scruple

pilot offer, card disputes
draft, November 2026

Verifiers and a release gate for software that decides card disputes.

Dispute logic changes often. Outcomes arrive weeks later. An average hides the reason code that got worse, and agents report their own wins. Scruple scores each change from ledger postings and policy constraints, never from what the agent says about itself, and blocks the release when it regresses.

WHAT EXISTS TODAY

Tested, on published fixtures and synthetic histories.

- Replays a change on held-out cases and on your imported history, and checks **every reason code separately**.
- Scores cost from **balance movements**; checks deadlines, evidence and the agent's own report against the ledger.
- **Blocks the pull request** in GitHub or GitLab, with a scorecard comment and a status check.
- A named approver who is not the rule's owner; a **hash-chained audit trail**; an evidence pack per decision.
- Stripe, Adyen and CSV ingest with redaction, card tokenisation and field encryption before storage.

None of it has yet run on a partner's data. That is what the pilot is for.

REAL MODELS, MEASURED

They know the rules. They misjudge the cases.

```
CFPB complaints, 100 held out, graded by code
Sonnet named the right regulation 93% of the time,
Haiku 90%, a TF-IDF classifier 87%.
Predicting relief, Sonnet scored AUROC 0.61, below
the classifier's 0.70.

12 simulated disputes each, every call the model's
own
Sonnet cut the merchant's loss most, -$32.77 a case
against an evidence policy [95% -64.01 to -4.33].
4 of its 10 wins were on claims the cardholder was
right about.

open models through the API, same 12 cases
gpt-oss-120b claimed a net cost of 0 on 5 cases that
lost money.
gpt-oss-20b passed the gate on 0 of 12: it
represented on documents it never read.
```

engine/docs/results/REAL_MODELS.md, MODEL_EPISODES.md. Small samples, wide intervals. Cases are synthetic; the model behaviour is real.

THE PILOT

Six weeks, one workflow, one change. Criteria signed before any data.

WEEK 1	WEEK 2	WEEK 3	WEEK 4	WEEK 5	WEEK 6
Kickoff	**Export**	**Backtest**	**Shadow**	**Gate**	**Readout**
Success criteria signed. Export agreed.	Ingest report: rows read, rejected and why.	The candidate on your history, per reason code.	Live traffic. Scruple acts on nothing.	In your CI, scorecards on real pull requests.	Pass, fail or not enough data. Your data deleted.

YOU PROVIDE

- A few thousand resolved disputes, read-only: Stripe API, Adyen or CSV.
- One change: a new rule, a prompt edit or a model swap.
- An owner, an engineer, a dispute expert, a reviewer: about 45 hours in all.

YOU GET

- A backtest, a shadow summary and a gate on one repository.
- A self-report check: claimed wins against the balance transactions.
- Evidence packs and a readout, reported whether it passes or fails.

Security. Runs inside your environment, or on a scruple machine with tokenised, redacted data only. Card numbers are tokenised and chosen fields encrypted before anything is stored. No SOC 2 yet; the security answers say what is not built.
Pricing is set per pilot.

[dhegde@umich.edu]

Dhruv Hegde, founder
thedhruvhegde.github.io/scruple/pilot/