

Season 2: Retrieval and RAG

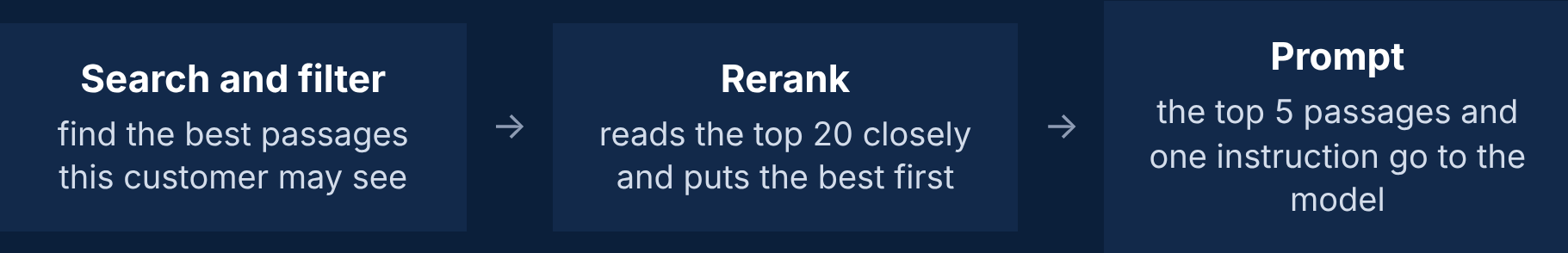
Building the prompt.

Search has found the passages. Now: how many to send, in what order, and what the model should say when none of them hold the answer. Tested against the 520 test questions.

The model answered a question the library could not

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, WITH NO ANSWER IN THE LIBRARY

How long is a refurbished part fitted to the Vale fridge freezer covered for? The library only covers replacement parts

The prompt said **replacement** parts are covered for 356 days. The model answered: a refurbished part, 356 days. Across the 120 questions with no answer to find:

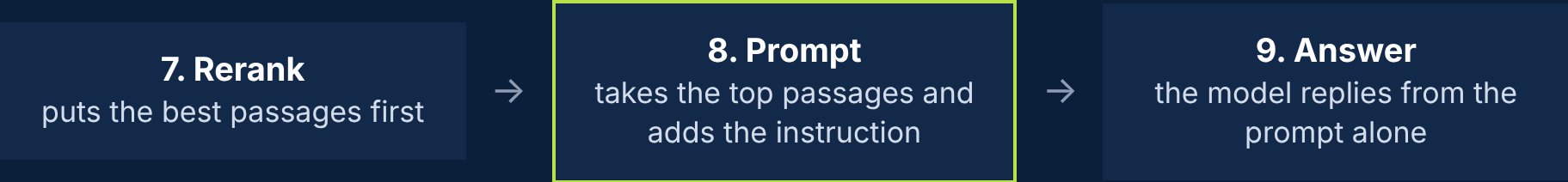
83 declined, as they should	37 answered anyway
Right	Wrong

This episode

The prompt is the last thing we control before the model answers. How many passages, in what order, and what do we tell it?

Where stage 8 sits

Stage 8 runs live, on every question. It turns the ranked passages into the text the model reads.



Until now	The top 5 passages, best first, under one plain instruction.
------------------	--

What can change	How many passages, their order, and the instruction. Search, the filter and the reranker stay as they are.
------------------------	--

Why it matters	The model sees nothing else. If the answer is not in the prompt, the model cannot use it. If the prompt says nothing about missing answers, the model fills the gap.
-----------------------	--

Stage 8: Prompt

WHAT THIS STAGE BUILDS

One instruction, then the passages, numbered best first, then the question. This is everything the model sees.

what the model received, shortened	the slide 2 question
<pre>instruction Answer the question using the passages below. Passages: [P1] Warranty: Vale fridge freezer, cover period The Vale fridge freezer is covered for 77 months from the date of delivery. ... Replacement parts fitted to the Vale fridge freezer are covered for a further 356 days. ... [P2] ... [P5] four more passages Question: How long is a refurbished part fitted to the Vale fridge freezer covered for?</pre>	

The gap	Nothing in the prompt says what to do when the passages do not hold the answer.
---------	---

Three things stage 8 decides

Each one is a single setting. Each one trades something away.

How many

k, the number of passages

More passages: more chance the answer is there, but more for the model to read, and more tokens paid for on every question.

