

METHOD DOCUMENT

Evaluation method *and report outline.*

How EvalAnchor reviews a RAG assistant: from agreeing scope to a finding your team can reproduce, with the evidence behind every step.

Version 0.2 Prepared 24 September 2026 Status Draft, pending human review

Corpus Original fictional product policies

30

draft evaluation cases
in the launch kit

20 / 10

publication candidates
/ private holdout

2

independent human
reviewers per response

3

review dimensions per
response

01 Purpose

This document describes the method behind an EvalAnchor evaluation and the structure of the report a pilot produces. The launch kit demonstrates traceable RAG evaluation using original, fictional product policies: 30 questions and 60 authored fixture responses. No model API calls were made to create it.

The goal is not a single score. It is a set of findings that show **where** an assistant fails, **why**, and **what evidence** proves it.

Draft. This is a method document, not a reviewed sample report. Authored fixtures are not outputs from a tested customer system, and no benchmark scores are published.

02 Five-stage workflow

Every engagement follows the same connected stages, so each finding can be traced back to the question and evidence that produced it.

01	Agree on scope Use cases, authorized documents, review criteria, and data-handling terms are agreed before any material is shared.	OUTPUT Scope record
02	Create representative questions Everyday requests, edge cases, and questions whose answer is absent from the documents, each with a reference answer.	OUTPUT Case set with references
03	Inspect retrieval Did the system find the right documents, and select passages that actually address the question?	OUTPUT Retrieved vs. relevant evidence
04	Review answers and citations Each important claim is checked against the passage it cites.	OUTPUT Claim-level judgments
05	Classify failures Human reviewers assign a failure category and record their reasoning.	OUTPUT Reviewed findings

03 Review dimensions

DIMENSION	REVIEWER QUESTION
Answer correctness	Does the response meet the reference answer and expected behavior?
Evidence support	Do the cited passages support the response's claims?
Citation validity	Are citations resolvable, relevant, and attached to the right claim?

PassFailUncertain

For every judgment, reviewers record the evidence they relied on, their reasoning, any critical defect, and time spent.

04 Failure categories

RETRIEVAL

Missed document

A relevant source existed but was never retrieved, so the answer could not use it.

RETRIEVAL

Irrelevant passage

A retrieved passage does not address the question, yet shapes the answer.

ANSWER

Unsupported claim

A claim in the answer is not backed by any retrieved or cited evidence.

CITATION

Citation mismatch

The citation points to a passage that does not support the claim it is attached to.

05 Human review and adjudication

Reviewer A

Scores independently, records evidence and reasoning.

Reviewer B

Scores the same response without seeing A's judgment.

Adjudication

Disagreements and "uncertain" judgments are resolved explicitly before a case is accepted.

Automated checks verify file integrity and evidence references. They cannot approve a case.

06 Example finding record

Case EX-01 · Refund policy		ILLUSTRATIVE, FICTIONAL
QUESTION	How long do customers have to request a refund?	
RETRIEVED PASSAGE	Shipping policy · delivery times	
GENERATED ANSWER	"Sixty days. [1]"	
CITATION	[1] → Shipping policy	
FAILURE CATEGORY	Unsupported claim · Citation mismatch	
REVIEWER NOTE	The refund policy (not retrieved) states thirty days. The cited passage does not mention refunds.	

07 Report structure

- | | |
|---|--|
| <p>1 Corpus and configuration
Versions of documents and system under test</p> <hr/> <p>2 Case inventory
All cases, with exclusions and reasons</p> <hr/> <p>3 Metrics by system and split
Numerators and denominators shown separately</p> <hr/> | <p>4 Failure examples
Each linked to the exact passages involved</p> <hr/> <p>5 Disagreements and resolutions
How reviewer conflicts were settled</p> <hr/> <p>6 Limitations and reproduction
Steps to rerun the check yourself</p> <hr/> |
|---|--|

Reviewed cases · CSV + JSONL

Failure report with evidence

Reproduction steps

Prioritized recommendations

08 Limits of this method

- Source-ID checks do not prove semantic support; that requires human judgment.
- Authored fixtures are not a benchmark of any real model.
- A small, purposive case set does not estimate production reliability.
- Performance claims require human review and a run against a real, frozen system.