TIER120 PBC · GENERATIVE AI CONSULTING

RAG vs. Fine-Tuning

A Decision Framework for Generative AI Implementations

"Should we use RAG or fine-tuning?" is one of the most common questions in generative AI program planning — and one of the most frequently miscast. The two techniques solve different problems. Choosing between them is less about model preference and more about whether the work depends on **content** (what the system should know) or **behavior** (how the system should respond). This one-pager provides a side-by-side comparison, a decision tree, and worked examples to support the conversation.

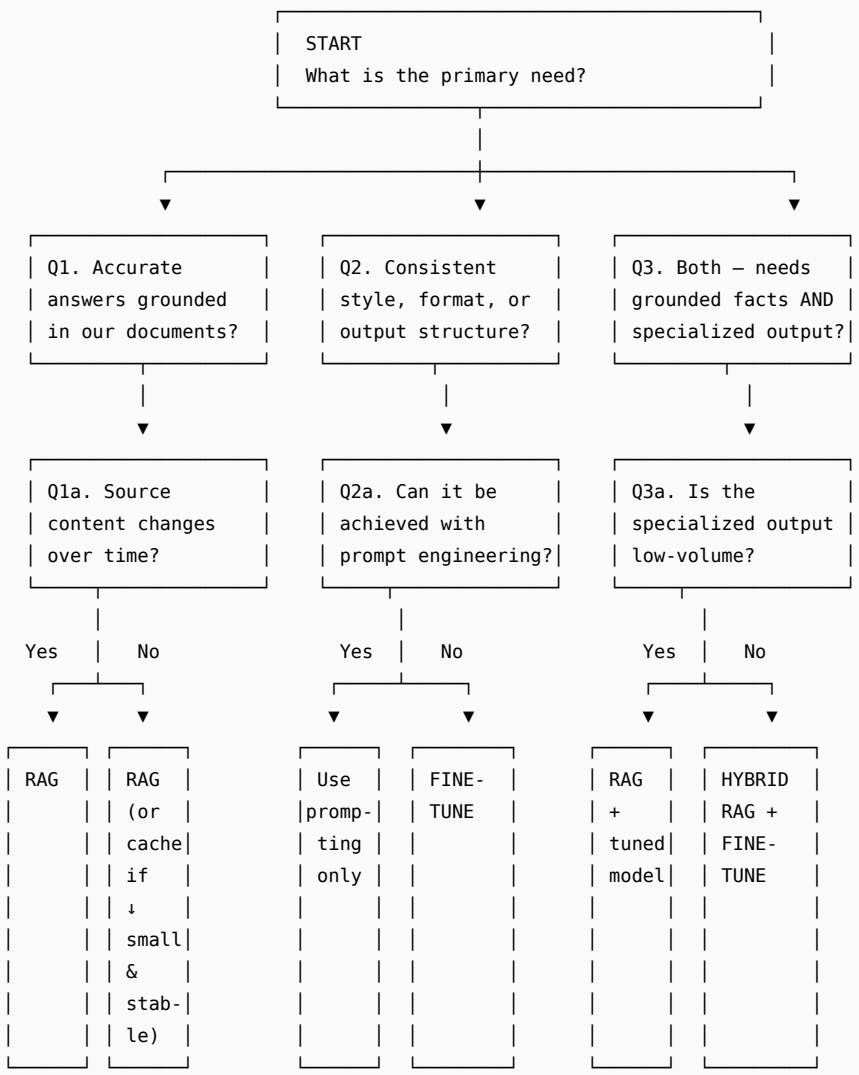
Side-by-Side Comparison

Dimension	RAG (Retrieval-Augmented Generation)	Fine-Tuning
What it does	Retrieves relevant content from your data at query time and provides it to the model as context.	Adapts the model's weights through additional training on your data.
Best for	Tasks that depend on factual content from a defined knowledge base; content that changes; auditability.	Tasks that depend on style, format, tone, structured output, or specialized vocabulary.
Data shape	Documents, knowledge base articles, policy manuals, structured records, FAQs.	Input-output pairs demonstrating the desired behavior; thousands of examples typical.
Update cycle	Update by adding or replacing documents. Changes visible immediately.	Update by retraining. New training runs cost time, money, and version management.
Auditability	High. Sources are retrieved and can be cited in the response.	Low. The model produces output without a traceable source for individual facts.
Hallucination risk	Reduced when retrieval is good; the model is constrained to retrieved context.	Unchanged. Fine-tuning teaches behavior, not new factual grounding.
Cost profile	Lower upfront. Storage and retrieval costs; per-query token costs.	Higher upfront training cost; lower per-query inference cost if the tuned model is smaller.
Time to first value	Days to a few weeks.	Weeks to months, including dataset preparation.
Operational burden	Embedding refresh, retrieval quality monitoring, index maintenance.	Training pipeline, model versioning, evaluation harness, drift management.
Multi-tenant fit	Strong — tenant-scoped retrieval isolates data at the query layer.	Weaker — a tuned model carries learned content; isolation requires per-tenant models.

