AI STRATEGY PORTFOLIO · EMPLOYER BRIEF

Earnings Call Interrogator

Purpose: measure when a small, task-specific open-weight model is good enough versus paying for frontier APIs — using a real equity-research workflow as the controlled test bed.

EXECUTIVE OVERVIEW

July 2026

Alex Behm, CFA, MBA

ENGINES	CALLS / ENGINE	LOCAL API COST	BEST PAID OPTION	TRAINING GPU
3	10	$0	GPT-5.5 · 11¢	GCP L4

DOCUMENT PURPOSE

SHAREABLE EMPLOYER BRIEF · BUILD-VS-BUY EVIDENCE FOR APPLIED GENAI

What this project is. Public companies hold quarterly earnings calls. Analysts care about what was asked, how management answered, where answers were partial, and how much of the call was scripted remarks versus live Q&A. I built a system that turns a raw transcript into a structured research PDF — then used that system as a controlled experiment to compare a **fine-tuned open-weight 8B model I trained myself** against **two frontier APIs** on cost, speed, reliability, and output behavior.

1. Business question

> *When is a small, owned model good enough for a narrow, repeatable task — and when is the frontier premium worth paying?*

I answered that with evidence. The experiment held **everything constant except the extraction model**: same transcripts, same prompt schema, same post-processing, same report template. That isolates economics and behavior so the comparison is fair.

Why this matters for employers: most GenAI debates are abstract. This project forces a decision framework with measured $/call, wall time, reliability, and flag density — the same tradeoffs a pursuit or product team faces when scoping AI services for a customer.

2. What the report contains

Section	What it tells an analyst
Bottom-line verdict	One-paragraph read on management transparency
Credibility score	0–100 heuristic of answer directness
Evasion flags	Questions with partial / deflected answers
Filibuster score	Q&A share of call vs history
Analyst scorecard	Who asked sharp questions (name/firm)
Metrics & guidance	Numbers + forward-looking statements

Scores are decision-support tools, not audited investment research.

3. Pipeline (product under test)

Stage	What happens	Why it matters
Ingest	Load full earnings-call transcript (PDF or text).	Real-world inputs, not toy prompts
Preprocess	Detect FactSet vs Motley Fool; split prepared remarks vs Q&A; parse speaker turns.	Model only sees the right section
Extract	Model returns structured JSON (questions, responses, evasion candidates, metrics, tone, guidance).	Comparable outputs across engines
Score	Deterministic layer: credibility, filibuster, analyst attribution, flags.	Scoring not reinvented per model
Report	Professional multi-page PDF watermarked with the engine used.	Stakeholder-ready artifact

4. Data provenance

Training data
Hugging Face `kurry/sp500_earnings_transcripts` (S&P 500 earnings calls, 2005–2025). Supervised fine-tuning examples focused on the **Q&A section** of the call — not the entire prepared-remarks monologue.

Evaluation data (held out)
10 recent **2026** calls collected after training (Motley Fool + FactSet): NVDA, PEP, LEN, GIS, FDS, WLY, AIOT, MSM, KRUS, HELE. Out-of-sample vs training corpus.

5. Measured snapshot (headline results)

	Qwen v2 (CPU)	Qwen GPU (est.)	GPT-5.5	Fable 5
API cost / call	$0	$0	~11¢	~31¢
Wall time / call	~205s	~34s (est.)	~67s	~55s
Evasion-flag density	~11%	~11%	~39%	~33%
First-pass success	8/10*	same model	10/10	10/10
Data leaves machine?	No	No	Yes	Yes

*Local misses recovered on retry. GPU latency is illustrative (~6× vs measured CPU), not a re-run. Detail and charts on pages 2–4.

6. Fine-tuning — two iterations (same base model)

Base model: **Qwen3-8B**, 4-bit QLoRA on a **GCP NVIDIA L4** (Unsloth). I did not ship the first attempt.

	v1 — failed	v2 — production path
Labels	Regex / pattern heuristics	Structured JSON from GPT-4o, aimed at earnings Q&A
Examples	4,448	499
Final loss	3.67 (did not learn the task)	0.016 (~230× better)
Train time	~6.8 hours	~9.4 hours (~500 steps)
Output quality	Unusable (greetings as “questions,” politeness as “evasions”)	Real analyst questions, contextual metrics, usable reports
Serving	—	GGUF export; local Ollama on mini PC (CPU)

