

Week 6: Recap buffer

 by Alwin Lin

A small reminder

Category	Weight	Criteria
Functionality	50%	- The bot operates without errors and fulfills its intended purpose.
		- Accurate integration of RAG (e.g., demonstrates retrieval of relevant external data sources and effectively grounds responses).
		- Smooth interactions (clear responses, low latency, logical outputs).
Innovation & Creativity	25%	- Unique or creative application of the bot for a specific subject or user group.
		- Effective use of Gemini and embedding techniques to tailor the project.
		- Demonstrates originality in features or design.
Presentation Quality	25%	- Clear and structured presentation explaining the bot's goals, design choices, and technical implementation.
		- Demonstrates a solid understanding of RAG principles and embedding models.
		- Engages the audience with visuals (e.g., slides, live demo, or flowcharts) and answers questions effectively.

A small reminder

Category	Weight	Criteria
Functionality	50%	- The bot operates without errors and fulfills its intended purpose.
		- Accurate integration of RAG (e.g., demonstrates retrieval of relevant external data sources and effectively grounds responses).
		- Smooth interactions (clear responses, low latency, logical outputs).
Innovation & Creativity	25%	- Unique or creative application of the bot for a specific subject or user group.
		- Effective use of Gemini and embedding techniques to tailor the project.
		- Demonstrates originality in features or design.
Presentation Quality	25%	- Clear and structured presentation explaining the bot's goals, design choices, and technical implementation.
		- Demonstrates a solid understanding of RAG principles and embedding models.
		- Engages the audience with visuals (e.g., slides, live demo, or flowcharts) and answers questions effectively.

So far in the course

- Prompt Engineering
- Tokens temperatures and Cost?
- What are embeddings?
- What are vector DBs?
- How does it all fit together?



Week 2: Prompt Engineering

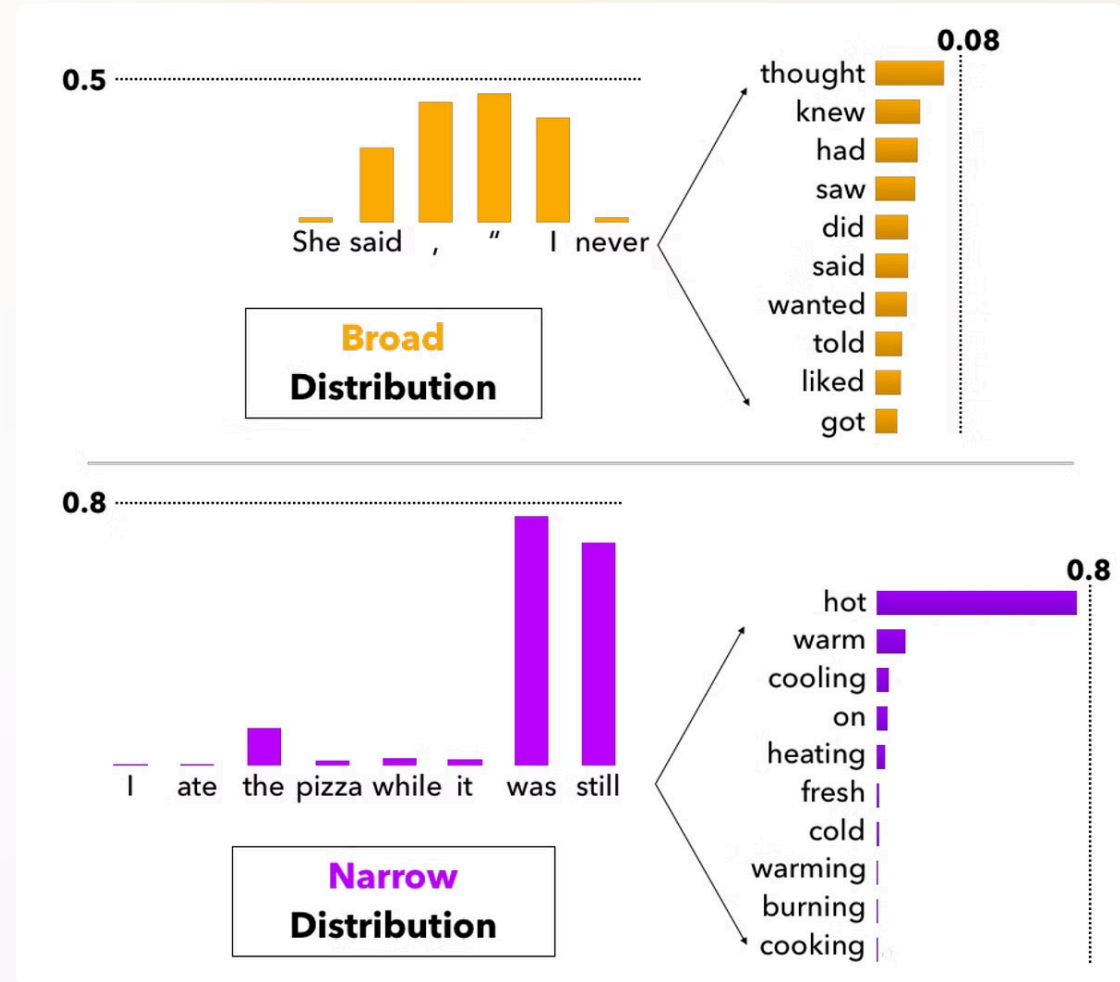
- [In context learning](#)
- [System Instruction](#)
- [Directional stimulus prompting](#)

