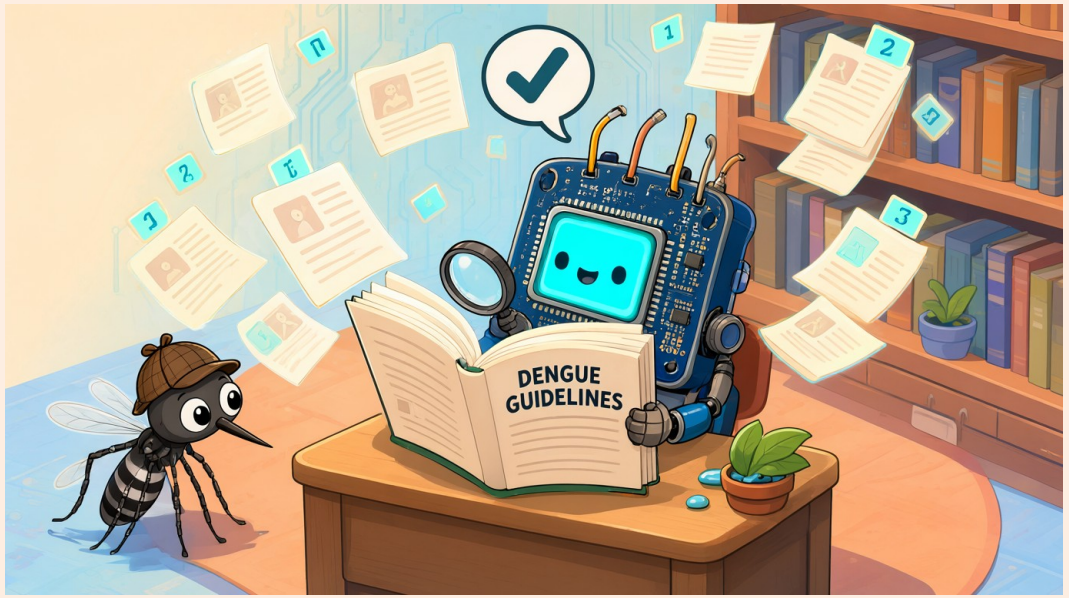
CHAPTER 6 · IESTI05 EDGE AI ENGINEERING

RAG at the Edge

Giving a small model a library



> Qwen3.5 · llama.cpp · bge-small · SQLite

Marcelo Rovai · UNIFEI

AIEng4D

THE ARC SO FAR

Each chapter gave the same small model something new

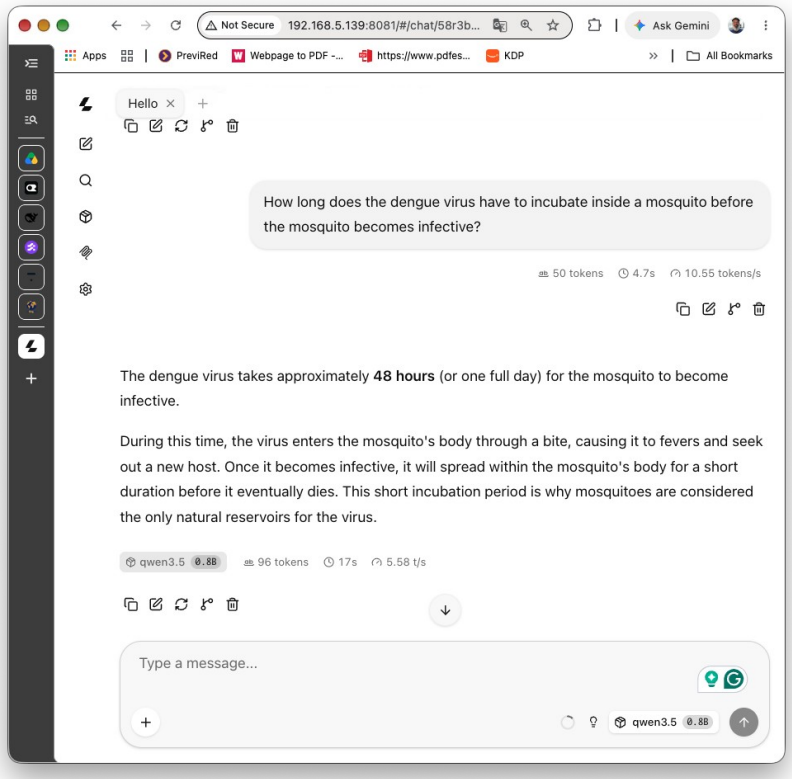
3	4	5	6
Multimodal	**Real World**	**Agentic**	**RAG**
eyes	*sensors*	*tools*	*a library*

Same Qwen3.5-0.8B. Today it learns to look things up before it answers.

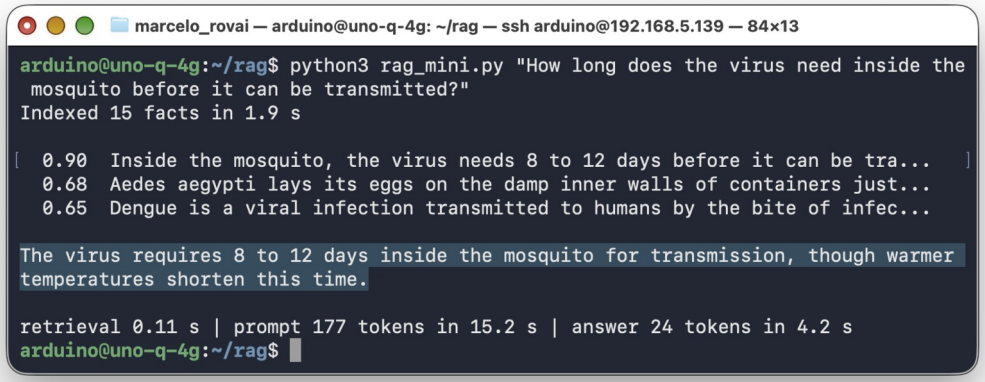
THE DESTINATION

Same model, same question, two answers

FROM MEMORY: 48 HOURS



WITH ONE RETRIEVED SENTENCE: 8 TO 12 DAYS



THE ROUTE

Seven stages, two classes

PART 1 · CLASS 27

1. **Why RAG** why RAG
2. **Embeddings** meaning as numbers
3. **Minimal RAG** 100 lines

PART 2 · CLASS 28

4. **The index** a 160-page PDF
5. **Asking** cited answers
6. **Tuning** measure first
7. **Agentic** and model size

WHAT YOU NEED

Two servers. Zero pip installs.

llama-server :8081

Qwen3.5-0.8B

~1.0 GB

reads and answers

llama-server :8082

bge-small-en-v1.5

97 MB

turns text into vectors

Python

requests

sqlite3

already on the board

No ChromaDB, no LangChain, no NumPy: the vector store is SQLite.

STAGE 1 OF 7 · Part 1 · Class 27

Why a small model needs a library

1

Three ways to add knowledge, and why RAG wins on a board.

