

SAMPLE GO-LIVE EVIDENCE PACK · 01

Meridian Health Systems

Retrieval-grounded question answering over eight public Nigerian health-regulation documents. One clean run, 50 questions, every number traceable to a committed file.

SPECIMEN · FICTIONAL CLIENT · NO REAL DATA

PREPARED AS A SAMPLE, NOT A CLIENT ENGAGEMENT

IN THIS PACK

- | | | | | | |
|----|--------------------------|----|-----------------------|----|------------------|
| 01 | Executive summary | 02 | One-page scorecard | 03 | Failure register |
| 04 | Fix list, ranked by risk | 05 | Method, limits, scope | | |

PREPARED FOR

Meridian Health
Systems
fictional client

MEASUREMENT

16 Oct 2026
single clean run

CORPUS

8 documents
1,218 passages

VERSION

v1.1.1
pack 1.0

01 The retrieval layer is the finding

Document-level hit rate reads 100% (25/25), but that metric is saturated and hides the real number. Chunk-level recall@5 against a graded gold set is **0.379** and MRR is **0.454** (n=23). Passage-level retrieval, not the model, is what costs answers.

The safety layer held. On 25 deliberately unanswerable questions the system refused all 25 — refusal correctness **1.000**, hallucination rate **0.000**. Before guardrails, the raw model produced two fabricated answers (Q27, Q48); both were intercepted before display. Citation correctness is **0.907** (49 of 54 inline citations point at something in the retrieved context) — measured, published, and not yet enforced at runtime.

Per-claim groundedness is scored for every answer shown to the user: mean **0.794 over 18 of 18**. The NHIA Operational Guidelines body was never recovered by OCR, so Q20–Q21 are correct refusals on a corpus gap and Q22 is the half-answered case, counted as an over-answer with its 0.7 score flagged as an open question.

0.379	0.907	1.000	0.794
RECALL@5 n=23 · the weakest measured layer	CITATION CORRECTNESS 49/54 · not enforced at runtime	REFUSAL CORRECTNESS 25/25 unanswerable refused	GROUNDEDNESS 18/18 scored by hand

MEASURED PERFORMANCE • EVERY BAR IS ONE PUBLISHED CELL

Recall@5 n=23		0.379
MRR n=23		0.454
Grounded answer rate 25		0.720
Per-claim groundedness 18		0.794
Citation correctness 54		0.907
Refusal correctness 25		1.000

How to read the chart. The two retrieval bars are the ones a buyer should care about: they say the right passage is often not in the five chunks the model is allowed to read. Everything above 0.7 measures the answer layer, which is doing its job on the passages it receives.

02 Every number, with its *n* and its source

Generated from committed artefacts by `analysis/render_results.py` and traceable to a row in `eval_results/b1_v1.csv`. No cell is computed by hand. *n* is stated for every row.

METRIC	VALUE	N	SOURCE
recall@5	0.379	23	b1_v1_metrics.json
context precision@5	0.174	23	b1_v1_metrics.json
MRR	0.454	23	b1_v1_metrics.json
document-level hit rate (saturated)	1.000	25	b1_v1_metrics.json
citation correctness	0.907	54	b1_v1_metrics.json
grounded answer rate	0.720	25	b1_v1_metrics.json
refusal correctness	1.000	25	b1_v1_metrics.json
hallucination rate, shown to user	0.000	25	b1_v1_metrics.json
hallucination rate, raw model pre-guardrail	0.080	25	b1_v1.csv (Q27, Q48)
per-claim groundedness (human-scored)	0.794	18	v1_scoring_sheet.csv
median end-to-end latency	7.48 s	50	b1_v1.csv
mean tokens per call (in + out)	1,493	50	b1_v1.csv

SIDE-BY-SIDE ON THE ORIGINAL 30-QUESTION SET

METRIC	VALUE	N	NOTE
recall@5	0.379	23	identical – retrieval is deterministic
citation correctness	0.898	49	44/49
grounded answer rate	0.680	25	17/25
refusal correctness	1.000	5	5/5
hallucination rate	0.000	5	0/5
groundedness	not scored	–	a score belongs to one answer text, never to a question id

The 30-question set shares Q1–Q30 with the 50-question run, so its retrieval numbers are identical by construction. The groundedness row reads *not scored* rather than zero: the human sheet was written against the 50-run's answer texts, and those scores are never attributed across runs.

03 Counted, not described

Each class carries its count and the question ids, so any claim can be opened against a row. Generated block in `notes/failure_modes.md`, which reads `eval_results/b1_v1.csv`.

LAYER	CLASS	COUNT	QUESTION IDS
retrieval	gold chunk outside top-5 (answerable, has gold)	7	Q3, Q5, Q8, Q10, Q11, Q14, Q16
retrieval	diversity-cap violation (>3 chunks from one document)	0	none
retrieval	cap engaged (exactly 3 chunks from one document)	46	not enumerated
answer	over-refusal by a guardrail (answerable, in corpus)	0	none
answer	over-refusal by the model itself (answerable, in corpus)	5	Q3, Q5, Q10, Q11, Q14
answer	refusal of a documented corpus-gap question	2	Q20, Q21
answer	over-answer on a corpus-gap question	1	Q22
answer	correct refusal (unanswerable)	25	Q26–Q50
answer	visible hallucination (unanswerable, answer shown)	0	none
answer	raw hallucination caught before display	2	Q27, Q48
answer	citation target missing (≥ 1 citation not retrieved)	4	Q1, Q6, Q16, Q24

Act ↔ guidance rank flips

On 19 of 50 questions an Act and its sibling guidance both appear in the top five, and the order is not stable in either direction: the Act outranks its guidance on 8 questions, the guidance outranks the Act on 11. The dangerous instance is the Q20–Q22 group, where the NHIA guidance wins top ranks on front matter alone because its body was never recovered by OCR.

