SNGULAR

Grounding Language Models.
RAG in Action

Retrieval-Augmented Generation.

Contents:

What is Retrieval-Augmented Generation(RAG)?

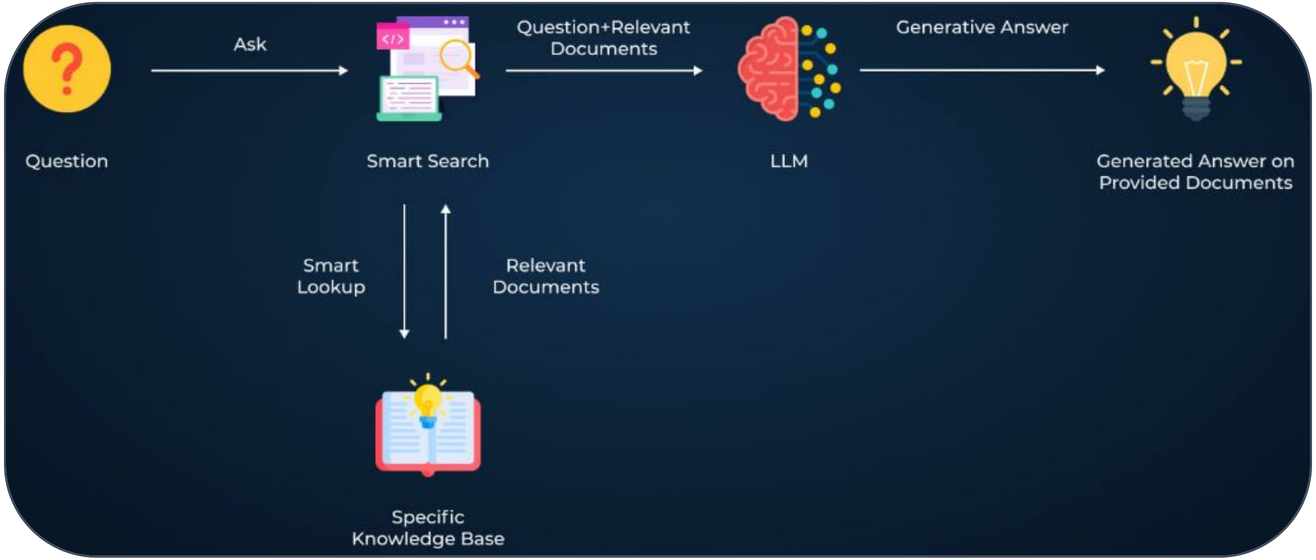
- **Retrieval:** Find relevant documents/passages from an external knowledge source (e.g., vectorial database).

- **Augmented:** Enriche prompt through providing context to language models.

- **Generation:** Use a language model like GPT or BERT-based models to generate a response based on the retrieved information.

SNGULAR

What is Retrieval-Augmented Generation(RAG)?

