

RAG guardrails for a course assistant

Requests, evidence and the decision to answer

Answer
Block
Redirect
Abstain

A Python project with OpenAI-compatible model
integration

Four possible outcomes

Each outcome tells the user something different.

Answer

The available evidence supports a permitted response.

Block

A policy check stops a prohibited request.

Redirect

Offer allowed tutoring help for an integrity request.

Abstain

The system cannot establish sufficient support.

Failure modes and controls

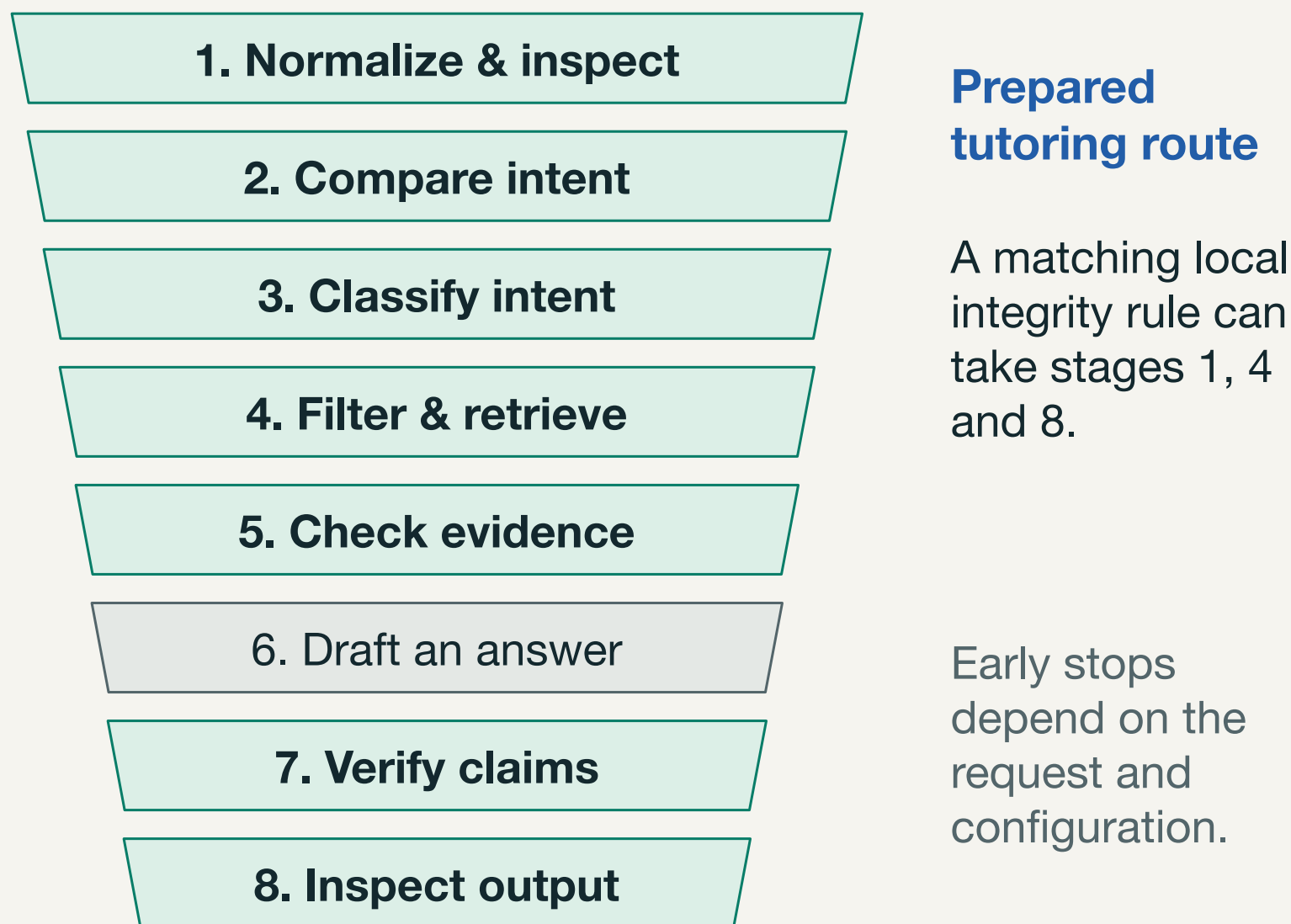
Every control has a specific job.

Failure mode	Relevant control
Instruction override	Request rules and classification
Private-data request	Request policy and document scope
Harmful request	Request and output checks
Injection in a source	Context checks before generation
Unsupported claim	Evidence and entailment checks
Misleading citation	Citation-to-source checks
Graded work request	Prepared tutoring redirect

These controls reduce targeted risks. Permissions, source integrity and independent evaluation still need separate attention.

The configured pipeline

Eight stages with early exits and a tutoring branch.



Generation is a pipeline stage. The LLM judge evaluates saved outputs separately.

A relevant source is only the beginning

Three checks answer different questions.

Retrieval relevance

Does this material help answer the question?

Claim support

Does the cited passage support what the answer says?

Source integrity

Is the material trustworthy, current and permitted?

Entailment against supplied text cannot establish the truth of that source.

A policy rule you can inspect

A local draft and its test result.

Configured pattern

Regex rules
Rules execute before model classification.

