

Season 2: Retrieval and RAG

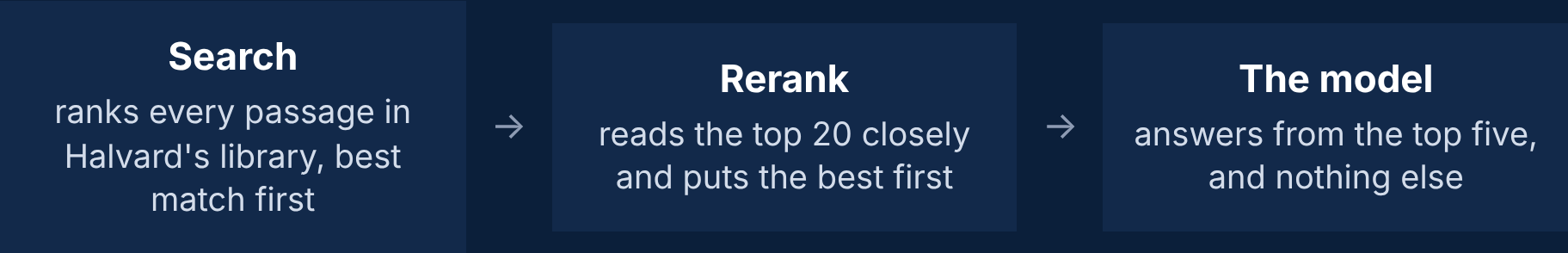
Filters.

Search ranks by meaning. It does not know who is asking, or which version of a policy is in force. A filter on two labels fixes both. Tested against the 520 test questions.

Search showed a customer a staff-only document

1 Load	2 Chunk	3 Embed	4 Rewrite	5 Search	6 Filter	7 Rerank	8 Prompt	9 Answer	10 Cite	11 Score
-----------	------------	------------	--------------	-------------	-------------	-------------	-------------	-------------	------------	-------------

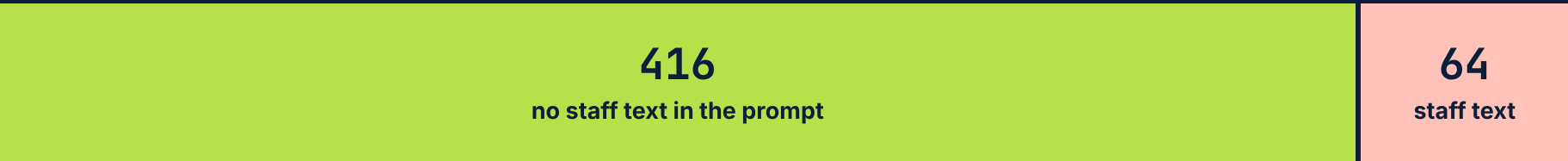
HOW OUR SYSTEM ANSWERS A QUESTION, SO FAR



A REAL TEST QUESTION, FROM A PUBLIC CUSTOMER

How much can engineers spend on parts for the Mira robot vacuum without asking for approval? Only a staff document holds the answer

Search ranked a page of the staff service manual **1st**. The model answered from it: "up to £163.00 in total". Across the 480 public questions:



Safe

A staff-only passage reached the prompt

This episode

Can a filter on two labels keep out staff text and old policy versions, without losing right answers?

Where stage 6 sits

Stage 6 runs live, on every question, between search and the reranker.



Until now	Stage 6 passed everything through. Any document could be found.
What changes	Chunks are removed before ranking, based on two labels on each document, never its text.
Security	This is access control . Without it, any customer can pull internal limits, pricing rules and known faults out of the system just by asking.
Why not let the model sort it out? In E2, the model answered from a staff passage 31 times out of 34. Whatever reaches the prompt gets used. The safe place to stop it is before the prompt. This is a security control, not only a quality fix.	

Two labels on every document

Each of Halvard's 344 documents carries two labels, set when it is added. Neither is in the text.

LABEL 1: ACCESS

Who may see it

public 340

staff 4

↓ a public customer asks

public documents only

The 4 staff documents: the engineer service manual, the escalation playbook, the known-faults register, and the goodwill and pricing rules.