Neuron activation



Week 3: Tokens, Temperatures, Costs

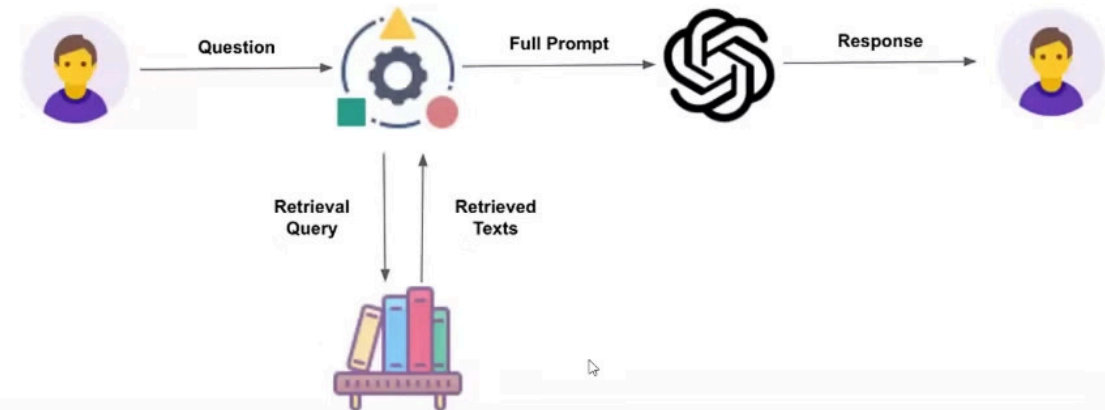
- [How much money are we burning?](#)
- [What are temperatures?](#)
- How to solve LeetCode questions with 200 tokens?



