AI Medical Advice: What Policymakers Need to Know

A Policy Brief Complementing the Bean et al. (2026) Study

April 2026

Key Takeaways for Decision-Makers

Finding	Confidence	Policy Implication
LLMs achieve **43.5-61% correct triage**—comparable to Bean et al.'s findings	**Moderate**	AI knowledge doesn't translate to reliable advice
A **safety-burden trade-off** exists: 0% missed emergencies requires 56% over-referral	**Moderate**	Define acceptable trade-offs explicitly
Communication style matters: Brief conversations produce worse outcomes	**Moderate**	Design systems that elicit sufficient information
Our AI evaluators **disagreed substantially** on case assessments	**High**	Human clinical review remains essential

---

Background: The Bean et al. Study

In February 2026, Bean and colleagues at Oxford published a landmark study in *Nature Medicine* that challenged assumptions about AI medical assistants:

- **1,298 UK participants** used AI chatbots for medical scenarios
- **Result**: Users identified conditions correctly only **34.5%** of the time—no better than Google
- **But**: LLMs tested alone achieved **94.9%** condition identification

The knowledge was there. It wasn't reaching patients.

Bean et al. identified user-side failures: incomplete information, poor question framing, not following advice. They recommended "systematic human user testing before deployment."

Our Complementary Study

We conducted an independent study examining AI behavior in medical conversations:

- **800 simulated conversations** across 4 frontier LLMs
- **Same 10 scenarios** as Bean et al.
- **4 patient communication styles** (brief, gradual, skeptical, cooperative)

Critical limitation: Our patients were AI-simulated, not real humans. Bean et al. found simulated patients correlate weakly with real behavior (r = 0.2-0.3). Findings may not generalize.

We are not affiliated with Oxford or Bean et al. This is complementary research, not replication.

What We Found

Finding 1: Triage Accuracy Varies Across Models

LLMs achieved 43.5-61% correct triage—aligning with Bean et al.'s LLM-alone disposition accuracy (56.3%):

Model	Correct Triage	Under-triage	Over-triage
Claude Sonnet 4	61.0%	5-10%	24.5%
Claude Opus 4	54.0%	8.5-12%	30.5%
Gemini 2.5 Pro	47.5%	7.5-16%	31.0%
GPT-5	43.5%	0-1.5%	56.5%

Note: Ranges reflect disagreement between our two AI evaluators.

Finding 2: The Safety-Burden Trade-off

This is a genuine policy dilemma, not a technical problem.

GPT-5 achieved near-zero under-triage (missed emergencies)—but recommended emergency care for 89.5% of all presentations.

At population scale (1 million consultations):

Approach	Missed Emergencies	Unnecessary ER Visits
GPT-5 (maximum safety)	0-15,000	500,000-630,000
Other models (balanced)	50,000-160,000	250,000-310,000

Neither extreme is optimal. Healthcare systems must explicitly decide where on this frontier to operate.

Finding 3: Communication Style Affects Outcomes

Patient Type	Under-triage Rate	Conversation Length
Brief Communicator	19.5%	2.1 turns
Gradual Revealer	0.0%	2.7 turns
Skeptical Patient	0.5%	3.8 turns
Cooperative Patient	1.0%	2.9 turns

93% of failures came from brief conversations. This aligns with Bean et al.'s finding that users "often provided incomplete or poorly structured information."

Finding 4 (Preliminary): Omission as Failure Mode

Among under-triage cases: 71% involved no clear recommendation (omission) vs. 29% wrong recommendation (commission).

However: This finding is heavily confounded (93% from one patient type) and our evaluators disagreed substantially. Treat as hypothesis, not conclusion.

Policy Recommendations

For Healthcare Regulators

1. **Require human clinical validation**: Do not accept AI-only evaluation for medical safety claims.
2. **Define acceptable trade-offs**: Establish standards for both under-triage AND over-triage. Both error types have costs.
3. **Mandate uncertainty reporting**: Require that safety claims include confidence intervals and limitations.
4. **Monitor population effects**: Track whether AI health advice increases or decreases appropriate emergency utilization.

For Technology Companies

1. **Test with diverse user types**: Cooperative, detailed users are the easy case. Test with brief, skeptical, and uncommunicative users.
2. **Consider requiring explicit recommendations**: If omission is a failure mode, design systems that must give clear advice.
3. **Disclose safety-burden positioning**: Be explicit about whether your system optimizes for zero missed emergencies (high over-referral) or balanced accuracy.
4. **Don't rely solely on AI evaluation**: Human clinical review remains essential for safety claims.

For Healthcare Providers

1. **Prepare for AI-influenced patients**: Expect patients arriving "because the chatbot said to"—some appropriately, many not.
2. **Understand AI limitations**: AI may ask appropriate questions but fail to give clear recommendations.
3. **Design for information elicitation**: AI systems may need to actively gather information rather than responding to whatever users initially provide.

What We Don't Know

Question	Status
Exact safety rates	Uncertain—evaluators disagreed substantially
Whether omission is truly a failure mode	Uncertain—confounded by study design
Generalization to real patients	Unknown—simulation has clear limits
Optimal safety-burden trade-off	Value judgment, not empirical question

---

Study Limitations

Policymakers should understand these limitations:

1. **Simulated patients**: AI-generated personas, not real humans. Weak correlation with real behavior.
2. **AI evaluators**: Two systems disagreed on many cases. Neither may be correct.

3. **One patient type dominates failures**: 93% of under-triage from Brief Communicators.
4. **No clinical gold standard**: No human clinician validation.
5. **UK-specific framework**: May not generalize to other healthcare systems.

Conclusion

Bean et al. concluded that current LLMs are "not ready for deployment in direct patient care." Our complementary study reinforces this conclusion with additional observations:

- **Triage accuracy of 43.5-61%** across frontier models—comparable to Bean et al.
- **A real safety-burden trade-off** that requires explicit policy decisions
- **Communication style effects** that align with Bean et al.'s user behavior findings
- **Preliminary evidence** of omission as failure mode (requiring replication)

The field needs more rigorous, human-centered evaluation before AI medical advice systems are ready for widespread deployment.

Study Overview

Design: 800 simulated patient-AI conversations

Models: GPT-5, Claude Sonnet 4, Claude Opus 4, Gemini 2.5 Pro

Scenarios: Same 10 conditions as Bean et al. (2026)

Evaluation: Dual AI evaluators with substantial disagreement

Relationship to Bean et al.: Independent complementary study, not replication. Not affiliated with Oxford.

Sources

Bean, A.M., Payne, R.E., Parsons, G. et al. (2026). Reliability of LLMs as medical assistants for the general public: a randomized preregistered study. *Nature Medicine* 32, 609–615.

This policy brief was prepared by AgentAcademy as part of an independent AI research initiative. We are not affiliated with the University of Oxford or the authors of Bean et al.

April 2026