

Season 2 intro

Retrieval and RAG.

Build an assistant that answers from your own documents. Then measure every stage until you know why it is right, and why it is wrong.

Would you ship these three answers?

A CUSTOMER ASKS

How much can engineers spend on parts for the Mira robot vacuum without asking for approval?

ASSISTANT

Engineers can fit parts worth up to **£163.00 in total**, using current internal parts prices, without asking for approval. [P1]

A CUSTOMER ASKS

What should I do if I have a wall bracket from batch 4867 for my Brio kettle?

ASSISTANT

Stop using the wall bracket from batch 4867 and return it for a refund. ... Contact Halvard Home customer support for return arrangements. [P1]

A CUSTOMER ASKS

How long should I wait for the Tessa dishwasher to cool down before I clean it?

ASSISTANT

The passages give conflicting cooling times: **11 minutes in P1** and **7 minutes in P2**. To be cautious, wait **at least 11 minutes**.

All three were wrong. None raised an error.

Real answers from this season's first system (S2 E2), on an invented appliance maker's documents.

PROBLEM Leaked

The £163 parts limit

It comes from a **staff-only engineer manual**. The customer should have been told nothing.

PROBLEM Withdrawn

The batch 4867 instructions

That notice was **withdrawn on 19 May 2026** and marked deleted. It was still in the index, so it was still an answer.

PROBLEM Out of date

The 11-minute cooling time

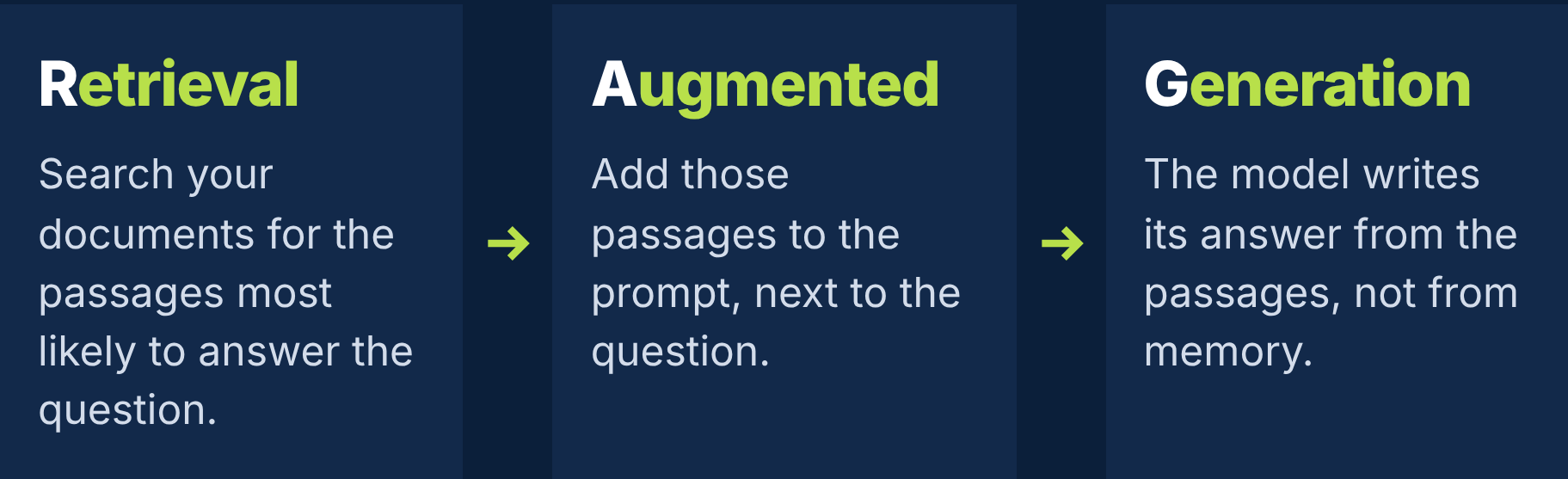
It is from the **superseded safety policy**. The current one says 7 minutes. Both versions were searchable.

Notice what is missing

No crash, no warning, no low-confidence flag. Each answer is fluent, and two cite a source. This is how a retrieval system fails.

RAG: retrieval-augmented generation

A model knows only its training data. Your documents were never in it. RAG hands the model the right pages at the moment of the question.

