OPTIMIZED PAYMENTS

WHITEPAPER · SEPTEMBER 2026

The Data Tax

What unready data costs enterprise AI, and how to stop paying it

On raw payments data with code access, only 5% of questions met the study's success criteria. With governed calculations, accuracy reached 98% while cost per question fell by 84%.

“When payments data isn't ready for AI, every question carries a hidden cost.”

5% → 98%	84%	75%
ACCURACY, RAW + CODE TO CERTIFIED TOOLS	LOWER COST PER QUESTION	FASTER RESPONSE

optimizedpayments.com

Enterprise AI does not become reliable simply because it can access your data. In our study, the same AI model went from meeting the success criteria on 5% of payments questions to 98% as raw data was standardized, business definitions were agreed, and calculations were enforced through governed tools. At the same time, model cost per question fell 84% and response time fell 75%. The implication for finance and payments leaders is straightforward: make data, definitions, and calculations reusable and governed before scaling AI.

80%	125×	4 in 5
silent failure rate on clean data without agreed definitions; wrong numbers with no signal	lower cost per correct answer, $18.00 to $0.14	of the accuracy gain came from definitions and governed tools

What the data tax is

The data tax is what an AI pays, in tokens, dollars, seconds, and wrong answers, when the data it is given is not ready for the question. Each part is preparation that could have been done once and reused. There are three parts, each paid in a different currency.

- **Structure tax.** The work AI must do to reconcile different processor files, formats, and fields before it can answer the question. Paid mostly in tokens and seconds.
- **Semantic tax.** The risk that AI calculates the metric using the wrong business definition. Paid in wrong numbers.
- **Enforcement tax.** The risk that AI knows the agreed definition but does not apply it correctly. Paid in wrong numbers that look right. In this paper, 'certified tools' means predefined, tested calculations with governed metric definitions and known behavior, so the AI calls a computation instead of reproducing it.

KEY POINTS

01 **What brings the tax down is turning structure, definitions, and calculations into reusable capability, not asking the model to redo them each time.**

Accuracy and cost improved together because the same repeated work left every question.

02 **The most dangerous place to stop is the middle of the ladder.**

A curated semantic layer makes the AI fluent before it makes it correct: at that level it answered 97% of numeric questions, and 80% of all numeric questions came back as a wrong number with nothing to signal it.

03 **Experts do not leave; they move.**

The correctness inside a certified tool comes from people who set the definition, validated the implementation, and maintain it as the business changes. Experts move from re-explaining and re-checking the same basis to building the rules, maintaining the capability, and judging material business action.

SECTION 01

Why the data tax arises

Start with a question every finance team asks: what did we pay in processing fees last month?

It sounds simple. In payments, answering it means deciding which fee lines count, applying the agreed scope, and computing on the basis the business uses. One case from the study shows what happens when those decisions are left to the AI. Asked for one acquirer's June 2026 processing fees by category, with the data structured into one queryable layer, the AI returned a total 15% below the signed-off figure. Given the written fee definitions to read, it returned the same total again, identical to the cent. Only when the calculation ran through a certified tool did the total match the book.

EXHIBIT 1 **One question, three levels of data readiness**

WHAT THE AI HAD	PROCESSING FEES	RESULT
Data structured into one queryable layer	$191,246.32	Wrong, 15% short
Same layer, plus the written definition documents	$191,246.32, identical to the cent	Wrong, and unchanged
Same layer, calculation run through a certified tool	$224,719.49	Correct, within tolerance
Signed-off ground truth	$225,000.00	Reference

Acquirer B, June 2026, total processing fees by category. Dollar figures are scaled to keep the merchant anonymous; ratios between rows are exact. Full grading detail is in Appendix B, case 02.

The data was queryable, the definition was in hand, and the answer was still 15% short. That gap, between connecting an AI to enterprise data and getting an answer the business can act on, is what this paper measures.

Payments is the hard case for enterprise AI. Many large merchants operate across multiple payment processors and acquirers, each with its own file formats, field labels, and fee structures. Many of the metrics the business runs on - effective rate, downgrade rate, interchange savings, authorization decline rate and dispute rate - are calculated metrics. They do not simply appear in a processor file, and their

meaning depends on agreed definitions and business rules. Independent research shows the same pattern on other enterprise data, and shows that people rarely catch it. What the research cannot separate is how much of the failure is the data and how much is the model; this study is designed to do that.

What the research says

- **The failure is an enterprise-scale one.** Models that solve 91% of an academic text-to-SQL benchmark solve 21% of Spider 2.0, whose real enterprise databases often hold more than a thousand columns and require metadata, dialect documentation, and multi-step SQL [1].
- **It fails on questions that span the business.** On a 13-table enterprise schema GPT-4 answered 17% of business questions correctly and none of the questions that needed more than five tables; a semantic representation of the same database lifted accuracy to 54% [2].
- **People do not catch it, and checking costs time.** Two thirds of employees who use AI at work rely on its output without evaluating accuracy [3], and in a randomized trial experienced developers using AI tools were 19% slower on real tasks while believing they were faster [4].

Three taxes behind three failures

The research describes the symptoms. The data tax names the three causes. Each is a piece of preparation that was not done once and reused, so the AI redoes it on every question, and each produces a characteristic failure.

TAX	WHAT IS MISSING	WHAT THE AI REDOES ON EVERY QUESTION	THE FAILURE IT PRODUCES	PAID IN
Structure tax	One queryable, structured representation of every acquirer's data	Finds, joins, and reconciles files, fields, formats, and fee bases	The retry loop and the token burn: the AI spends its budget rebuilding structure rather than answering	Tokens and seconds
Semantic tax	Agreed business definitions, including basis and scope	Decides what the metric means and which interpretation the question needs	The plausible wrong number: a defensible calculation on a basis the business does not use	Wrong numbers
Enforcement tax	A validated implementation of the agreed calculation	Translates the definition into a computation, with the option of getting it wrong	The definition read but not applied: the right document is in context and the number does not change	Wrong numbers that look right

