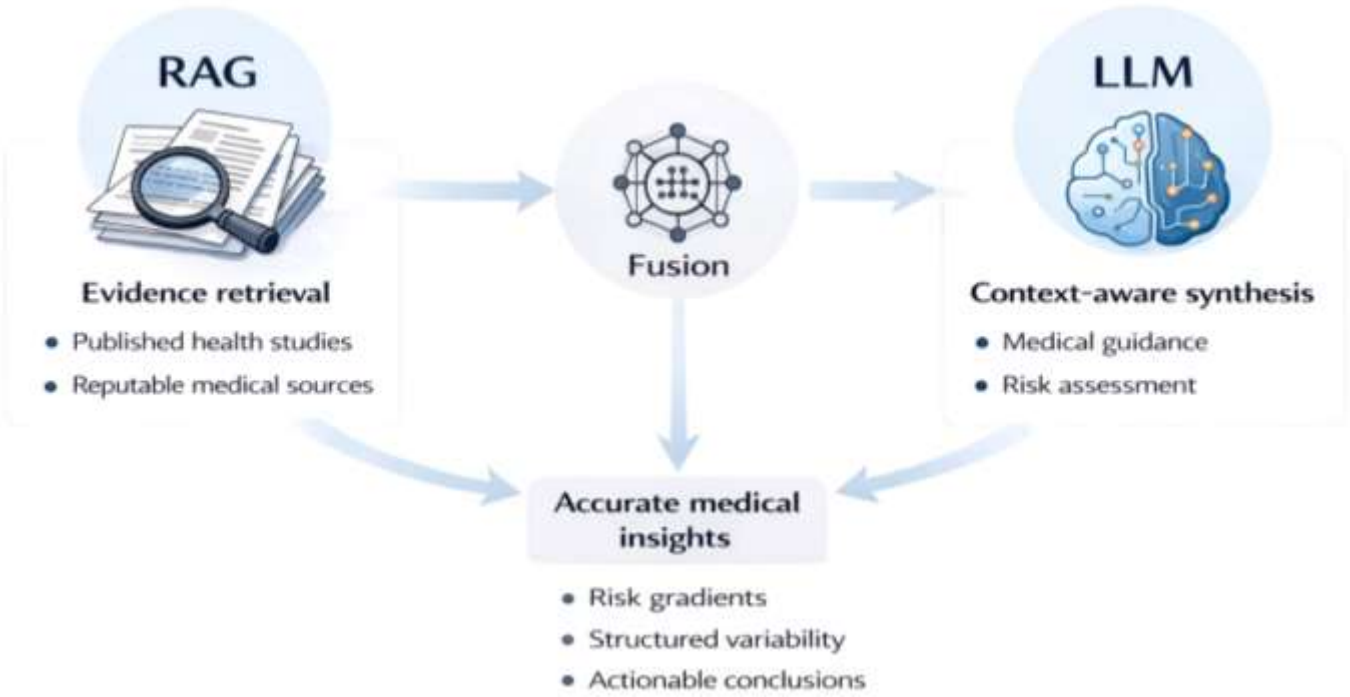
Evaluation Framework for Medical and Research Applications



