Getting Started with Llama3

Lucas Soares

12-03-2025

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

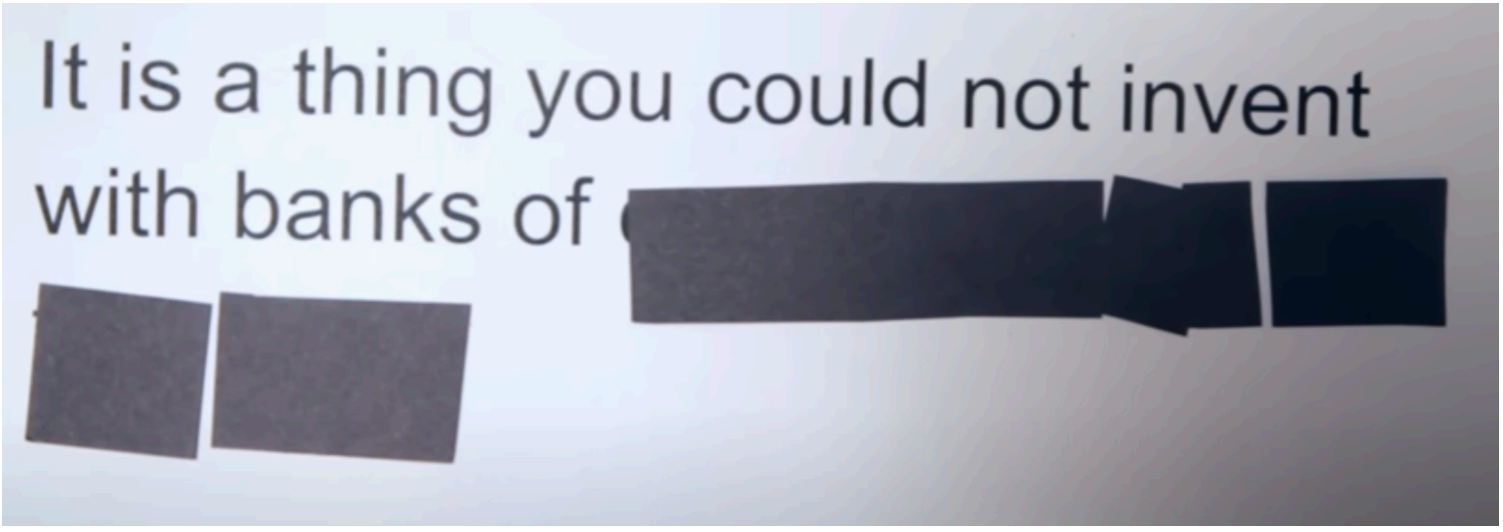
Methodology Notes

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

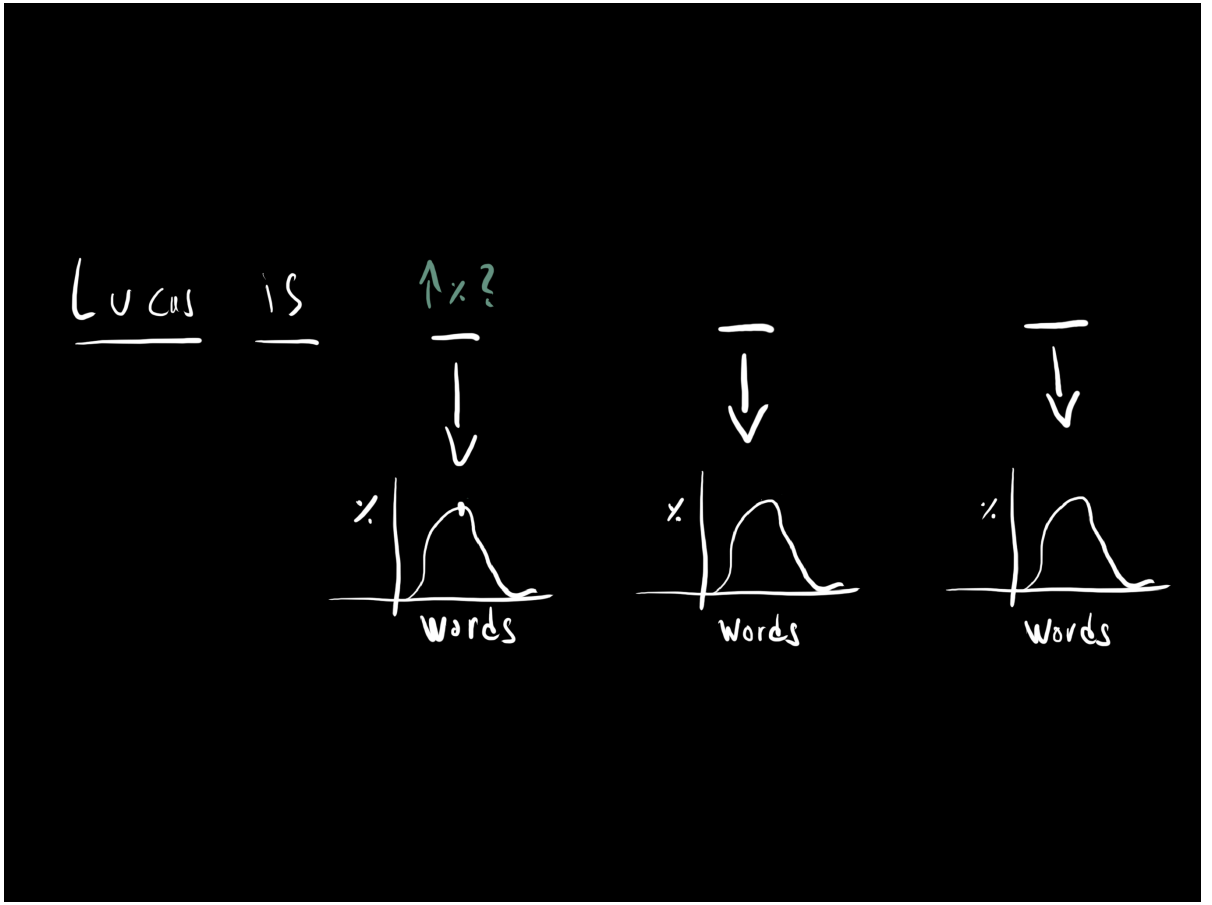
Methodology Notes

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

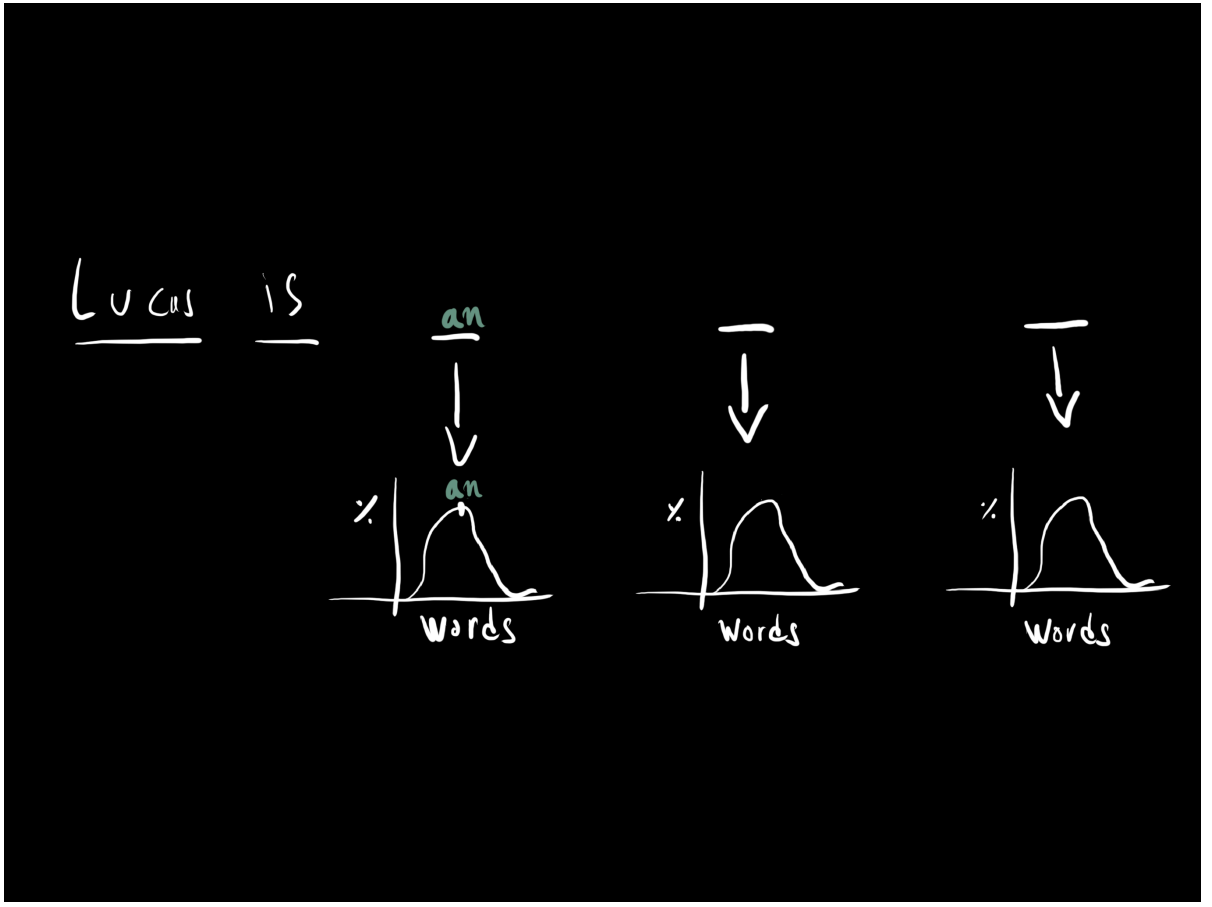
LLMs Predict the Next Word



