



Claude AI from basic to advance





Page

Messages

Notifications 3

Insights

Publishing Tools

Settings

Help ▾



somkiat.cc

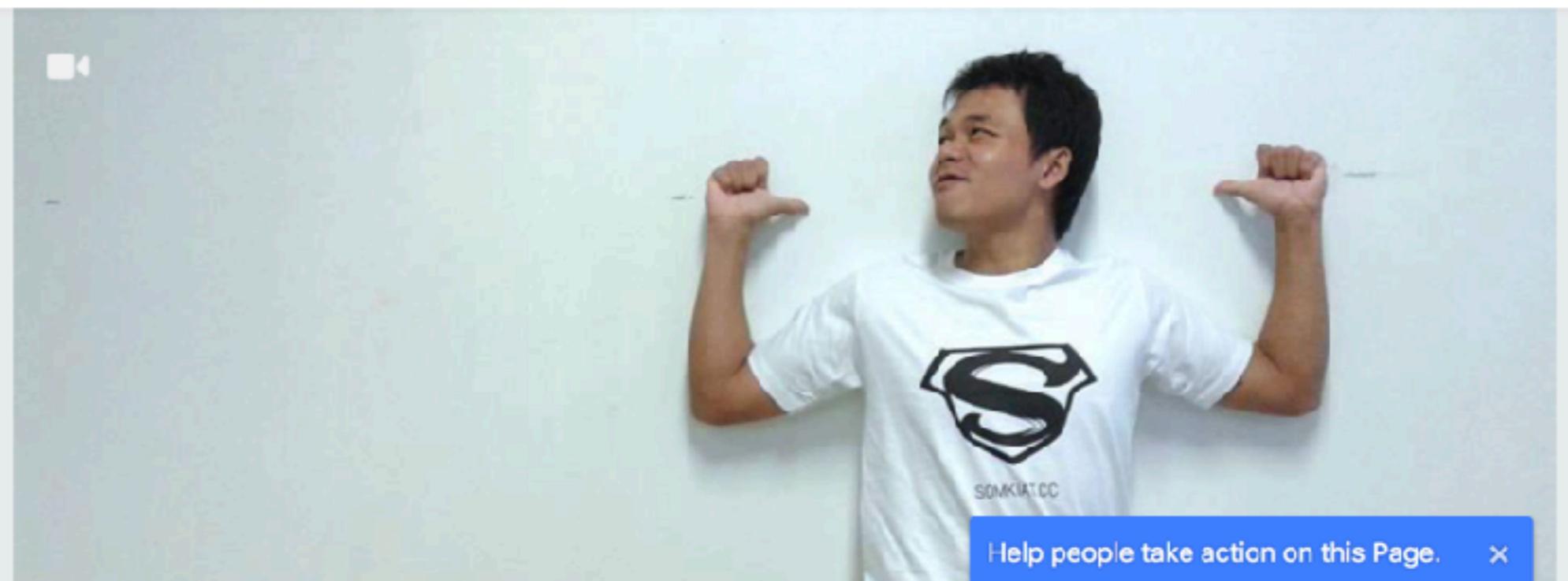
@somkiat.cc

Home

Posts

Videos

Photos



AI

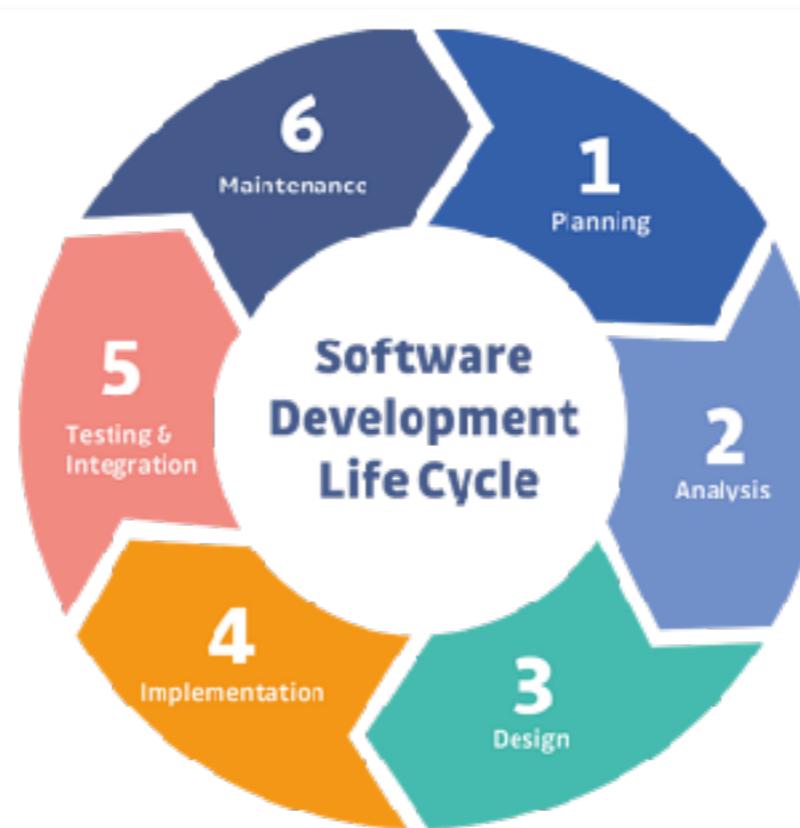
3

**[https://github.com/up1/
workshop-ai-with-technical-team](https://github.com/up1/workshop-ai-with-technical-team)**



Goals

Integrate Generative AI in Software Development
Optimize code quality
Team up with AI on coding tasks
Develop innovative solutions



Software Development

Requirement

Design

Develop

Testing

Deploy

Generative AI

Improve Productivity ... (Replace human !!)



AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

Learning Path

AI/ML

LLM, SLM

Prompt Engineer

AI in Software development

Develop AI/LLM app

RAG

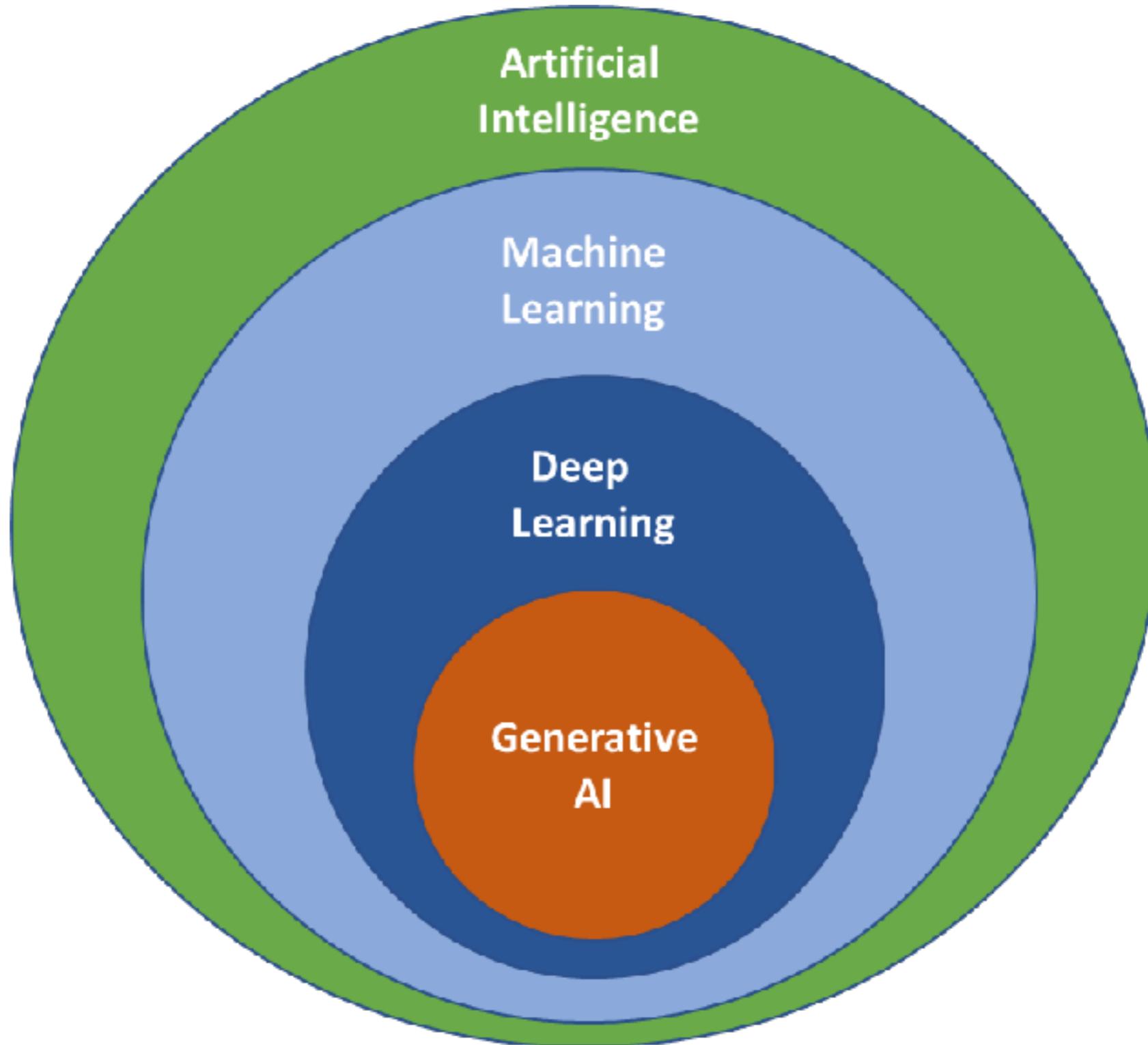
Use cases and workshops with Claude AI

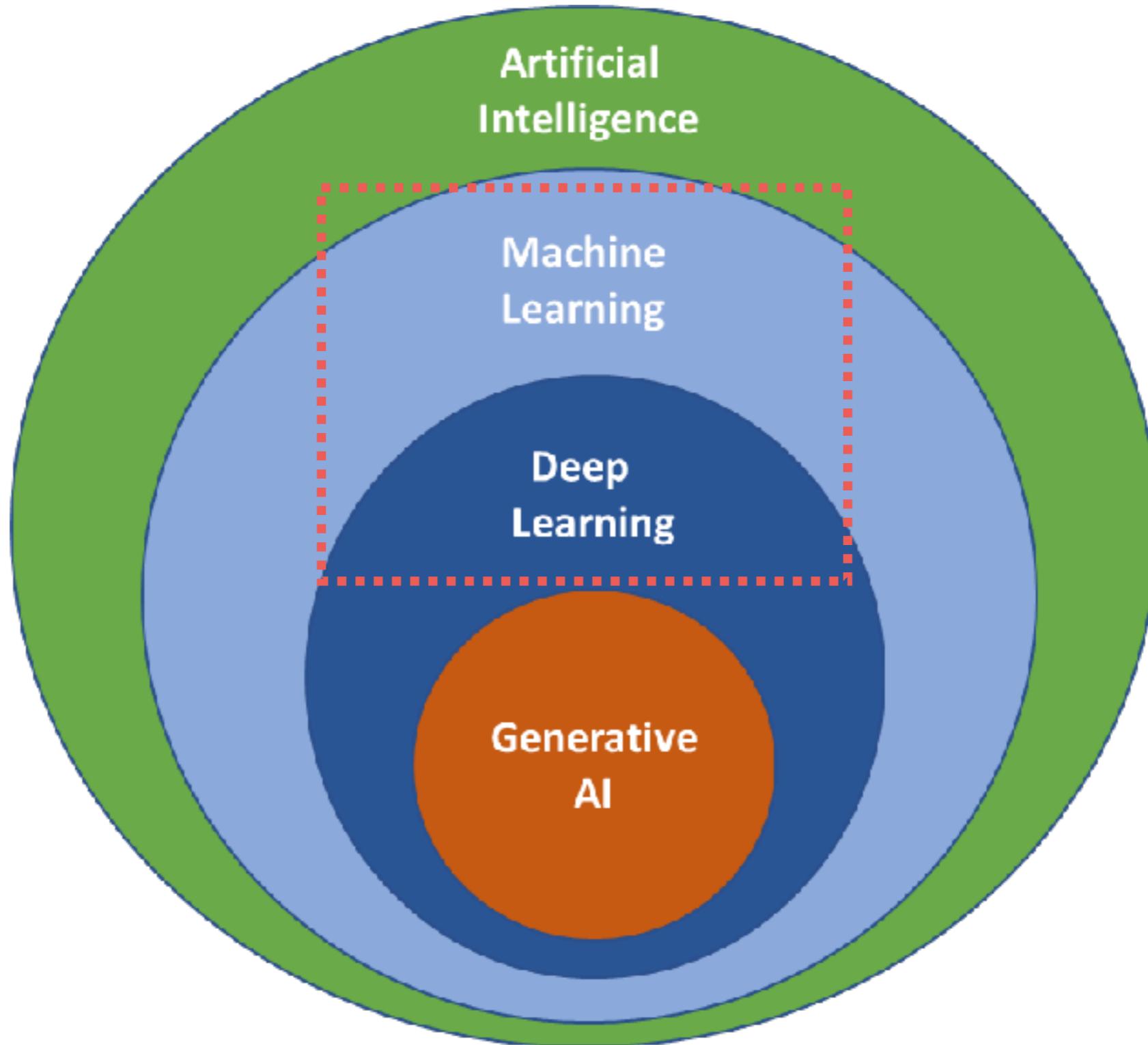