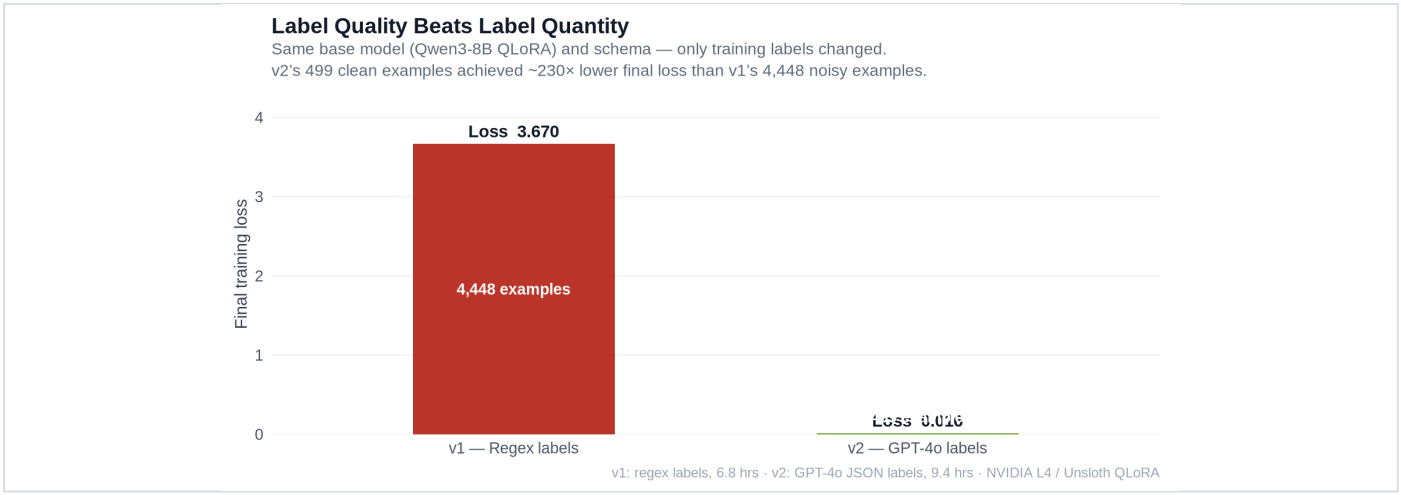

Figure 1. Label quality beat label quantity — the central AI investment lesson from this project.

> *If I could put one sentence in front of every executive budgeting an AI project: the money is in the data quality, not the model size.*

7. Three extraction engines compared

OWNED SPECIALIST
Qwen3-8B v2
Fine-tuned for this task only. Offline on my desk. $0 API. Sample: NVDA_…_Finetuned.pdf

FRONTIER — ECONOMIC
GPT-5.5
OpenAI API. No task fine-tune. ~11¢ · ~67s avg. Sample: NVDA_…_GPT.pdf

FRONTIER — PREMIUM
Claude Fable 5
Anthropic via Nous. ~31¢ · ~55s avg. Sample: NVDA_…_Fable.pdf

DESIGN RULE Same transcripts · same JSON schema · same scoring · same PDF template · only the model call changes. Reports watermarked by engine.

8. Latency dispersion

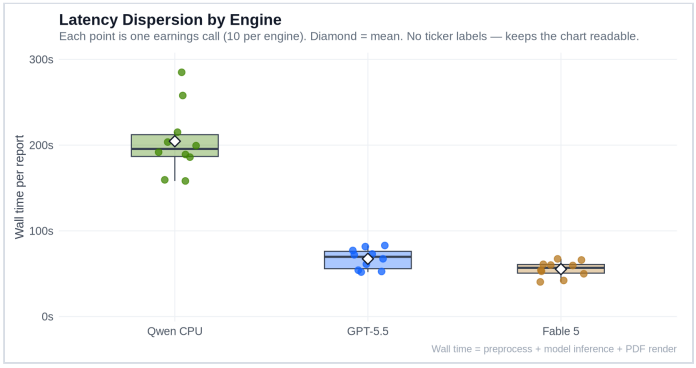

Figure 2. Per-call wall times (10 each). Diamond = mean.

