01 · EXECUTIVE SUMMARY

Halden AI

Document intelligence for insurance claims · Series B · Prepared for a sample investor

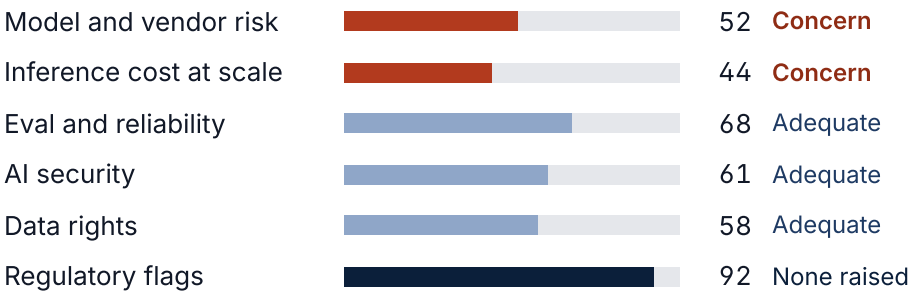
Assess engagement · 6 business days of fieldwork · Read-only access to repository, AWS account and staging

AI ASSET RATING

B 63 / 100

Real, production AI with a margin problem at scale. Supports the product thesis; discounts the unit economics until fixed.

MODULE SCORES (0–100)



Bands: Concern under 55 · Adequate 55–74 · Strong 75+ · Grade B = 60–69

Core findings

#	FINDING	EVIDENCE	CLASS	IMPACT
F1	**Gross margin falls from 71% today to 46% at 10x usage**	Provider invoices, cost model, load tests	Discount	−25 pts GM at 10x
F2	**81% of inference spend sits with one frontier-model provider**	Billing export, architecture trace	Discount	Margin set by vendor pricing
F3	**Extraction accuracy measures 91.4%, against 97% claimed in the deck**	1,200-document held-out eval	Discount	Claim overstated by 5.6 pts
F4	**Throughput breaks at about 6x current load on a single-node vector store**	Staged load test to failure	Fix	$90–140k to remove
F5	**9 of 60 prompt-injection payloads succeeded; 3 live keys in repository history**	Injection suite, history scan	Fix	$40–50k, before close

Implications for the deal

RAISE IN NEGOTIATION

Model gross margin at 46% at 10x usage, not the 70%+ in the plan. Treat the stated 97% accuracy as 91%.

SET AS CLOSING CONDITIONS

Rotate the exposed keys and patch the injection paths. Counsel to review the 22% of training data scraped from the web (p. 4).

FIX AFTER CLOSE

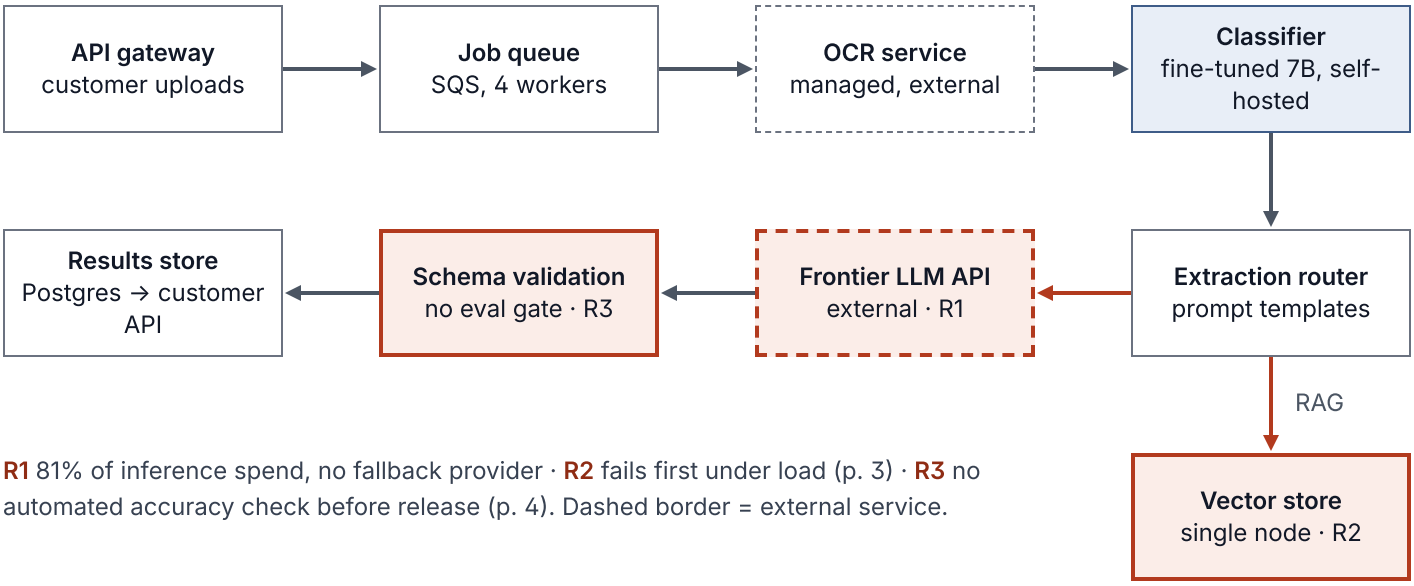
A 100-day plan takes gross margin at 10x back to about 66% and removes the 6x ceiling, for $380–540k (p. 5).