THREE WAYS TO ADD KNOWLEDGE

Only one of them fits on the board

Bigger model

more parameters

2B already halves the speed

Fine-tuning

change the weights

not on four A53 cores; redo for every update

RAG

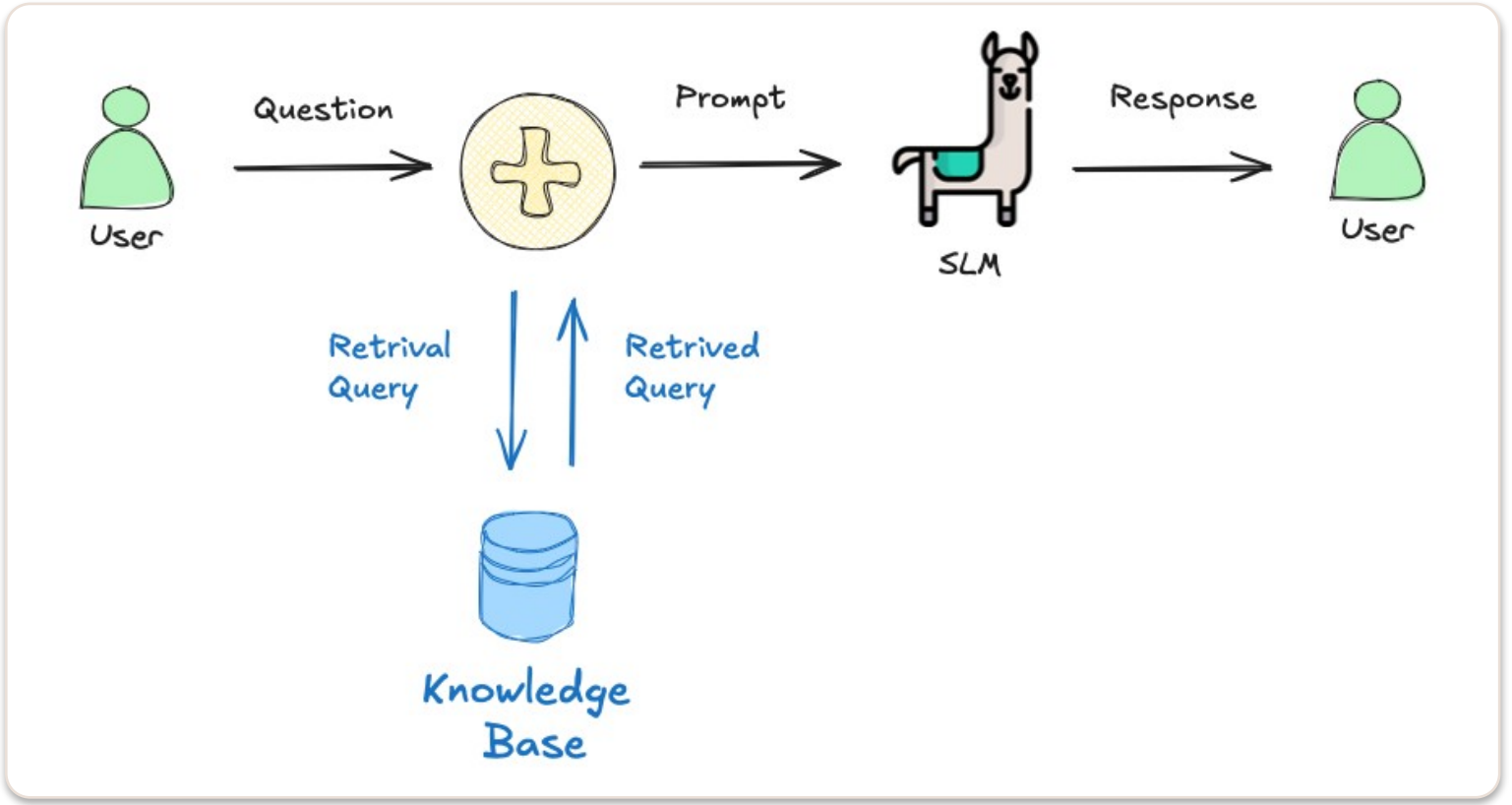
change only the prompt

~100 MB embedding model and a search

Bonus: RAG answers can cite the passages they came from.

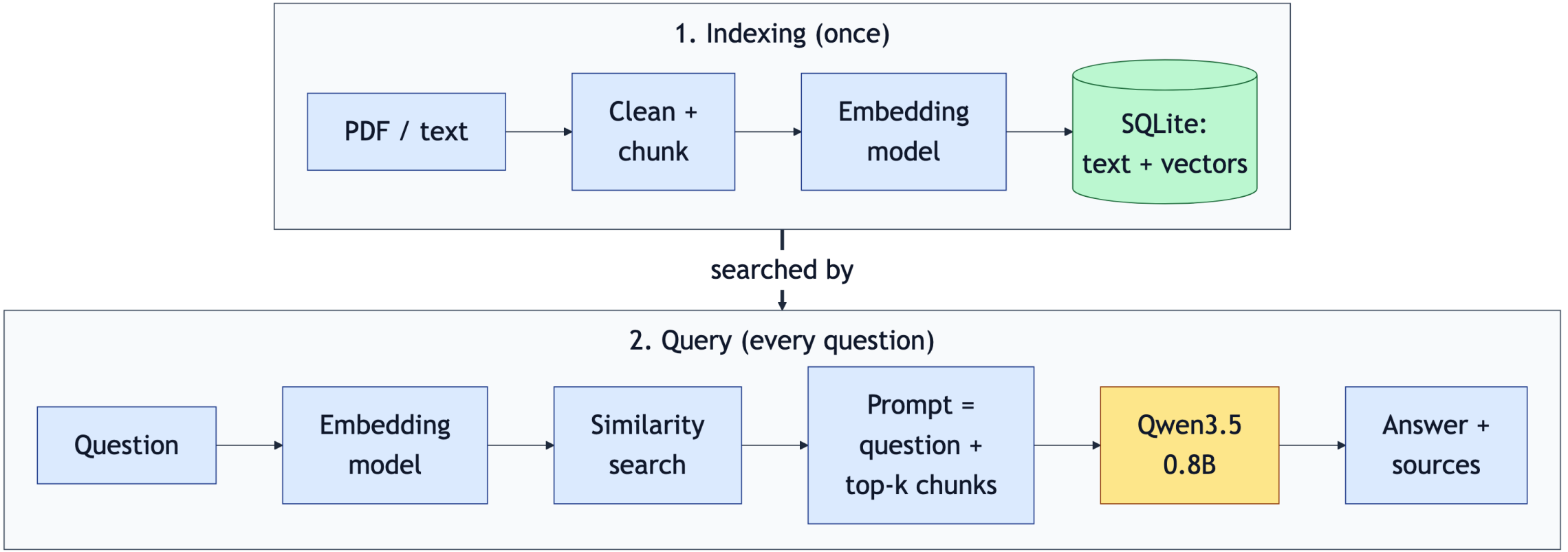
THE IDEA

Search first, then answer from what was found



HOW RAG WORKS

Index once. Query every time.



STAGE 2 OF 7 · Part 1 · Class 27

Embeddings: meaning as numbers

2

A second, tiny model on a second port.

RUN IT ON THE BOARD

Download and start the embedding server

1 · Download bge-small-en-v1.5

```
mkdir -p ~/models/embed && cd ~/models/embed
wget -c https://huggingface.co/CompendiumLabs/bge-small-en-v1.5-gguf/resolve/main/bge-small-en-v1.5-q8_0.gguf
```

2 · Start it with --embedding, on port 8082

```
~/llama.cpp/build/bin/llama-server \
  --model ~/models/embed/bge-small-en-v1.5-q8_0.gguf \
  --embedding \
  --port 8082 \
  --threads 4 \
  --ctx-size 2048 --batch-size 2048 --ubatch-size 2048 \
  --alias bge-small
```

RUN IT ON THE BOARD

Test it, then compare four sentences

3 · In a second terminal: one sentence in, 384 numbers out

```
curl -s http://127.0.0.1:8082/v1/embeddings \
  -H "Content-Type: application/json" \
  -d '{"input": ["Dengue is spread by Aedes mosquitoes."]}' \
  | python3 -c "import sys, json; d = json.load(sys.stdin); \
print(len(d['data'][0]['embedding']), d['usage'])"
# expected: 384 and the token count
```

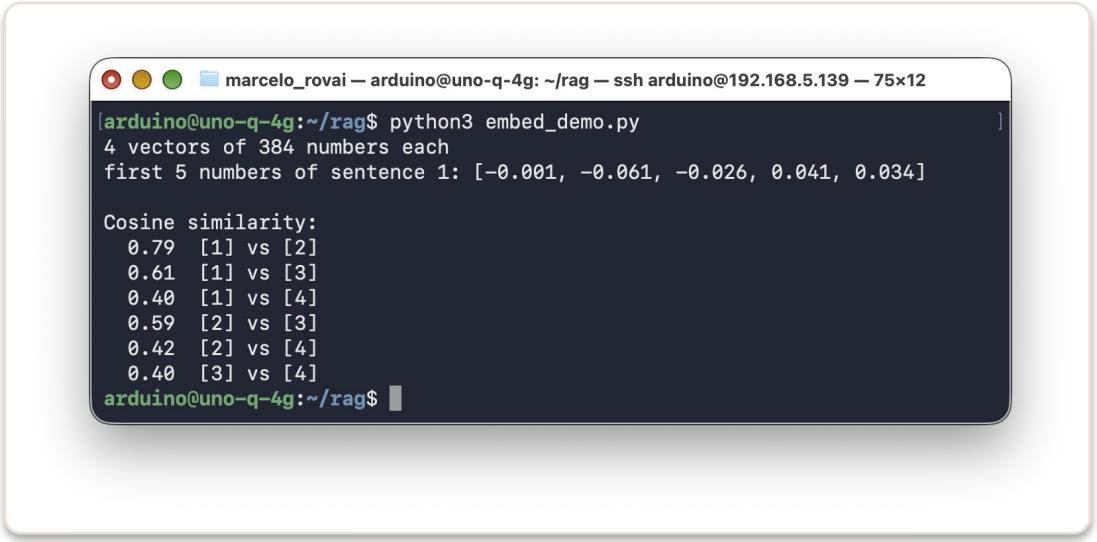
4 · The embedding demo (script in the book chapter, saved in ~/rag)

```
cd ~/rag
python3 embed_demo.py
```

