Reveel

REVEEL — INDEPENDENT STUDY 2026

We put a frontier AI model up against real parcel audit data.

Here's why "just use ChatGPT" doesn't work for parcel spend management — and what actually does.

~47%
Accuracy when prompting an LLM directly

99.5%
Accuracy with a deterministic rules engine + validated inputs

52,289
Real FedEx charge lines scored against known-correct output

4-week feasibility study

Independent AI engineering firm

Real FedEx invoices & negotiated agreement terms

Frontier LLM (Claude Sonnet 4.6)

EXECUTIVE SUMMARY

Every shipper has heard the pitch. We decided to test it.

"Why pay for a parcel spend platform? Just feed your invoices and your carrier agreement into an AI model and ask it what you're owed."

We commissioned an independent AI engineering firm to spend four weeks attempting to replicate our invoice repricing and audit output using a frontier large language model, real FedEx invoices, and a customer's actual negotiated agreement terms. Two approaches were tested: prompting the model directly, and using the model to help build deterministic audit software.

> *"A coin flip's accuracy with an average error of $7.80 per charge line is not an audit. On a typical invoice file of 50,000+ charge lines, that error is larger than the discrepancies being audited."*

The errors were biased in the worst direction: the model systematically understated discrepancies, silently leaving recovery on the table. The approach that did work required four weeks of expert engineering for a single carrier, still left material gaps unsolved, and was only possible because the team had Reveel's verified audit output to test against. That last point is the one most DIY teams miss.

THE EXPERIMENT

They had every advantage a well-resourced internal team could hope to assemble.

52,289 Real FedEx charge lines across 8,000 unique shipments from a national consumer brand — 3 service types, 21 charge descriptions, 8 zones	**21,266** Rows of canonical FedEx 2025 list rates, plus the full published surcharge schedule (187 rows)
~4,950 Negotiated agreement terms: discounts, earned-discount tiers, minimum charges, rate caps, and surcharge overrides	**1** Answer key — Reveel's production audit output for every charge line, so the team could measure accuracy precisely and debug against known-correct results

The question: with all of that in hand, can an AI-powered system reliably reprice every charge line and flag billing discrepancies — the core of a parcel rate audit?

TEST 1

TEST 1

Ask the LLM directly.

Give a frontier model the invoice row, the relevant rate-table excerpts, the surcharge schedule, and the negotiated agreement terms, and ask it to compute the correct net charge.

~47%	$7.80	Biased low
NOT AN AUDIT	LARGER THAN THE GAPS	WORST DIRECTION
Charge lines within $0.01 of the correct answer — a coin flip	Average error per charge line — bigger than most discrepancies being audited	The model understated discrepancies, manufacturing false confidence

> The errors hide money rather than flag it. A tool that under-reports overcharges doesn't just fail to help — it actively confirms that your invoices are clean when they aren't.

The team also tested retrieval-augmented strategies — assembling the most relevant rate rows, agreement terms, and worked examples into the prompt. Results improved but never approached audit-grade accuracy. Retrieval can surface the right evidence, but the model still has to perform layered arithmetic and conditional logic that a rules engine executes perfectly every time.

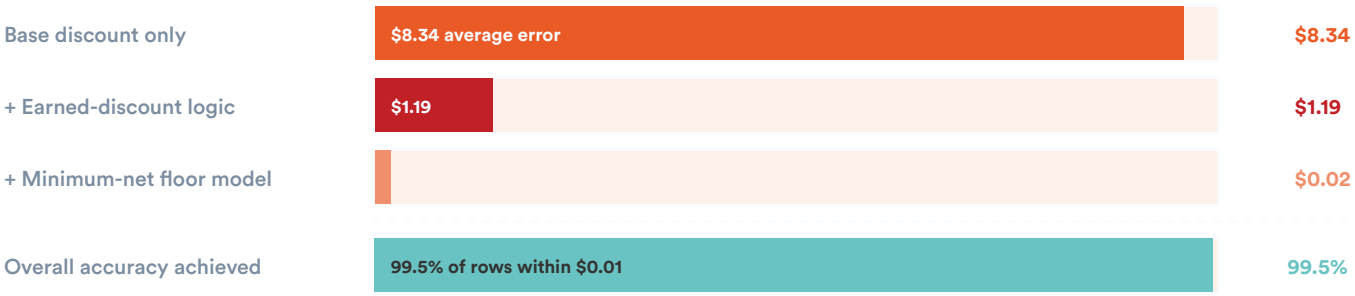
> *"A prompt-only system is not supported by the current evidence." — Independent AI engineering firm*

TEST 2

Use AI to build deterministic audit software.