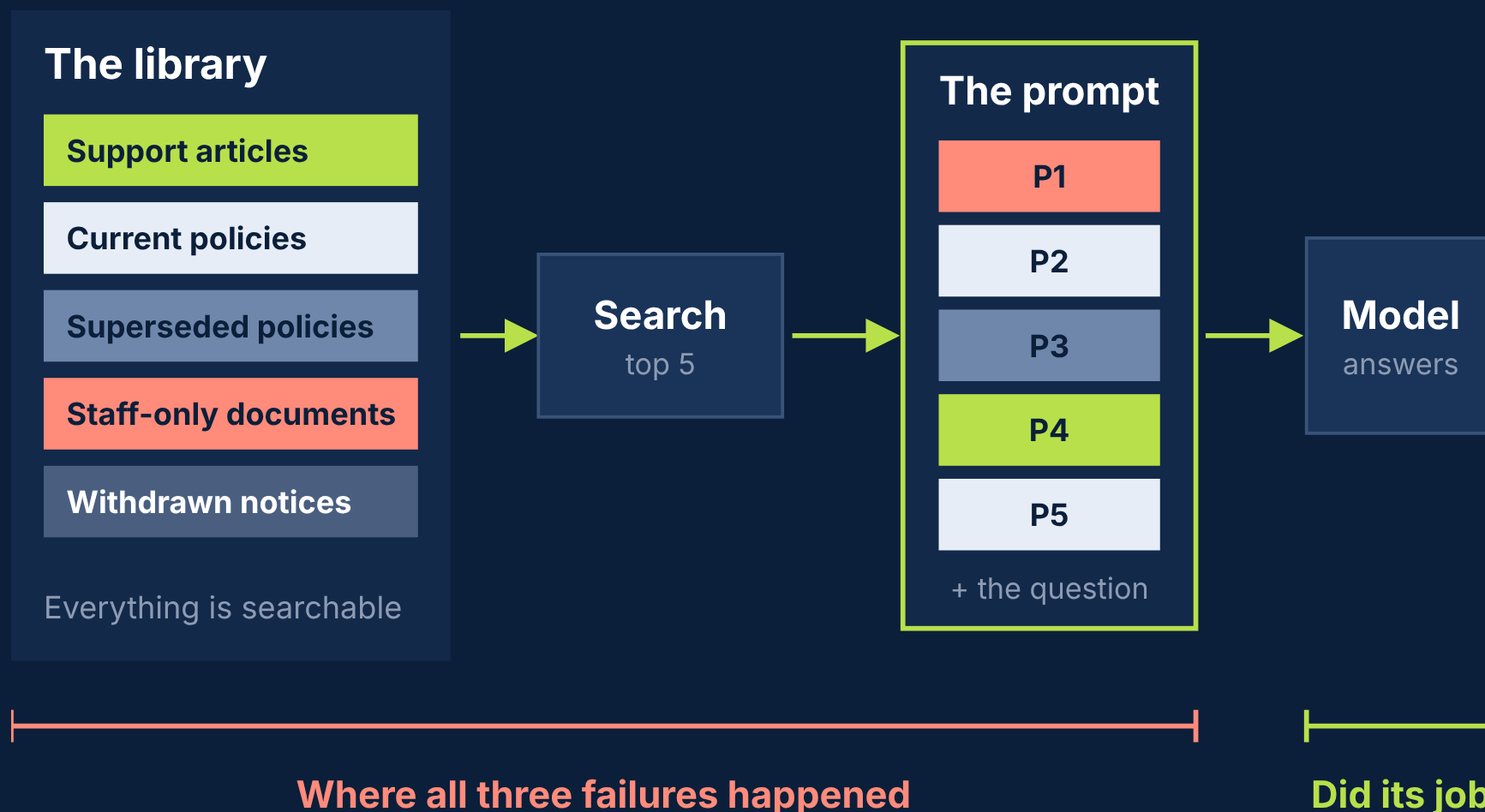


An open-book exam, where **search picks the pages** the model is allowed to read.

Example: the season's final system, on a test question	
QUESTION	How long should I wait for the Tessa dishwasher to cool down before I clean it?
RETRIEVED	Safety policy, current > Tessa dishwasher > Safe use "Let the Tessa dishwasher cool for 7 minutes before cleaning it."
ANSWER	Wait 7 minutes before cleaning the Tessa dishwasher. [P1]

The model did its job. It was handed the wrong pages.

A second model checked each of the three answers against the passages it was given. Every claim was supported: all three were **faithful**.



The season in one sentence

An answer can only be as good as the passages it is given. In the first system, most wrong answers lost the right passage before the model ever saw it.