Decision Tree

Work top-to-bottom. The questions are ordered by how often they distinguish the correct answer; most cases resolve in two or three steps.

Figure 1 — Decision Tree



Recommendation Legend

RAG	Use Retrieval-Augmented Generation. Lowest risk and fastest time to value when the answer must come from your documents.
FINE-TUNE	Train a model on input-output examples. Use when consistent format, tone, or specialized language is the primary goal.
HYBRID	RAG provides grounding; fine-tuning provides style or domain vocabulary. Higher complexity; reserve for cases where both are required.
NEITHER	Prompt engineering, traditional ML, or rule-based automation may fit better. Generative AI is not the right tool for every problem.

Worked Questions & Examples

The Six Questions That Decide

Q1. Does the answer need to come from a specific body of documents you control?

If yes, you need retrieval. The model alone cannot reliably answer factual questions about your specific policies, contracts, or knowledge base. **Choose RAG.** Fine-tuning will not solve this — a fine-tuned model still hallucinates facts; it just hallucinates them in your house style.

Q2. Does the source content change?

If your documents update monthly, quarterly, or in response to policy changes, fine-tuning becomes a treadmill: every meaningful update demands a retraining cycle. **Choose RAG.** If the corpus is small, static, and stable for years, a fine-tuned model is feasible — though usually still not preferable.

Q3. Is the goal a specific output format, style, or vocabulary?

If you need the model to produce a particular structured output (e.g., a standardized clinical note format, a specific JSON schema, a domain vocabulary the base model does not use), fine-tuning earns its place. **Choose FINE-TUNE** — but only after testing whether prompt engineering and few-shot examples in the prompt already get you there. Many "we need to fine-tune" requirements dissolve under good prompting.

Q4. Do you have enough labeled training data?

Fine-tuning needs hundreds to thousands of input-output pairs that demonstrate the desired behavior. If you cannot produce them — and producing them is harder than people expect — fine-tuning is not yet viable. **Fall back to RAG plus prompt engineering.**

Q5. Do you need to audit which source produced each answer?

Regulated, legal, and high-stakes use cases require traceability. A fine-tuned model cannot tell you which training example produced a specific phrase in its output; a RAG system can return the source document and passage. **If auditability is a hard requirement, RAG is the only defensible choice.**

Q6. Will the system serve multiple tenants or organizations on shared infrastructure?

RAG isolates tenants at the retrieval layer using per-tenant indexes or filters. A fine-tuned model carries learned content in its weights — true isolation requires a separate tuned model per tenant, which is operationally heavy. **For multi-tenant platforms, default to RAG.**

Worked Examples

Example Use Case	Choose	Why
Internal policy Q&A — "What is our travel reimbursement policy for international trips?"	**RAG**	Answer must come from a specific authoritative document and may change with policy updates.
Customer support assistant grounded in product documentation	**RAG**	Documentation evolves; need source citations; multi-product or multi-tenant isolation matters.
Generating financial summaries in a specific format with consistent section headers and tone	**FINE-TUNE**	The task is about output structure and style, not novel facts. The financial data comes from another system.
Coding assistant for an internal DSL or API	**FINE-TUNE**	Specialized vocabulary and

		patterns the base model has not seen; output structure matters.
Regulatory compliance Q&A across changing federal and state rules	**RAG**	Sources change; auditability is a hard requirement; every answer must cite a regulation.
Drafting case notes in a specific clinical or legal voice using current patient or matter records	**HYBRID**	Records are current and tenant-specific (RAG); voice and structure are domain-standardized (fine-tune).
Categorizing inbound email into one of 12 known categories	**NEITHER**	A classifier or rule set is faster, cheaper, more explainable, and easier to govern.
Generating creative marketing copy with no specific source material	**Prompting**	Neither RAG nor fine-tuning is needed; prompt engineering plus a strong base model is sufficient.

Common Anti-Patterns

"Fine-tune the model on our policy manual." Policy manuals change. Fine-tuning will encode an obsolete version. Use RAG.

"RAG will fix our model's style problems." Retrieval changes what the model sees, not how it writes. If the issue is tone or format, prompt engineering or fine-tuning is the answer.

"We need both, so let's start with hybrid." Start with RAG plus prompt engineering. Add fine-tuning only when measured evaluations show prompting cannot close the style or format gap.

"We'll fine-tune to reduce hallucinations." Fine-tuning does not reduce hallucinations in general — and can make them worse on topics outside the tuning distribution. Use RAG to ground responses in retrievable content.

About Tier120 PBC

Tier120 PBC is a Florida-based consulting firm helping government agencies and businesses modernize through generative AI, cloud, and compliance solutions. Our Generative AI practice partners with clients to assess readiness, choose the right architecture, and stand up production AI capabilities with the governance to support them.

Talk to us: tier120pbc.com · Schedule a consultation through the website.