The second approach used the LLM as a coding assistant. The team used agentic AI tooling to write deterministic software — code that parses invoices, normalizes charge descriptions, looks up published rates, and applies the negotiated discount stack in sequence. Accuracy came from stacking rule layers, each one cutting the error dramatically:

AVERAGE ERROR PER ROW, BY RULE L AYER ADDED



Surcharge auditing reached 99.94% accuracy within $0.01 (20,317 of 20,330 scored rows). Transportation auditing reached 98.25% — but only after layering in the earned-discount and floor logic above.

SO THE DIY CASE IS PROVEN? LOOK AT THE FINE PRINT.

What it took to get here: four weeks of specialist AI engineering, scoped to one carrier, one agreement, two invoices, and three service types, with continuous scoring against Reveel's verified answer key throughout. And after all that, these gaps remained:

TRANSPORT FLOOR LOGIC
Had to be reverse-engineered from outcomes — 47 candidate floor values inferred from patterns in the correct answers — not derived from the agreement itself.

EARNED-TIER SEQUENCING
Earned-discount tier selection, minimum-reduction sequencing, and rate-cap interaction — the order-of-operations rules for express pricing — remained incomplete.

FUEL SURCHARGE SCHEDULES
Date-indexed fuel surcharge schedules were not yet incorporated at all.

SILENT TERM-MAPPING ERROR
A surcharge interpretation applied a 50% discount where the agreement delivered 65% — silently mispricing every affected shipment.

Now multiply by every carrier, amendment, general rate increase, and mid-year rate change in a real shipping operation.

THE CRITICAL FINDING

None of it works without an answer key.

Throughout the study, the engineering team measured itself against Reveel's verified audit output — 52,289 known-correct charge lines. Every accuracy figure in this document exists because there was ground truth to compare against. The hardest logic was discovered by debugging against the answer key, not by reading the agreement.

A team building this in-house has no answer key:

You cannot measure your own accuracy
If your engine says an invoice is clean, is it clean — or is your engine wrong? Without verified output, those two situations are indistinguishable.

You cannot find your own blind spots
The expert team's biggest gaps were rules they didn't know existed until the answer key disagreed. An internal team hits the same gaps and never finds out.

Your errors look exactly like savings
Under-applied discounts report phantom overcharges. Over-applied ones hide real recovery. Both failure modes are invisible without ground truth.

The answer key wasn't a convenience. It was the load-bearing asset — and it represents what Reveel actually sells: pricing logic validated across thousands of carrier agreements and billions of dollars in parcel spend.

RESULTS AT A GLANCE

Head to head.

	PROMPT A FRONTIER LLM	DETERMINISTIC ENGINE (4-WEEK EXPERT BUILD)	REVEEL
Accuracy within $0.01	~47%	99.5%	Production-validated at scale
Avg. error per row	~$7.80	$0.007 (~1,000x lower)	Continuously verified
Error direction	Biased low — understates recovery	Residual misses small and traceable	Cites exact rate row and term
Surcharge accuracy	Unreliable on complex families	99.94% within $0.01	Included
Time to deploy	Instant — but wrong	4 weeks for 1 carrier, 1 agreement	Live in days
Audit trail	Cannot show derivation	Cites rate row and term	Cites rate row and term
Ground truth	None	Needed Reveel's answer key	Is the answer key
Maintenance	N/A	Your engineers, forever	Included — GRIs, fuel, amendments

WHAT THIS MEANS

The study didn't debunk AI. It clarified where AI belongs.

The conclusion wasn't "AI doesn't work." It was more precise: AI is the right tool for reading documents and the wrong tool for repeating math. The study's own recommendation: *"The next dollar of effort should go into deterministic transport rules, not prompt tuning."*

The justified architecture is a hybrid — AI-assisted extraction of agreement terms up front, a deterministic rules engine for millions of repeated pricing calculations, and AI again only for residual edge cases. That is the architecture Reveel has spent years building and validating.

AI

01

Agreement ingestion and term extraction

Reading messy, semi-structured carrier agreements. Extracting discounts, tiers, minimums, and caps into structured data. Interpreting ambiguous language.

DETERMINISTIC ENGINE

02

Every rate lookup and discount stack

Base discounts, earned tiers, minimums, caps, fuel, surcharge overrides — applied in the correct sequence, with a citation trail for every result.

AI + GROUND TRUTH

03

Rule libraries no DIY team has

Floor logic, tier sequencing, and surcharge mapping validated across thousands of agreements. GRIs, fuel schedules, and amendments updated continuously.

The four-week experiment, run by AI specialists with every advantage, validated the category Reveel created: parcel rate auditing is a deterministic-software problem with an AI assist — not a chatbot problem.

