RESERVE BANK OF AUSTRALIA

RBA PubCHAT (**C**onversational **H**ub for **A**nalysis and **T**hought)
-- Unlocking Institutional Knowledge with RAG

Max Zang, Data Science Hub, Economic Research Department
Reserve Bank of Australia

Motivation

Because lots of people asked

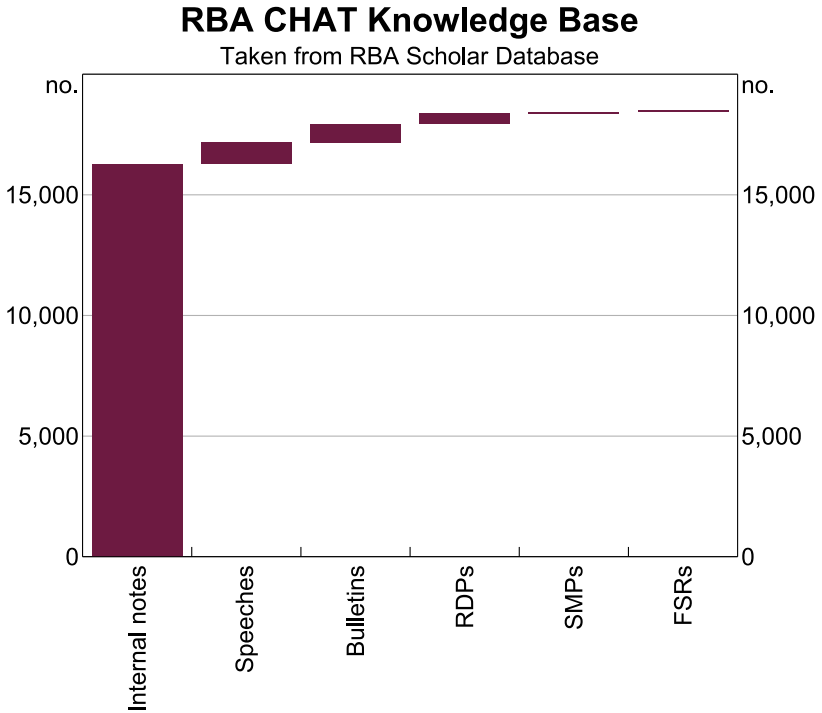
Keys into Bankwide priorities

- *Smarter, Simpler and Faster*
- AI maturity uplift

More specifically

- It's empowering
  - Puts the Staff's history of thought at people's fingertips
- It unlocks institutional knowledge
  - Surfaces valuable insights that might be buried or siloed
- It accelerates early-stage research
- It supports onboarding and knowledge transfer

The data



Built on robust knowledge management foundations

- All documents accessible from the Bank's record management system

Database automatically updates

Split documents into paragraphs (or 'chunks'). Store the chunks in a "vector database":

- 200,000 rows (paragraphs)
- 1,000 columns (numerical embedding of the paras)

Demo

- RBA Chat

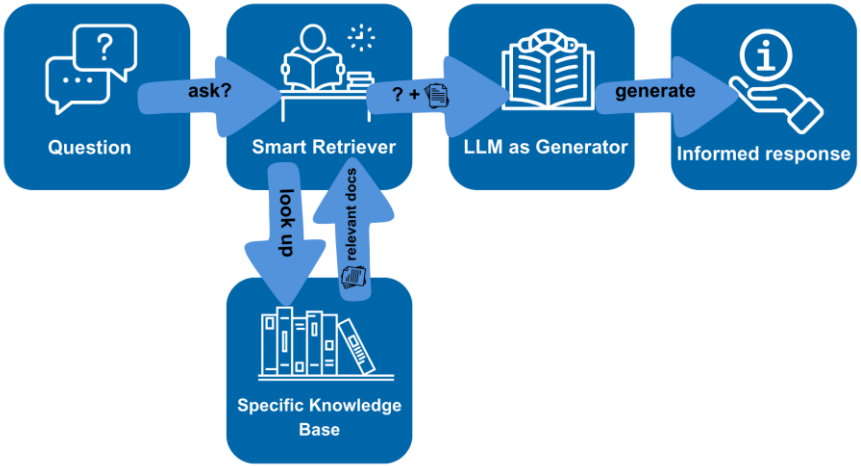
- “*Summarise the research done on the impact of changes in housing wealth on household consumption -- so called housing wealth effects. Divide your summary into (1) findings from household-level survey data; and (2) findings from aggregate time series data.* ”