LABEL 2: STATUS

Is it still in force

current 308

old 36

↓ anyone asks

current versions only

The 36 old ones: 4 policies replaced by a new version, and 32 notices and advice pages since withdrawn.

So the pattern is
Decide who may see what, and what is in force, when a document is added. Then search only among what is left.

Filter before ranking, or after?

The same two rules can run in two places. Where they run decides what the model is left with.

BEFORE RANKING

Remove, then rank

search: best 50 allowed chunks



rerank the top 20



top 5 into the prompt

A refused chunk never takes a place. The next allowed one moves up, so the prompt always gets 5.

AFTER RANKING

Rank, then remove

search: best 50 of everything



rerank, then remove refused chunks



top 5 of what survives

Refused chunks still use up places in the ranked list. If too many are removed, the prompt can come up short.

The risk to watch

A prompt with fewer passages can lose right answers. The test runs both, and counts every prompt that comes up short.

Stage 6: Filter

WHAT CHANGES IN THIS STAGE

Before search ranks, every chunk this asker may not see, or from a document no longer in force, is taken out of the running. Search ranks only what is left.

```
E7 filter: {mode: passthrough}
```

```
E8 filter: {mode: before, access: true, status: true}
```

```
src/lab/rag/filter.py
```

```
the whole rule
```

```
def refusal(chunk, asker, config):
    """Why the rules refuse this chunk to this asker."""
    if config.access and chunk.access == "staff" \
        and asker != "staff":
        return "access"
    if config.status and chunk.status != "current":
        return "status"
    return None
```

access, status

Each rule can be switched off on its own. The test does exactly that.

How the filter stops a staff-only answer

THE PUBLIC QUESTION FROM SLIDE 2

How much can engineers spend on parts for the Mira robot vacuum without asking for approval? Answer: not for a customer to know

E7	Search ranks a page of the staff service manual first. The model quotes the internal limit, £163.00.	1st
E8	The access rule removes that page before ranking. Five current, public passages fill the top 5.	gone
Reply	"The passages do not specify a parts-spending limit." The right answer for a customer.	right

Two problems, one prompt

The E7 prompt also held two superseded policy chunks, old copies of the Servicing and Warranty policies. The status rule removed those too.

How the filter stops an out-of-date answer

A VERSION QUESTION

How much does it cost to have an engineer come out for my Keel washing machine once the warranty runs out? Answer: £61.00

Warranty, version 1

superseded, ranked 1st

"An engineer visit for the Keel washing machine after the warranty has ended costs **£119.00**."

Warranty, version 2

current, ranked 2nd

"An engineer visit for the Keel washing machine after the warranty has ended costs **£61.00**."

Both copies reached the prompt, and the model answered from the first: **£119.00**. With the status rule, only version 2 is left to rank. The answer: **£61.00**.

40 of 40, from 20

On the 40 version questions, right answers went from 20 to 40. The right passage was in the prompt every time. The old copy beside it was the problem.

The test: five ways to set up the filter

Stage 6 is the filter, the step between search and the reranker. The system ran five times, changing only how the filter was set up.

The access rule

A public customer never sees staff pages.

The status rule

Only the current version of a document is used.

THE FIVE SET-UPS

1 No filter [the reference](#)

Nothing is removed. Staff pages and old versions can reach the prompt. The system as it stood after E7.

2 Both rules, before ranking

Staff pages and old versions are removed before the reranker sees the list.

3 Both rules, after ranking

The same two rules, applied after the reranker has put the list in order.

4 Access rule only

Staff pages are removed. Old versions stay in.

5 Status rule only

Old versions are removed. Staff pages stay in.

How the five set-ups were measured

Two rounds, on Halvard's 520 test questions: 480 asked by public customers, 40 by staff.

1 Round 1: search only, all five

No model answers yet. Code checks three things for every question: is the right passage in the top 5? Did a staff passage reach a public customer's prompt? Did an old version reach any prompt?

2 Round 2: the full system, two of the five

No filter, and both rules before ranking, chosen in advance. The model answers all 520, and a judge model marks each answer.