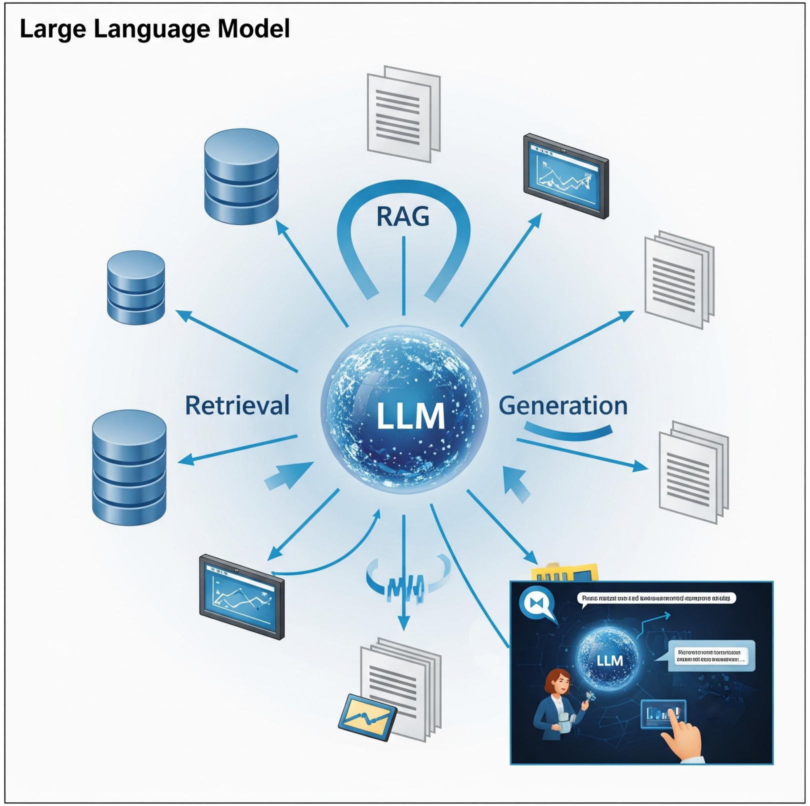


SNGULAR

Why RAG?

Large language models (LLMs) like GPT are trained on huge datasets but:

- They can hallucinate or produce inaccurate facts.
- They don't know anything newer than their training cutoff date.
- They can’t access private, specific, or dynamic data (e.g., your company docs, current news).



SNGULAR

Grounding & Context

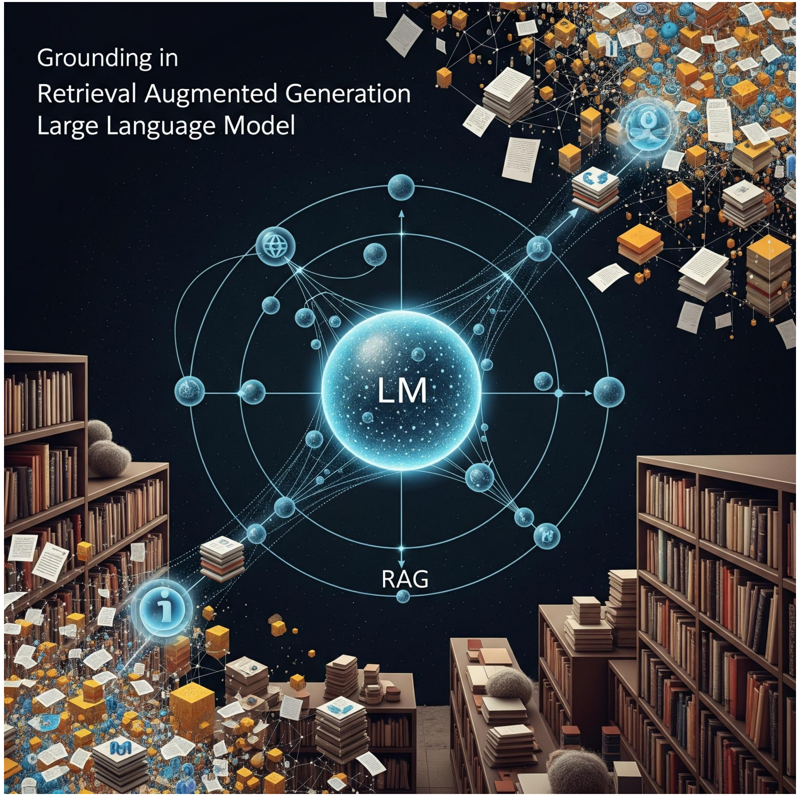
Grounding

Grounding means ensuring that a model's **output is based on real, verifiable data**:

- Instead of "**hallucinating**" facts, the model refers to **retrieved documents**.

- The response is traceable to **actual sources** (like a passage in a **document**, a **database** entry, etc.).