02 · ARCHITECTURE AND MODEL DEPENDENCY · SCORE 52

The "proprietary model" is a fine-tuned open-weights classifier and prompts on a third-party API.

The defensible asset is the 4.2M labelled claim documents, not the model. Extraction, which carries the product's value, runs entirely through one external frontier-model API with no fallback path.

EXHIBIT 2.1 · REQUEST PATH, AS TRACED IN THE REPOSITORY AND AWS ACCOUNT



R1 81% of inference spend, no fallback provider · **R2** fails first under load (p. 3) · **R3** no automated accuracy check before release (p. 4). Dashed border = external service.

EXHIBIT 2.2 · INFERENCE SPEND BY MODEL, LAST 90 DAYS



EXHIBIT 2.3 · THIRD-PARTY MODEL AND SERVICE DEPENDENCIES

Dependency	Role	Switching estimate	Risk
Frontier LLM API	Field extraction	6–8 weeks to re-prompt and re-evaluate; no abstraction layer today	High
Open-weights 7B base	Classification	Weights held in-house; licence permits commercial use	Low
Managed OCR service	Text extraction	Two drop-in alternatives tested within 2% on accuracy	Low

IMPLICATION

Discount: one vendor's pricing sets the extraction margin (F2). **Fix:** an abstraction layer and a qualified second provider, $40–60k (p. 5). **Support:** the labelled dataset is a real moat, if its rights hold (p. 4).

03 · INFERENCE COST AT SCALE · SCORE 44

Gross margin falls from 71% to 46% at 10x usage, and throughput breaks at about 6x.

EXHIBIT 3.1 · GROSS MARGIN BY USAGE MULTIPLE (LOG SCALE)

