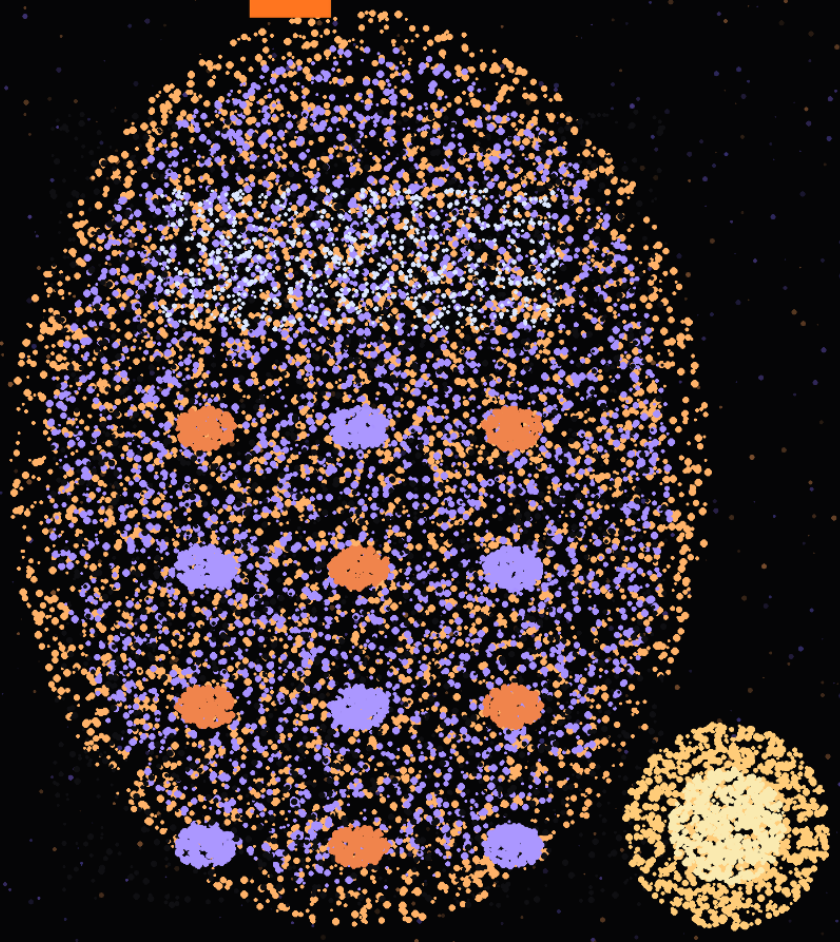
HEBBIAN



Hebbian Research Notes · Paper #001

Comparative Tax Benchmark: Qwen 2.5 3B and DeepSeek 7B

Evaluaciones 2026-04-24 · 16 pruebas · 5 categorías · 4 corridas

Comparative Tax Benchmark: Qwen 2.5 3B and DeepSeek 7B

Hebbian Research Notes · APA 7th Edition Format

Santiago Saltaren

Research Team, Hebbian SAS

Bogotá, Colombia

April 25, 2026

Abstract

This paper presents a comparative evaluation of four model runs focused on Colombian tax and accounting tasks. The study uses a 16-prompt benchmark across five categories, with a 0-5 keyword-coverage score and loop penalties for repetitive outputs. We compare Qwen 2.5 3B baseline vs its fine-tuned variant, and DeepSeek 7B without vs with thinking mode. Results show that, in this snapshot, Qwen fine-tuning does not outperform baseline in global score and increases latency, while DeepSeek thinking mode yields a modest quality gain with minimal latency impact. The study establishes a transparent baseline for subsequent improvement cycles.

Keywords: fine-tuning, tax benchmark, small models, comparative evaluation, Hebbian.

Introduction

Deploying domain-specialized assistants in tax workflows requires balancing regulatory precision, response safety, and operational speed. In this context, Hebbian designed a reference evaluation to compare baseline models, fine-tuned variants, and thinking-mode configurations under a single protocol. The goal is to quantify performance differences and communicate findings in a replicable, auditable format aligned with IMRyD structure and APA criteria (Hebbian SAS, 2026a; Hebbian SAS, 2026b).

Methodology

Evaluation date: April 24, 2026.

Benchmark: 16 prompts across 5 categories (VAT, IFRS, SIMPLE TAX REGIME, SAFETY, UGPP).

Score: 0-5 (higher is better) using keyword coverage in golden_set.json.

A -2.0 penalty is applied when loop-like repetitive behavior is detected.

The 600 iterations are reported as LoRA training context (--iters 600), not as a native field in evaluations.db.

