

SIMULATED CASE STUDY

Evaluation Report

AI-assisted audit evidence review prototype

Project	AI-assisted audit evidence review (Project 5)
Company	Northstar Financial Technologies (fictional)
Model	Deterministic mock (no language model)
Run date	2026-10-16

This document is part of a portfolio exercise. Northstar Financial Technologies, its people, systems, records and findings are invented. No real audit was performed, no assurance is given and nothing here is a certification or compliance claim.

1. Summary

A prototype reviews 108 synthetic evidence cases in three ways: deterministic rules on the structured record, a mock model on the snippet text, and the two combined. On this corpus the combined pipeline detected 40 of 43 seeded exceptions (recall 0.93) against 30 for rules alone (recall 0.70), with similar precision (0.95 and 0.94). It abstained on all 18 genuinely ambiguous or contradictory cases. The mock model alone scored as well on exceptions (F1 0.95 against 0.94).

What this does not show: the model is a deterministic mock, not a language model; the corpus and the mock were written by the same person from related templates; and the labels are the author's judgement. The numbers demonstrate that the evaluation harness works. They are not evidence that AI improves audit quality, and they say nothing about any real model.

2. What was run

Run date (UTC)	2026-10-16
Model	mock-heuristic 1.0 (mode mock): deterministic regular-expression extraction and keyword classification; no language model was called
Prompt version	evidence-review-v1.0 (describes the interface a live model would receive; the mock ignores it)
Rules version	rules-1.0
Temperature	not applicable (deterministic mock)
External calls made	0
Seeds	development 20261101, evaluation 20261102
Python	3.12.13

3. Evaluation design

Two corpora were generated from different seeds. A development set of 44 cases was used while writing the mock. The held-out evaluation set of 108 cases was run once for this report; the mock was frozen first and only a rule bug (malformed dates crashed instead of abstaining) and a metric definition (field accuracy excludes cases where the true value deliberately differs from the snippet) were fixed afterwards, both recorded in the change log.

Ground-truth label	Cases
Exception	43
No exception	35
Needs review	18
Not relevant	12

Cases cover standard exceptions and passes, boundary near-misses (completed on the due date, disabled on business day 2, emergency approval on day 2, evidence period exactly equal), records where the text reveals what the structured data hides (failed tests, re-enabled account, excluded accounts), stale structured data and mislabelled change type (rule false positives), ambiguous numeric dates (4/5 could be 4 May or 5 April), contradictory sources, and unrelated evidence.

Definitions and denominators

- Positive class: Exception. Decidable cases are those whose ground truth is Exception or No exception; precision, recall, F1, false-positive and false-negative rates use only these.
- A prediction of Needs review or Not relevant is an abstention, not a positive. Recall counts abstentions on true exceptions as not detected (recall excluding abstentions is also shown).

- Precision = TP / predicted Exceptions. False-positive rate = FP / ground-truth No exception cases. False-negative rate = 1 - recall.

- Abstention rate = share of all cases predicted Needs review. Correct-abstention rate = Needs review predicted on ground-truth Needs review cases. Intervals are 95% Wilson intervals.

- Classification threshold: none. The mock returns a label directly. Its confidence value is an uncalibrated heuristic (share of expected fields found) and is not a probability; no calibration was performed.

- Reviewer acceptance and correction rates: not measured. No human reviewed these cases for this report. The reviewer queue records them when a person uses it.

4. Results

Variant	Precision	Recall	F1	False-positive rate	False-negative rate	Abstention (all cases)	Overall agreement
Deterministic rules only	0.94 (30/32; 95% CI 0.80-0.98)	0.70 (30/43; 95% CI 0.55-0.81)	0.80	0.06 (2/35; 95% CI 0.02-0.19)	0.30 (13/43; 95% CI 0.19-0.45)	0.33 (36/108; 95% CI 0.25-0.43)	0.58 (63/108; 95% CI 0.49-0.67)
Mock model alone (text only)	0.95 (41/43; 95% CI 0.84-0.99)	0.95 (41/43; 95% CI 0.84-0.99)	0.95	0.06 (2/35; 95% CI 0.02-0.19)	0.05 (2/43; 95% CI 0.01-0.15)	0.08 (9/108; 95% CI 0.04-0.15)	0.88 (95/108; 95% CI 0.81-0.93)
Rules + mock model (assisted)	0.95 (40/42; 95% CI 0.84-0.99)	0.93 (40/43; 95% CI 0.81-0.98)	0.94	0.06 (2/35; 95% CI 0.02-0.19)	0.07 (3/43; 95% CI 0.02-0.19)	0.19 (21/108; 95% CI 0.13-0.28)	0.94 (101/108; 95% CI 0.87-0.97)
Variant	Correct abstention (ambiguous or contradictory)		Irrelevant evidence identified		Overconfident on ambiguous or contradictory		
Rules + mock model (assisted)	1.00 (18/18; 95% CI 0.82-1.00)		1.00 (12/12; 95% CI 0.76-1.00)		0.00 (0/18; 95% CI 0.00-0.18)		
Mock model alone (text only)	0.50 (9/18; 95% CI 0.29-0.71)		1.00 (12/12; 95% CI 0.76-1.00)		0.50 (9/18; 95% CI 0.29-0.71)		
Deterministic rules only	0.50 (9/18; 95% CI 0.29-0.71)		0.00 (0/12; 95% CI 0.00-0.24)		0.50 (9/18; 95% CI 0.29-0.71)		