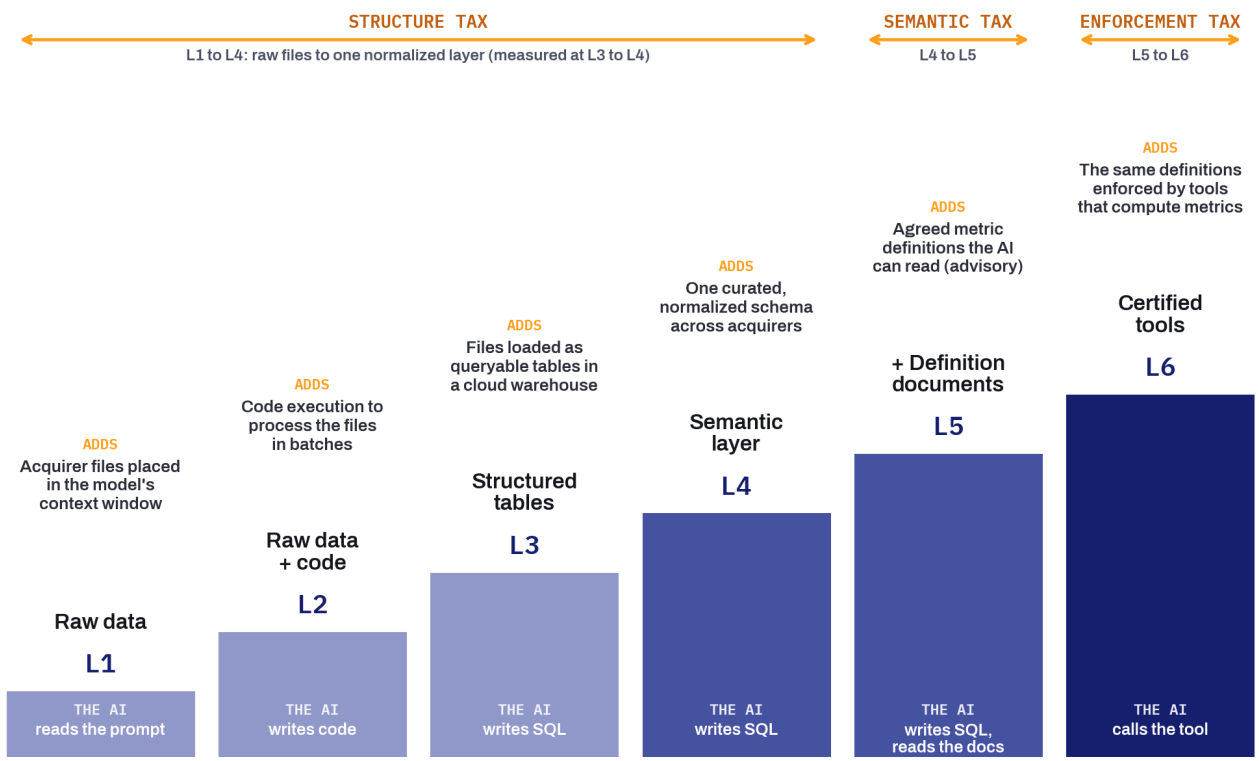
The structure tax shows up mainly in cost and latency; the semantic and enforcement taxes show up mainly in accuracy. The alternative to paying them is not a one-time project that never needs attention again. It is a reusable, maintained capability: source mappings that have owners, definitions that can be reviewed, and computations that can be tested when the data or the business rules change. What the study measures is what the capability refunds, one step at a time on a six-level ladder from raw files to certified tools. The next section defines the ladder.

SECTION 02

Six levels of data readiness

The study defines six levels of data readiness. Each level adds exactly one kind of upfront work, and the task left to the AI changes with it. The three taxes are the differences between levels: the structure tax is the distance from raw files to one structured layer (L1 to L4, measured on the same questions at L3 to L4), the semantic tax is definitions added (L4 to L5), and the enforcement tax is definitions enforced (L5 to L6).

EXHIBIT 2 **The readiness ladder: what each level adds and what the AI does**



L1 places raw files directly in the model's context and is a separate feasibility test on one acquirer's book: a week of files exceeded the 1M-token context, and at the largest slice that fit the AI returned cannot-answer on 29 of 30 questions. L2 to L6 form the main 80-question comparison.

How the study measured

HELD FIXED

One model (Claude Sonnet 5, thinking disabled). 80 pre-registered questions, 60 numeric and 20 root-cause, across authorization, fees, sales, interchange, and disputes. Two acquirers and one enterprise merchant. May to June 2026 data. One rubric with signed-off ground truth.

ACCURACY

Numeric answers are graded correct or not within 0.5% of ground truth, with no partial credit; where an acquirer supplies no data, saying so counts as correct. Root-cause answers are scored 0 to 5 with partial credit for the numeric and causal parts, and count as correct only at 5 of 5.

COST

Model usage per question at published Claude Sonnet 5 rates ($2 per million input tokens, $10 per million output; five-minute cache writes $2.50 and reads $0.20), the rate in effect when the runs were executed in July 2026 and unchanged at publication. All cost figures in this paper use these rates. Every model call within a question is counted, including self-corrections. Build and maintenance of each level are excluded.

LATENCY

Mean wall-clock seconds from question to final answer, including tool calls and SQL execution. Latency is reported alongside accuracy and cost at every level because a slow correct answer still delays a decision.

The design asks one focused question: with the model held fixed, how much does data preparation and execution change answer quality, model usage, and response time? It does not compare models, and it does not claim that model choice is unimportant. It isolates the part of the result that data work alone can move.

SECTION 03

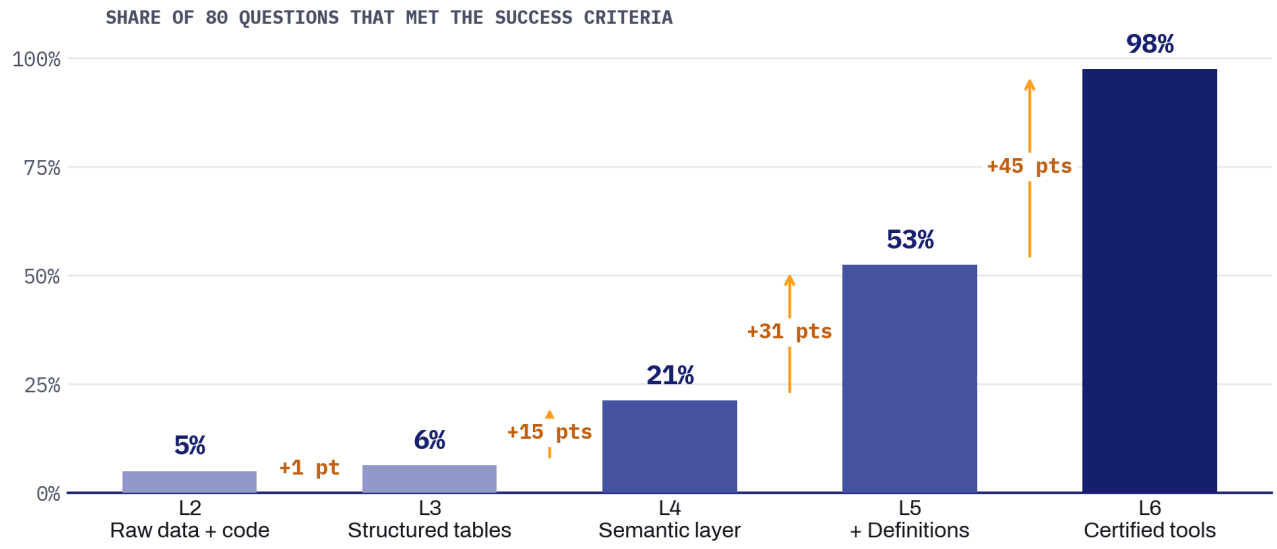
What we found

TAKEAWAY 1

Clean data helps AI answer. Agreed definitions and enforcement help it answer correctly.

