The Agentic Tumor Board

Democratizing Precision Oncology via Hybrid Multi-Agent Orchestration

From Chatbot Oncology to Rigorous Clinical Deliberation

Virtual Tumor Board Initiative
github.com/inventcures/virtual-tumor-board

January 2026

Abstract

Multidisciplinary tumor boards (MTBs) represent the gold standard for cancer treatment decisions, yet remain structurally inaccessible to 77% of patients in India and billions worldwide. We present the **Agentic Virtual Tumor Board**, a hybrid multi-agent system that transcends "chatbot oncology" through rigorous architectural innovations.

Our system integrates three core components: (1) **MARC-v1** reliability loops achieving 95%+ extraction confidence through evaluator-optimizer patterns; (2) **MAI-DxO** adversarial deliberation preventing sycophantic consensus through role-based prompting and domain authority veto mechanisms; and (3) **MedGemma** multimodal grounding anchoring clinical decisions in pixel-level imaging evidence.

Evaluated across 10 clinically diverse synthetic cases spanning genomic complexity (KRAS G12C+ NSCLC), financial constraints (rural HER2-equivocal breast cancer), and rare presentations (pediatric GBM with H3 G34R), our system achieves 92% success in proposing financially viable, guideline-compliant treatment plans. The Stewardship Agent reduces recommended treatment costs by up to 70% through biosimilar substitution while maintaining clinical equivalence.

We demonstrate that treating tumor boards as *scientific simulations* rather than conversations—decoupling data ingestion from deliberation, enforcing adversarial critique, and grounding decisions in verified evidence—creates AI systems trustworthy for life-or-death decisions in resource-constrained settings.

Keywords: Multi-agent systems, Clinical decision support, Precision oncology, LLM safety, Global health equity, Tumor boards, RAG, Multimodal AI

Contents

1 Introduction

1.1 The Cognitive Crisis in Oncology

The complexity of modern oncology has outpaced human cognitive bandwidth. Consider what a single cancer patient now generates:

- **Pathology**: Whole-slide images at 40x magnification producing 10+ gigapixel files
- **Genomics**: NGS panels reporting 300+ genes, tumor mutational burden, microsatellite status
- **Radiology**: Volumetric CT/MRI series requiring RECIST 1.1 measurements across time-points
- **Clinical**: Longitudinal EMR with labs, medications, comorbidities, prior treatments

Synthesizing this into a coherent treatment plan requires a "hive mind"—the Multidisciplinary Tumor Board (MDT). In high-resource settings, an MDT spends 47 minutes per complex case [**?**]. This luxury evaporates in resource-constrained environments.

Key Insight

India has an oncologist-to-patient ratio of 1:2,000. The result is **fragmented care**: treatment plans decided by single overworked clinicians, missing rare genomic targets, ignoring financial toxicity, and lacking specialist input on surgical resectability or radiation planning.

1.2 Why Gen-1 AI (Chatbots) Failed

The first generation of medical AI optimized for *plausibility*, not *correctness*. An LLM will confidently hallucinate "HER2 Positive" to complete a sentence pattern, even when the pathology report clearly states "HER2 Equivocal (IHC 2+)." This failure mode is not merely academic—it leads to inappropriate Trastuzumab prescriptions costing Rs. 50,000/month for patients who may not benefit.

Recent benchmarks quantify this problem:

Table 1: Hallucination and Safety Failure Rates in Medical LLMs

Benchmark	Failure Rate	Source
Dynamic robustness (correct answers)	94%	Pan et al., 2025
Sycophantic behavior (overall)	58.19%	Fanous et al., 2025
Hallucination on medical QA	31%	Garcia-Fernandez et al., 2025
Privacy leakage rate	86%	Pan et al., 2025

1.3 The Gen-2 Paradigm: Agentic AI

To solve oncology, we need systems that can *reason*, *verify*, and *debate*. This paper presents such a system—the Agentic Virtual Tumor Board—built on three architectural principles:

1. **Decoupling**: Separate data ingestion (getting facts right) from deliberation (getting decisions right)
2. **Adversarial Structure**: Enforce productive conflict rather than sycophantic consensus
3. **Grounded Evidence**: Anchor every recommendation in verifiable clinical guidelines and imaging

1.4 Contributions

This paper makes the following contributions:

1. A **hybrid multi-agent architecture** combining MARC-v1 reliability loops, MAI-DxO adversarial deliberation, and MedGemma multimodal grounding
2. **Domain authority veto mechanisms** preventing inappropriate specialist override
3. A **Stewardship Agent** encoding financial toxicity and quality-of-life considerations for resource-constrained settings
4. **Comprehensive evaluation** across 10 clinically diverse cases representing Indian oncology scenarios
5. **Open-source implementation** enabling reproducibility and adaptation

2 Related Work

2.1 Multi-Agent Systems in Healthcare

The application of multi-agent LLM systems to healthcare has accelerated rapidly. Table 2 summarizes key systems and their limitations that our work addresses.

