Getting Started with Llama3.2

Lucas Soares

12-02-2024

Methodology Notes

Methodology Notes

1. Presentation Block

Methodology Notes

1. Presentation Block
2. Notebook Demo

Methodology Notes

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

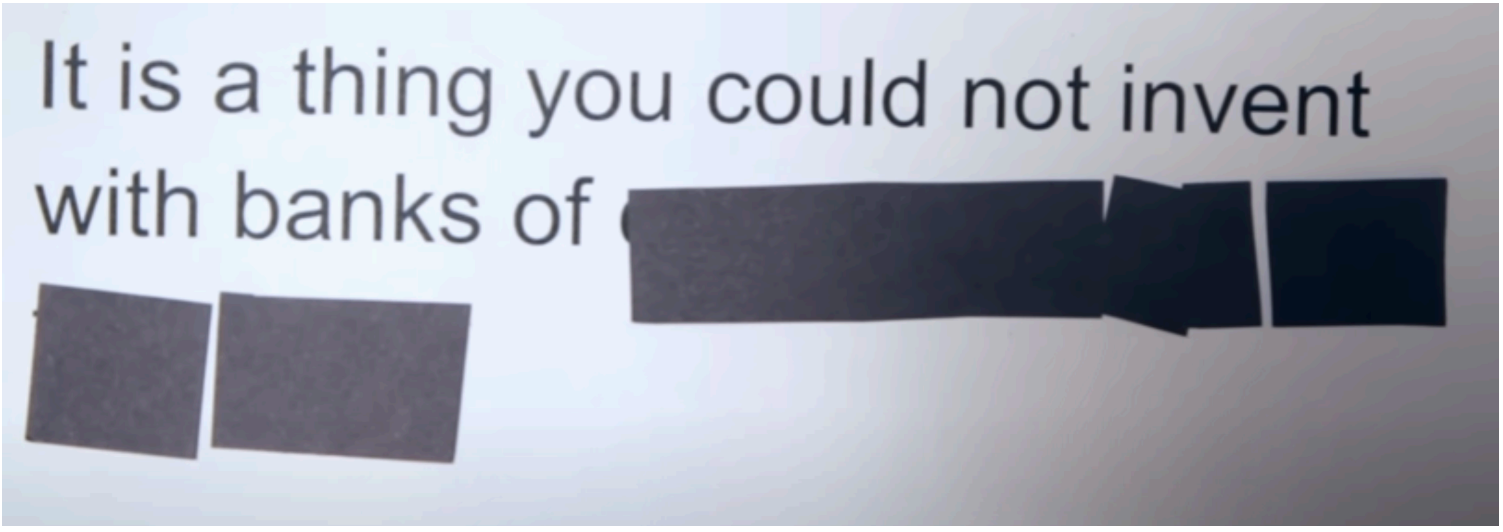
Methodology Notes

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

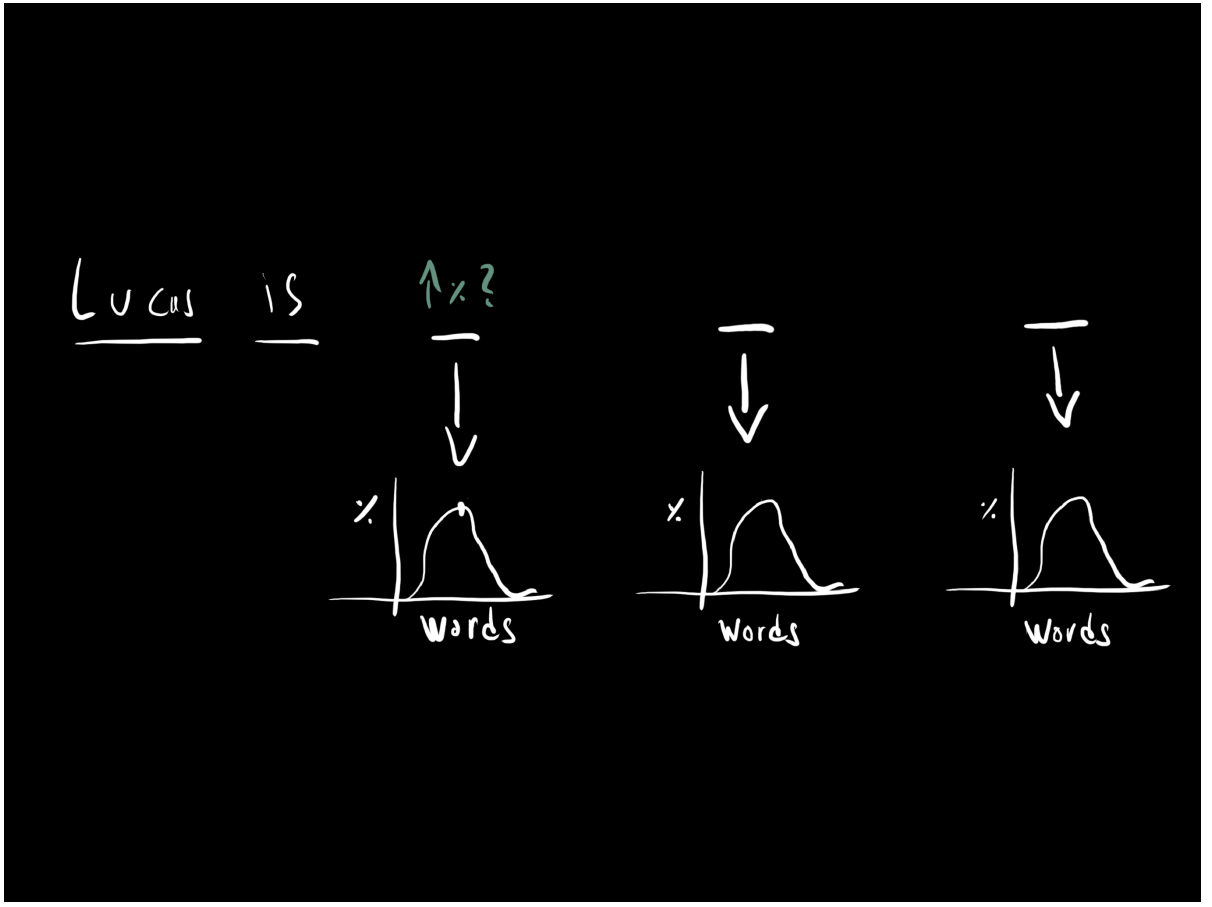
Methodology Notes

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

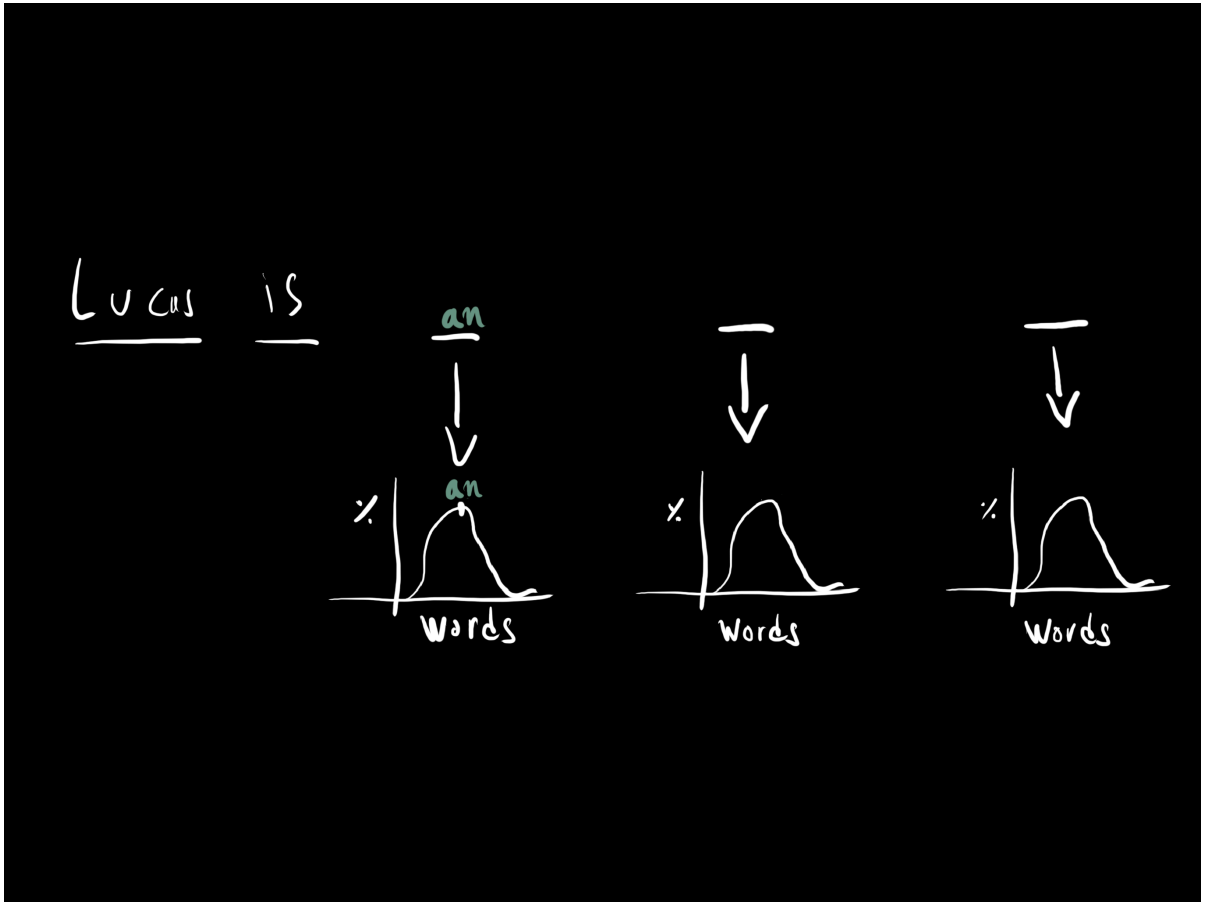
LLMs Predict the Next Word



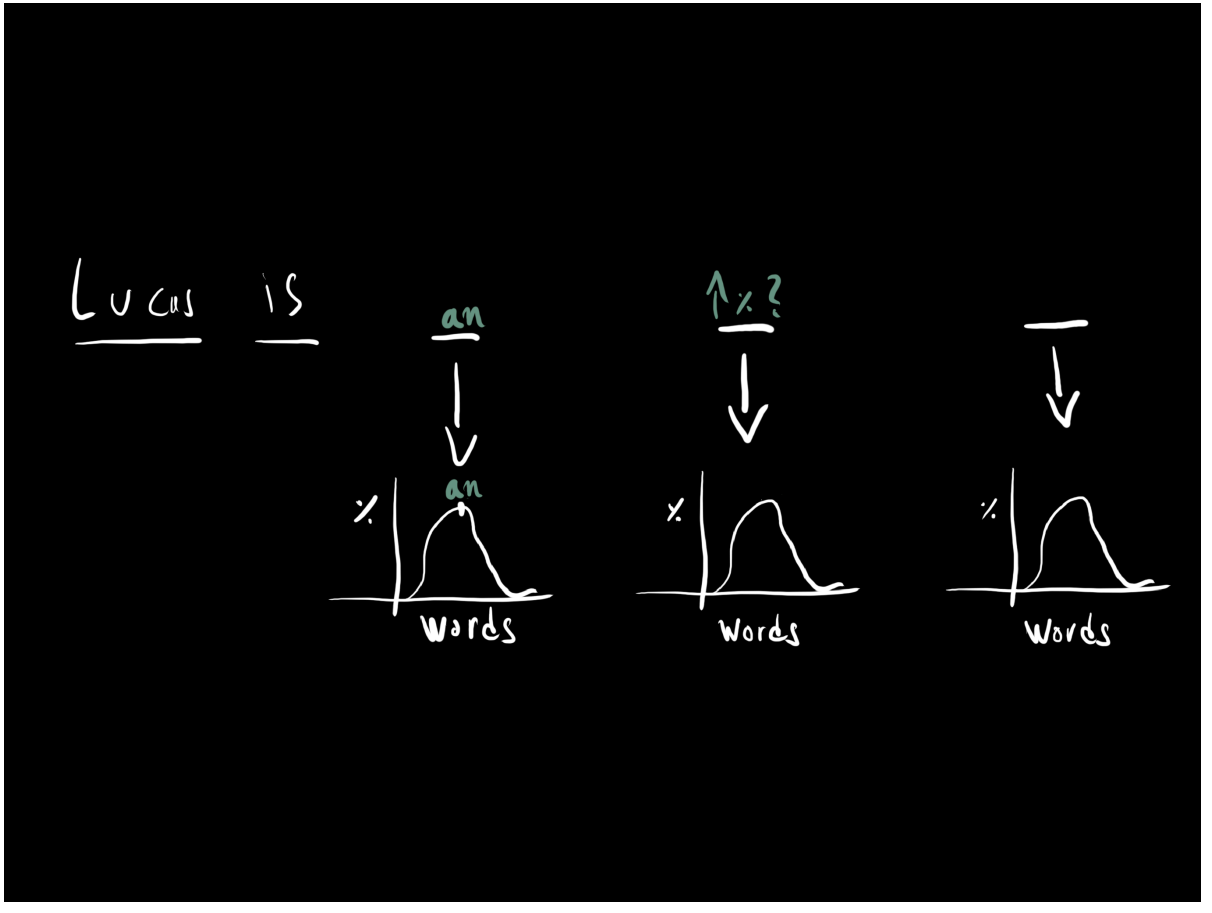
LLMs Predict the Next Word



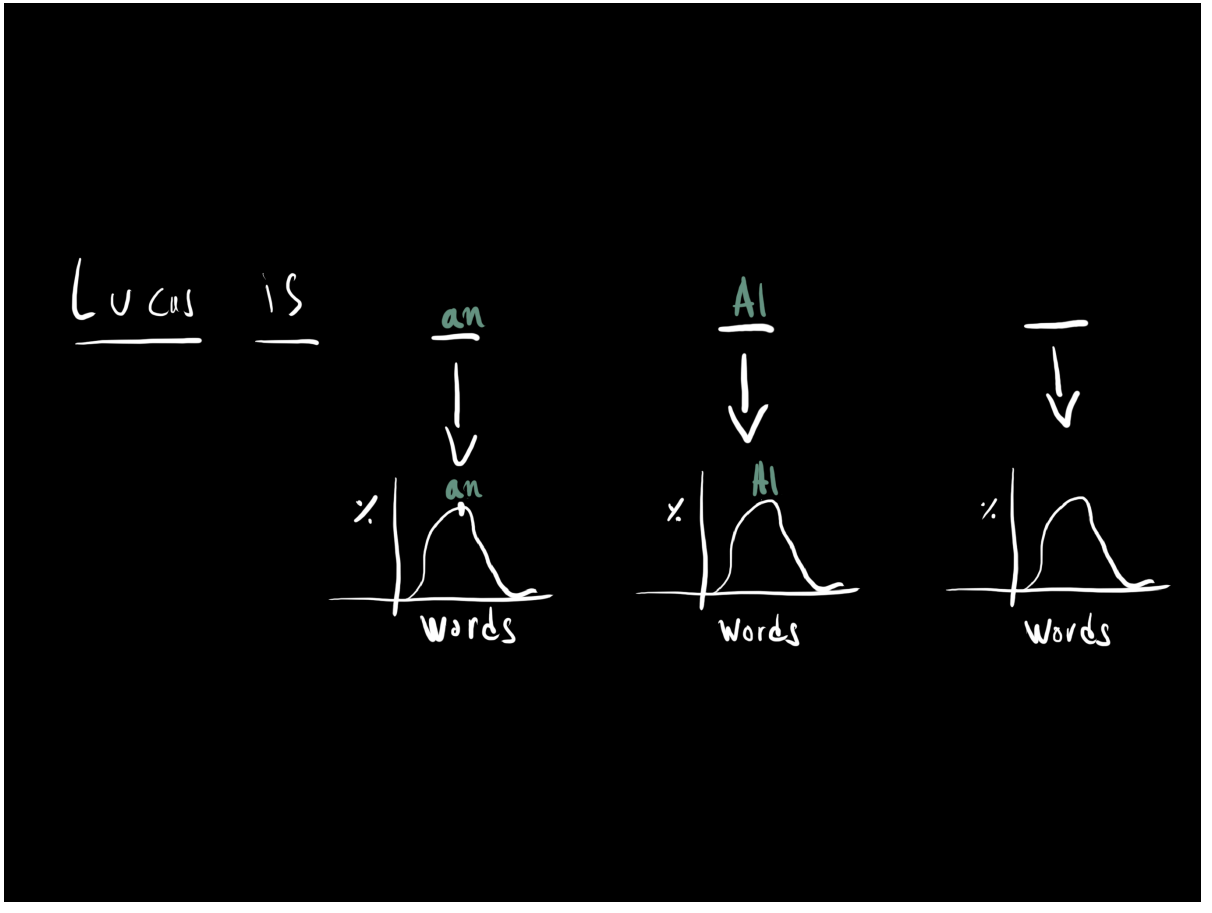
LLMs Predict the Next Word



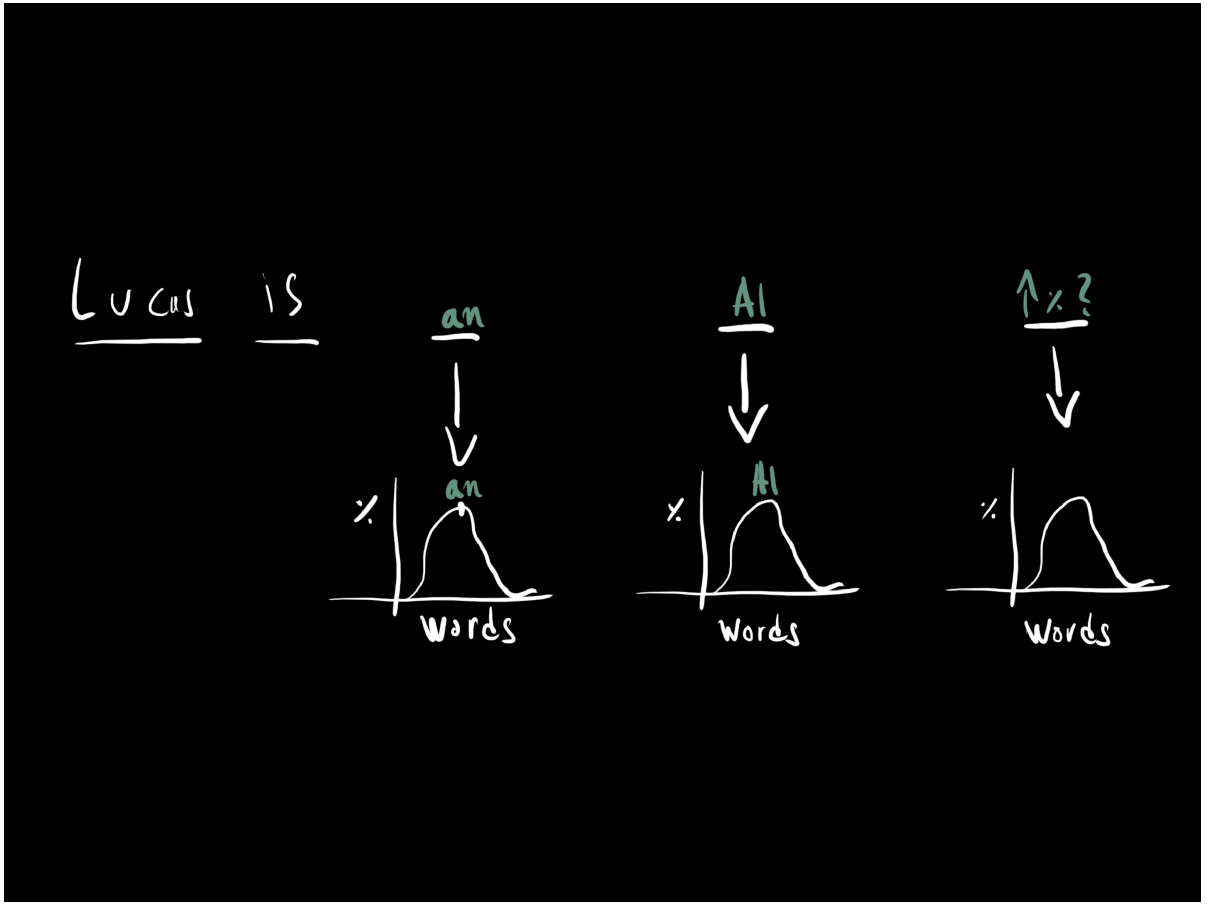
LLMs Predict the Next Word



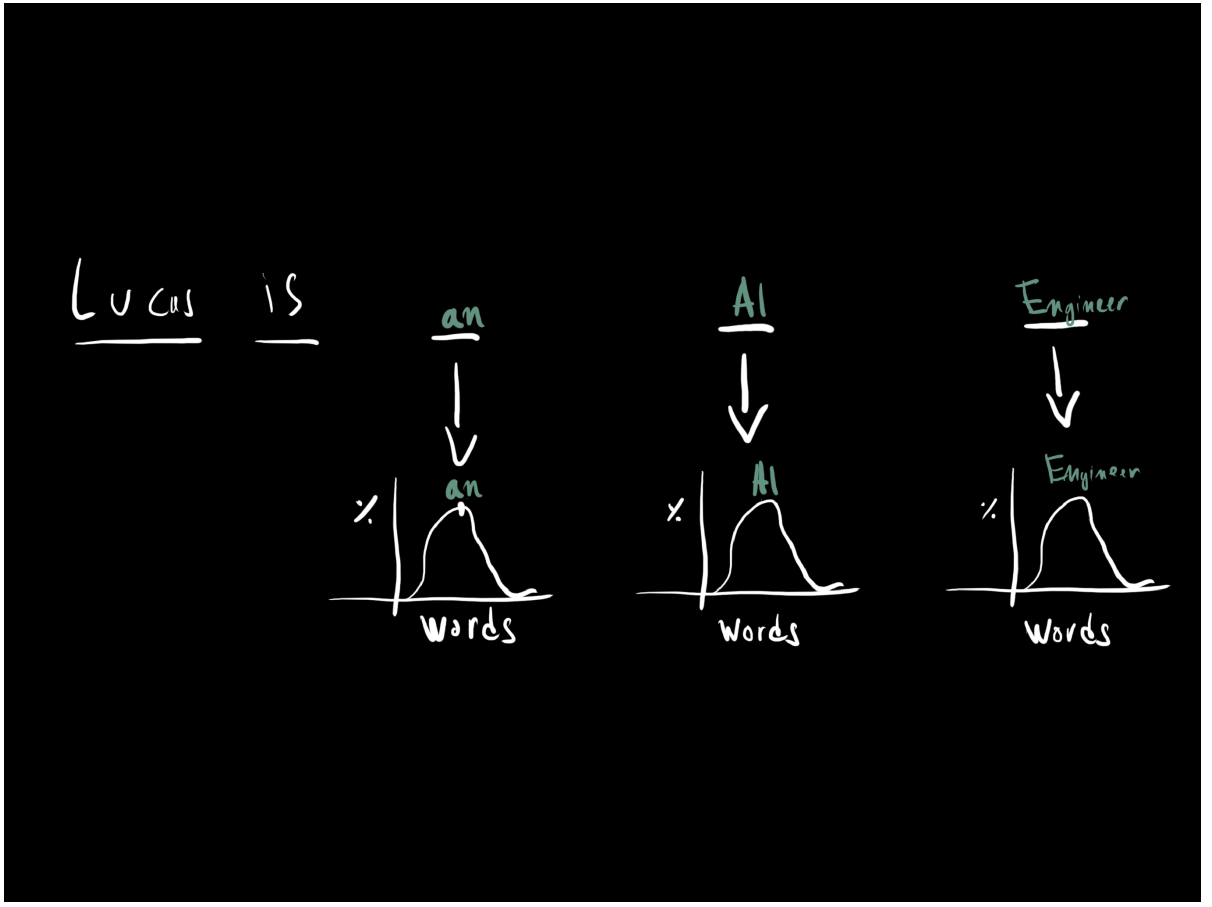
LLMs Predict the Next Word



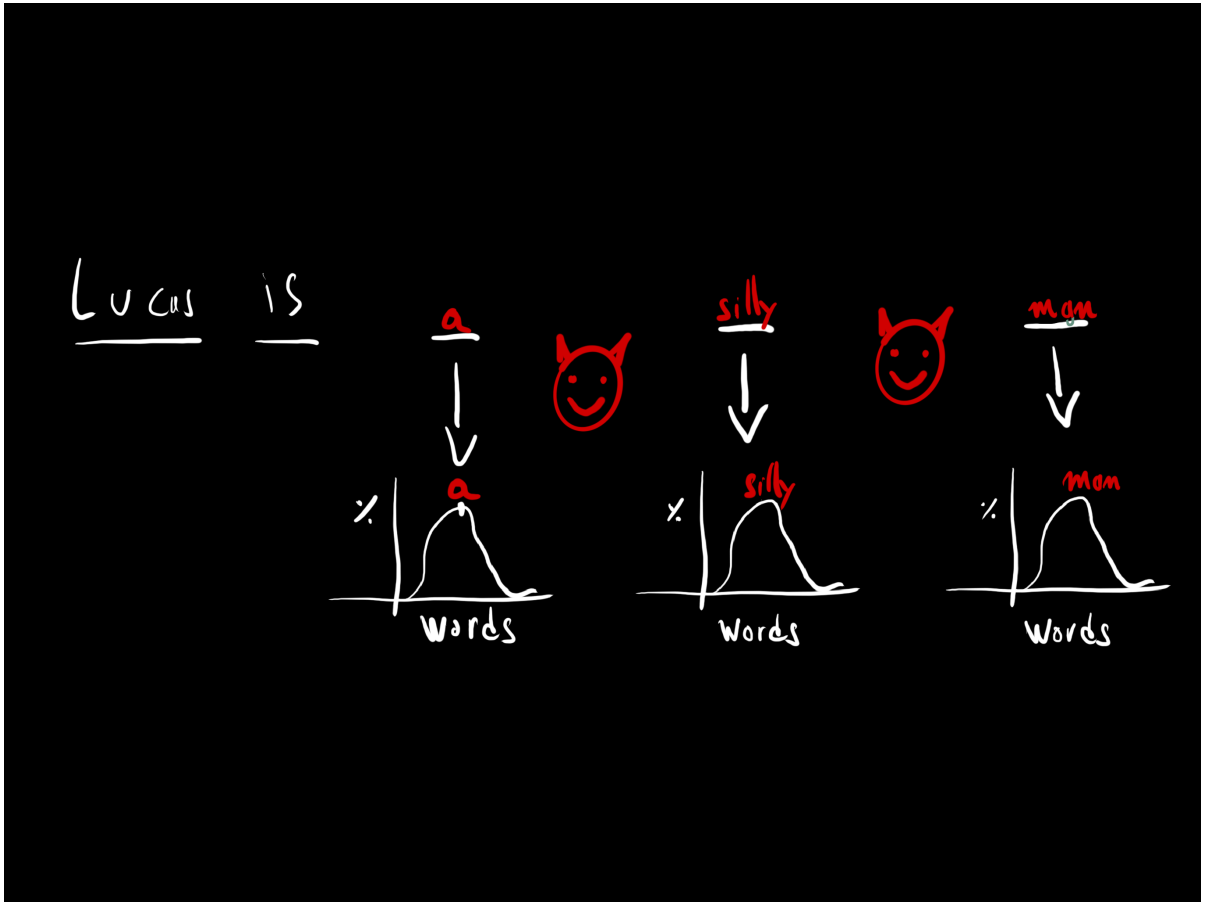
LLMs Predict the Next Word

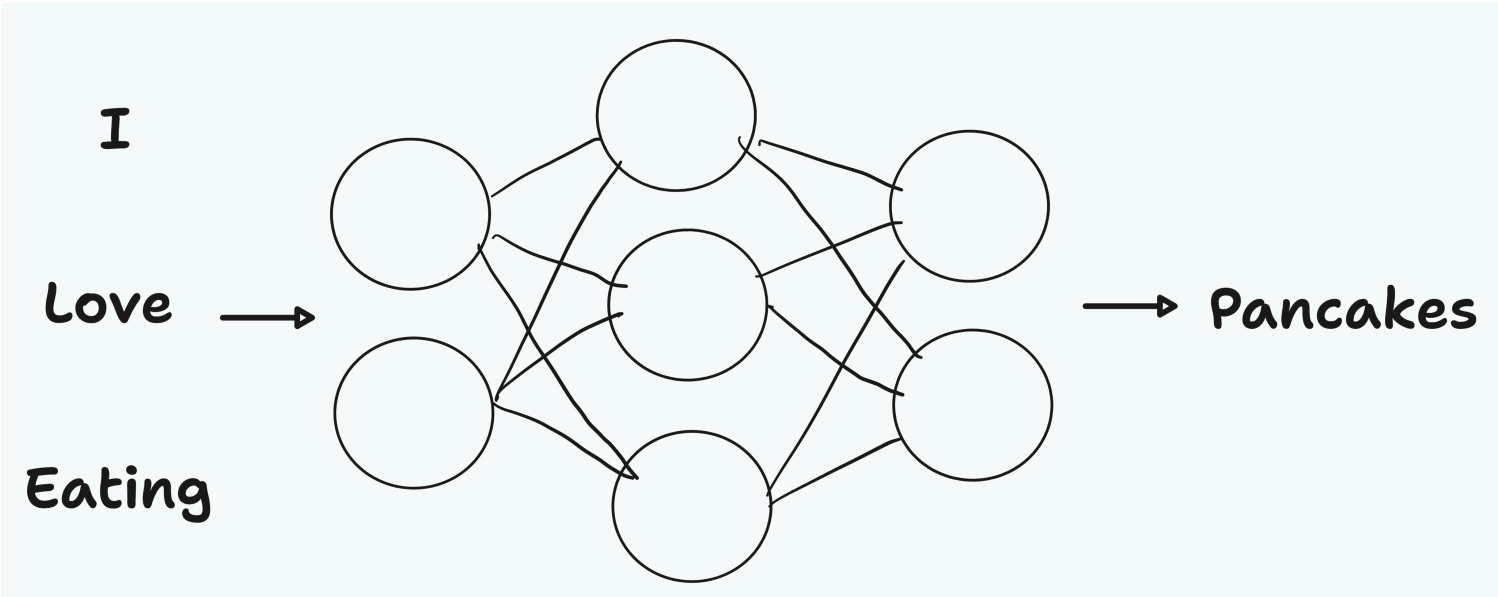