Structure alone barely moved accuracy. Raw data with code met the study's success criteria on 5% of the 80 questions; loading the same files into a cloud warehouse added one percentage point; curating them into one semantic layer (a standardized data model with consistent fields across processors) added 15. The two large steps are the two that add meaning: agreed definition documents added 31 points, and enforcing those definitions through certified tools added 45 more, finishing at 98%. The same shape holds on the 60 numeric questions alone, where certified tools answered 59 of 60.

EXHIBIT 3 **Accuracy by readiness level**



Share of 80 pre-registered questions that met the success criteria: a numeric answer within 0.5% of ground truth (or a correct statement that the data does not exist), or a root-cause answer scored 5 of 5. Step gains in percentage points.

Reading the ladder step by step shows what each kind of preparation buys. Structuring cuts the structure tax in dollars and seconds, and only modestly in accuracy: one curated schema cut cost per question from $1.10 to $0.47, cut response time from 40 to 18 seconds, and cut SQL errors from 46 to 9, while accuracy

moved from 6% to 21%. Definitions reduce the semantic tax almost entirely in accuracy, for four cents. Enforcement improves accuracy and lowers cost together, and it removes an error class that definitions alone could not, which the next finding traces.

EXHIBIT 4 **The three taxes, step by step**

TAX AND STEP	ACCURACY	COST PER QUESTION	RESPONSE TIME	WHAT THE STEP BOUGHT
Structure: one curated schema (L3 to L4)	6% → 21%	$1.10 → $0.47 (−57%)	40 s → 18 s	Dollars and seconds; SQL errors fell from 46 to 9
Semantic: definitions (L4 to L5)	21% → 53%	$0.47 → $0.51 (+4¢)	18 s → 16 s	Accuracy; the AI queries the right field first time, SQL errors 9 to 1
Enforcement: certified tools (L5 to L6)	53% → 98%	$0.51 → $0.14 (−73%)	16 s → 15 s	Accuracy and dollars; a 64K-token tool surface replaces a 249K-token context

Accuracy is share of 80 questions meeting the success criteria; response time is mean seconds per answer.

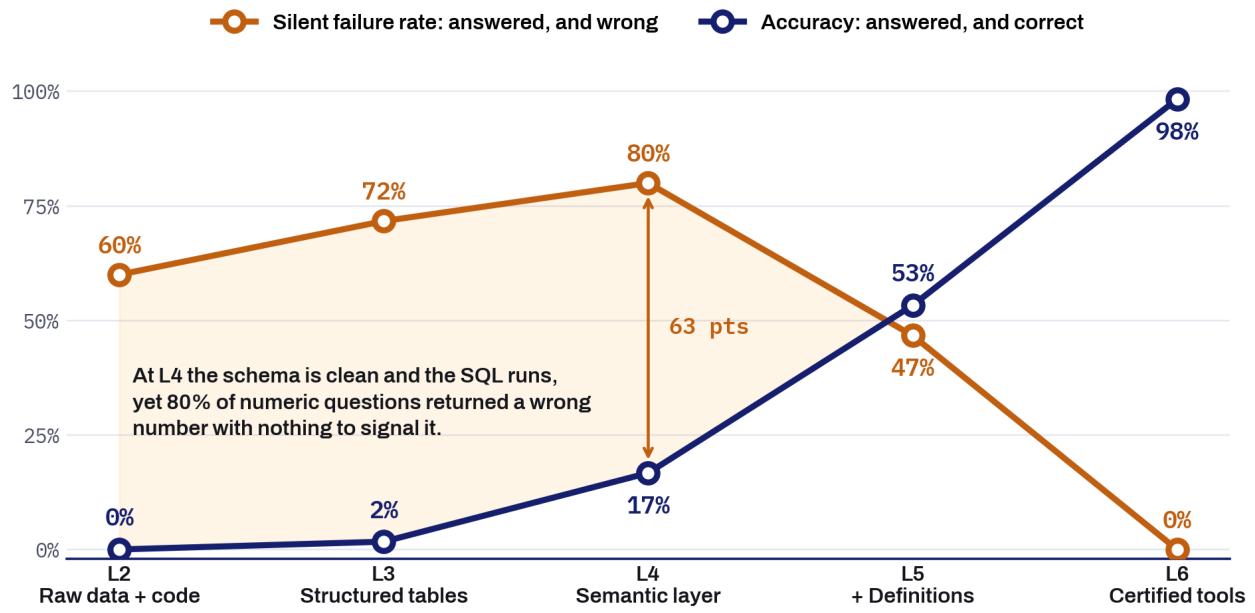
Structure is the floor, not the result. Structuring the data made it easier to use. It did not tell the AI which answer the business meant: at the semantic layer, four numeric answers in five were still wrong.

A definition the AI can read is not a definition it will apply. Definitions did most of the work: on Acquirer A's effective rate the AI was wrong on all levels except for L5 and L6, because once the fee-base definition was in context it applied it (Appendix B, case 01). The fee question that opened this paper shows the remaining gap: for Acquirer B the same definitions were in context, and the answer did not change (Section 01; Appendix B, case 02). A definition document tells the model how a metric should be computed, and the model decides each time whether to follow it. A certified tool turns the computation itself into a tested, governed capability the model calls instead of reproducing, and the choice disappears. That is the enforcement tax: wrong numbers that look right.

TAKEAWAY 2

The most dangerous AI answer is a confident wrong answer. Structured data without definitions produces the most of them.

EXHIBIT 5 **Silent failure rate and accuracy by readiness level**



Silent failure rate is coverage times selective risk; a partially correct multi-part answer counts as wrong because the business received a number that did not meet the standard. Sensitivity in Appendix C: if partial answers are excluded from the wrong count, the L4 rate is 62% and the peak moves to L3 (72%).

Accuracy measures how often the AI was right. It does not measure how often the AI was wrong without saying so, and for a business that acts on the number, that is the risk. The study therefore also reports the silent failure rate: the share of all 60 numeric questions that came back as a number that did not meet the standard.