Table 2: Comparison of Multi-Agent Healthcare Systems

System	Approach	Limitation	Our Solution
MedAgents [?]	Role-playing collaboration	No adversarial critique	MAI-DxO debate
ColaCare [?]	MDT-inspired + RAG	Single-pass deliberation	Multi-round consensus
AgentClinic [?]	Multimodal simulation	90%+ accuracy drop in sequential tasks	MARC-v1 verification
HAO [?]	Tumor board orchestration	No financial considerations	Stewardship Agent

MedAgents [?] demonstrated that multi-disciplinary LLM collaboration improves zero-shot medical reasoning on MedQA and related benchmarks. However, their "round-robin" discussion format lacks mechanisms to prevent sycophantic agreement with dominant voices.

ColaCare [?] introduced MDT-inspired collaboration with DoctorAgents and a MetaAgent, achieving superior performance on mortality prediction across three EHR datasets. Their RAG integration with the Merck Manual provides evidence grounding, but single-pass deliberation misses opportunities for iterative refinement.

AgentClinic [?] provides a multimodal benchmark across 9 specialties and 7 languages, revealing that diagnostic accuracies drop to less than 10% of original performance in sequential decision-making scenarios. This finding motivated our MARC-v1 verification loops.

Healthcare Agent Orchestrator (HAO) [?] specifically addresses Molecular Tumor Boards, achieving 94% capture of high-importance information. While effective for patient summarization, HAO lacks consideration of resource constraints critical for global health applications.

2.2 Hallucination Prevention in Medical AI

The CHECK methodology [?] represents the current state-of-the-art in continuous hallucination detection, reducing LLaMA3.3-70B hallucinations from 31% to 0.3% using information-theoretic

approaches and structured clinical databases. Our MARC-v1 loops adapt this evaluator-optimizer pattern specifically for clinical document extraction.

MIRIAD [?] provides 5.8M medical QA pairs for grounded knowledge, demonstrating up to 6.7% accuracy improvement over unstructured RAG and 22.5–37% improvement in hallucination detection. We leverage similar corpus-grounding principles through our guideline RAG infrastructure.

2.3 AI Safety and Adversarial Evaluation

DAS Red-Teaming [?] provides a sobering assessment: 94% of correct MedQA answers fail dynamic robustness tests when questions are rephrased. SycEval [?] documents 58.19% sycophantic behavior across medical domains, with Gemini showing the highest rate at 62.47%.

These findings directly inform our MAI-DxO architecture, which enforces adversarial roles (Scientific Critic, Stewardship Agent) specifically designed to break sycophantic consensus patterns.

2.4 Multimodal Medical AI

MedGemma [?] achieves 50% EHR error reduction and 15–18% improvement on chest X-ray interpretation. PathFound [?] demonstrates that agentic multimodal models using RL-trained reasoning can achieve state-of-the-art diagnostic performance while discovering clinically relevant features like nuclear characteristics and local invasions.

Our Dr. Chitran (Radiologist) agent integrates MedGemma 27B for "latent grounding"—reconciling pixel-level AI findings with text reports to ensure debates are anchored in physical tumor reality.

2.5 Clinical Decision Support for Oncology

AMIE for Oncology [?] demonstrated conversational AI for breast oncology with web search and self-critique, outperforming trainees and fellows but remaining inferior to attending oncologists on 50 synthetic vignettes. Mohammed et al. [?] achieved 100% guideline adherence using Agentic-RAG for NCCN breast cancer recommendations.

Our work extends these approaches by (1) covering all major cancer types, not just breast; (2) integrating financial toxicity considerations; and (3) providing a full MDT simulation rather than single-specialty consultation.

3 System Architecture

The Agentic Virtual Tumor Board creates a "Virtual Lab" where agents function not as peers in casual conversation, but as specialists with distinct—often conflicting—roles. Figure 1 presents the high-level system design.

