Getting Started with Llama3

Lucas Soares

11-06-2024

Methodology Notes

The presentation will be organized into the following structure:

Methodology Notes

The presentation will be organized into the following structure:

1. Presentation Block

Methodology Notes

The presentation will be organized into the following structure:

1. Presentation Block
2. Notebook Demo

Methodology Notes

The presentation will be organized into the following structure:

1. Presentation Block
2. Notebook Demo
3. Quick Q&A + Summary

Methodology Notes

The presentation will be organized into the following structure:

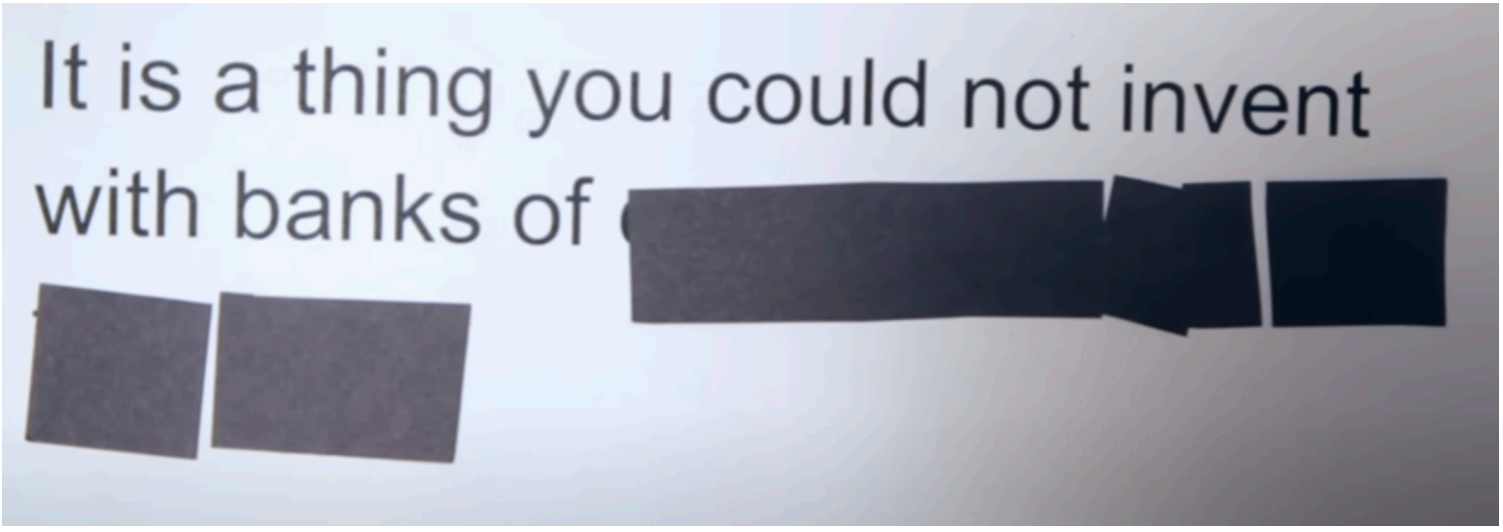
1. Presentation Block
2. Notebook Demo
3. Quick Q&A + Summary
4. Optional Exercise During Q&A

Methodology Notes

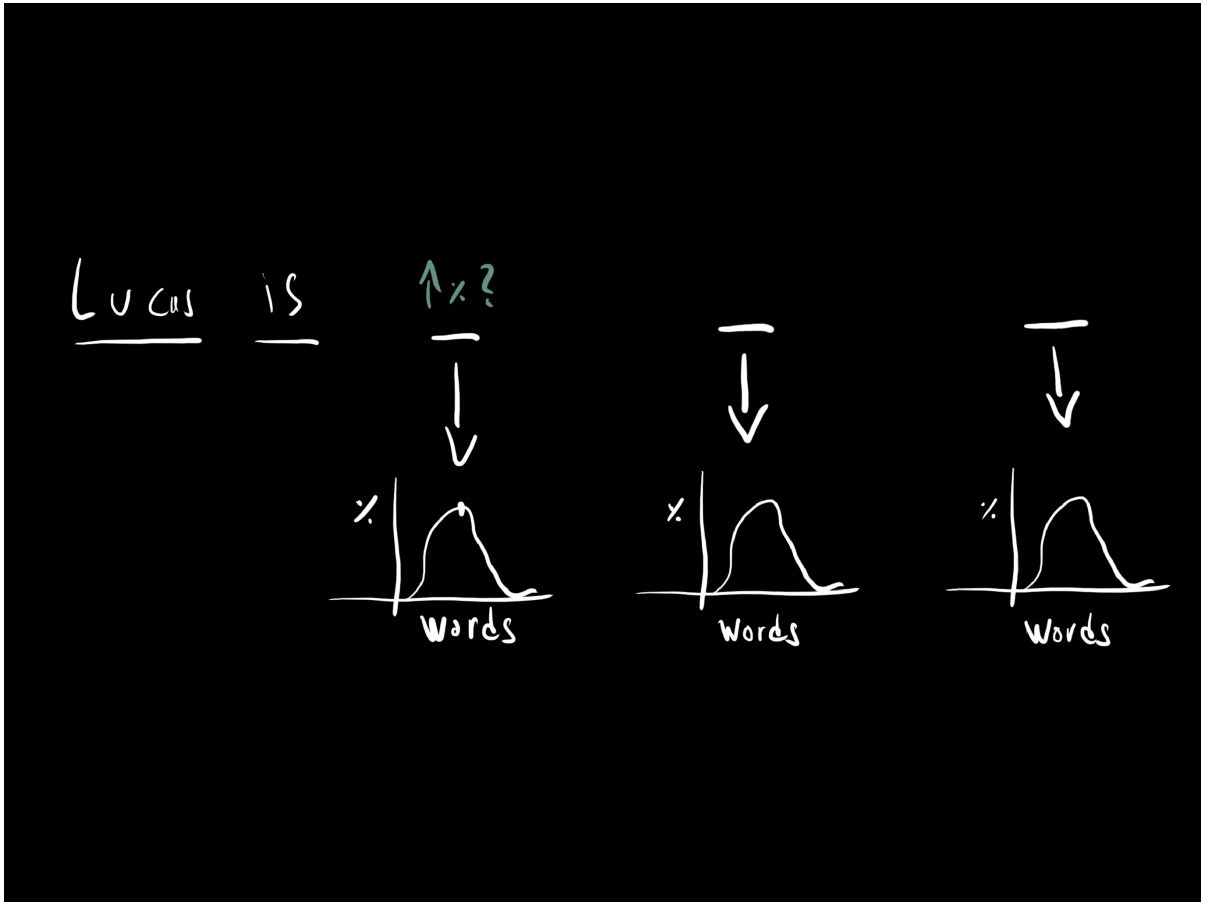
The presentation will be organized into the following structure:

1. Presentation Block
2. Notebook Demo
3. Quick Q&A + Summary
4. Optional Exercise During Q&A
5. Repeat

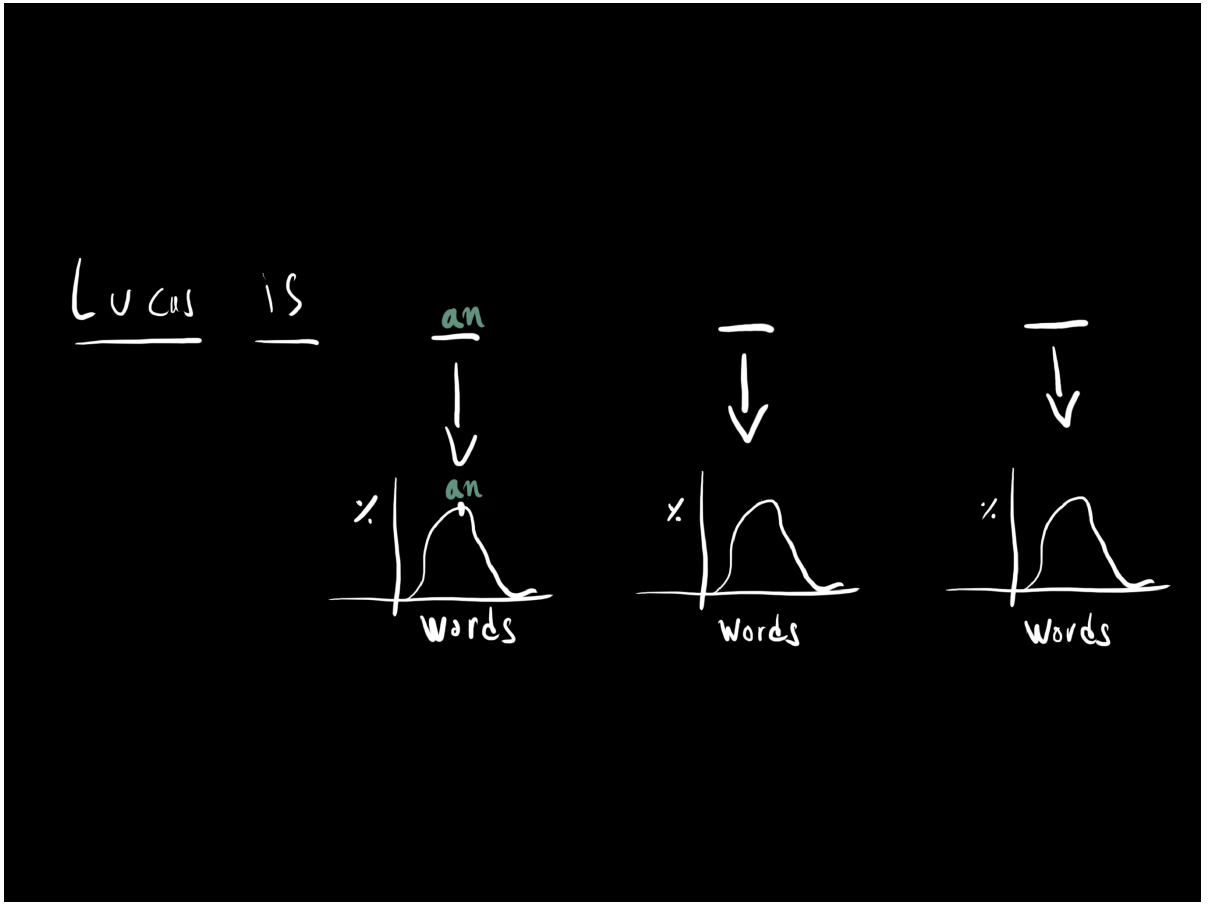
LLMs Predict the Next Word



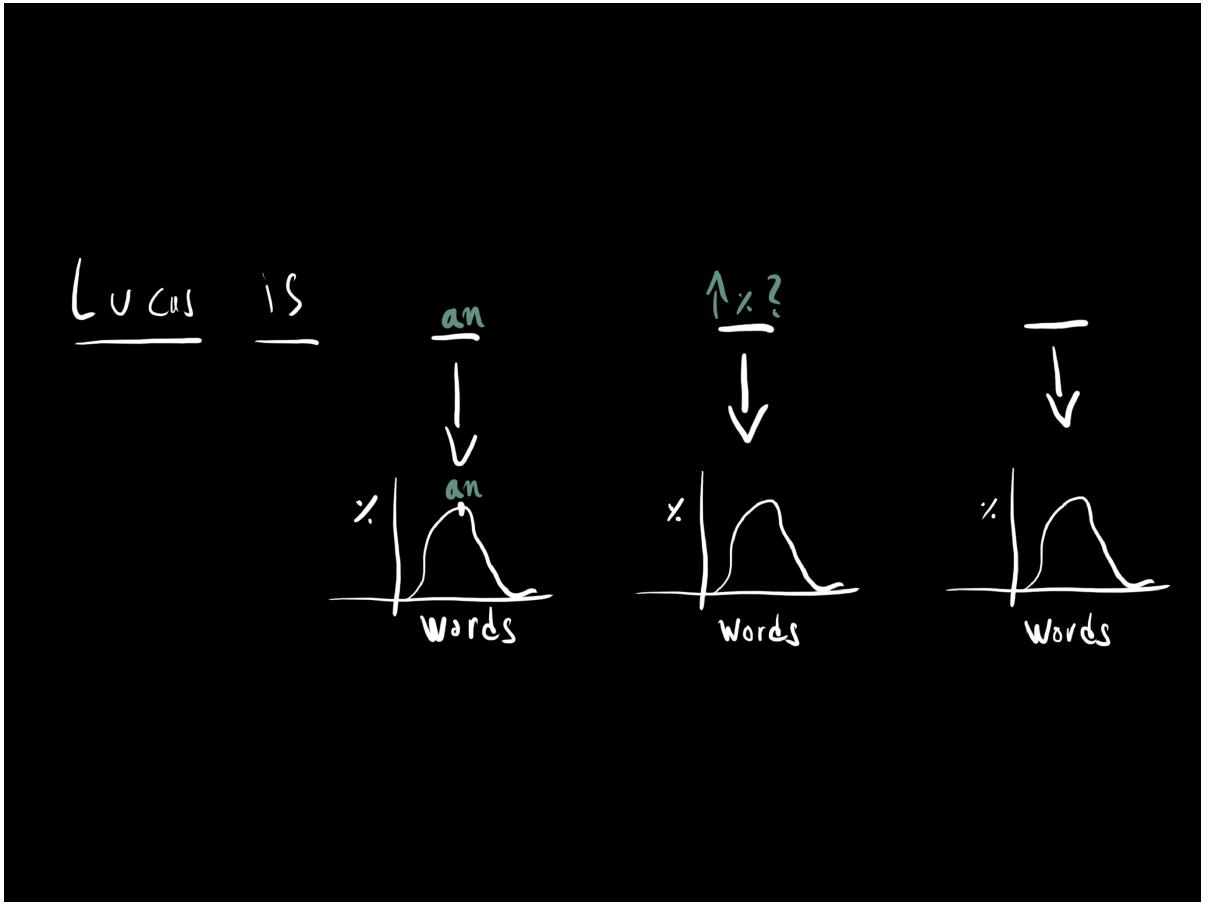
LLMs Predict the Next Word



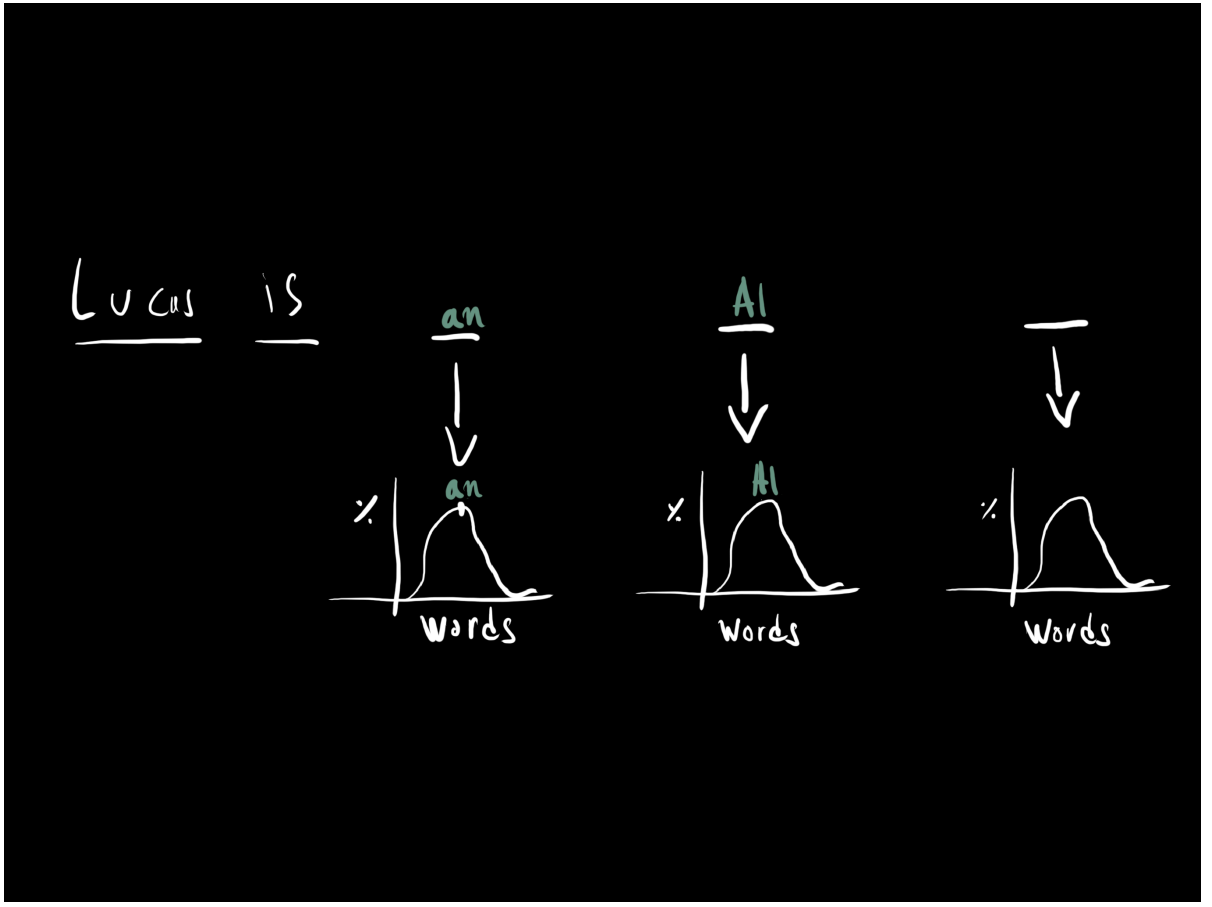
LLMs Predict the Next Word



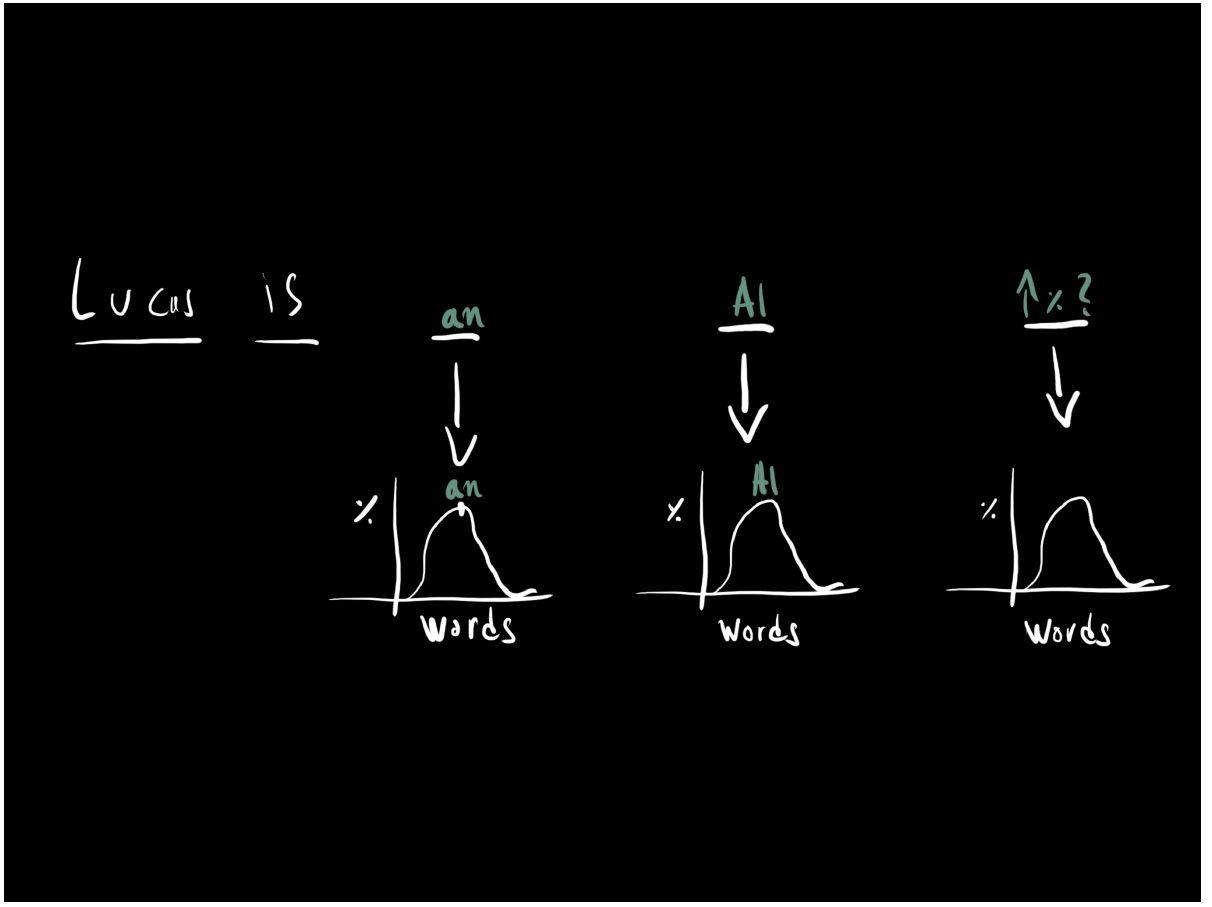
LLMs Predict the Next Word



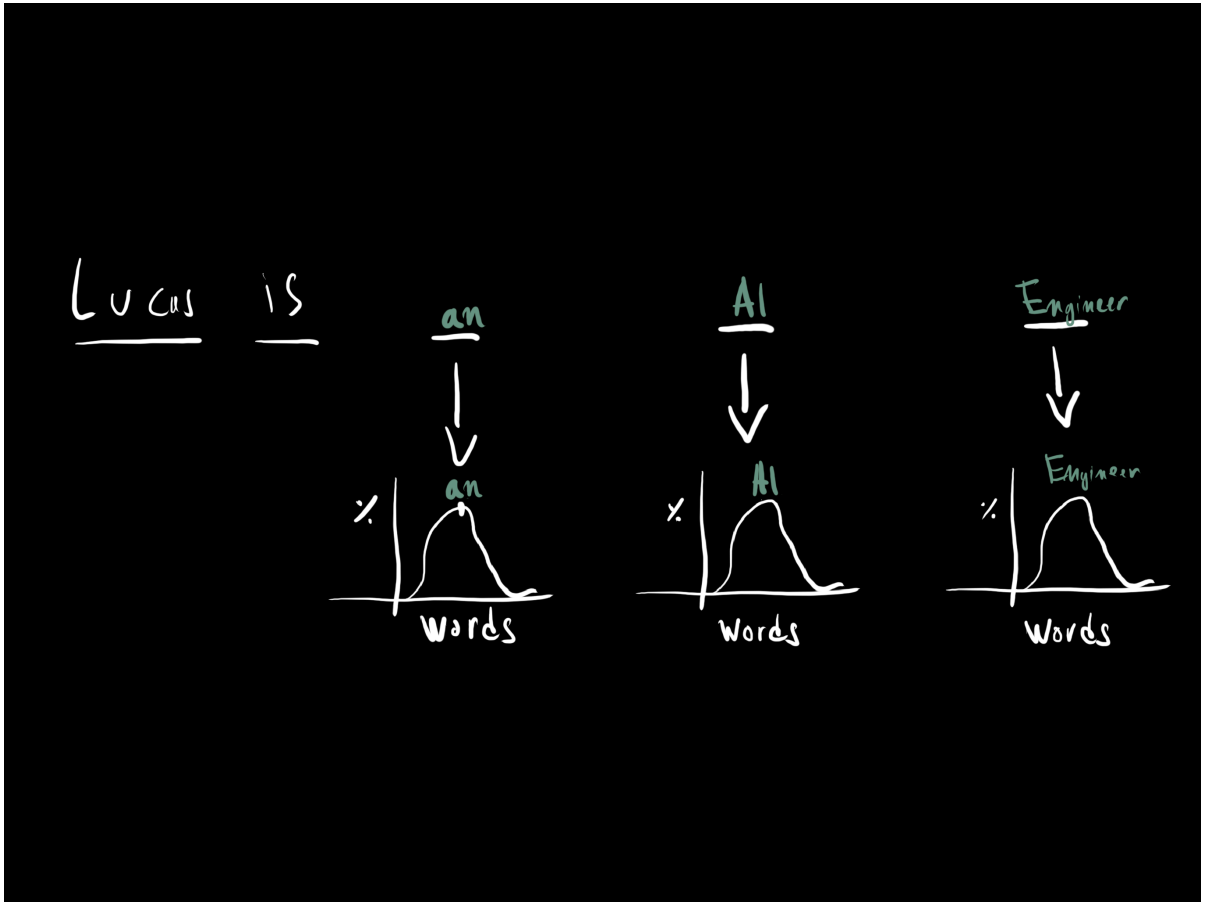
LLMs Predict the Next Word



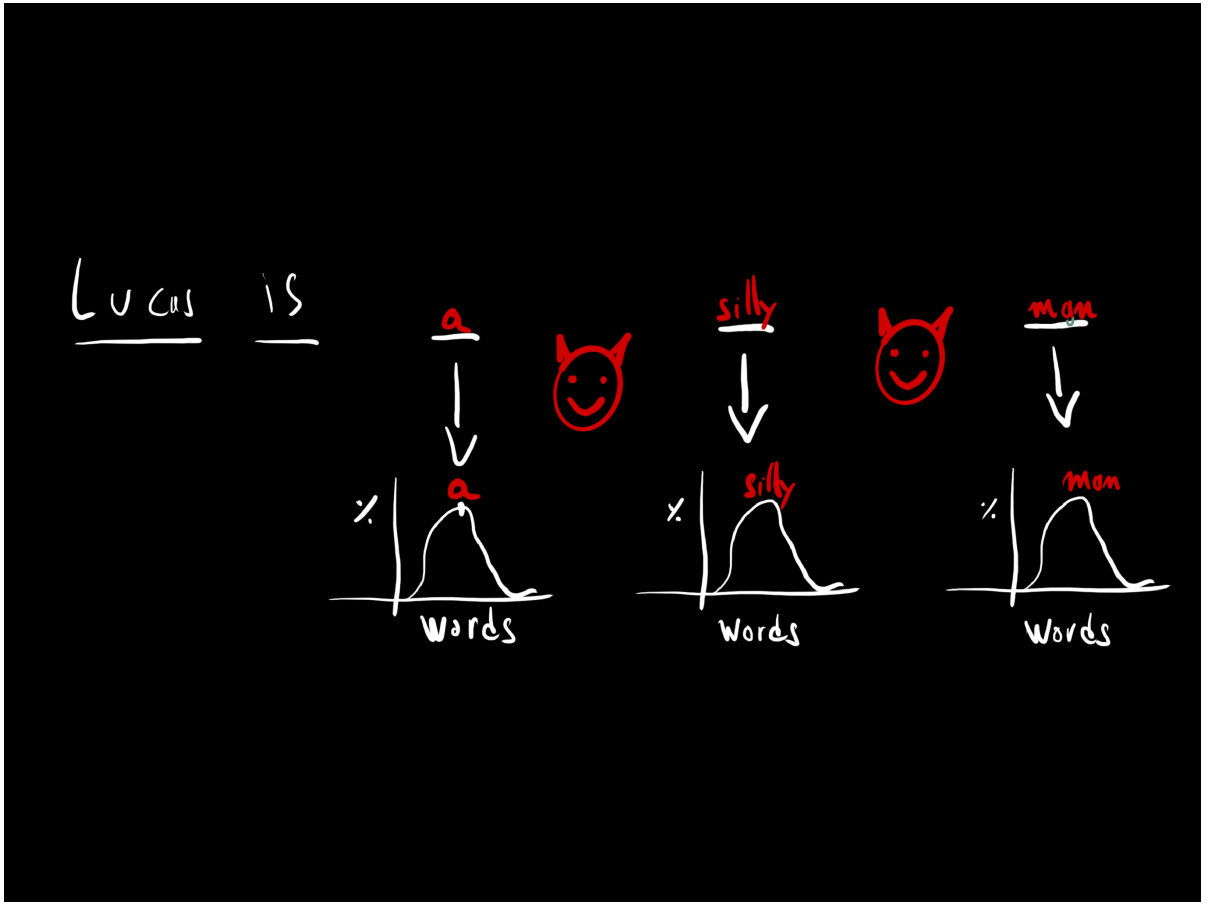