Module 1

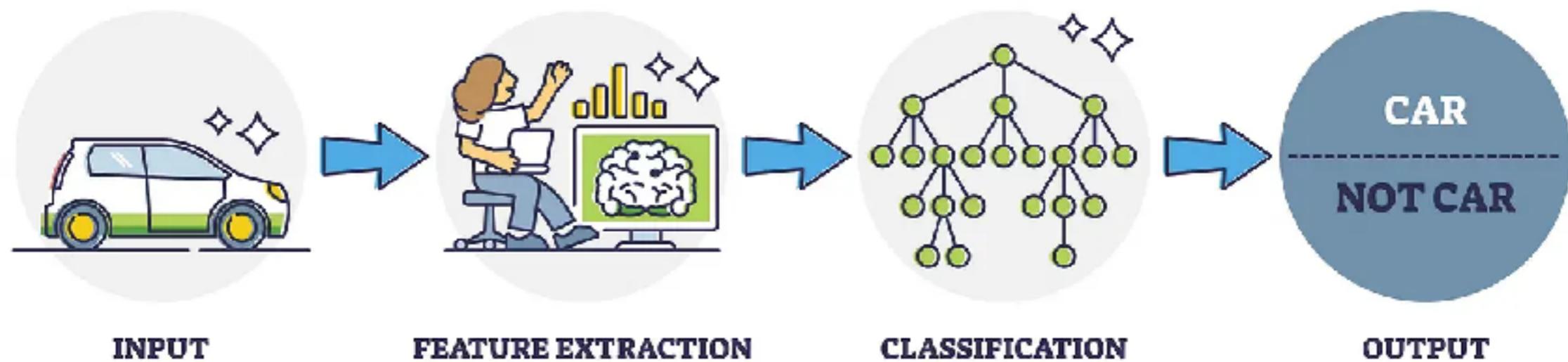
AI and Machine Learning (ML)



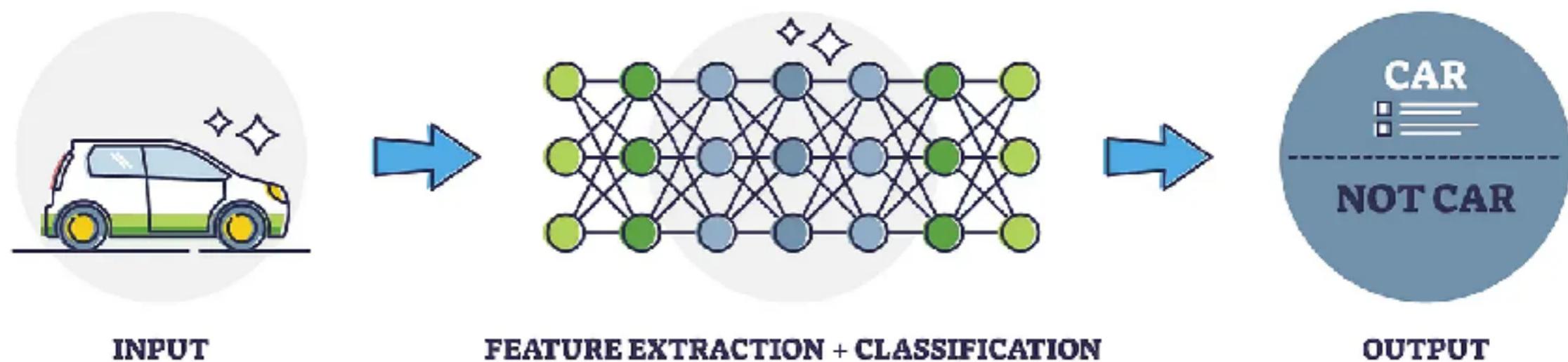




MACHINE LEARNING

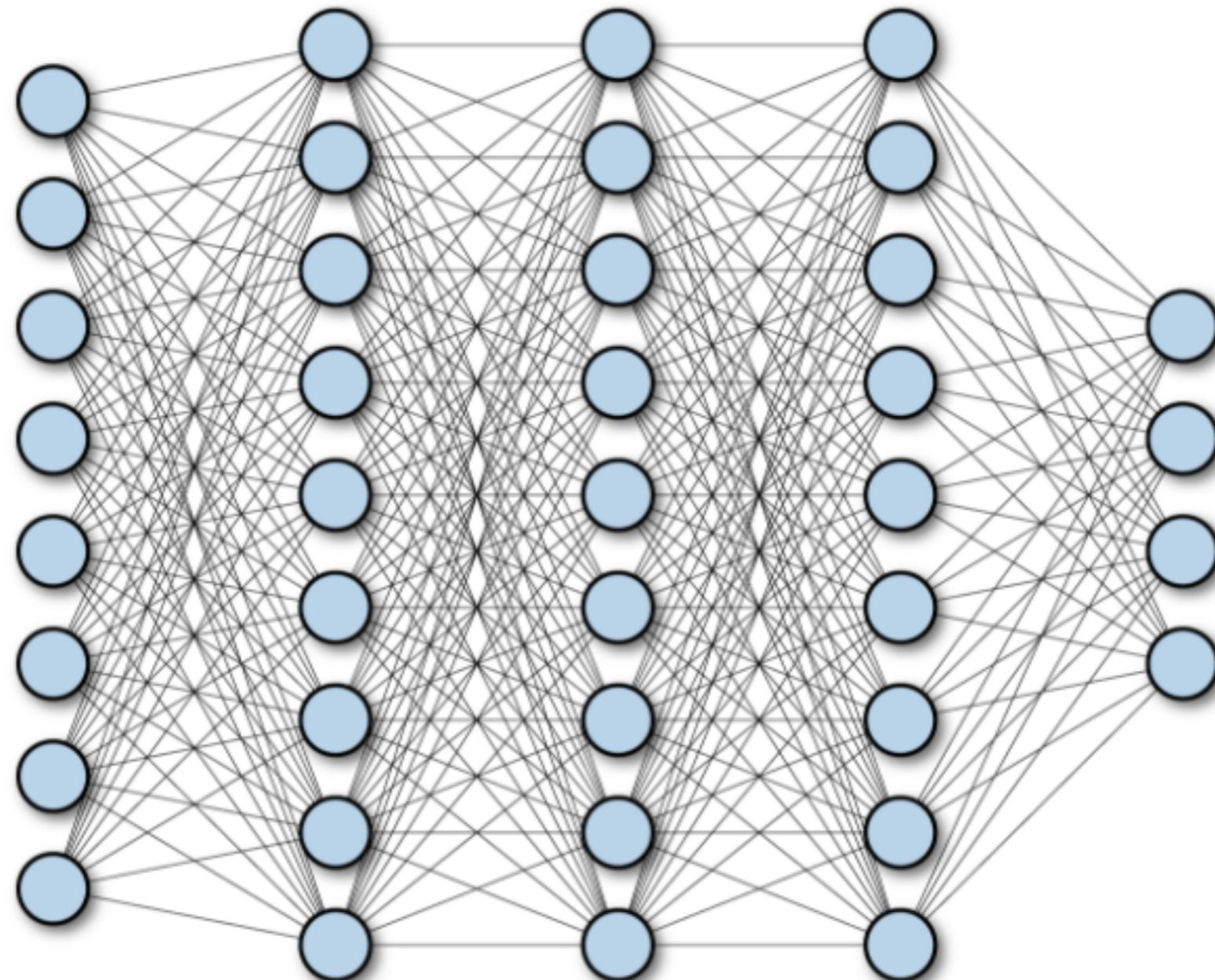


DEEP LEARNING



AI

Multilayer Deep Fully Connected Network



Neural Network !!