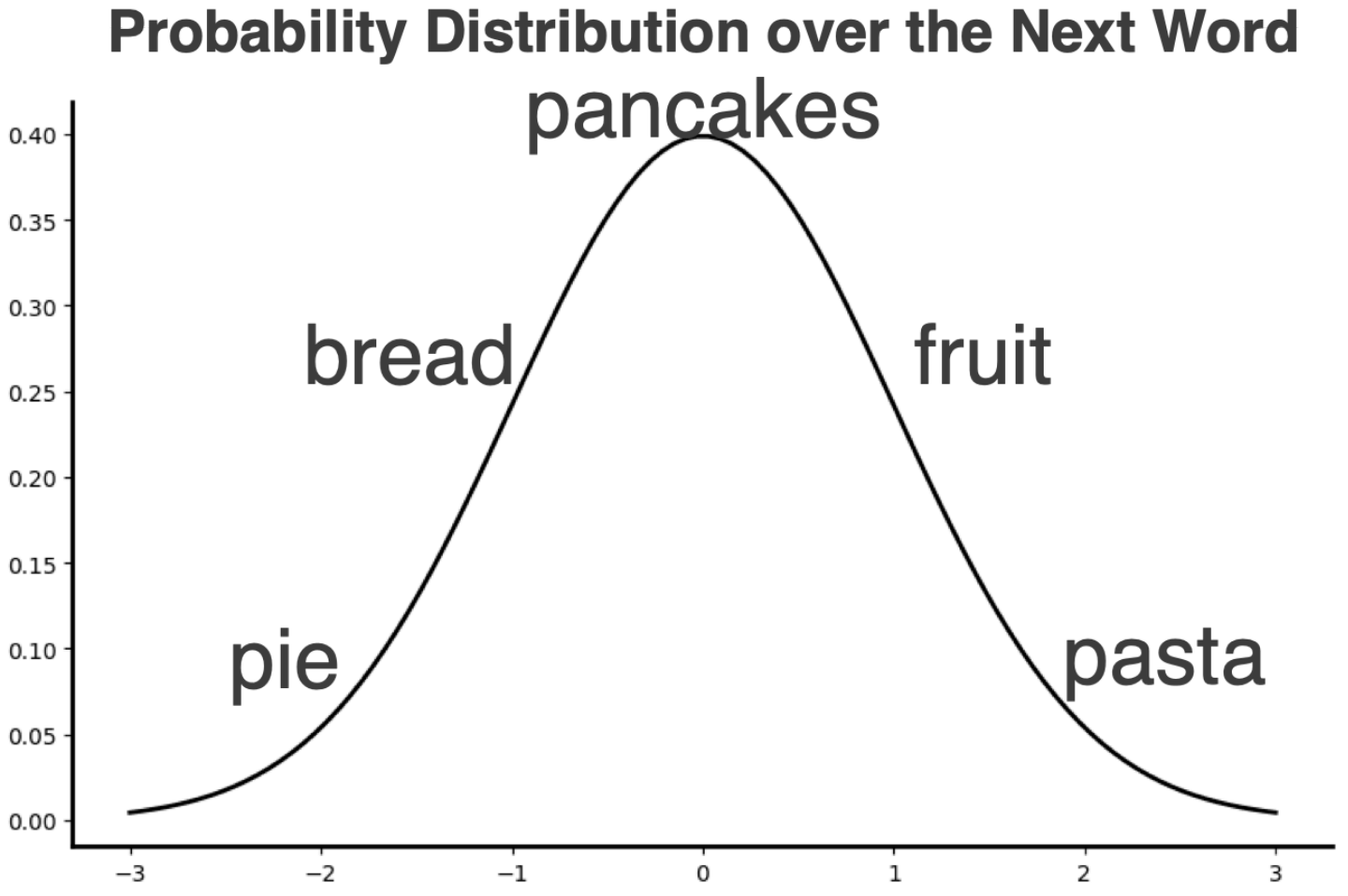
LLMs Predict the Next Word



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 Llama 3.2

Llama 3.2: Revolutionizing edge AI and vision with open, customizable models

September 25, 2024 • 15 minute read

Llama3.2 Release

- LLM Released by Meta in July of 2024

Llama3.2 Release

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

Llama3.2 Release

- LLM Released by Meta in July of 2024
- Open source with a Commercial license (just like Llama2 & Llama3)
- Meta Llama3.2 Resources

Llama3.2 Release

- LLM Released by Meta in July of 2024
- Open source with a Commercial license (just like Llama2 & Llama3)
- Meta Llama3.2 Resources
- Llama 3.2 Repo

Incredible Performance in 4 sizes

Incredible Performance in 4 sizes

- Llama3.2 is OPEN SOURCE

Incredible Performance in 4 sizes

- Llama3.2 is OPEN SOURCE
- Released in 4 sizes: 1B, 3B for text and 11B & 90B for the vision models.

Incredible Performance in 4 sizes