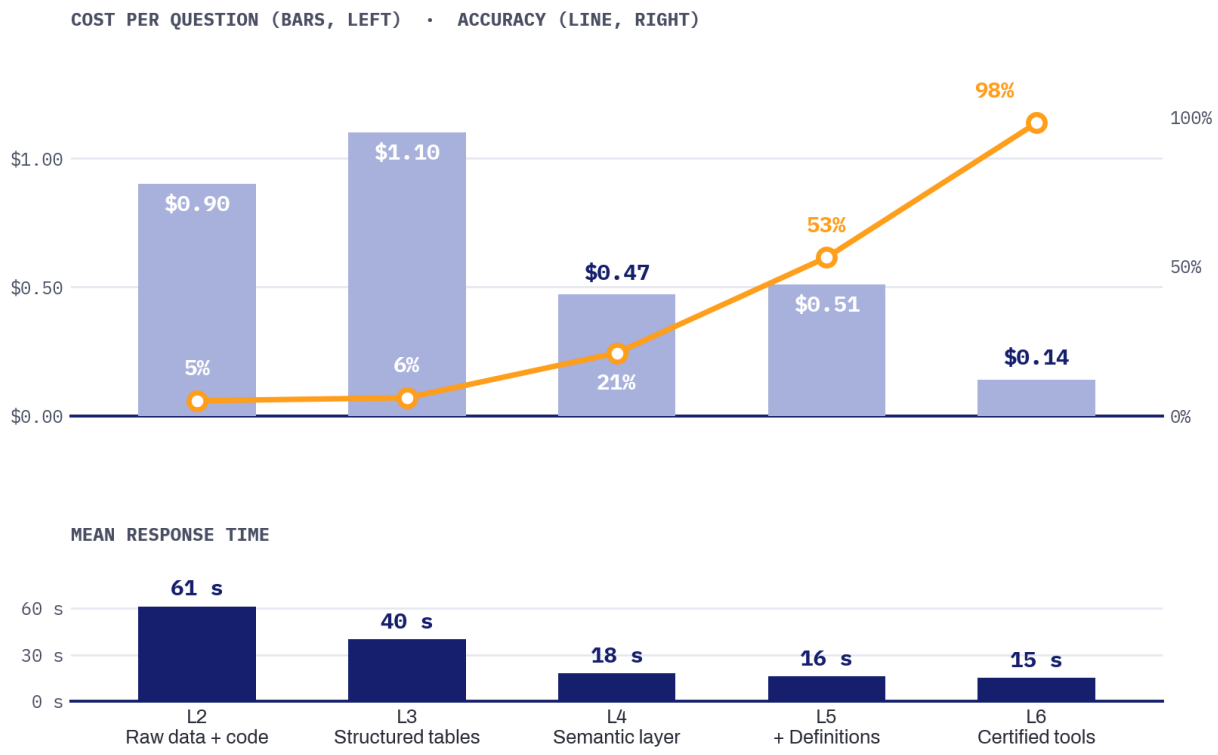
The silent failure rate did not fall as the ladder rose. At the semantic-layer stage, AI returned a number for 97% of numeric questions, but 80% of all numeric questions were wrong without any warning. Once definitions were added, the silent failure rate fell to 47% and reached 0% with certified tools. The lesson is the gap between answering and answering correctly. As data is prepared, an AI becomes willing to return a number before it becomes reliable at returning the right one, so the share of questions answered is not a measure of quality. **Measure the answers, not the answer rate.**

TAKEAWAY 3

Accuracy improved while model cost per successful answer fell about 125-fold.

A wrong answer has little business value, so a more useful unit is cost per correct answer. In the study, that fell from $18.00 on raw data with code to $0.14 with certified tools, about 125 times lower. Cost per question fell about six times, from $0.90 to $0.14, while questions meeting the study's success criteria increased from 4 of 80 to 78 of 80. The ratio uses the same model and pricing at every level.

EXHIBIT 6 **Cost per question, accuracy, and response time by readiness level**



Accuracy as in Exhibit 3. Latency is mean wall-clock seconds per answer.

Accuracy and cost moved together, not against each other. The tokens the AI spends rebuilding structure, deciding definitions, and reproducing computations are both the cost and the source of error, so removing that work lowers both at once, and the ladder never traded one for the other. The one step that raised cost was loading raw files into queryable tables without curating them, from $0.90 to $1.10 per question, because the AI still carried an 863K-token file specification into every call to reconcile them. Definition documents added four cents and 31 accuracy points; certified tools then cut cost 73%, to $0.14,

while lifting accuracy to 98%, because a 64K-token tool surface replaced a 249K-token context of schema and definitions.

LEVEL	COST PER QUESTION	CORRECT OF 80	COST PER CORRECT ANSWER
L2 Raw data + code	$0.90	4	**$18.00**
L3 Structured tables	$1.10	5	**$17.60**
L4 Semantic layer	$0.47	17	**$2.21**
L5 + Definition documents	$0.51	42	**$0.97**
L6 Certified tools	$0.14	78	**$0.14**

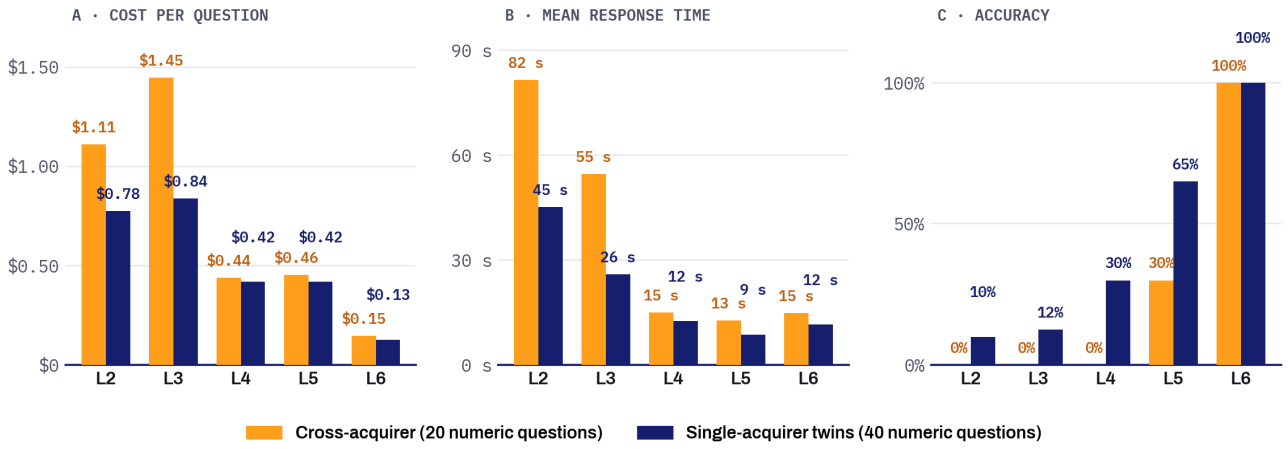
Total cost is the displayed cost per question times 80. This is cost per success under the rubric, not the cost of retrying until a usable answer arrives. L2's four successes are attributed abstentions; on the 60 numeric questions alone L2 has no matching answers and the ratio is undefined.

TAKEAWAY 4

The pattern holds where it is hardest: reconciling across acquirers is where the tax is highest and where preparation pays off most

Twenty of the 60 numeric questions ask for a figure across both acquirers: Acquirer A, Acquirer B, and the combined book. Each is paired with two single-acquirer versions of the same metric, so the study can isolate what combining acquirers adds. On raw data, cross-acquirer questions cost 43% to 73% more and took 81% to 111% longer than their single-acquirer versions, and the AI answered none of them correctly. The semantic layer nearly eliminated the cost premium, but accuracy stayed low.

