DeLTA 2026 · 7th Int'l Conference on Deep Learning Theory & Applications · SHORT PAPER · POSTER

A Decision Pipeline for RAG Architecture Selection in Medical Education

Empirical Analysis and Cost–Performance Guidelines

Jovana Dobreva[1,3], Tadej Horvat[2], Matjaž Gams[2], Igor Mishkovski[1], Kostadin Mishev[1], Monika Simjanoska Misheva[1]

[1]Ss. Cyril and Methodius University, Skopje • [2]Jožef Stefan Institute, Ljubljana • [3]Aalborg University, Aalborg

Scan → RAGCare-QA
huggingface.co/datasets/ChatMED-Project/RAGCare-QA

THE PROBLEM

- Educators deploy complex RAG and pay 3× more — for no accuracy gain
- One-size RAG: either overkill or too slow for real classroom use
- No evidence-based guide exists for choosing RAG in medical education

RESEARCH QUESTION

Four RAG architectures, same accuracy — so what actually separates them?

CONTRIBUTIONS

- Three-phase RAG selection decision pipeline
- Empirical comparison: 4 architectures, 420 MCQs
- Accuracy (Wilson CIs), latency & cost benchmarks
- Honest null results + stated limitations

KEY FINDING All 4 architectures **tie at 96% accuracy — pick by cost and speed, not accuracy.**

RAGCARE-QA DATASET

- 420 medical MCQs
- 6 specialties, 3 difficulty levels
- Reproducible exact-match scoring (no LLM judge)
- GPT-4o answers, GPT-4o-mini auxiliary
- R2R framework, text-embedding-3-small

In · Define Your Use Case & Requirements

- What type of questions will your app handle?
- What are your latency, budget & accuracy needs?

QUESTION-TYPE MIX (RAGCARE-QA)

- Vanilla RAG: 315 Qs (75%)
- Vector RAG: 82 Qs (19.5%)
- Graph RAG: 23 Qs (5.5%)
- Difficulty: Basic 150 · Intermediate 181 · Advanced 89

EVALUATION SETUP

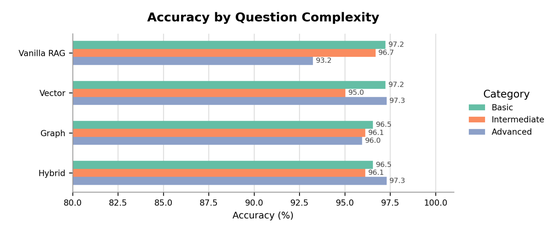
- 4 architectures on shared R2R framework
- Generator: GPT-4o (temp 0.1) — same for all
- Embeddings: text-embedding-3-small, dim 512
- Exact option-label match (A–E), no LLM judge

1 · Phase 1 · Requirements Analysis

Classify question type, domain, and accuracy targets.

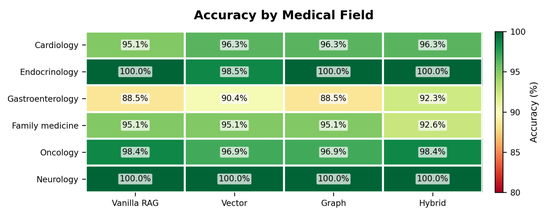
- Question type classification
- Domain specificity assessment
- Accuracy requirements
- User & performance needs

ACCURACY BY COMPLEXITY



Accuracy by question difficulty level across architectures.

SPECIALTY × ARCHITECTURE



Accuracy heatmap: specialty by architecture.

2 · Phase 2 · Constraint Evaluation

Map your latency, budget, and infrastructure constraints.

- Latency tolerance tiers
- Compute & resource availability
- Budget & cost limits
- Technical expertise level

CONSTRAINT TIERS

- Real-time (< 2 s) → Vanilla RAG only
- Interactive (2–5 s) → Vanilla, Vector, most Hybrid
- Batch (> 5 s) → all four architectures
- Cost tier: 1× (constrained) · 2× (standard) · 3×+ (rich)

RESULTS (N = 420, 95% WILSON CI)

ARCHITECTURE	ACCURACY [CI]	LATENCY	COST
Vanilla RAG	96.25% [93.90, 97.64]	1–2 s	1.0×
Vector RAG	96.25% [93.90, 97.64]	2–4 s	1.5×
Graph RAG	96.25% [93.90, 97.64]	4–8 s	3.0×
Hybrid RAG	96.50% [94.19, 97.82]	2–6 s	2.2×