Learn it, test it in code, keep what works

Advice about RAG is easy to agree with and hard to trust. So every idea this season is tested on one realistic case study, in code you can run.

1

Learn the idea

What a stage does, and how it can fail, as a mental model before any numbers.

**2**

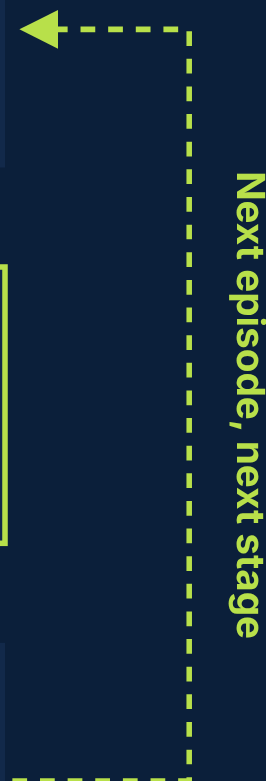
Test it on Halvard Home, in code

Claims and pass marks fixed first. Then a real run on the same library and the same 520 questions.

**3**

Keep the rule, take it home

A change joins only if it clears its mark. The decision rule is what you apply to your own data.



Your role this season

You are the engineer building Halvard Home's assistant. The brief: answer customers and staff from the company's own documentation. Wrong answers cost money; leaks cost trust.

Halvard Home is invented, on purpose

We created the company and wrote all of its documentation for this course, so the theory can be practised on something that behaves like real work.

01

It looks like real work

Short support articles, long policies, outdated versions, staff-only manuals and withdrawn notices: the mix real companies keep.

02

No model has seen it

It was written for this course, so it is in no model's training data. A right answer can only come from retrieval, never memory.

03

Every answer is known

Each fact was planted on purpose, and every planted value is unique. Code, not opinion, checks whether it was found.

04

Anyone can rerun it

Public, MIT licensed, and frozen on 12 October 2026, before any episode ran. Nothing was tuned to the results.

The trade-off, stated up front

A model wrote the prose around planted facts, so the library is cleaner and more uniform than real documents. Each episode's "What this does not prove" says where that matters.

Meet Halvard Home

A home appliance maker with ten products and two kinds of people asking its assistant questions.

- Ashby tumble dryer
- Brio kettle
- Corra induction hob
- Fenwick cooker hood
- Keel washing machine
- Lumen smart thermostat
- Mira robot vacuum
- Orrin heat pump
- Tessa dishwasher
- Vale fridge freezer

Customers	Staff
May be told anything in the public documents, and nothing else.	May also be told what is in the internal documents.

A000@1.mdone of 344 documents

title: Resetting your Ashby tumble dryer
products: Ashby tumble dryer
access: public status: current

On the Ashby tumble dryer, E-2194 means a factory reset was interrupted before the saved settings could be cleared. Your previous preferences may still be active. To try the reset again: 1. Switch the dryer off, unplug it and wait five minutes. ...

Highlighted: a planted fact. The labels above it (access, status) are what E8 filters on.

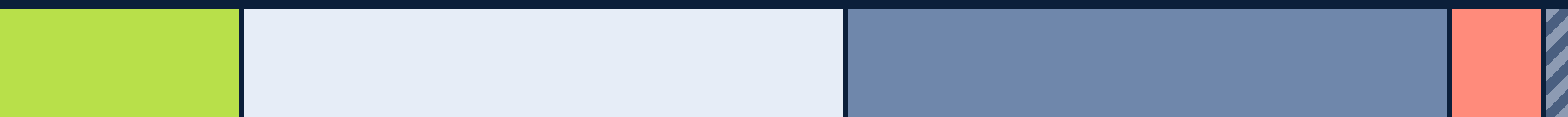
Its documentation

The library the assistant searches. The current public part alone is about 155,000 tokens: more than a 128,000-token context window holds.

By count 344 documents



By length about 289,000 tokens



Most documents are short; most of the text is in eight long ones. Both shapes need handling.

-  **300 Support articles.** Short, one fact each, many carrying error codes and part numbers.
-  **4 Policies:** warranty, returns, safety, servicing. About 27,900 tokens each.
-  **4 Superseded versions** of those policies, with some terms changed.
-  **4 Staff-only documents** a customer must never be shown.
-  **32 Withdrawn notices,** marked deleted.

520 questions with known answers

Sixteen types, in five families. Each tests a different way to fail.