Question

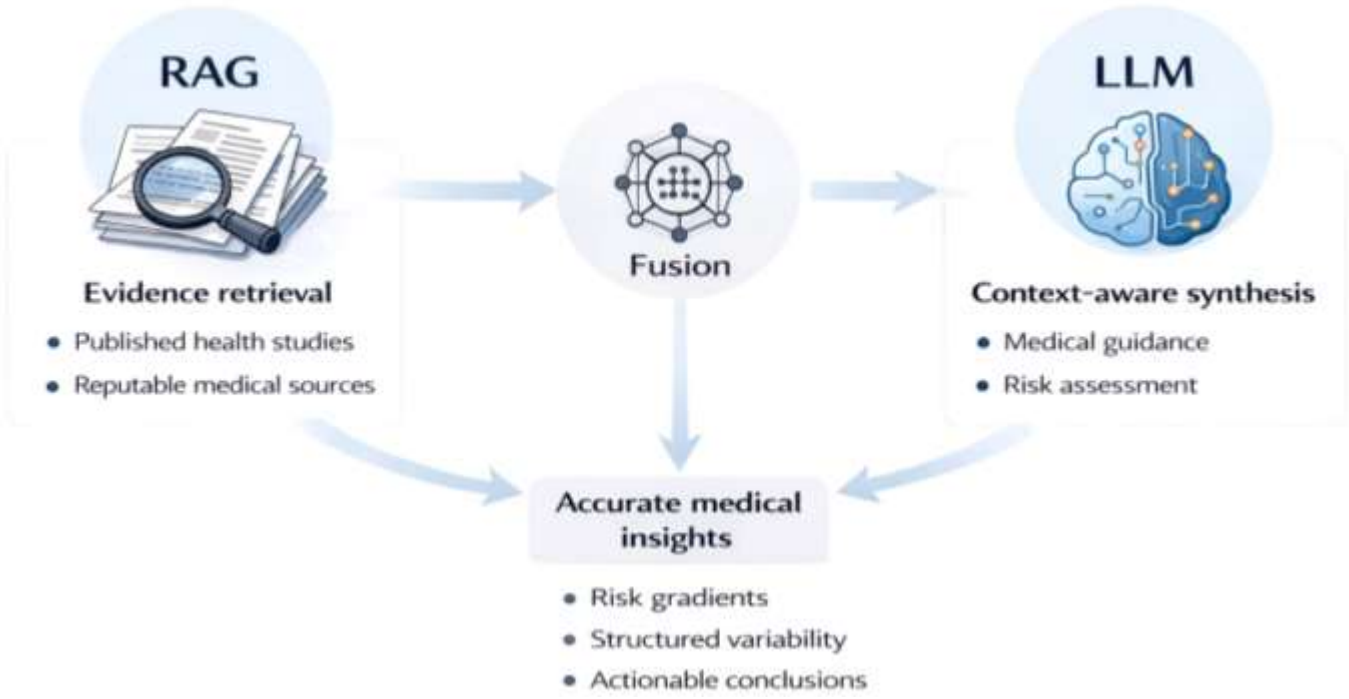
1. **Is "sounding right" enough in high-stakes domains?**
2. Where Are We Now?
3. If conclusions agree, does the reasoning still matter?
4. How much does retrieval alone improve accuracy?

Why Research & Medical Use case ?

- Failure cost increases across domains.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Evaluation Framework for Medical and Research Applications

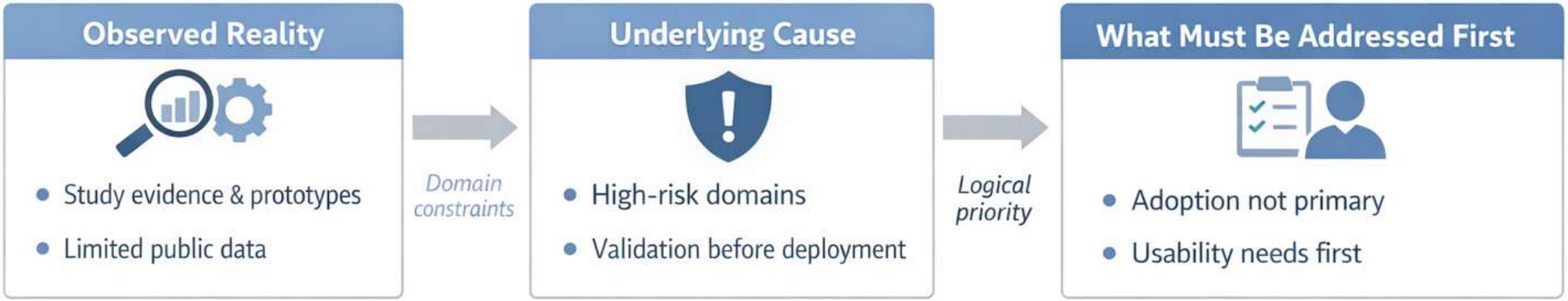


Question

1. Is “sounding right” enough in high-stakes domains?
2. **Where Are We Now?**
3. If conclusions agree, does the reasoning still matter?
4. How much does retrieval alone improve accuracy?