Week 4: Embeddings, VectorDB, workflows

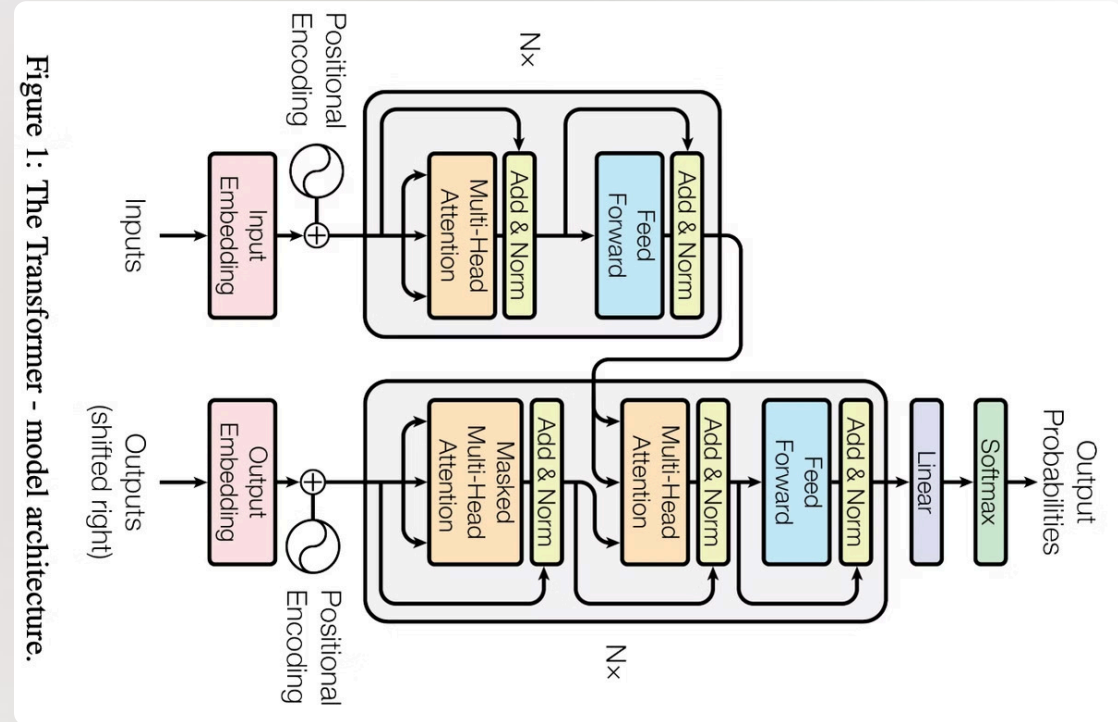
- [What is this RAG?](#)
- [What are my options in embedding models?](#)
- What are my options in vector database?
- Why do I need a vector database?