FIRST CONTACT

Similar meaning, similar direction

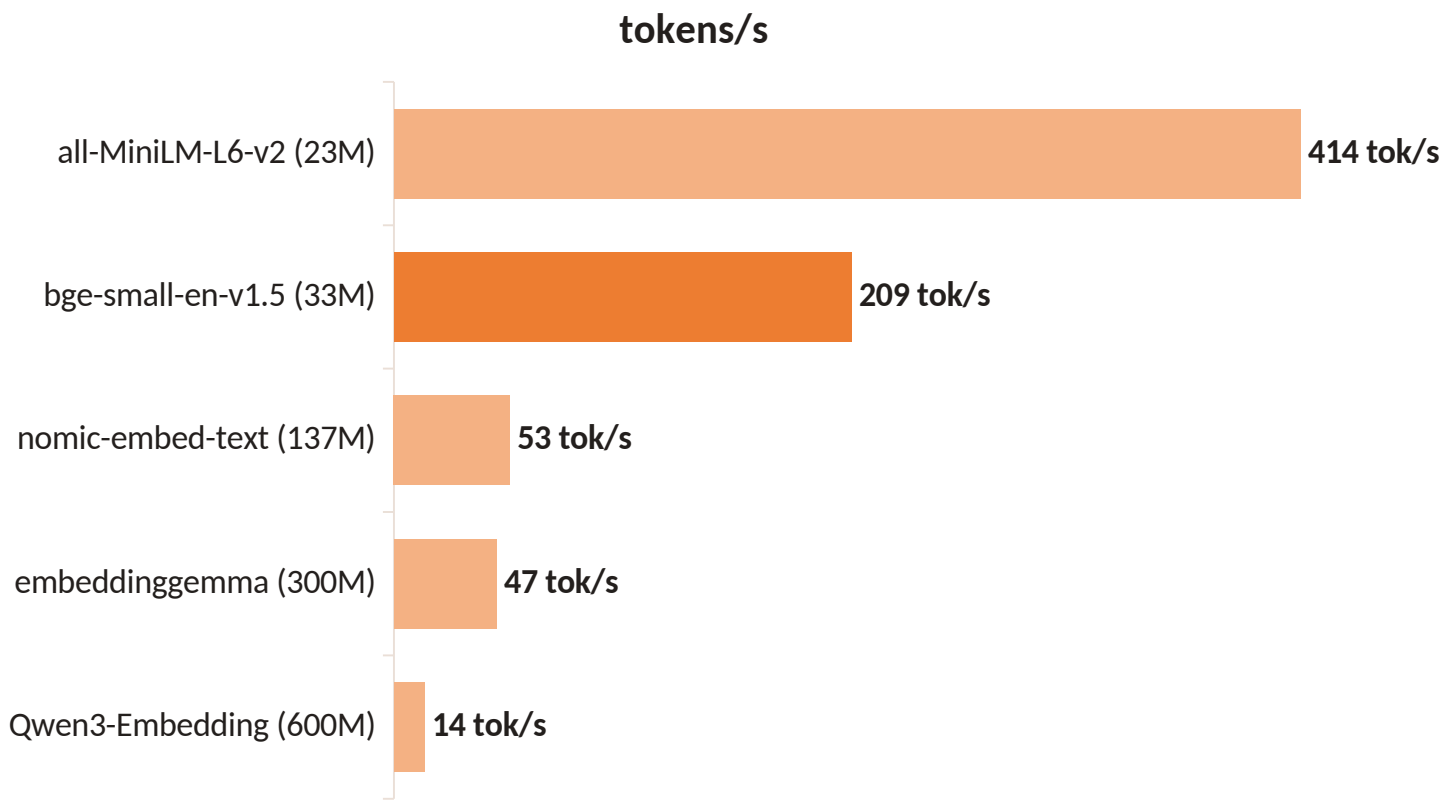
[1]: "Aedes aegypti mosquitoes lay their eggs in standing water.",
[2]: "Mosquito larvae develop in water-filled containers such as old tyres.",
[3]: "Dengue fever causes high fever and severe joint pain.",
[4]: "The Arduino UNO Q has a quad-core Cortex-A53 processor.",



384 numbers per sentence. The score is a dot product.

WHICH EMBEDDING MODEL?

Cost grows much faster than size



~6 min

to index the whole book with bge-small (our choice)

~2 hours

with Qwen3-Embedding: and 1 GB of RAM

STAGE 3 OF 7 · Part 1 · Class 27

A minimal RAG in 100 lines

3

Fifteen facts, one question, and where the time goes.

RUN IT ON THE BOARD

Both servers with one script, then the minimal RAG

1 · Stop the manual server (Ctrl+C), then start both in the background

```
# start_rag_servers.sh is in the book chapter: save it in ~/rag
cd ~/rag
chmod +x start_rag_servers.sh
./start_rag_servers.sh
# port 8081 ready
# port 8082 ready
```

2 · The same question, without RAG and with it

```
python3 rag_mini.py --no-rag "How long does the virus need inside the mosquito before it can be transmitted?"
python3 rag_mini.py "How long does the virus need inside the mosquito before it can be transmitted?"
```

RAG_MINI.PY

Without RAG, then with

--NO-RAG: RABIES AND HIV, 28.8 S

```
marcelo_rovai — arduino@uno-q-4g: ~/rag — ssh arduino@192.168.5.139 — 71×20
arduino@uno-q-4g:~/rag$ python3 rag_mini.py --no-rag "How long does the
 virus need inside the mosquito before it can be transmitted?"
Indexed 15 facts in 1.8 s

There is no single "time limit" for how long a virus needs to be inside
 a mosquito, as the duration depends entirely on **which specific virus
** is being transmitted. Different viruses have vastly different transm
ission windows:

*   **Rabies Virus**: This is an acute, highly contagious virus that ca
n be transmitted from infected animals (like dogs or cats) to humans wi
thin **24 hours**. The virus replicates rapidly in the mosquito's saliv
ary glands and saliva during feeding, making it extremely dangerous if
the bite occurs immediately after exposure.
*   **HIV**: This is a retrovirus with an extremely short transmission
window. It can be transmitted from infected individuals to others withi
n **12–36 hours** of contact (

prompt 27 tokens, answer 150 tokens in 28.8 s
arduino@uno-q-4g:~/rag$
```

