

SAMPLE DELIVERABLE / EVALUATION HEALTH CHECK

A relevant answer can still be wrong.

Insurance question answering | Evidence selection and abstention

An offline comparison on public NFIP guidance. The revised retriever passes more checks, but still misses a policy exception. Further review is needed before relying on these response decisions.

7 / 14

BASELINE PASSES

50.0% of assessment checks

10 / 14

CANDIDATE PASSES

71.4% of assessment checks

4

CHECKS STILL FAIL

1 evidence, 1 abstention, 2 routing

Measured behaviour	Baseline	Candidate
Expected passage selected Answerable questions	7 / 11	9 / 11
Explicit abstention Insufficient-evidence questions	0 / 3	1 / 3
Combined evidence / routing contract	7 / 14	10 / 14

What the result tells us

Three additional checks passed. Vocabulary handling and a relevance gate help, but the mortgage exception is still missed and the gate rejects one answerable car-coverage question. Two personal-data questions do not enter the required abstention state.

Scope of this sample

20 authored questions: 6 development + 14 reserved assessment checks. 18 paraphrased passages from 4 public pages. Extractive responses only; no hosted LLM was called. This is not a client case study.

TRACE THE SCORE BACK TO AN ACTUAL RESPONSE

Three checks worth reading

A07 / WRONG EVIDENCE / REMAINS UNRESOLVED

The exception loses to the default

Question: Do I need to wait thirty days if I buy cover while taking out a mortgage?

Candidate, P05: Returns the normal thirty-day waiting period.

Expected, P06: The mortgage-linked purchase exception has no waiting period. [S1, waiting-period FAQ]

Next action: Test exception-first retrieval using new default-versus-exception contrast cases.

A03 / IMPROVED EVIDENCE SELECTION

Hotel costs reach the correct exclusion

Question: Will flood insurance pay the hotel bill while my house is repaired?

Baseline, P10: Returns instructions to report a claim.

Candidate, P04: Selects the exclusion for temporary accommodation during repairs. [S1, exclusions FAQ]

Next action: Keep this as a regression check; source citations alone would not catch the old mismatch.

A04 / FALSE ABSTENTION / REMAINS UNRESOLVED

The relevance gate blocks an answer

Question: Does this flood cover include my car?

Candidate: Abstains: evidence score 0.1694 is below the configured 0.18 threshold.

Expected, P04: Retrieve the car exclusion from the supplied guidance. [S1, exclusions FAQ]

Next action: Calibrate the answer-versus-abstain tradeoff on a separate labelled set.

What to change next

Priority	Proposed change	Acceptance check
P0	Retrieve and compare relevant exceptions before selecting a general rule.	New paired questions select the correct default or exception.
P0	Define explicit answerability routing and acceptable qualified refusals.	Review A12/A13 with a domain owner; decide the intended response contract.
P1	Calibrate the relevance gate across short questions and paraphrases.	Measure missed answers and unsupported routing on a fresh set.

Method and limits

Baseline: word-count cosine retrieval. Candidate: word and character TF-IDF, vocabulary expansions, topic bonuses and a 0.18 relevance gate. Both use the same 18 passages and return the selected text.

A pass means the authored expected passage was selected, or the explicit abstention state was used where required. This does not measure generative answer correctness, hallucinations, model cost or production latency.

The questions, evidence labels and scoring design were drafted with AI assistance and prepared by the same author. Inputs were hash-locked before the first run; no configuration was changed after the assessment result. The reserved set is not an independent blind benchmark. No domain expert has validated the labels.

Rubric nuance: A13 returns a passage that already says an individual settlement cannot be calculated. Its failure is the missing explicit abstention state, not an invented payout. A domain review may accept that qualified response. The 10/14 result is specific to this strict routing rubric.

Sources and reproducibility

- [S1] [Buying flood insurance](#) - FEMA / NFIP, accessed 30 October 2026.
- [S2] [Starting a flood insurance claim](#) - FEMA / NFIP, accessed 30 October 2026.
- [S3] [What is a flood?](#) - FEMA / NFIP, accessed 30 October 2026.
- [S4] [NFIP glossary](#) - FEMA / NFIP, accessed 30 October 2026.

The accompanying pack contains the source map, all 20 cases, frozen settings, raw outputs, CSV results and the runner. Public US guidance is used only as an evaluation fixture; no personal coverage decision is made.

See this level of detail for your own workflow.

Discuss an Evaluation Health Check | [skorani.github.io](#)
Prepared by Samira Korani - Mileaway AI / MILEAWAY AI SYSTEMS LIMITED