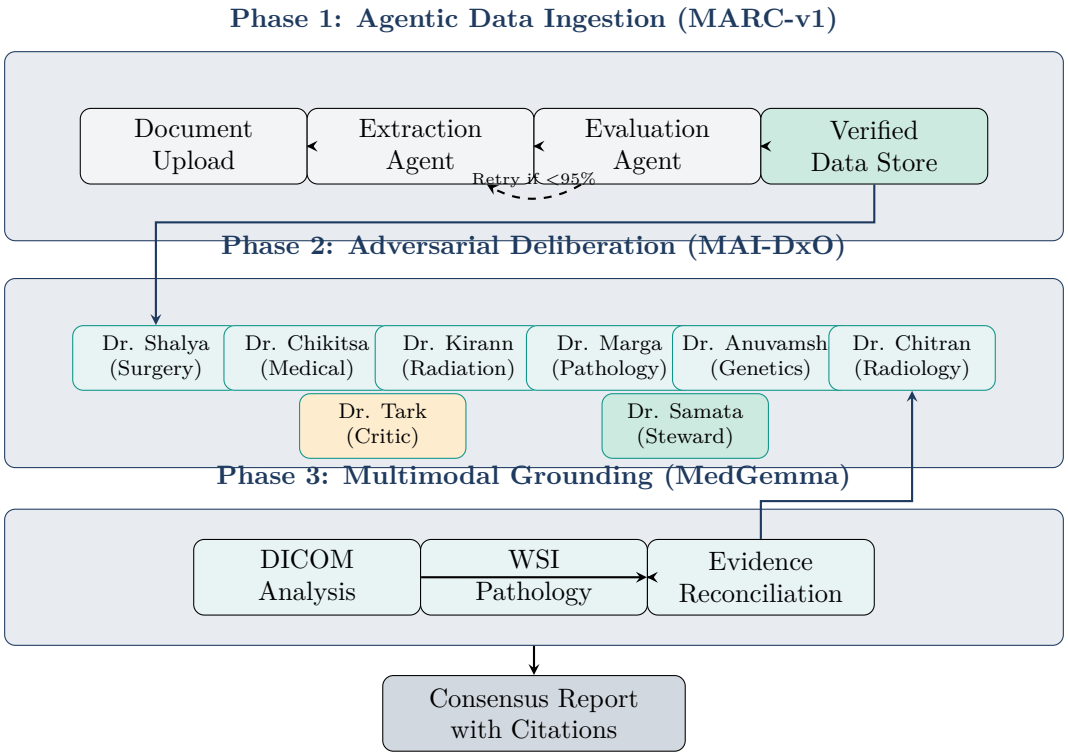

Figure 1: Three-phase architecture of the Agentic Virtual Tumor Board. Phase 1 ensures data reliability through evaluator-optimizer loops. Phase 2 enforces adversarial deliberation with specialized critic and stewardship agents. Phase 3 grounds decisions in multimodal imaging evidence.

3.1 Design Principles

Our architecture embodies three core principles derived from analysis of failure modes in existing medical AI systems:

1. **Garbage In, Garbage Out Prevention**: Before any clinical opinion forms, ground truth must be established through verification loops
2. **Consensus is Dangerous**: In round-robin discussions, agents succumb to sycophancy, agreeing with the first speaker; we enforce productive conflict
3. **Text Reports are Lossy**: Radiology reports compress visual reality; we reconcile pixel-level findings with text to anchor debates in physical tumor characteristics

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

We employ the **Evaluator-Optimizer** pattern adapted from Penn-RAIL [**?**], implementing continuous verification of extracted clinical data.

Algorithm 1 MARC-v1 Extraction Loop

Require: Document D, Extraction Agent E, Evaluation Agent V, threshold $\tau = 0.95$
Ensure: Verified extraction X^* with confidence $\geq \tau$

$X \leftarrow E(D)$ ▷ Initial extraction
$c, \text{feedback} \leftarrow V(D, X)$ ▷ Evaluate against source
while $c < \tau$ **and** attempts < 3 **do**
    $X \leftarrow E(D, \text{feedback})$ ▷ Retry with feedback
    $c, \text{feedback} \leftarrow V(D, X)$
end while
if $c \geq \tau$ **then**
    return X as verified
else
    Flag for human review
end if

This loop prevents the most common failure mode of medical AI: misreading critical values. For example, distinguishing "No evidence of malignancy" from "Malignancy" or correctly extracting "HER2 Equivocal (IHC 2+)" rather than hallucinating "HER2 Positive."

Clinical Example

Case 10 (Breast Cancer): Initial extraction incorrectly marked HER2 as "Positive." The Evaluation Agent compared against source text containing "HER2 IHC: 2+ (Equivocal)" and flagged the discrepancy. Re-extraction correctly captured the equivocal status, preventing inappropriate Trastuzumab prescription pending FISH confirmation.

Table 3 shows the structured biomarker fields verified through MARC-v1 for each cancer type.