LLMs Predict the Next Word

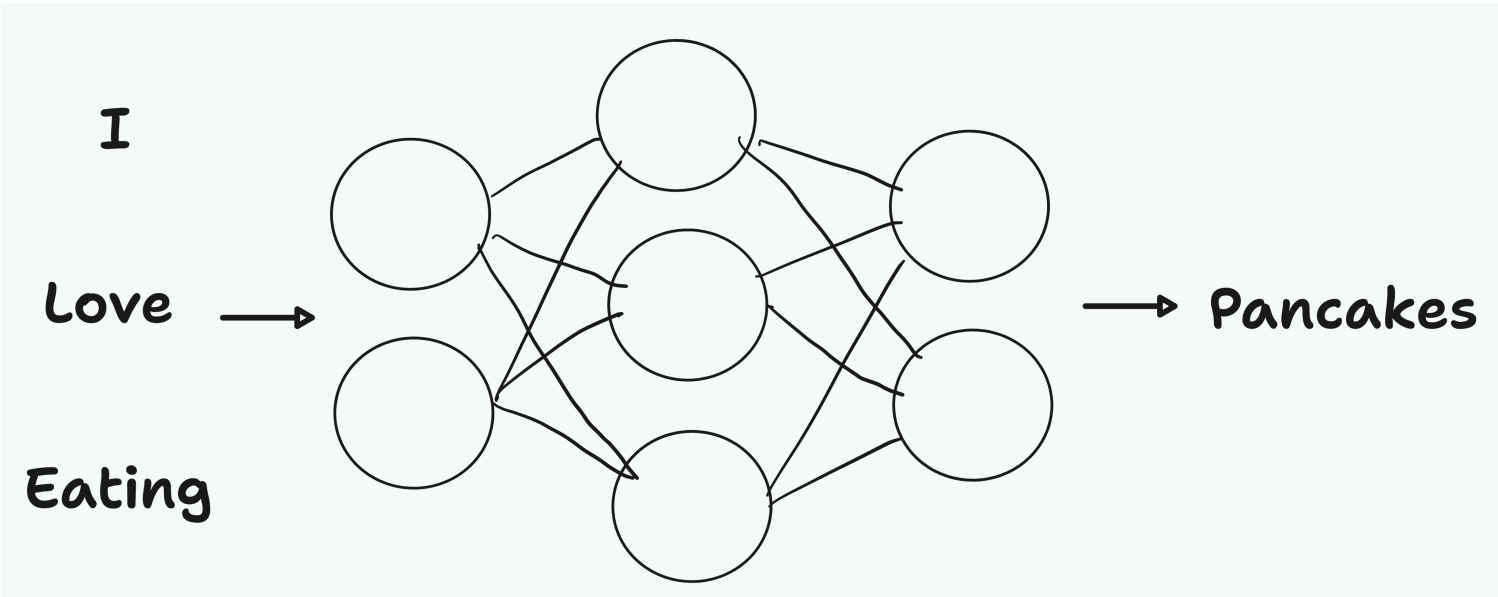


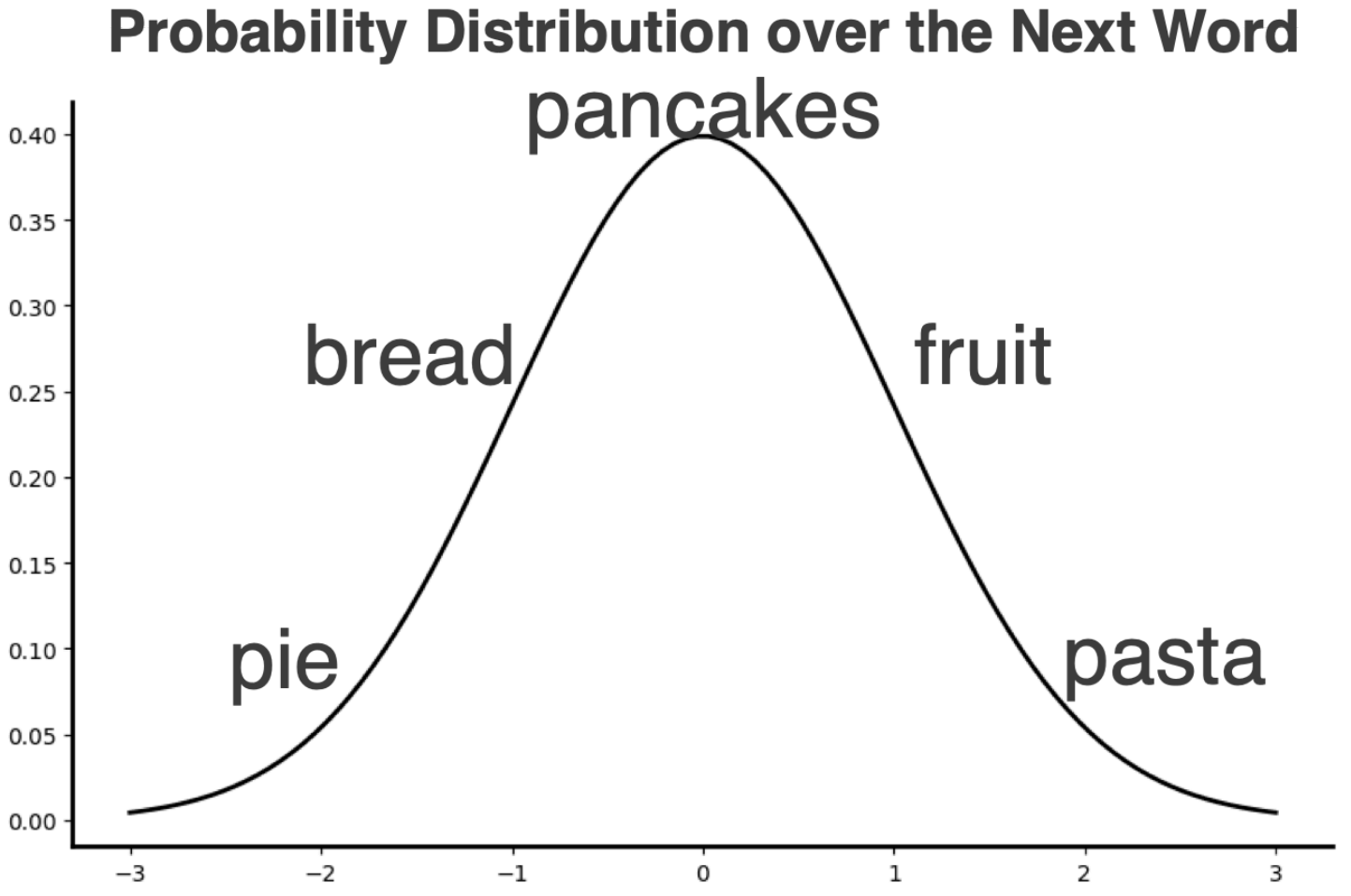
LLMs Predict the Next Word



LLMs Predict the Next Word




I
Love
Eating
Pancakes


Probability Distribution over the Next Word
pancakes
bread
fruit
pie
pasta
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
−3
−2
−1
0
1
2
3

Introduction to Llama3

Llama3 Release

- LLM Released by Meta in April of 2024

Llama3 Release

- LLM Released by Meta in April of 2024
