



PRACTICAL AI

Evaluate Retrieval with Evidence

Measure ranking quality without hiding missing labels.



Sanjaay Gowda T S

FREE SAMPLE / INTRODUCTION + CHAPTER 1 / EDITION 1.0.0

sanjaaygowdats.ai

About this free sample

Evaluate Retrieval with Evidence

Sanjaay Gowda T S

First edition 1.0.0, 2026-10-13

Copyright 2026 Sanjaay Gowda T S. This is a labelled free sample, not the complete ebook. You may share this sample unchanged. It contains the cover, sample permissions, contents, introduction, Chapter 1 and a closing product link. The remaining chapters, worked solutions and companion files are not included. The full ebook text, diagrams and artwork have separate single-reader terms; this sample permission does not authorize sharing the full purchased PDF or EPUB.

The companion source code, tests and fictional data are separately MIT licensed. You may adapt or share that project with its included licence notice. MIT does not apply to the ebook text or book artwork. The books and companion projects have no technical activation requirement or buyer-specific watermark.

This first edition was prepared with AI assistance. The case studies are fictional teaching examples. Any measured outputs come from the supplied examples, not production customer data. The teaching scope, assumptions and limitations are explained in the chapters. Independent human technical, editorial and assistive-technology reader review is pending. Automated tests and accessibility checks do not establish certification or suitability for every use.

Run examples in a separate working folder using the supplied synthetic data. Review inputs and calculations before adapting the techniques to real operations. Model software, service access, subscriptions, ongoing updates and consulting are not included.

Author website: sanjaaygowdats.ai

Ebook terms: sanjaaygowdats.ai/ebooks/license

Inside this free sample

Start here.....	4
1. Define the retrieval decision	5
Continue with the full guide	7

This free sample contains the introduction and Chapter 1 only. The closing link opens the complete guide. The companion project and worked solutions are supplied with the complete ebook.

Start here

A retrieval system can improve an average while becoming less useful for the question that matters most. A new configuration might find more summaries, push the current policy below an old copy, and return something for a question the collection cannot answer. A single success percentage hides these changes. This book teaches you to expose them before retrieval feeds a generated answer.

You will evaluate ordered evidence identifiers against explicit judgments. The companion receives rankings as input; it does not contain an embedding model, search service or language model. This separation lets you test the evaluator independently of whichever retriever eventually supplies its lists. Every question, identifier and ranking in the example is an invented fixture for the fictional Harbor Equipment Library. None is an observed customer result or a real model output.

Use Python 3.10 or later and basic lists, dictionaries and functions. Open a terminal in the extracted project directory. Confirm that `main.py` and `example-input.json` are siblings, then run:

SHELL

```
python --version
python main.py --input example-input.json
python -m unittest discover -s tests -v
```

Windows users may replace `python` with `py -3`; macOS and Linux users may need `python3`. No packages, accounts or API keys are required. The output should match `expected-output.json` as JSON data. If Python cannot find `main.py`, check the working directory. Run tests from that directory too, because the tests load the example fixture.

Your deliverable is a reviewable retrieval decision: versioned cases, justified grades, complete rankings, defined metrics and visible regressions. Passing unit tests verifies specified evaluator behavior on those tests. It does not establish a retriever's accuracy, the representativeness of your labels or the support of a later generated answer. Keep those claims separate throughout the exercises.

1. Define the retrieval decision

Begin with an evidence need

Harbor Equipment Library lends tools to community members. Its fictional documents include a current renewal policy, a summary, an obsolete policy, deposit rules, deposit exceptions and opening hours. A member asking whether a borrowed drill can be renewed needs the current rule. A page containing drill and renewal may discuss the subject without supplying the applicable condition. Your evaluation must represent that difference.

Write a sentence before assigning labels: "Find the current rule and any exception needed to determine whether this loan can be renewed." This makes the evidence need concrete. If the task instead asks staff to locate any introduction to renewals, the summary might deserve the highest grade. The same ranking can suit one task and fail another. Relevance relates a question and task to a particular source version.

Decide whether you judge files, passages or complete evidence bundles. File labels are convenient when every file contains one short rule. They mislead when a handbook contains many unrelated sections. A matching filename can count as success even when the returned paragraph omits the rule. Passage labels reduce that ambiguity, but require stable passage identity and an annotation process that survives corpus changes.

Separate four decisions

First, did the experiment run the planned case? Second, did returned evidence contain relevant material? Third, was the strongest evidence early enough to fit the context budget? Fourth, could a reviewer answer the complete question from that material? The companion calculates selected ranking properties and exposes missing rankings. It does not resolve the final question automatically.



Figure 1. Versioned cases and judgments meet a recorded ranking before metrics and case review.

Read Figure 1 as accountable inputs. A case without judgments cannot produce a meaningful relevance score. A ranking without a case identifier cannot be compared reliably. A report without a corpus version cannot establish which policy the assessor saw. Keeping these dependencies visible is more useful than adding a complicated aggregate before basic records exist.

Declare local conventions

Here, k means the maximum returned positions evaluated. A positive grade means relevant. Grades distinguish evidence strength, but every positive grade contributes equally to binary recall. Unknown cases contain no positive judgments. An absent ranking is treated as empty for arithmetic and separately marked missing. These are explicit project conventions rather than universal requirements for evaluation software.

The project evaluates evidence availability, not policy truth. A grade-three document can itself contain a mistake. Source-owner review determines whether the corpus is suitable; retrieval evaluation determines whether approved evidence reaches the relevant position. You need both responsibilities when an assistant will influence decisions. Do not interpret an authoritative-looking identifier as permission to skip source review.

Exercise: write an evidence contract

Choose a fictional renewal question. Write the user, required source version, evidence unit and what would make it unanswerable. Explain why finding the obsolete policy should not count as success. Keep this contract beside the question when labeling it. A reviewer should be able to challenge your grade using the contract rather than guessing your preferences.

Continue with the full guide

This free sample ends after the introduction and Chapter 1. The complete guide contains nine teaching chapters, worked exercise solutions, primary references and a separate Python companion project. The full PDF, reflowable EPUB and project files are supplied together.

Use the complete guide to work through the remaining exercises and inspect the supplied examples. The companion code and fictional data are separately MIT licensed; the complete ebook has the single-reader terms described on its product page.

The examples are teaching fixtures, not production customer results. Independent human technical, editorial and assistive-technology reader review remains pending. Software access, subscriptions, consulting and guaranteed results are not included.

[View the complete guide and purchase options.](#)

You may share this labelled free sample unchanged. Keep the full purchased PDF and EPUB private. Thank you for reading.