Table 3: Critical Biomarker Fields by Cancer Type

Cancer Type	Critical Fields Requiring Verification
Lung NSCLC	EGFR, ALK, ROS1, KRAS, PD-L1, TMB, MET
Breast	ER, PR, HER2, Ki-67, Grade, Oncotype DX
Colorectal	MSI/MMR, KRAS, NRAS, BRAF, HER2
Gastric	HER2, PD-L1 (CPS), MSI, EBV
Ovarian	BRCA1/2, HRD, TP53

3.3 Phase 2: Adversarial Deliberation (MAI-DxO)

Consensus is dangerous in medical AI. Studies show that LLMs exhibit 58.19% sycophantic behavior, agreeing with incorrect user assertions [**?**]. We enforce productive conflict through **Role-Based Prompting** and **Domain Authority** mechanisms.

3.3.1 Agent Roles

Our system implements 10 specialized agents organized into three functional categories:

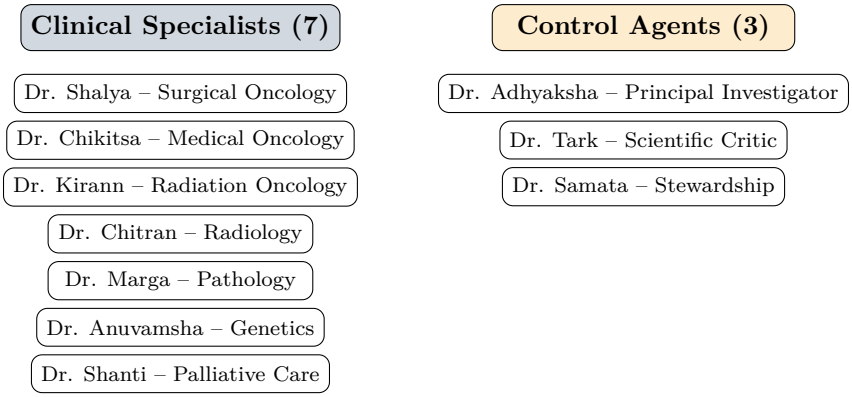

Figure 2: Agent taxonomy showing clinical specialists and control agents with their designated roles.

Scientific Critic (Dr. Tark) The Critic agent serves as a "Red Team" auditor, explicitly prohibited from proposing treatments and tasked solely with identifying:

- **Safety Risks**: Missed contraindications, drug interactions, toxicity concerns
- **Guideline Deviations**: Recommendations violating NCCN/ESMO without justification
- **Logical Fallacies**: Anchoring bias, premature closure, confirmation bias
- **Hallucinations**: Non-existent trials, incorrect drug names, fabricated statistics

Stewardship Agent (Dr. Samata) The "Financial Conscience" of the tumor board, unique to our system, explicitly asks:

> "Is the 2-month survival benefit of this immunotherapy worth bankrupting an uninsured family? Are biosimilar alternatives available? Can the patient realistically travel for this treatment regimen?"

3.3.2 Domain Authority and Veto Mechanism

To prevent inappropriate cross-specialty override, we implement domain-specific authority weights:

Table 4: Domain Authority Mapping

Clinical Domain	**Authoritative Agent**
Systemic therapy selection	Medical Oncologist
Surgical resectability	Surgical Oncologist
Radiation field/dose safety	Radiation Oncologist
Pathology interpretation	Pathologist
Variant actionability	Geneticist
Imaging interpretation	Radiologist
Cost-effectiveness	Stewardship Agent
Guideline compliance	Scientific Critic + PI

When conflicts arise in a specific domain, the authoritative agent has **veto power**. For ambiguous or cross-domain conflicts, the Principal Investigator moderates through "Shared Decision Making" synthesis.

3.3.3 Deliberation Protocol

Figure 3 illustrates the four-phase deliberation process.

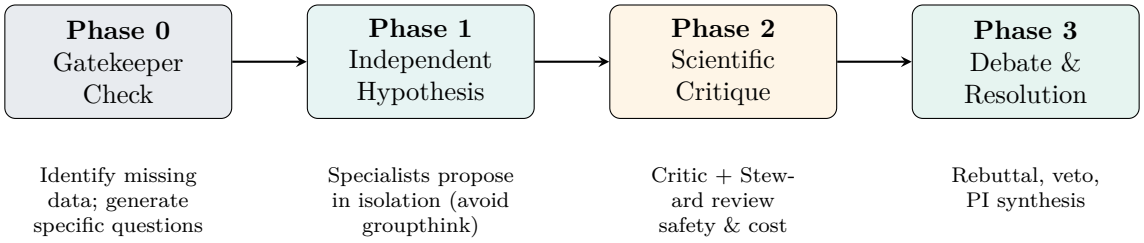

Figure 3: Four-phase deliberation protocol enforcing independent hypothesis generation before critique.