160 questions	Find the passage Error codes, paraphrases, look-alikes, shared words "What does error E-2194 mean?"
120 questions	Read the long policies One sentence, two sentences, needs its heading, which version "How often should I change the filter in my Corra induction hob?"
80 questions	Join it up Two questions in one, two passages to connect, follow-ups "What about the Mira robot vacuum?"
40 questions	Asked by staff The answer is in an internal document, and this asker may see it "How long does the Tessa dishwasher's built-in test take?"
120 questions	Should be declined Staff-only, detail not covered, topic not covered, withdrawn "What is the cooking capacity of the Halvard air fryer?"

Each answerable question is labelled with the exact characters that answer it, so code, not opinion, decides whether search found it.

One test question, run for real

Every question is a record like this. The pipeline runs it; code and a judge score it.

```
"id": "IDENT-000",
"question": "What does error E-2194 mean?",
"asker": "public", "answerable": true,
"evidence": A000, characters 35 to 150
```

	First system E2	Final system E16
SEARCH	Right article ranked 11th . Not in the top 5.	Right article ranked 1st .
ANSWER	"The provided passages don't mention error E-2194, so I can't determine what it means."	"E-2194 means the Ashby tumble dryer's factory reset was interrupted before its saved settings could be cleared... [P1]"
SCORED	Search miss. Wrong answer. Lost at ranking.	Search hit. Right answer. Cited and faithful.

Multiply by 520 questions and you have the scoreboard. Same item, same library; only the pipeline changed.

One pipeline, eleven stages

E2 builds it with the simplest choice at every stage. Each later episode changes one stage.

INDEXING	1	Load documents	E2
	2	Chunk cut into passages	E4
	3	Embed and index	E2 E12
EVERY QUESTION	4	Rewrite the question	E6
	5	Search	E5
	6	Filter permissions, versions	E8
	7	Rerank	E7
	8	Build the prompt	E9
	9	Answer	E2
	10	Cite	E10 E11
	11	Score	E3

Filled chip: the episode that changes the stage. Outlined: a field note that tests it.

Four numbers, every episode

Search and answer are scored apart, so a wrong answer can be traced to the stage that caused it.

Search

HIGHER IS BETTER

of 400 answerable

The right passage is in the **top 5** handed to the model.

Right answers

HIGHER IS BETTER

of 400 answerable

Judged **correct**, and **faithful**: every claim supported by the passages.

Not found

HIGHER IS BETTER

of 120 unanswerable

The system **declines** when the library cannot answer, or this asker may not be told.

Leaks

TARGET: ZERO

of 480 public questions

A **staff-only passage** reached a customer's prompt.

Each episode closes on this board: the stage it changed, the four numbers before and after, and whether the change joined the system.

A change has to earn its place

1

Write the claim

What the change should improve, and on which questions.

2

Fix the pass mark

The score the change must reach, decided before any result exists.

3

Commit both to the code repo

Dated, public, and never edited after the run.

4

Run it on the full test set

Same library, same questions, same scoring as every other episode.

Clears the mark

It joins the reference system, and every later episode builds on it.

Falls short

It stays out, and the episode says so and shows why.

Why so strict

A pass mark set after the run can be bent to fit any result. Set before, it keeps noise out. The price: some real gains stay out too. You will see which.

Every episode runs in the same order

So you always know where you are, and where to look for what you need.

01	The question	One question the episode answers
02	Mental model	The idea, drawn before it is measured
03	The experiment and its claims	What was tested, and the pass marks, fixed in advance
04	A worked example	One real test question, start to finish
05	Findings	Charts with 95% intervals; "better" only when they do not overlap
06	What this does not prove	The limits, stated plainly
07	Run it yourself	The code that produced the result
08	For engineers, for leaders	What to build, and what to ask your team
09	Scoreboard and decision rule	The four numbers, and one sentence to keep

Sixteen episodes, four arcs

ARC 1 Foundations

- E1 Why retrieval exists
- E2 Your first RAG system
- E3 Measuring it

ARC 2 Finding the passage

- E4 Chunking
- E5 Search: keyword, vector, hybrid
- E6 Rewriting the question
- E7 Reranking
- E8 Filters, permissions and fresh documents

ARC 3 Using the passage

- E9 Building the prompt
- E10 Grounding and citations
- E11 Does a citation rule change the answer? FIELD NOTE

ARC 4 Running it and deciding

- E12 Swapping the embedding model FIELD NOTE
- E13 Long context or retrieval? FIELD NOTE
- E14 Tracing a wrong answer
- E15 When RAG is the wrong answer
- E16 Season 2 recap

Build it, measure it, then improve the stages that **find** the passage before the ones that **use** it.