Retrieval Augmented Generation



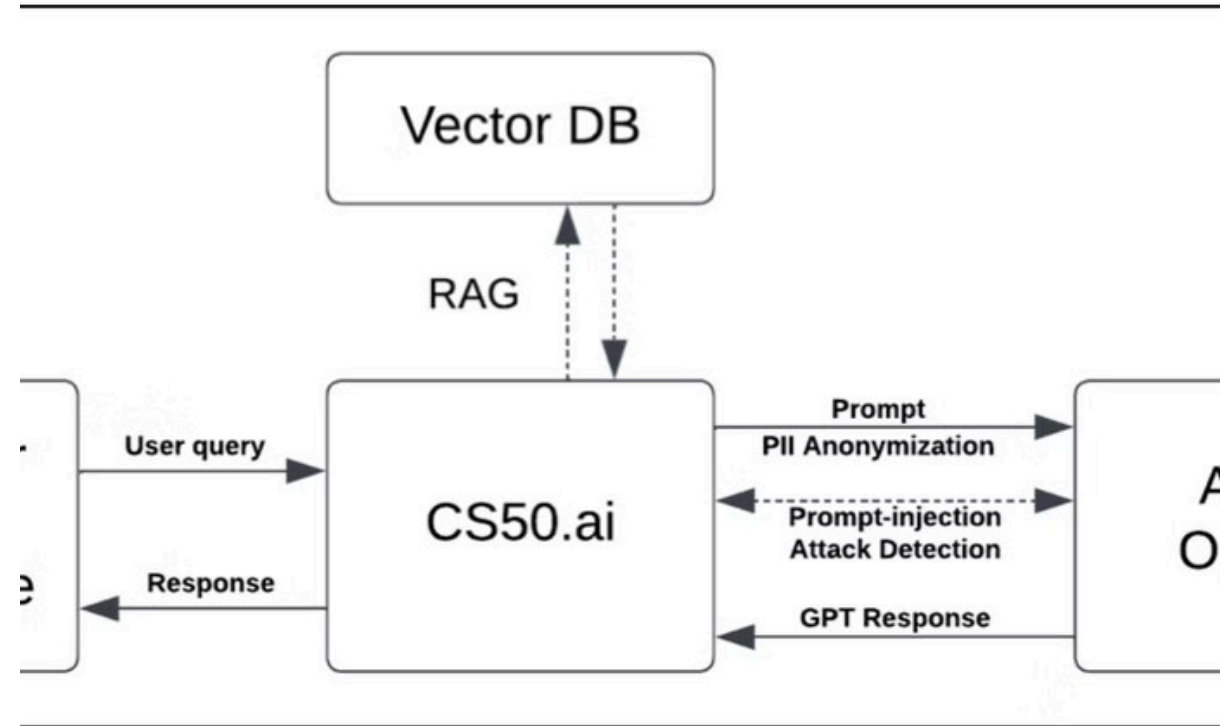
Week 5: Behind the working of embedding

- What is a vector? Why vectors?
- How does it know what I'm looking for?
- How does that help us with RAG?



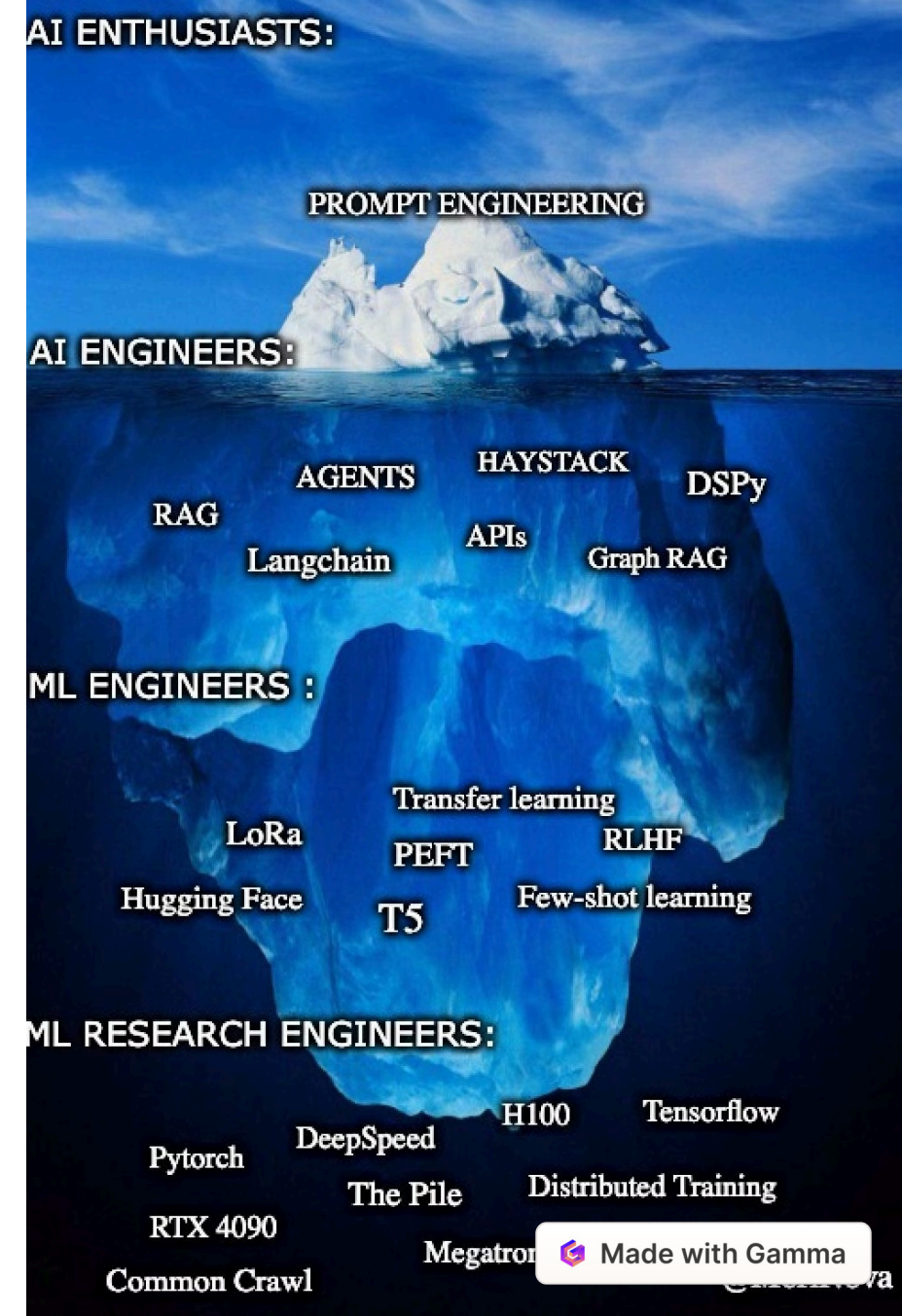
What does the high level look like?