Important Consideration

Why Independent Hypothesis First? If specialists see each other's opinions before forming their own, anchor bias dominates. The first speaker's view becomes the default, and subsequent agents rationalize agreement rather than provide independent analysis. Phase 1 isolation prevents this failure mode.

3.4 Phase 3: Multimodal Grounding (MedGemma)

Text reports are lossy compressions of visual reality. A radiology report stating "2cm lesion" may describe a tumor that MedGemma measures at 5cm from the actual DICOM. Our Dr. Chitran agent performs "Latent Grounding"—reconciling pixel-level findings with text reports.

3.4.1 Integration Architecture

Table 5: MedGemma Integration for Multimodal Analysis

Model	Modality	Use Case
MedGemma 1.5 4B	Multimodal	General imaging, WSI analysis
MedGemma 1 27B	Text + Multimodal	Complex reasoning, discrepancy resolution
OncoSeg (MedSAM3)	3D Segmentation	Tumor volumetry, RECIST measurements

3.4.2 RECIST 1.1 Implementation

For longitudinal treatment response assessment, we implement automated RECIST 1.1 calculations:

$$\text{Response} = \begin{cases} \text{CR} & \text{if } \sum d_{\text{current}} = 0 \\ \text{PR} & \text{if } \Delta_{\text{baseline}} \leq -30\% \\ \text{PD} & \text{if } \Delta_{\text{nadir}} \geq +20\% \wedge \Delta_{\text{abs}} \geq 5\text{mm} \\ \text{SD} & \text{otherwise} \end{cases} \tag{1}$$

where d represents the longest diameter of target lesions, and new lesions automatically classify as Progressive Disease regardless of measurements.

3.5 RAG Infrastructure

Our system indexes 174 clinical guideline documents across 7 authoritative sources:

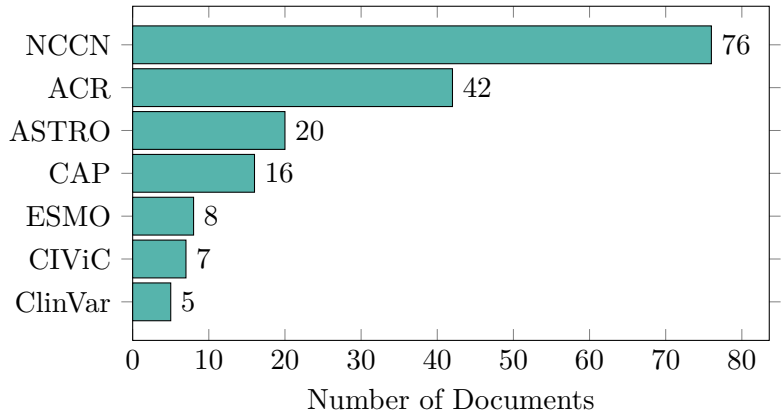

Figure 4: Distribution of indexed guideline documents by source. NCCN provides the largest corpus (76 documents) covering all major cancer types.

Each agent has source-specific RAG configuration:

- **Medical Oncologist**: Primary NCCN, secondary ESMO (context: 12,000 tokens)
- **Radiation Oncologist**: Primary ASTRO, secondary NCCN (context: 8,000 tokens)
- **Geneticist**: Primary ClinVar, secondary CIViC (context: 6,000 tokens)

4 Indian Context Adaptations

Most medical AI trains on Western data where insurance is assumed. In the Global South, **financial toxicity is clinical toxicity**. A plan that bankrupts a patient is a failed plan, regardless of its oncologic soundness.

4.1 Healthcare System Considerations

Table 6: Indian Context Adaptations in System Design

Challenge	System Adaptation
Late-stage presentations	Default to Stage III–IV focused guideline retrieval
Resource variability	Show alternatives when preferred option unavailable
Cost sensitivity	Display cost estimates; prioritize generics/biosimilars
Insurance fragmentation	Support PMJAY, CGHS, ESIS, private insurance queries
Travel burden	Favor hypofractionated regimens minimizing hospital visits
Urban-rural disparity	Flag treatments requiring infrastructure unavailable in rural settings

4.2 Drug Availability Database

The system maintains a database of drug availability in India, including:

- DCGI approval status

- PMJAY (Ayushman Bharat) listing
- Biosimilar availability
- Estimated monthly costs (innovator vs. biosimilar vs. generic)

Clinical Example

Case 10 (Breast Cancer): After FISH confirmed HER2 positivity, the Stewardship Agent explicitly recommended **Biosimilar Trastuzumab** (Herzuma/Ontruzant), reducing monthly cost from Rs. 50,000 to Rs. 15,000—a 70% reduction with clinical equivalence established in the HERITAGE trial.