Design Parameters

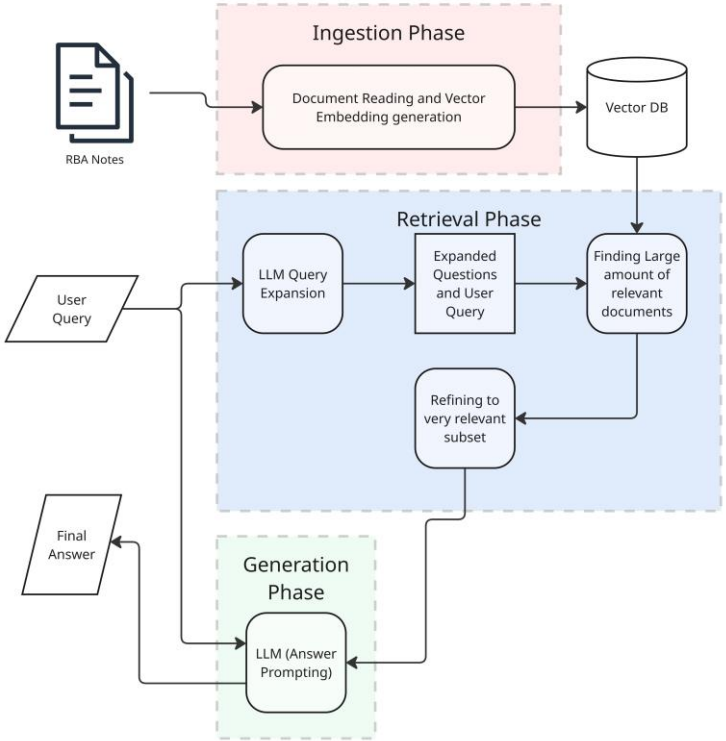
Limitations

- × No ChatGPT / other proprietary models
  - Data security concerns
  - No knowledge of internal records
  - Responses untraceable
- × Restrictive hardware
  - Single GPU (Nvidia T4)
    - 16GB VRAM