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

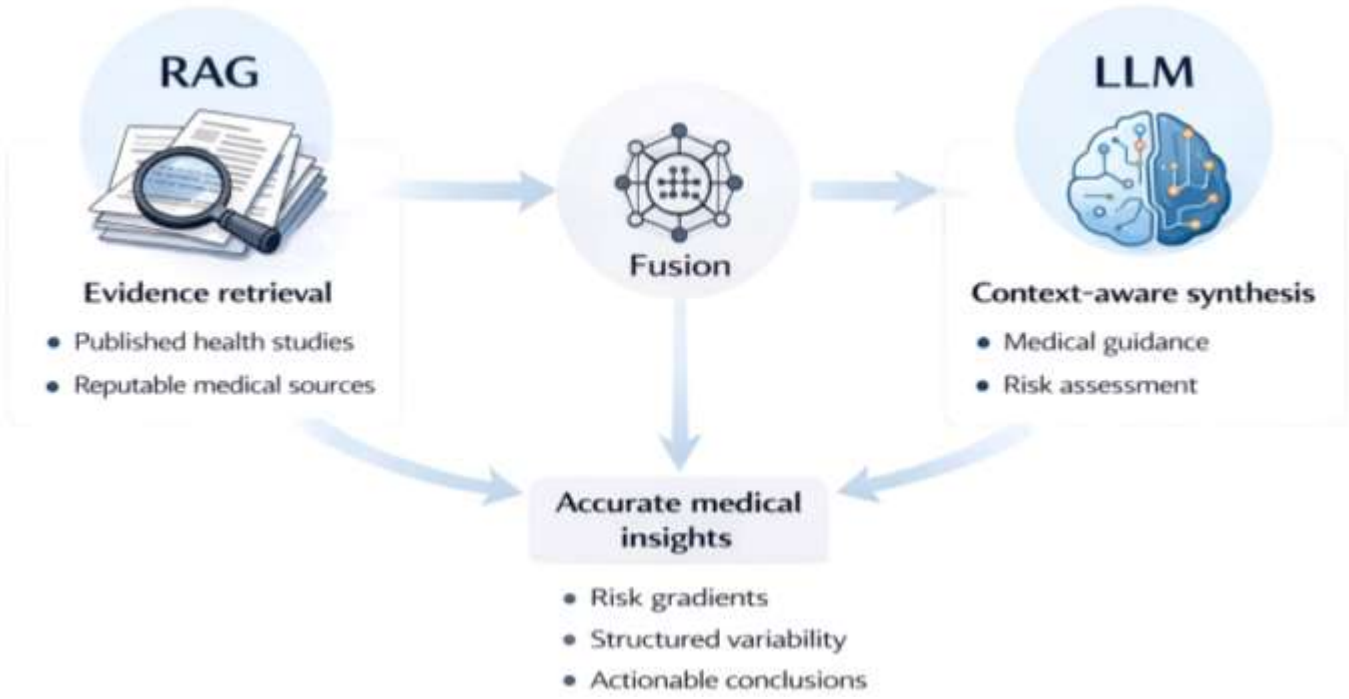
Why isn't this widely used yet?



> Given limited real-world deployment, this work focuses on clarifying what is necessary before adoption.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Evaluation Framework for Medical and Research Applications



Question

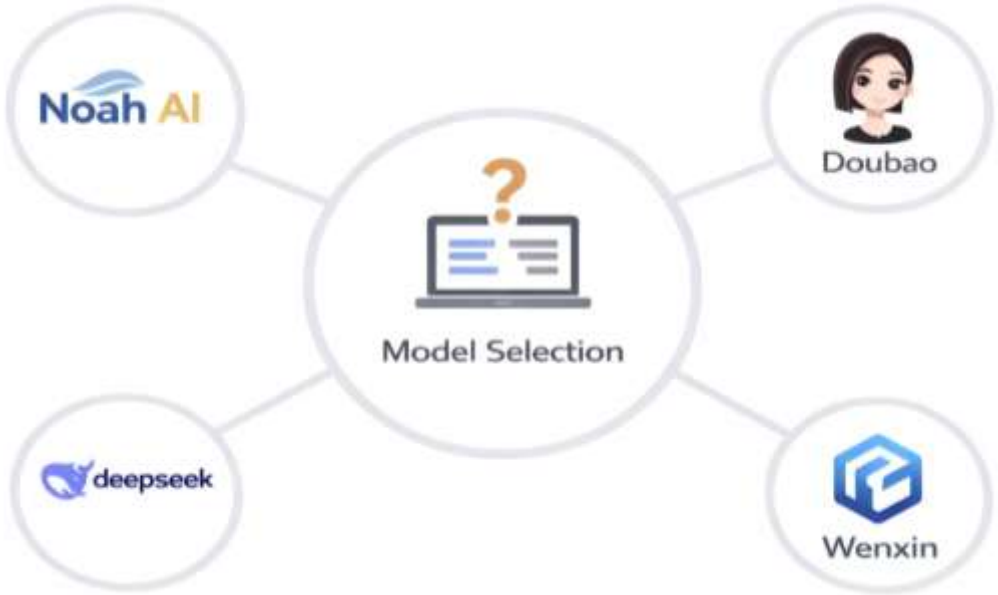
1. Is "sounding right" enough in high-stakes domains?
2. Where Are We Now?
3. **If conclusions agree, does the reasoning still matter?**
4. How much does retrieval alone improve accuracy?