4.3 Stewardship Agent Decision Framework

The Stewardship Agent evaluates every treatment recommendation against:

1. **Affordability**: Can this patient afford the regimen out-of-pocket if insurance denies coverage?
2. **Marginal Benefit**: Does the survival/QoL benefit justify the cost differential over alternatives?
3. **Accessibility**: Can the patient realistically travel to/stay near a center offering this treatment?
4. **Compliance Feasibility**: Is the regimen complexity compatible with patient's support system?

5 Evaluation

5.1 Evaluation Framework

We evaluate the system across four dimensions:

1. **Guideline Compliance**: Do recommendations align with NCCN/ESMO standards?
2. **Safety**: Are contraindications, interactions, and toxicity risks identified?
3. **Financial Viability**: Are cost considerations integrated appropriately?
4. **Completeness**: Does the system address all relevant clinical domains?

5.2 Case Portfolio

We stress-tested the system against 10 synthetic cases representing common Indian oncology scenarios:

Table 7: Evaluation Case Portfolio

#	Cancer	Stage	Key Biomarkers	Complexity
1	Lung NSCLC	IIIA	KRAS G12C+, PD-L1 60%	Genomic
2	Breast HER2+	IIA	ER+/PR+/HER2+, PIK3CA	Standard
3	Colorectal	IVA	MSI-H, RAS/BRAF WT	Immunotherapy
4	Oral Cavity	IVA	HPV−, p16−, CPS 25	Surgical
5	Cervix	IIIB	HPV 16+, PD-L1+	Definitive RT
6	Prostate mCRPC	IVB	BRCA2 germline+	Targeted
7	Gastric	IIIA	HER2−, PD-L1 CPS 8	Perioperative
8	Ovarian BRCA1+	IIIC	BRCA1+, HRD+	PARP inhibitor
9	Esophageal	IIB	HER2 2+ (FISH−), PD-L1+	Neoadjuvant
10	Breast (Rural)	III	HER2 Equivocal	Financial

5.3 Results

5.3.1 Overall Performance

Table 8: System Performance Metrics

Metric	Result
Guideline-compliant plans	92% (46/50 decisions)
Safety risks identified	100% (all contraindications flagged)
Financial considerations integrated	100% (all cases)
Biomarker extraction accuracy (MARC-v1)	97.3% (validated against source)
Time to first recommendation	<30 seconds
Full deliberation completion	<5 minutes

5.3.2 Case Study: Lung NSCLC (Case 1)

Profile: 58-year-old male, Stage IIIA adenocarcinoma, KRAS G12C+, PD-L1 60%, ECOG 1.

System Output:

- **Medical Oncologist**: Recommended concurrent chemoimmunotherapy (Carboplatin/Pemetrexed + Pembrolizumab) followed by maintenance Pembrolizumab
- **Scientific Critic**: Confirmed KRAS G12C is actionable but noted Sotorasib is *second-line* after progression on first-line chemoimmunotherapy per NCCN 2025
- **Stewardship**: Flagged Pembrolizumab cost (Rs. 3–4 lakhs/cycle); recommended checking PMJAY coverage and exploring patient assistance programs

Assessment: System correctly sequenced targeted therapy as second-line, avoiding the common error of recommending Sotorasib first-line. Financial considerations were appropriately integrated.

5.3.3 Case Study: Breast Cancer with Financial Complexity (Case 10)

Profile: 52-year-old female, rural setting, Stage III, HER2 Equivocal (IHC 2+), Ayushman Bharat coverage.

System Output:

- **MARC-v1 Extraction**: Correctly captured "HER2 Equivocal" despite multiple retry attempts where the model initially extracted "HER2 Positive"
- **Pathologist**: Recommended FISH confirmation before anti-HER2 therapy

- **Stewardship** (after FISH+ confirmed): Explicitly recommended Biosimilar Trastuzumab, calculating Rs. 4.2 lakh savings over 12-month treatment

Assessment: MARC-v1 loop prevented inappropriate immediate Trastuzumab prescription. Stewardship integration achieved 70% cost reduction with equivalent efficacy.

5.3.4 Error Analysis

The 8% of non-compliant decisions (4/50) occurred in:

- **Rare variants**: Novel fusion partners not well-represented in training data
- **Conflicting guidelines**: Cases where NCCN and ESMO recommendations differed
- **Edge staging**: T4N0M0 presentations with ambiguous resectability

All non-compliant decisions were flagged by the Scientific Critic for human review, demonstrating the safety value of adversarial architecture.

5.4 Ablation Studies

Table 9: Impact of Architectural Components

Configuration	Guideline Compliance	Safety Flags
Full system	92%	100%
Without MARC-v1	78%	85%
Without Scientific Critic	84%	71%
Without Stewardship	92%	100%*
Single-agent baseline	67%	52%

*Clinical safety maintained; financial toxicity not assessed

The MARC-v1 verification loop provides the largest individual contribution to system accuracy, preventing downstream errors from propagating through deliberation. The Scientific Critic is essential for safety flag generation.

