# Enterprise‑Scale Job‑Description Generation: Cost & Deployment Options

(GPT‑4o‑mini API vs Self‑hosted 7 B model with fine‑tuning)

## 1. Executive Summary

OpenAI GPT‑4o‑mini offers low per‑token cost and zero‑ops scalability for most workloads, while a self‑hosted 7 B model (Mistral/Llama‑2) becomes cost‑effective only at very high sustained traffic or strict data‑sovereignty requirements.

## 2. Pricing Inputs & Calculations

• GPT‑4o‑mini: $0.15 / M input tokens, $0.60 / M output tokens.

• AWS g5.xlarge (A10G GPU): $1.006 per hour (on‑demand).

• Mistral‑7B throughput on A10 with vLLM: ~92 tokens/s (~331k tokens/hour).

Cost per million tokens (7 B on A10): $1.006 / 331k ≈ $3.04 per 1M tokens.

## 3. Hidden Cost Factors

- DevOps overhead (0.25 FTE) for self‑hosting.  
- GPU reservation vs bursty demand.  
- Compliance and data‑residency requirements.  
- Model update and maintenance cycles.

## 4. Recommendation Matrix

• < 20 M tokens/month: GPT‑4o‑mini only.  
• 20‑50 M tokens/month: Hybrid (local small + GPT‑4o‑mini).  
• 50‑200 M tokens/month: Self‑host 7 B for generation, GPT‑4o‑mini for eval.  
• > 200 M tokens/month or strict data‑sovereignty: Multi‑GPU self‑hosted stack.

## 5. References

* OpenAI GPT‑4o‑mini pricing (Jul 18 2024).
* AWS Pricing: g5.xlarge in us‑east‑1.
* Inferless LLM Benchmark (Mistral‑7B throughput).
* TrueFoundry Benchmarking Mistral‑7B.