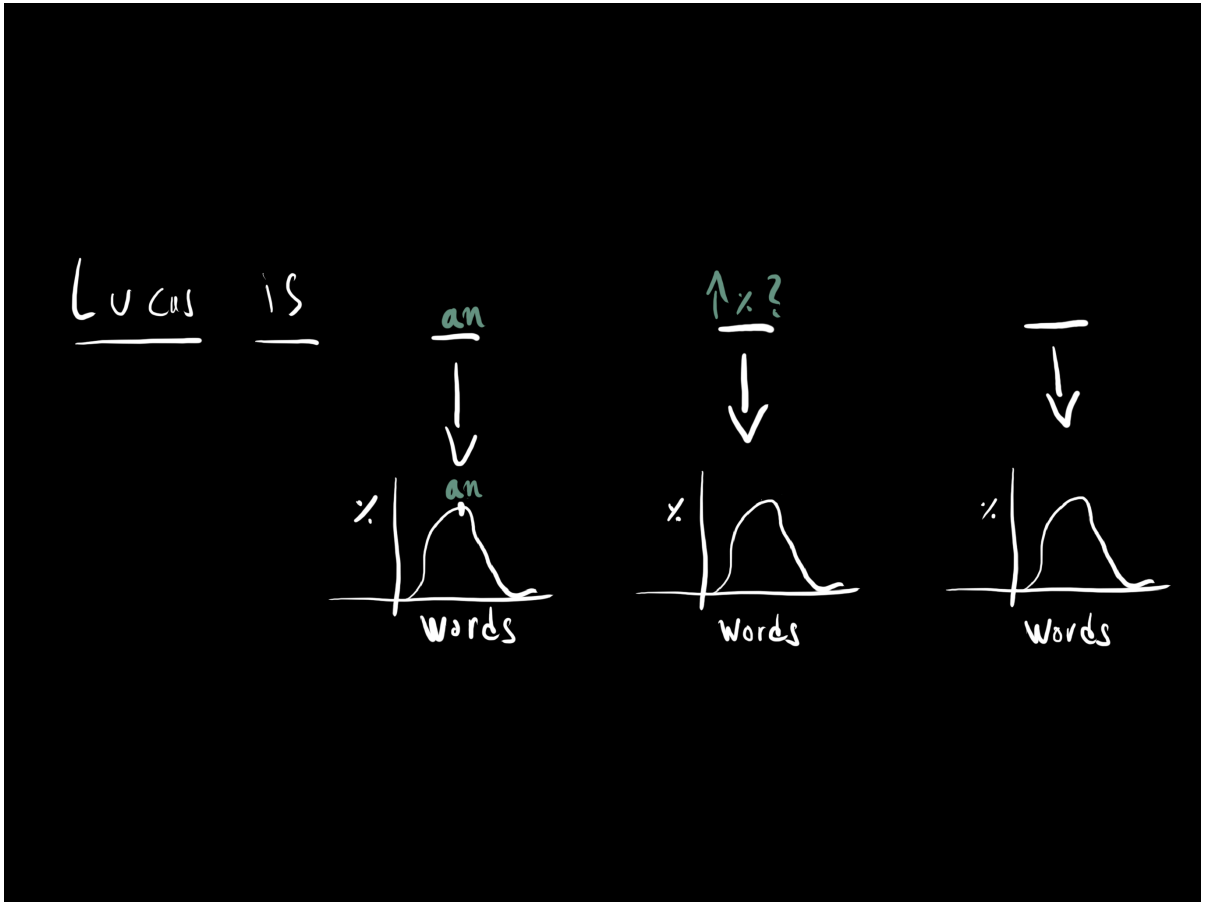
LLMs Predict the Next Word



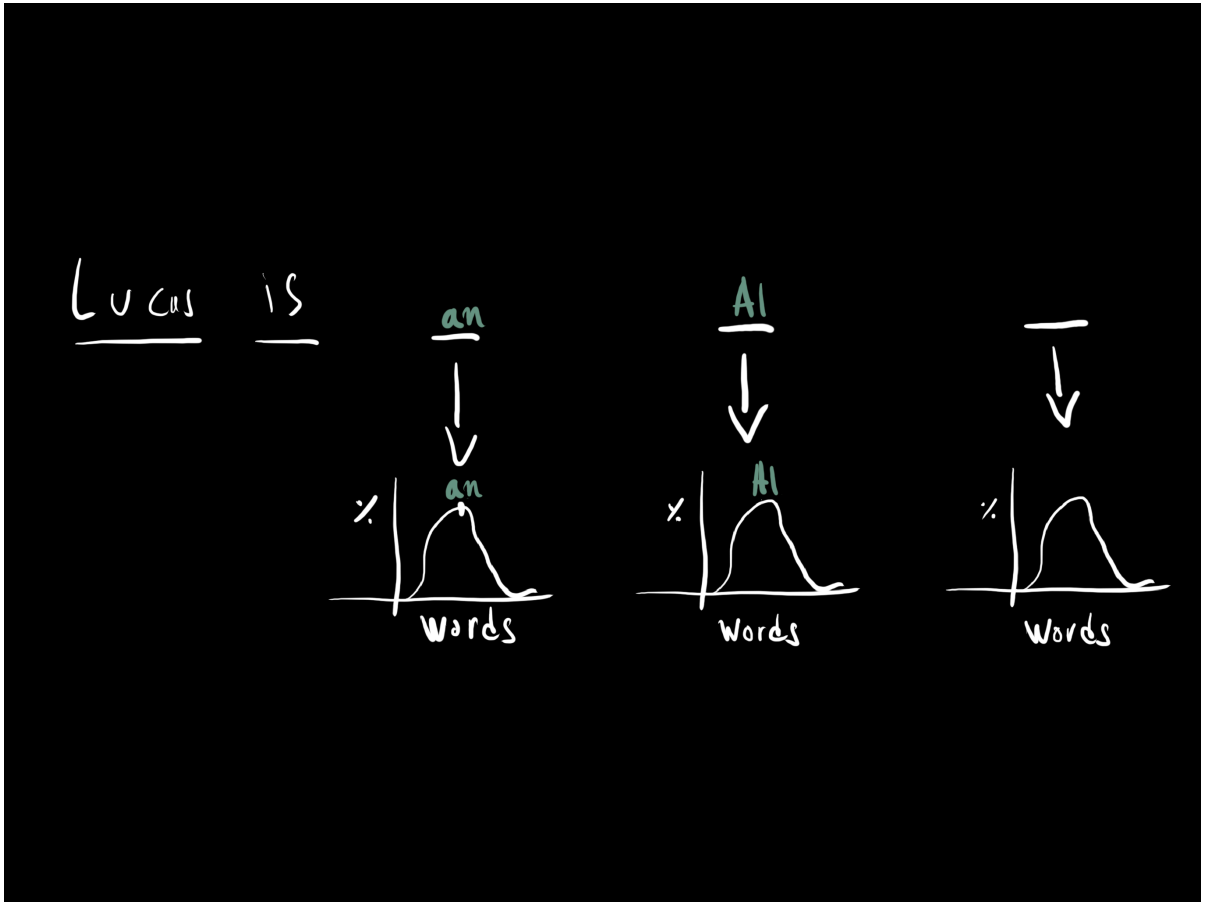
LLMs Predict the Next Word



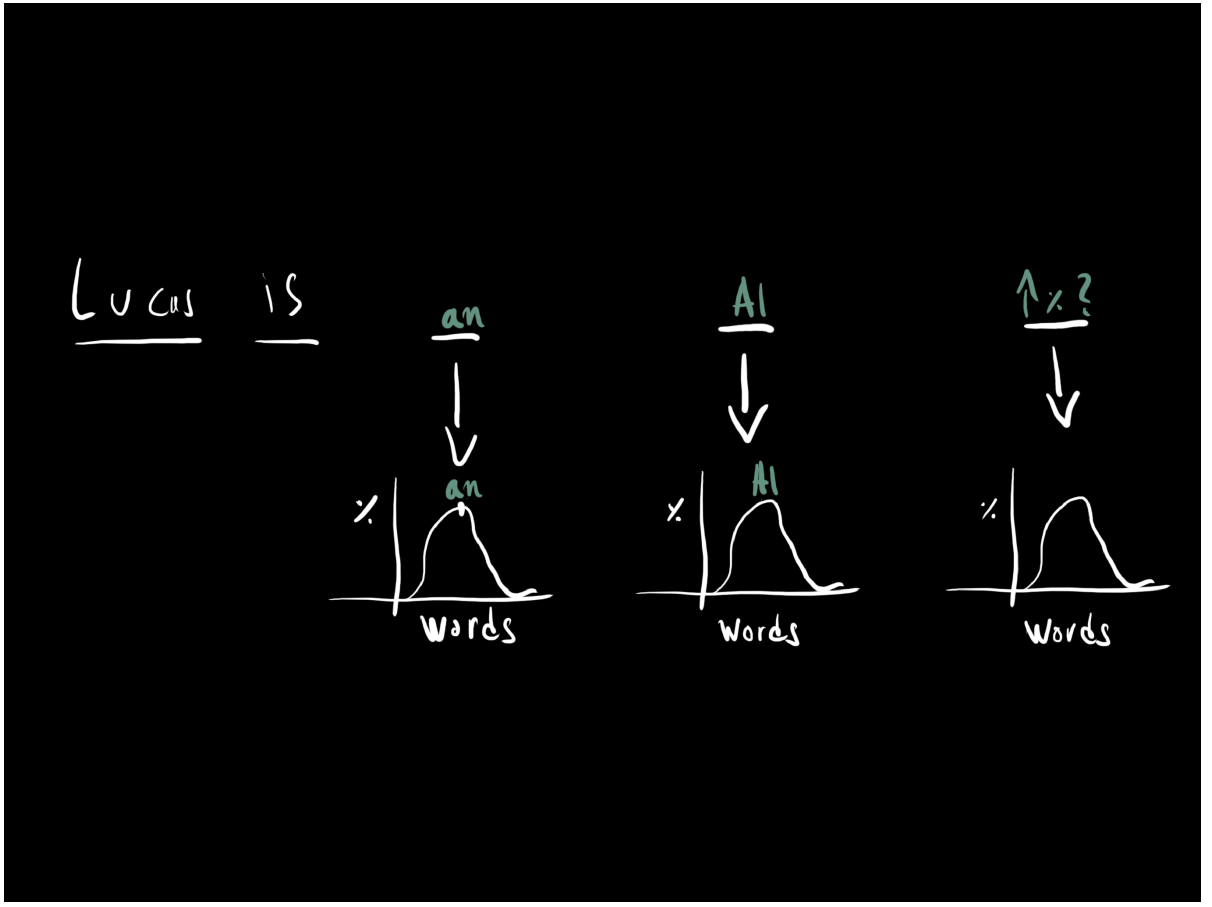
LLMs Predict the Next Word



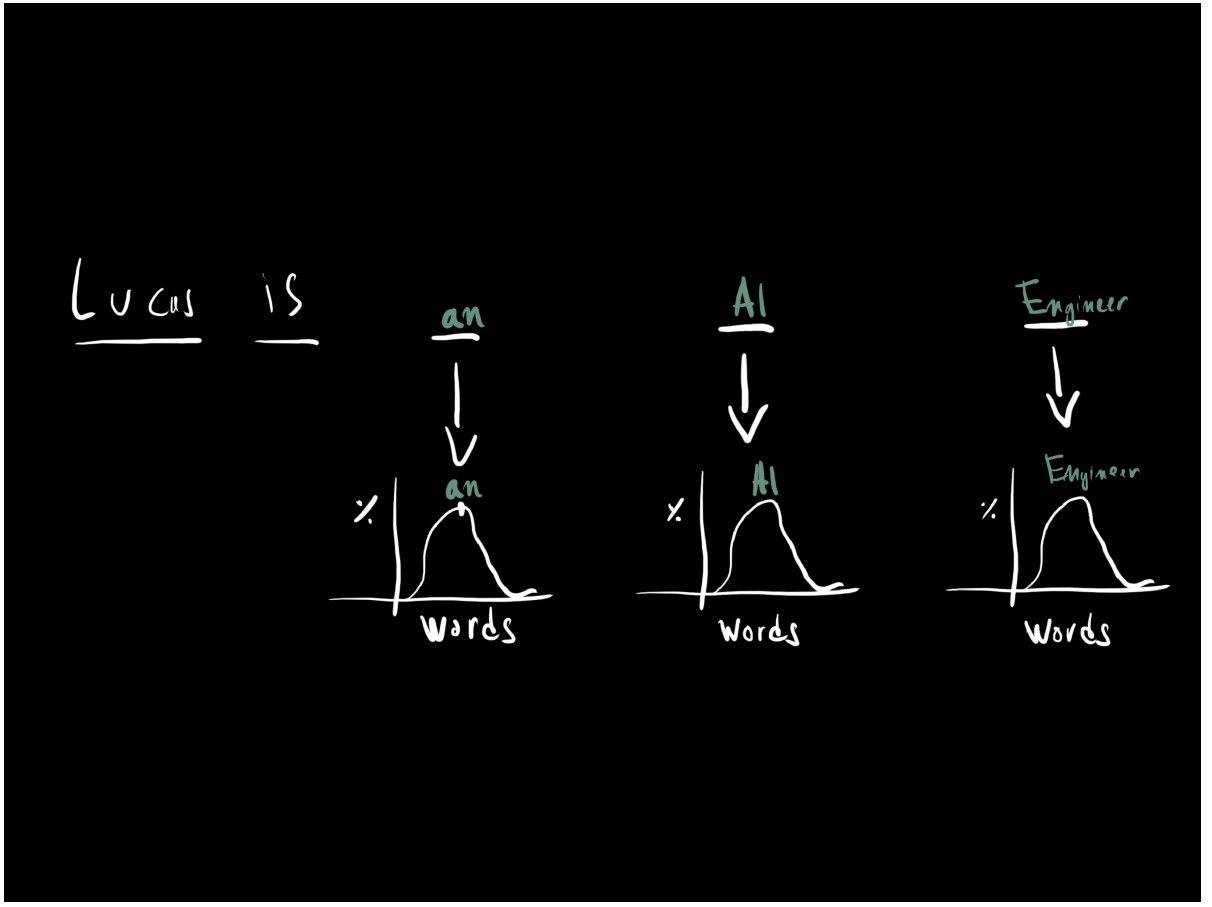
LLMs Predict the Next Word



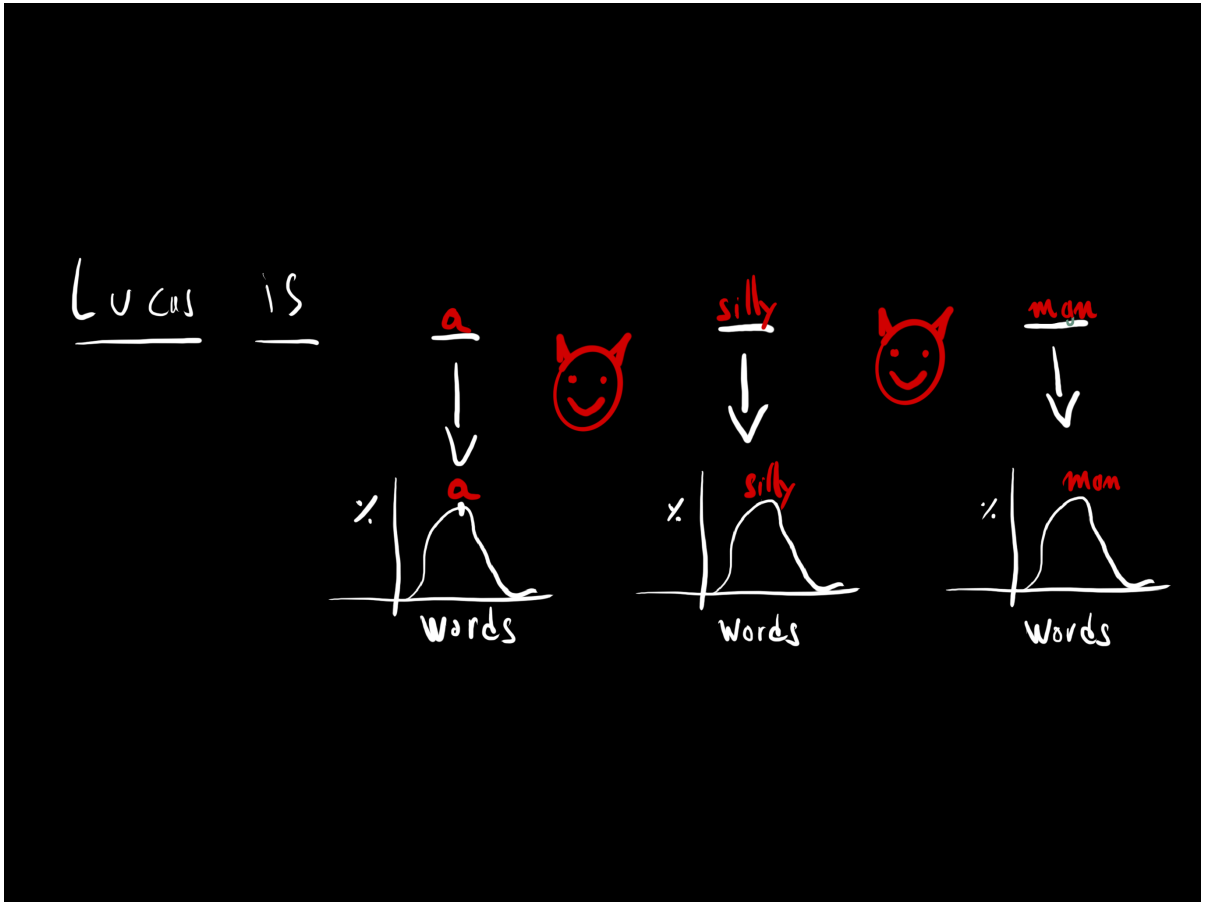
LLMs Predict the Next Word

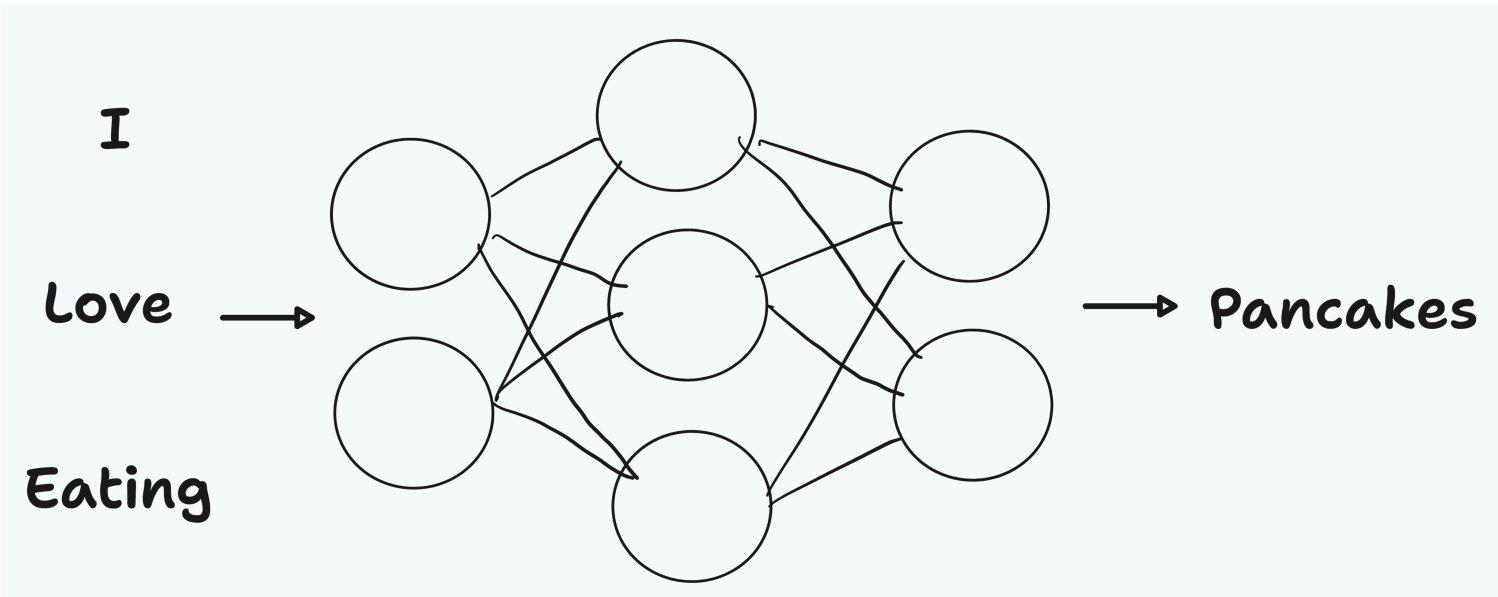


