

LITERATURE REVIEW · SOURCE-BOUND SYNTHESIS

Evaluating retrieval-augmented generation

What the foundational literature establishes, where evidence is thin, and how to design an evaluation.

CORE PAPERS

3

read in full

CLAIMS
MAPPED

18

to sources

OPEN
QUESTIONS

5

evaluation design

COVERAGE

**2020-
2023**

selected

EXECUTIVE READ

The selected literature establishes the value of combining retrieval with generation and the practical strength of dense retrieval. It does not support a single universal quality metric; evaluation must separate retrieval, grounding, answer quality, latency, and cost.

01

**Lewis et al.
(2020)**

02

**Karpukhin et al.
(2020)**

03

**Gao et al.
(2023)**

LITERATURE REVIEW · SOURCE-BOUND SYNTHESIS

Evidence, decisions, and a usable handoff

The complete file keeps the working structure editable and the review points visible.

DIMENSION	QUESTION	MEASURE	FAILURE TO INSPECT
Retrieval	Was useful evidence found?	Recall · rank	Corpus gaps
Grounding	Does the answer follow evidence?	Citation support	Contradictions
Answer	Is it useful and correct?	Task rubric	Abstention
Operations	Can it run reliably?	Latency · cost	Tail behaviour

DECISION / NEXT REVIEW

Use a layered evaluation: retrieval set, grounded-answer rubric, abstention cases, and operational thresholds. Do not collapse them into one score.