Front-matter and TOC dilution

9 of 50 questions have at least one retrieved chunk shorter than 200 characters in their top five. Heading-quality guards removed most TOC ghost-headings; what survives is concentrated on the scanned Acts.

Q22 is the honest middle of a corpus gap, not a miscount. Its funding half is grounded in the Act's Part VII (retrieved at rank 5, now a graded gold chunk). Its accountability half is extrapolated from a recovered section heading whose body was never OCR'd. A human scored the answer 0.7 — partly grounded — and this pack flags that score as an open question because it sits on the rubric's band boundary.

04 What to fix first, and how you will know

Ranked by measured defect, not by effort. Each row names the verification that moves on a re-run — a fix that cannot be checked is not a fix.

#	FIX	WHAT IT UNBLOCKS	HOW IT IS VERIFIED
1	Passage-level retrieval. Raise the internal fetch before the diversity cap; add section-label-aware re-ranking; or hard-filter front-matter chunks out of a slot.	The 7 gold-miss questions (Q3, Q5, Q8, Q10, Q11, Q14, Q16) and, through them, most of the 5 model over-refusals. This is the single largest measured defect.	Re-run <code>run_eval.py --prefix b1_v1 --strict --resume; recall@5</code> and the gold-miss count move, or they do not.
2	Citation-target verification at runtime. After generation, confirm each cited title and section exists in the retrieved set; strip or re-prompt on a miss.	The 5 citations in 4 displayed answers that point outside the provided context. Citation correctness is 0.907; one of the five is a chunk-label artefact rather than a real miss.	Citation correctness rises toward 1.0 on the same run; <code>citations_missing</code> falls.
3	Acquire a text-complete copy of the NHIA Operational Guidelines 2023. The current PDF is a phone capture whose body was never recovered.	Q20–Q21 become answerable instead of correct refusals; Q22's accountability half becomes answerable instead of extrapolated; the front-matter failure on that group disappears.	The gold set gains chunks for Q20 and Q21; validation passes; the corpus-gap refusal class drops to 0.
4	Human second pass on the gold chunks. Currently agent-graded, with every evidence string verified verbatim, but not audited by a second grader.	Raises the trustworthiness of every retrieval metric built on the gold set.	The second grader's disagreements are recorded; metrics are recomputed against the audited set.
5	Citation-sufficiency verification. Does the cited passage actually support the claim built on it?	Closes the blind spot that let Q22's extrapolated half through: distance checks see a relevant chunk, target matching sees a present citation, and neither can read the passage.	A sufficiency metric separates Q22 from the 17 fully grounded answers on the same frozen run; only then wire it into the guardrail stack.

FIX 1 MOVES

Recall@5 and the gold-miss count. Both are on the scorecard.

FIX 2 MOVES

Citation correctness and `citations_missing`. Both are on the scorecard.

FIX 3 MOVES

Two refusal classes and one over-answer class. All three are on the scorecard.

What this list is not. It is not a commitment to implement, and it is not a guarantee that any metric will improve. Meridian's engineers own the fixes; the pack names what the evidence points at and how a re-run would falsify it.

05 How it was measured, and what it cannot tell you

Gold set

50 hand-built questions: 25 answerable, 25 unanswerable across four strata — clinical, out-of-corpus, near-miss, mixed. The unanswerable half is synthetic by design, because a question the documents cannot answer is exactly where a chatbot is most tempted to invent something.

Gold chunks

For the 25 answerable questions, the passages that contain the answer were graded by an AI agent, with every evidence string verified as a verbatim substring of its chunk. Not yet audited by a second human — fix 4 on the previous page.

What recall@5 does not capture

It measures whether graded gold chunks appear in the retrieved top five. It does not measure whether the model used them correctly, whether a non-gold chunk also carried a usable answer, or whether the grading itself is complete.

Citation correctness

An inline citation is correct when it matches a retrieved chunk under one of four auditable rules, in order: exact title and section; whitespace-insensitive pair; normalised title containment with section containment; and the whitespace-insensitive form of that. Multi-comma Act titles are resolved against known document titles before any split. It catches citations pointing outside the provided context. It does not judge whether a matched citation supports the adjacent claim.

Groundedness rubric

1.00 — every factual claim supported by a retrieved chunk. 0.70–0.95 — correct and grounded, with a minor omission or a mis-scoped detail. 0.60–0.70 — partly grounded, with inferred claims. 0.0 — refusal on an answerable question. Groundedness measures faithfulness, not correctness, and not citation accuracy. One score, Q22's 0.7, sits on a band boundary and is recorded as an open question rather than a settled number.

Honest limitations

Fifty questions; synthetic unanswerables; agent-graded gold, not second-audited; groundedness scored by a single human grader with no second-grader agreement number; evaluation on a rate-limited free-tier endpoint, with retries and resume in the harness. A 0% hallucination rate on 25 synthetic unanswerable questions is a floor at 4-point resolution, not a guarantee.

Every number in this pack can be regenerated from the delivered scripts by a third party in under 30 minutes: build the corpus and index, run the evaluation, render the tables. The per-claim groundedness score is the one number a machine cannot produce.

REPRODUCIBILITY

A stranger test was performed twice from a fresh clone: one documented command sequence, zero manual steps, **20 minutes 7 seconds** wall-clock, ending in 8/8 content-verified documents, 1,218 chunks, and a passing drift check on the bare clone.

OUT OF SCOPE

This pack is an evidence report, not legal advice, not a compliance opinion, not a certification, and not a guarantee of any future result. Nothing here states that any facility complies with any regulation. The harness is delivered as-is; the ranked fix list is a recommendation, not a commitment.