Three claims worth testing on their own

A field note takes one piece of common advice and runs it against the season's system, with its own pass marks fixed in advance.

E11

Does requiring a citation change what a model says, or only how it says it?

The same questions, with and without a citation rule.

E12

What changes when the embedding model does, and what does re-embedding cost?

A smaller model against a larger one, question by question.

E13

When does sending every document beat a retrieval pipeline?

Every current public document, about 160,000 tokens, in one prompt, against the pipeline.

The same three questions, sixteen episodes later

Same library, same questions. Different pages reach the prompt.

How much can engineers spend on parts for the Mira robot vacuum without asking for approval?

I could not find an amount that engineers may spend on parts without approval. The passages say chargeable work requires your agreement, but do not specify a parts-spending limit. [P1][P2]

WHAT CHANGED Customers no longer search staff-only documents (E8), and "not found" is an allowed answer (E10).

What should I do if I have a wall bracket from batch 4867 for my Brio kettle?

I could not find information about a wall bracket from batch 4867 in the passages.

WHAT CHANGED Withdrawn notices are filtered out of search (E8).

How long should I wait for the Tessa dishwasher to cool down before I clean it?

Wait 7 minutes before cleaning the Tessa dishwasher. [P1]

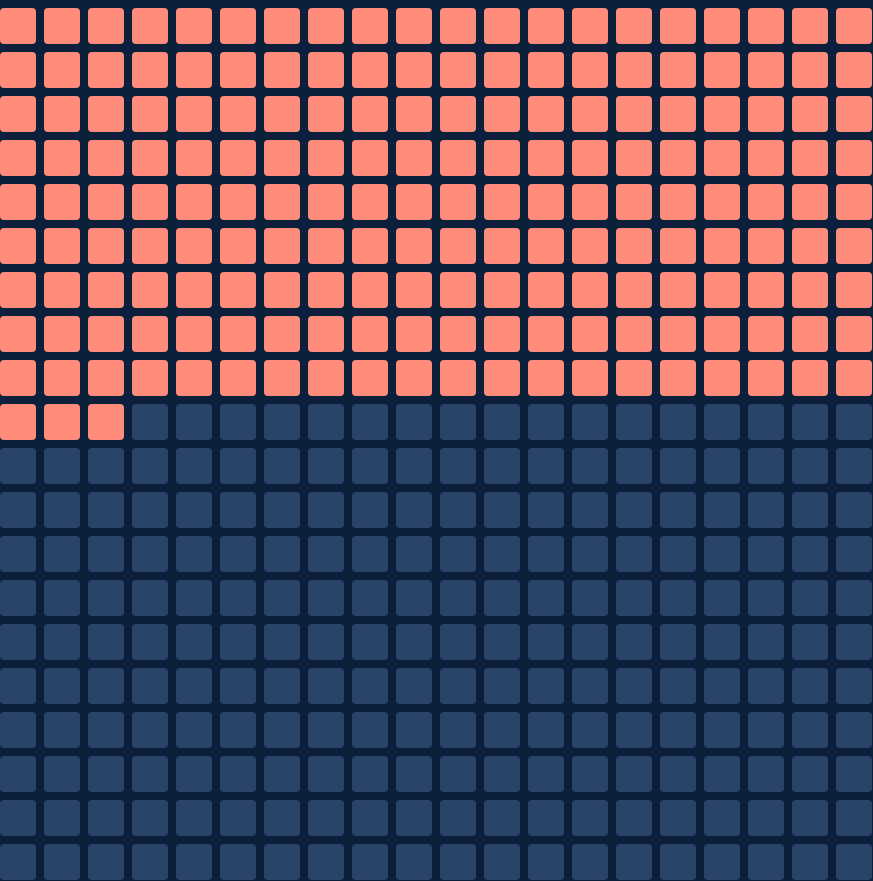
WHAT CHANGED Only the current policy is searched (E8), and the answer cites it (E10).

Same questions. Far fewer wrong answers.

Each square is one of the 400 answerable questions. Coral marks a wrong answer.