Field extraction accuracy of the mock: 1.00 (379/379; 95% CI 0.99-1.00), measured on cases where the true values equal the snippet's statements. A perfect score on short fixed templates is expected and should not be read as a prediction for real documents.

Confusion matrix: Deterministic rules only (rows ground truth, columns prediction)

Ground truth	Exception	No exception	Needs review	Not relevant
Exception	30	7	6	0
No exception	2	24	9	0
Needs review	9	0	9	0
Not relevant	0	0	12	0

Confusion matrix: Mock model alone (text only) (rows ground truth, columns prediction)

Ground truth	Exception	No exception	Needs review	Not relevant
Exception	41	2	0	0
No exception	2	33	0	0
Needs review	0	9	9	0
Not relevant	0	0	0	12

Confusion matrix: Rules + mock model (assisted) (rows ground truth, columns prediction)

Ground truth	Exception	No exception	Needs review	Not relevant
Exception	40	2	1	0
No exception	2	31	2	0
Needs review	0	0	18	0
Not relevant	0	0	0	12

5. Deterministic only versus AI-assisted

Rules alone abstained on 36 of 108 cases because structured fields were missing, missed 7 exceptions that the text revealed, and could not tell that 12 snippets were unrelated evidence. The assisted pipeline fixed most of these. Recall rose from 0.70 to 0.93, but the intervals (0.55-0.81 and 0.81-0.98) overlap at this sample size. The gain exists partly because the corpus was built with text-only information that structured rules cannot see, so it is partly by design.

Adding the structured record to the mock did not improve exception detection (F1 0.95 alone, 0.94 assisted); it helped on contradictions, where the mock alone cannot know the structured record disagrees (correct abstention 0.50 alone, 1.00 assisted), at the cost of unnecessary abstentions (3 cases). Overall agreement with ground truth is highest for the assisted pipeline.

6. Error analysis

Missed exception (2 cases)

Case	Type	Snippet	Truth	Prediction	Reason given
EVR-070	text issue	...tle. Approved by Devon Larkfield on June 8, 2026. Implemented 2026-06-12. Test result: Pass (TST-3114). Comment: the test evidence TST link belongs to a different change.	Exception	No exception	rules: no rule fired
EVR-085	text issue	... in scope 70; accounts reviewed 70. Date completed: April 10, 2026. Deadline: April 30, 2026. Accounts created in the last quarter were not included in the list provided.	Exception	No exception	rules: no rule fired

False alarm (stale or wrong structured data) (2 cases)

Case	Type	Snippet	Truth	Prediction	Reason given
EVR-011	stale fp	... 24/07/2026 (due 30/07/2026). Note: Morgan Stonebrook's Admin role on SYS-05 was revoked on 2026-07-10, before the review; the entitlement export predates the revocation.	No exception	Exception	rules: Reviewer holds Admin access on the reviewed system
EVR-044	stale fp	...brook (IT Service Desk Manager). Note: Sam Stonebrook's Admin role on SYS-01 was revoked on 2026-04-10, before the review; the entitlement export predates the revocation.	No exception	Exception	rules: Reviewer holds Admin access on the reviewed system

Unnecessary abstention (3 cases)

Case	Type	Snippet	Truth	Prediction	Reason given
EVR-010	text issue	... Source Control and CI/CD account disabled on 12 May 2026. Update: the account was re-enabled on request of the manager three days later and was still enabled at extract.	Exception	Needs review	model flagged: Snippet and structured record disagree on account_status.
EVR-052	mislabel fp	... approval given 16 Apr 2026. Developer: Devon Winterbourne. (The ticket was filed as type Normal in the tool in error; it is an emergency change per the incident record.)	No exception	Needs review	model flagged: Snippet and structured record disagree on change_type.
EVR-066	mislabel fp	...pective approval recorded 4 May 2026 by Morgan Stonebrook. (The ticket was filed as type Normal in the tool in error; it is an emergency change per the incident record.)	No exception	Needs review	model flagged: Snippet and structured record disagree on change_type.