- Llama3.2 is OPEN SOURCE
- Released in 4 sizes: 1B, 3B for text and 11B & 90B for the vision models.
- Incredible evaluation performances for all models.

Lightweight instruction-tuned benchmarks

Category / Benchmark	Llama 3.2 1B	Llama 3.2 3B	Gemma 2 2B IT (measured)	Phi-3.5-mini IT (measured)
General — MMLU (5-shot)	49.3	63.4	57.8	**69.0**
Open-rewrite eval (0-shot, rougeL)	**41.6**	40.1	31.2	34.5
TLDR9+ (test, 1-shot, rougeL)	16.8	**19.0**	13.9	12.8
IFEval	59.5	**77.4**	61.9	59.2
Tool Use — BFCL V2	25.7	**67.0**	27.4	58.4
Nexus	13.5	**34.3**	21.0	26.1
Math — GSM8K (8-shot, CoT)	44.4	77.7	62.5	**86.2**
MATH (0-shot, CoT)	30.6	**48.0**	23.8	44.2
Reasoning — ARC Challenge (0-shot)	59.4	78.6	76.7	**87.4**
GPQA (0-shot)	27.2	**32.8**	27.5	31.9
Hellaswag (0-shot)	41.2	69.8	61.1	**81.4**
Long Context — InfiniteBench/En.MC (128k)	38.0	63.3	—	39.2
InfiniteBench/En.QA (128k)	20.3	19.8	—	11.3
NIH/Multi-needle	75.0	84.7	—	52.7
Multilingual — MGSM (0-shot, CoT)	24.5	**58.2**	40.2	49.8

