LLM Training & Safety Architecture

Executive Intent
This document summarizes the Large Language Model (LLM) training and deployment strategy for the Resume Parsing & Screening Platform. The design prioritizes recruiter trust, explainability, legal defensibility, and cost control, while safely leveraging modern LLM capabilities.

Strategic Design Positioning

- LLMs are not decision-makers; they are constrained extraction and interpretation engines.
- All hiring signals are computed deterministically and are fully reproducible.
- AI is used only when rule-based systems fail (unstructured text and semantic ambiguity).

LLM Training Strategy (What We Train)

We fine tune LLMs exclusively for structured resume parsing. Supervised Fine tuning (SFT) is performed using QLoRA adapters on a vetted dataset where the target output is TOON (Typed Object-Oriented Notation), not JSON. This enforces schema correctness at generation time and significantly reduces hallucination risk.

What We Explicitly Do NOT Train

The following components remain permanently deterministic and are never learned by LLMs:

- Experience calculation (timeline-based only)
- Salary extraction (regex-based)
- Skill verification for scoring
- Final score computation and weighting

AI Safety & Trust Controls

- TOON is enforced at all LLM generation boundaries; raw JSON generation is prohibited.
- A validation firewall overrides any incorrect or fabricated AI output.
- Deterministic matchers always precede AI-assisted scoring.
- Full fallback to baseline parsing models is retained for instant rollback.

Vector Database Role

A vector database is used strictly as semantic memory. Validated resume representations are embedded for scalable semantic matching and context retrieval. The vector store does not train models and cannot influence factual extraction or scores.

Business & Engineering Impact

- Predictable and auditable hiring decisions (compliance ready).
- Lower LLM operating costs via fine−tuned, task specific models.
- Clear separation of concerns enables independent evolution of AI and business logic.
- Architecture scales from API based models to on premise or hybrid deployments.

Summary

The platform applies fine−tuned LLMs only for constrained extraction, while preserving deterministic scoring and explainability as the system of record for hiring decisions.