02 · PROCAPITA GROUP · PRODUCTION · PRIVATE

Enterprise LLM Fine-Tuning & Self-Hosted Inference

Domain-specialist LLMs for HR products — competencies, job descriptions and role similarity — trained on proprietary data and served on private infrastructure with no per-token cost.

ROLE AI & Data Engineer — Nov 2025 to present

STACK Python · Pandas · JSONL · Qwen models · QLoRA · PEFT · Hugging Face Transformers · vLLM · CUDA · Ubuntu · Tailscale (WireGuard)

10+	65K	10,000+	6	5,000+	$0
proprietary datasets	JSONL rows · flagship set	raw files cleaned	QLoRA fine-tuned LLMs	enterprise documents	per-token API spend

THE PROBLEM

Generic models don't know an organisation's competency frameworks, and sending HR data to a cloud API isn't acceptable. Earlier attempts across 6+ models were unstable because the training data was messy.

WHAT I BUILT

- **Dataset engineering.** A Python pipeline that standardises, deduplicates and validates 10,000+ raw source files into retraining-ready JSONL, including a flagship 65,000-row job-title-to-competency dataset.
- **Fine-tuning.** 6 domain-specialist models on Qwen base models with QLoRA and PEFT, for competency generation and role-similarity reasoning, with evaluation and overfitting mitigation on limited GPUs.
- **Self-hosted serving.** A multi-server vLLM fleet on Ubuntu with OpenAI-compatible endpoints, connected over a private Tailscale mesh with no public ports: sub-second latency and full data sovereignty.
- **Continuous ingestion.** Incoming records are cleaned and structured automatically for the next retraining cycle, replacing manual preprocessing.

PIPELINE

1. 10,000+ raw files
2. Clean · dedupe · validate
3. JSONL datasets (65K-row flagship)
4. QLoRA / PEFT on Qwen
5. Eval & checkpoint
6. vLLM on Ubuntu servers
7. Private Tailscale mesh
8. HR product features