Question

- ✓ Differences in residents' physical health status after COVID-19 infection: a comparison between those who were infected and those who were not.

Result

- ✓ There **is no substantive contradiction** in conclusions across the four models.



Shared conclusion:

COVID-19–infected residents consistently show worse post-acute physical health outcomes (e.g., fatigue, respiratory symptoms, reduced functional capacity), while non-infected residents are mainly affected by indirect pandemic effects.

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Evidence Anchoring Comparison (Traceability)

Traceable evidence

- Explicit study linkage
- Quantified outcomes (RR / *excess %*)
- Claim → Paragraph → Publication

Traceable evidence

Plausible but unverifiable

- Authority-style language
- Vague quantifiers (*"studies suggest..."*)
- No direct citations

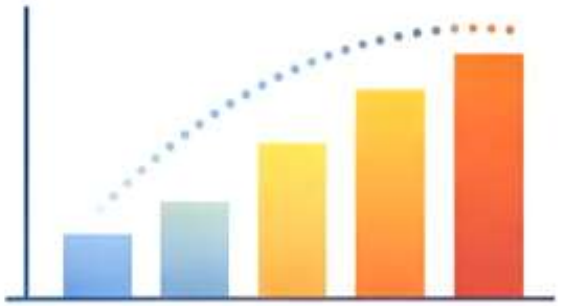
Plausible but unverifiable

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Evidence Anchoring Comparison (Traceability)

Structured variability

- Subgroup differentiation
- Time & severity thresholds
- Explicit uncertainty separation