In what order

best first, or best at the ends

Models can overlook the middle of a long prompt (S1 E7). Putting the best passages at the start and end is the usual fix.

The instruction

what the model is told

Whether the model may say "I could not find it" when the passages do not hold the answer.

The trade-off

Every passage costs tokens on every question. Fewer passages are cheaper, but only if the answer is still among them.

Best at the ends: what the reorder does

The same five passages, in two orders. Only their places in the prompt change.

RANK ORDER

Best first

1	best passage
2	2nd best
3	3rd best
4	4th best
5	5th, the weakest

The weakest passages sit at the end, where the question follows.

BEST AT THE ENDS

Weakest in the middle

1	best passage
2	3rd best
3	5th, the weakest
4	4th best
5	2nd best

The two strongest sit at the start and the end, the places a model reads most reliably.

What S1 E7 found

Position did not matter for simple lookups up to 64,000 tokens. Our prompts are about 2,800. The test checks whether the fix helps here.

The test: five ways to build the prompt

The system ran five times, changing only stage 8. Search, the filter and the reranker stayed the same.

The plain instruction

"Answer the question using the passages below."

The added sentence

"If the passages do not contain the answer, reply that you could not find it."

1 Top 5, plain the reference

The top 5 passages, best first, the plain instruction. The system after E8.

2 "Not found" allowed

The same top 5, with the added sentence.

3 Top 3

Only the top 3 passages, plain instruction.

4 Best at the ends

The top 5, reordered with the weakest in the middle.

5 Top 8 search check only

No model answers: would 8 passages catch more right passages than 5?

How the five set-ups were measured

Halvard's 520 test questions: 400 that the library can answer, and 120 it cannot.

1 Round 1: search check, all five

Code checks one thing: was the right passage among the passages sent? For top 3, that means the top 3; for top 8, the top 8.

2 Round 2: the full system, the first four

The model answers all 520, and a judge model marks each answer right, wrong, or declined.

Handled right

A new score for this episode: answerable questions answered right, plus unanswerable questions declined. Out of 520.

The rule to join, set in advance

Handled right clearly above the reference's (the likely ranges must not overlap).

Two kinds of question to watch

These two show what each change to the prompt can gain, and what it can cost.

No answer

120 questions

The library does not hold the answer: 60 never covered, 40 in staff documents, 20 about withdrawn notices. The right reply: it cannot be found.

"How long is a refurbished part fitted to the Vale fridge freezer covered for?"

Follow-ups

40 questions

A second question that only makes sense after the first. Search sees only the last line, so the right passage often ranks low in the top 5.

Earlier: "How far can I lean the Orrin heat pump to one side while carrying it?"

Then: "What about the Mira robot vacuum?"

Why these two

The new sentence is aimed at the first kind. Cutting to 3 passages is a risk for the second.

One added sentence fixed a wrong answer

THE QUESTION FROM SLIDE 2

How long is a refurbished part fitted to the Vale fridge freezer covered for? Answer: the library does not say

THE ONLY CHANGE: ONE SENTENCE ADDED TO THE INSTRUCTION

Answer the question using the passages below.
If the passages do not contain the answer, reply that you could not find it.

Before	"A refurbished part fitted to the Vale fridge freezer is covered for 356 days from the date it is fitted..."	wrong
After	"The passages do not specify how long a refurbished part fitted to the Vale fridge freezer is covered for. They state that replacement parts fitted during an authorised repair are covered for 356 days."	right

Same passages, different instruction

Told only to answer, the model stretched a figure for replacement parts to fit. Told it may say "not found", it said so, and still passed on what the library does say.

The model now says “not found” when it should

Round 2. For the 120 questions the library cannot answer, the right reply is "I could not find it".

	SAID NOT FOUND	RIGHT ANSWERS	HANDLED RIGHT
Plain the reference	83 of 120	359 of 400	442 of 520
With the sentence	105 +22	357 -2	462 +20

No answer to find	Of the 120: 25 more now say "not found", 3 no longer do.
Answerable	Of the 400: 9 gained, 11 lost. Right answers held.
How it changed	19 of the 25 had already hinted the passages lacked the answer, then gave advice. Now the model says it plainly, first.

The new sentence made the model too cautious 4 times

4 of the 400 answerable questions were declined, with the right passage in the prompt.