ANN

Artificial Neural Network

CNN

Convolutional Neural Network

RNN

Recurrent Neural Network

General purpose

Spatial data

Sequencial data

Variety of data types

Image

Time-series data

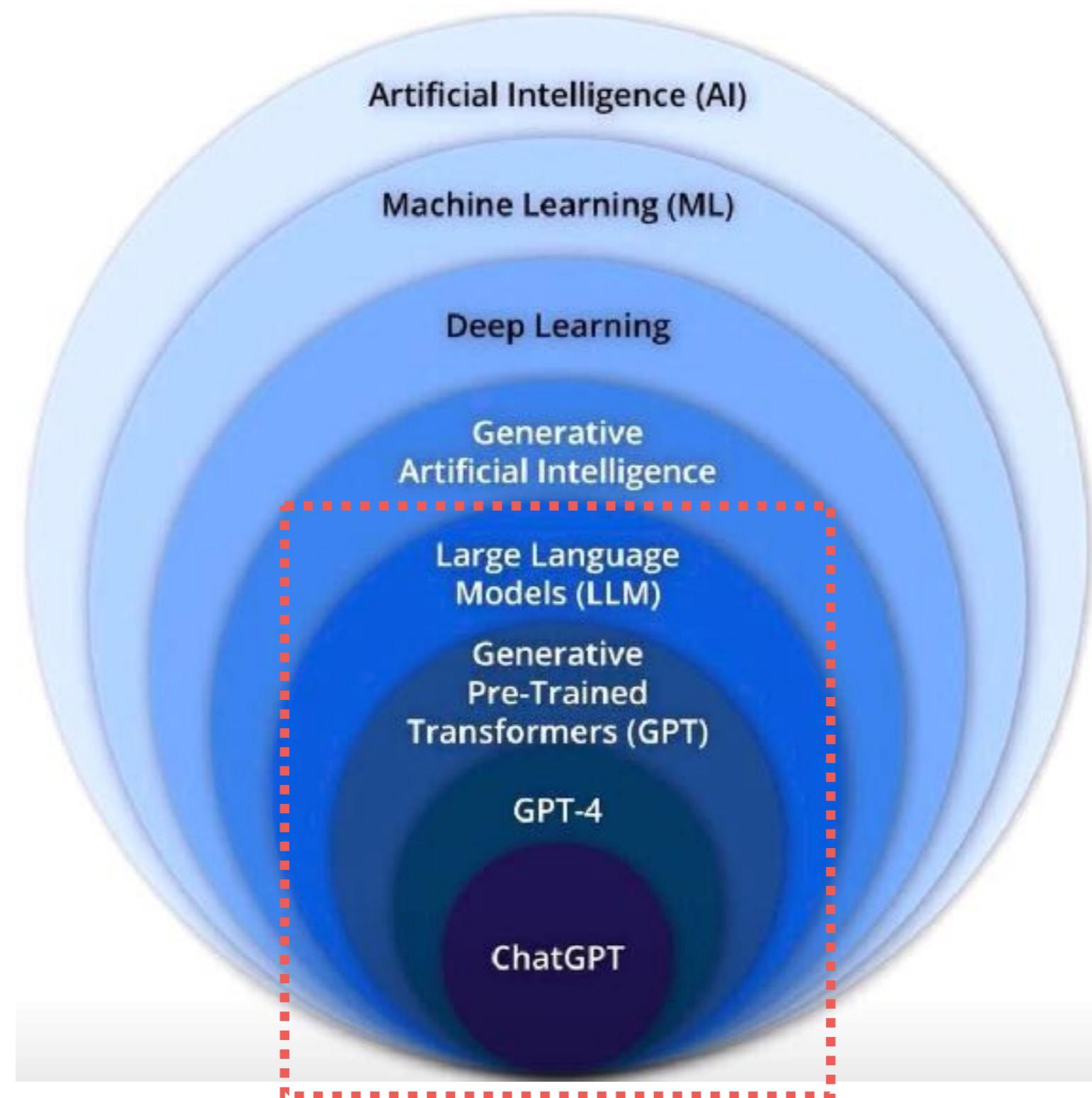
Simple data