BY QUESTION TYPE

TYPE	VANILLA	VECTOR	GRAPH	HYBRID
Basic (n=315)	96.8%	95.9%	95.6%	96.5%
Multi-vec (n=82)	95.1%	**97.6%**	**97.6%**	96.3%
Graph (n=23)	91.3%	91.3%	**100%**	95.7%

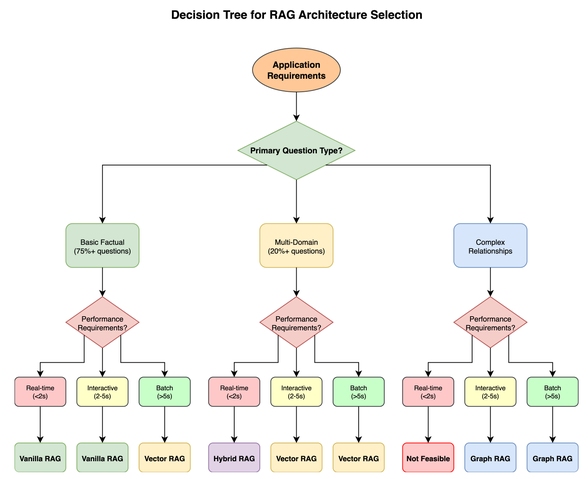
No pairwise difference is significant (Fisher's exact p ≥ 0.49).

3 · Phase 3 · Architecture Assessment

Validated on 420 MCQs × 4 architectures — apply scores to your constraints.

- Performance-to-question mapping
- Cost–benefit analysis
- Risk assessment
- Architecture scoring vs requirements

DECISION TREE



Routes on question mix and budget; thresholds are indicative.

Four RAG architectures — the candidates evaluated

Vanilla RAG

96.25%
1.0× cost · 1–2 s

Dense bi-encoder, top-k cosine, concat context

- Best for basic factual
- Lowest cost & latency

Vector RAG

96.25%
1.5× cost · 2–4 s

Multi-representation embeddings, hierarchical retrieval

- Best for multi-domain
- Semantic-cluster reranking

Graph RAG

96.25%
3.0× cost · 4–8 s

Knowledge graph, entity-relation, multi-hop

- Best for complex relations
- Multi-hop over graph

Hybrid RAG

96.50%
2.2× cost · 2–6 s

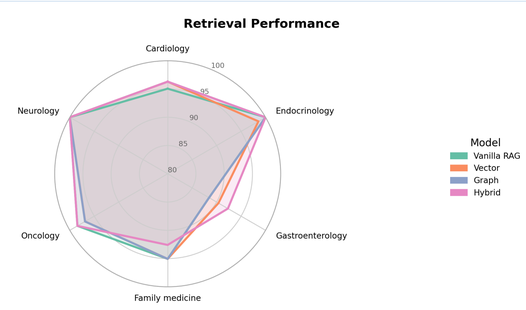
Query-adaptive selection plus ensembling

- Best for mixed workloads
- Confidence-based routing

STATISTICAL HONESTY

- No architecture beats another — every comparison p ≥ 0.49
- Graph RAG's 100% on 23 questions doesn't survive small-sample scrutiny (Wilson lower bound: 85.7%)
- Accuracy won't decide this for you — cost and latency will

ACCURACY BY SPECIALTY



Accuracy across 6 specialties (n=54–87); CIs overlap.

Out · Output · Architecture Recommendation

- Budget-constrained? Vanilla RAG: same accuracy, 3× cheaper than Graph
- Need multi-hop reasoning? Graph RAG — but budget 4–8 s per query
- Always validate cutoffs on your own patient-population data

WORKED EXAMPLES

- Terminology lookup tool → Vanilla RAG
- Clinical case-study site (multi-domain) → Vector RAG
- Research assistant (complex relations) → Graph RAG
- High-volume mixed platform → Hybrid RAG

CHOOSE BY WORKLOAD

Terminology lookup → **Vanilla RAG**	1.0× · 1–2 s	
Case-study tool → **Vector RAG**	1.5× · 2–4 s	
Research assistant → **Graph RAG**	3.0× · 4–8 s	
High-volume platform → **Hybrid RAG**	2.2× · 2–6 s	

LIMITATIONS No RAG-free baseline · Graph-enhanced subset: only n=23 · Costs modelled, not measured · Single exact-match judge · Single inference per question

Paper: to appear — DeLTA 2026 Proceedings (SCITEPRESS) • **Dataset:** huggingface.co/datasets/ChatMED-Project/RAGCare-QA • **Contact:** jdo@cs.aau.dk

Funded by the European Union — Horizon Europe project **ChatMED** (grant 101159214). Views expressed are the authors' only.