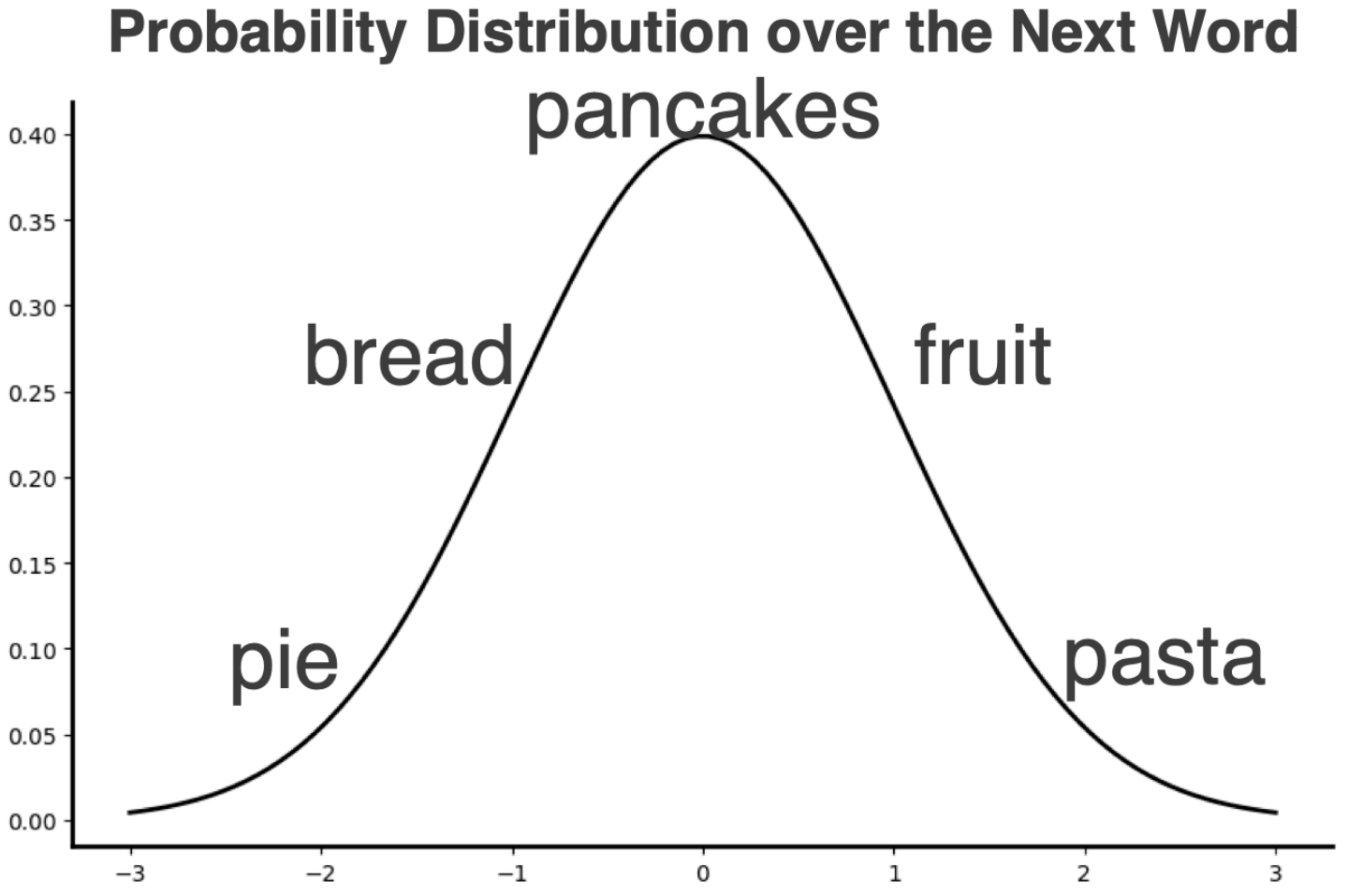
LLMs Predict the Next Word



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

The future of AI: Built with Llama

December 19, 2024 • 8 minute read

Llama3 Releases

- Meta Released Llama 3 in April of 2024
- Llama 3.1 was released in July of 2024
- Llama 3.2 was released in September of 2024
- Llama 3.3 was released in December of 2024

Llama 3 Series

- Open source with a Commercial license

Llama 3 Series

- Open source with a Commercial license
- The latest and greatest: Llama3.3

	Training Data	Params	Input modalities	Output modalities	Context length	GQA	Token count	Knowledge cutoff
Llama 3.3 (text only)	A new mix of publicly available online data.	70B	Multilingual Text	Multilingual Text and code	128k	Yes	15T+	December 2023