RAG: 8 TO 12 DAYS, 19.4 S

```
marcelo_rovai — arduino@uno-q-4g: ~/rag — ssh arduino@192.168.5.139 — 84×13
arduino@uno-q-4g:~/rag$ python3 rag_mini.py "How long does the virus need inside the
 mosquito before it can be transmitted?"
Indexed 15 facts in 1.9 s

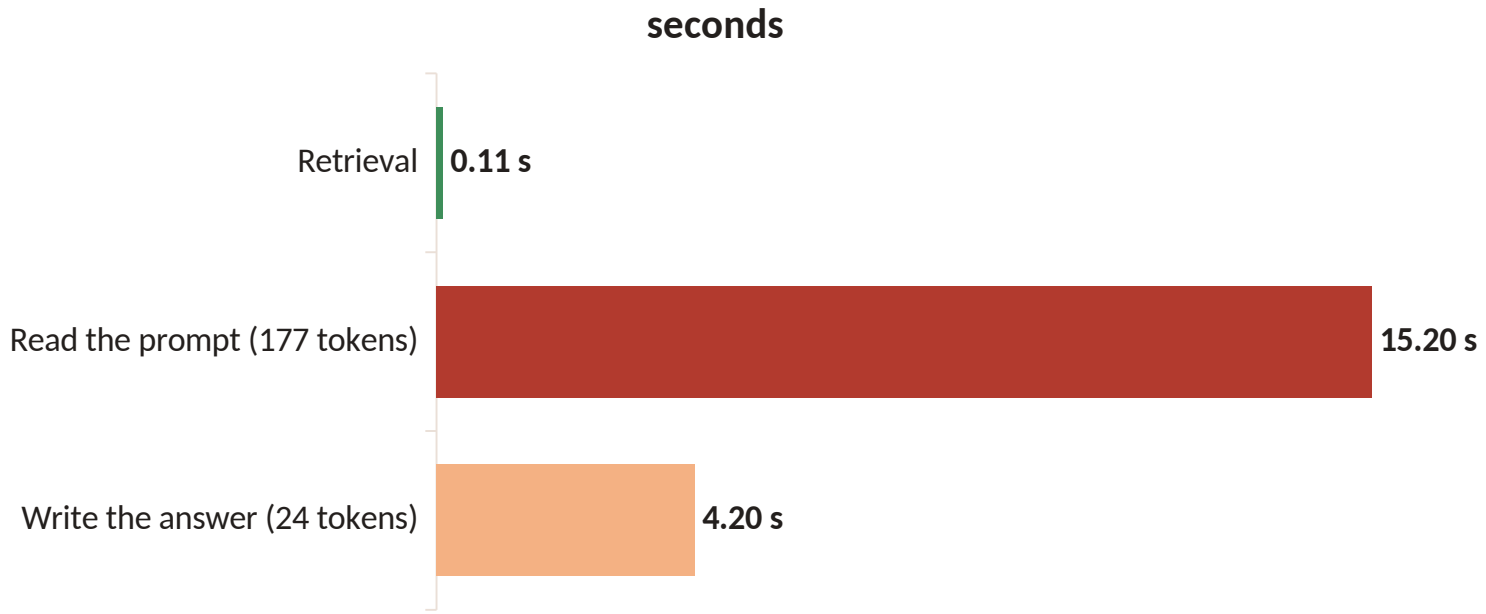
[  0.90  Inside the mosquito, the virus needs 8 to 12 days before it can be tra...  ]
  0.68  Aedes aegypti lays its eggs on the damp inner walls of containers just...
  0.65  Dengue is a viral infection transmitted to humans by the bite of infec...

The virus requires 8 to 12 days inside the mosquito for transmission, though warmer
temperatures shorten this time.

retrieval 0.11 s | prompt 177 tokens in 15.2 s | answer 24 tokens in 4.2 s
arduino@uno-q-4g:~/rag$
```

WHERE THE TIME GOES

Reading the prompt is the bottleneck



~11 tok/s

prefill on the A53: every word in the prompt costs ~0.1 s

This single number shapes every design decision in Part 2.

STAGE 4 OF 7 · Part 2 · Class 28

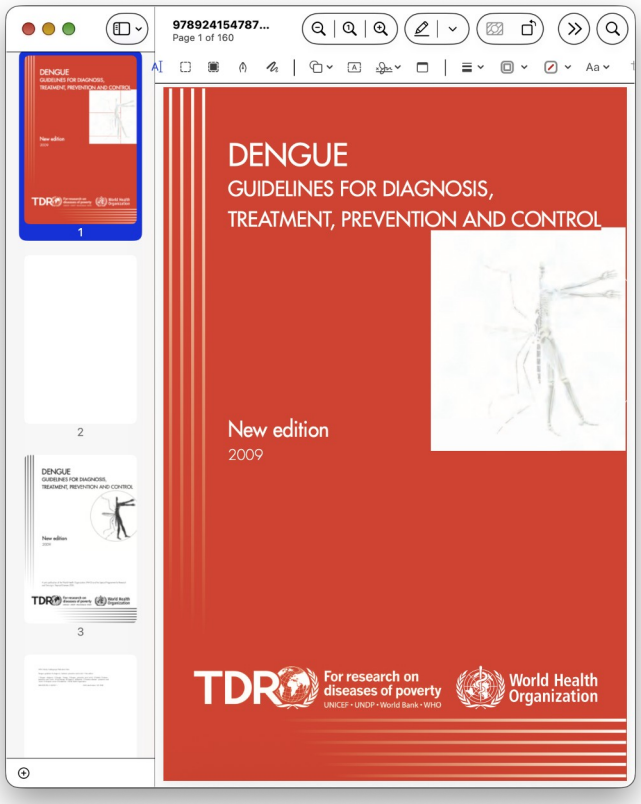
A real document

4

A 160-page WHO guideline becomes a 1.7 MB index.

THE KNOWLEDGE BASE

A real, long, technical document



160

pages: WHO dengue guidelines, 2009

54,984

words after pdftotext

10 s

to extract the text on the board

2009

parts are out of date, on purpose

RUN IT ON THE BOARD

Download, extract, and index the guidelines

1 · Download the PDF (about 1.5 MB)

```
mkdir -p ~/rag/docs && cd ~/rag/docs
wget -O who_dengue_2009.pdf "https://iris.who.int/server/api/core/bitstreams/b1db05d8-0e97-469e-b529-464dca6de540/content"
```

2 · Extract the text (sudo once, for poppler)

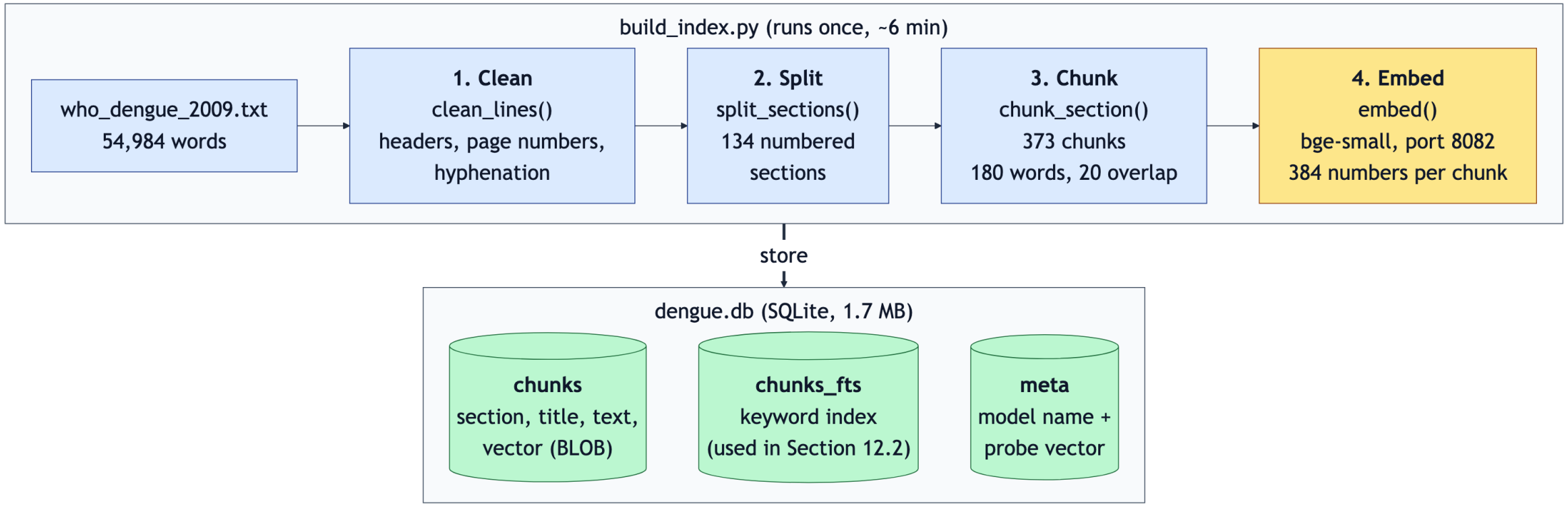
```
sudo apt install -y poppler-utils
pdftotext who_dengue_2009.pdf who_dengue_2009.txt
wc -w who_dengue_2009.txt        # 54,984 words
```

3 · Build the index (about 6 minutes)

```
cd ~/rag
python3 build_index.py docs/who_dengue_2009.txt
```