Let's not forget Llama 3.1!

Llama 3.1 8B & 70B

Category / Benchmark	Llama 3.1 8B	Gemma 2 9B IT	Mistral 7B Instruct	Llama 3.1 70B	Mixtral 8x22B Instruct	GPT 3.5 Turbo
General MMLU (0-shot, CoT)	73.0	72.3 (5-shot, non-CoT)	60.5	86.0	79.9	69.8
MMLU PRO (5-shot, CoT)	48.3	-	36.9	66.4	56.3	49.2
IFEval	80.4	73.6	57.6	87.5	72.7	69.9
Code HumanEval (0-shot)	72.6	54.3	40.2	80.5	75.6	68.0
MBPP EvalPlus (base) (0-shot)	72.8	71.7	49.5	86.0	78.6	82.0
Math GSM8K (8-shot, CoT)	84.5	76.7	53.2	95.1	88.2	81.6
MATH (0-shot, CoT)	51.9	44.3	13.0	68.0	54.1	43.1
Reasoning ARC Challenge (0-shot)	83.4	**87.6**	74.2	94.8	88.7	83.7
GPQA (0-shot, CoT)	32.8	-	28.8	46.7	33.3	30.8
Tool use BFCL	76.1	-	60.4	84.8	-	**85.9**
Nexus	38.5	30.0	24.7	56.7	48.5	37.2
Long context ZeroSCROLLS/QuALITY	81.0	-	-	90.5	-	-
InfiniteBench/En.MC	65.1	-	-	78.2	-	-
NIH/Multi-needle	98.8	-	-	97.5	-	-
Multilingual Multilingual MGSM (0-shot)	68.9	53.2	29.9	86.9	71.1	51.4

