Function Calling and Structured Outputs

We will begin 8:40 PM

Calvin and Hobbes
by WATTERSON ©1993
I WISH WE COULD STOP SUMMER, RIGHT HERE AND HAVE THE DAYS STAY JUST THE WAY THEY ARE.
THAT'S THE PROBLEM WITH LIFE. IT ROLLS ALONG WITH SPEED, YOU CAN'T CONTROL. YOU CAN'T GO FASTER OR SLOWER.
FUN EXPERIENCES ALWAYS GO ROARING BY...
...WHILE BAD EXPERIENCES NEVER PASS QUICKLY ENOUGH.
I WISH WE COULD CHOOSE HOW FAST AND SLOW EVENTS GO.
FOR EXAMPLE, I'D LIKE TO SPEED UP CHILDHOOD AND GET TO DRIVING AGE.
IT'S NOT THE PACE OF LIFE I MIND. IT'S THE SUDDEN STOP AT THE END.
distributed by universal press syndicate 7-6

Agenda

1. What is RAG, why do LLMs need RAG?
2. (Recap) What is embedding? and how does RAG use it?
3. How to implement RAG and use it for … retrieving data.

1. Create the tools scheme

2. Provide this to model

3. Create the tools

4. Hit model -> tools if wants to us

5. Backend : Execute tools -> Return response to model

6. No tools -> Model gives output

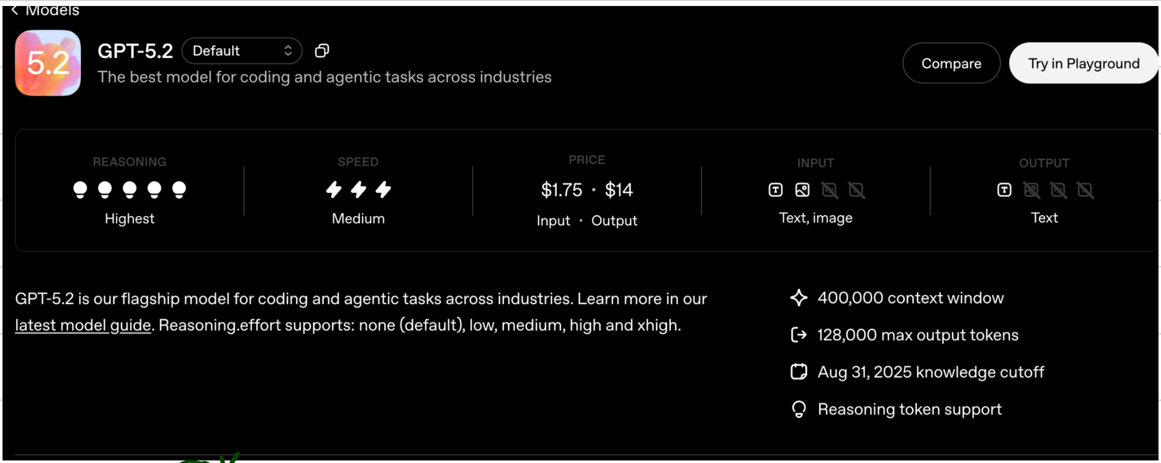
Why do we need RAG?

my AI when I ask for a source

Customer Service

5000 pdf, doc ; query : " How do I paint on transparent pixels"

" Content Window" → 2M



GPT-5.2

The best model for coding and agentic tasks across industries

REASONING	SPEED	PRICE	INPUT	OUTPUT
● ● ● ● ● Highest	⚡ ⚡ ⚡ Medium	$1.75 · $14 Input · Output	Text, image	Text

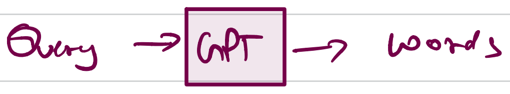
GPT-5.2 is our flagship model for coding and agentic tasks across industries. Learn more in our latest model guide. Reasoning.effort supports: none (default), low, medium, high and xhigh.

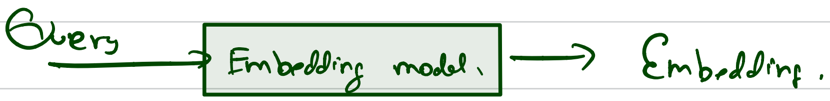
- 400,000 context window
- 128,000 max output tokens
- Aug 31, 2025 knowledge cutoff
- Reasoning token support

$\frac{50K}{4\text{ Lakh}}$ → $\frac{S1}{80}$ - ; $\frac{3\text{ Lakh}}{4\text{ Lakh}}$ → $\frac{S2}{65}$

S1 >> S2

Satisfaction score.