Example: Citizen request costs - No grounding **LLM will guess.**



SNGULAR

Grounding & Context

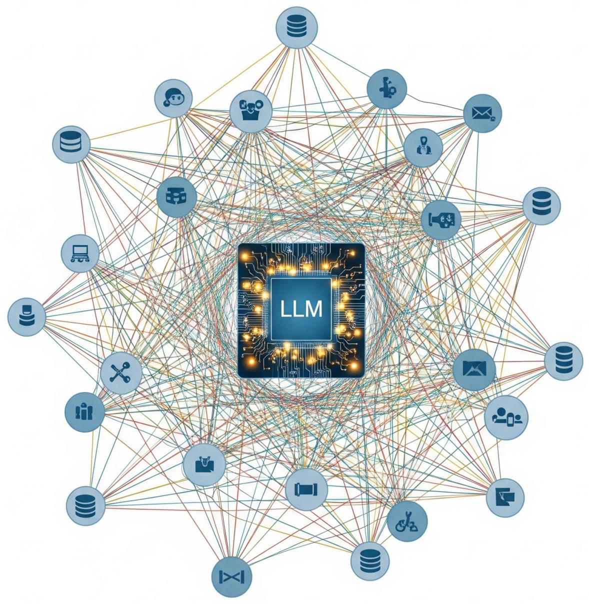
Context

Information available to the model during generation that **helps** it **understand** and respond **accurately**.

It's important because:

- Improves factual accuracy
- Increases trust (traceability to sources)
- Enables domain-specific responses (like medical, legal, corporate)

Retrieval Augmented Generation



SNGULAR

Vectorization

Tokenization

```
["What", "is", "RAG", "?"]
```

Let:

- $T = [t_1, t_2, ..., t_n]$ be the token sequence.

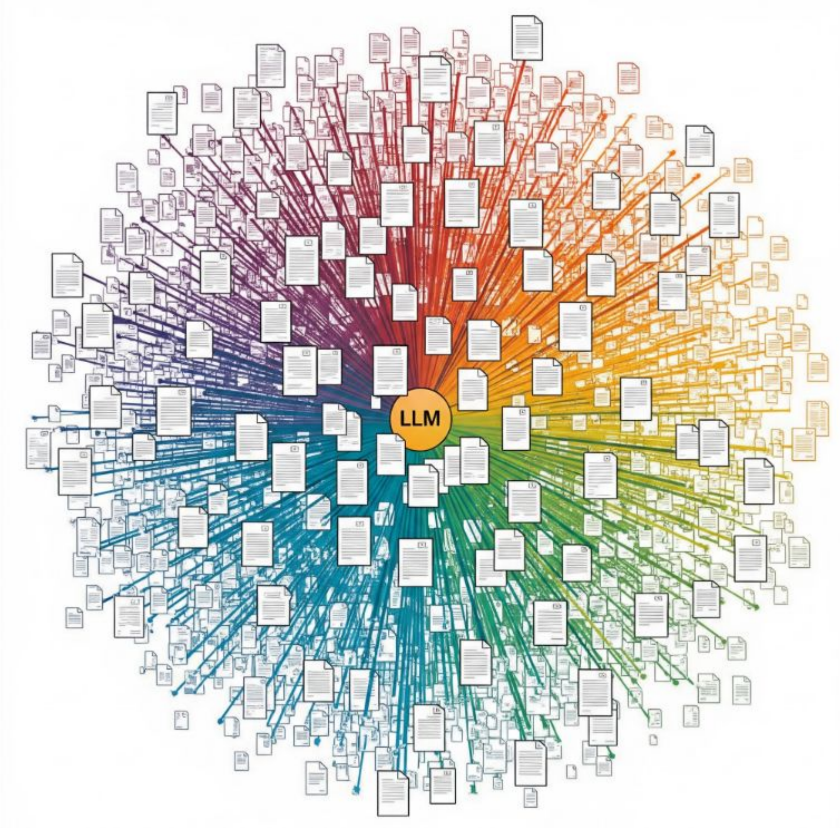
Token Embedding Lookup

- V: vocabulary size
- d: embedding dimension (e.g., 768)

$$\mathbf{e}_i = E[t_i]$$

So the sequence becomes:

$$[\mathbf{e}_1, \mathbf{e}_2, ..., \mathbf{e}_n] \in \mathbb{R}^{n \times d}$$

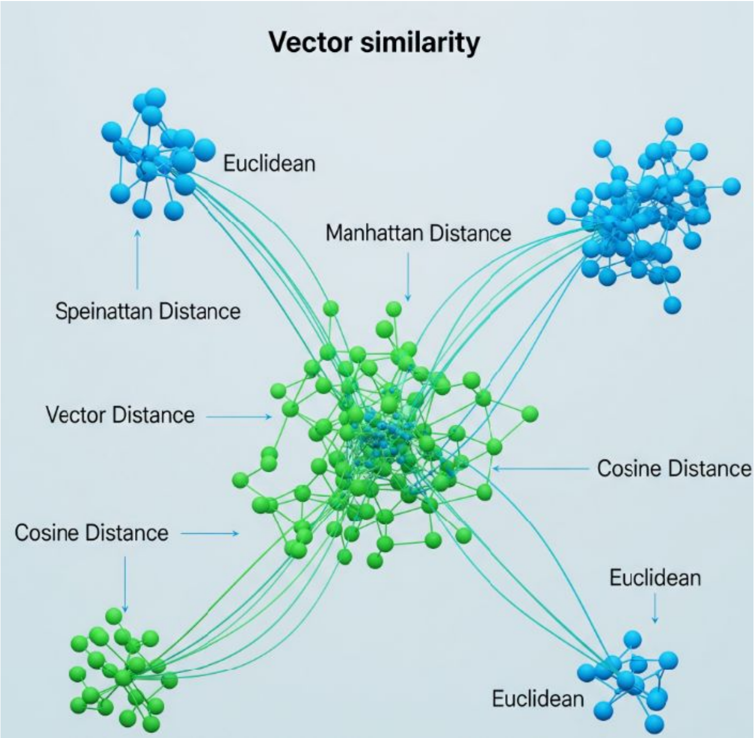


https://gptforwork.com/tools/tokenizer

SNGULAR

Distance Measurements

- Understanding distance metrics is crucial for RAG.
- Various methods exist to calculate vector similarity.
- Common measures include Cosine and Euclidean Distance.
- Selecting the right distance metric impacts retrieval quality.

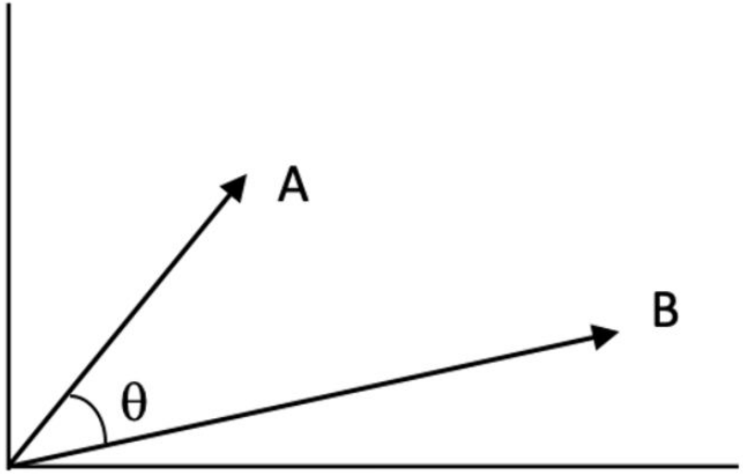


SNGULAR

Distance Measurements.

Cosine Similarity

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$



- **Range**: $[-1, 1]$
- **Use case**: Text embeddings, where direction matters more than magnitude.

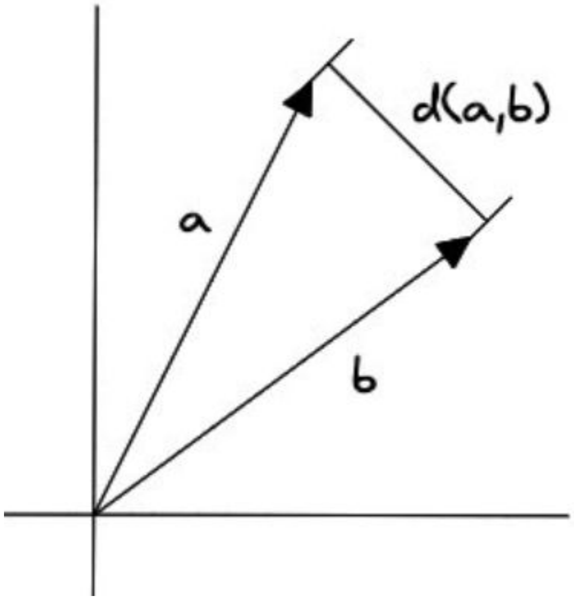
SNGULAR

Distance Measurements.