Risk-weighted evidence

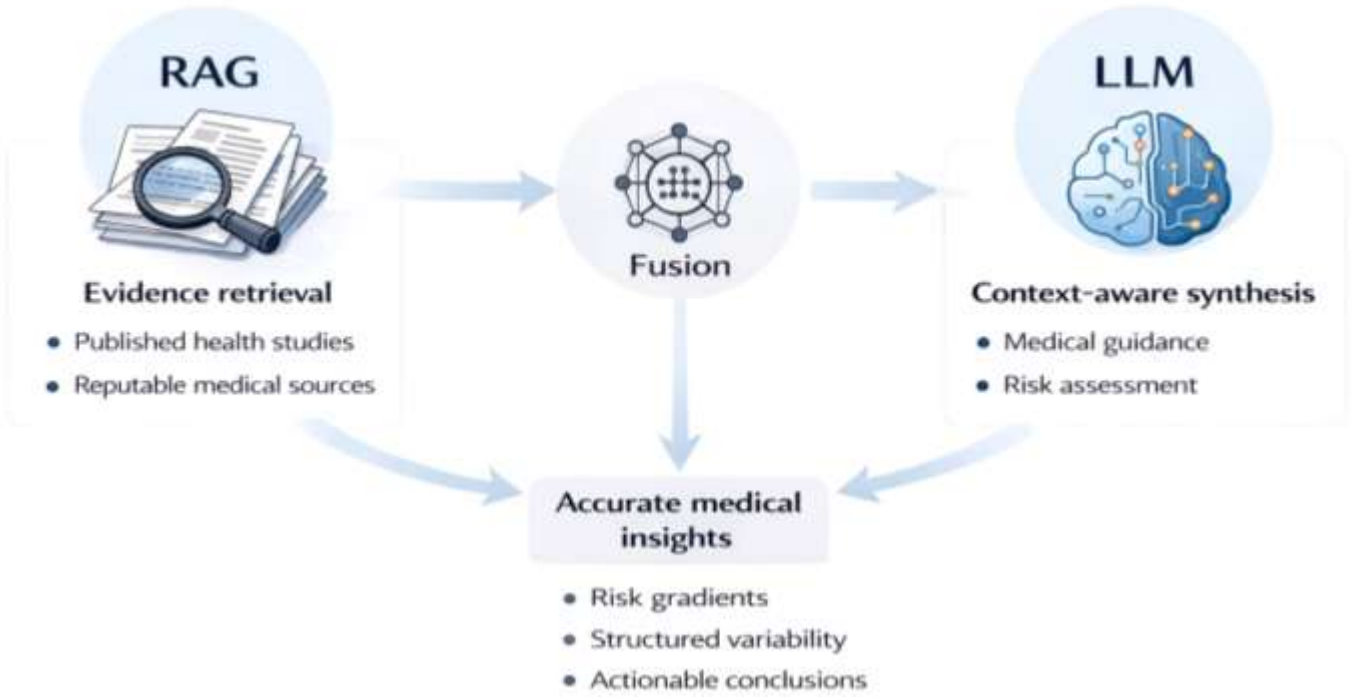
Flattened story–flattened

- Collapses into “all infected”
- No uncertainty separation
- Smooth but misleading coherence

Oversimplified narrative

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

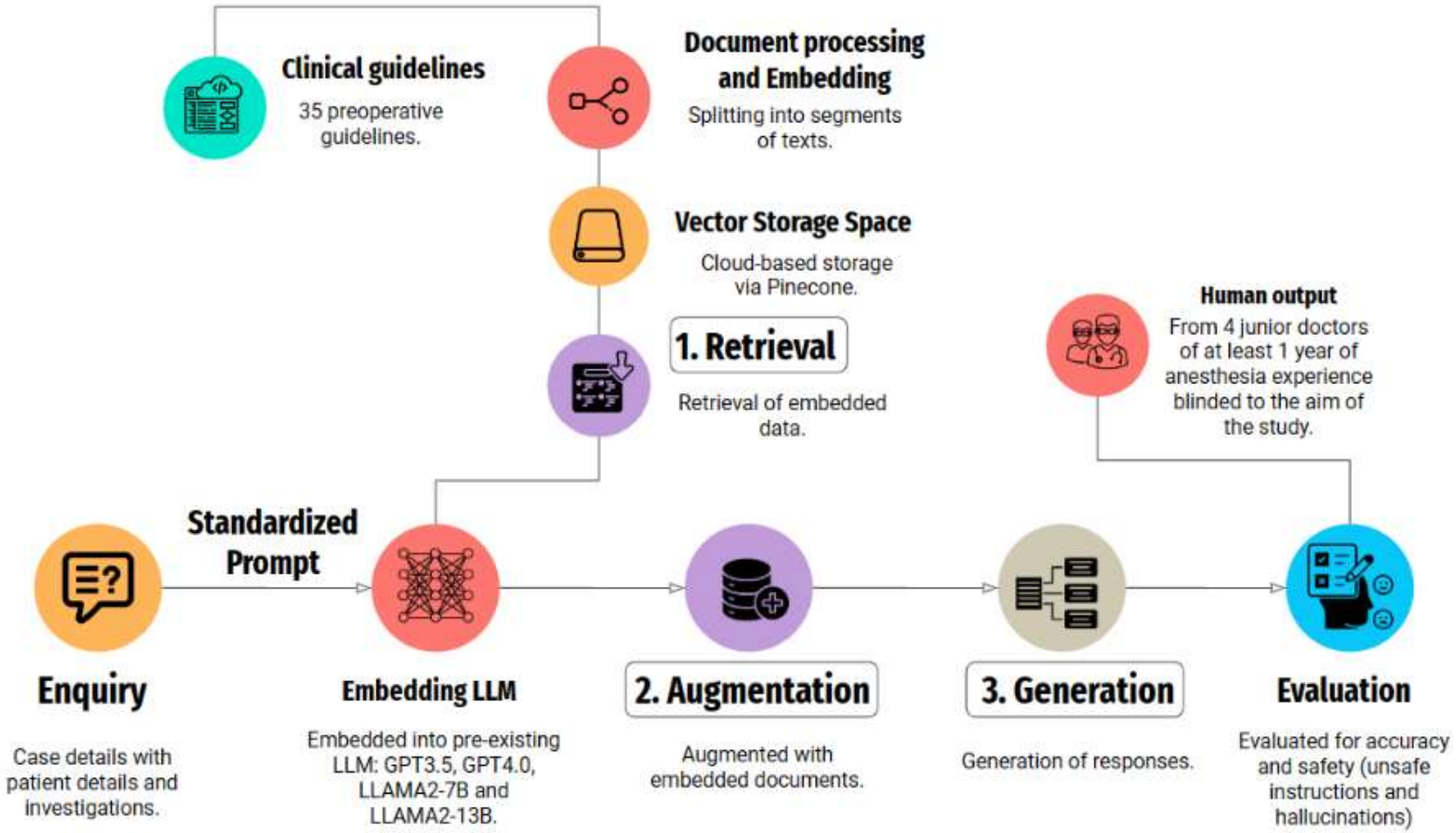
Evaluation Framework for Medical and Research Applications