Hugging Face - Llama 3.3 70B Instruct

Llama 3.3 Evaluation Results

Category	Benchmark	# Shots	Metric	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama-3.3 70B Instruct
	MMLU (CoT)	0	macro_avg/acc	73.0	86.0	86.0
	MMLU Pro (CoT)	5	macro_avg/acc	48.3	66.4	68.9
Steerability	IFEval			80.4	87.5	92.1
Reasoning	GPQA Diamond (CoT)	0	acc	31.8	48.0	50.5
Code	HumanEval	0	pass@1	72.6	80.5	88.4
	MBPP EvalPlus (base)	0	pass@1	72.8	86.0	87.6
Math	MATH (CoT)	0	sympy_intersection_score	51.9	68.0	77.0
Tool Use	BFCL v2	0	overall_ast_summary/macro_avg/valid	65.4	77.5	77.3
Multilingual	MGSM	0	em	68.9	86.9	91.1

Hugging Face - Llama 3.3 70B Instruct

Llama 3.3 Evaluation Results

Category	Benchmark	# Shots	Metric	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama-3.3 70B Instruct
	MMLU (CoT)	0	macro_avg/acc	73.0	86.0	86.0
	MMLU Pro (CoT)	5	macro_avg/acc	48.3	66.4	68.9
Steerability	IFEval			80.4	87.5	92.1
Reasoning	GPQA Diamond (CoT)	0	acc	31.8	48.0	50.5
Code	HumanEval	0	pass@1	72.6	80.5	88.4
	MBPP EvalPlus (base)	0	pass@1	72.8	86.0	87.6
Math	MATH (CoT)	0	sympy_intersection_score	51.9	68.0	77.0
Tool Use	BFCL v2	0	overall_ast_summary/macro_avg/valid	65.4	77.5	77.3
Multilingual	MGSM	0	em	68.9	86.9	91.1

good at general problem solving

Llama 3.3 Evaluation Results

Category	Benchmark	# Shots	Metric	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama-3.3 70B Instruct
	MMLU (CoT)	0	macro_avg/acc	73.0	86.0	86.0
	MMLU Pro (CoT)	5	macro_avg/acc	48.3	66.4	68.9
Steerability	IFEval			80.4	87.5	92.1
Reasoning	GPQA Diamond (CoT)	0	acc	31.8	48.0	50.5
Code	HumanEval	0	pass@1	72.6	80.5	88.4
	MBPP EvalPlus (base)	0	pass@1	72.8	86.0	87.6
Math	MATH (CoT)	0	sympy_intersection_score	51.9	68.0	77.0
Tool Use	BFCL v2	0	overall_ast_summary/macro_avg/valid	65.4	77.5	77.3
Multilingual	MGSM	0	em	68.9	86.9	91.1

follows instructions well

Llama 3.3 Evaluation Results

Category	Benchmark	# Shots	Metric	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama-3.3 70B Instruct
	MMLU (CoT)	0	macro_avg/acc	73.0	86.0	86.0
	MMLU Pro (CoT)	5	macro_avg/acc	48.3	66.4	68.9
Steerability	IFEval			80.4	87.5	92.1
Reasoning	GPQA Diamond (CoT)	0	acc	31.8	48.0	50.5
Code	HumanEval	0	pass@1	72.6	80.5	88.4
	MBPP EvalPlus (base)	0	pass@1	72.8	86.0	87.6
Math	MATH (CoT)	0	sympy_intersection_score	51.9	68.0	77.0
Tool Use	BFCL v2	0	overall_ast_summary/macro_avg/valid	65.4	77.5	77.3
Multilingual	MGSM	0	em	68.9	86.9	91.1

good at thinking
GPT-4o is 54%

Llama 3.3 Evaluation Results

Category	Benchmark	# Shots	Metric	Llama 3.1 8B Instruct	Llama 3.1 70B Instruct	Llama-3.3 70B Instruct
	MMLU (CoT)	0	macro_avg/acc	73.0	86.0	86.0
	MMLU Pro (CoT)	5	macro_avg/acc	48.3	66.4	68.9
Steerability	IFEval			80.4	87.5	92.1
Reasoning	GPQA Diamond (CoT)	0	acc	31.8	48.0	50.5
Code	HumanEval	0	pass@1	72.6	80.5	88.4
	MBPP EvalPlus (base)	0	pass@1	72.8	86.0	87.6
Math	MATH (CoT)	0	sympy_intersection_score	51.9	68.0	77.0
Tool Use	BFCL v2	0	overall_ast_summary/macro_avg/valid	65.4	77.5	77.3
Multilingual	MGSM	0	em	68.9	86.9	91.1

good at coding

Whiteboard - Llama 3 Offering Breakdown

Q&A / Break

Notebook Demo - Introduction to Using The Llama Series Models

Q&A / Break

Query Your Docs Locally with Llama 3

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

Query Your Docs Locally with Llama 3

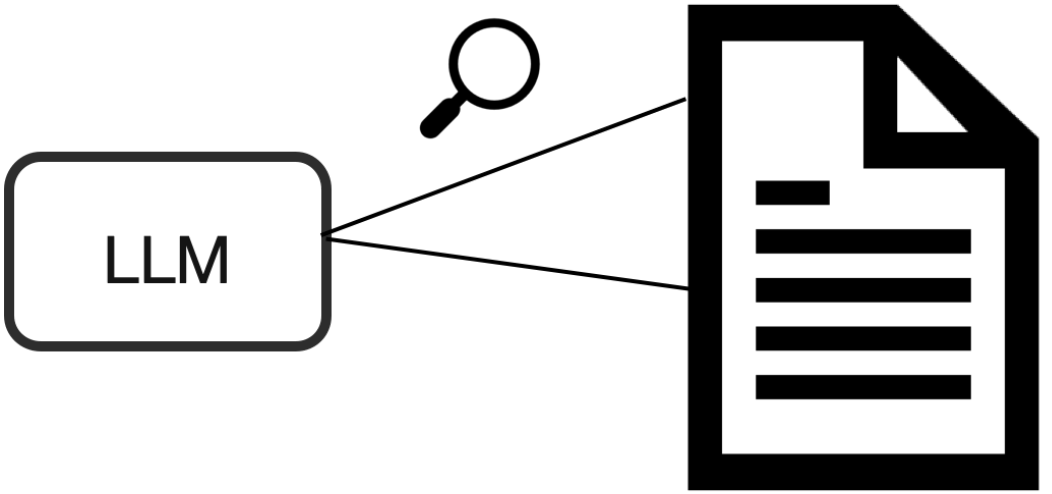
- Privacy concern with closed source LLMs.

Query Your Docs Locall with Llama 3

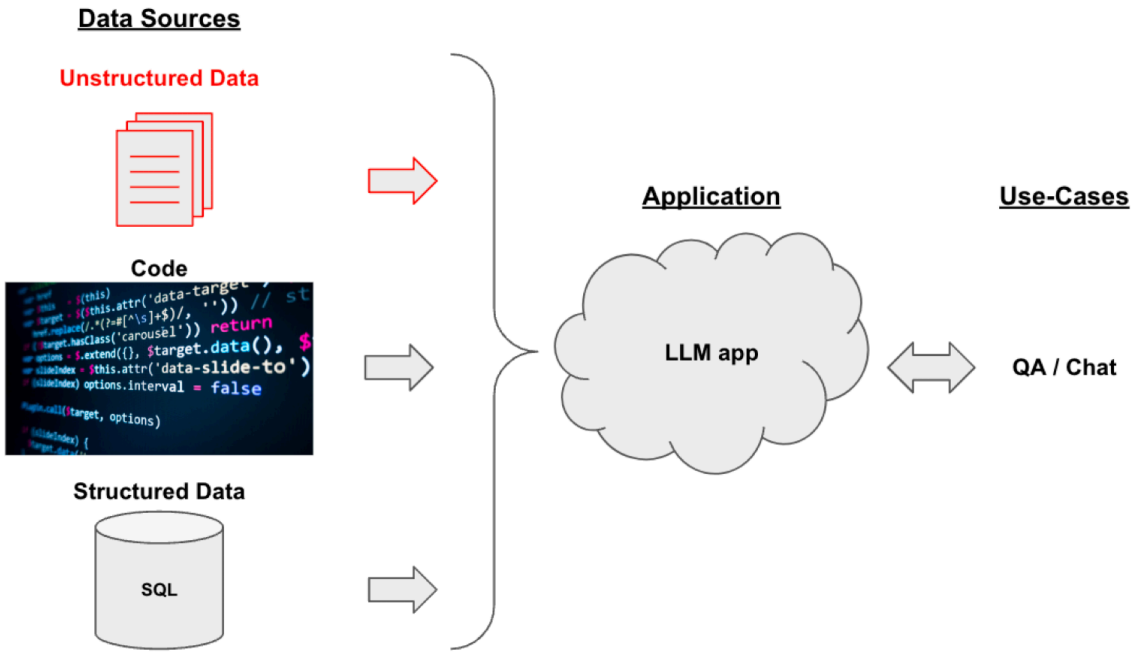
- Solution? Llama 3!