RAG – In House

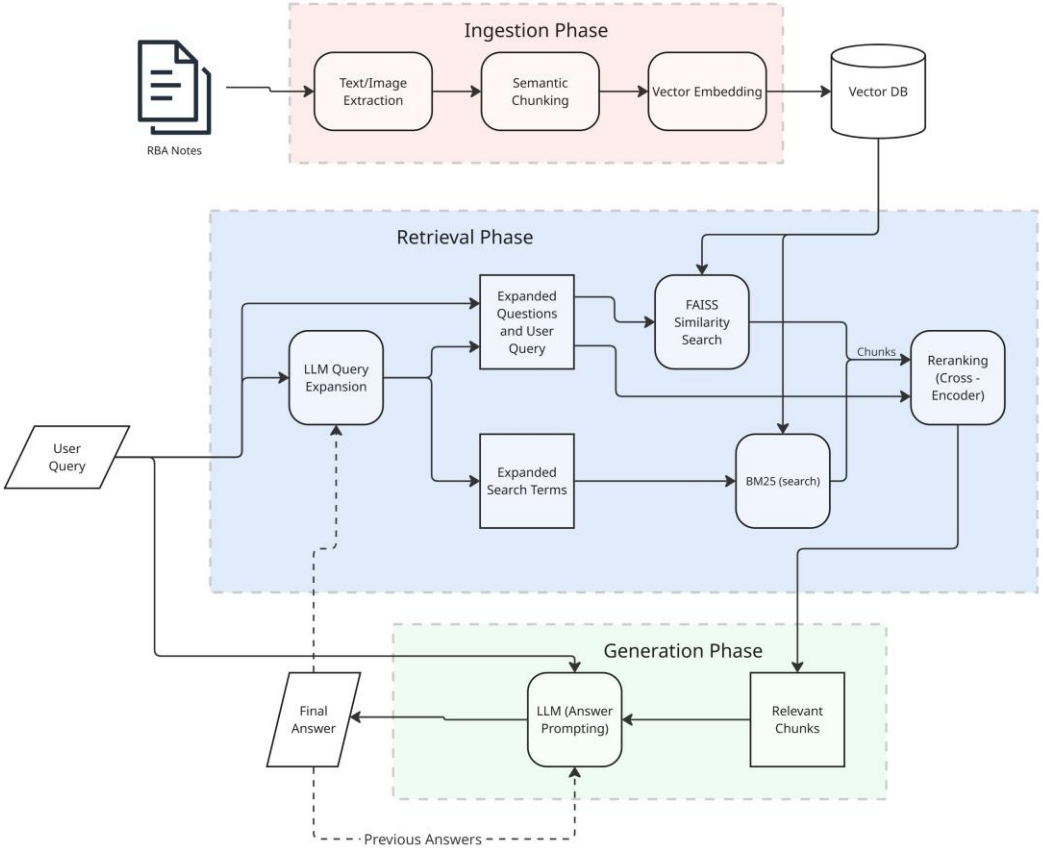


RBA PubCHAT



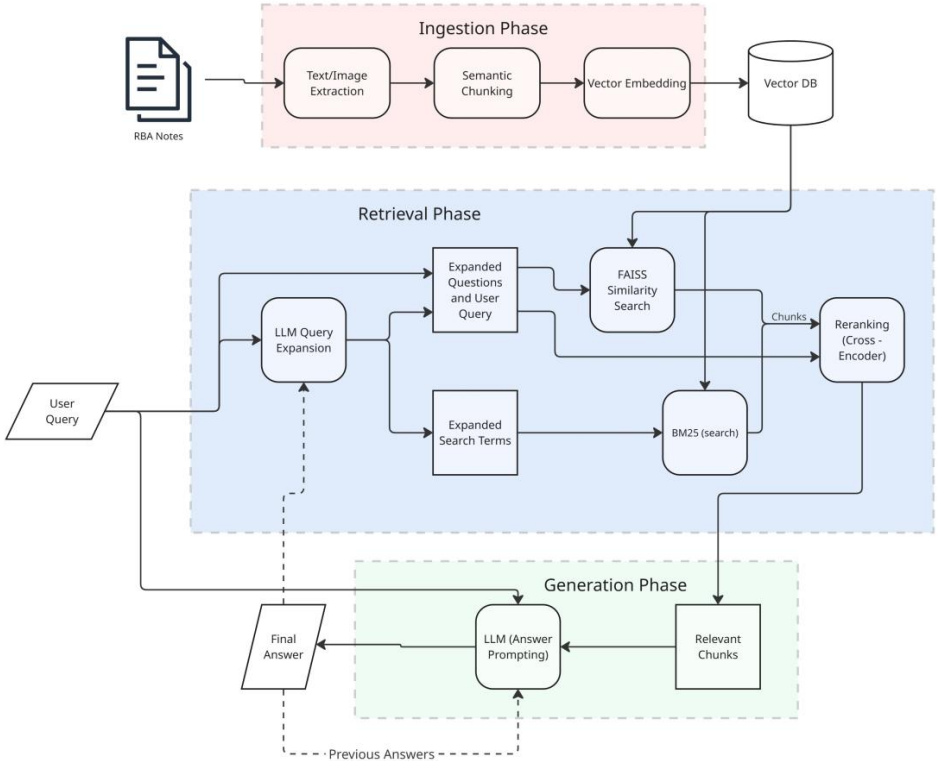
- **Ingestion:** AI models ‘encode’ documents = searchable by **meaning**
- **Retrieval:** Discover documents: high **relevance**; comprehensive **coverage**.
- **Generation:** Reasoning model: clear, well-structured answer; **precisely** and **logically** summarizes retrieved text.