EXHIBIT 7 **Cross-acquirer premium versus single-acquirer twins**



Cross-acquirer is the 20 numeric questions asking for Acquirer A, Acquirer B, and the combined figure; twins are the 40 numeric questions asking the same metric for one acquirer. Root-cause questions are excluded from both groups so the two sets have the same composition.

Adding definitions lifted accuracy to 6 of 20. Certified tools answered all 20 correctly, at $0.15 per question, nearly the same cost as the single-acquirer questions. The gain came from a single, consistent way to interpret and calculate across acquirers, not from access to their records. The study tested two acquirers. Each additional acquirer likely adds to the cost of reconciling unprepared data, but the mechanism that removed the premium still applies.

SECTION 04

What it means for your roadmap

The results map to a roadmap, and the order of the work matters as much as the work itself. Structure first, agree definitions second, certify third is the sequence we recommend. It is not the only technical order possible, but each step's payoff in the study depended on the one before it: definitions were applied to a curated schema, and tools enforced definitions that had already been agreed. Three things follow: who supplies the correctness inside the tools, what the reuse is worth at scale, and where to begin.

Where the correctness comes from: experts above the tools

Moving a calculation out of the model raises the question a buyer should ask next: who establishes that the calculation inside the tool is the right one? It is not the tool. The mapping from every acquirer's raw fields to one curated schema, the agreed definitions with their fee bases and scopes, and the tested computations are encoded expertise, and the study's own results show where that encoding stops.

Interchange was the only domain not fully solved, at 14 of 16 at the certified-tools level. The two remaining misses are root-cause questions on which the numbers were right and the explanation was not. The acquirer's downgrade reasons and qualification rules are not in any file, any definition document, or any tool; they are in the heads of interchange specialists. Closing that gap is expert work of a new kind: building the knowledge base of rules, reasons, and exceptions that the AI can draw on for deep root-cause analysis, so that the same expertise that certifies a metric also explains why it moved. A tool computes what has been certified into it and nothing more.

That is the operating model this paper recommends, and it places experts at three points.

WHERE EXPERTS PARTICIPATE	THEIR RESPONSIBILITY	WHAT BECOMES REUSABLE
Before execution	Agree the metric's meaning, basis, scope, sources, and valid exceptions	A definition that can be read, reviewed, and owned
In validation and maintenance	Verify source mappings and tool behavior, investigate discrepancies, assess the effect of changes in data or rules	A tested analytical capability with an owner
At interpretation and action	Review material findings, weigh explanations and options, approve consequential action, feed new knowledge back into the definitions	Better decisions, and a capability that improves

This is a recommended operating model, not a condition the study measured. The goal is to move repeatable work into a controlled capability while keeping people responsible for meaning, exceptions, and action. Experts would remain necessary even if all 80 questions had passed: someone still decides what a metric should mean, validates changes, and judges whether a finding warrants action. Certified tools make agreed expertise reusable; they do not make it unnecessary. The domain results show where that expertise earns the most.

EXHIBIT 8 **Accuracy by domain and readiness level (correct of 16)**

DOMAIN	L2	L3	L4	L5	L6	READING
Sales	0	0	6	11	16	Fully resolved by certified tools
Authorization	0	1	2	14	16	Definitions do most of the work
Fees	0	0	0	6	16	Zero until definitions arrive; 10 of 16 still missed at L5, all resolved at L6
Interchange	4	4	6	8	14	Improves level by level; the only domain not fully solved, where expertise earns its keep
Disputes	0	0	3	3	16	Enforcement drives the result, on a taxonomy experts had to correct first

Each domain holds 16 questions, 12 numeric and 4 root-cause.

Fees and interchange carry the heaviest tax because they are the most derived: an effective rate needs a fee base and a downgrade rate needs interchange reference data, and neither sits in an acquirer file. Those are the domains where definitions and certified tools return the most, and where the experts who write and maintain them matter most.

What the tax costs at scale: the machine pays first

The study measured the tax the machine pays: model usage per question. That figure is the same unit an enterprise is billed in, so it scales directly. The panel applies the measured difference between raw data with code ($0.90) and certified tools ($0.14), $0.76 per question, to hypothetical monthly volumes. These are modeled illustrations, not measured savings, and a question means one business question processed end to end, including every model call within it.

EXHIBIT 9 **Modeled annual model-usage difference by monthly volume**

AI QUESTIONS PER MONTH	MONTHLY COST, RAW DATA + CODE	MONTHLY COST, CERTIFIED TOOLS	ANNUAL DIFFERENCE, STANDARD RATES (UPPER BOUND)	ANNUAL DIFFERENCE, 50% DISCOUNT (LOWER BOUND)
1,000	$900	$140	$9,120	$4,560
5,000	$4,500	$700	$45,600	$22,800
10,000	$9,000	$1,400	$91,200	$45,600
50,000	$45,000	$7,000	$456,000	$228,000

Modeled, not measured.

Use the right baseline. Not every organization starts at raw data, and the whole raw-to-certified difference should not be claimed by default. From structured tables the difference to certified tools is $0.96 per question ($57,600 a year at 5,000 questions); from a semantic layer $0.33 ($19,800); from definition documents $0.37 ($22,200). Use the row that resembles the workflow today.

Discount scenario. Enterprises rarely pay standard rates, and negotiated volume pricing is not published, so two published prices bracket the figure: standard rates, which every figure in this paper uses, and the Batch API at 50% of standard [7], the deepest published discount. At 5,000 questions a month from a raw-data start, the annual difference is between $22,800 and $45,600. Any uniform discount scales both levels by the same factor, so the six-to-one ratio per question and the accuracy gap hold at any price.

The organization pays second: the human tax

Model cost is the tax the study measured. The tax it did not measure may be larger: at the semantic layer, 80% of numeric answers were wrong numbers delivered with no signal. Every one of those either gets caught by a person or gets acted on. Independent surveys suggest the catching is uneven: about a quarter of organizations using generative AI review every output before use while a similar share review a fifth or less [8], and two thirds of employees who use AI at work rely on its output without checking it [3]. Either way the organization pays: in analyst hours spent reconciling, or in decisions made on wrong numbers. This study did not measure that cost. The panel shows the order of magnitude on stated assumptions.

At 5,000 questions a month, every one-minute reduction in average additional manual effort per question is 1,000 analyst hours a year.

EXHIBIT 10 **Illustrative human reconciliation sensitivity**

AVERAGE ADDITIONAL MANUAL EFFORT AVOIDED PER QUESTION	HOURS RELEASED PER YEAR	AT $75 PER HOUR	AT $100 PER HOUR
0.5 minute	500	$37,500	$50,000
1 minute	1,000	$75,000	$100,000
2 minutes	2,000	$150,000	$200,000
2.25 minutes (the 50% → 5% scenario)	2,250	$168,750	$225,000
3 minutes	3,000	$225,000	$300,000

At 5,000 questions a month. Labor rates of $75 and $100 an hour are assumptions. "Manual effort" means additional validation, reconciliation, and correction, averaged across all questions. The 50% → 5% scenario assumes 50% of answers in a raw-data environment and 5% in a governed one need reconciliation at five minutes each; neither rate was measured, and neither is a recommended review policy. Formulas in Appendix D.

