TokenWake

TIER ZERO SOLUTIONS · TOKENWAKE

AI Deployment Forecast

Credit Policy to Tested Business Rules

What this workflow costs to run over twelve months, which model configuration to run it on, and its diagnostic characteristics — derived from the structure of the work, against the model catalog on the run date below.

PREPARED FOR	R Star Bank · Mary Jones
WORKFLOW	Credit Policy to Tested Business Rules
RUN DATE	19 August 2026 · model catalog July 2026 · 64 independent runs
RUN IDENTITY	b47db0e3ccf16f65 · replay verified

CONTENTS

R Star Bank is not a real customer. This intake came through the tokenwake.io form as an end-to-end test. Every figure is output from a certified twelve-month run that replays bit-identical — the company is invented, the run is not.

START HERE

How to read this report

Five things worth knowing before the numbers.

What this is

A forecast of a workflow that has not run yet, derived from the structure of the work rather than from a usage history you do not have. Every figure comes from one certified engine run that replays bit-identical, and the model selection comes from a sweep in which every configuration was scored.

Check the picture before you read a number

The next page is your workflow as we modeled it. If an arrow is wrong, or a step is missing, everything after it is wrong too. That page is the one thing we need you to confirm.

Some workflows are expensive because of how they are built

Not because of what they cost per token — because of their shape. A step that loops on its own output, a gate that sends work back further than it needs to go, a spawned child with no ceiling on what it can start. Where the design itself is wasting money in yours, we name what is doing it and say what would close it. Often there is nothing to find, and that is a result too.

There is no single right configuration

What a wrong answer costs you decides which one wins, and that depends on your domain rather than on your workflow. Section 3 gives a ladder: every configuration that wins somewhere, and the point at which the answer changes. You do not have to name a number — you have to know which side of the line you are on.

What is yours, and what is ours

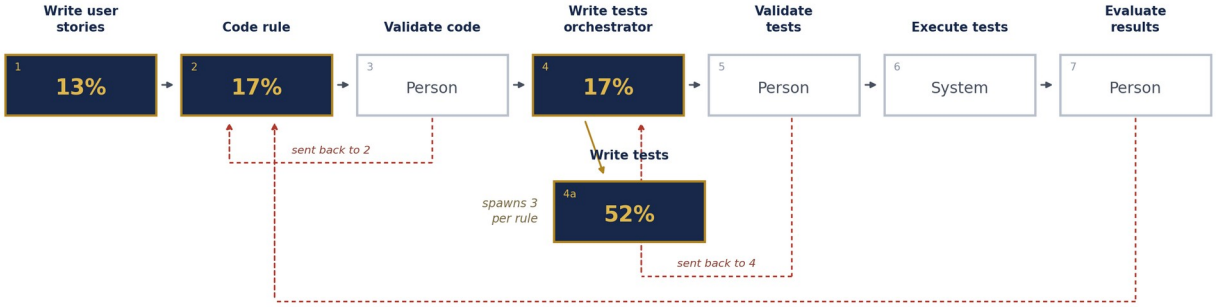
Section 5 lists every input with where it came from: your intake, an independent benchmark, a vendor's posted rate, or our own default. Confidence in a forecast is set by the weakest load-bearing input, not by the average, so those tags are worth reading.

BEFORE THE NUMBERS

The workflow we modeled

If anything here is wrong, stop and tell us.

20 seats × 2.5 sessions a workday = **50 rules started a day**



Behind the numbers in this report: 640,000 simulated runs of your workflow.

SECTION 1

Highlights

The four answers.

WHAT IT WILL COST

$23.98 a rule, all in

$298,100 a year — $99,460 in tokens and $198,640 in the reviewer hours rework created. The token bill is 33% of it.

WHAT REWORK COSTS IN REVIEW

One review in five is a repeat

A review is one person looking at one rule at one gate. The run measured 181 a business day; 38 of them are repeat visits by work that came back — 3,485 reviewer hours a year.