RBA PubCHAT

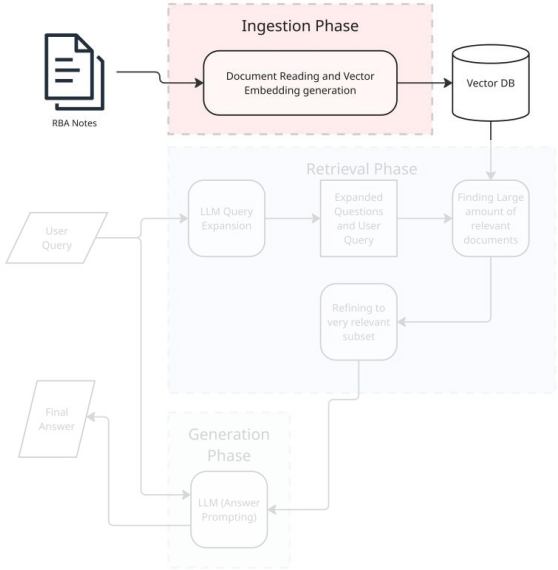

Ingestion Phase
RBA Notes
Text/Image Extraction
Semantic Chunking
Vector Embedding
Vector DB
Retrieval Phase
Expanded Questions and User Query
FAISS Similarity Search
LLM Query Expansion
Chunks
Reranking (Cross - Encoder)
User Query
Expanded Search Terms
BM25 (search)
Generation Phase
Final Answer
LLM (Answer Prompting)
Relevant Chunks
Previous Answers

Key Techniques

- Semantic Chunking
  - "all-mpnet-base-v2"
- Embedding Model
  - "Qwen3-Embedding-0.6B"
- Query Expansion
  - “Qwen3-14B-Q4”
- Re-ranking
  - “ms-marco-MiniLM-L6-v2”
- Reasoning Model
  - “Qwen3-14B-Q4”



Document Embedding



Semantic Chunking

- Split text based on semantic similarity
- Implemented with Langchain
  - text_splitter.SemanticChunker()

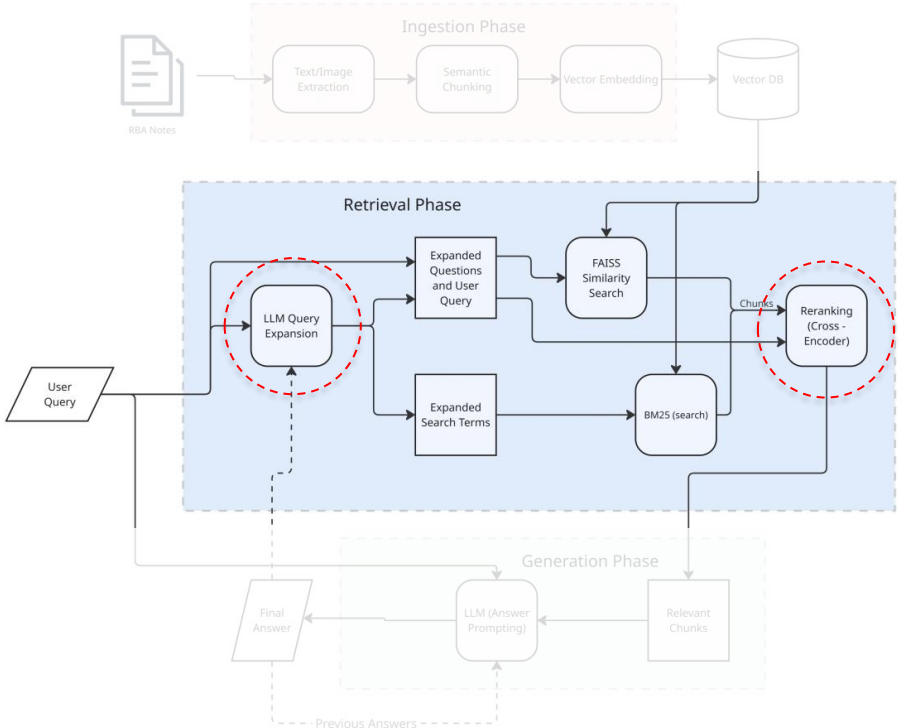
- ✓ Meaningful units of information
- ✓ Aligns with human understanding
- ✓ Improved embedding quality
- ✓ Better retrieval