A leak

A public customer's prompt holds at least one staff-only passage. Counted out of the 480 public questions.

The rule to join, set in advance

Leaks clearly below the reference's (the likely ranges must not overlap), and right answers not lower.

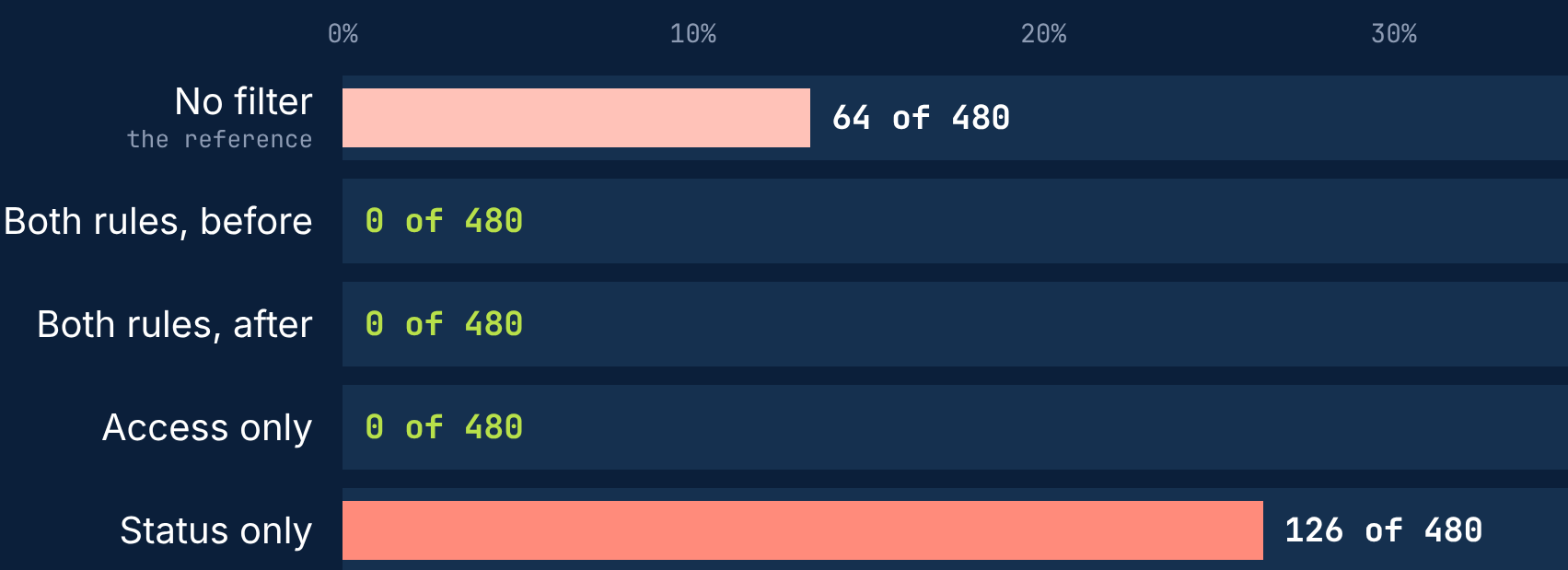
Four kinds of question to watch

Halvard's test questions come in kinds. These four show what a filter changes, and what it must not.

Staff-only 40 questions	A public customer asks something only a staff document answers. The right reply: it cannot be found. "How much can engineers spend on parts for the Mira robot vacuum without asking for approval?"
Withdrawn 20 questions	About a notice that has since been withdrawn. The right reply: it cannot be found. "What should I do if I have a wall bracket from batch 4867 for my Brio kettle?"
Versions 40 questions	The library keeps an old and a current copy of a policy. The answer must come from the current one. "How much does it cost to have an engineer come out for my Keel washing machine once the warranty runs out?"
Staff askers 40 questions	Asked by staff, who may see staff documents. The filter must not take these away. "How long does the Tessa dishwasher's built-in test take?"

No staff text reached a public prompt

Round 1: public questions whose prompt held a staff-only passage, out of 480, for each of the five filter set-ups.