Question

1. Is “sounding right” enough in high-stakes domains?
2. Where Are We Now?
3. If conclusions agree, does the reasoning still matter?
4. **How much does retrieval alone improve accuracy?**

Quantitative Evidence: Experimental Setup for Evaluating RAG



Source: Ke et al., 2024

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Quantitative Results: Accuracy Comparison (LLM vs. RAG)

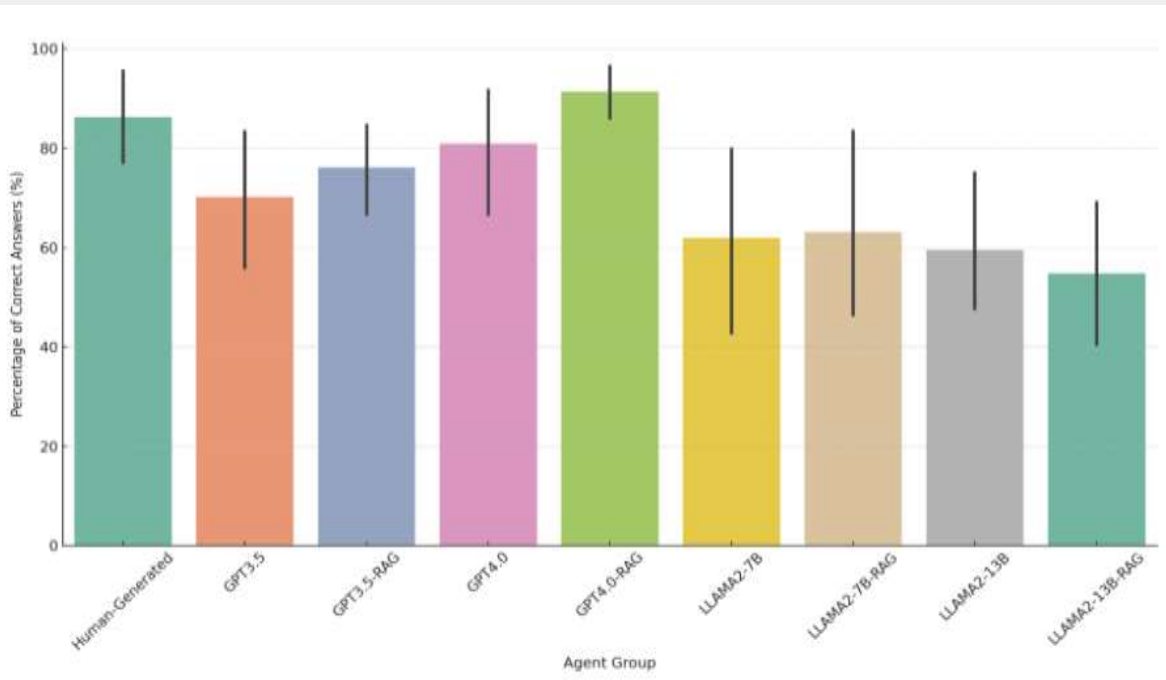


Figure 2: Percentage of accurate answers across the different groups.

Retrieval consistently improves answer accuracy.

- Standalone LLM ≈ 80% accuracy
- LLM + RAG ≈ 91% accuracy
- Comparable to junior doctors

→ Accuracy gains are attributable to retrieval alone.

Source: Ke et al., 2024

Summary

- In high-stakes domains, reliability matters more than plausibility.
- Models may agree on conclusions, but differ in epistemic standards.
- RAG doesn’ t change answers — it makes reasoning accountable.