- Open source with a Commercial license (just like Llama2)

Llama3 Release

- LLM Released by Meta in April of 2024
- Open source with a Commercial license (just like Llama2)
- Meta Llama3 Resources

Incredible Performance in 2 sizes

- Llama3 is OPEN SOURCE
- Released in 2 sizes: 8B and 70B parameters.
- Incredible evaluation performances for both 70B and 8B models

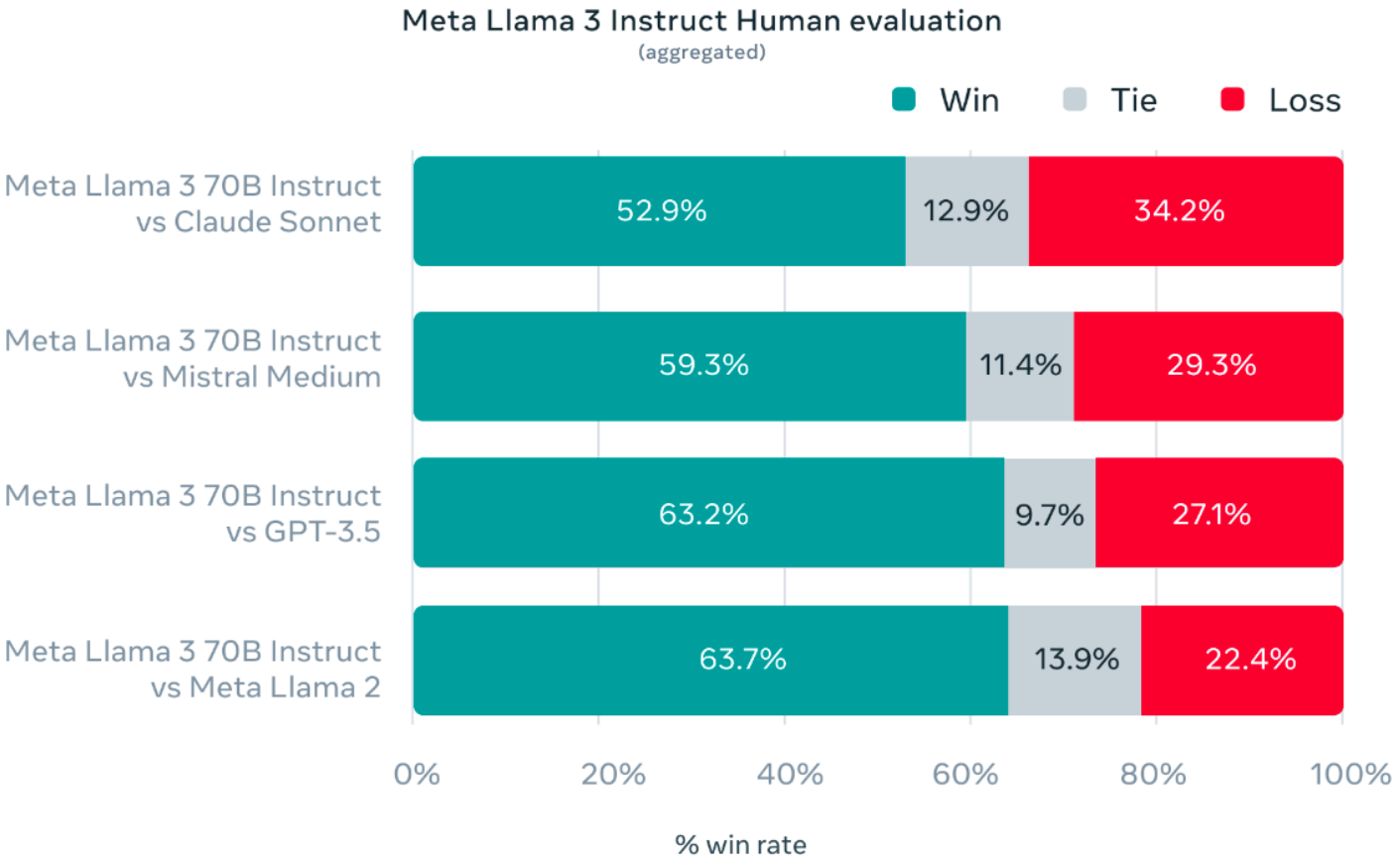
Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