Embedding Model: Qwen3-Embedding-0.6B

Rank (Bo…	Model	Zero-shot	Memory U…	Number of P…	Embedding D…	Max Tokens	Mean (T…	Mean (TaskT…	Bitext …	Classification	Clustering	Instru
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82	54.59	5.18
2	Qwen3-Embedding-8B	99%	28866	7B	4096	32768	**70.58**	**61.69**	**80.89**	**74.00**	**57.65**	10.06
3	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33	57.15	**11.56**
4	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83	52.33	5.09
5	Linq-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24	50.60	0.94
6	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55	52.77	4.94
7	multilingual-e5-large-instruct	99%	1068	560M	1024	514	63.22	55.08	80.13	64.94	50.75	-0.40
8	embeddinggemma-300m	99%	578	307M	768	2048	61.15	54.31	64.40	60.90	51.17	5.61
9	SFR-Embedding-Mistral	96%	13563	7B	4096	32768	60.90	53.92	70.00	60.02	51.84	0.16
10	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64	47.84	4.08
11	GritLM-7B	99%	13813	7B	4096	32768	60.92	53.74	70.53	61.83	49.75	1.45

Query Expansion and Reranking

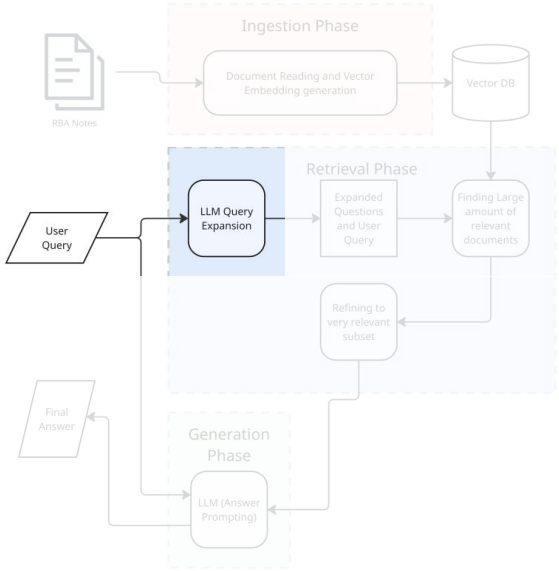
- Query Expansion:
  - completeness of retrieval
- Re-ranking:
  - Ensures high relevance



Query expansion

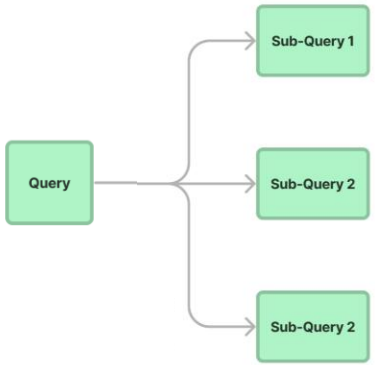
We want to extract some nuance.