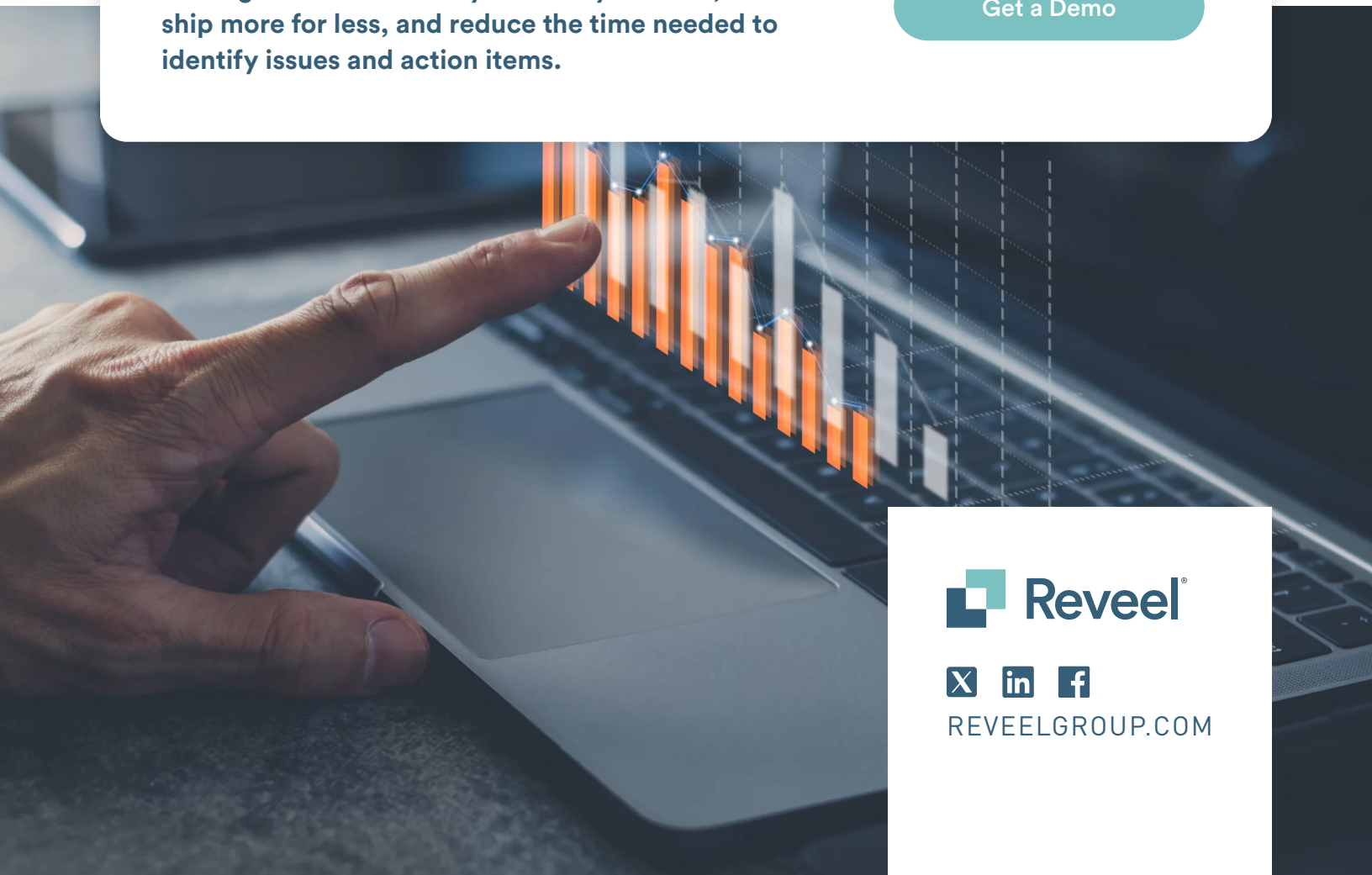
The question for a shipper isn't whether frontier models are impressive. It's whether you want to spend engineering quarters rediscovering pricing rules Reveel has already verified — with no way to check your work.

Reveel

Reveel is the leading Shipping Intelligence™ Platform that enables companies to level the playing field with their carriers across all modes of transportation. With over 18 years of parcel agreement management expertise and over $8bn in spend intelligence, the company provides actionable insights to help customers make smarter business decisions and have peace of mind. Reveel helps shippers leverage the power of data science and peer comparison data to capture significant ROI and improve their competitive advantage.

Request a demo today to see how you can leverage automation to synthesize your data, ship more for less, and reduce the time needed to identify issues and action items.

Get a Demo



Reveel

REVEELGROUP.COM