Llama 3.1 405B Rivals Closed Source Models

Category / Benchmark	Llama 3.1 405B	Nemotron 4 340B Instruct	GPT-4 (0125)	GPT-4 Omni	Claude 3.5 Sonnet
General MMLU (0-shot, CoT)	88.6	78.7 (non-CoT)	85.4	**88.7**	88.3
MMLU PRO (5-shot, CoT)	73.3	62.7	64.8	74.0	**77.0**
IFEval	**88.6**	85.1	84.3	85.6	88.0
Code HumanEval (0-shot)	89.0	73.2	86.6	90.2	**92.0**
MBPP EvalPlus (base) (0-shot)	88.6	72.8	83.6	87.8	**90.5**
Math GSM8K (8-shot, CoT)	**96.8**	92.3 (0-shot)	94.2	96.1	96.4 (0-shot)
MATH (0-shot, CoT)	73.8	41.1	64.5	**76.6**	71.1
Reasoning ARC Challenge (0-shot)	**96.9**	94.6	96.4	96.7	96.7
GPQA (0-shot, CoT)	51.1	-	41.4	53.6	**59.4**
Tool use BFCL	88.5	86.5	88.3	80.5	**90.2**
Nexus	**58.7**	-	50.3	56.1	45.7
Long context ZeroSCROLLS/QuALITY	**95.2**	-	**95.2**	90.5	90.5
InfiniteBench/En.MC	**83.4**	-	72.1	82.5	-
NIH/Multi-needle	98.1	-	**100.0**	**100.0**	90.8
Multilingual Multilingual MGSM (0-shot)	**91.6**	-	85.9	90.5	**91.6**

Llama3.1 Technical Details

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

Llama3.1 Technical Details

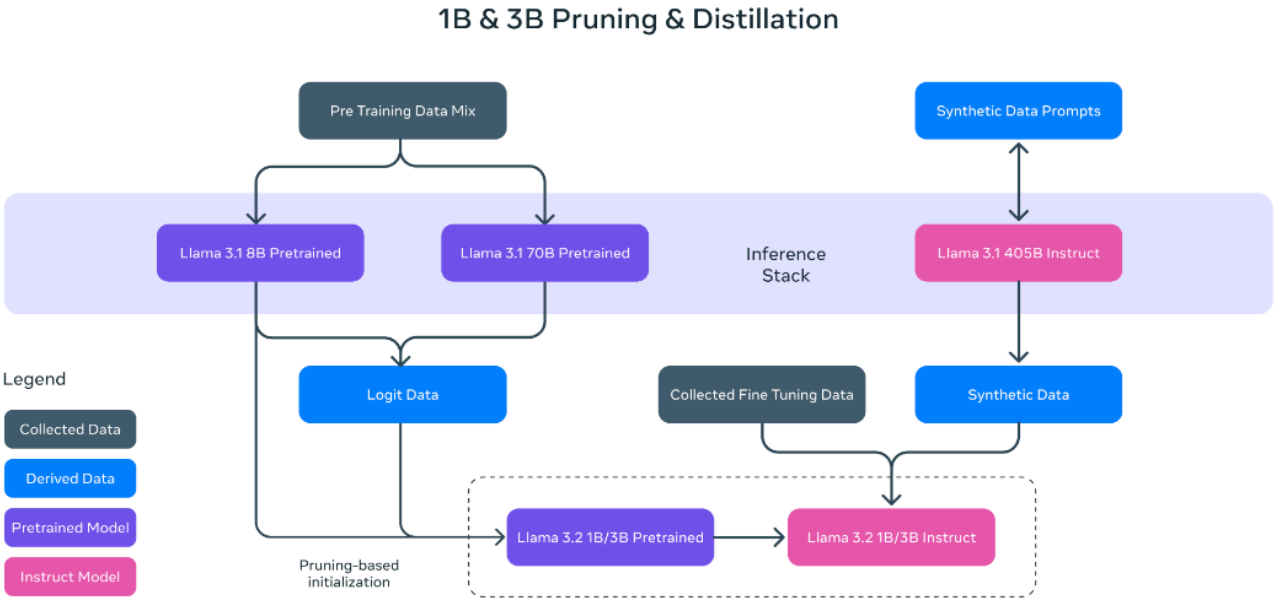
- Data: Trained on over 15 trillion tokens of text data
- Context length: 128k tokens

Llama3.1 Technical Details

- Data: Trained on over 15 trillion tokens of text data
- Context length: 128k tokens
- System level approach for responsible development

Meta AI Llama3.2 Release Blog Post

From Llama 3.1 to Llama 3.2 with Knowledge Distillation & Pruning



[Meta AI Llama3.2 Release Blog Post]

Training Llama 3.2 Vision Models

Training Llama 3.2 Vision Models

1. Architecture Adaptation:
    - Integrated image processing capabilities while preserving text abilities through adapter weights, cross-attention layers, and frozen language model parameters.

Training Llama 3.2 Vision Models

1. Architecture Adaptation:
    - Integrated image processing capabilities while preserving text abilities through adapter weights, cross-attention layers, and frozen language model parameters.
2. Training Pipeline:
    - Built upon Llama 3.1 models through pretraining on image-text pairs followed by domain-specific fine-tuning.

Training Llama 3.2 Vision Models

1. Architecture Adaptation:
    - Integrated image processing capabilities while preserving text abilities through adapter weights, cross-attention layers, and frozen language model parameters.
2. Training Pipeline:
    - Built upon Llama 3.1 models through pretraining on image-text pairs followed by domain-specific fine-tuning.
3. Post-Training Alignment:
    - Enhanced model safety and performance through supervised fine-tuning with rejection sampling and safety-focused data.

Notebook Demo - Introduction to Llama3.2

Query Your Docs Locally with Llama3.2

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

Query Your Docs Locall with Llama3.2

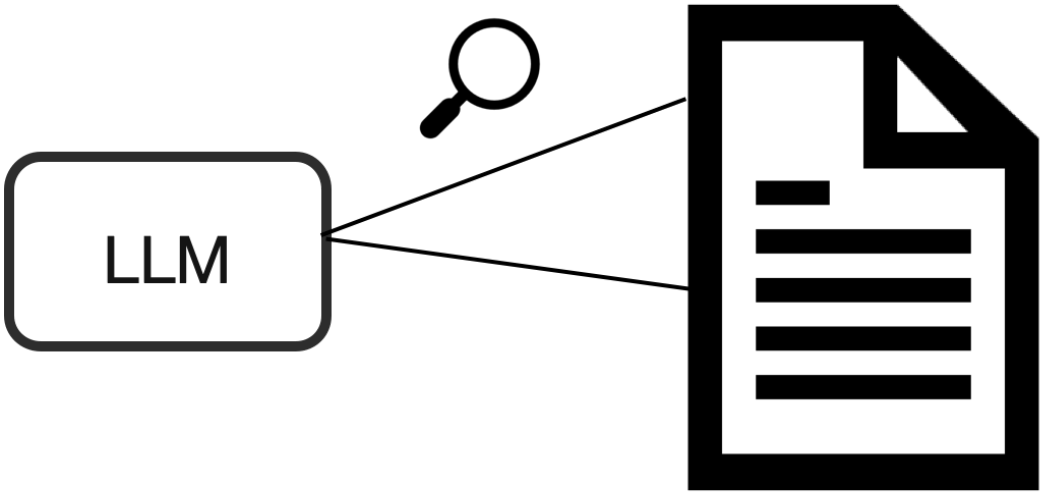
- Privacy concern with closed source LLMs.

Query Your Docs Locall with Llama3.2

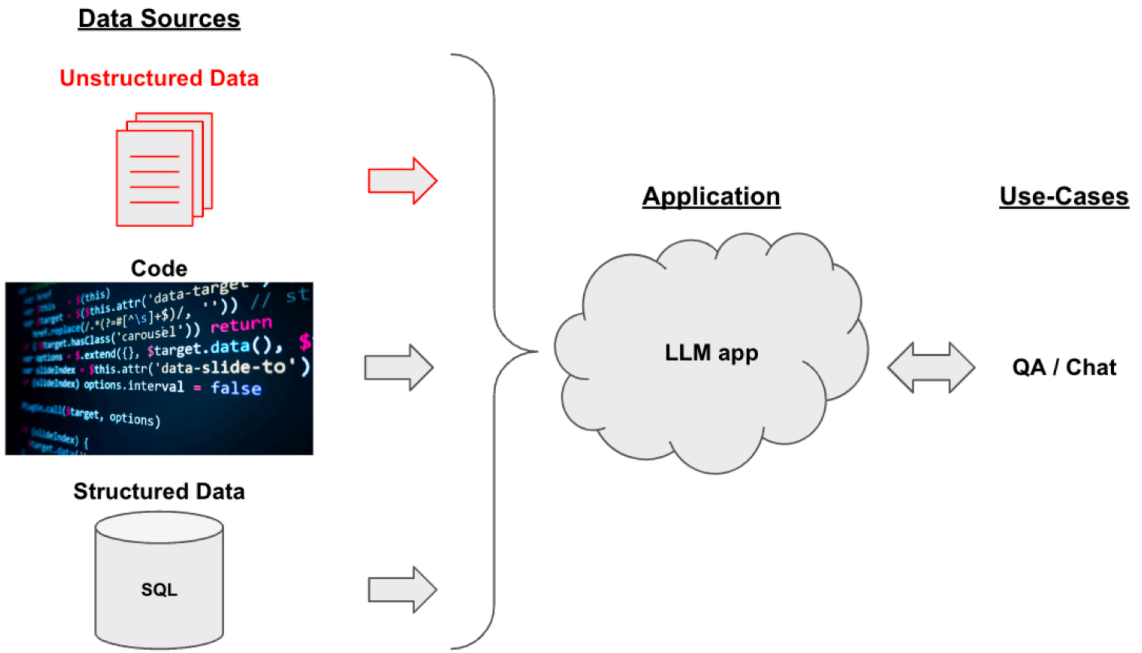
- Solution? Llama3.2!