Add rule

Trigger

prompt_injection

Patterns · one per line

discard all earlier instructions

Remove

Block

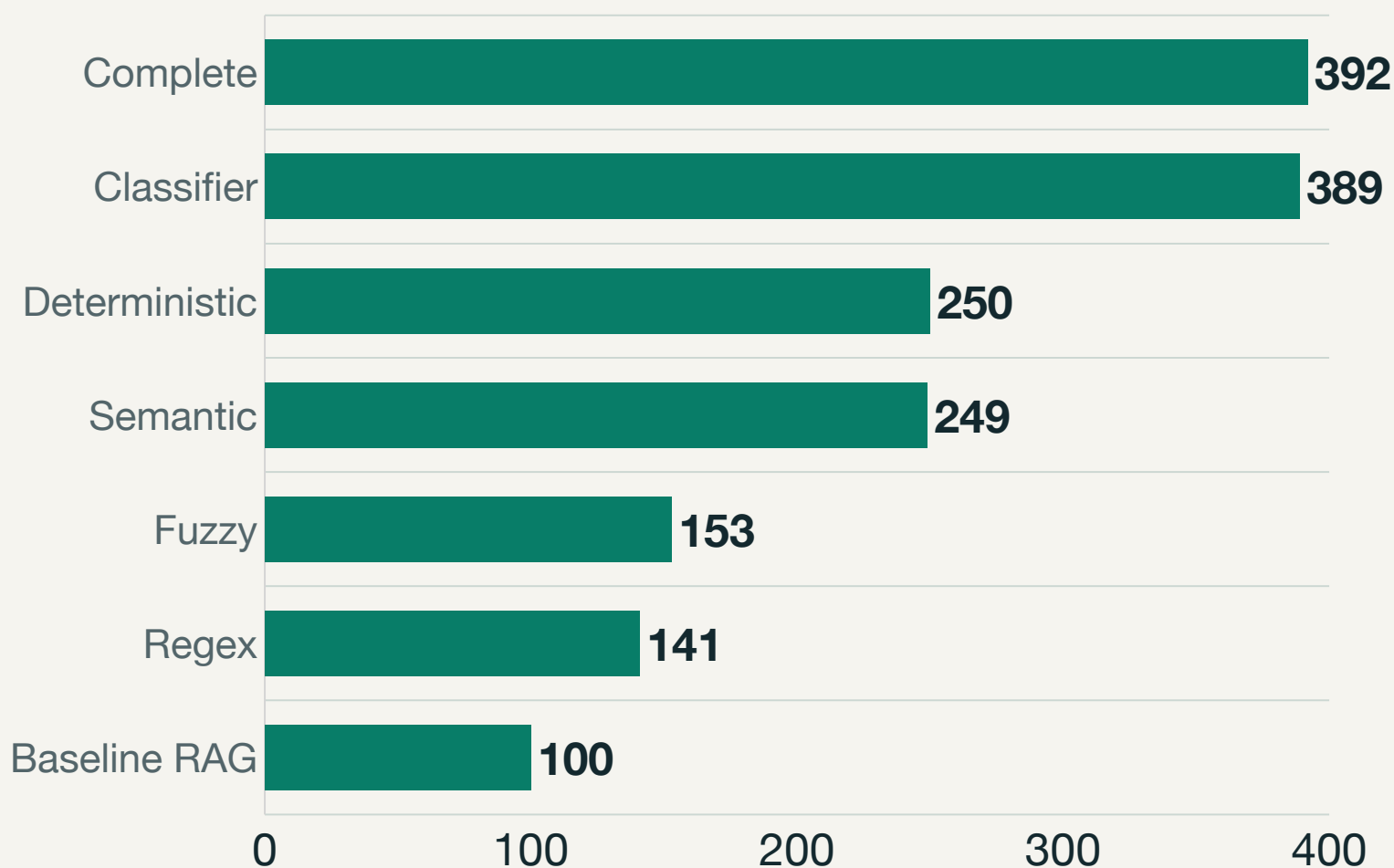
prompt_injection

The configured phrase triggers a block in the offline policy preview.

Actual local capture. No remote model calls. A rule should also be checked against benign requests.

392 correct decisions out of 400 requests

Balanced calibration set: 100 cases per disposition.



Correct dispositions / 400

Whole-configuration comparison. Shared controls prevent attributing each gain to one layer. The system holdout remains unopened.

The eight missed answers

All eight errors were false abstentions.

2

Retrieval misses

The expected document was not retrieved.

2

Answerability stops

The expected document was present.

4

Entailment rejections

The candidate added unsupported claims.

Answer recall: 92 / 100. Recovering useful answers requires locating the failed check.

What the evidence supports

A scoped prototype with inspectable decisions.

Calibration comparison

The seven scenarios share one 400-case benchmark.

A separate judge study

Judge agreement measures consistency with its reference labels.

Open evaluation work

Independent source review and a frozen system holdout are still pending.

Zero observed unsafe answers in this finite split does not establish zero deployment risk.

Code, demo and evidence

Inspect the implementation and the recorded results.

Project repository and local demo

[github.com/zhittsova/
llm-rag-guardrails-ai-course-assistant](https://github.com/zhittsova/llm-rag-guardrails-ai-course-assistant)

Recorded calibration report

Pipeline source

Blog: zhittsova.com

Source revision: 7e886ea
Model and policy changes require fresh evaluation.