Text

LSTM

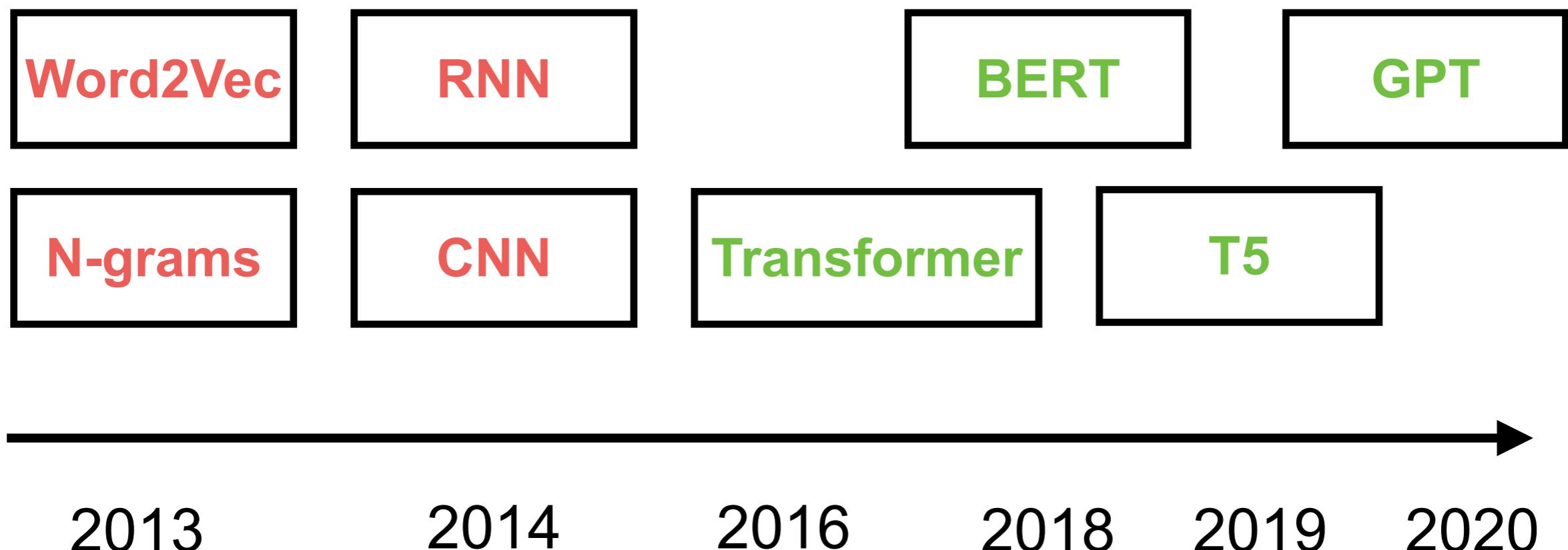
Long Short-Term Memory



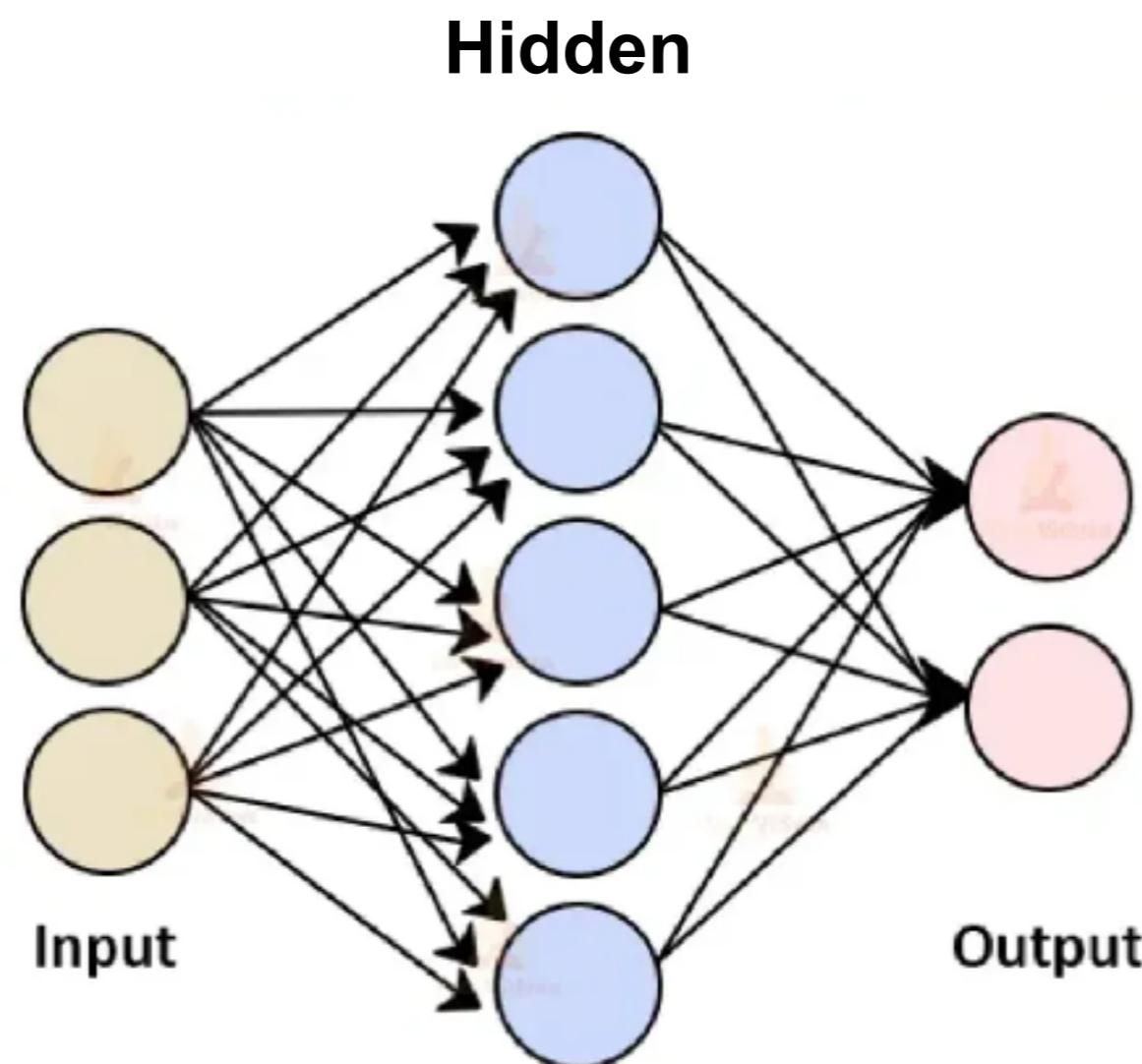


Language Model History

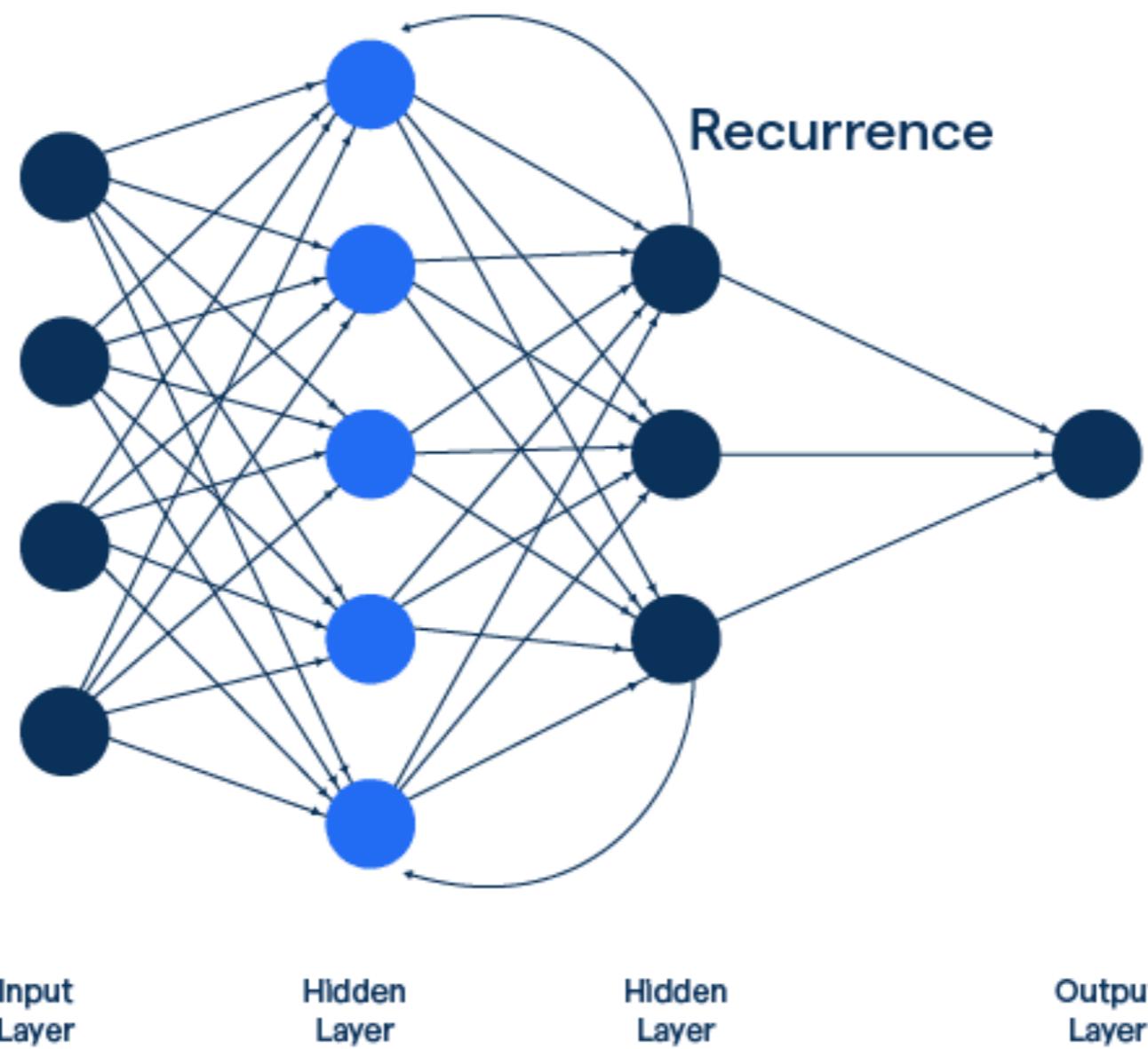
Trained to understand and generate human language