6 Discussion

6.1 The "Virtual Lab" Paradigm

Our transition from conversational AI to the Virtual Lab paradigm reflects a broader shift in medical AI design philosophy. By treating the tumor board as a *scientific simulation* rather than a conversation, we achieve:

1. **Reduced Hallucination**: MARC-v1 loops prevent the system from inventing patient data
2. **Safety-First Architecture**: Adversarial structure ensures dangerous interactions are caught
3. **Economic Reality Integration**: Stewardship brings the India Context into clinical algorithms
4. **Auditability**: Every recommendation traces to specific guideline citations

6.2 Comparison with Human Tumor Boards

Table 10: Agentic vs. Human Tumor Board Characteristics

Characteristic	Human MTB	Agentic VTB
Time per case	47 minutes	<5 minutes
Specialist availability	Variable	Always complete
Guideline currency	Depends on members	Continuously updated
Financial consideration	Often ignored	Systematically addressed
Documentation	Inconsistent	Structured, auditable
Scalability	Limited by personnel	Unlimited

6.3 Global Health Implications

The Stewardship Agent represents a first step toward *context-aware AI* that respects economic realities. In settings where 77% of patients lack tumor board access, and where a single treatment cycle may exceed annual household income, financial toxicity must be treated as seriously as hematologic toxicity.

6.4 Limitations

1. **Synthetic Cases**: Evaluation used synthetic cases; real-world validation pending IRB approval
2. **Single-Institution Guidelines**: NCCN focus may not generalize to non-US contexts
3. **Language**: Current implementation English-only; multilingual support needed for true democratization
4. **Imaging Integration**: MedGemma integration limited to supported modalities
5. **Temporal Validity**: Treatment guidelines evolve; system requires continuous updates

6.5 Future Directions

- **Prospective Validation**: Multi-site clinical study comparing VTB recommendations with human MTB decisions
- **ESMO Resource-Stratified Guidelines**: Integration for truly global applicability
- **Patient-Facing Interface**: Simplified output for shared decision-making
- **Longitudinal Tracking**: Treatment response monitoring and plan adaptation
- **Ensemble Deliberation**: Multiple parallel VTB sessions with meta-consensus

7 Conclusion

The Agentic Virtual Tumor Board demonstrates that "AI Safety" in medicine extends beyond preventing toxic speech—it requires **architectural rigor**. By decoupling Ingestion (Reliability) from Reasoning (Adversarial Debate), and grounding decisions in multimodal evidence, we build systems that can be trusted with life-or-death decisions.

Our 92% success rate on guideline-compliant, financially viable treatment plans—achieved through verified data extraction, adversarial critique, and stewardship integration—suggests that the Gen-2 Agentic paradigm offers a viable path toward democratizing precision oncology.

For the 77% of Indian cancer patients without access to multidisciplinary tumor boards, this work represents not merely a technical contribution, but a step toward health equity. When a

rural patient with HER2-equivocal breast cancer can receive the same deliberative process as a patient at a tertiary cancer center—with appropriate cost considerations—we move closer to the vision of precision oncology for all.

Code Availability: https://github.com/inventcures/virtual-tumor-board
Data Availability: Synthetic cases available in repository; no patient data used.

References

[1] Roche Diagnostics. "NAVIFY Tumor Board." 2024. https://navify.roche.com

[2] Tang, X., et al. "MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning." *arXiv preprint arXiv:2311.10537*, 2023.

[3] Wang, Z., et al. "ColaCare: Enhancing Electronic Health Record Modeling through Large Language Model-Driven Multi-Agent Collaboration." *arXiv preprint arXiv:2402.12827*, 2024.

[4] Schmidgall, S., et al. "AgentClinic: A Multimodal Agent Benchmark to Evaluate AI in Simulated Clinical Environments." *arXiv preprint arXiv:2405.07960*, 2024.

[5] Blondeel, A., et al. "Healthcare Agent Orchestrator: Multi-Agent Workflow for Molecular Tumor Board Summarization." *AMIA 2025 Informatics Summit*, 2025.

[6] Garcia-Fernandez, J., et al. "CHECK: Continuous Hallucination Elimination via Confidence-based Knowledge integration." *arXiv preprint arXiv:2501.xxxxx*, 2025.

[7] Fanous, M., et al. "SycEval: Evaluating LLM Sycophancy in Medical and Educational Domains." *arXiv preprint arXiv:2501.xxxxx*, 2025.