RAG with Llama3.2

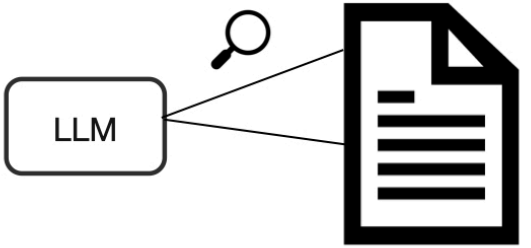
- RAG - Retrieval Augmented Generation



RAG with Llama3.2

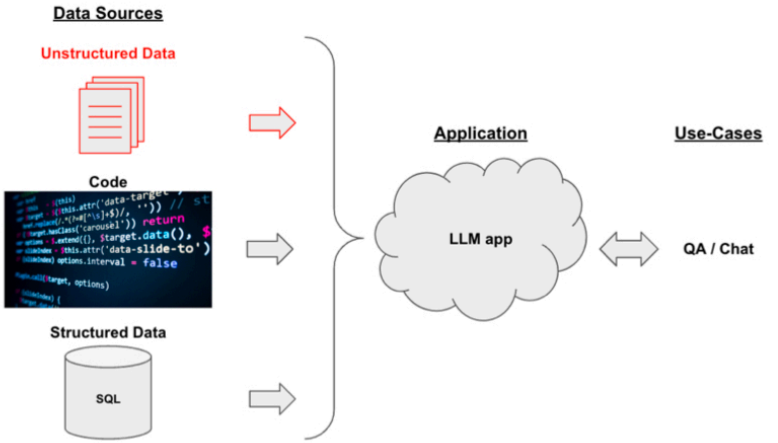


RAG - Retrieval Augmented Generation



Langchain Docs

RAG - Retrieval Augmented Generation

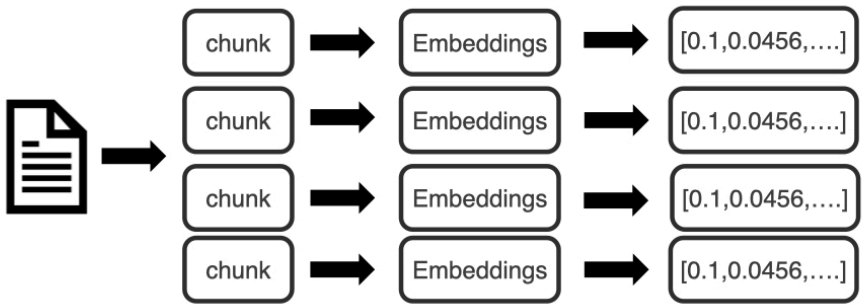


[Langchain Docs]

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

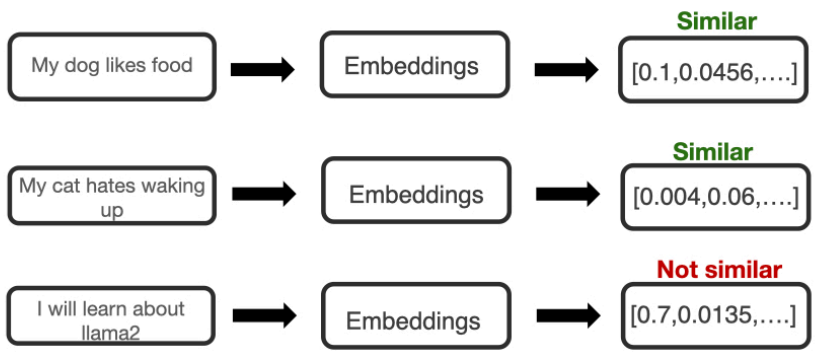
Langchain Docs

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



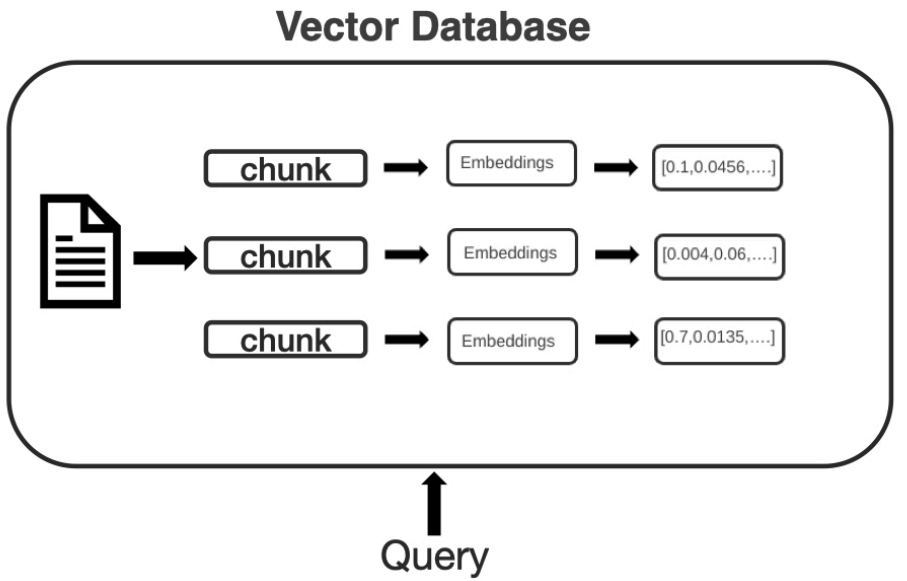
Langchain Docs

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

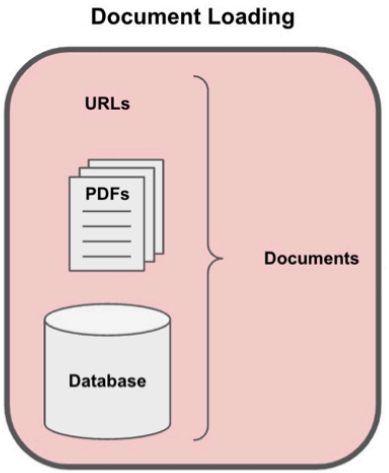


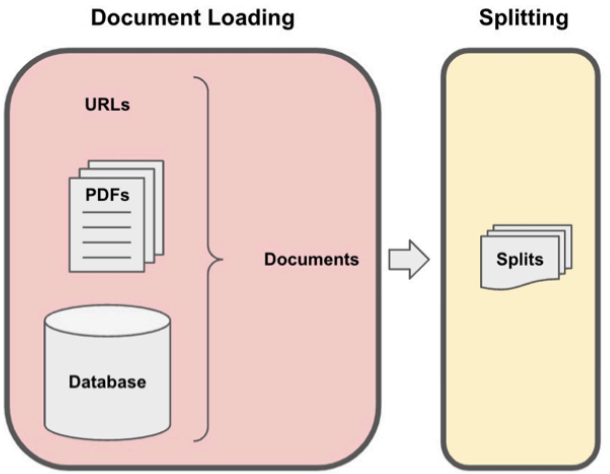
Langchain Docs

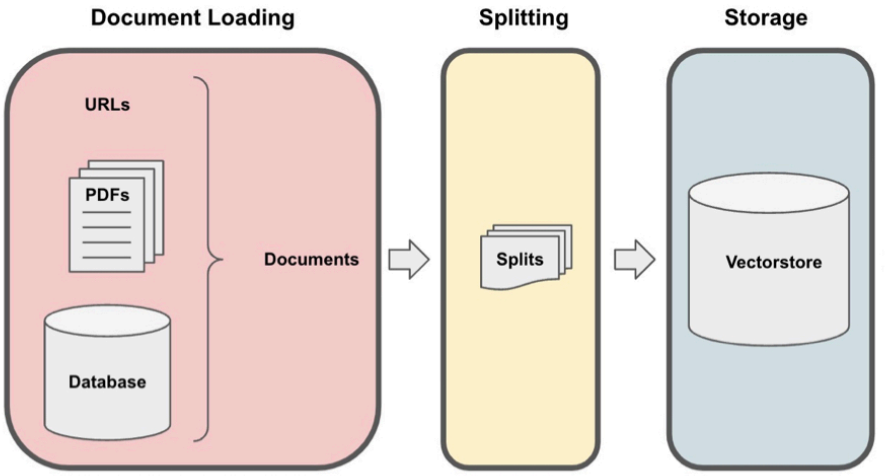
- RAG - Retrieval Augmented Generation
- LLMs have a limited context length
- **Embeddings**: capture content and meaning