Results

Quantitative model summary (friendly name + technical ID):

Qwen 2.5 3B (Baseline) | score=1.1562 | latency=1.6150s | loops=0 | ID: mlx-community/Qwen2.5-3B-Instruct-4bit

Qwen 2.5 3B (Fine-Tuned) | score=0.7438 | latency=8.2344s | loops=1 | ID: mlx-community/Qwen2.5-3B-Instruct-4bit_fine_tuned

DeepSeek 7B (Direct) | score=0.8438 | latency=7.5750s | loops=0 | ID: mlx-community/DeepSeek-R1-Distill-Qwen-7B-4bit

DeepSeek 7B (Thinking) | score=0.9500 | latency=7.6075s | loops=0 | ID: mlx-community/DeepSeek-R1-Distill-Qwen-7B-4bit_thinking

Figure 1. Average score by model.

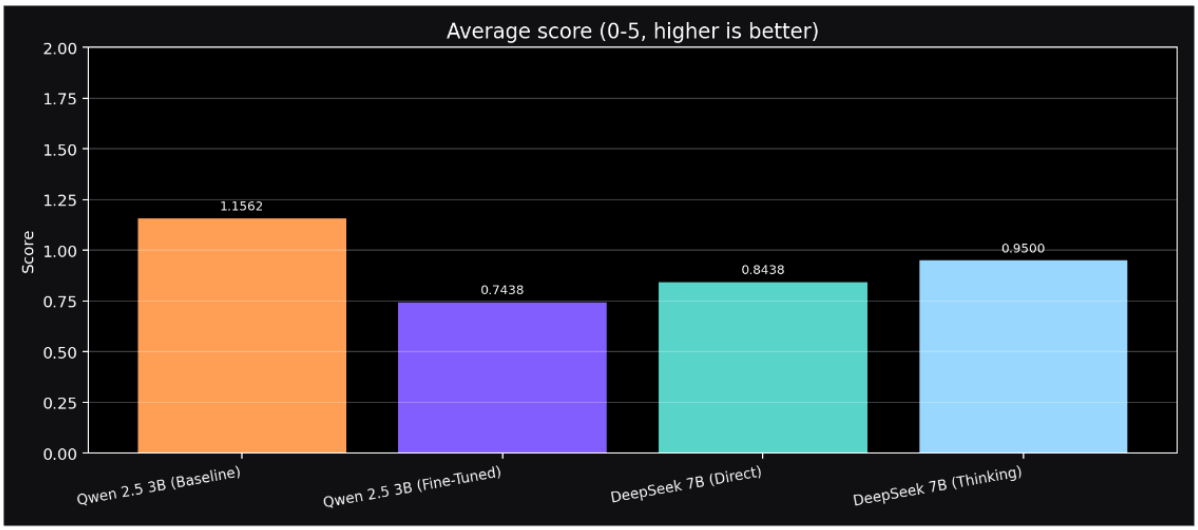


Figure 2. Average latency by model.