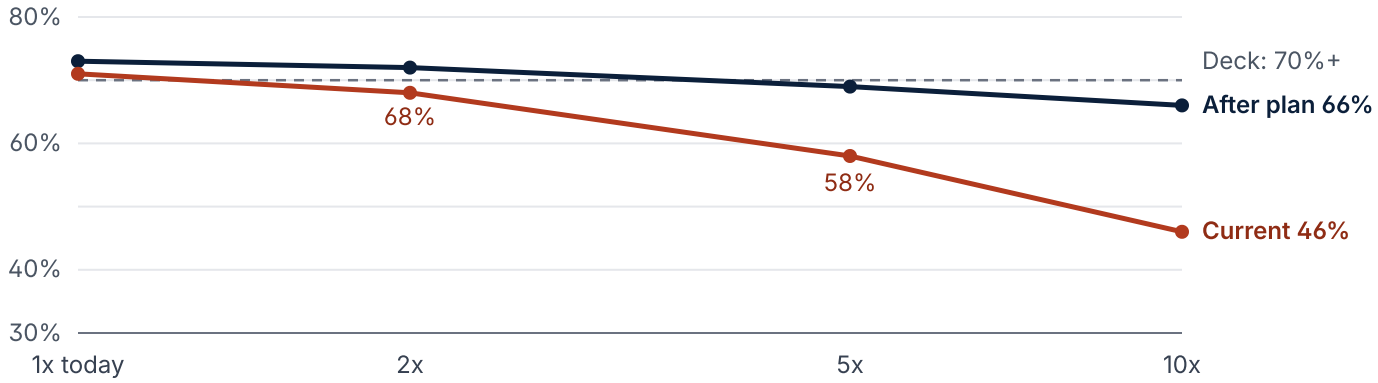


EXHIBIT 3.2 · P95 LATENCY UNDER LOAD

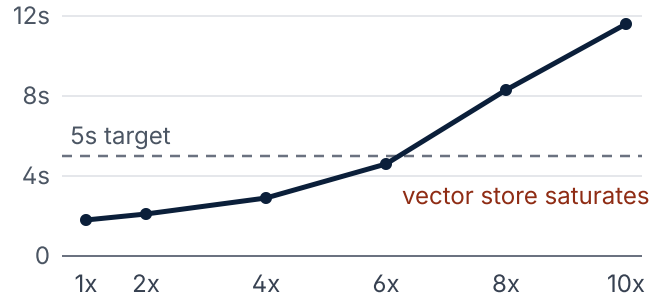


EXHIBIT 3.3 · ERROR RATE UNDER LOAD

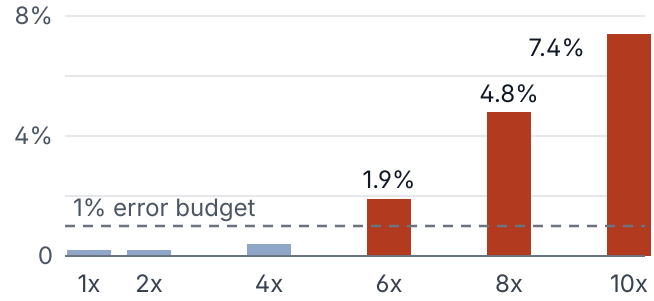
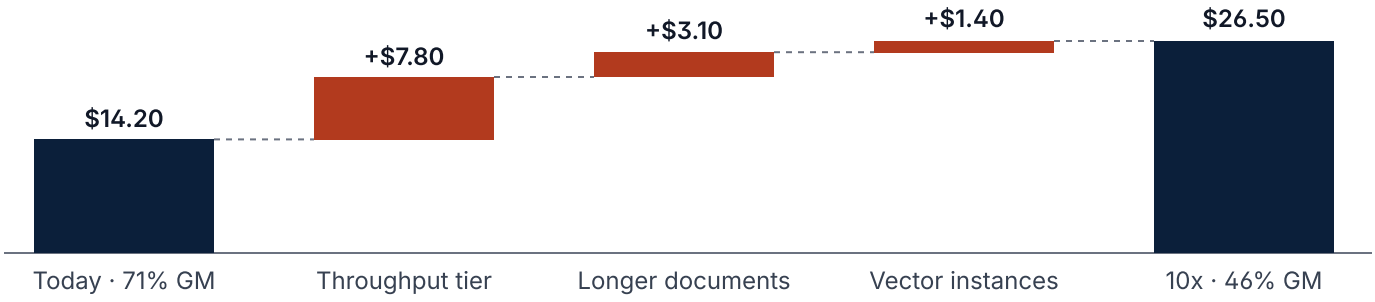


EXHIBIT 3.4 · COST PER 1,000 DOCUMENTS, TODAY TO 10X (REVENUE: $49 PER 1,000)



IMPLICATION

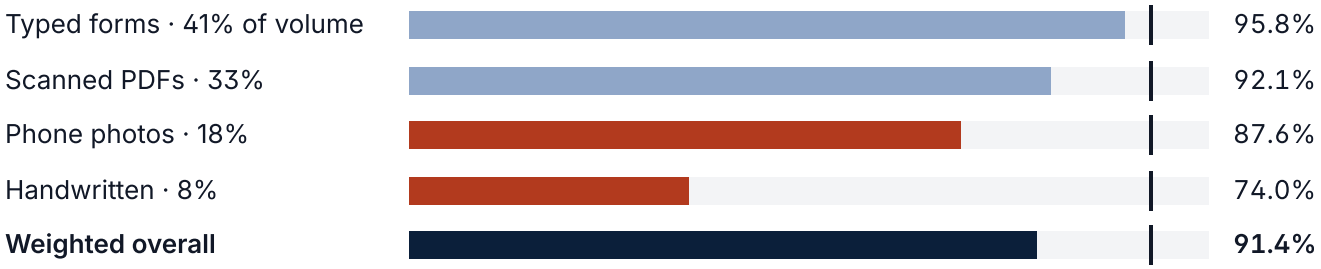
Discount: the plan's 70%+ gross margin does not survive scale on today's architecture; model 46% at 10x. **Fix:** routing, caching and a second provider recover about 20 points (p. 5). The 6x ceiling must be removed before the growth plan's year-two volumes.

Method: cost model from 90 days of invoices; load stepped to failure on staging. Figures illustrative.

04 · RELIABILITY 68 · SECURITY 61 · DATA RIGHTS 58 · REGULATORY 92

Accuracy is 91.4%, not 97%. Two security findings should close before signing.

EXHIBIT 4.1 · EXTRACTION ACCURACY, 1,200 HELD-OUT DOCUMENTS (AXIS 60–100%)



Black tick = 97% claimed in the deck. 2.8% of extracted fields had no supporting text in the source (hallucinated). Release 3.5 cut photo accuracy by 3.1 points and shipped unnoticed: there is no eval gate (R3).