RAG with Llama 3

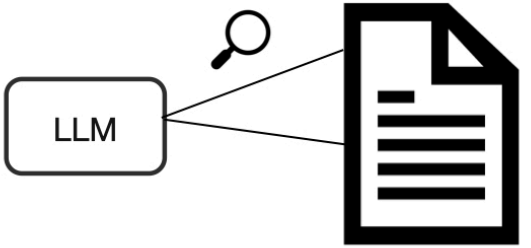
- RAG - Retrieval Augmented Generation



RAG with Llama 3

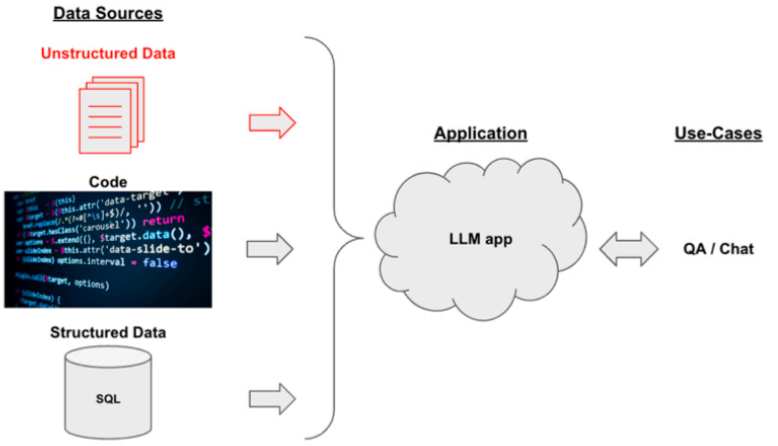


RAG - Retrieval Augmented Generation



Langchain Docs

RAG - Retrieval Augmented Generation

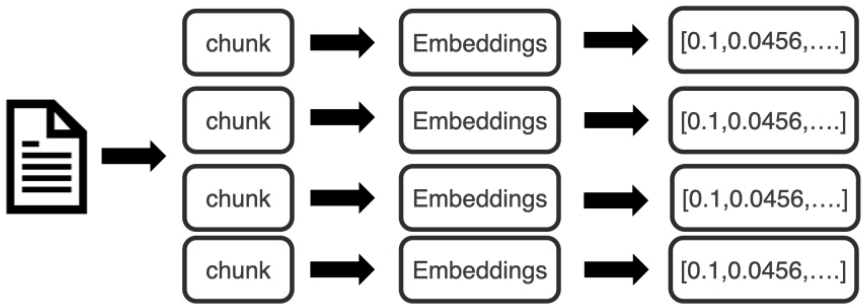


[Langchain Docs]

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

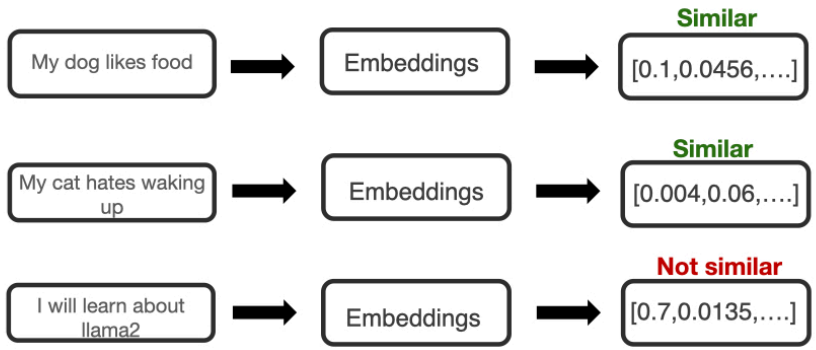
Langchain Docs

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



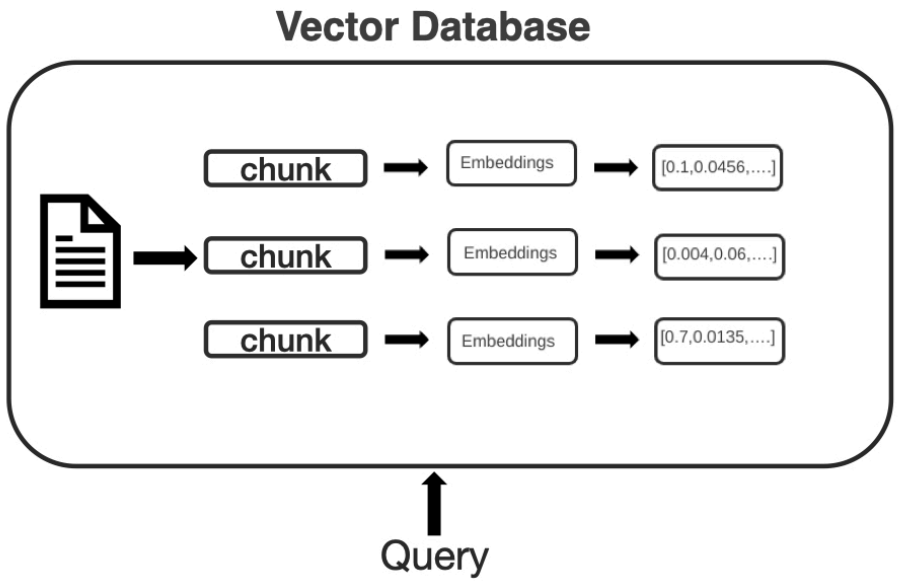
Langchain Docs

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

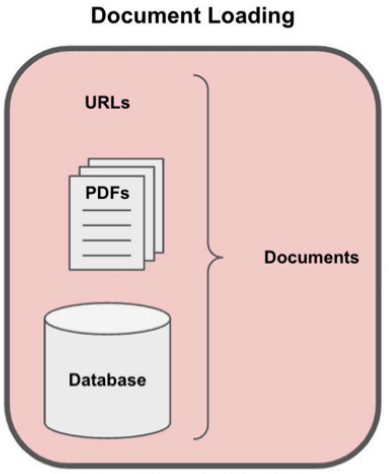


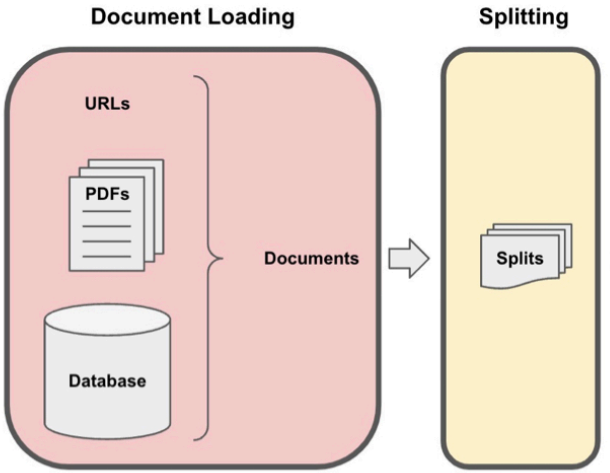
Langchain Docs

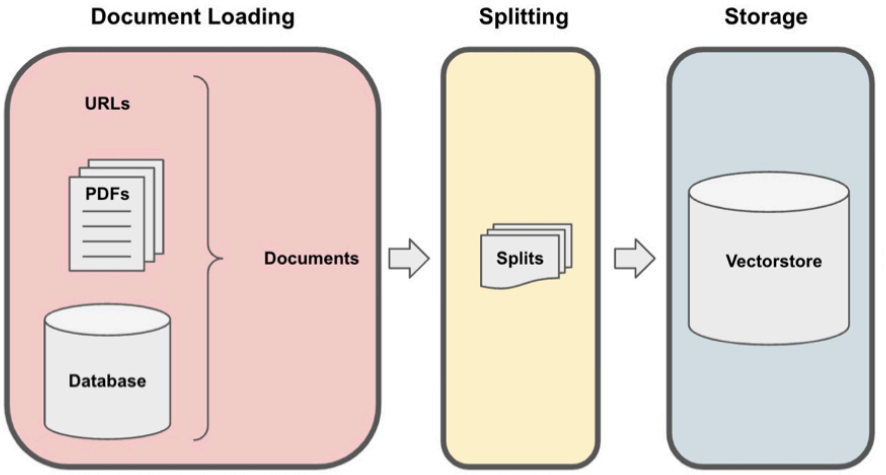
- RAG - Retrieval Augmented Generation
- LLMs have a limited context length
- **Embeddings**: capture content and meaning