*Please see evaluation details for setting and parameters with which these evaluations are calculated.

Meta AI Llama3 Release Blog Post



[Meta AI Llama3 Release Blog Post]

Meta Llama 3 Pre-trained model performance

	Meta Llama 3 8B	Mistral 7B		Gemma 7B	
		Published	Measured	Published	Measured
MMLU 5-shot	**66.6**	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	**45.9**	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	**61.1**	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 0-shot	**79.1**
DROP 3-shot, F1	**58.4**	--	54.4	--	56.3

	Meta Llama 3 70B	Gemini Pro 1.0	Mixtral 8x22B
		Published	Measured
MMLU 5-shot	**79.5**	71.8	77.7
AGIEval English 3-5-shot	**63.0**	--	61.2
BIG-Bench Hard 3-shot, CoT	**81.3**	75.0	79.2
ARC-Challenge 25-shot	**93.0**	--	90.7
DROP 3-shot, F1	**79.7**	74.1 variable-shot	77.6

*Please see evaluation details for setting and parameters with which these evaluations are calculated.

Llama3 Technical Details

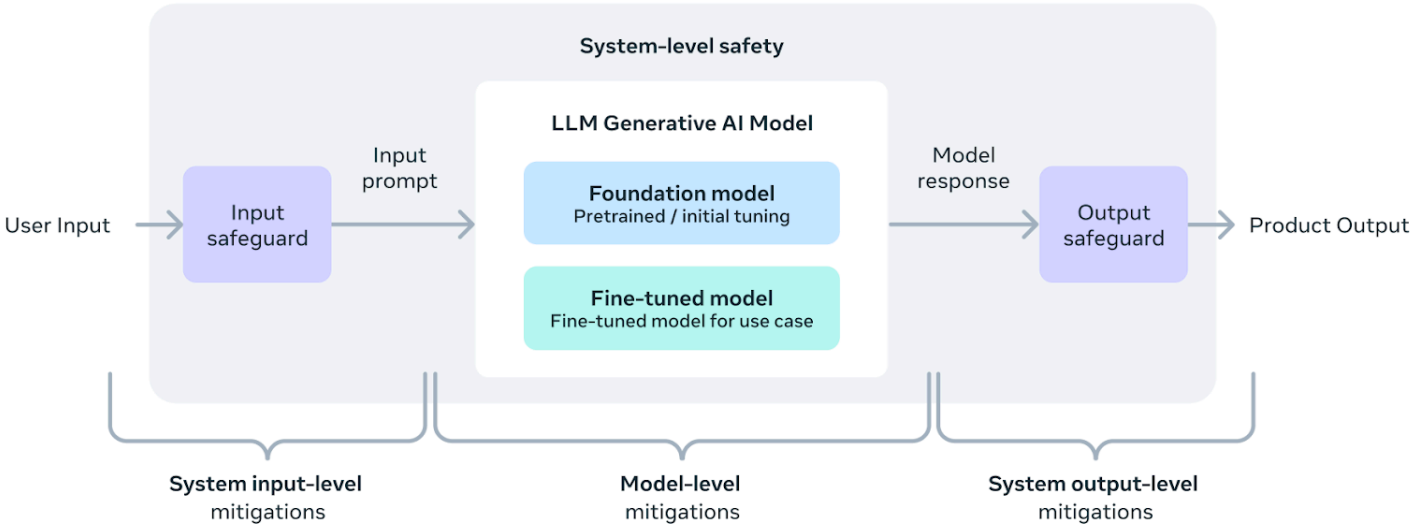
- Data: Trained on over 15 trillion tokens of text data

Llama3 Technical Details

- Data: Trained on over 15 trillion tokens of text data
- Context length: 8192 tokens

Llama3 Technical Details

- Data: Trained on over 15 trillion tokens of text data
- Context length: 8192 tokens
- System level approach for responsible development



[Meta AI Llama3 Release Blog Post]

Notebook Demo - Introduction to Llama3

Query Your Docs Locally with Llama3

- Need for LLMs with access to context-relevant data.

Query Your Docs Locall with Llama3

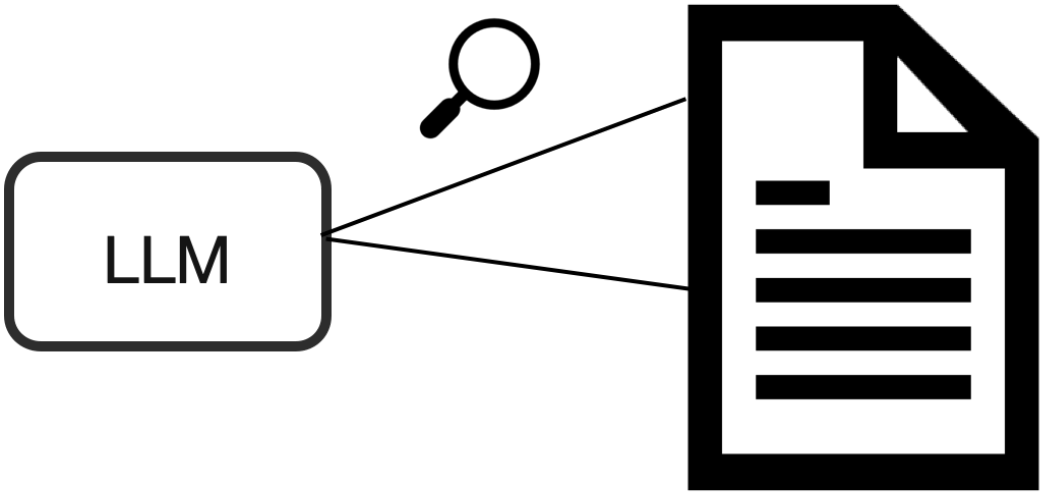
- Privacy concern with closed source LLMs.

Query Your Docs Locall with Llama3

- Solution? Llama3!