BUILD_INDEX.PY

Four stages, one SQLite file



STAGE 5 OF 7 · Part 2 · Class 28

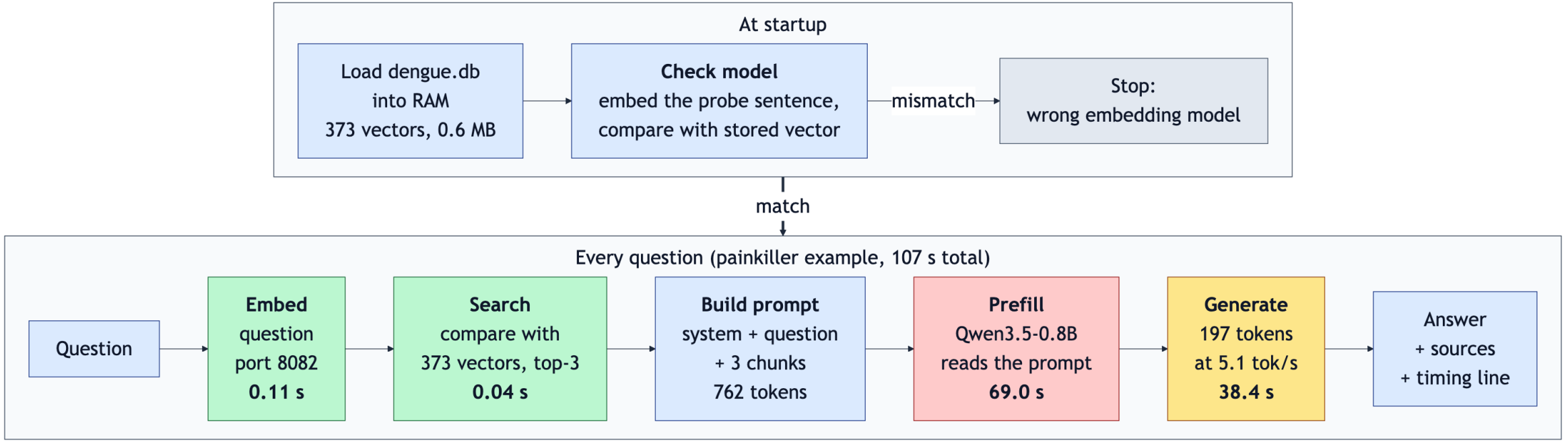
Asking the guidelines

5

Cited answers, and the first honest failure.

ASK_DENGUE.PY

The search is instant. The reading is not.



RUN IT ON THE BOARD

Ask the first question

Both servers must be running: ./start_rag_servers.sh

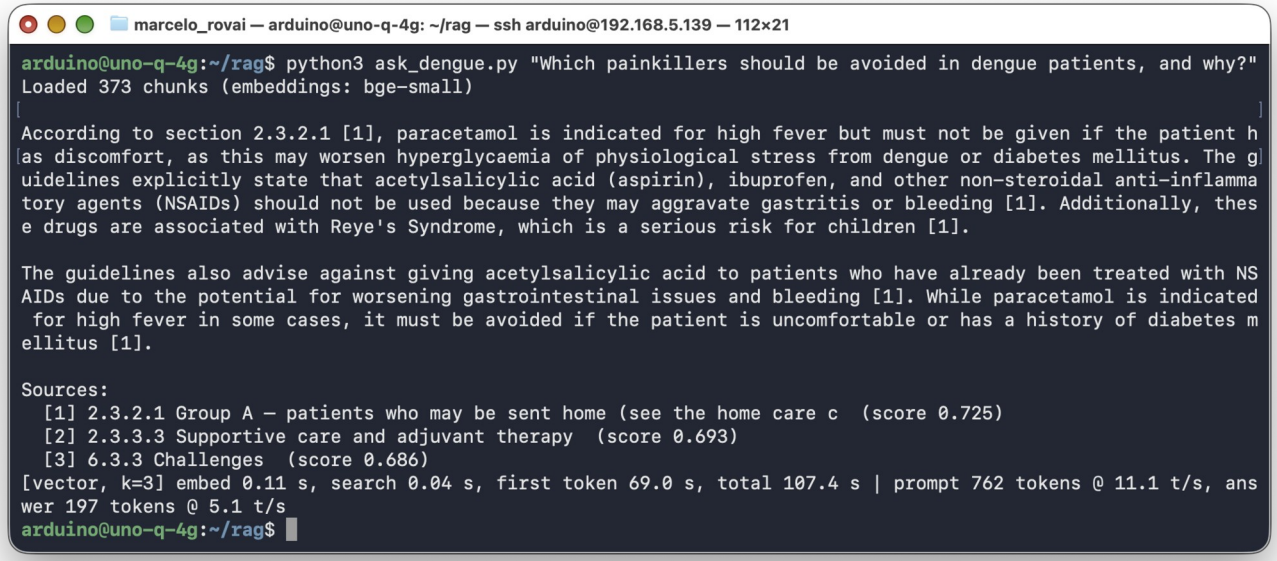
```
cd ~/rag
python3 ask_dengue.py "Which painkillers should be avoided in dengue patients, and why?"
```

The output, in three parts

```
# 1. the answer, with [1] [2] [3] citations
# 2. Sources: section number, title, and similarity score
# 3. a timing line: embed, search, first token, total, prompt tokens
```

FIRST QUESTION

Right source. Right citation. Wrong sentence.



Inverted

paracetamol: give it, not avoid it

Misplaced

the diabetes warning is about sugary drinks

Invented

a rule that is not in the text

STAGE 6 OF 7 · Part 2 · Class 28

Measure, then tune

6

Retrieval quality, chunk size, hybrid search, thresholds, reranking.

RUN IT ON THE BOARD

Measure retrieval, then latency

1 · Retrieval quality: 17 questions, no SLM call, under a minute

```
python3 eval_retrieval.py --verbose
python3 eval_retrieval.py --mode hybrid --verbose
```

2 · Latency for k = 1, 2, 3, 5, with an empty prompt cache (slow: ~6 minutes)

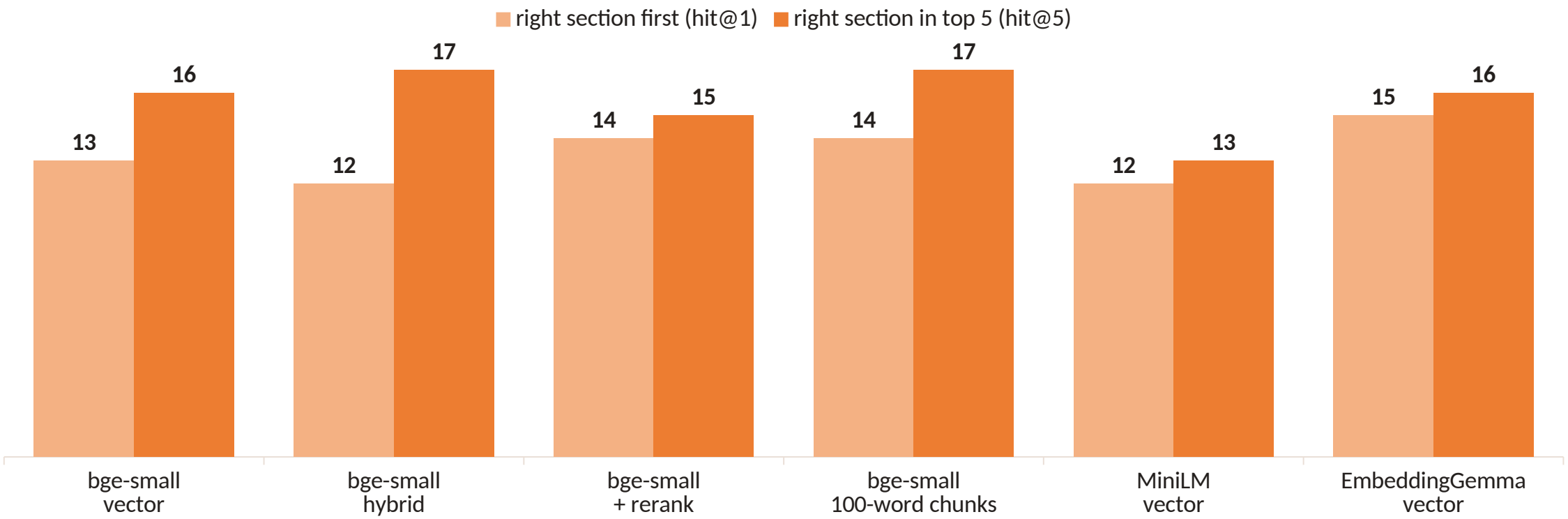
```
python3 rag_bench.py --db dengue.db --k 1 2 3 5 --cold
```

3 · A second index with 100-word chunks

```
python3 build_index.py docs/who_dengue_2009.txt --db dengue_100.db --max-words 100 --overlap 15
python3 ask_dengue.py --db dengue_100.db --mode hybrid "..."
```