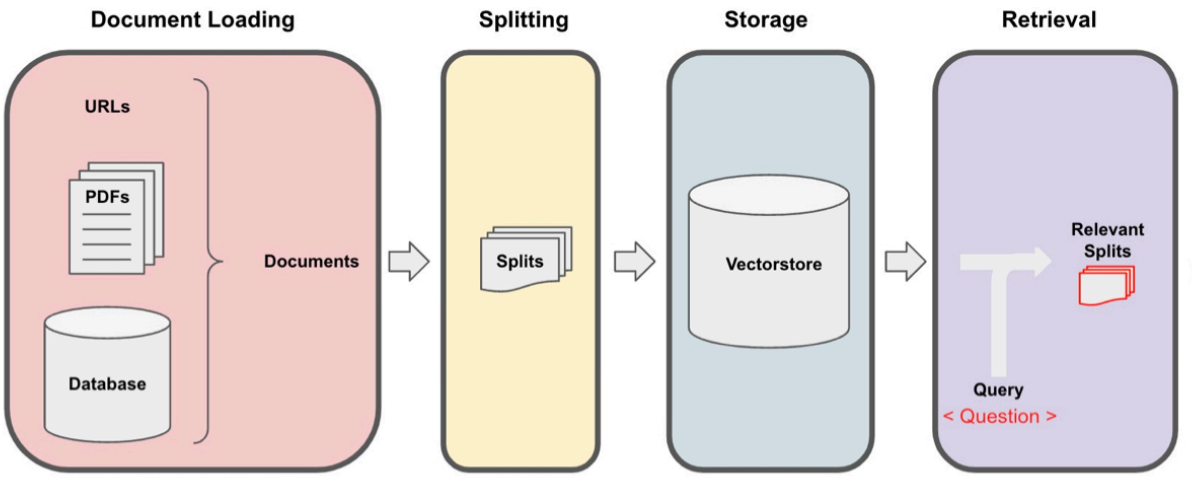


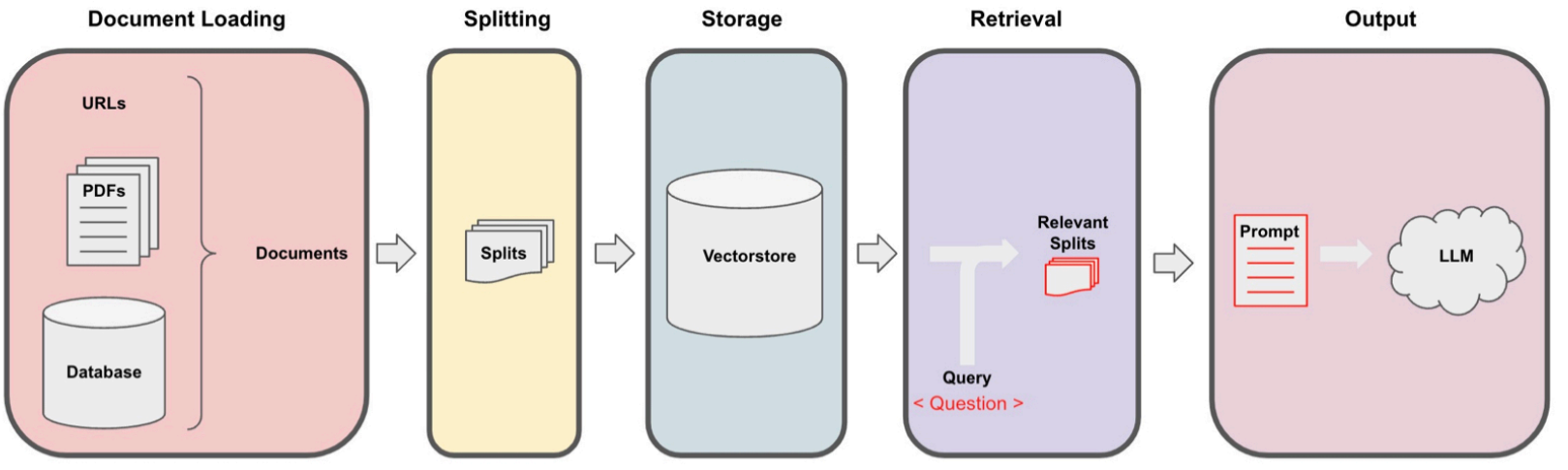
Langchain Docs

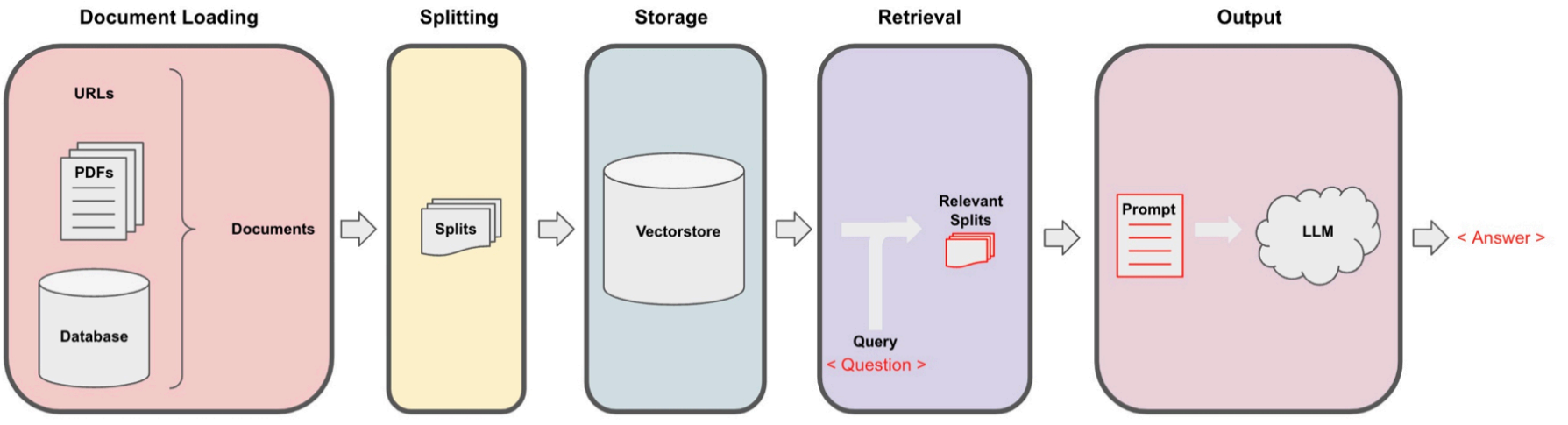

Document Loading
URLs
PDFs
Database
Documents


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

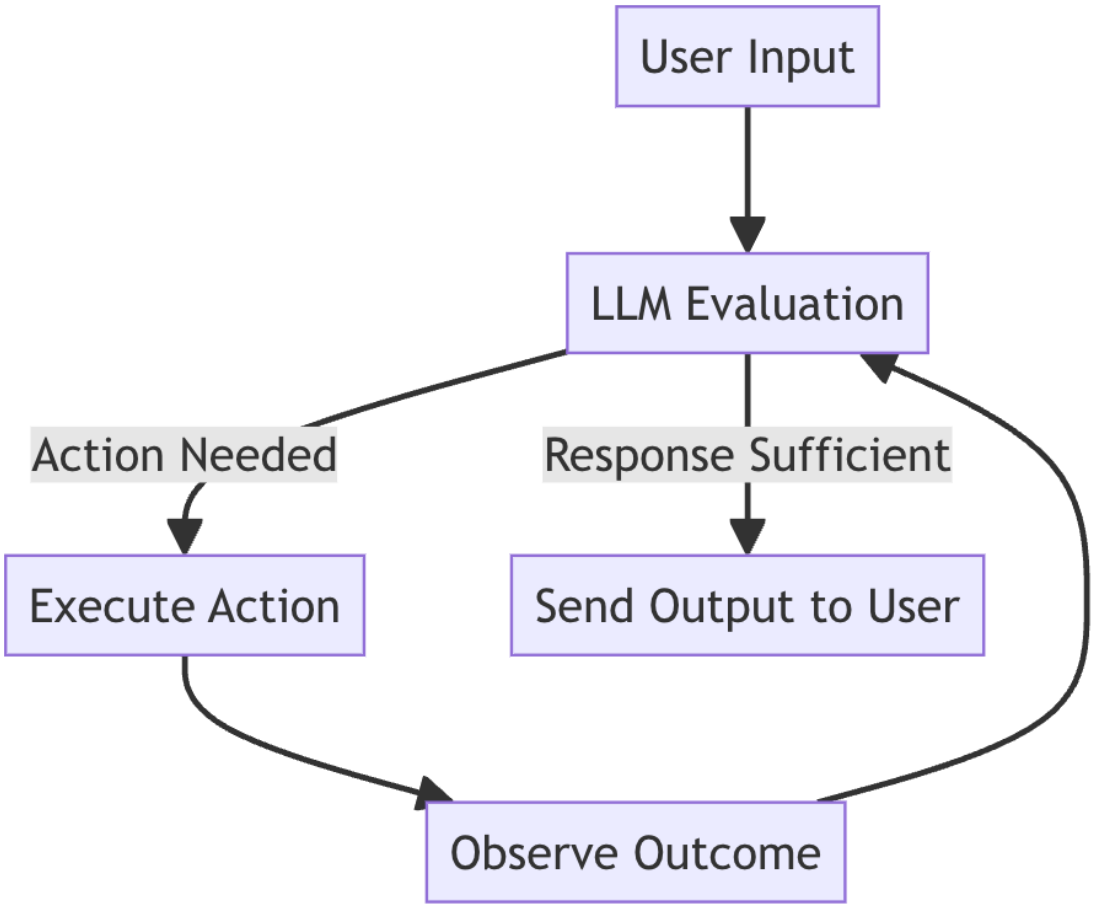

Document Loading
URLs
PDFs
Database
Documents
Splitting
Splits
Storage
Vectorstore
Retrieval
Relevant Splits
Query
< Question >
Output
Prompt
LLM
< Answer >

Notebook Demo - Local RAG with Llama 3

Q&A / Break

Local Agents with Llama 3

Local Agents with Llama 3



Explanation of the agent loop in cognitive architectures.

