TeamMedAgents: Consolidated Review Summary

Ratings: 4 (Reject) | 7 (Accept) | 5 (Marginal) | 4 (Reject) | AI (Critical) **Outcome:** 3 rejections, 1 marginal, 1 accept

---

CRITICAL ISSUES

1. Statistical Validity (All reviewers except R2)

Problems:

- Only 50 questions/config insufficient for conclusions with ±3-5% variance
- **Data discrepancy:** Main paper claims PubMedQA 76.6%, appendix shows 68-70%; MedQA 92.6% vs 91.3-91.6%
- No significance testing despite overlapping error bars
- Appendix shows extreme instability (±15% SE, MDAgents 75%→56% performance swings)
- Some accuracies mathematically impossible (not multiples of 2/150)

Required: Reconcile discrepancies, run full test sets, add bootstrap/McNemar tests

2. Unfair Baselines (R1, R3, R4, AI)

Problems:

- MDAgents uses GPT-4, yours uses GPT-4o (confounds comparison)
- Missing: multi-agent debate, MedAgents, MMedAgent, MedChat, ColaCare, self-consistency baselines, ensemble methods

Required: Re-run all with GPT-4o, add competitive baselines

3. Under-Specified Methodology (AI, R3)

Missing details preventing replication:

- Trust update function (formula, bounds, decay)
- Aggregation formula beyond leader weight
- Recruitment policy (features, thresholds)
- Shared mental models (representation, checks)

- Closed-loop triggers/termination
- "Special Set" selection process (no validation set = overfitting risk)

Required: Formulas, pseudocode, validation methodology

4. Unsupported "Optimal" Claims (AI, R1)

Problems:

- Table 2 lists configs without numeric results
- Contradictions: MMLU-Pro single component 84% > TeamMedAgents 82%; PathVQA 76% > 74.67%
- MedQA Table 2 says "Shared Mental Model" but Table 1 shows "Team Orientation" highest

Required: Report all Table 2 numbers, explain contradictions

MAJOR CONCERNS

5. Novelty & Positioning (R4, AI)

- Concept already explored (MedAgents, MedChat, MMedAgent-RL)
- Incomplete related work discussion
- Framework appears similar to MDAgents
- Need clear differentiation

6. Limited Evaluation (All reviewers)

- Only static QA, no clinical validation
- One main table insufficient
- Missing: error analysis, failure modes, qualitative examples, team size/rounds ablations
- Wrong metrics: VQA needs soft-accuracy, differential diagnosis needs top-k ranking
- No cost/latency data (3 rounds × 2-5 agents = 6-15× cost)

7. Implementation Gaps (R2, R3, AI)

- Backup behavior & adaptability not implemented
- Prompt details limited (need examples)
- No domain expert validation

- No guardrails for persuasive-but-wrong leaders or trust echo-chambers

---

PRESENTATION ISSUES

8. Clarity Problems (All)

- Figure 1 cluttered, avatar needs fix (R3)
- Dataset name inconsistencies (Medbullets/MedBullets, PMC-VQA/PmcVQA)
- "Tchangho" typo (should be Tchango)
- Narrative contradicts data (claims all 8 datasets improved, actually 7)
- Need sample outputs showing clinical use (R2)

ACKNOWLEDGED STRENGTHS

- Theory-grounded, principled design (all reviewers)
- Modular, interpretable components enabling ablations
- Broad evaluation (8 benchmarks, text + multimodal)
- Key insight: "All Features" not universally optimal
- Notable gains on visual tasks (PathVQA +9.37%)

PRIORITY ACTIONS

CRITICAL (for acceptance):

1. Fix data discrepancies, increase sample size, add significance tests
2. Re-run baselines fairly (same model), add competitive methods
3. Specify all formulas/mechanisms with pseudocode
4. Report Table 2 numbers, explain contradictions

HIGH: 5. Expand related work, clarify novelty vs. existing work 6. Add error analysis, appropriate metrics, cost/latency, ablations 7. Fix figures, typos, inconsistencies, add sample outputs

MEDIUM: 8. Analyze failure modes, add guardrails 9. Add prompt examples, justify missing components

DESIRABLE: 10. Test with clinicians, multi-turn simulations

BOTTOM LINE

Strong conceptual foundation undermined by statistical issues, unfair baselines, under-specification, and contradictory claims. With rigorous revision addressing data integrity, fair comparisons, and complete methodology, this could be impactful.