- A RAG powered LLM
- A Vector DB
- A functioning UI



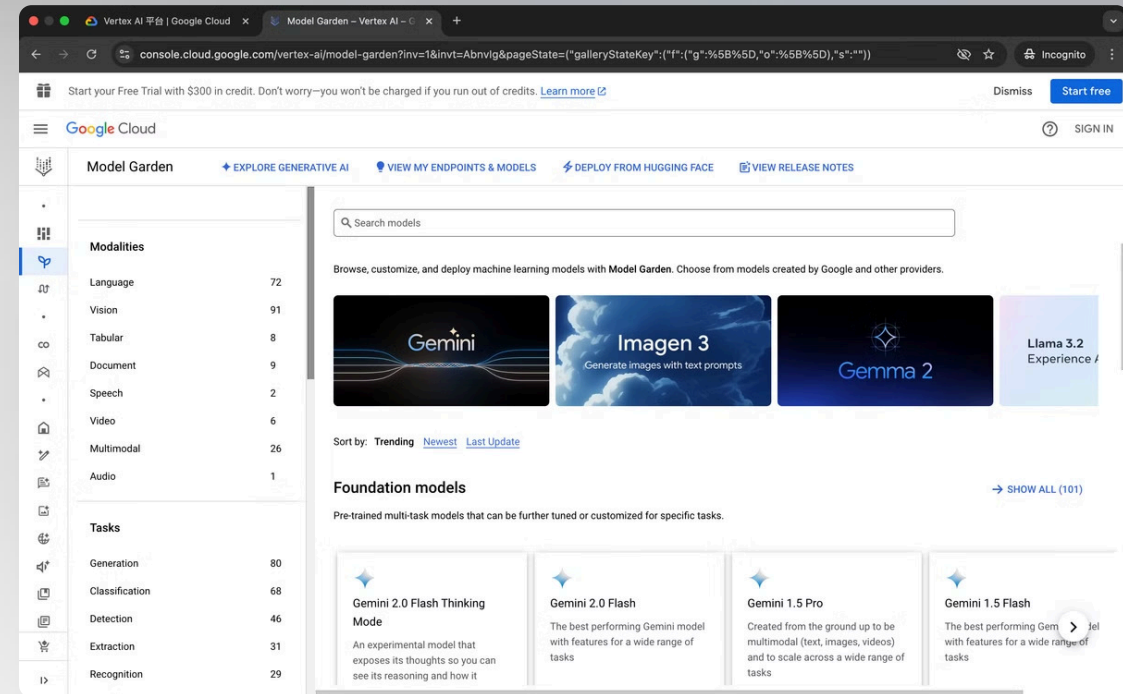
he system architecture of CS50.ai. G
ises to student queries and optional
-augmented generation technique t
ccuracy by incorporating facts from

What is outside of what we taught?



Vertex AI: Free stuff

- What's the catch?
- What do I need?
- How much stuff can I do with it?