Read across	Every version with the access rule had no leaks at all: 0 of 480.
The odd one out	The status rule alone leaked twice as often as no filter. Next slide: why.

The version rule alone doubled the leaks

A REAL TEST QUESTION, FROM A PUBLIC CUSTOMER

Why is my Ashby tumble dryer making a rattling, thumping or scraping noise?

NO FILTER	STATUS RULE ONLY
1 Ashby dryer: strange noises	1 Ashby dryer: strange noises
2 Ashby dryer: unusual vibration	2 Ashby dryer: unusual vibration
3 Servicing policy, old	3 Servicing policy, current
4 Safety policy, old	4 Safety policy, current
5 Safety policy, current	5 Known-faults register, staff

Two old copies left. The current Servicing policy took one place. The next passage in line was a page of the staff known-faults register.

126 of 480, from 64

Old versions had filled 761 of the 2,600 prompt places (29%). Remove them, and what moves up is often staff-only. The two rules only work together.

With the wrong passage gone, the model declines

Round 2: the 60 questions a public customer should get no answer to. Declined:

	STAFF-ONLY	WITHDRAWN	ALL 60
No filter the reference	3 of 40	0 of 20	3 of 60
Both rules before ranking	29 +26	13 +13	42 +39

Why it works	The model answers from whatever it is given. Take the wrong passage away, and it says it cannot find the answer.
The other 18	Some answer from a nearby public passage. Others say the passages do not give the answer, then add advice. The judge did not count those as declines.

The filter stopped leaks and added right answers

Round 2: both rules before ranking against no filter, the whole system, all 520 test questions.

	SEARCH	RIGHT	NOT FOUND	LEAKS
No filter the reference	316 of 400	316 of 400	41 of 120	64 of 480
Both rules, before	344 +28	359 +43	83 +42	0 -64

Question by question	Of the 400, the filter got 98 right that no filter got wrong, and 8 the other way.
Why search rose	Old copies stopped taking places in the top 5. 28 more right passages got in; none dropped out.
Ranked too low	Wrong answers caused by a passage ranked below 5th: 55 before, 34 after.

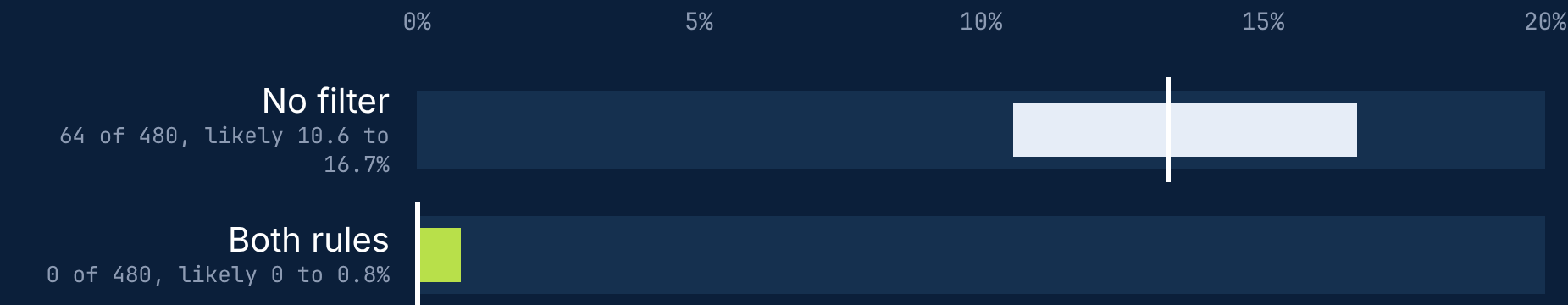
Filtering after ranking did no harm here

Expected	Filtering the ranked list would leave some prompts with fewer than 5 passages.
Found	Not one. Search returns 50, and even after removals at least 5 allowed passages were always left. Before and after gave identical results.
Not tested	A shallow search, such as the top 10. There, removing after ranking could leave real gaps.
The lesson After ranking was only safe here because the list was long. Before ranking cannot come up short at any depth, so it stays the default.	