MEASURING RETRIEVAL

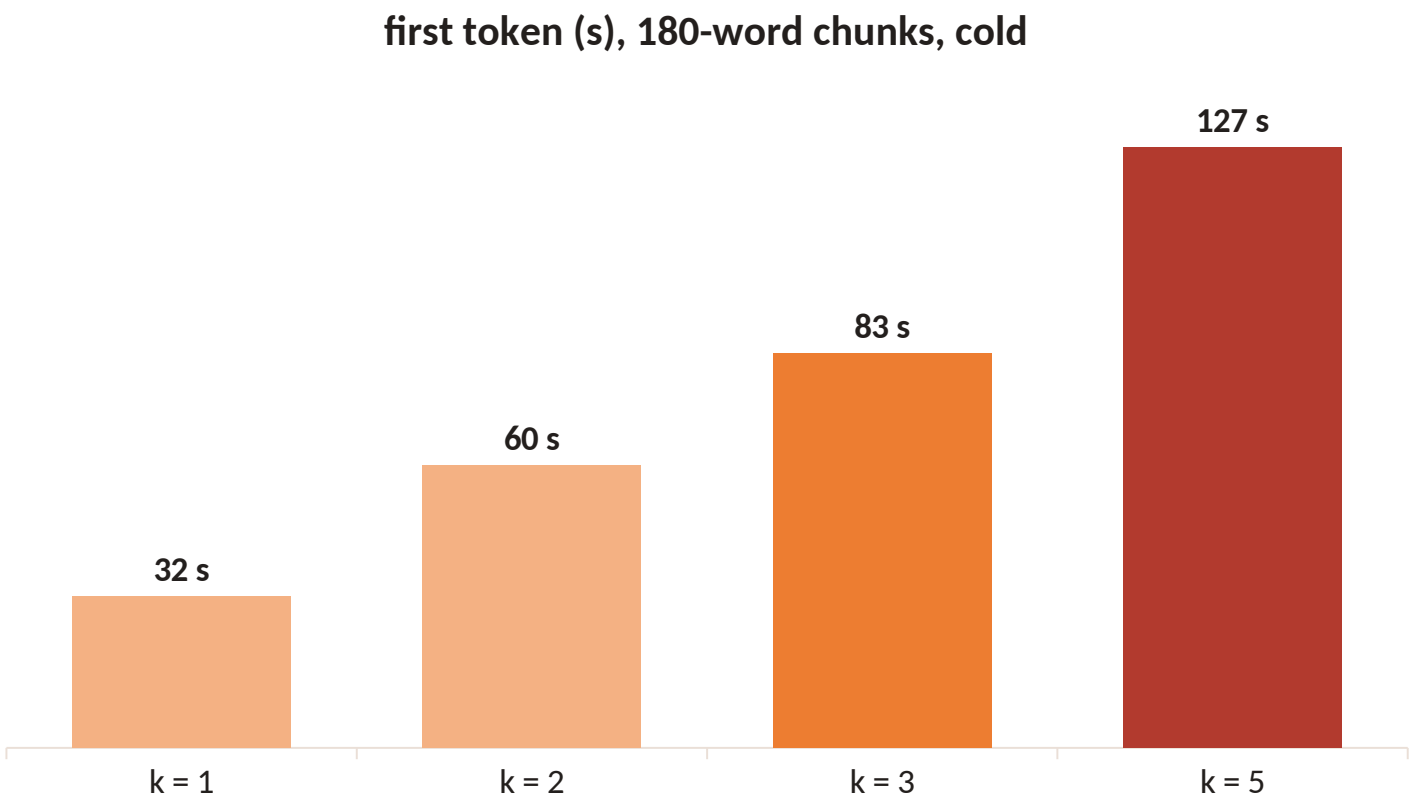
17 questions with known answers



Measure before you tune: every change is judged against this chart.

CHUNK SIZE AND TOP-K

Every extra chunk costs about 25 seconds



50 s

k = 3 with 100-word chunks, instead of 83 s

13.7 s

k = 2 when the prompt cache already holds most of it

HYBRID SEARCH

Keywords and meaning, merged by rank

Embeddings

match meaning

tires = tyres

BM25 keywords

match exact words

plasma = plasma, in any language

17 / 17

right section in the top 5, for 0.01 s extra

But a better retrieval score does not guarantee a better answer: next slide.

THE PAINKILLER QUESTION AGAIN

Better retrieval, worse answer

Setup	Answer	Time
0.8B, vector search	right source, misread	107 s
0.8B, hybrid search	"The guidelines do not cover this."	~90 s
0.8B, hybrid, k = 2	says the drugs are not listed	95 s
0.8B, + rerank	pins the diabetes warning on aspirin	94 s
2B, hybrid, temperature 0	**correct, every rule in place**	**257 s**

RUN IT ON THE BOARD

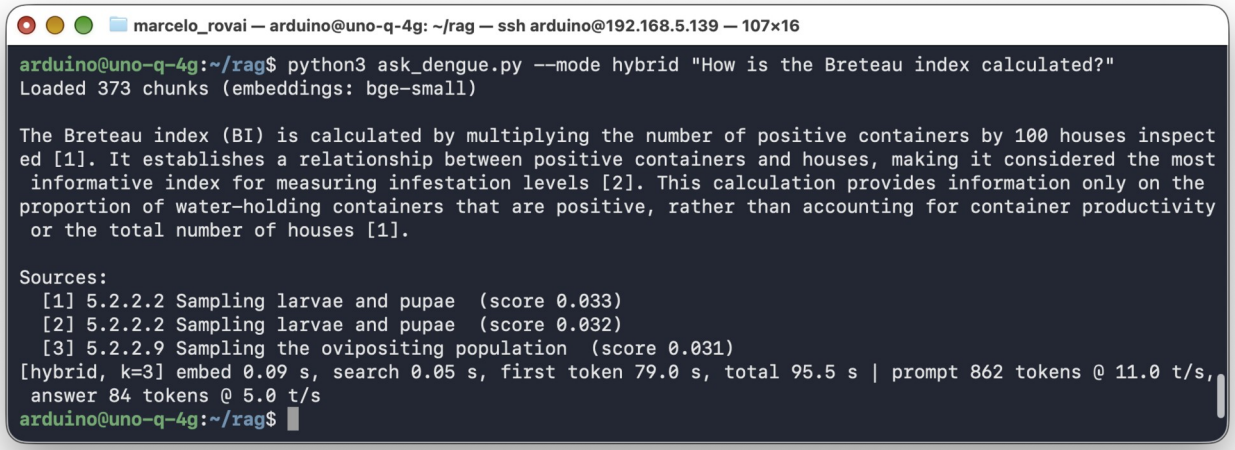
Try the limits

Questions that test extraction, thresholds, and the age of the document

```
python3 ask_dengue.py --mode hybrid "How is the Breteau index calculated?"
python3 ask_dengue.py --threshold 0.55 "Give me a recipe for sourdough bread."
python3 ask_dengue.py "How is malaria treated?"
python3 ask_dengue.py --mode hybrid "Is there a vaccine against dengue?"
# each takes one to two minutes, except the one the threshold rejects
```

WHEN EXTRACTION BREAKS

No retrieval method can fix a lost fraction bar



BI = Number of positive containers X 100 houses inspected

The PDF had a fraction:
positive containers × 100 ÷ houses inspected

Check the chunks that hold formulas, tables, and figures.

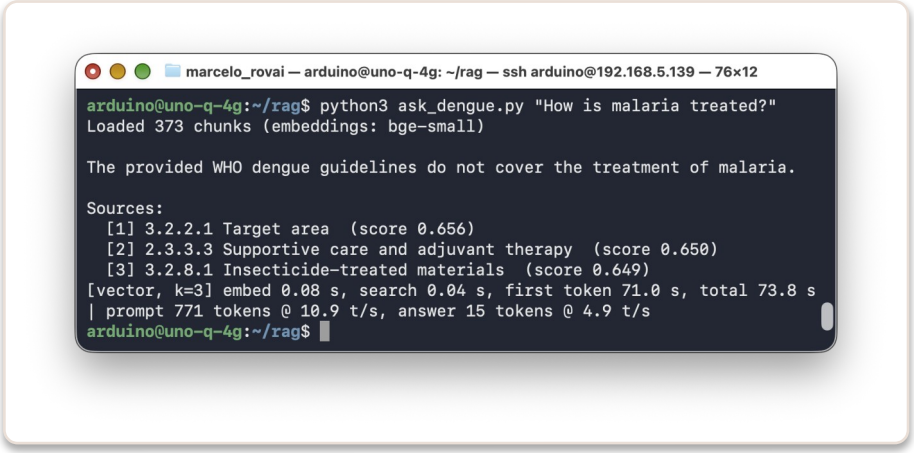
SAYING "I DON'T KNOW"

