Sample deliverable - worked example

This is a worked example of the audit deliverable, built entirely from the public QuantBench data so you can see the structure and the standard of evidence before you buy. The "client" here is hypothetical; every number is real and traceable to a row in the public dataset (https://huggingface.co/datasets/Mohaaxa/quantbench-leaderboard-data).

A PDF copy (quantbench-sample-audit.pdf) of this document is available to download and forward.

Engagement summary

Scenario: a small team serving a 1.5B instruct model for internal document Q&A, batch size 1, currently on FP16, evaluating whether to quantize and whether to move from an A10 to a cheaper T4 tier.

Verdict up front: AWQ at 4-bit is the recommended path on both tiers. It costs about +0.72 perplexity against the FP16 baseline and cuts peak VRAM from 6.48 GB to 4.57 GB. Do not move this workload to T4 - see finding 3.

1. Quantization trade-off table

Measured on Qwen/Qwen2.5-1.5B-Instruct, calibration wikitext2-train n=128 seed 0, greedy decode, batch 1, 64->256 tokens, median of 3, prefill included.

Config	delta PPL vs FP16	tok/s	Peak VRAM	Verdict
FP16 baseline (A10, AWQ stack)	-	45.38	6.48 GB	reference
AWQ 4-bit (A10)	+0.7852	19.74	4.57 GB	recommended
GPTQ 4-bit (A10)	+0.4940	17.45	7.69 GB	better quality, worse memory
FP16 baseline (T4, AWQ stack)	-	32.64	6.48 GB	reference
AWQ 4-bit (T4)	+0.7208	8.75	4.57 GB	viable but slow
GPTQ 4-bit (T4)	+0.4323	4.58	7.92 GB	not recommended

Read the quality column as "perplexity added on wikitext2-test." Lower is better; the FP16 references are 9.6394 (A10) and 9.0767 (T4), so +0.78 is roughly an 8% relative increase.

2. Prioritized fix list

1. Adopt AWQ 4-bit on A10 - high value, low effort. 4.57 GB peak versus

6.48 GB FP16 frees enough headroom to raise concurrency or drop to a smaller instance. Quality cost is +0.79 PPL. Median quantization wall-time was 6.2 minutes, so this is an afternoon of work, not a project.

2. Do not adopt GPTQ on the evidence in this table - medium confidence.

GPTQ shows better quality here (+0.49 vs +0.79) but every GPTQ row in this benchmark loaded via a torch-fallback path rather than an optimized kernel, so both its speed and its 7.69 GB memory figure are pessimistic and not a fair read of GPTQ's ceiling. Recommended follow-up before deciding: re-run GPTQ with a working optimized kernel. This is precisely the kind of caveat the audit surfaces rather than hides.

3. Do not migrate this workload to T4 - high confidence. AWQ throughput

falls from 19.74 to 8.75 tok/s (?56%) for a tier that is roughly 46% cheaper per hour. On a throughput-bound workload that is a worse deal per query, not a better one.

4. Calibrate on data that matches your evaluation distribution - high value

if you use GPTQ. For GPTQ the calibration corpus mattered far more than its size: median delta PPL was +0.7174 with wikitext2 calibration versus +1.3776 with chat-style OpenHermes calibration at the same n=128, while raising n from 32 to 512 moved it only 0.05-0.11. AWQ was comparatively indifferent (+0.6798 vs +0.7106).

3. Reproduction

Every row above is reproducible without trusting this document:

```
# 1. get the measurements
curl -L https://huggingface.co/datasets/Mohaaxa/quantbench-leaderboard-data/resolve/main/rows.csv -o rows

# 2. get the exact weights behind the recommended row
huggingface-cli download Mohaaxa/quantbench-artifacts \
  --include "qwen25-1p5b__awq__wikitext2__n128__s0/*"

# 3. re-run the protocol (pinned config in the methodology page)
```

4. Limitations of this worked example

Stated as they would be in a real deliverable:

- Two model sizes only (1.5B, 1.7B). Nothing here extrapolates to 7B+ and

this document does not claim it does.

- Perplexity is a proxy. It is not task accuracy, and it is not your users'

satisfaction. A real audit for a production workload should add a task-specific eval on your own held-out prompts.

- Throughput is one point on the workload surface: batch 1, 64->256 tokens,

greedy. Different batching or prompt lengths can reorder these rows.

- The FP16 baselines are single measurements, so every delta inherits one

unreplicated reference number.

- Absolute perplexity is not comparable across GPU tiers here (different

evaluation token caps), which is why the table pairs each tier with its own reference.