





Topic

- Regular chatbot problems
- What RAG offers
- How it works
- How to use it
- Why to use it
- Conclusion
- Related technologies

Emerging trends: a gentle introduction to RAG

Kenneth Ward Church , Jiameng Sun, Richard Yue , Peter Vickers , Walid Saba and Raman Chandrasekar 

Regular chatbot problems

Timeliness: Informations are **OUT OF DATE**

Hallucination: Informations are **IMPROPER**

Timeliness: (Asked in 2024)

user: Who won the most recent world series?

Response: The Atlanta Braves won the most recent World Series in 2021.

Hallucination: (Not specifying what paper)

user: Please summarize the paper on psycholinguistics.

Response: Sure! Psycholinguistics is a field that blabla...

Regular chatbot problems

What is the source of these problems?

Timeliness: LLM was trained at a certain time.

Hallucination: No context, poor data, lack of infos, etc..

How to solve them?

Basic method (No additional infos): Using guard rails.

user: Who won the world series in 2035?

Response: I'm sorry, but I am unable to provide real-time information

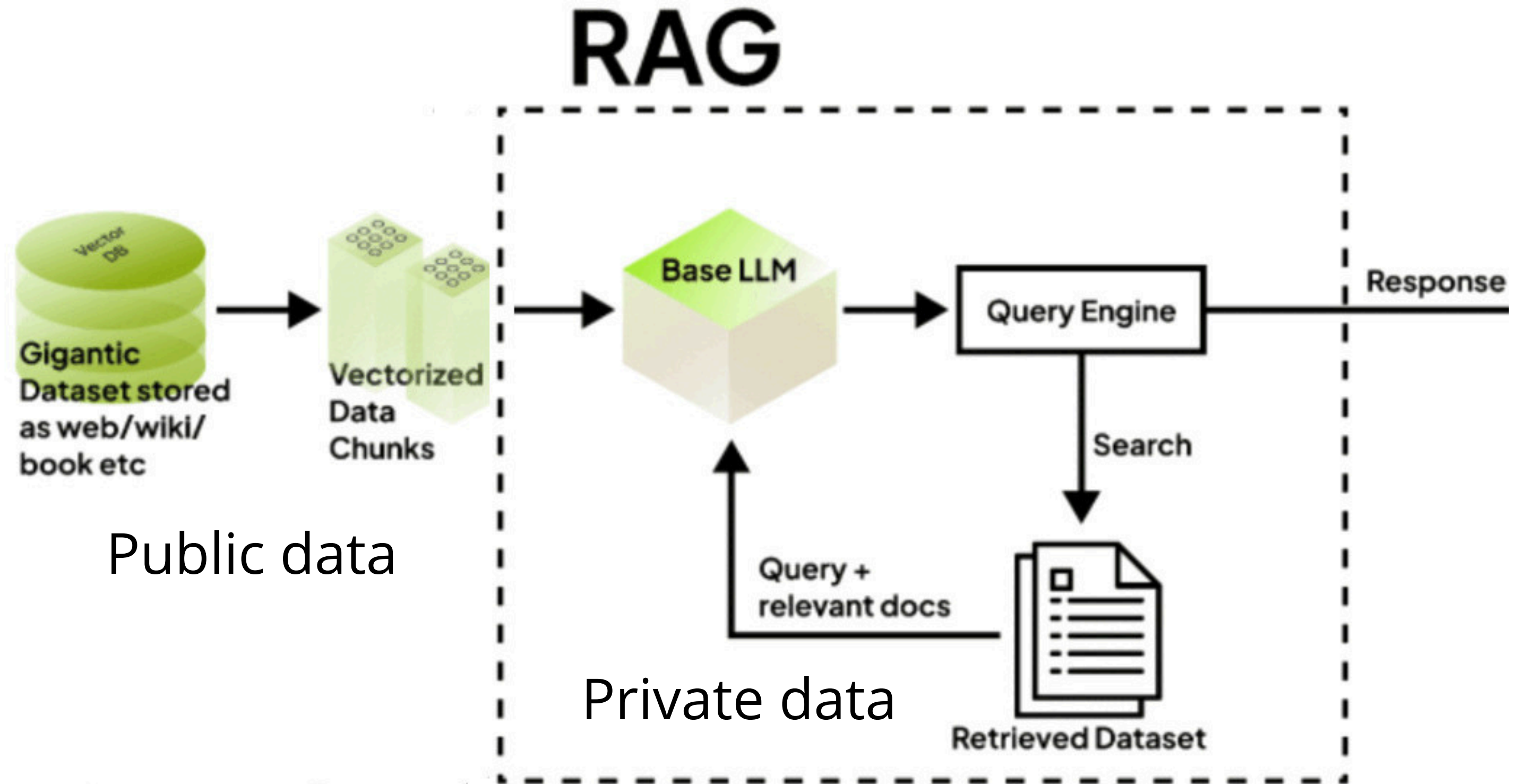
user: Please summarize the paper on psycholinguistics?