Two cautions apply. Review rates are set by policy, not error rates, so a well-governed system may review more answers, and reduce time per review instead. Released capacity becomes cash only where it replaces specific outsourcing, overtime, or hiring.

Reading the two panels together. At 5,000 questions a month from a raw-data starting point, the measured model-usage difference is $45,600 a year. Every minute of average reconciliation time saved per question adds $75,000 to $100,000 of analyst capacity at the stated rates. The model tax is measured, the human tax is modeled, and both exclude the cost of building and maintaining the governed layer. Reducing repeated preparation and calculation work improves model economics on the study's evidence, and reducing repeated reconciliation may add organizational value that has to be measured at its own boundary.

Start with one workflow, not the whole enterprise

An enterprise does not sit on one rung. Finance may query a warehouse while operations runs scripts over exports, and different metrics have different gaps. Pick one question that matters to the business and is reconciled by hand today, a processing-fee change, an effective-rate comparison, an approval-rate analysis, and choose it because it has a clear owner and a real decision behind it. Find which tax that workflow pays, close that gap, and check whether answer quality, model usage, and reconciliation hours actually moved. Then scale. The ladder is a diagnostic, not a maturity score.

Three questions place any workflow. The first question answered "no" is the tax being paid today, and the row to act on.

EXHIBIT 11 **Which tax is this workflow paying?**

	ASK	IF NO, IT IS PAYING · HOW IT SHOWS UP	DO THIS NEXT
1	Are every acquirer's files structured into one curated schema, with one field for each concept?	**Structure tax** Exports pasted into prompts; scripts over raw files; a warehouse of raw tables. Slow answers, high token bills, SQL that hunts for the right field.	Structure into one curated layer. In the study: cost per question $1.10 → $0.47, response time 40 s → 18 s.
2	Are the business definitions of every metric written down and agreed, including fee base and scope?	**Semantic tax** Clean schema, fast SQL, and answers that do not match the dashboard. Two analysts, two effective rates. Nobody can say which definition the AI used.	Agree and publish definitions before opening the layer to AI questions. In the study: accuracy 21% → 53%; silent failures 80% → 47%.
3	Are those definitions enforced by tools that compute the metric, rather than read by the model and applied at its discretion?	**Enforcement tax** Definitions exist, but numbers still move between runs and between analysts. Right document, wrong application.	Certify the definitions as tools the AI calls. In the study: accuracy 53% → 98%, cost $0.51 → $0.14, silent failures 47% → 0%.

Study figures are for one enterprise merchant, two acquirers, and 80 questions; they show the operating impact of each step, not the price of building it.

Close the loop from question to action. Return to the fee question that opened this paper. The right total is the beginning of a useful workflow, not the whole of it. The workflow establishes the fee total on the agreed basis, compares it with the prior period on the same basis, finds where the change is concentrated, has an expert assess whether the finding is material, and only then decides whether to investigate further or act. Any new definition question, missing-data case, or recurring exception feeds back into the maintained capability.

The study did not measure the outcome of that workflow. It measured the analytical foundation the workflow requires: whether the number on the agreed basis can be produced reliably, cheaply, and fast. A dependable calculation is a better basis for investigation and judgment; it replaces neither.

Does your domain pay the same tax?

Payments is the evidence here, but the mechanism is not specific to payments. Ask one question: does your headline metric sit in a column, or is it computed from a basis that lives in a dashboard or an analyst's head? If the second, the mechanism this study measured applies to you. Effective rate is to payments what loss ratio is to insurance, net revenue retention to software, and cost per case to healthcare claims. None of them is a column.

Limitations and disclosures

- **Vendor-authored.** Optimized Payments builds and sells the curated semantic layer and the certified tools measured at the higher-readiness levels. The study was designed to test that investment case, and readers should weigh the findings with that interest in mind.
- **Root-cause grading.** Root-cause answers are consultant-graded. At the top level the numeric component was auto-verified for all 20 and the causal component was carried from the consultant grading of the earlier run. The 78 of 80 headline is therefore a composite of numeric matches, attributed abstentions, and root-cause answers graded that way.
- **Scope.** One merchant book, one model, one two-month window. Cost is model usage only at published Claude Sonnet 5 rates and excludes the initial build and ongoing maintenance of each level; user re-asks are not modeled. The structure tax is defined as L1 to L4, and only its final step, L3 to L4, is measured on the same 80 questions as the rest of the ladder. Scale-up and labor figures are modeled on stated assumptions and are gross scenarios, not net returns; no production review-time saving, decision outcome, or return on investment was measured.

APPENDIX A

Methodology

Controls held fixed

Every level ran the same model (Claude Sonnet 5, thinking disabled), temperature and scaffold, question wording, May to June 2026 data window, and grading rubric. Only data readiness changed between levels, so the differences between levels can be read as the price of that readiness. L2 to L6 form the main comparison; L1 is a separate feasibility test, described below.

Pre-registration

The 80 questions, the ground-truth answer for each, and the grading rubric with numeric tolerances were fixed before any run. Comparison and root-cause questions compare June 2026 against May 2026. Corrections to ground truth made after pre-registration were logged with the date and reason, and every level was graded against the corrected values in the final graded matrix, so no level was scored against a different reference from another.

Grading rubric

- Numeric match within 0.5% relative tolerance, or exact match to ground truth rounded at the precision the model reported; where an answer states its own underlying figures at higher precision, grading uses the highest-precision stated value for the asked quantity. Unit-aware (fraction or percent, basis points or percent). No partial credit for a single-figure answer.
- Cross-acquirer questions require Acquirer A, Acquirer B, and the combined figure to match to count as correct; a multi-part answer with at least one component matching is classed partial and does not count. Breakdown questions require the stated total or the sum of components to match.
- Data that does not exist. Acquirer B supplies no interchange data, so for its interchange questions an explicit zero and an attributed cannot-answer are both counted as correct. This is a grading convention for the one acquirer without an interchange feed, not an observed rate of zero.
- Root cause. The 20 root-cause questions were consultant-graded on numeric accuracy (0 to 3) and causal correctness (0 to 2), with partial credit on each part; a question counts as correct in the headline only at 5 of 5. At L6 the numeric component was auto-verified against ground truth for all 20 and causal correctness was carried from the consultant grading.
- Other outcome classes. Answers were also classed incorrect, declined (the AI said it could not answer without a valid attribution to missing data), no answer (three L3 questions where the AI reached its tool-step limit), and log truncated (seven L2 questions where the run log was cut off before a final figure was captured). All count as not correct and remain in every denominator; their token usage is included in cost as recorded.