A threshold for the obvious, the prompt for near misses

SOURDOUGH BREAD: BEST SCORE BELOW 0.55, REJECTED IN UNDER 0.2 S

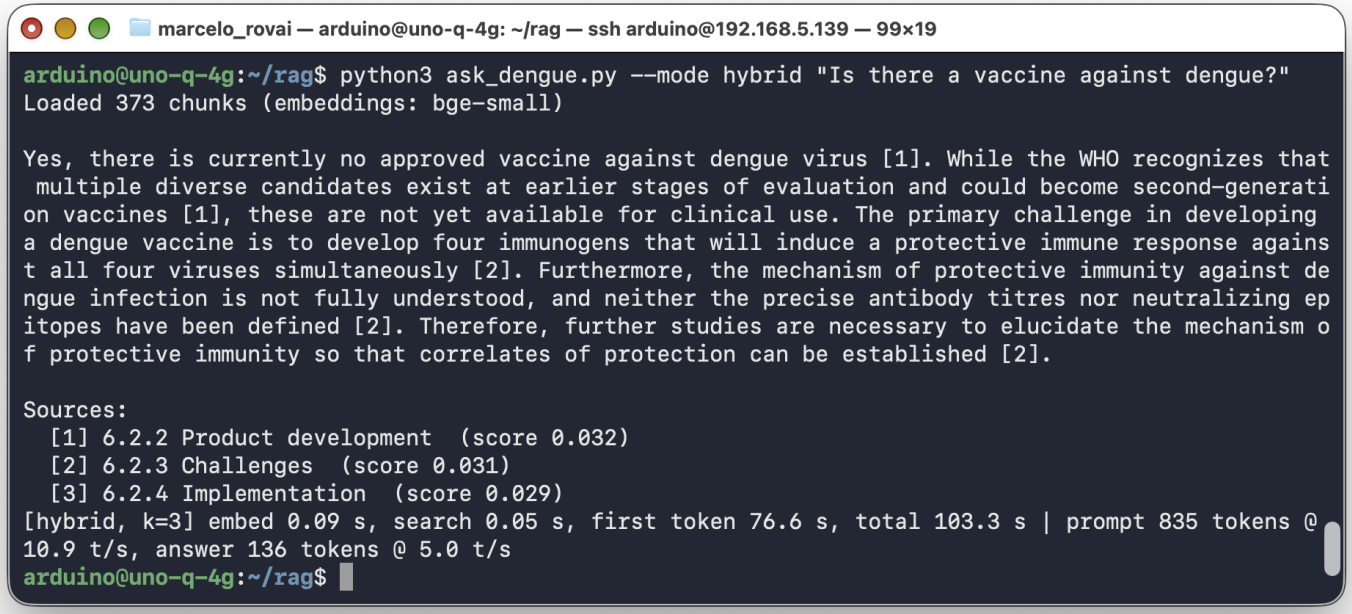


MALARIA: SCORES LIKE A REAL QUESTION; THE MODEL SAYS NO, IN 74 S



WHAT THE KNOWLEDGE BASE DOES NOT KNOW

Faithful to a 2009 document



2009

"vaccines are being developed"

today

dengue vaccines are licensed and in use

RUN IT ON THE BOARD

Add a reranker on port 8083

1 · Download jina-reranker-v1-tiny-en and start it with --reranking

```
cd ~/models/embed
wget -c https://huggingface.co/gpustack/jina-reranker-v1-tiny-en-GGUF/resolve/main/jina-reranker-v1-tiny-en-Q8_0.gguf
~/llama.cpp/build/bin/llama-server --model ~/models/embed/jina-reranker-v1-tiny-en-Q8_0.gguf \
  --reranking --port 8083 --threads 4 --ctx-size 4096 --batch-size 2048 --ubatch-size 2048 &
```

2 · Measure it, then compare one question with and without it

```
cd ~/rag
python3 eval_retrieval.py --rerank --verbose
python3 ask_dengue.py "Quais repelentes devem ser usados contra o mosquito?"
python3 ask_dengue.py --rerank "Quais repelentes devem ser usados contra o mosquito?"
```

RERANKING

A second model reads question and passage together

VECTOR SEARCH: REPELLENTS SECTION RANKED 6TH

```
marcelo_rovai — arduino@uno-q-4g: ~/rag — ssh arduino@192.168.5.139 — 100×13
arduino@uno-q-4g:~/rag$ python3 ask_dengue.py "Quais repelentes devem ser usados contra o mosquito?"
Loaded 373 chunks (embeddings: bge-small)

Os repelentes que devem ser usados contra o mosquito são os citados na seção 3.2.2.2 do contexto, es
pecificamente a tabela de larvicidas para água não potável [3].

Sources:
  [1] 3.3.2 Advocacy, social mobilization and legislation  (score 0.634)
  [2] 3.3.3 Social mobilization and communication  (score 0.621)
  [3] 3.2.2.2 Insecticides  (score 0.617)
[vector, k=3] embed 0.11 s, search 0.04 s, first token 49.1 s, total 58.0 s | prompt 542 tokens @ 11
.1 t/s, answer 45 tokens @ 5.0 t/s
arduino@uno-q-4g:~/rag$
```

+ RERANK: MOVED TO 1ST, CORRECT ANSWER

```
marcelo_rovai — arduino@uno-q-4g: ~/rag — ssh arduino@192.168.5.139 — 99×15
arduino@uno-q-4g:~/rag$ python3 ask_dengue.py --rerank "Quais repelentes devem ser usados contra o
mosquito?"
Loaded 373 chunks (embeddings: bge-small)

Os repellentes que devem ser usados são aqueles contendo DEET (N, N-diethyl-3-methylbenzamide), IR3
535 (3-[Nacetyl-N-butyl]-aminopropionic acid ethyl ester) ou Icaridin. A aplicação deve ser feita e
m locais expostos e seguir as instruções da etiqueta do produto [1].

Sources:
  [1] 3.2.6 Individual and household protection  (score 0.064)
  [2] 1.4.2 The vectors  (score 0.044)
  [3] 3.2.1 Environmental management  (score 0.039)
[vector, k=3] embed 0.11 s, search 0.04 s, rerank 5.37 s, first token 85.4 s, total 100.7 s | promp
t 929 tokens @ 10.9 t/s, answer 78 tokens @ 5.0 t/s
arduino@uno-q-4g:~/rag$
```

One rescued, two lost on our 17 questions, and 5–7 s per query. Measure first.