Euclidean Distance

$$\text{euclidean\_distance}(\vec{a}, \vec{b}) = \sqrt{\sum_i (a_i - b_i)^2}$$

- **Use case**: Clustering, spatial distances.
- **Downside**: Sensitive to scale and vector length.



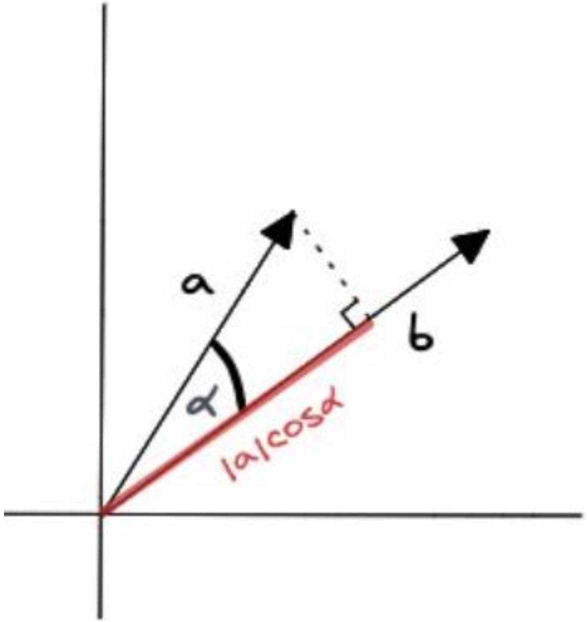
SNGULAR

Distance Measurements.

Dot Product

$$\vec{a} \cdot \vec{b} = \sum_i a_i b_i$$

- **Use case**: Sometimes used as a fast approximation of cosine similarity (especially when vectors are normalized).



SNGULAR

Real-world Use Case

01 Context of Project - Current Flow

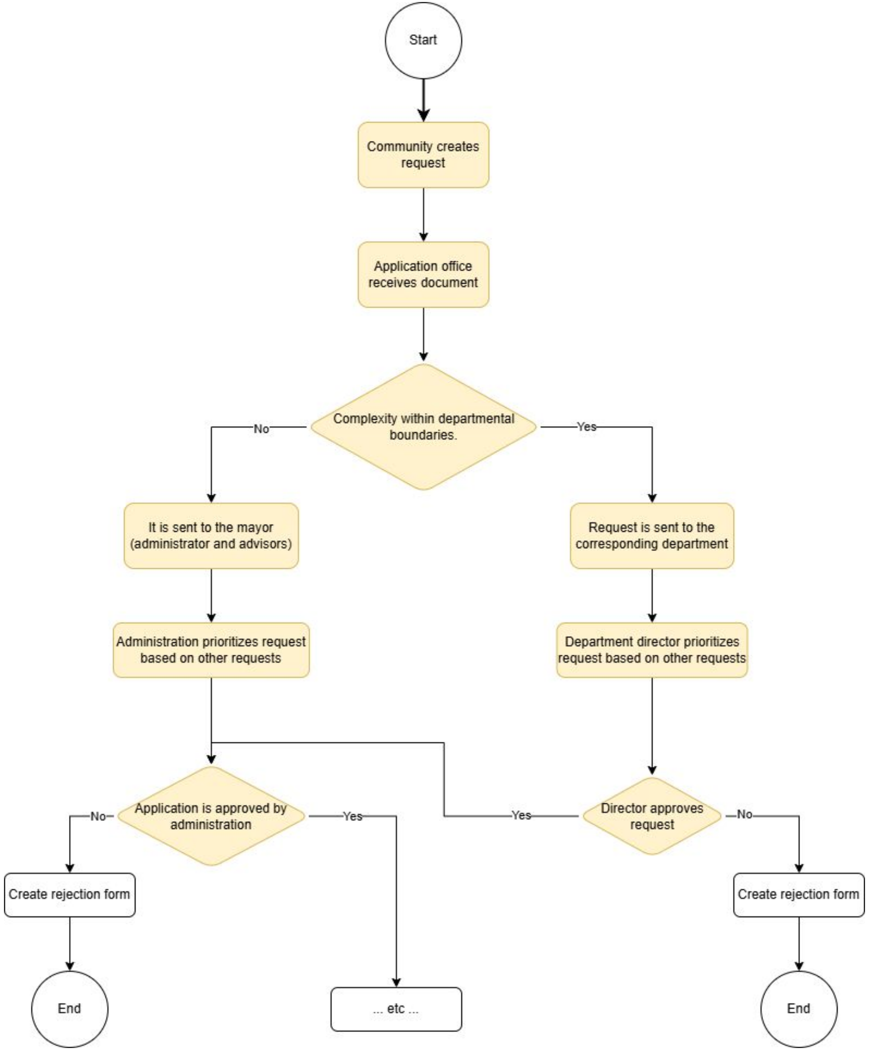
02 Context of Project - Benefit to the community