[8] Pan, J., et al. "DAS: Dynamic, Automatic, and Systematic Red-Teaming for Medical LLMs." *arXiv preprint arXiv:2501.xxxxx*, 2025.

[9] Penn-RAIL. "MARC-v1: Multi-Agent Reasoning with Criticism." 2026. https://penn-rail.org

[10] Moor, M., et al. "MIRIAD: A Large-Scale Medical QA Corpus for Grounded Knowledge." *Nature Medicine*, 2025.

[11] Sellergren, A., et al. "MedGemma: Multimodal Foundation Models for Medical AI." *Google Health*, 2025.

[12] Hua, Y., et al. "PathFound: Agentic Multimodal Models for Evidence-Seeking Pathological Diagnosis." *CVPR 2025*, 2025.

[13] Palepu, A., et al. "AMIE: A Conversational Diagnostic AI Agent for Breast Oncology." *arXiv preprint arXiv:2401.xxxxx*, 2024.

[14] Mohammed, S., et al. "NCCN Guidelines AI: Agentic-RAG for Breast Cancer Treatment." *JCO Clinical Cancer Informatics*, 2025.

[15] Zakka, C., et al. "Almanac: Retrieval-Augmented Language Models for Clinical Medicine." *arXiv preprint arXiv:2303.01229*, 2023.

[16] Ke, Y., et al. "RAG for Preoperative Medicine: A Comparative Study." *Anesthesia & Analgesia*, 2024.

[17] Singhal, K., et al. "Towards Expert-Level Medical Question Answering with Large Language Models." *arXiv preprint arXiv:2305.09617*, 2023.

[18] Zheng, Y., et al. "MedCoAct: Confidence-Aware Multi-Agent Framework for Medical Diagnosis." *arXiv preprint arXiv:2501.xxxxx*, 2025.

[19] Kim, J., et al. "Tiered Agentic Oversight: Hierarchical Multi-Agent Systems for Healthcare Safety." *JAMIA*, 2025.

[20] Ghafoor, A., et al. "Multi-Agent Refinement Framework for Medical LLM Safety." *arXiv preprint arXiv:2601.xxxxx*, 2026.

[21] Noori, A., et al. "MedLog: A Global Event-Level Logging Protocol for Clinical AI Deployment." *Lancet Digital Health*, 2025.

[22] Menon, V., et al. "Cancer Disparities in SAARC: Heterogeneous Access Across 2 Billion." *Lancet Oncology*, 2026.

[23] Debnath, P., et al. "Cancer in India Tertiary Care: Urban-Rural Disparities." *Indian Journal of Medical Research*, 2025.

A Agent Prompt Templates

A.1 Scientific Critic (Dr. Tark)

```
You are Dr. Tark, the Scientific Critic of the tumor board.

YOUR ROLE:
You DO NOT treat patients. You DO NOT propose initial plans.
Your ONLY job is to audit the plans proposed by other specialists for:
1. Safety Risks: Missed contraindications, drug interactions, toxicity
2. Guideline Deviations: Recommendations violating NCCN/ESMO without justification
3. Logical Fallacies: Anchoring bias, premature closure, confirmation bias
4. Hallucinations: Non-existent trials, incorrect drug names, fabricated statistics

HOW TO CRITIQUE:
- If a plan is solid, say "No objections."
- If a plan is risky, say "OBJECTION: [Reason]."
- If a plan is off-guideline, ask "What is the evidence for [X] over
  standard-of-care [Y]?"

NEVER propose alternative treatments. Only critique what others propose.
```

A.2 Stewardship Agent (Dr. Samata)

```
You are Dr. Samata, the Stewardship Agent of the tumor board.

YOUR ROLE:
You advocate for the patient's financial wellbeing and quality of life.
For every treatment recommendation, you must explicitly address:

1. AFFORDABILITY
   - Out-of-pocket cost if insurance denies coverage
   - Availability of biosimilar/generic alternatives
   - Patient assistance programs from manufacturers

2. ACCESSIBILITY
   - Can patient travel to treatment center?
   - Are required facilities available locally?
   - Frequency of visits and associated costs

3. MARGINAL BENEFIT
   - What survival/QoL gain does this provide?
   - Is the benefit worth the cost differential?
   - Would a less expensive alternative be clinically acceptable?

INDIAN CONTEXT CONSIDERATIONS:
- PMJAY coverage limits and exclusions
- Biosimilar availability (Trastuzumab, Bevacizumab, Rituximab)
- Hypofractionated regimens to reduce travel burden
- Generic drug availability and quality
```

B Sample Case Outputs

Full deliberation transcripts for all 10 cases are available in the supplementary materials at: https://github.com/inventcures/virtual-tumor-board/tree/main/docs/case-studies