Token and cost accounting

Input, output, cache-write, and cache-read tokens were read from the API usage object for every call and priced at the published Claude Sonnet 5 rates: $2 per million input tokens, $10 per million output tokens, $2.50 per million five-minute cache writes, and $0.20 per million cache reads [7]. The runs were executed in July 2026 at these rates, then the model's introductory pricing; Anthropic made them the standard rates in August 2026, so the run-time and publication-time prices coincide. Token counts, not dollars, are the measured quantity, and the per-level totals in Appendix C can be repriced at any rate. Every model call within a question's recorded attempt is included, so self-corrections after SQL or tool errors are already in the per-question figure; user re-asks were not simulated. L2 cost is the last recorded attempt per question; 22 questions were re-run during the L2 rung and the earlier attempts are excluded. L3 tokens are 99.4% cache reads and are priced as such. The initial build and ongoing maintenance at each level are excluded.

Token figures in the body are mean uncached input tokens per question summed across the calls within each question (about 249K at L5 and 64K at L6), except the raw-file specification sizes (about 863K tokens for both acquirers at L3 and 337K for Acquirer A at L1), which are the lengths of the specification documents held in context, not a further cost category to add to the per-question figure. Mean total tokens per question (Appendix C) combine differently priced token categories and should not be multiplied by a single rate.

Treatment of L1

The raw-files-in-context level is a context-breakpoint measurement on the Acquirer A book only (30 questions, three repeats). A seven-day slice of files reached 1,130,227 input tokens and a six-day slice 1,045,485, both above the 1M-token limit; a five-day slice fit. At that slice the AI returned cannot-answer with a specification-grounded diagnosis on 29 of 30 questions, consistently across repeats. L1 is excluded from ladder averages.

APPENDIX B

Three worked examples

Three questions traced through the ladder: what the AI answered at each level, how the rubric graded it, and which tax the case exposes. Dollar figures in case 02 are scaled to keep the merchant anonymous; ratios between rows are exact. All rates and percentages are as recorded.

01 How definitions correct the fee basis

Question: Acquirer A effective rate for June 2026.

LEVEL	ANSWER	GRADED
Ground truth	240.1 bps	**Reference**
L2 Raw data + code	10.7 bps	**Wrong**
L3 Structured tables	58.8 bps, computed on net sales	**Wrong: wrong fee base**
L4 Semantic layer	308 bps	**Wrong**
L5 + Definition documents	240.1 bps	**Correct**
L6 Certified tools	240.1 bps	**Correct**

This is the semantic tax paid and refunded. Fees over gross sales and fees over net sales are both defensible computations; only the definition says which one the business uses. Once the definition was in context, the AI applied it, and the certified tool computed the same figure.

02 The definition read but not applied

Question: Acquirer B total processing fees by category for June 2026.

LEVEL	PROCESSING FEES	GRADED
Ground truth	$225,000.00	**Reference**
L2 Raw data + code	Declined the category split	**Wrong**
L3 Structured tables	$146,180.00, blended	**Wrong**
L4 Semantic layer	$191,246.32	**Wrong, 15% short**
L5 + Definition documents	$191,246.32, identical to L4	**Wrong, unchanged**
L6 Certified tools	$224,719.49	**Correct, within tolerance**

Dollar figures scaled to preserve anonymity; ratios between rows are exact.

The definition document changed nothing about which fee lines the AI counted for this acquirer. Reading a definition is not enforcing it. This is the case that opens Section 01. Across the fees domain, 10 of 16 questions were still missed at L5, and every one resolved at L6.

03 The cross-acquirer twin pair

Question: overall approval rate for June 2026 for Acquirer A, Acquirer B, and the combined book.

LEVEL	ACQUIRER A	ACQUIRER B	COMBINED	GRADED
Ground truth	32.88%	64.54%	33.40%	**Reference**
L2 Raw data + code	48.6%, reinvented basis	n/a	n/a	**Wrong**
L3 Structured tables	36.59%	64.53%	37.17%	**Wrong: one of three**
L4 Semantic layer	36.3%	64.5%	36.6%	**Wrong: one of three**
L5 + Definition documents	32.88%	64.54%	33.36%	**Correct**
L6 Certified tools	32.88%	64.54%	33.36%	**Correct**

At L2 only the Acquirer A figure is recorded. A cross-acquirer question counts as correct only when all three figures match.

The single-acquirer twins tell the same story: Acquirer A's approval rate was wrong at L2, L3, and L4 and right from L5, while Acquirer B's was right from L3. The combined answer inherits the harder acquirer's difficulty. Structuring the data removed most of the cost premium but not the definitional error; definitions and enforcement removed the error.

The three cases share one pattern. At the raw levels the AI is wrong or silent for structural reasons: it cannot find, join, or trust the fields it needs. At the semantic layer it computes fluently on the wrong basis, the most dangerous outcome because the answer looks finished. Definitions correct the basis when they are followed, and case 02 shows they are not always followed. Certified tools remove the choice.

APPENDIX C

Summary results matrix

One row per readiness level, L2 to L6. Correct of 80 counts numeric answers within tolerance, attributed abstentions where the data does not exist, and root-cause answers at 5 of 5. Cost is mean per question at published Claude Sonnet 5 rates; latency is mean seconds; tokens are mean total per question; SQL errors are totals over 80 questions.

LEVEL	CORRECT OF 80	NUMERIC CORRECT OF 60	ROOT CAUSE 5 OF 5, OF 20	COST PER QUESTION	LATENCY (S)	TOKENS PER QUESTION	SQL ERRORS
L2 Raw data + code	4	4	0	$0.90	61	3.42M	n/a
L3 Structured tables	5	5	0	$1.10	40	5.11M	46
L4 Semantic layer	17	12	5	$0.47	18	229K	9
L5 + Definition documents	42	32	10	$0.51	16	250K	1
L6 Certified tools	78	60	18	$0.14	15	65K	n/a

SQL errors do not apply to L2 (code) or L6 (certified tools); L6 recorded zero tool errors. L3 tokens are 99.4% cache reads. Numeric correct of 60 includes one attributed abstention at L6 (59 without it), four at L2 and L3, and two at L4. L1 (raw files in context) is excluded from the ladder: on the Acquirer A book it returned cannot-answer on 29 of 30 questions at the largest slice that fit the 1M-token context.

Outcome classes and root-cause scores

MEASURE	L2	L3	L4	L5	L6
Root-cause average score (0 to 5)	1.55	2.20	2.85	4.00	4.80
Root-cause numeric-perfect (3 of 3, of 20)	3	3	7	13	20
Numeric correct (of 60)	0	1	10	32	59
Numeric partial (of 60)	0	0	11	11	0
Numeric incorrect (of 60)	36	43	37	17	0
Numeric declined (of 60)	13	9	0	0	0
Numeric no answer (of 60)	0	3	0	0	0

Selective prediction measures, numeric questions

MEASURE (OF 60 NUMERIC)	L2	L3	L4	L5	L6
Answered with a number (coverage)	36 (60%)	44 (73%)	58 (97%)	60 (100%)	59 (98%)
Declined, abstained, or no final answer	17	16	2	0	1
Wrong among answered (selective risk)	100%	98%	83%	47%	0%
Silent failure rate (coverage × risk)	60%	72%	80%	47%	0%
Abstention recall	32%	27%	4%	0%	100%

Coverage and selective risk follow El-Yaniv and Wiener (2010) and Geifman and El-Yaniv (2017) [5, 6]; partial answers count as wrong; abstention recall is abstentions over abstentions plus wrong answers. At L4, 48 of 60 is 80% of all numeric questions; among the 58 answered it is 83% (the selective risk row). Both are reported so the two denominators are not confused. Excluding L2's seven truncated logs from the denominator gives 68% at L2, still below L4. Treating partial answers as neither correct nor wrong gives silent failure rates of 60%, 72%, 62%, 28%, 0% (peak at L3) and abstention recall of 32%, 27%, 5%, 0%, 100%. The three classes mirror the correct, incorrect, and not-attempted grading of SimpleQA [9].