03 RAG

SNGULAR

Context of Project

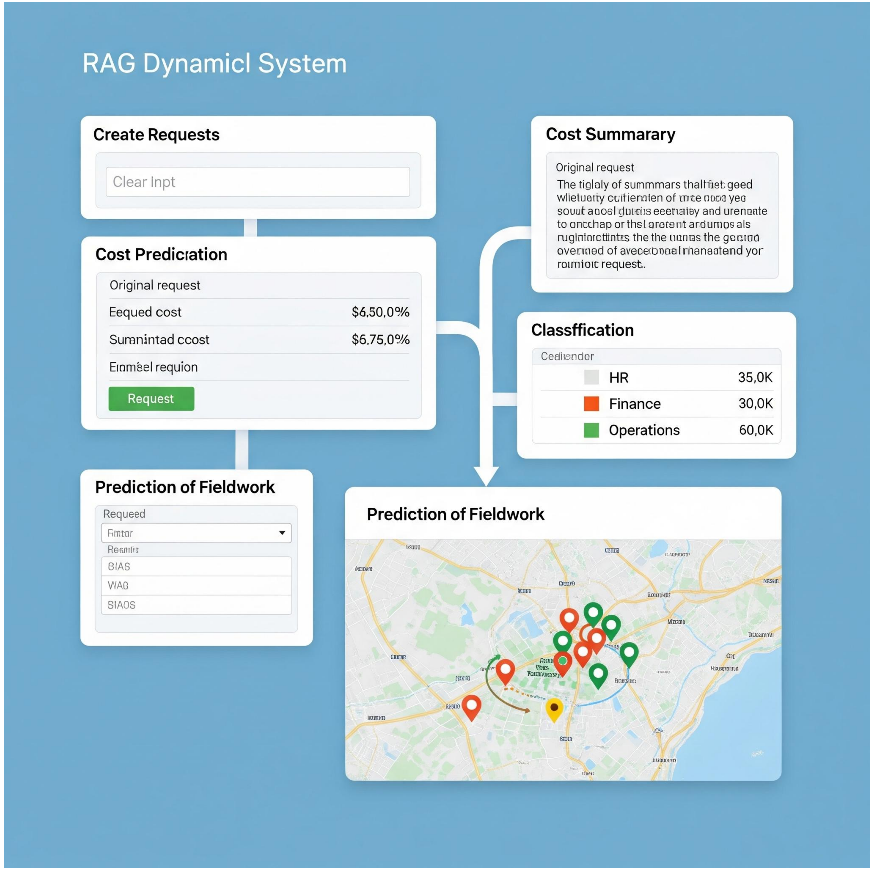
Current Flow



Context of Project

Functionalities of RAG

- Create requests
- Classification by priority
- Classification by department
- Summarization of requests
- Predict costs
- Predict fieldwork
- Geographic location