Response: I would need more specific information to provide an accurate answer.

Advanced method (Additional infos): ✨RAG✨

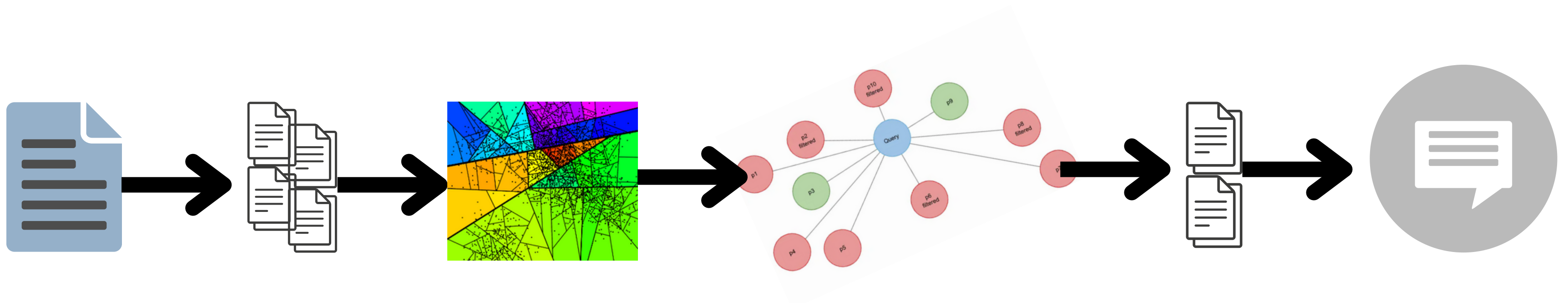
What RAG offers

“Retrieval-augmented generation (RAG) adds a simple but powerful feature to chatbots, the ability to upload files just-in-time.”



How it works

1. Upload one or more files
2. Parse files into chunks
3. Embed chunks as vectors in a vector space
4. Index vectors with an ANN algorithm such as ANNOYk or FAISS
5. Retrieve chunks near prompt (using ANN)
6. Generate responses

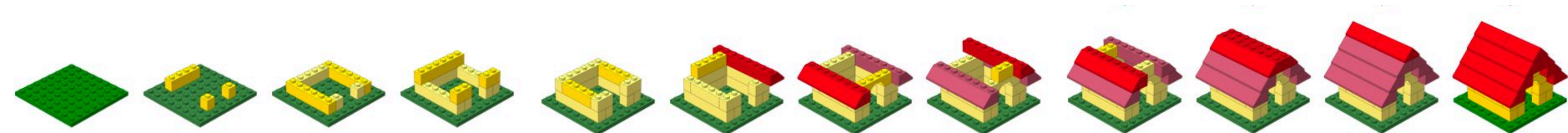


How to use it

Chatbots often lost in the context so its good to pave the way.

Chain-of-Thought (CoT): Decomposing larger tasks into smaller subtasks.

“In short, chatbots are not magic. Chatbots will be more successful if we spoon-feed them. RAG spoon-feeds them by inserting the relevant content into the input before invoking the response, and CoT Prompting spoon feeds them by cutting up prompts into bite-size pieces to prevent hallucinations.”



Why to use it

Robust guard rails: Too easy to crash through existing guard rails

Pivots/Hallucinations: Chatbots tend to pivot and/or hallucinate when asked to discuss content that goes beyond the training set such as an unspecified paper

Timeliness: Training time = inference time

Conclusion

“Chatbots are trained on massive amounts of public data.

By adding the ability to upload files just-in-time, RAG addresses a number of gaps in the chatbot’s knowledge base such as timeliness, references to background knowledge, private data, etc. Gaps in the knowledge base can lead to hallucinations.

By filling in many of these gaps just-in-time, RAG reduces the chance of hallucinations.”

Related technologies

Tools: LangChain, HuggingFace, and VecML

Benchmark: CRAG, CLAPNQ

Source:

<https://www.cambridge.org/core/journals/natural-language-engineering/article/emerging-trends-a-gentle-introduction-to-rag/4FF461F4066A0C16135F2D2849E3356A>

Thank you!