CONFIGURATION TO RUN

Claude Opus 5 (max)

On all four model-bearing steps. Credit decisioning sits in the top downstream impact band. Section 3 shows what the cheaper fleets save, and what they let through.

WHAT SHIPS WRONG

3 rules in 12,431

0.02% of delivered rules carry an undetected error — a failure the workflow did not catch.

SECTION 2

The forecast

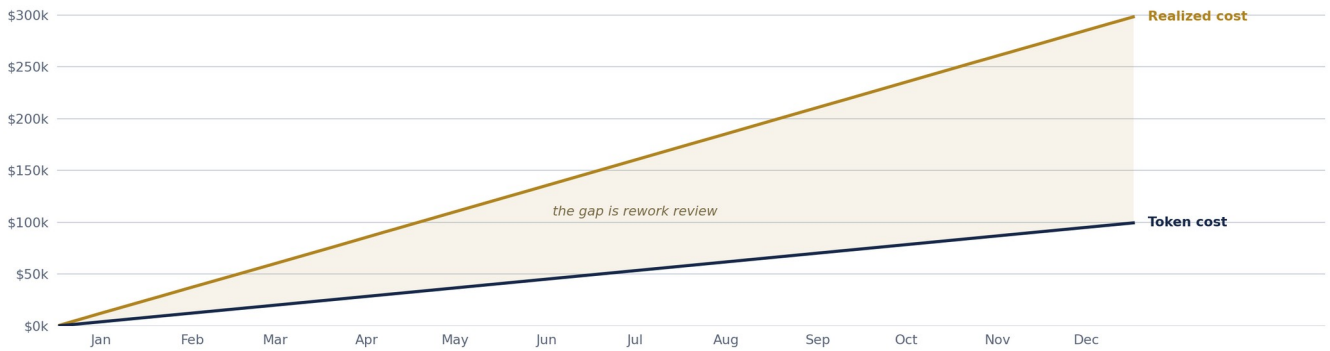
Twelve months of the configuration your downstream impact level selects — Claude Opus 5 (max) at every model-bearing step.

PER YEAR	VALUE
Work token cost	$85,506
Rework token cost	$13,954
Total token cost	**$99,460**
Human rework review cost	$198,640
Total realized cost	**$298,100**
Total token cost per finished rule	$8.00
Total realized cost per finished rule	**$23.98**
Rework share of token spend	14.0%
Rework share of realized cost (tokens + review)	71%
Time to a finished rule	1.61 hours

Realized cost is what you spend: first-pass tokens, the tokens spent redoing work that came back, and the reviewer hours that rework created. **Review is 67% of it — twice the token bill.** Designed review is excluded throughout; only the review rework caused is priced, at the professional tier, $57 an hour loaded. We assumed a review time. On an engagement it is replaced with yours, and this figure moves with it.

Time to a finished rule is elapsed time from the first model call to accepted results. It does not include time the rule sits in a queue for a human, so real elapsed time in an organization will be longer — by an amount that is a property of the organization rather than of the workflow.

Cumulative cost, twelve months



$298,100 ± 2% across 12,431 rules delivered. The lower line is what the vendor invoices. The upper line is what the workflow costs you.

What your actuals will impact

OURS, NOT YOURS	WHAT IT MOVES
First-pass quality at each gate	Drives the rework share — 14.0% of token spend in this workflow.
Time at each gate	Drives the reviewer hours, and so the largest line in the forecast.
Downstream impact level	Selects the configuration in section 3, and so the configuration section 2 forecasts.
Step size → input tokens	Scales the token bill. Changes no band; it applies to every configuration equally.
Share of rules treated as complex	Shifts failure and rework rates together.
Distribution of volume	Moves the daily figures. Annual totals do not change.

These are what the actuals calibration tier replaces — your telemetry instead of our assumption.