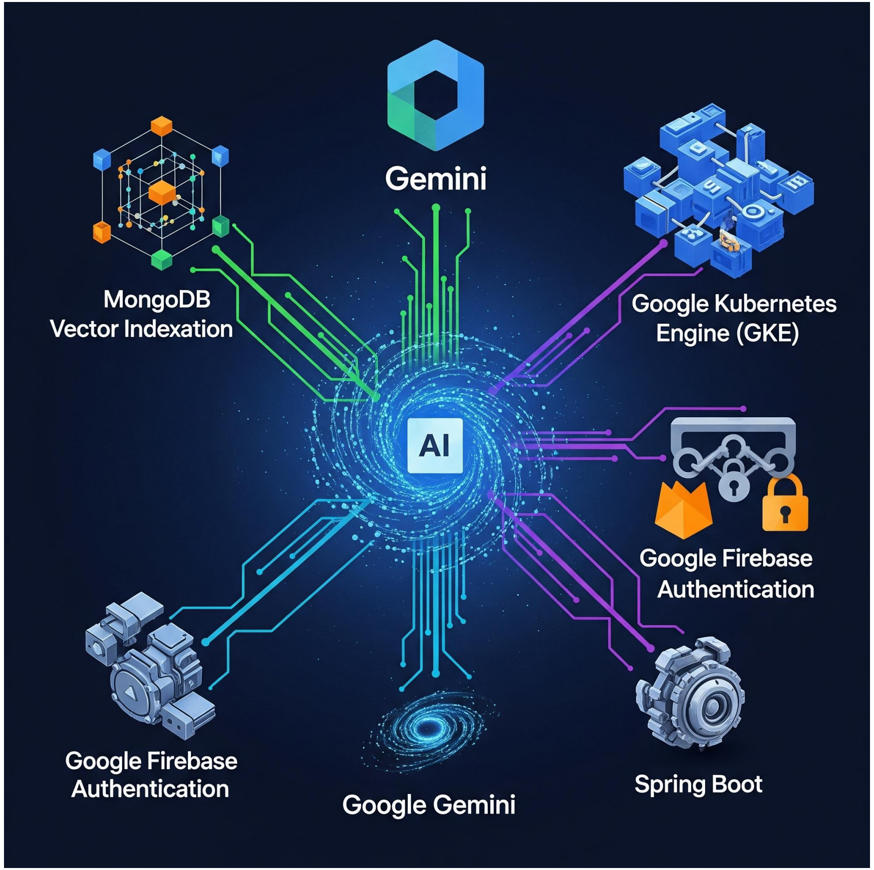


SNGULAR

Context of Project

Technologies used in RAG

- Google Gemini
- MongoDB - Vector Indexation
- Google Kubernetes Engine (GKE)
- Google Firebase Authentication
- Spring Boot



SNGULAR

Bibliography

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

- Yu, H., Gan, A., Zhang, K., Tong, S., Liu, Q., & Liu, Z. (2024, August). Evaluation of retrieval-augmented generation: A survey. In *CCF Conference on Big Data* (pp. 102-120). Singapore: Springer Nature Singapore.

- Wu, S., Xiong, Y., Cui, Y., Wu, H., Chen, C., Yuan, Y., ... & Xue, C. J. (2024). Retrieval-augmented generation for natural language processing: A survey. *arXiv preprint arXiv:2407.13193*.

- Jeyaraman, J. (2025). *Vector Databases Unleashed: Isolating Data in Multi-Tenant LLM Systems*. Libertatem Media Private Limited.

- Yu, J., Amores, J., Sebe, N., & Tian, Q. (2006, July). A new study on distance metrics as similarity measurement. In *2006 IEEE international conference on multimedia and expo* (pp. 533-536). IEEE.

Questions

Thanks for your attention!!!