EXHIBIT 4.2 · AI SECURITY TESTS, RUN ON STAGING UNDER THE TARGET'S WRITTEN AUTHORIZATION

Test	Result	Severity	Remediation
Prompt injection, 60 payloads in uploaded documents	9 succeeded; 4 rewrote extracted payout fields	High	Input isolation, output allow-list
Secrets in repository history	3 live keys: cloud, model provider, database replica	High	Rotate now; add secret scanning
Dependencies with high CVEs	4 packages, none reachable from the public API	Medium	Upgrade in first sprint
Model and package supply chain	Weights downloaded unpinned at build time	Medium	Pin versions and verify checksums
Cross-tenant access, 40 probes	0 succeeded; row-level isolation holds	Pass	None needed

EXHIBIT 4.3 · TRAINING DATA BY SOURCE, 4.2M DOCUMENTS



Customer terms permit training on de-identified data (clause 7.2). The web-scraped share has no recorded licence. An ablation run without it costs 1.2 points of accuracy, so removal is affordable if counsel requires it. Technical evidence for counsel, not a legal opinion.

REGULATORY · 92 — Document processing for insurers. No sensitive-technology indicators under Investment Canada Act or CFIUS lists; no filing trigger identified for counsel.

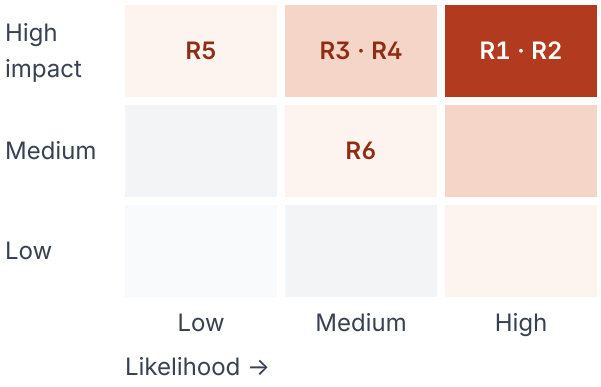
IMPLICATION

Discount: treat the accuracy claim as 91%, and weaker on photos and handwriting, which the growth plan targets. **Condition:** rotate keys and close the injection path before signing. **Fix:** an eval gate on every release, $50–70k (p. 5).

05 · RISK REGISTER AND VALUE-CREATION PLAN

A 100-day plan recovers 20 points of margin at 10x and removes the scale ceiling, for $380–540k.

EXHIBIT 5.1 · RISK REGISTER: LIKELIHOOD OVER THE HOLD PERIOD × IMPACT ON THE THESIS



R1 Provider raises prices or changes terms on 81% of spend

R2 Throughput ceiling at ~6x stalls year-two growth

R3 Silent accuracy regressions reach customers

R4 Injected documents alter claim payout fields

R5 Web-scraped training data must be removed

R6 Exposed keys used before rotation

Every risk is re-scored in each quarterly Monitor run.

EXHIBIT 5.2 · VALUE-CREATION OPPORTUNITIES, SIZED

#	Opportunity	Addresses	Cost	Effect at 10x	Days
1	Route about 60% of documents to the fine-tuned 7B model	F1, F2	$120–160k	**+9 pts gross margin**	31–60
2	Semantic cache for repeat claim templates	F1	$40–60k	**+6 pts gross margin**	31–60
3	Provider abstraction and a qualified second provider	F2, R1	$40–60k	**+5 pts; one vendor under 50%**	61–100
4	Shard the vector store; autoscale extraction workers	F4, R2	$90–140k	**Ceiling removed; p95 under 5s**	61–100
5	Eval harness that gates every release	F3, R3	$50–70k	**Accuracy measured per release**	0–30
6	Injection defences, key rotation, pinned supply chain	F5, R4, R6	$40–50k	**Both High findings closed**	0–30
	Total		**$380–540k**	**Gross margin 46% → 66%**	**100**

EXHIBIT 5.3 · 100-DAY SEQUENCE, DELIVERED WITH HALDEN'S ENGINEERING TEAM

Days 0–30 · Secure and measure	Days 31–60 · Cut cost	Days 61–100 · Scale and de-risk	Quarter 2
6 · Close High findings 5 · Eval gate in CI	1 · Model routing 2 · Semantic cache	4 · Shard and autoscale 3 · Second provider live	Monitor re-run proves gains

Scope and limits. Read-only repository, AWS account and staging; no production testing. Load tests replay the production traffic mix on staging. Cost ranges are engineering estimates, ±25%. Data-rights findings are technical evidence for counsel. Fadak provides technical assessment and guidance, not investment, legal or formal valuation advice.