STAGE 7 OF 7 · Part 2 · Class 28

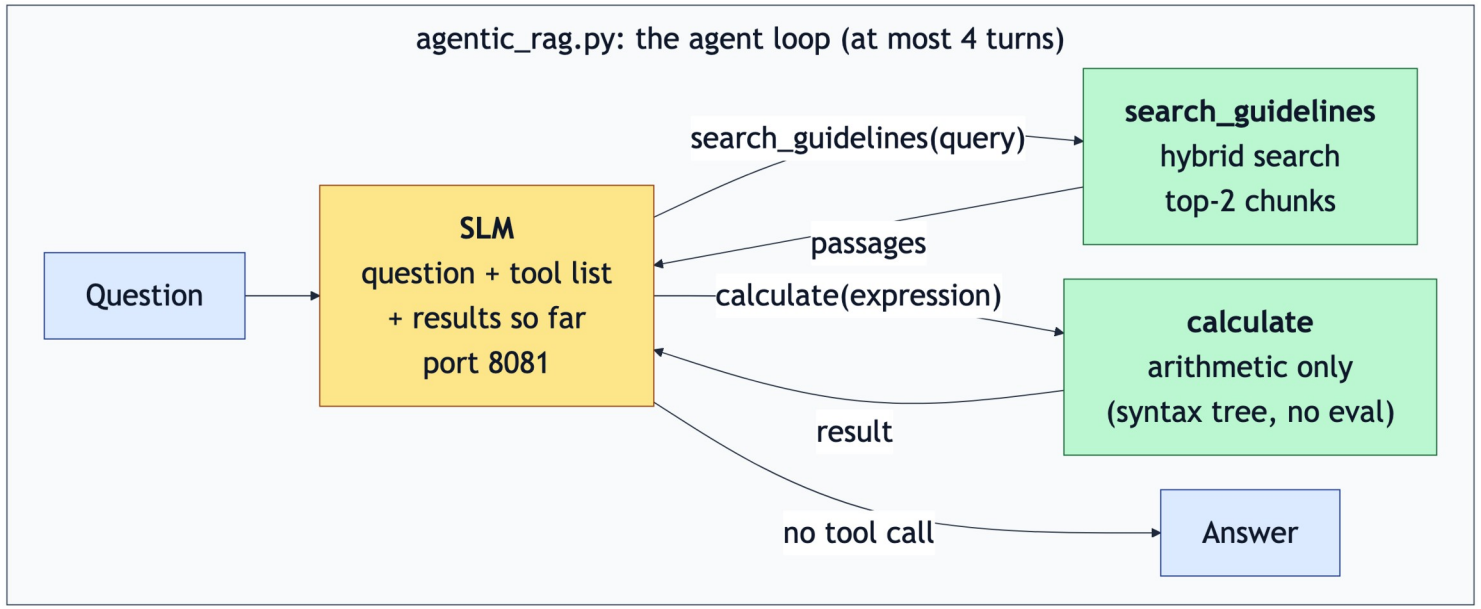
Agentic RAG and model size

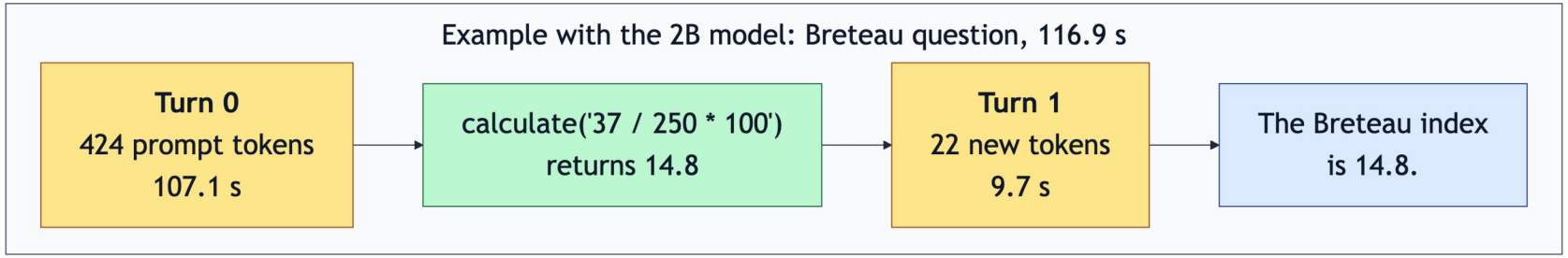
7

Letting the model decide when to search, and what a bigger model buys.

AGENTIC_RAG.PY

Retrieval becomes one tool among others





RUN IT ON THE BOARD

Run the agent, then swap in the 2B model

1 · Agentic RAG with Qwen3.5-0.8B

```
python3 agentic_rag.py "A survey found 37 positive containers in 250 houses. What is the Breteau index?"
```

2 · Restart port 8081 with the 2B model (same alias: no Python changes)

```
pkill -f "[a]lias qwen3.5-0.8b"
nohup ~/llama.cpp/build/bin/llama-server --model ~/models/Qwen3.5-2B-UD-Q4_K_XL.gguf \
  --port 8081 --host 0.0.0.0 --jinja --reasoning off --reasoning-budget 0 \
  --ctx-size 4096 --threads 4 --batch-size 512 --ubatch-size 64 \
  --predict 1024 --alias qwen3.5-0.8b > ~/rag/logs/slm2b.log 2>&1 &
```

3 · Ask the same question again (about 2 minutes, cold)

```
python3 agentic_rag.py "A survey found 37 positive containers in 250 houses. What is the Breteau index?"
```

WHO DECIDES?

The 0.8B never called a tool

0.8B: NO TOOL CALL, ANSWER FROM MEMORY

```
marcelo_rovai — arduino@uno-q-4g: ~/rag — ssh arduino@192.168.5.139 — 101×17
arduino@uno-q-4g:~/rag$ python3 agentic_rag.py "A survey found 37 positive containers in 250 houses.
What is the Breteau index?"
  [turn 0] prompt 424 new tokens, 87 generated, 53.2 s

The Breteau index is calculated as the number of positive containers per 100 houses surveyed.

Given:
*   Number of positive containers = 37
*   Total houses surveyed = 250

Calculation:
$$ \frac{37}{250} \times 100 = 14.8 $$

The Breteau index is **14.8**.

Total: 53.4 s
arduino@uno-q-4g:~/rag$
```

2B: CALCULATE() → 14.8

```
marcelo_rovai — arduino@uno-q-4g: ~/rag — ssh arduino@192.168.5.139 — 101×10
arduino@uno-q-4g:~/rag$ python3 agentic_rag.py "A survey found 37 positive containers in 250 houses.
What is the Breteau index?"
  [turn 0] prompt 4 new tokens, 37 generated, 15.8 s
  [turn 0] tool call: calculate({'expression': '37 / 250 * 100'}) -> '14.8'
  [turn 1] prompt 22 new tokens, 12 generated, 9.7 s

The Breteau index is **14.8**.

Total: 25.5 s
arduino@uno-q-4g:~/rag$
```

> Need the search to happen? Do it in code, and let the model only read.

WHAT A BIGGER MODEL BUYS

Reads correctly, costs about 3.5× the time

Qwen3.5-0.8B

~11 tok/s	reading the prompt
misreads	paracetamol, Breteau
63–70 s	per answer

Qwen3.5-2B

~4.5 tok/s	reading the prompt
correct	paracetamol, Breteau
212–216 s	per answer

RUN IT ON THE BOARD

Try a third model: LFM3-1.7B

1 · Download the official GGUF (1.1 GB, released October 27, 2026)

```
cd ~/models
wget -c https://huggingface.co/LiquidAI/LFM3-1.7B-GGUF/resolve/main/LFM3-1.7B-Q4_K_M.gguf
```

2 · Restart port 8081 with it (reasoning is on by default: turn it off)

```
pkill -f "[a]lias qwen3.5-0.8b"
nohup ~/llama.cpp/build/bin/llama-server --model ~/models/LFM3-1.7B-Q4_K_M.gguf \
  --port 8081 --host 0.0.0.0 --jinja --reasoning off --reasoning-budget 0 \
  --ctx-size 4096 --threads 4 --batch-size 512 --ubatch-size 64 \
  --predict 1024 --alias qwen3.5-0.8b > ~/rag/logs/lfm3.log 2>&1 &
```

3 · The prompt the 0.8B refused

```
python3 ask_dengue.py --mode hybrid "Which painkillers should be avoided in dengue patients, and why?"
```

A THIRD CANDIDATE

LFM3-1.7B reads well, but no faster than the 2B

Qwen3.5-0.8B

~11 tok/s

reading the prompt

refused: "The guidelines do not cover this."

~90 s

LFM3-1.7B

~4.9 tok/s

reading the prompt

correct, paracetamol in the right place

243 s

Qwen3.5-2B

~4.5 tok/s

reading the prompt

correct, every rule in its place

257 s

Same hybrid prompt and chunks, temperature 0, cold cache. Also: every answer starts with a stray </think>.

PERFORMANCE

Measured on the UNO Q, 4 GB, no fan

0.15 s
retrieval: embed + search

~11 tok/s
prefill, 0.8B

83 s
first token, k = 3, cold

1.7 MB
index for 160 pages

6 min
to build the index

97 MB
RAM for the embedding server

FIVE THINGS

To take away from this chapter

1. **Prefill is the bottleneck** retrieval is cheap; reading is not
2. **Measure retrieval** with questions whose answers you know
3. **Judge by answers** a better score can give a worse answer
4. **Check the extraction** formulas and tables break first
5. **Keep the sources visible** a RAG answer is a pointer, not a verdict

RESOURCES

Where to read further

Resource	Where
Book chapter: RAG at the Edge	mjrovai.github.io › GenAI_with_Q › RAG at the Edge
Scripts for this chapter	github.com/Mjrovai/ARDUINO-UNO-Q › RAG_Edge
WHO dengue guidelines (2009)	iris.who.int/handle/10665/44188
llama.cpp server (embeddings, reranking)	github.com/ggml-org/llama.cpp › tools/server

Questions?

Prof. Marcelo J. Rovai

rovai@unifei.edu.br

UNIFEI — Federal University of Itajubá, Brazil
AiEng4D — Academic Network Co-Chair
EdgeAIP — Academia-Industry Partnership Co-Chair