9. Evasion-flag density

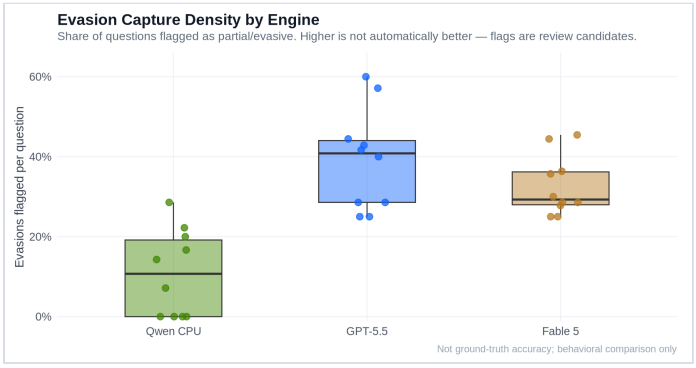

Figure 3. Flags per question — review candidates, not ground truth.

10. Results — AI efficiency frontier

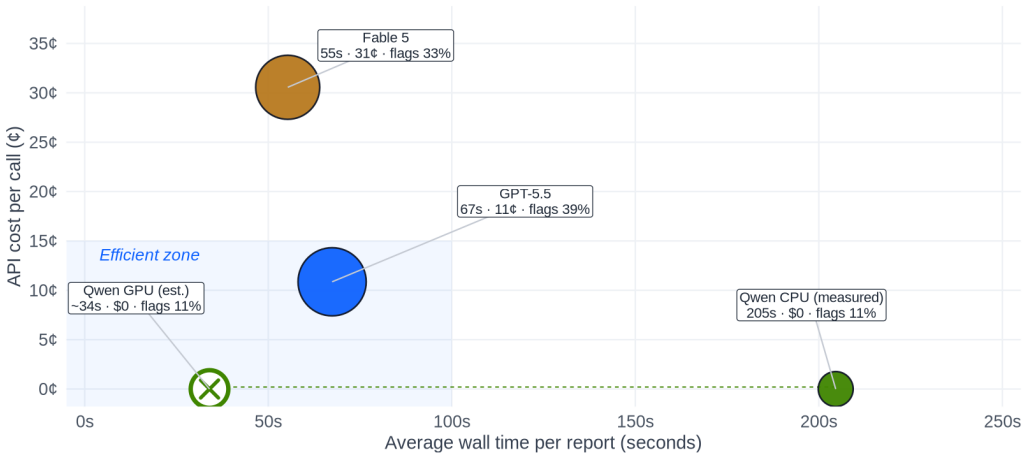

Figure 4. X = wall time · Y = API ¢/call. Labels show time, cost, and evasion-flag density. Solid = measured; hollow × = Qwen GPU estimate (~6× vs CPU). Bubble size also encodes flag density.

GPU callout (illustrative, not re-benchmarked). Measured Qwen wall time (~205s) is CPU inference on a mini PC. Applying ~6× model throughput for a mid-range consumer GPU implies roughly **~34s wall time at $0 API**. Shown as a hollow × so it is never confused with measured data.

Metric	Qwen v2 (CPU)	Qwen v2 (GPU est.)	GPT-5.5	Fable 5
API cost / report	$0	$0	~11¢	~31¢
Wall time / report	~3.4 min	~34 s (est.)	~67 s	~55 s
Evasion-flag density	~11%	~11% (same model)	~39%	~33%
First-pass reliability	8/10*	same model	10/10	10/10
Data leaves machine?	No	No (if local GPU)	Yes	Yes
Best use case	High volume + privacy	High volume + speed	Default paid path	Premium speed/flags

*Local misses recovered on retry (timeout / format). Full recovery = 10/10. Higher flag density ≠ proven accuracy.

Local model
Wins on marginal cost and data control. CPU is slow; GPU estimate closes most of the latency gap without API spend.

GPT-5.5
Best measured paid economics — fast, cheap, reliable. Default if you will not host a model.

Fable 5
Fastest frontier option and aggressive flagging at ~3× GPT price. Premium only if human QA proves better capture.

11. How to read the scores (quick glossary)

Term	Meaning in this project
Wall time	Clock time start→finish for one report (preprocess + model + PDF).
Success rate	Share of runs that produced valid JSON + PDF on first pass.
Credibility score	Internal 0–100 transparency heuristic after extraction — decision support, not ground truth.
Evasion-flag density	Flags ÷ questions. Candidate rate for human review; engines disagree by design.
Filibuster	Q&A words ÷ total call words (150 WPM duration estimate).