Practical Use Case: Private Docs Chatbot

Practical Use Case: Private Docs Chatbot

- **Scenario**: An LLM-powered chatbot designed to assist users by retrieving information from private documents.

Practical Use Case: Private Docs Chatbot

- **Scenario**: An LLM-powered chatbot designed to assist users by retrieving information from private documents.
- **User Input**: User asks about specific details from an internal document.

Practical Use Case: Private Docs Chatbot

- **Scenario**: An LLM-powered chatbot designed to assist users by retrieving information from private documents.
- **User Input**: User asks about specific details from an internal document.
- **LLM Decision**: Determines if the requested information requires search the document repository.

Practical Use Case: Private Docs Chatbot

- **Scenario**: An LLM-powered chatbot designed to assist users by retrieving information from private documents.
- **User Input**: User asks about specific details from an internal document.
- **LLM Decision**: Determines if the requested information requires search the document repository.
- **Action Taken**: If a search is required, the agent retrieves relevant document sections and provides a concise response based on the content.

Practical use case of LLM agents in customer support.

Notebook Demo - Tool Calling/Structured Outputs and local agents with Llama 3

Q&A / Break

Fine Tunning Llama 3

Whiteboard - Fine Tunning Llama 3: what, when, why and beyond

Fine Tunning Llama 3

- What is fine-tuning?

Fine Tunning Llama 3

- What is fine-tuning?
- It's the process of training a smaller model on a specific task leveraging a domain-specific dataset.

Fine Tunning Llama 3

- What is fine-tuning?
- It's the process of training a smaller model on a specific task leveraging a domain-specific dataset.
- What does it do?

Fine Tunning Llama 3

- What is fine-tuning?
- It's the process of training a smaller model on a specific task leveraging a domain-specific dataset.
- What does it do?
- Fine-tuning transfers the pre-trained model’s learned patterns and features to new tasks, improving performance and reducing training data needs.

Fine Tunning Llama 3

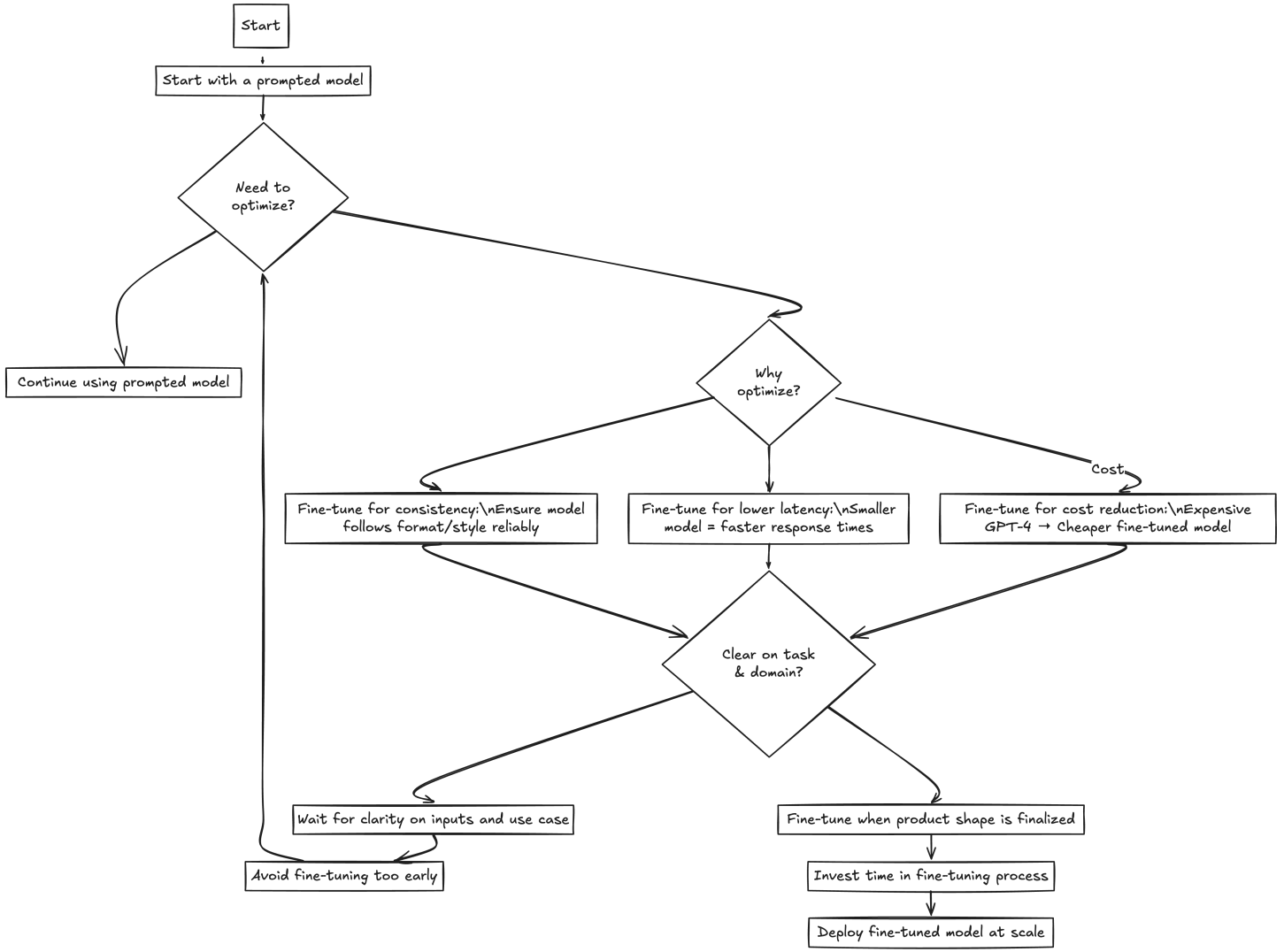
- What is fine-tuning?
- It's the process of training a smaller model on a specific task leveraging a domain-specific dataset.
- What does it do?
- Fine-tuning transfers the pre-trained model’s learned patterns and features to new tasks, improving performance and reducing training data needs.
- Why fine-tune?

Fine Tunning Llama 3

- What is fine-tuning?
- It's the process of training a smaller model on a specific task leveraging a domain-specific dataset.
- What does it do?
- Fine-tuning transfers the pre-trained model’s learned patterns and features to new tasks, improving performance and reducing training data needs.
- Why fine-tune?
- To improve the model's performance on a specific task.

Fine Tunning Llama 3

- What is fine-tuning?
- It's the process of training a smaller model on a specific task leveraging a domain-specific dataset.
- What does it do?
- Fine-tuning transfers the pre-trained model’s learned patterns and features to new tasks, improving performance and reducing training data needs.
- Why fine-tune?
- To improve the model's performance on a specific task.
- When to fine-tune?


Start
Start with a prompted model
Need to optimize?
Continue using prompted model
Why optimize?
Cost
Fine-tune for consistency:\nEnsure model follows format/style reliably
Fine-tune for lower latency:\nSmaller model = faster response times
Fine-tune for cost reduction:\nExpensive GPT-4 → Cheaper fine-tuned model
Clear on task & domain?
Wait for clarity on inputs and use case
Avoid fine-tuning too early
Fine-tune when product shape is finalized
Invest time in fine-tuning process
Deploy fine-tuned model at scale

Notebook Demo - Fine-Tuning Llama 3 - Walkthrough

Q&A / Break

Notebook Demo - End To End Example with Llama 3

References

1. Llama 3.3 Model Card
2. Fine Tuning Video
3. High Quality Structured Report Generation Demo
4. Llama Cookbook – Finetuning
5. Llama Recipes
6. Generating Synthetic Data
7. Llama Cookbook – End-to-End Use Cases
8. Structured Outputs with Ollama
9. Chunk Size Explanation
10. Embeddings + Llama by Simon Willison
11. Embeddings + Parquet for Small Projects
12. Awesome Local LLMs
13. LLM-Driven Data Engineering
14. Text Chunking in RAG
15. Llama Resources – AI.Meta
16. Llama 3 Paper (arXiv)