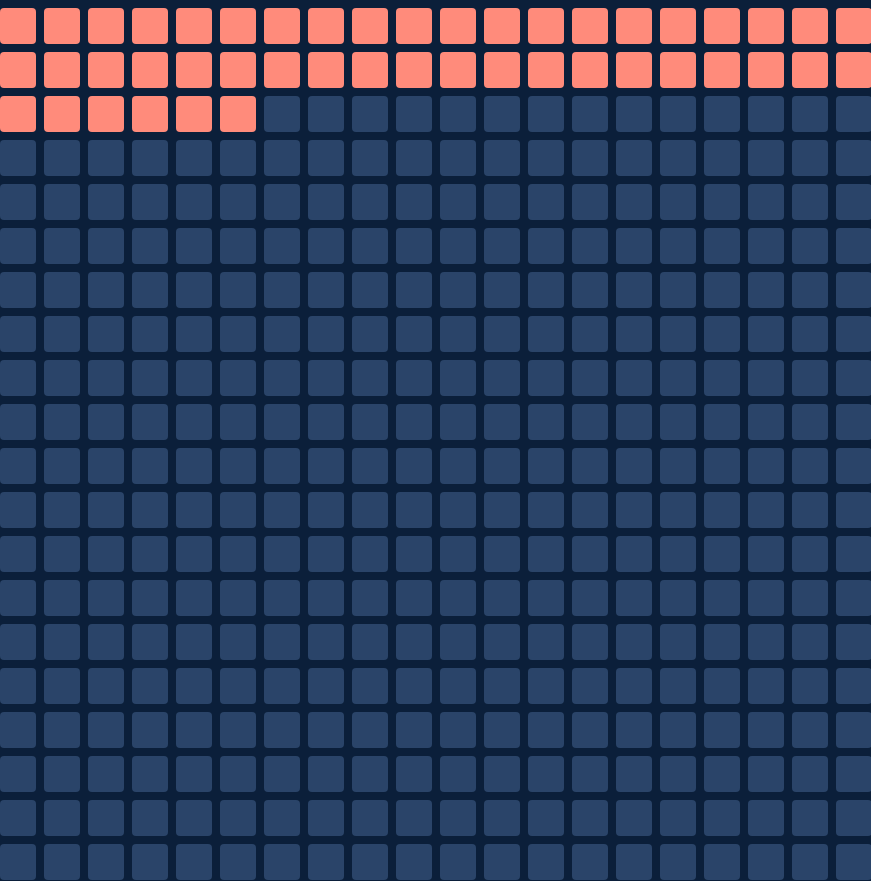
The first system

E2: simplest choice at every stage



The final system

E16: only the changes that earned a place



■ Wrong answer ■ Right answer

And the leaks

By the end, no staff-only passage reaches a customer's prompt. Each episode shows which change did the work, and which ones did not.

By the end of the season, you can

- ✓ Build a complete RAG pipeline and explain every stage E1, E2
- ✓ Build a test set, including questions your documents cannot answer E3
- ✓ Choose chunking, search and reranking by testing, not by default E4 to E7
- ✓ Keep staff-only and outdated documents out of the prompt E8
- ✓ Make every claim cite its source, then check the citation in code E9 to E11
- ✓ Trace a wrong answer to the stage that lost the passage E14
- ✓ Decide when RAG is the wrong tool for the job E13, E15

Every number, rerunnable

The library, the questions, the pipeline and every result are in one public repository.

```
# rebuild an episode's analysis and charts from
# the committed results: no key, no spend
$ uv run lab rag analyse episodes/<ep>/config.yaml

# count calls and tokens before a paid run
$ uv run lab rag estimate episodes/<ep>/config.yaml
```

Answers	gpt-6-luna , reasoning off
Judge	gemini-3.6-flash , from another provider, so the system never grades itself
Embeddings	text-embedding-3-small

E2 and E3 ran on larger models (gpt-6.1-sol, claude-sonnet-5-5). From E4 the season switched to cheaper ones, for cost, and reran the baseline under them.

The season's method

Build the simplest complete system. Measure it. Change one stage at a time, and keep only what clears a bar set in advance.

Next: S2 E1 Why retrieval exists

What retrieval fixes, what it does not, and the new failure it introduces.

Where this fits

The series map

S1 LLMs	S2 Retrieval	S3 Strategy
S4 Adapting	S5 Agents	S6 Economics
S7 Data	S8 Evaluation	S9 AI teams
S10 Reliability	S11 Security	S12 Delivery
S13 Performance	S14 Resilience	S15 Governance

☐ Leadership season

 **Follow**

linkedin.com/in/brendan-kelly-irl

 **Episode guide**

brendankellyai.github.io

 **Code**

github.com/BrendanKellyAI/applied-ai-lab