Comparative Evaluation of Knowledge-Grounded LLM Strategies for Manufacturing Safety Q&A

SIGHT's safety assistant must satisfy accuracy, low latency, and low operational cost, especially because answers are safety-critical and may be used in real-time VR/AR coaching contexts. The project has already produced expert-validated Q&A benchmarks for three machines and demonstrated that retrieval strategy and model configuration meaningfully affect correctness, latency, and per-query cost. This plan extends that foundation by comparing three *deployment paradigms*:

1. **RAG** (this has already been done).
2. **Long-context** (placing all of the relevant documentation directly in the model context – compression strategies?).
3. **Fine-tuning** (supervised fine-tuning on the benchmark + task data; optionally preference tuning).

Research questions and hypotheses

RQ1 — Accuracy

Which paradigm yields the highest safety-critical factual accuracy on the existing machine benchmarks?

- **H1:** Properly configured RAG achieves equal or higher accuracy than long-context for most questions due to targeted grounding and reduced distraction from irrelevant passages.
- **H2:** Fine-tuning improves consistency on frequent/templated questions but may underperform on rare/novel questions unless paired with retrieval.

RQ2 — Speed (latency)

Which paradigm provides lowest end-to-end latency under realistic deployment constraints (cloud LLM API; potential edge constraints)?

- **H3:** Long-context is often slowest when context is large (token processing dominates).
- **H4:** RAG with lightweight retrieval and small top-k can be fastest for many queries.

RQ3 — Cost

Which paradigm provides the best cost-performance frontier at projected query volumes?

- **H5:** Long-context incurs the highest marginal cost per query (large input tokens).
- **H6:** RAG yields lower marginal cost than long-context when retrieval narrows context; fine-tuning shifts spend from per-query to up-front training.

RQ4 — Robustness & maintainability (??)

How do approaches compare in **updateability, auditability, and compliance** when documentation changes (new OEM revisions, OSHA updates)?

- **H7:** RAG is easiest to update (replace corpus/index) and easiest to audit (citations). Fine-tuning requires retraining to incorporate changes.

Scope and artifacts

Datasets (already available)

- Expert-validated Q&A benchmarks for:
  - **Bridgeport manual mill**
  - **Haas TL-1 CNC**
  - **Universal Robots UR5e** s

Evaluation design

Common evaluation dimensions

All approaches are evaluated on three primary dimensions:

1. **Accuracy**: correctness against gold answers
2. **Speed**: end-to-end latency and time-to-first-token (streaming metric).
3. **Cost**: marginal per-query cost (tokens + retrieval infra), plus amortized training costs for fine-tuning.

Platform

Google Cloud's Vertex AI or OpenAI's Playground
Google's Gemini has a much bigger context window