APPENDIX D

Economic formulas and the reconciliation scenario

D1. Model usage. Let Q be monthly business questions and Cb, Cg the model cost per question at the baseline and governed configurations. Annual model-cost difference = 12 × Q × (Cb − Cg). At Q = 5,000 with Cb = $0.90 and Cg = $0.14: $45,600. With Cb from L3 ($1.10): $57,600; from L4 ($0.47): $19,800; from L5 ($0.51): $22,200.

D2. Average manual effort. Let Δt be the assumed reduction in average additional manual minutes per question and W the assumed hourly labor cost. Annual hours released = 12 × Q × Δt ÷ 60. Annual capacity value = hours released × W. These formulas value time; they do not establish that the time will be released or turned into cash.

D3. The 50% → 5% scenario. Assume 5,000 monthly questions, additional reconciliation on 50% of raw-environment answers and 5% of governed-environment answers, and five minutes per reconciliation. Fewer reconciliations per month = 5,000 × (0.50 − 0.05) = 2,250. Hours per month = 2,250 × 5 ÷ 60 = 187.5; per year 2,250, an average of 2.25 minutes per question. At $75 an hour $168,750; at $100 an hour $225,000. Where review duration differs between environments, the monthly formula is Q × (rb × tb − rg × tg) ÷ 60, with r the reconciliation incidence and t minutes per reconciliation. None of the behavioral assumptions was measured.

D4. Investment boundary. For a cash case, compare model spend actually avoided and identifiable labor spend actually avoided with the incremental cost of implementation, licensing, infrastructure, data maintenance, tool validation, and expert governance. For a capacity case, report released hours and their value separately and say how they will be used. Do not treat response-time reductions as recoverable analyst hours, infer review incidence from benchmark error rates, or add hypothetical avoided errors and faster decisions without separate evidence.

REFERENCES

Sources cited

External sources are cited for the pattern they establish. Their figures come from other models, schemas, and question sets and were not re-run by Optimized Payments. All study figures come from the Data-Tax Study graded matrix and run files.

[1] Lei, F., Chen, J., Ye, Y., et al. (2025). Spider 2.0: Evaluating language models on real-world enterprise text-to-SQL workflows. International Conference on Learning Representations (ICLR 2025). arXiv:2411.07763. Reports 21.3% task success for the authors' o1-preview code-agent framework on Spider 2.0 against 91.2% on Spider 1.0 and 73.0% on BIRD.

[2] Sequeda, J. F., Allemang, D., and Jacob, B. (2024). A benchmark to understand the role of knowledge graphs on large language model's accuracy for question answering on enterprise SQL databases. ACM GRADES-NDA 2024. arXiv:2311.07509. Reports 16.7% zero-shot accuracy for GPT-4 on an enterprise insurance schema, 0% on both high-schema-complexity question quadrants, and 54.2% over a knowledge-graph representation of the same database.

[3] Gillespie, N., Lockey, S., et al. (2025). Trust, attitudes and use of artificial intelligence: A global study 2025. University of Melbourne and KPMG. 48,340 respondents in 47 countries, November 2024 to January 2025. Reports that 66% rely on AI output without evaluating accuracy and 56% have made mistakes in their work due to AI.

[4] Becker, J., Rush, N., Barnes, E., and Rein, D. (2025). Measuring the impact of early-2025 AI on experienced open-source developer productivity. METR. arXiv:2507.09089. Randomized controlled trial, 16 developers, 246 tasks; developers were 19% slower with AI tools while believing they had been faster.

[5] El-Yaniv, R., and Wiener, Y. (2010). On the foundations of noise-free selective classification. Journal of Machine Learning Research, 11, 1605 to 1641.

[6] Geifman, Y., and El-Yaniv, R. (2017). Selective classification for deep neural networks. Advances in Neural Information Processing Systems 30 (NeurIPS 2017).

[7] Anthropic (2026). Pricing. Claude Platform documentation, platform.claude.com/docs/en/about-claude/pricing, accessed 17 September 2026. Claude Sonnet 5 at $2 per million input tokens and $10 per million output tokens; five-minute cache writes $2.50 and cache reads $0.20 per million; Batch API at 50% of standard rates; volume discounts negotiated case by case. Announced at launch as introductory pricing through 31 August 2026 and since confirmed as the standard price; a scheduled increase to $3 and $15 per million input and output tokens on 1 September 2026 did not take effect.

[8] McKinsey & Company (2025). The state of AI: How organizations are rewiring to capture value. McKinsey Global Survey, published 12 March 2025; fielded 16 to 31 July 2024, 1,491 respondents in 101 nations. Reports that 27% of organizations using generative AI review all AI-generated content before use, while a similar share review 20% or less.

[9] Wei, J., Karina, N., Chung, H. W., et al. (2024). Measuring short-form factuality in large language models (SimpleQA). OpenAI. arXiv:2411.04368. Grades answers as correct, incorrect, or not attempted.

OPTIMIZED PAYMENTS

Enterprise payments AI is only as good as the data beneath it.

See what governed payments data can do for your AI.

Pick one payments question that matters to your business. We'll show you the data, definitions, and calculations behind the answer - and how Harmonize makes them reusable for AI. Request a demo to see governed metric tools on your own payments data.

ABOUT THIS STUDY

Authors

Valerie Chan, Technical Product Manager, Optimized Payments
Fei Hou, Chief Data Scientist, Optimized Payments

Contributors

Yuanda Zhu, AI/ML Data Scientist, Optimized Payments
Sandeep Suresh Date, Operations Research Analyst, Optimized Payments

Reviewed by

Anand Goel, CEO, Optimized Payments

Citation

Optimized Payments (2026). The Data Tax: what unready data costs enterprise AI, and how to stop paying it. Whitepaper, September 2026.

Disclaimer

Results reflect one merchant book, one model, and one two-month window under the conditions described in Appendix A and are not a guarantee of results in other environments. Modeled figures are illustrations on stated assumptions, not measurements. Acquirer and merchant identities are anonymized. Optimized Payments builds and sells the products measured at the higher-readiness levels of the study.

CONTACT

datatax@optimizedpayments.com
optimizedpayments.com