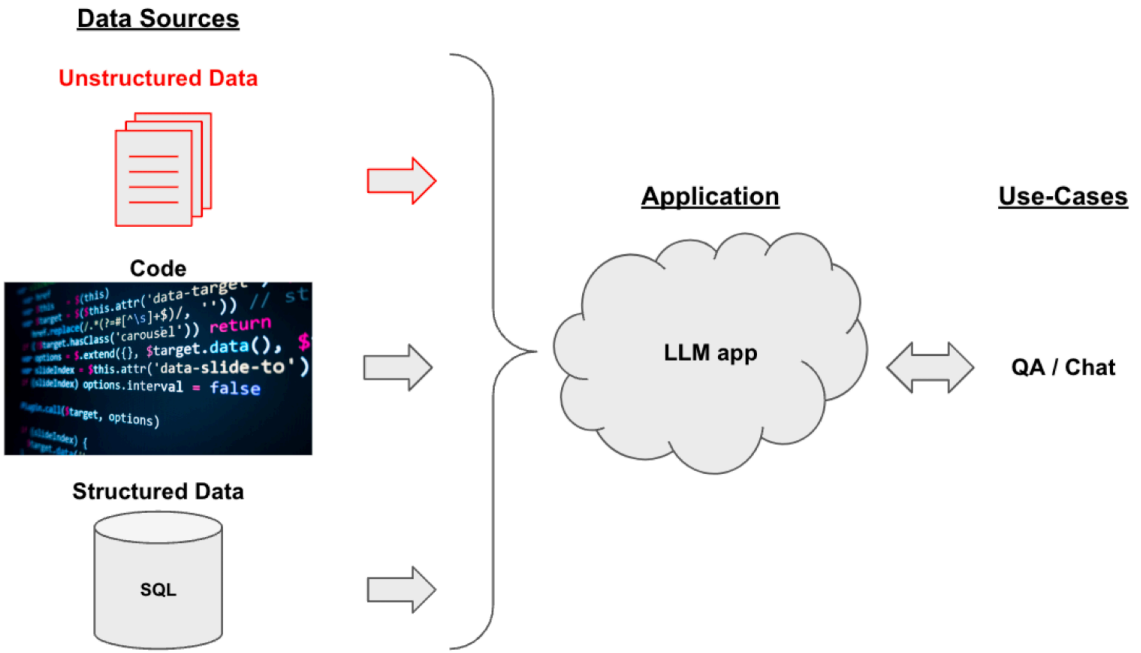
RAG with Llama3

- RAG - Retrieval Augmented Generation



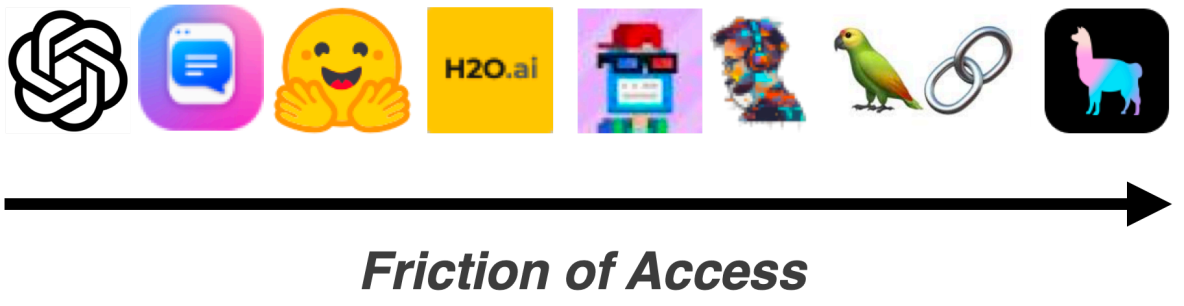
RAG with Llama3

- LLMs have a limited context length



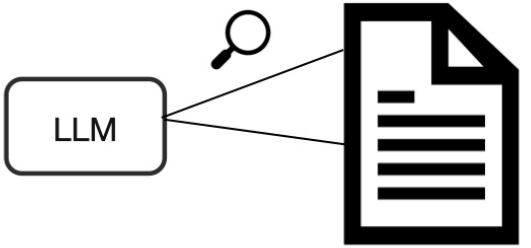
Q&A RAG Tech Friction of Access

- Framework for RAG Systems
- Friction of Access



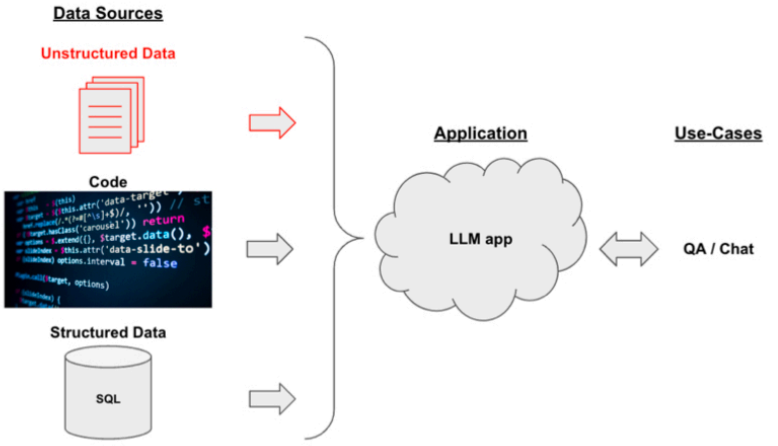
Langchain Docs

RAG - Retrieval Augmented Generation



Langchain Docs

RAG - Retrieval Augmented Generation

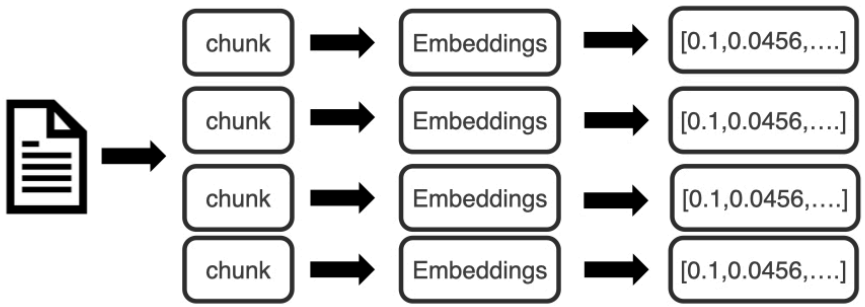


[Langchain Docs]

- RAG - Retrieval Augmented Generation
- LLMs have a limited context length

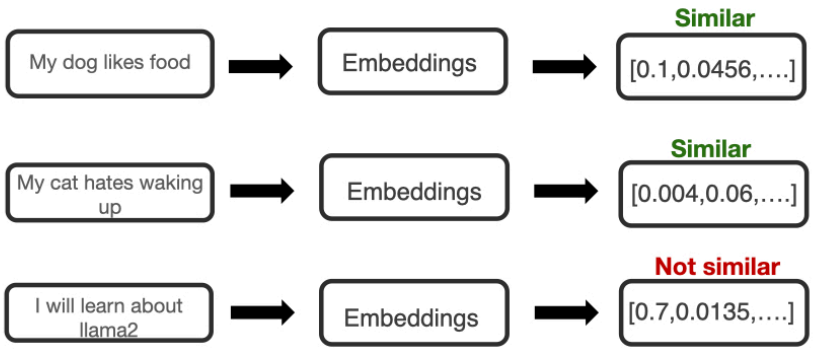
Langchain Docs

- RAG - Retrieval Augmented Generation
- LLMs have a limited context length
- **Embeddings**: capture content and meaning



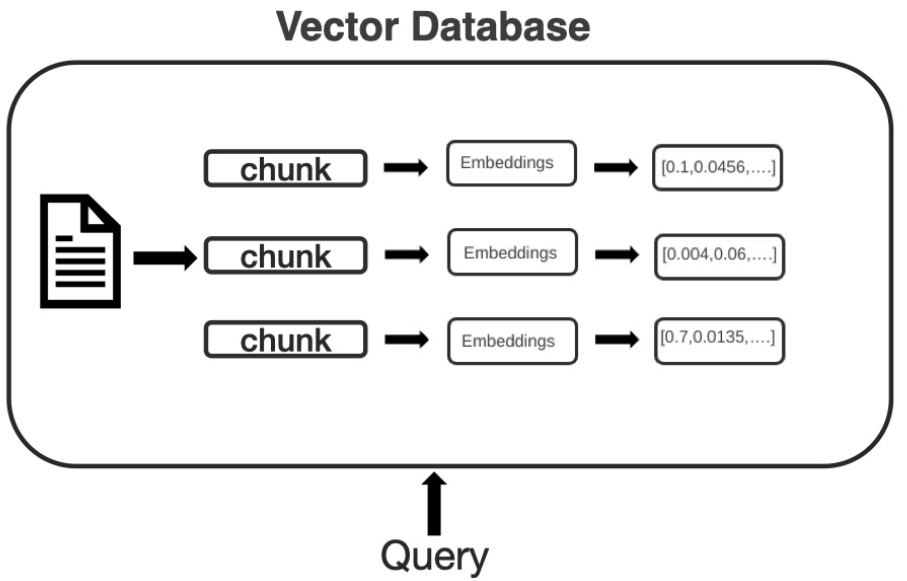
Langchain Docs

- RAG - Retrieval Augmented Generation
- LLMs have a limited context length
- **Embeddings**: capture content and meaning

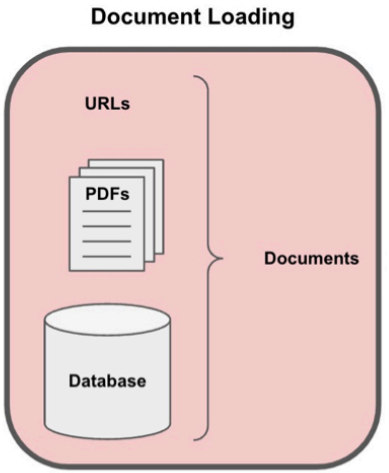


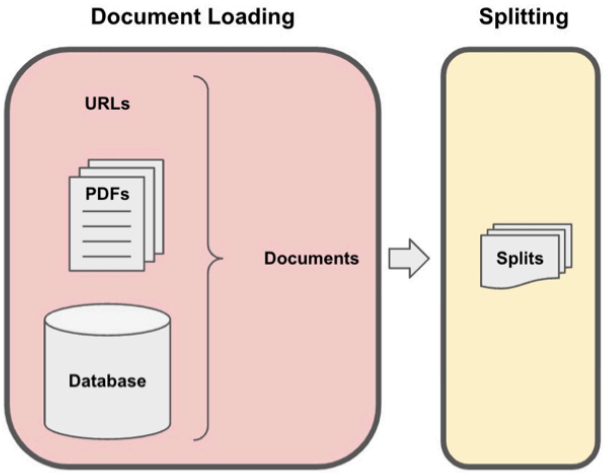
Langchain Docs

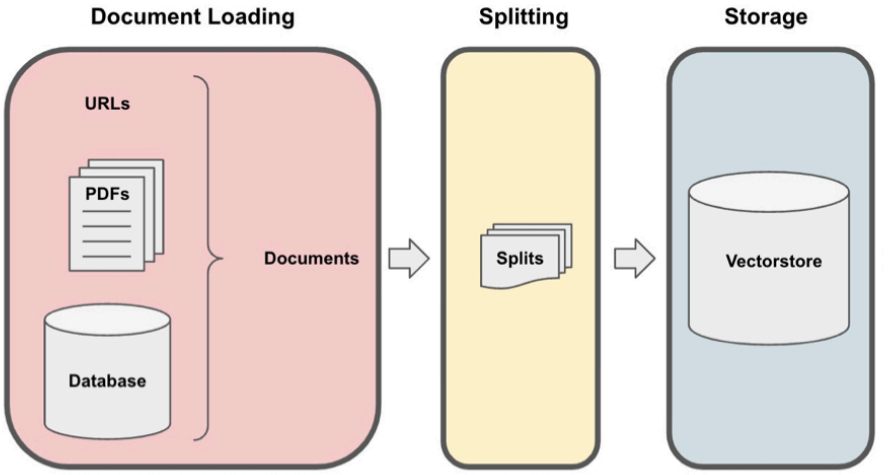
- RAG - Retrieval Augmented Generation
- LLMs have a limited context length
- **Embeddings**: capture content and meaning