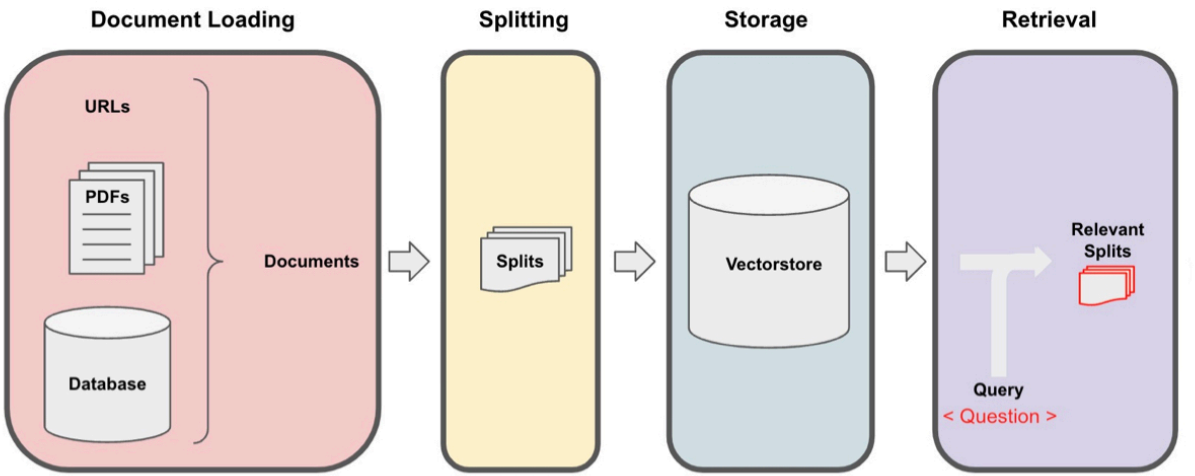


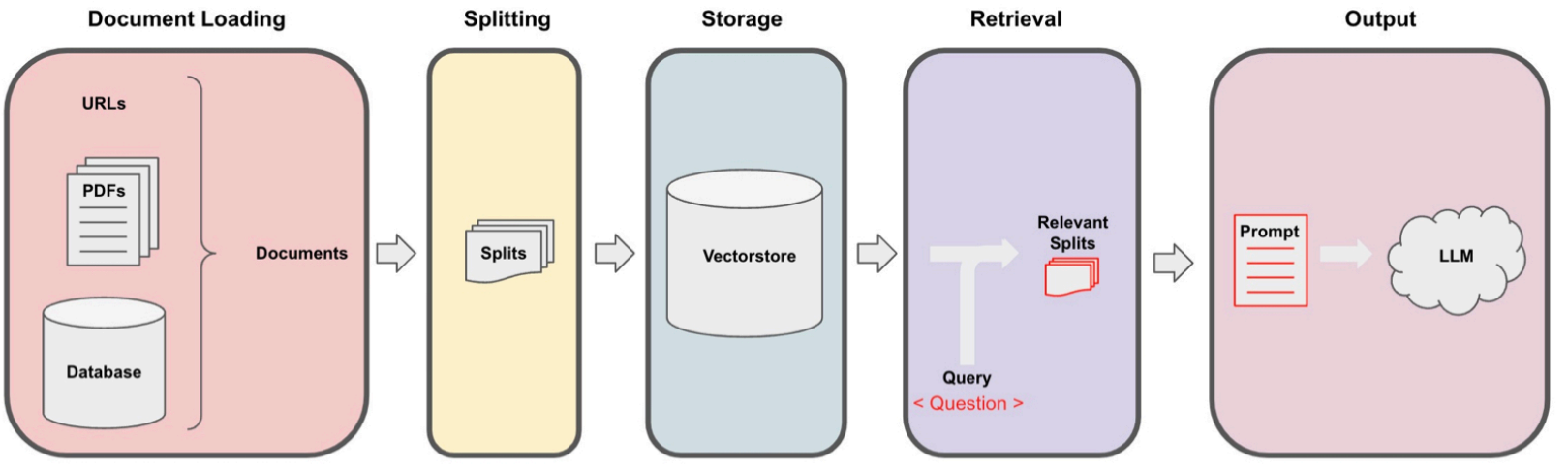
Langchain Docs

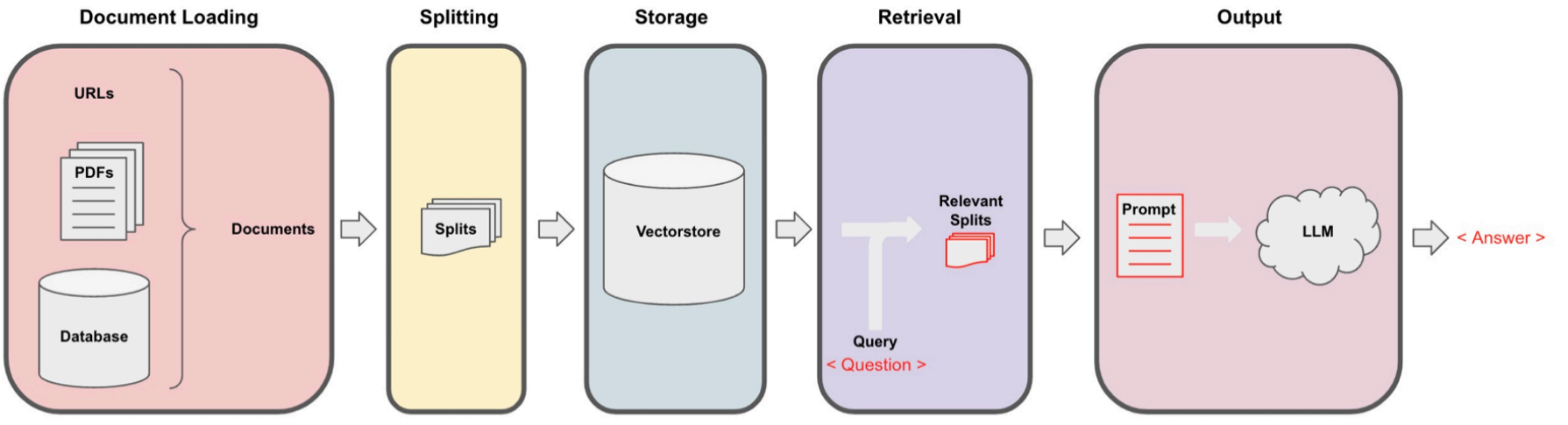

Document Loading
URLs
PDFs
Database
Documents


Document Loading
Splitting
URLs
PDFs
Database
Documents
Splits


Document Loading
Splitting
Storage
URLs
PDFs
Documents
Database
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

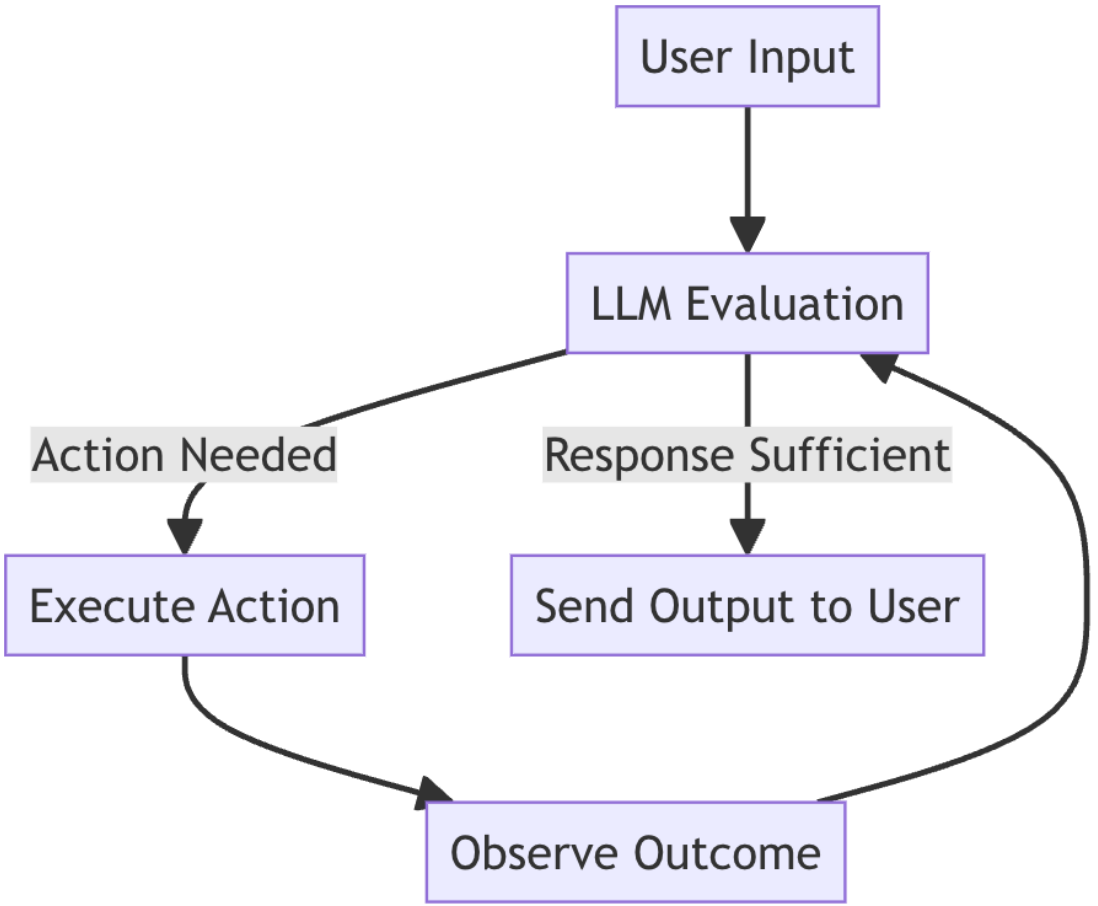

Document Loading
URLs
PDFs
Database
Documents
Splitting
Splits
Storage
Vectorstore
Retrieval
Relevant Splits
Query
< Question >
Output
Prompt
LLM
< Answer >

Notebook Demo - Local RAG with Llama 3.2

Q&A / Break

Local Agents with Llama 3.2

Local Agents with Llama 3.2



Explanation of the agent loop in cognitive architectures.

Practical Use Case: Customer Support Agent

Practical Use Case: Customer Support Agent

- **Scenario**: An LLM-powered customer support agent.

Practical Use Case: Customer Support Agent

- **Scenario**: An LLM-powered customer support agent.
- **User Input**: Customer asks about order status.

Practical Use Case: Customer Support Agent

- **Scenario**: An LLM-powered customer support agent.
- **User Input**: Customer asks about order status.
- **LLM Decision**: Determines if it can provide the status directly or if it needs to fetch data from the database.

Practical Use Case: Customer Support Agent

- **Scenario**: An LLM-powered customer support agent.
- **User Input**: Customer asks about order status.
- **LLM Decision**: Determines if it can provide the status directly or if it needs to fetch data from the database.
- **Action Taken**: If data fetch is needed, the agent queries the database and updates the user with the order status.

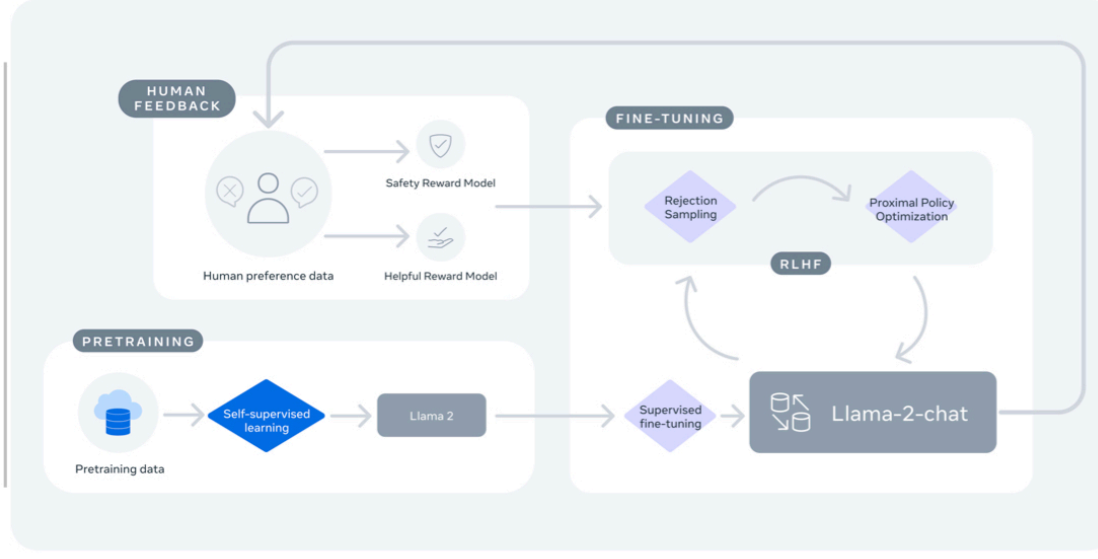
Practical use case of LLM agents in customer support.