SECTION 3

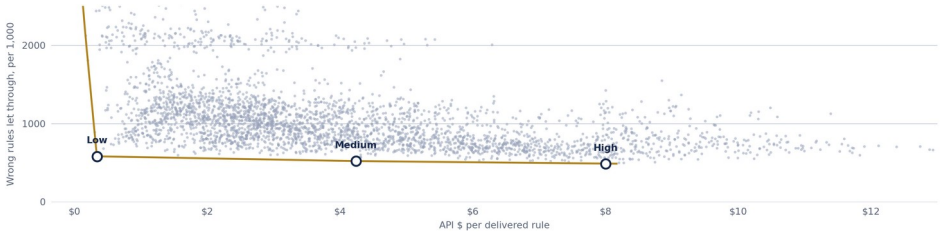
The configuration

Every configuration is ranked on true cost per delivered rule — API spend plus the modeled cost of the failures it lets through — not on API spend alone. Your downstream impact level decides which configuration wins. Cheap models spend less and let more wrong rules through, so the answer turns on your domain rather than on the workflow — regulatory exposure, safety, money, reputation: where a wrong rule lands decides how much it matters. Every configuration below wins somewhere; nothing else wins at any price.

DOWNSTREAM IMPACT	RUN THIS	TOKENS AGAINST THE FORECAST
Low impact	Grok 4.5 (high) on all four steps	−88%
Medium impact	Grok 4.5 (high), then Opus 5 (max) on the last three steps	−12%
High impact	Opus 5 (max) on all four steps	—

Every configuration that wins somewhere is listed. The points where one overtakes another are computed, not chosen, and your downstream impact level places you between two of them. The highest level lets fewest wrong rules through, whatever it costs; the lowest is the best value before quality falls away. A workflow can have more levels than this one does. Token cost is measured against the configuration section 2 forecasts.

What each step up the ladder costs



The three ringed points are the three configurations in the table above. Every dot is one configuration — cost per delivered rule against the wrong rules it lets through. The gold line is the efficient frontier: nothing beats these on both counts at once. Every configuration was scored; this chart plots a sample of them.

SECTION 4

Diagnostics

Where the money sits, and where the work stops.

How a token cap affects your work

FORECAST IMPACT	CAP PER RULE	WORK LOST	WHERE THE WORK STOPS	API SPEND	RULES DELIVERED
none	—	—	—	$99,460	12,431
5%	$10.48	17%	Write tests (4a)	$94,487	10,294
10%	$8.05	26%	Write tests (4a)	$89,510	9,171
15%	$6.91	40%	Write tests (4a)	$84,538	7,508
20%	$5.38	95%	Write tests (4a)	$79,566	654

A cap is a kill, not a throttle: a rule stopped mid-flight has already spent what it spent. Every cap tested halts the work in the same place, because that is where the money is.

SECTION 5

Workflow structure and inputs

What you told us, and what we supplied.

What you told us about your workflow

STEP	WHAT IT DOES	PERFORMED BY	RETRIES	AFTER THE RETRIES
1	Write user stories	AI	2	moves on anyway
2	Code rule	AI	2	moves on anyway
3	Validate code	Person	2	back to step 2
4	Write tests orchestrator	AI	2	moves on anyway · spawns 3 scripts per rule
4a	Write tests	AI	2	the rule ships with fewer tests
5	Validate tests	Person	2	back to step 4
6	Execute tests	System	2	moves on anyway
7	Evaluate results	Person	2	back to step 2

Throughout this report a step that fails has produced a wrong answer, not stopped. A retry is the step re-doing its own work because the output was wrong. A review is one person looking at one rule at one gate. Fail and retry do not refer to a timeout, an API error or an outage — those are not forecast.

Each box on the flow shows that step's share of the annual API bill, or who performs it where no model is called. Step 4a is ours, not yours: you told us step 4 spawns three scripts, and we model each script as its own step so you can see what it costs.

Inputs, and where they came from

DECLARED you told us. **ESTIMATED** a considered estimate. **ASSUMED** a Tier Zero default. **BENCHMARK** measured by an independent evaluation, not by us and not by the vendor. **PUBLISHED** the vendor's own posted rate.