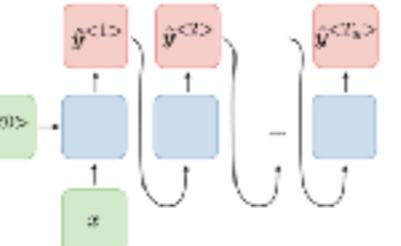
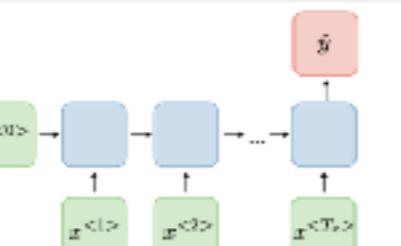
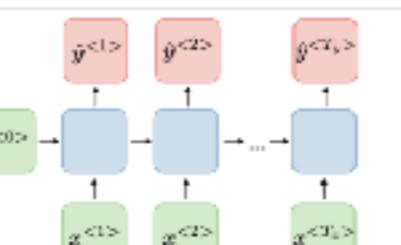
RNN + LSTM



Recurrent Neural Network



Types of RNN

Type of RNN	Illustration	Example
One-to-one $T_x = T_y = 1$		Traditional neural network
One-to-many $T_x = 1, T_y > 1$		Music generation
Many-to-one $T_x > 1, T_y = 1$		Sentiment classification
Many-to-many $T_x = T_y$		Name entity recognition
Many-to-many $T_x \neq T_y$		Machine translation

<https://aman.ai/primers/ai/dl-comp/>



How to apply RNN/LSTM in NLP (Natural Language Processing) task ?



How NLP Works

The machine responds with an audio file



A human talks to the machine



Data-to-audio conversion occurs



The machine captures the audio



The machine processes the text's data

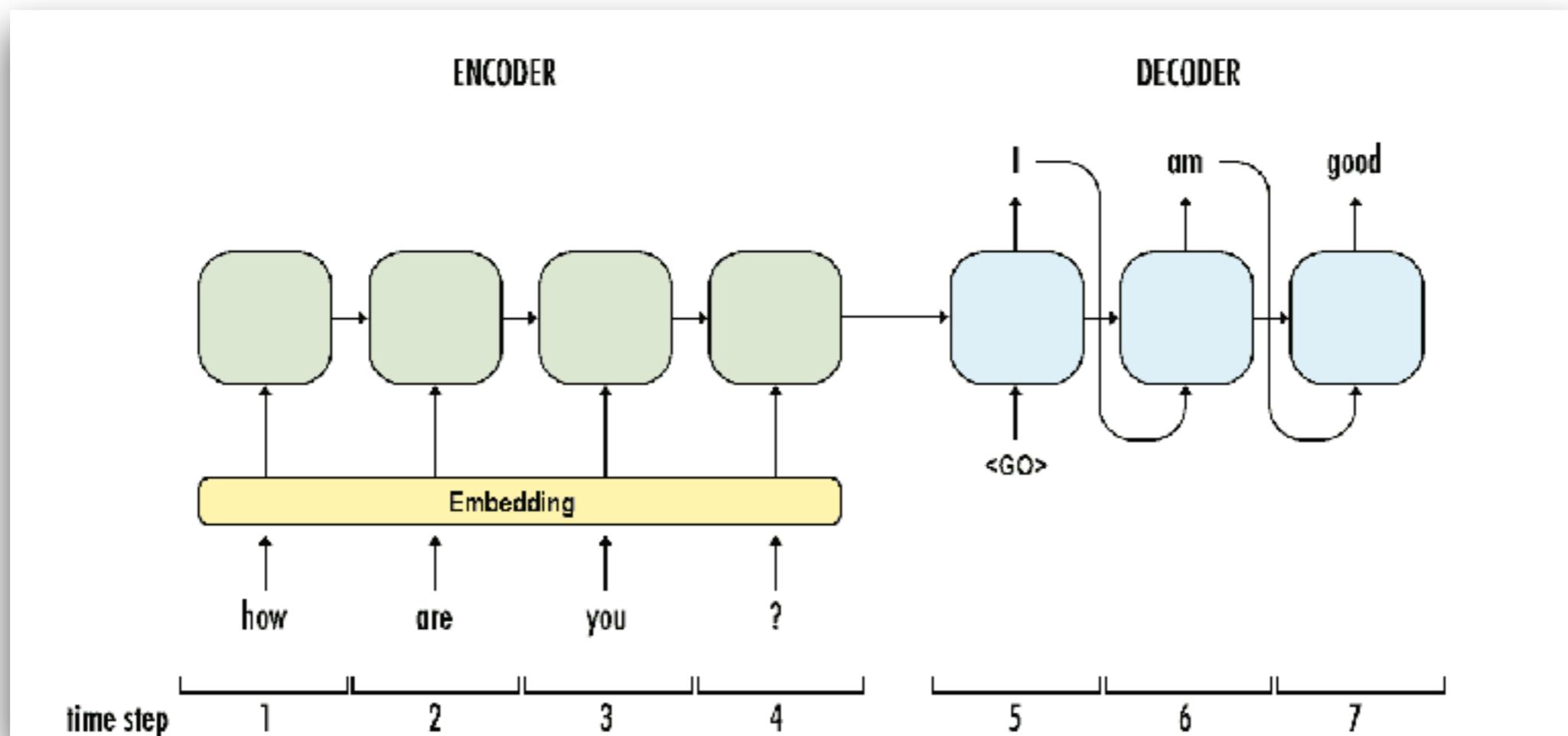


Audio-to-text conversion takes place



Sequence-to-Sequence model

Encoder-decoder model
Convert to **fixed length vector** data



Problems with RNN ?