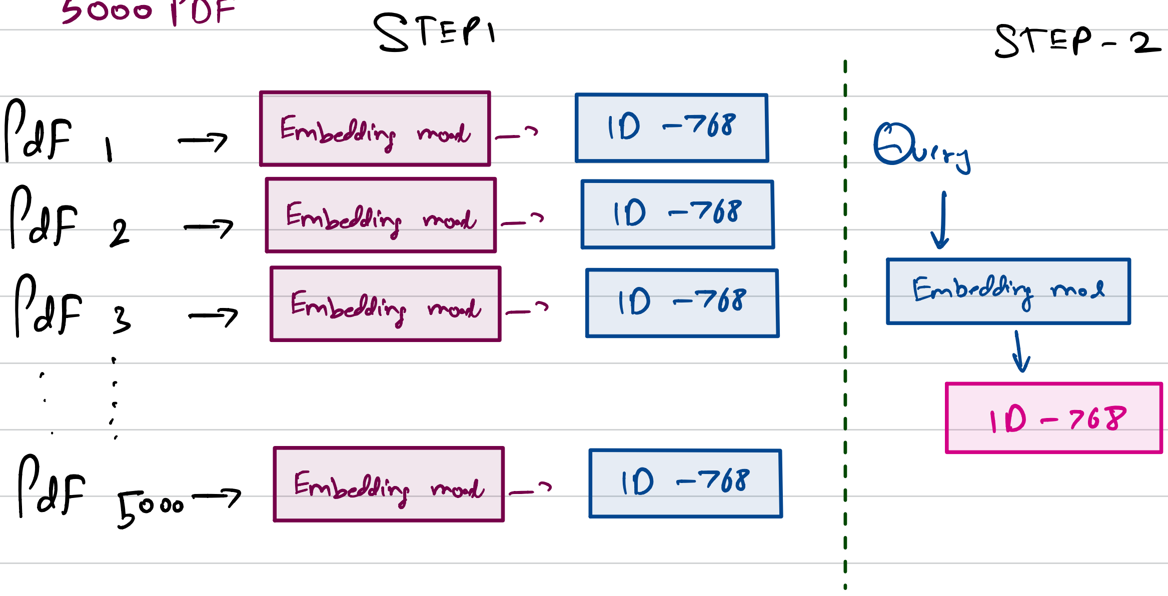




embedding-gemma → 768

1D vectors,

5000 PDF



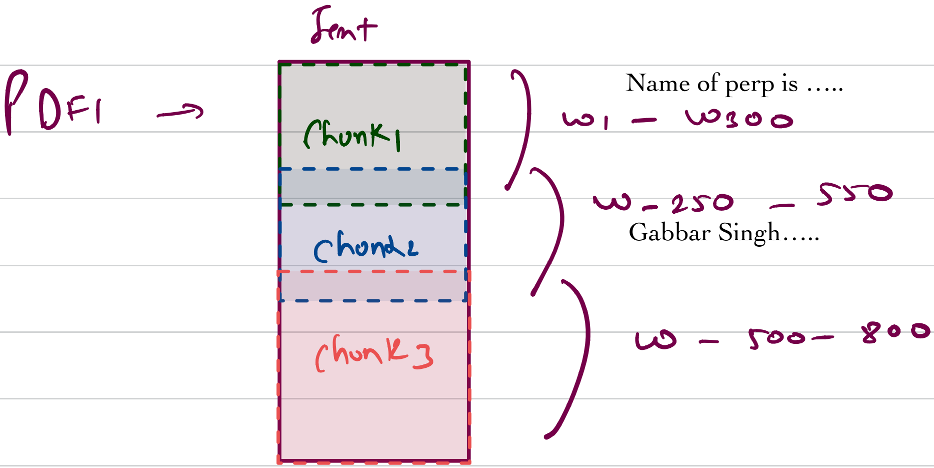
STEP-3

FAISS INDEX

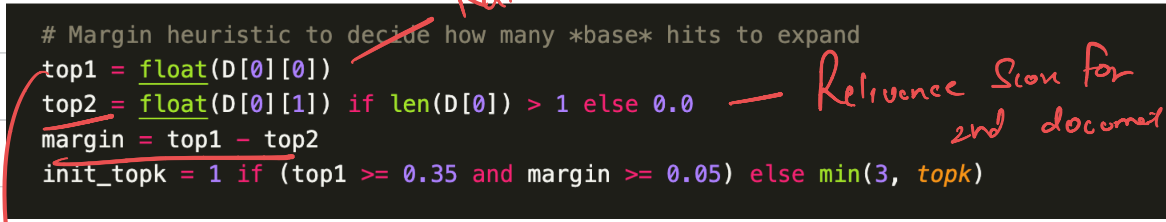
Cosine-Sim (PDF- Vector, Query)

PDF 3 → highest sim Score

↳ LLM as Content + Query → Answer.



Relevance Score.



```
# Margin heuristic to decide how many *base* hits to expand
top1 = float(D[0][0])
top2 = float(D[0][1]) if len(D[0]) > 1 else 0.0
margin = top1 - top2
init_topk = 1 if (top1 >= 0.35 and margin >= 0.05) else min(3, topk)
```

— Relivence Scon for 2nd documet

0.95
top1

0.5
top2

; Margin: top1 – top2
= 0.45