- Improves the recall of the search
- Generative and creative abilities of an LLM and leverage the knowledge inherent in the model
  - LLM: **Qwen3** (our best model available)



"What has happened to cash use over time?" sub-queries:

- *What are the key **economic**, **technological**, and **social** factors that have influenced changes in cash use over time?*
- *Are there significant **regional** or **demographic** differences in the decline or persistence of cash usage?*

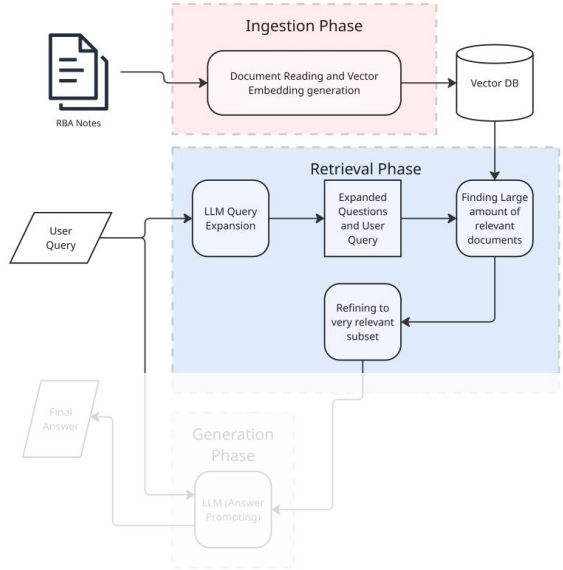


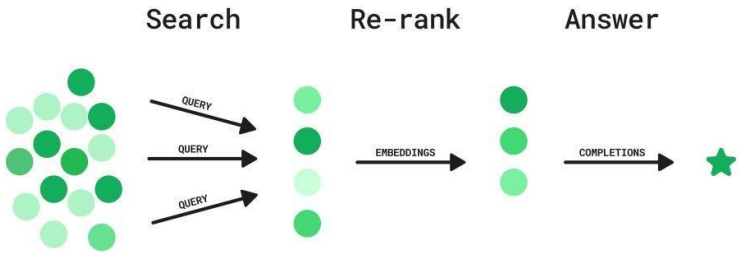
Re-rank

Re-ranking is a second, more accurate scoring step after the initial search.

It uses a small model to directly compare each retrieved chunk with the original query.

- “ms-marco-MiniLM-L6-v2”





- Initial retrieval casts a wide net.

- Re-ranking narrows it down to a high-quality subset.

- We cannot put every full note into the final model.

Other search enhancements

Specific term searching (BM25):
This adds a layer of search for **specific terms**
- Complements semantic vector search, which focuses on **meaning**.
- Balancing, Relevance (semantic meaning) and Precision (exact matches)

Maximal Marginal Relevance
Each query search for both also searches in a method called Maximal Marginal Relevance (MMR) to balance:
- Precision / Relevance
- Diversity

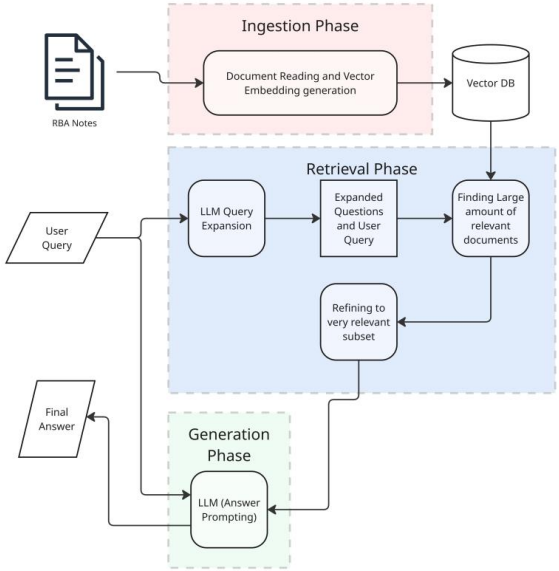
(Alters score to retrieve relevant chunks to a query and diverse from each other)

Time Decay
The returned scores from our reranking are adjusted to account for the age of the content, gradually reducing the relevance of older items to prioritize more recent and timely results.