ONE OF THE FOUR

How long is the warranty on the Brio kettle? Two passages give two periods

Plain

"The passages give conflicting warranty periods: P1 says two years from purchase, while P2 says 47 months from delivery..."

right

Allowed

"...one says two years from purchase, while another says 47 months from delivery. I couldn't determine which period applies."

wrong

Conflicts

3 of the 4 had two passages that disagree. Both answers report both periods; the second ends "I couldn't determine", and the judge marked it a decline.

The lesson

A small cost: 4 in 400. Permission to decline needs a test for declining too often, not only too rarely.

Three passages cost 13 right answers

Round 2: the top 3 against the top 5, the whole system, all 520 test questions.

	PASSAGE SENT	RIGHT	DECLINED	PROMPT
Top 5 the reference	344 of 400	359 of 400	83 of 120	2,767 tokens
Top 3	336 -8	346 -13	76 -7	1,807 -35%

Question by question	Of the 400 answerable: 19 lost, 6 gained. A split that uneven happens by chance about 1 time in 70.
Who lost	Follow-ups: 13 of the 19. Search found their passage only 8 times in 40 within the top 3, against 15 in the top 5.

Why three passages lost this answer

A FOLLOW-UP FROM THE TEST

"What about the Mira robot vacuum?" after "How far can I lean the Orrin heat pump to one side while carrying it?"

Search sees only "What about the Mira robot vacuum?", not the earlier question about carrying it. So the passage with the answer ranks only **5th**.

TOP 5 SENT: THE REFERENCE

- 1 Mira: safe use
- 2 Mira: warranty cover
- 3 Mira warranty article
- 4 Mira: servicing
- 5 **Mira: installing and moving**, the answer

5th is sent. "Do not tilt the Mira robot vacuum more than 51 degrees when carrying it." **Right.**

TOP 3 SENT

- 1 Mira: safe use
- 2 Mira: warranty cover
- 3 Mira warranty article
- 4 not sent
- 5 not sent: **the answer**

5th is cut. "The passages don't specify how far you can tilt the Mira robot vacuum..." **Wrong.**

The lesson

Follow-ups often find their answer low in the list. Cut the list short, and those answers go first.

Moving the best passages to the ends changed nothing

Round 2: best at the ends against rank order, the whole system, all 520 test questions.

RANK ORDER, THE REFERENCE

359

right answers of 400

BEST AT THE ENDS

359

right answers of 400. Question by question: 10 gained, 10 lost.

Declines

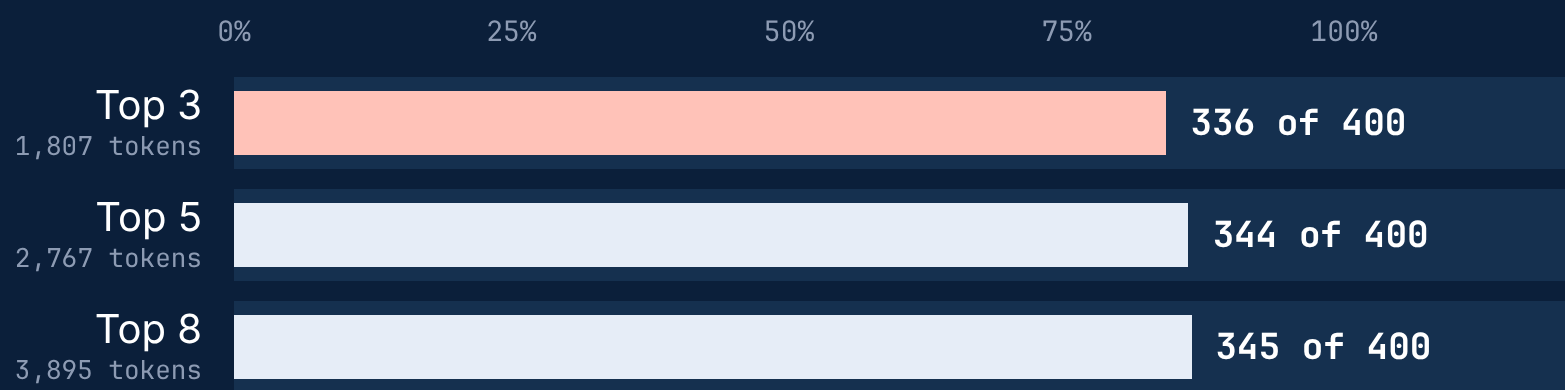
83 to 77 of 120. Question by question, 6 gained and 12 lost: too few to rule out luck.