Long input data

Not sequence data

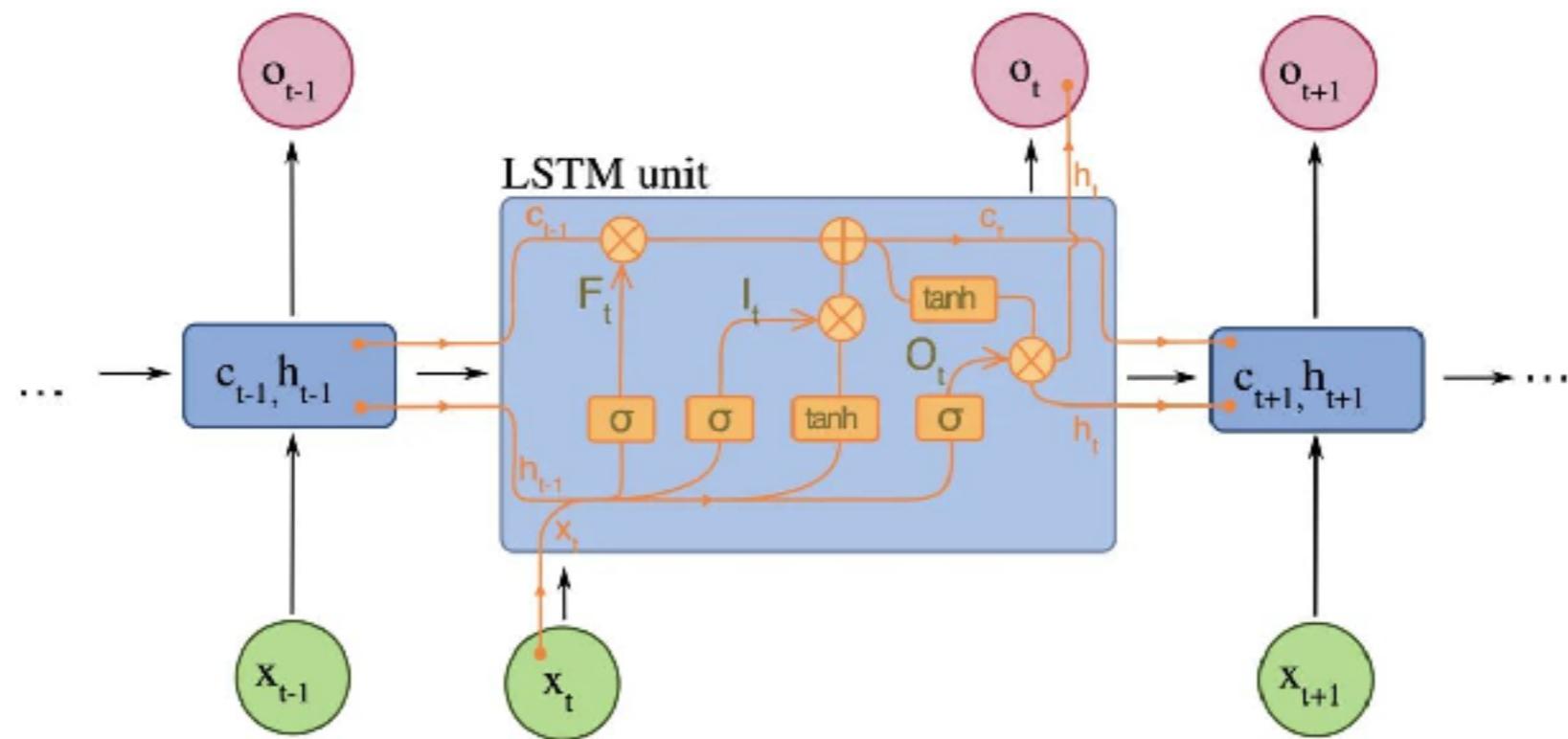
**Slow to train
(sequential processing)**

<https://jinglescode.github.io/2020/05/27/illustrated-guide-transformer/>



LSTM

Improve memory and deal with longer sequences than RNN



Slow

Complex



Drawback of RNN encoder-decoder

Fixed-length vector !!



