Decision Matrix: RAG vs Fine-Tuning vs Embeddings (VAST)

Use the **VAST** lens—**Volatility**, **Accountability**, **Specificity**, **Throughput**—to quickly route use cases to the right approach (or combo).

How to use this

- Score each criterion 1–5 (Low to High) for your use case.
- Use the “Approach Fit Map” to see which method aligns with your scores.
- sanity-check with the “Routing Rules” and set guardrails with the checklist.

VAST Scoring Grid

Criterion	**What it means**	**How to score (1–5)**
Volatility	How often the source-of-truth changes	1 = Rarely updates; 5 = Weekly/daily updates
Accountability	Need for citations, access control, audit	1 = Low; 5 = Strict governance required
Specificity	Task stability and format rigidity	1 = Ad-hoc/creative; 5 = Repetitive/templated
Throughput	Latency/QPS and unit cost pressure	1 = Flexible SLA; 5 = Tight SLA, high volume

Approach Fit Map

Approach	**Volatility**	**Accountability**	**Specificity**	**Throughput**	**Use when…**
RAG	High	High	Medium	Medium	Knowledge changes; must cite and respect ACLs
Fine-Tuning	Low	Medium	High	High	Stable tasks; strict formats;

Approach	Volatility	Accountability	Specificity	Throughput	Use when…
					low-latency at scale
Embeddings	Medium	Medium	Medium	High	“Find similar,” route/cluster/ dedupe; cheap scale

Quick Routing Rules (Executive Shortcuts)

- If **Volatility ≥ 4** or **Accountability ≥ 4** → lead with **RAG**.
- If **Specificity ≥ 4** and **Volatility ≤ 2** → add **Fine-Tuning**.
- If **Throughput ≥ 4** or high QPS → lean on **Embeddings** and small **Fine-Tuned** models; keep contexts short.
- Default hybrid for most enterprise workloads: **Embeddings** substrate + **RAG** for ground truth + light **Fine-Tuning** for consistent outputs.

Reference Outcomes and Trade-Offs

Dimension	RAG	Fine-Tuning	Embeddings
Freshness	Excellent (live retrieval)	Weak (requires retraining)	Good (if index updated)
Governance/Audit	Strong (citations, ACLs)	Medium (needs lineage/process)	Medium (treat vectors as derived data)
Cost per call	Medium–High (tokens, re-rank)	Low (once trained)	Low (search at scale)
Latency	Medium	Low	Low
Build complexity	High (pipelines + vector DB)	Medium (data/labeling + evals)	Low–Medium (index hygiene)

Guardrails Checklist (make this non-negotiable)

- **Data pipeline**: PII redaction, document lineage, role/field-level tags.
- **Access control**: Enforced in retriever and UI; log sources and prompts.
- **Evaluations**: Golden sets for quality, faithfulness, safety; trend dashboards.
- **Drift monitors**: Content, data, behavioral, and process drift alerts.
- **Change control**: Reindex on approved updates; version models; rollback path.

KPIs that keep you honest

- **Task Quality**: Faithfulness (RAG), format adherence (Fine-Tuning), precision@k (Embeddings).
- **Ops**: Latency p95, cost per successful task, deflection rate, human review rate.
- **Risk**: Unauthorized access attempts blocked, citation coverage, policy mismatches.

Worked Example (use in portfolio reviews)

- Volatility 5, Accountability 5, Specificity 3, Throughput 3 → Primary **RAG**; add **Embeddings** for recall; consider small **Fine-Tune** for summary format.
- Volatility 2, Accountability 3, Specificity 5, Throughput 5 → Primary **Fine-Tuning** + **Embeddings**; optional **RAG** for edge references.

Bottom line

Lead with the VAST scores. Treat **RAG** as your truth layer when things change and governance matters, compress cost and latency with **Embeddings**, and lock in predictable behavior with **Fine-Tuning** where tasks are stable.