Notebook Demo - Tool Calling and local agents with Llama 3.2

Q&A / Break

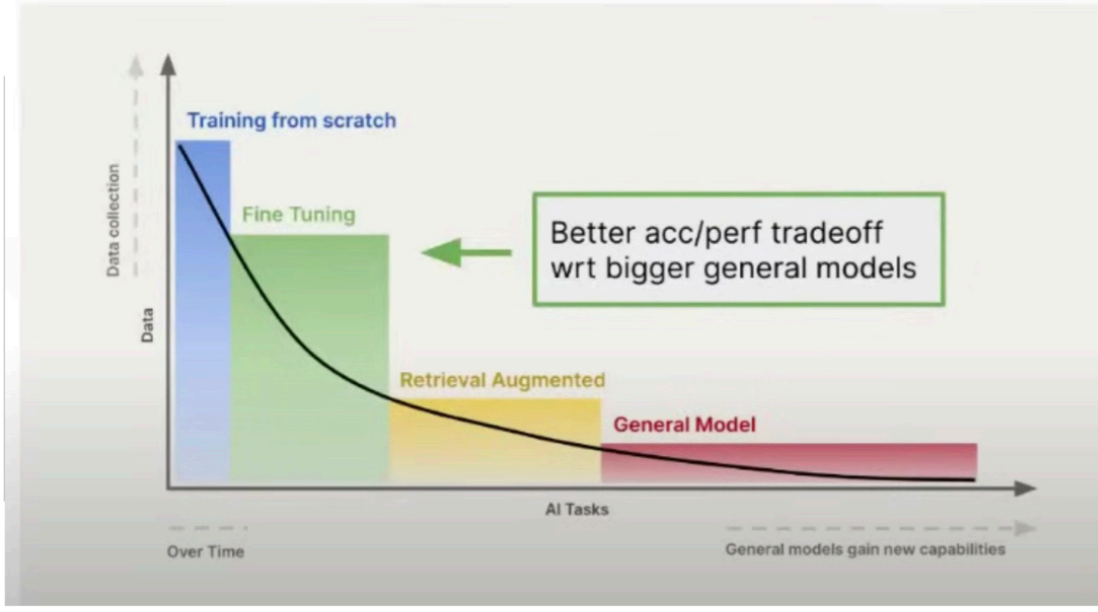
Fine Tunning Llama3.2

What is Fine Tuning?



LLama2 Paper

- What is Fine Tuning?
- Why Fine Tune?



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

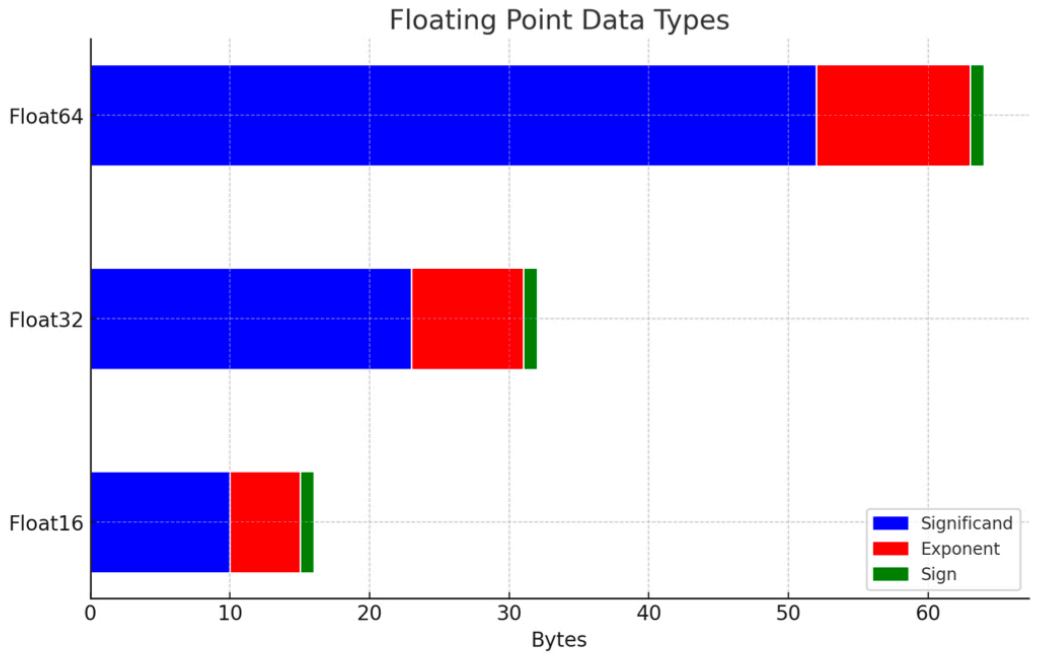
- What is Fine Tuning?
- Why Fine Tune?
- Memory cost of LLMs: parameters, gradients, optimiser states

The Memory Bottleneck: GPU comparison

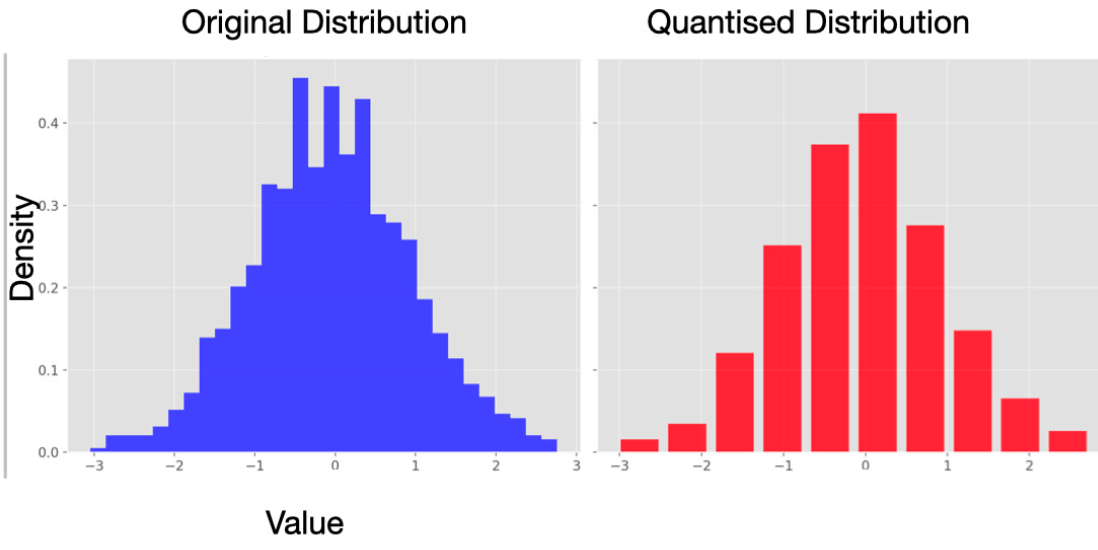
GPU	Tier	$ / hr (AWS)	VRAM (GiB)
H100	Enterprise	12.29	80
A100	Enterprise	5.12	80
V100	Enterprise	3.90	32
A10G	Enterprise	2.03	24
T4	Enterprise	0.98	16
RTX 4080	Consumer	N/A	16

Problem - Loading Params

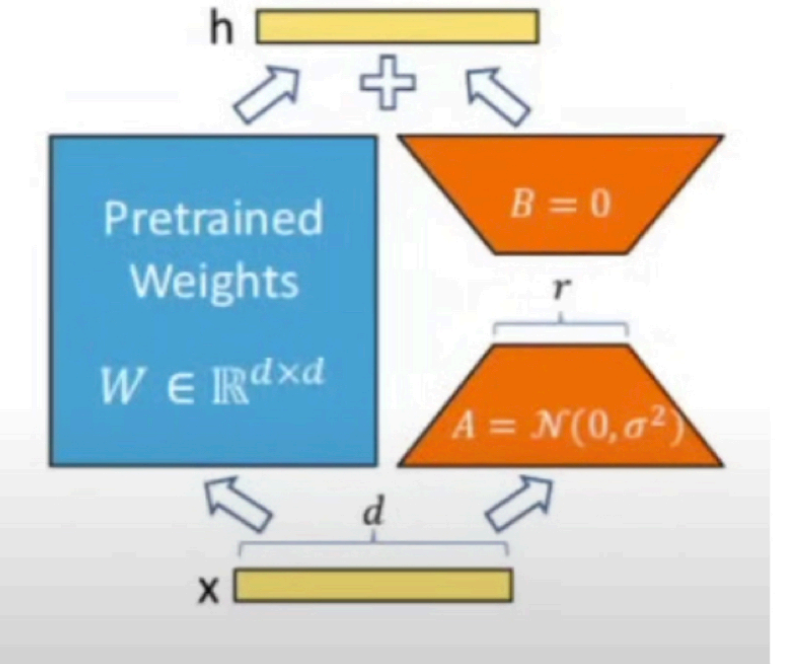
Solution - Half Precision



- 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

Notebook Demo - Fine-Tuning Llama3.1 - Walkthrough