MultiModel capabilities:

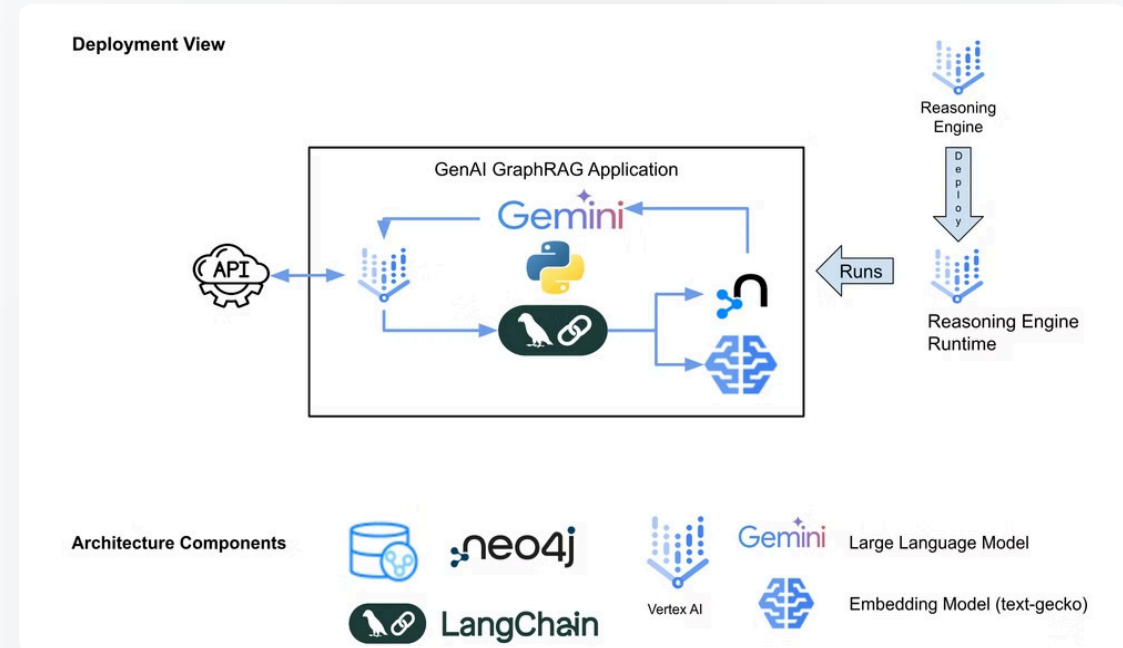
- How does it do it?
- What are the limitations
- How much does it cost?

its apply when you use the multimodalembdding model for text and image embedding

Value and description	
Text and image data	
of API e per	120
ngth	32 tokens (~32 words) The maximum text length is 32 tokens (approximately 32 words). If the input exceeds 32 tokens, the model internally shortens the input to this length.
	English
	BMP, GIF, JPG, PNG
	Base64-encoded images: 20 MB (when transcoded to PNG) Cloud Storage images: 20MB (original file format) The maximum image size accepted is 20 MB. To avoid increased network latency, use Cloud Storage. Additionally, the model resizes images to 512 x 512 pixel resolution. Consequently, you should provide higher resolution images.
Video data	
	N/A - The model doesn't consider audio content when generating video embeddings
	AVI, FLV, MKV, MOV, MP4, MPEG, MPG, WEBM, and WMV
ngth	No limit. However, only 2 minutes of content can be analyzed at a time.


GraphRAG: When words are just not enough


- What if we use graphs?
- When and why traditional RAG fails?
- Why don't we all use graphs?



RAGAS: How do I know RAG is working?

- Is it my data? My model?
- What could be evaluated?
- How do I go about evaluating?

 List of available metrics

 **Core Concepts**

Components

- General >
- Evaluation >

Metrics

- Overview
- Available Metrics ▾
 - Retrieval Augmented Generation >
 - Agents or Tool Use Cases >
 - Natural Language Comparison >
 - SQL >
 - General Purpose >
 - Other Tasks >

Test Data Generation

- RAG >
- Agents or tool use >

Feedback Intelligence

Retrieval Augmented Generation


- Context Precision
- Context Recall
- Context Entities Recall
- Noise Sensitivity
- Response Relevancy
- Faithfulness
- Multimodal Faithfulness
- Multimodal Relevance

Agents or Tool use cases

- Topic adherence
- Tool call Accuracy
- Agent Goal Accuracy

Natural Language Comparison

- Factual Correctness
- Semantic Similarity
- Non LLM String Similarity
- BLEU Score
- ROUGE Score

 Made with Gamma