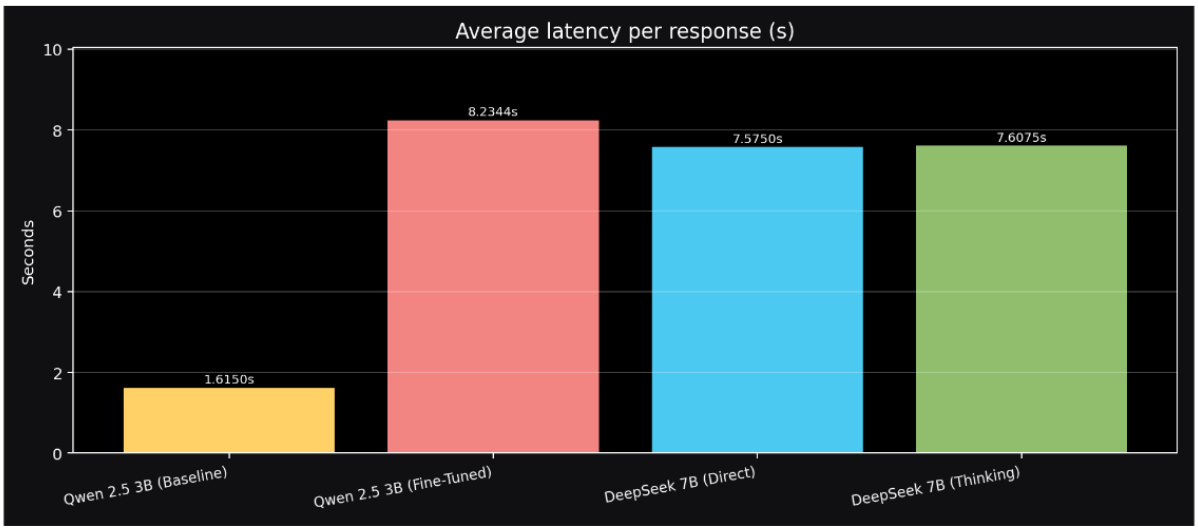
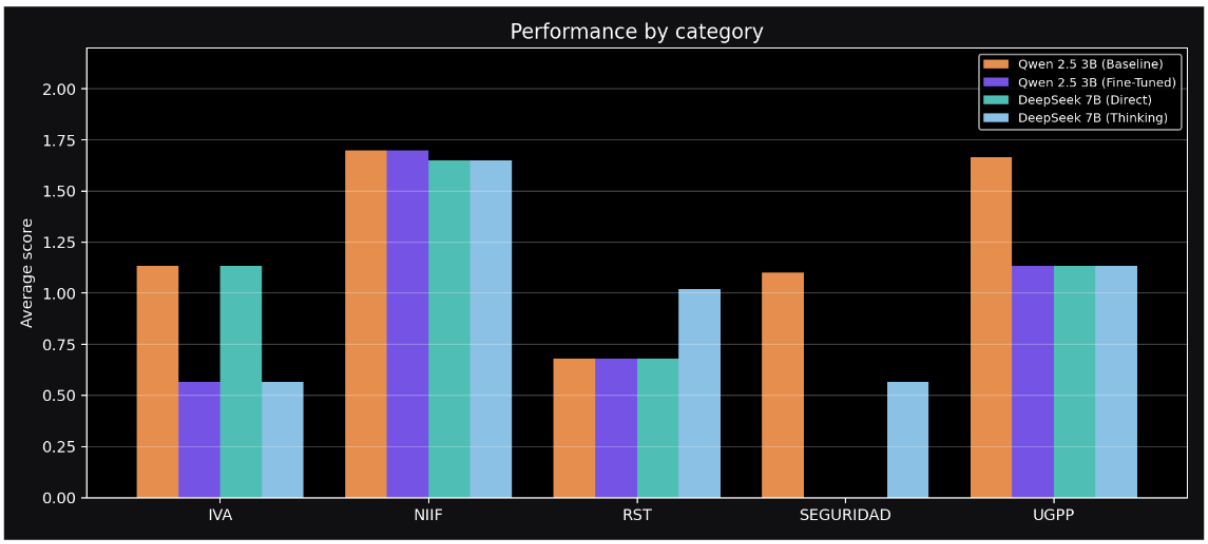


Figure 3. Performance by category.



Qwen fine-tuned vs baseline: Δscore -35.68% and Δlatency +409.87%; questions better/same/worse = 0/13/3.
DeepSeek thinking vs non-thinking: Δscore +12.59% and Δlatency +0.43%; questions better/same/worse = 2/13/1.

Discussion

Results indicate that fine-tuning does not automatically guarantee quality gains; outcomes depend on alignment between training data, evaluation criteria, and operational objectives. In this benchmark, expected improvements from Qwen fine-tuning do not materialize and latency increases significantly. Meanwhile, DeepSeek thinking mode provides a moderate score gain with almost no latency penalty, which can be valuable in workflows where accuracy outweighs strict real-time

constraints.

Strengths: consistent protocol, run-level traceability, and controlled comparisons.

Limitations: small sample size and keyword-based scoring.

Hebbian corporate note: this benchmark is building the foundation for the company's definitive fine-tuned model, whose name is still to be announced.

Learning hypothesis: fine-tuning very small models may provide more value as a strategic learning phase than as the final production version.

Next step: expand golden set, strengthen safety prompts, and rerun multi-round evaluations.

Conclusions

This paper establishes a transparent baseline for Hebbian's tax-model stack evolution.

In this snapshot, Qwen baseline keeps a better score-latency balance than its fine-tuned variant.

DeepSeek with thinking delivers an incremental quality gain with controlled latency impact.

From a corporate perspective, these results enable the roadmap toward Hebbian's definitive fine-tuned model (name not yet published).

Fine-tuning on smaller models is currently treated as a validated learning stage before scaling to larger-capacity models.

Further tuning and larger evaluation sets are recommended before high-impact deployment decisions.

References

Congress of Colombia. (1989). *Colombian Tax Code* (current amendments).

Hebbian SAS. (2026a). *evaluations.db* [Internal evaluation database, unpublished].

Hebbian SAS. (2026b). *golden_set.json* [Internal benchmark set, unpublished].

Hebbian SAS. (2026c). *run_training.sh* [LoRA training configuration with 600 iterations].

UGPP. (2025). *Guidelines and sanctions for social protection contributions*.