INPUT	VALUE	SOURCE	NOTE
Seats and volume	20 × 2.5 a workday	DECLARED	Person-triggered
Horizon	12 months	DECLARED	Given instead of a spending ceiling; the form accepts either
Maximum loop re-entries	5	DECLARED	Read at the top of the band you gave
Test scripts spawned per rule	3	DECLARED	A fixed number, as entered
Model eligibility	No Chinese-jurisdiction providers	DECLARED	A procurement rule; four of fourteen models excluded
Rejection rate at each gate	10% routine, 25% hard	ESTIMATED	Applied at all three gates
Downstream impact level	high	DECLARED	Selects the configuration in section 3. No dollar figure is asked for or used
Reviewer tier	professional, $57/h loaded	ASSUMED	Not declared. Drives the largest line in the forecast
Time at each gate	20 / 15 / 20 min	ASSUMED	Ours until you supply your own
Step size → input tokens	15k a call	ASSUMED	Tier Zero size bands. Output tokens come from the model catalog, per model
Share of rules treated as complex	25%	ASSUMED	Tier Zero default
Distribution of volume	even across the calendar	ASSUMED	Annual totals are right; daily peaks are smoothed
Accuracy, hallucination, tokens per task, latency	per model	BENCHMARK	AA-Omniscience v4.1 and Artificial Analysis, catalog July 2026
Model pricing	14 models	PUBLISHED	The vendors' posted rates

Every one of these inputs varies in real life, and the engine does not hold them fixed. It runs your workflow 64 times over, each run with a different draw of how the luck falls — which rules come in hard, which steps fail, how many tokens a call uses, how long a review takes. Every configuration we tested faced the same 64 draws, so none of them won on luck. Choosing your configuration meant doing that for all 10,000 eligible combinations of models across your steps: 640,000 runs of your workflow.

APPENDIX

Model reference metrics

Every model in the catalog, and what it is measured at.

MODEL	PROVIDER	IN $/MTOK	OUT $/MTOK	ACCURACY	HALLUC.	OUT TOK/TASK	MIN/TASK
Claude Fable 5	Anthropic	10.00	50.00	61%	55%	33,000	12.7
GPT-5.6 Sol (max)	OpenAI	5.00	30.00	59%	89%	15,000	5.9
Gemini 3.1 Pro Preview	Google	2.00	12.00	55%	50%	13,000	1.3
Claude Opus 5 (max)	Anthropic	5.00	25.00	54%	50%	36,978	6.6
Grok 4.5 (high)	xAI	2.00	6.00	52%	54%	14,000	3.2
Gemini 3.5 Flash	Google	1.50	9.00	52%	61%	28,000	5.8
Claude Opus 4.8	Anthropic	5.00	25.00	47%	36%	40,000	17.7
Kimi K3 — excluded	Kimi	3.00	15.00	46%	51%	23,397	6.0
GPT-5.6 Luna (max)	OpenAI	1.00	6.00	42%	90%	19,000	1.8
Claude Sonnet 5	Anthropic	3.00	15.00	38%	37%	69,000	15.4
DeepSeek V4 Flash (max) — excluded	DeepSeek	0.08	0.17	37%	96%	45,000	3.3
Kimi K2.6 — excluded	Moonshot	0.66	3.41	33%	39%	38,000	7.8
Claude Haiku 4.5	Anthropic	1.00	5.00	17%	26%	24,000	2.3
MiniMax-M3 — excluded	MiniMax	0.30	1.20	15%	16%	24,000	4.1

Four models are marked excluded: DeepSeek V4 Flash, Kimi K2.6, Kimi K3 and MiniMax-M3, all Chinese-jurisdiction providers, ruled out at your instruction. That is a procurement rule and not a quality judgment — their measured figures are shown so the exclusion can be seen rather than taken on trust. Ten models remained eligible and every configuration of them was scored.

Benchmarks from AA-Omniscience v4.1 and Artificial Analysis, captured 24 July 2026. Pricing is the vendors' posted rates. Hallucination is not a quality score. It is a failure split — the share of a model's failures that pass as correct instead of coming back as visible rework.

A forecast is a projection, not a guarantee. Actual costs and results will differ. See Terms of Service §2.2.