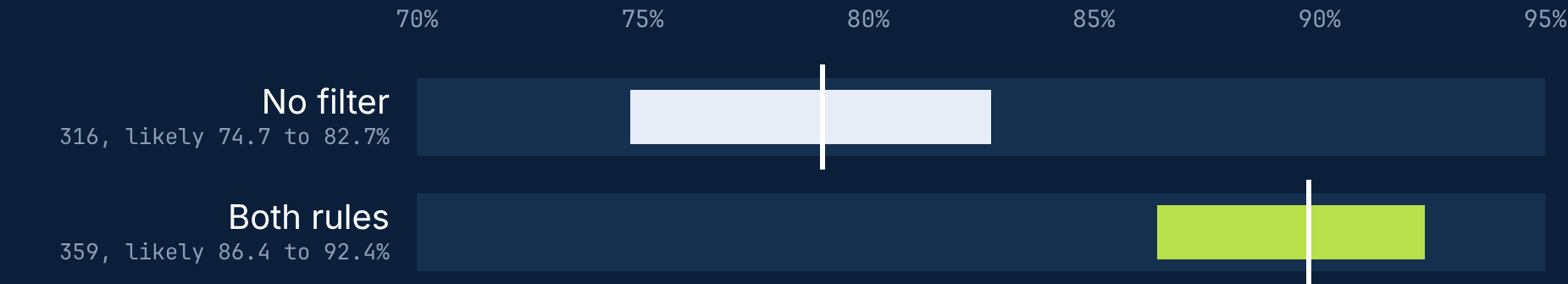
The filter joins the system by a wide margin

The rule, set in advance: leaks wholly below the old range, and right answers not lower.

LEAKS: PUBLIC PROMPTS WITH STAFF TEXT



RIGHT ANSWERS, OF 400



Not close, this time

E7 joined by 0.2 points. Here the ranges are far apart on both measures, and question by question it is 98 gained against 8 lost.

What this does not prove

- **That the labels are right.** Every rule here trusts the metadata. A staff document tagged public leaks under every filter.
- **Two access levels.** Public and staff only. Real libraries have teams, regions and customers.
- **Deletion.** A withdrawn notice stays in the index; the status rule hides it. A real deletion should also remove it.
- **The judge** is gemini-3.6-flash, checked against E3's sample.

Run it yourself

All five filter set-ups, and the two full runs.

```
# rebuild every result and chart: no key, no cost
$ uv run lab rag analyse episodes/s2-e8-filters/config.yaml

# the full run (keys needed)
$ uv run lab rag run episodes/s2-e8-filters/config.yaml
```

The filter	No model call. A check on two labels for each chunk.
Cost	\$2.22 for the episode: \$1.01 of OpenAI calls, mostly reranking for the four filter set-ups; \$1.21 of judging.

Check who can see what in your system

Here, a check on two labels stopped every leak and added 43 right answers. It worked because every document already had the labels.

Ask your team

Can a customer's question reach an internal document? Have we tested it, question by question?

Ask your team

When a policy changes or a notice is withdrawn, what happens to the old copy in our index?

Ask your team

Who sets the access label when a document is added, and who checks that it is right?

Stage 6 joins the reference system

1 Load	2 Chunk	3 Embed	4 Rewrite	5 Search	6 Filter	7 Rerank	8 Prompt	9 Answer	10 Cite	11 Score
-----------	------------	------------	--------------	-------------	-------------	-------------	-------------	-------------	------------	-------------

	SEARCH	RIGHT	NOT FOUND	LEAKS
After E7 LLM rerank, top 20	316 of 400	316 of 400	41 of 120	64 of 480
After E8 filter, before ranking	344 of 400	359 of 400	83 of 120	0 of 480

Lime: joined. Coral: tested, not joined. Every later episode starts from the bottom row.

Decision rule

Filter by permission and version before ranking, never after, and test that a restricted passage cannot reach the prompt.

Next: S2 E9 Building the prompt

The prompt now holds only passages this customer may see. How many to send, in what order, and what should the model say when none of them fit?

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