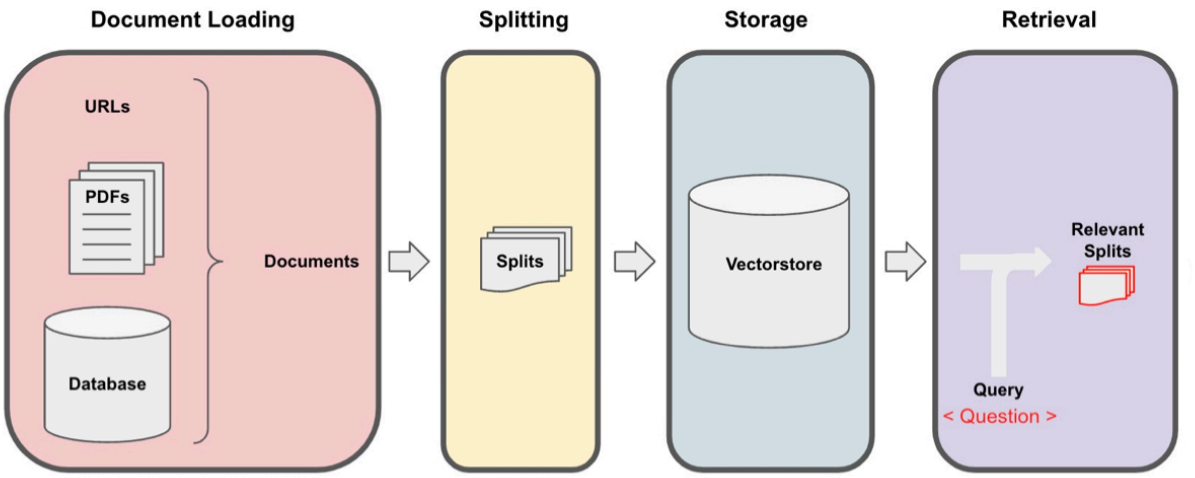


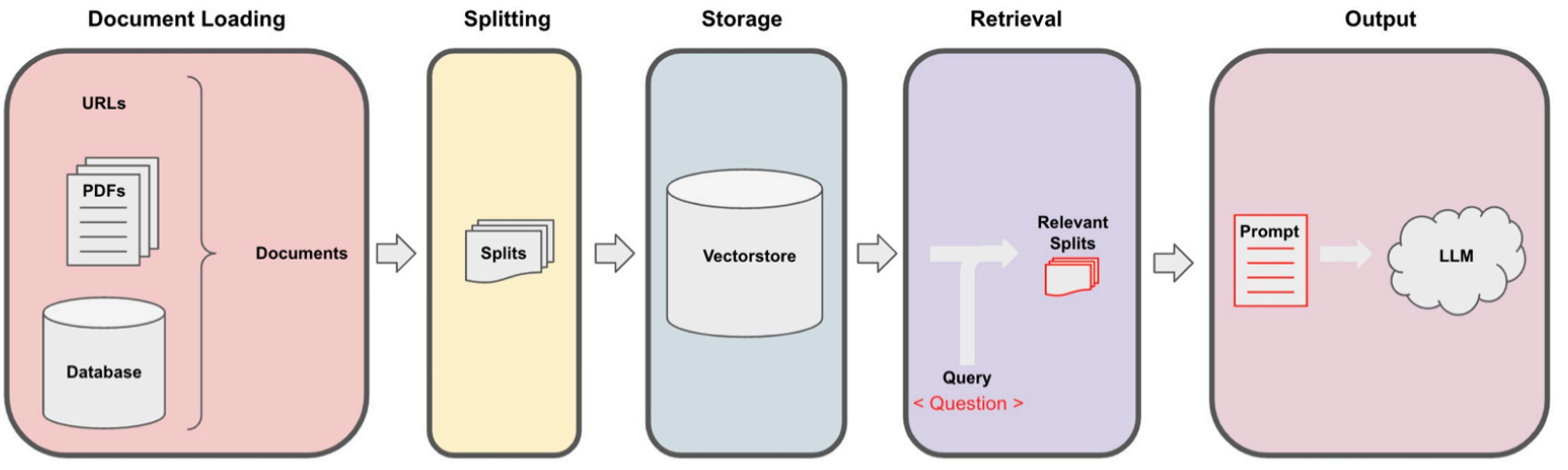
Langchain Docs

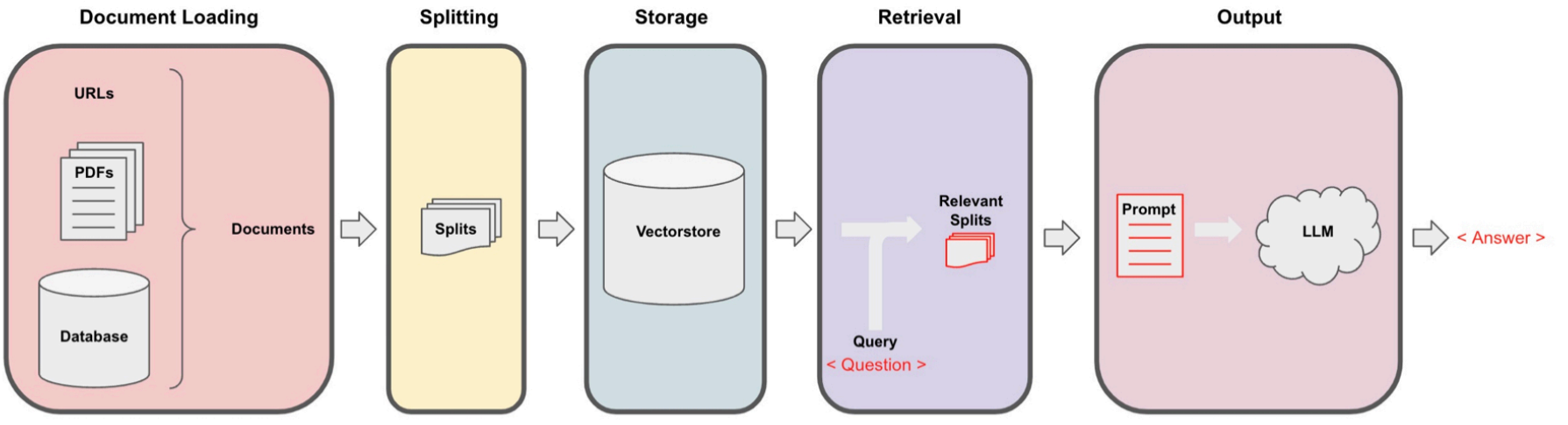

Document Loading
URLs
PDFs
Database
Documents


Document Loading
Splitting
URLs
PDFs
Database
Documents
Splits


Document Loading
Splitting
Storage
URLs
PDFs
Database
Documents
Splits
Vectorstore

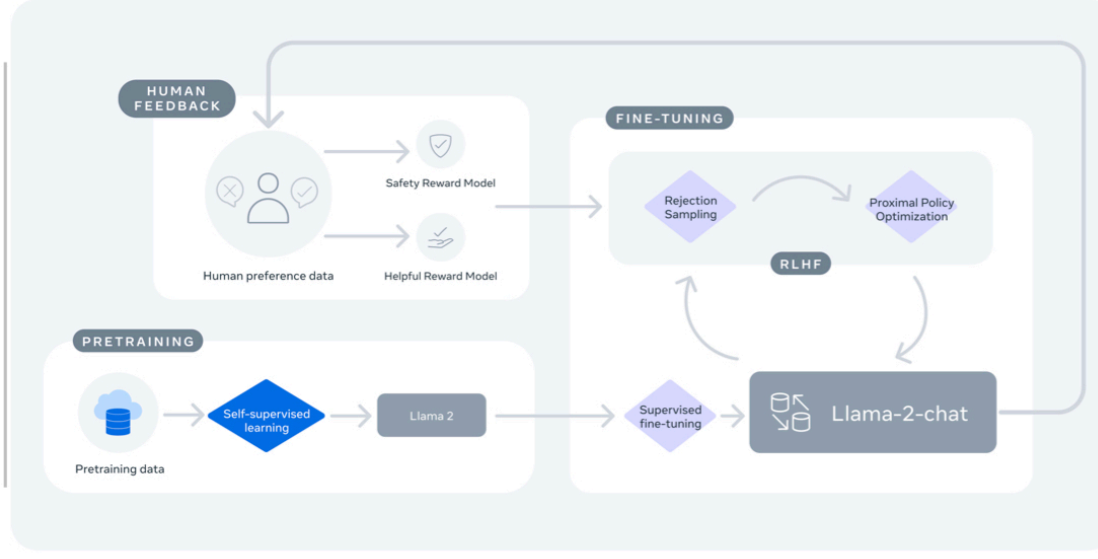

Document Loading
Splitting
Storage
Retrieval
URLs
PDFs
Documents
Database
Splits
Vectorstore
Relevant
Splits
Query
< Question >