The lesson

S1 E7 found no "lost in the middle" up to 64,000 tokens. At 2,800 tokens there is no middle to lose. Fix it only when your prompts are long.

Going past five passages finds almost nothing more

Round 1: how often the right passage was among the passages sent, for the 400 answerable questions.



Top 8 One more right passage found, for 41% more tokens on every question.

Top 3 8 fewer found, and 13 fewer right answers, for 35% fewer tokens.

So the pattern is
Here, 5 is the point where more passages stop paying for themselves. Find yours the same way.

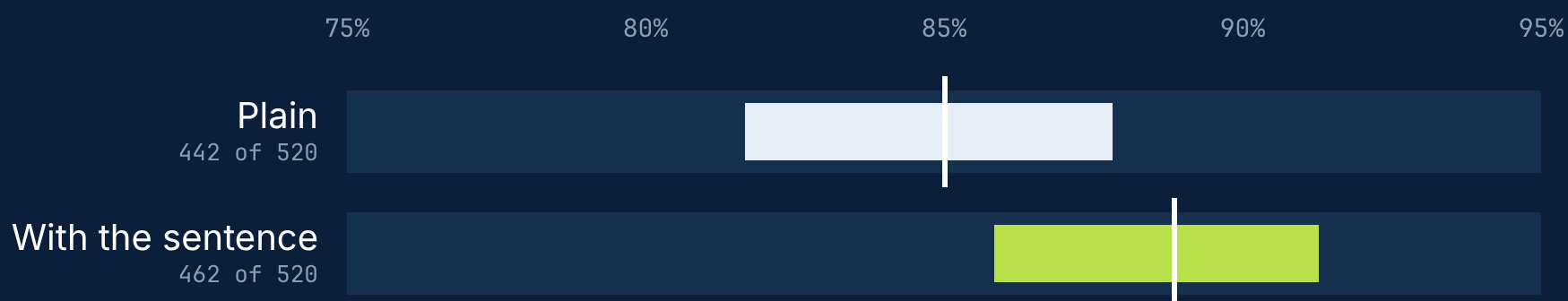
Why the best change still did not join the system

Before the run, one bar was set: a change joins only if it clearly beats the current system on all 520 questions.

Clearly beats

With 520 questions, luck moves a score by a few points. Each bar shows where the true score probably sits. The two bars must not overlap.

HANDLED RIGHT, OF 520



The bars overlap by 2 points, so it stays out. Top 3 and best at the ends scored lower, and were never close.

Why a real gain fell short

The gain came from 120 questions. The other 400 barely moved, so across all 520 it shrank inside the luck. The season recap (E16) asks whether this bar was set right.

What this does not prove

- **One wording per idea.** Another way to say "not found" might do better, or worse.
- **No combinations.** "Not found" with top 3, for example, was not tested.
- **Only up to 8 passages.** Longer prompts, where order may matter, were not tested.
- **The judge** is gemini-3.6-flash, checked against E3's sample.

Run it yourself

All five set-ups, the four full runs, and the two instruction files.

```
# rebuild every result and chart: no key, no cost
$ uv run lab rag analyse episodes/s2-e9-building-the-prompt/config.yaml

# the full run (keys needed)
$ uv run lab rag run episodes/s2-e9-building-the-prompt/config.yaml
```

The instructions	src/lab/rag/prompts/: plain.md and allow_not_found.md, one line each.
Cost	\$3.72 for the three new full runs: \$0.39 of answering, \$3.32 of judging.

Check what your system says when it does not know

Here, one added sentence raised correct declines from 83 to 105 of 120, and right answers held.

Ask your team

When the answer is not in our documents, what does our assistant say? Have we counted?

Ask your team

Does our prompt tell the model it may say "I could not find it"?

Ask your team

How many passages do we send, and has anyone measured what each one adds?

Stage 8 was tested. It did not join.

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 E9 unchanged since E8	344 of 400	359 of 400	83 of 120	0 of 480
Tested "not found" allowed	344 of 400	357 of 400	105 of 120	0 of 480

Lime: joined. Coral: tested, not joined. Stage 8 keeps its plain instruction, and every later episode starts from the top row.

Decision rule

Send the fewest passages that keep the answer score, and make "not found" an answer the model is allowed to give.

Next: S2 E10 Grounding and citations

An answer can sound sourced and still not be. How do you prove an answer came from the passage it cites?

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