Generation

Qwen3

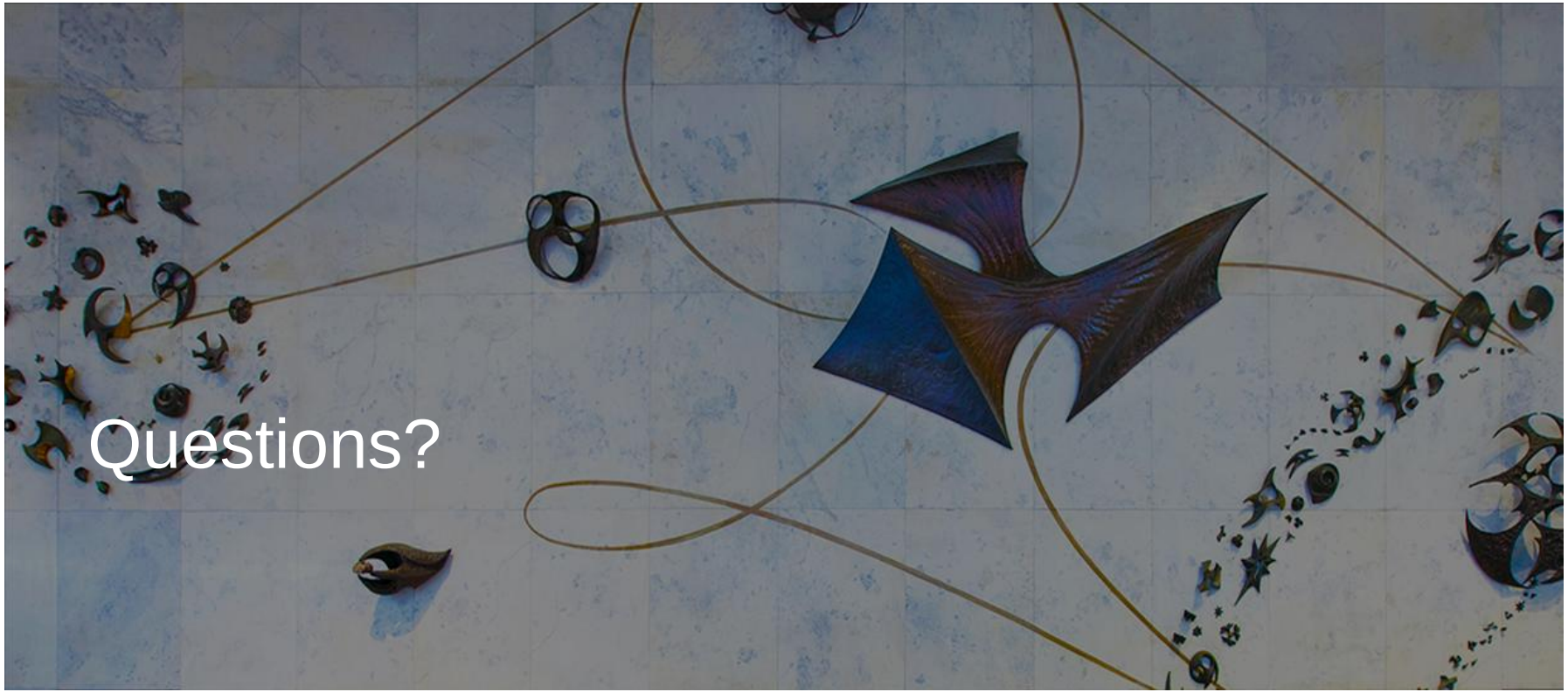
- Number of Parameters: 14.8B
- Quantized to Q4
  - ✓ Speed, memory efficient
  - × Loss of accuracy
- Theoretical context limit: ~ 25,000 words
  - On our hardware: ~ 12,000 words
    - Input + output



- ✓ Leading performance among open-source models
- ✓ Thinking mode (switchable)
- × Conversation mode is limited by context length
- × Concurrent users

Next steps

- Early-stage user testing: small scale; gather feedback; re-optimise tool. (We are here)
- Enhance compute: re-deploy tool on upcoming new hardware
- Expanded user testing & accuracy evaluation
- Cost-benefit analysis
- Risk assessment
- Release


Questions?

Search Performance Test – Re-ranking