Document Loading
URLs
PDFs
Database
Documents
Splitting
Splits
Storage
Vectorstore
Retrieval
Relevant
Splits
Query
< Question >
Output
Prompt
LLM


Document Loading
URLs
PDFs
Documents
Database
Splitting
Splits
Storage
Vectorstore
Retrieval
Relevant Splits
Query
< Question >
Output
Prompt
LLM
< Answer >

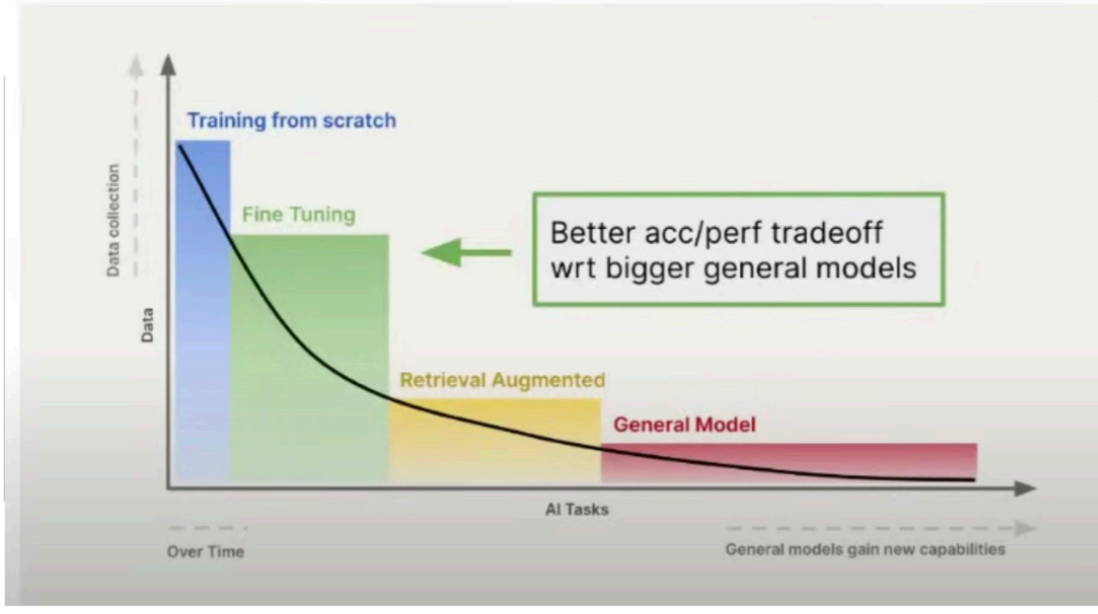
Q&A / Break

What is Fine Tuning?



LLama2 Paper

- What is Fine Tuning?
- Why Fine Tune?



[Efficient Fine-Tuning for Llama-v2-7b on a Single GPU]

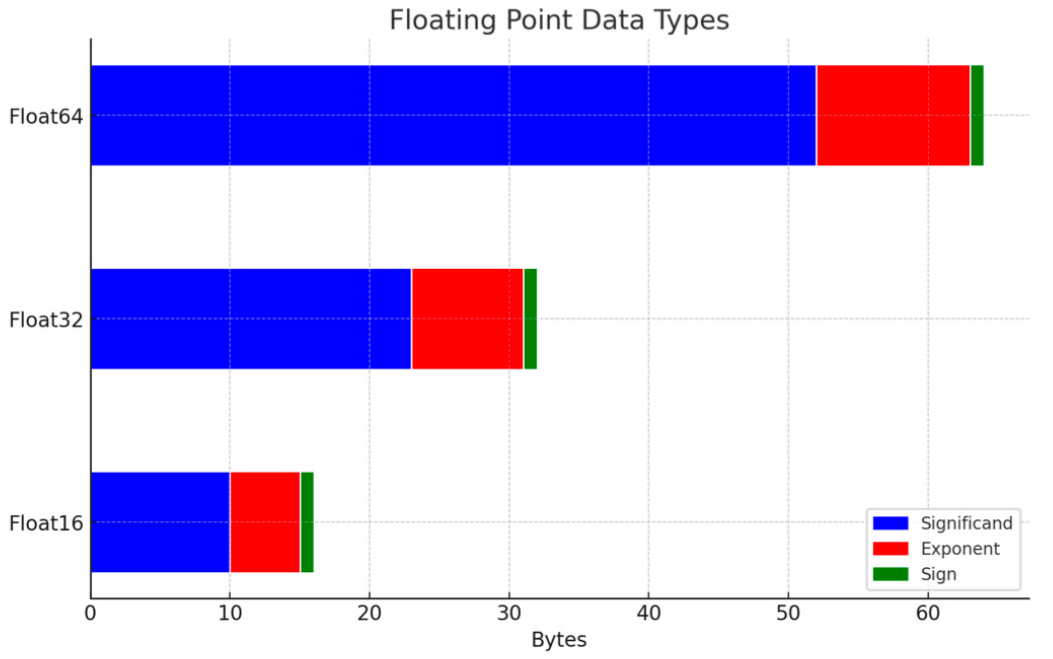
- What is Fine Tuning?
- Why Fine Tune?
- Memory cost of LLMs: parameters, gradients, optimiser states

The Memory Bottleneck: GPU comparison

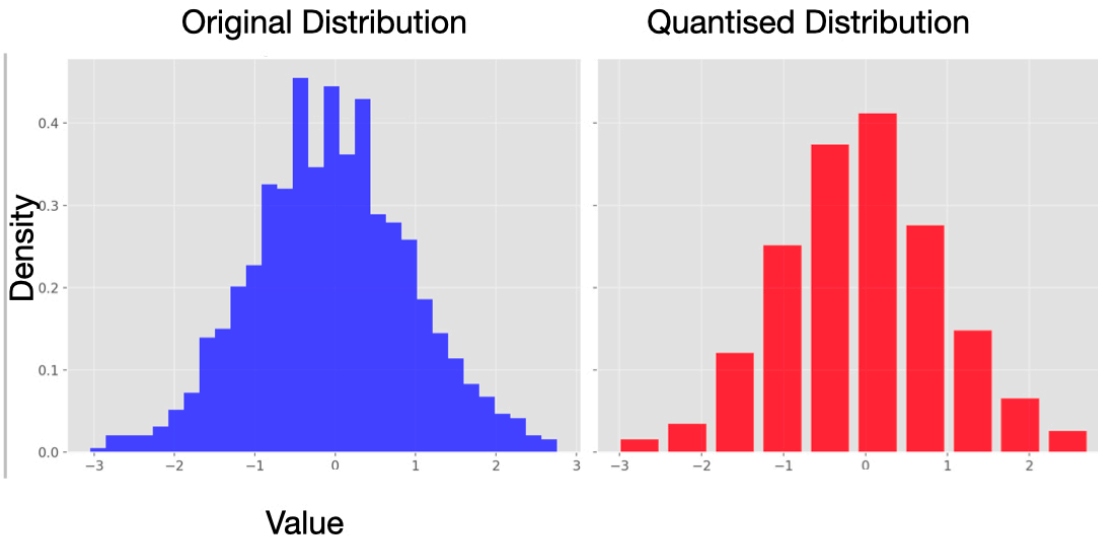
GPU	Tier	$ / hr (AWS)	VRAM (GiB)
H100	Enterprise	12.29	80
A100	Enterprise	5.12	80
V100	Enterprise	3.90	32
A10G	Enterprise	2.03	24
T4	Enterprise	0.98	16
RTX 4080	Consumer	N/A	16

Problem - Loading Params

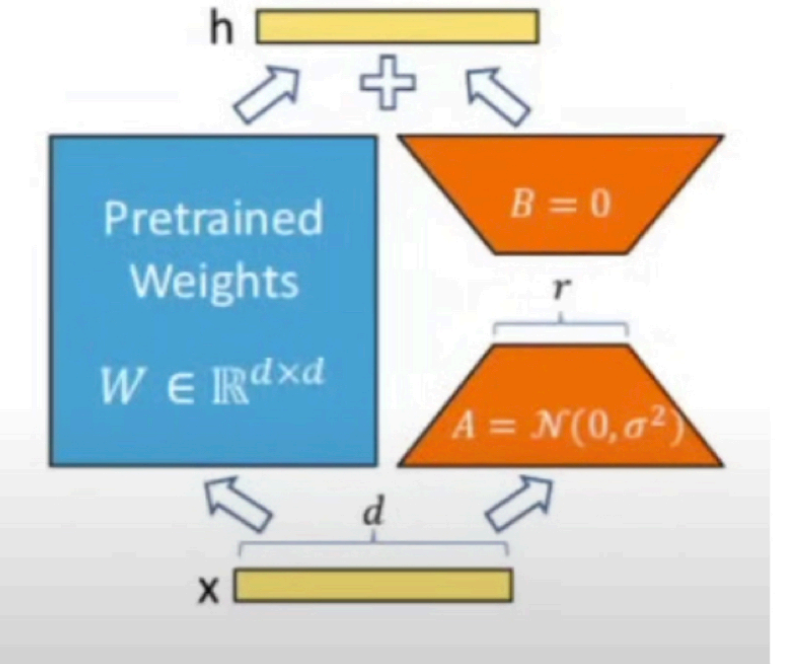
Solution - Half Precision



- Problem - Loading Params
  Solution - Half Precision
- Problem - Loading Gradients
  Solution - Quantization



- Problem - Loading Params
  Solution - Half Precision
- Problem - Loading Gradients
  Solution - Quantization
- Problem - Loading Optimizer States
  Solution - LoRA, QLora



Lora Paper