Correct abstention: ambiguous wording (9 cases)

Case	Type	Snippet	Truth	Prediction	Reason given
EVR-004	ambiguous date	...1), 2026-Q2. I have reviewed 66 of 66 accounts and confirm access is appropriate. Completed 04/08/2026. Due date 30 Jul 2026. Signed Riley Winterbourne (Head of Product).	Needs review	Needs review	model flagged: A numeric date could be read as dd/mm or mm/dd; the outcome depends on the reading.
EVR-015	ambiguous date	Evidence note: change export. Stated period 01/01/2026 to 30/06/2026. File content spans 03/01/2026 to 06/07/2026.	Needs review	Needs review	model flagged: A numeric date could be read as dd/mm or mm/dd; the outcome depends on the reading.
EVR-022	ambiguous date	Change for Core Ledger. Requested by Devon Hallam, approved by Morgan Stonebrook on 2026-05-01. Implemented 04/05/2026. Tests: passed (TST-6075).	Needs review	Needs review	model flagged: A numeric date could be read as dd/mm or mm/dd; the outcome depends on the reading.

Correct abstention: contradictory sources (9 cases)

Case	Type	Snippet	Truth	Prediction	Reason given
EVR-007	contradictory	Change request for Core Ledger (SYS-03). Requested by Mica Stonebrook. Approved by Morgan Redmond on 30/04/2026. Implemented 1 May 2026. Test result: Pass (TST-1561).	Needs review	Needs review	model flagged: Snippet and structured record disagree on approved_by.
EVR-019	contradictory	IT ticket: leaver E0137, termination date April 27, 2026. AWS Production Account account disabled on 27 Apr 2026.	Needs review	Needs review	model flagged: Snippet and structured record disagree on account_status.
EVR-038	contradictory	Collection record - identity extract. Coverage claimed: 1 Jan 2026 through 30 Jun 2026. File content spans 1 Jan 2026 to 2026-06-27.	Needs review	Needs review	model flagged: Snippet and structured record disagree on data_max.

Irrelevant evidence correctly identified (12 cases)

Case	Type	Snippet	Truth	Prediction	Reason given
EVR-001	irrelevant	Training schedule: annual first aid refresher sessions held on the third floor, 14 attendees.	Not relevant	Not relevant	model: No terms in the snippet relate to the control topic.
EVR-013	irrelevant	Procurement note: renewal of office coffee machine lease approved for 12 months at the existing rate.	Not relevant	Not relevant	model: No terms in the snippet relate to the control topic.
EVR-026	irrelevant	Facilities report: quarterly fire drill completed on all floors, evacuation time 4 minutes 10 seconds, no issues noted.	Not relevant	Not relevant	model: No terms in the snippet relate to the control topic.

Wrong extraction

0 extraction errors were found on the scored cases. The mock reads only a fixed set of phrasings. A numeric date that could be day/month or month/day is deliberately not guessed: it triggers abstention. Relative dates (for example 'two days after approval') and unusual phrasings are not read at all, but the corpus contains none, so this blind spot is untested here rather than shown to be harmless.

Findings from the error analysis: (1) the two missed exceptions are text cues the mock does not know ('belongs to a different change', 'not included in the list provided'), which shows that keyword cue lists miss paraphrases; (2) the false alarms are stale structured data (an Admin role revoked before the review) that the mock reads literally; (3) the unnecessary abstentions come from the structured record and the text disagreeing: two benign (a change filed with the wrong type) and one real exception (an account re-enabled after disablement) where the pipeline abstained instead of flagging it; (4) every ambiguous-date case was correctly sent to a human.

7. Limitations

- The model is a deterministic mock. Nothing here measures any real language model, and no live model was called. The live adapter exists but has only been tested against a stubbed transport.
- Corpus and mock share an author and related templates, so performance is optimistic and in-distribution. Real evidence is messier: scans, tables, free text, other languages.
- Ground truth is the author's judgement fixed by scenario design; several labels are policy interpretations, not facts.
- Small sample: 43 exceptions and 35 passes. Intervals are wide and several differences are within them.
- The irrelevant-evidence cases are easy (clearly unrelated topics). Subtly irrelevant evidence was not tested.
- No human reviewers took part, so reviewer acceptance, correction and time-saved rates are unmeasured. No claim is made about efficiency gains.
- Reproducibility: fixed seeds and a deterministic mock make results exactly repeatable here. A live model would add non-determinism, version drift and provider change; results would need re-running and the date and model version recording.

8. Conclusions

Supported: the pipeline runs end to end without a paid API; outputs validate against a strict schema; rules, abstention and the audit log behave as designed; on this synthetic corpus, adding text understanding to rules reduces missed exceptions and handles unrelated evidence. Not supported: that AI improves audit performance in general, that the mock's scores predict a real model's, or that any suggestion can replace a reviewer's judgement. Final test status and findings always require human confirmation.

9. Reproduce

From the repository root: `docker build -t assurance-build . ; docker run --rm -v "$PWD":/work -w /work -e PYTHONPATH=/work/src:/work/scripts assurance-build python -m evidence_review.cli generate ; ... python -m evidence_review.cli run`. Evaluation file hash 2edc160d9f9f3a2a..., labels hash fa53edf9b8d38ba9..., prompt hash afd7ac1b3f21585a....