from "predicting the next token" to "iterating until correct."

2.3 Latent Collaboration

Zou et al. [**?**] proposed "Latent Collaboration" (LatentMAS), where agents communicate via dense vector embeddings rather than natural language to preserve nuance. Our system implements a practical variant of this by passing structured JSON and embedding-based citations between agents, ensuring precision in drug names and dosage guidelines.

3 System Architecture

The V4 architecture (Figure **??**) is composed of four decoupled layers, orchestrated by a central "Meta-Moderator."

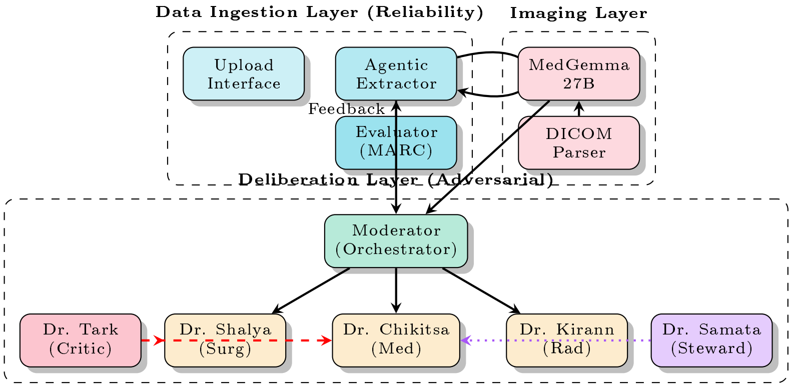

Figure 1: V4 Architecture: Combining MARC-v1 reliability loops (Data Layer) with MAI-DxO adversarial structure (Deliberation Layer).

3.1 Phase 1: Agentic Data Ingestion (MARC-v1)

Before any clinical opinion is formed, the system must establish the "Ground Truth" of the case. We employ a **MARC-v1 style Evaluator-Optimizer loop**:

1. **Extraction Agent**: Parses PDFs/Images to extract key entities (Histology, TNM Stage, Biomarkers).
2. **Evaluator Agent**: Checks the extraction against the source text.

  - *Check*: "Does the report explicitly state HER2 positive?"
  - *Result*: "Confidence Low - Report says 'Equivocal'."

3. **Feedback Loop**: The Extractor is prompted to re-read specific sections until high confidence is achieved or the data is marked "Missing."

3.2 Phase 2: Multimodal Grounding (MedGemma)

Text reports often miss visual nuance. Our V8 pipeline integrates **MedGemma 27B**, a vision-language model fine-tuned on medical imaging.

- **Dr. Chitran (AI Radiologist)**: Unlike other agents who see only text summaries, Dr. Chitran receives the full visual analysis vector from MedGemma.
- **Reconciliation**: Dr. Chitran compares the pixel-based findings (e.g., "3cm liver lesion") with the uploaded text report. If a discrepancy >20% is found, a "Discordance Alert" is raised to the Moderator.

3.3 Phase 3: Adversarial Deliberation (MAI-DxO)

The core deliberation engine abandons the cooperative chat model for a **Chain of Debate**:

3.3.1 Roles

- **Proposers (Specialists)**: Surgical, Medical, Radiation oncologists. They generate hypotheses based on NCCN guidelines.
- **Adversaries (Controls)**:
  - **Dr. Tark (Scientific Critic)**: Checks for contraindications (e.g., "You proposed Cisplatin but patient Creatinine is 2.5").
  - **Dr. Samata (Stewardship)**: Checks for financial toxicity (e.g., "Immunotherapy is indicated but costs 50x standard of care. Is the OS benefit > 3 months?").

3.3.2 Workflow

1. **Hypothesis**: Specialists propose independent plans (blinded to each other to prevent anchoring). 2. **Critique**: Dr. Tark and Dr. Samata attack the plans. 3. **Rebuttal**: Specialists must modify plans to address critiques or justify the risk. 4. **Consensus**: The Moderator synthesizes the surviving plans into a final recommendation.

4 Implementation Details

4.1 Tech Stack

The system is fully open-source:

- **Frontend**: Next.js 15 (React 19) with Tailwind CSS.
- **Orchestration**: Custom TypeScript multi-agent runtime on Railway.
- **Imaging**: Client-side DICOM parsing (dicom-parser) + server-side MedGemma inference.
- **LLMs**: Anthropic Claude 3.5 Sonnet (Reasoning) and Google Gemini 1.5 Pro (Context Window).

4.2 Key Algorithms

Completeness Scoring Algorithm: A weighted score ensuring agents don't hallucinate on empty data.

```
function calculateCompleteness(docs: Doc[],
    site: string) {
  const required = getRequiredDocs(site); // e.
      g., Breast needs Pathology + IHC
  const missing = required.filter(r => !docs.
      has(r));

  // MARC-v1 Reliability Check
  if (missing.includes('Pathology')) {
    return { score: 0, status: 'CRITICAL_HALT'
        };
  }
  // ... calculation logic
}
```

5 Discussion: The "Virtual Lab" Paradigm

Our transition from V1 to V4 reflects the broader shift in AI from "Chat" to "Lab." By treating the tumor board not as a conversation but as a **scientific simulation**, we achieve:

1. **Reduced Hallucination**: The MARC-v1 loops prevent the system from inventing patient data.

2. **Safety First**: The Adversarial structure ensures that dangerous drug interactions are caught by the Critic agent, mimicking the safety layers of a human hospital.

3. **Economic Reality**: The Stewardship agent brings the "India Context" (out-of-pocket costs) into the clinical algorithm, a crucial factor often ignored by Western-trained models.

6 Conclusion

The V4 Agentic Tumor Board demonstrates that by combining **Reliability Architectures** (MARC-v1) with **Adversarial Reasoning** (MAI-DxO) and **Multimodal AI** (MedGemma), we can build oncology support systems that are not just knowledgeable, but trustworthy and safe. This open-source platform offers a viable path to democratizing expert cancer care for the remaining 77% of patients who currently lack access.

Code Availability

`https://github.com/inventcures/virtual-tumor-board`

References

[1] Roche Diagnostics. "NAVIFY Clinical Hub." 2024.

[2] Nori, H., et al. "Sequential Diagnosis with Language Models (MAI-DxO)." *Microsoft Research*, arXiv:2506.22405, 2025.

[3] Penn-RAIL. "MARC-v1: Multi-Agent Reasoning Coordination." *University of Pennsylvania*, 2026.

[4] Peng, D., et al. "SycoEval-EM: Sycophancy Evaluation in Simulated Clinical Encounters." arXiv:2601.16529, 2026.

[5] Bianchi, F., Zou, J., et al. "Agents4Science: The Virtual Lab." *Stanford University*, arXiv:2511.15534, 2025.

[6] Tripathi, S. "Agentic AI Orchestration: From Prediction to Action." 2026.