12. Scale economics — quarterly screening volume

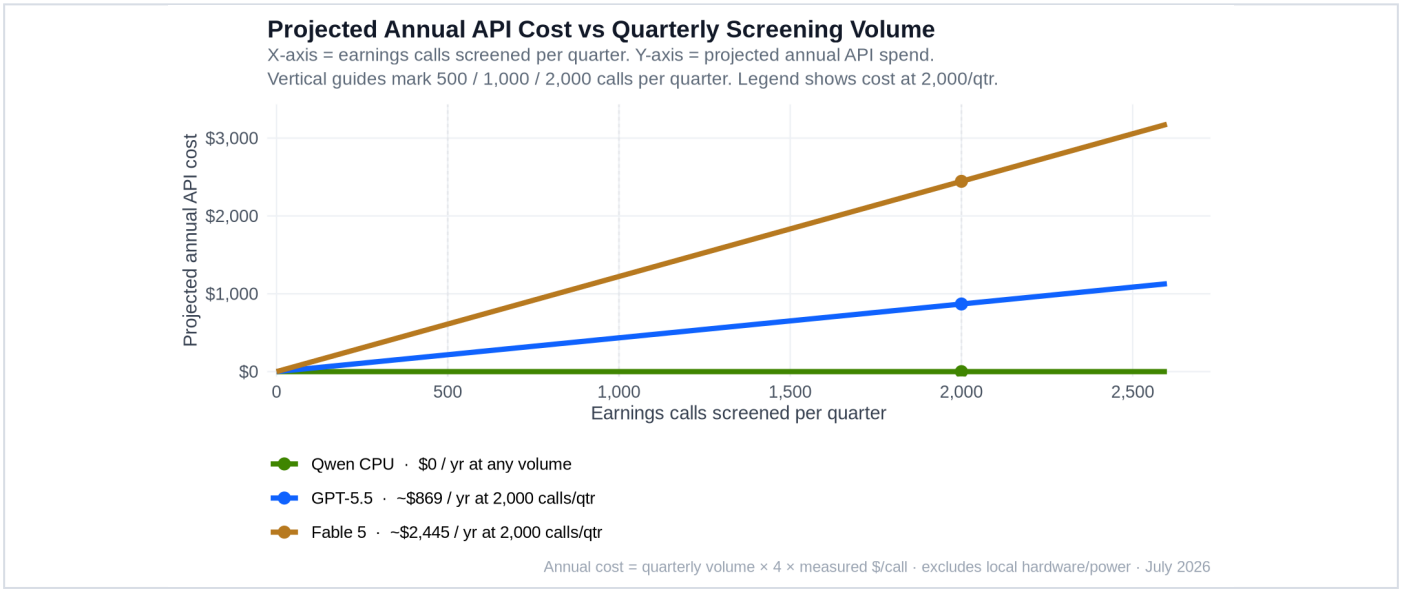

Figure 5. X = earnings calls screened **per quarter**. Y = projected **annual** API spend. At 2,000 calls/qtr: GPT ~$870/yr · Fable ~$2.4k/yr · owned model $0 API.

13. Decision framework

Situation	Lean toward
Low volume, need speed, OK sending data to vendor	**GPT-5.5**
Need maximum flag candidates + fastest frontier	**Fable 5** (if QA justifies premium)
High volume, privacy, or predictable unit cost	**Owned Qwen** (GPU if latency matters)
Unclear volume / early pilot	Start API; re-evaluate at ~500–1,000 calls/qtr

14. Strategy takeaways

- **Narrow tasks can be specialized** — 8B + excellent labels produced usable reports.
- **Evaluation is the product** — fixed template/scoring made build-vs-buy decision-grade.
- **Frontier premium buys speed and candidate judgment**, not automatic truth.
- **Architecture should outlive the model** — swap engines without rebuilding the report system.
- **Volume changes the answer** — quarterly scale chart makes the crossover concrete.

15. Honest limitations

- n = 10 companies — directional, not large-sample proof.
- Credibility / evasion scores are heuristics, not audited labels.
- GPU latency is estimated from throughput ratios, not a re-run on GPU.
- Portfolio case study — not a commercial SaaS claim.

16. Accompanying materials

Material	Location
This brief (PDF/ODT/DOCX)	`Employer Brief/`
Technical case study	`Employer Brief/02_…`
NVDA samples (3 engines)	`Sample Reports/`
Charts + CSVs + R	`Visuals/` · `Data/` · `Source/`

Under the hood: Qwen3-8B QLoRA on GCP L4 → Ollama GGUF; six-key JSON; deterministic scoring; WeasyPrint PDF; frontier via OpenAI + Nous with identical prompts.