Can't store large information



**Different sentence with similar words
BUT with different meanings !!**



Attention Mechanism

Improve performance of encoder-decoder

BUT still have bottleneck with RNN

<https://arxiv.org/pdf/1409.0473>

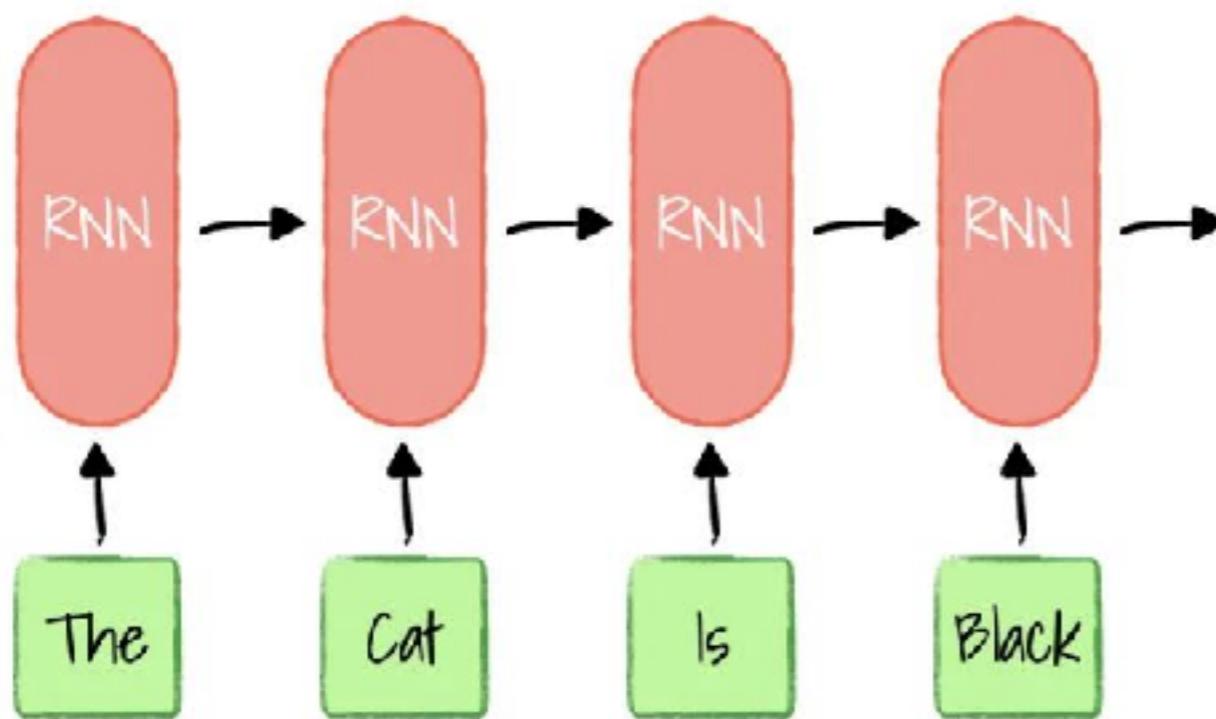


Can we remove RNN for sequential data ?

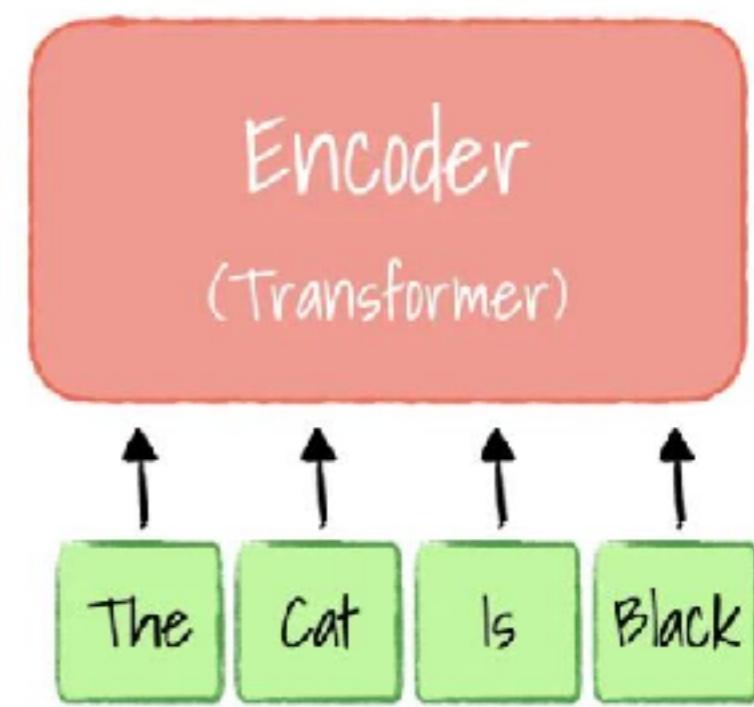


Parallel processing

RNN based Encoder



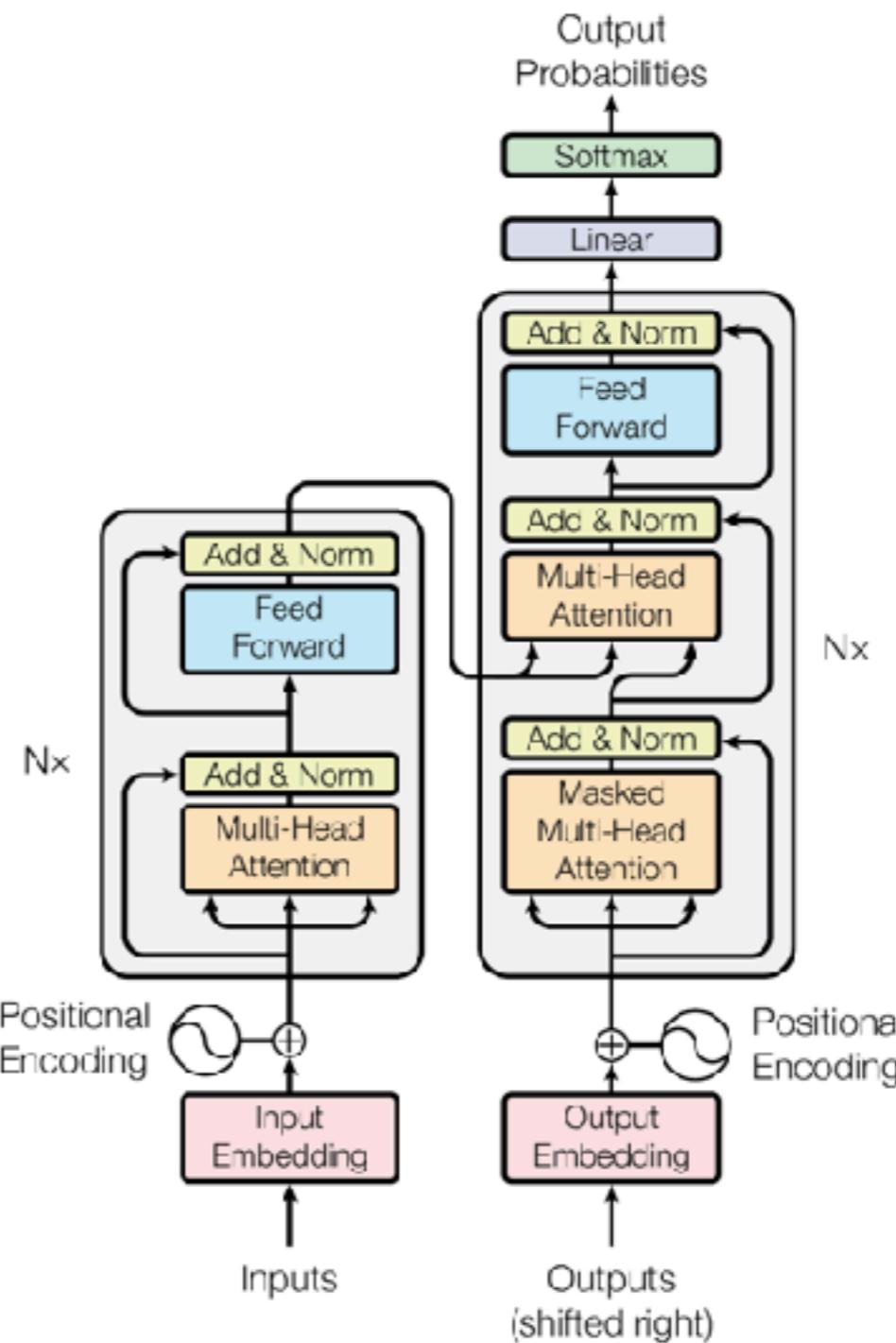
Transformer's Encoder



**Start using Transformer
apply attention mechanism**



Attention is all you need



<https://arxiv.org/abs/1706.03762>

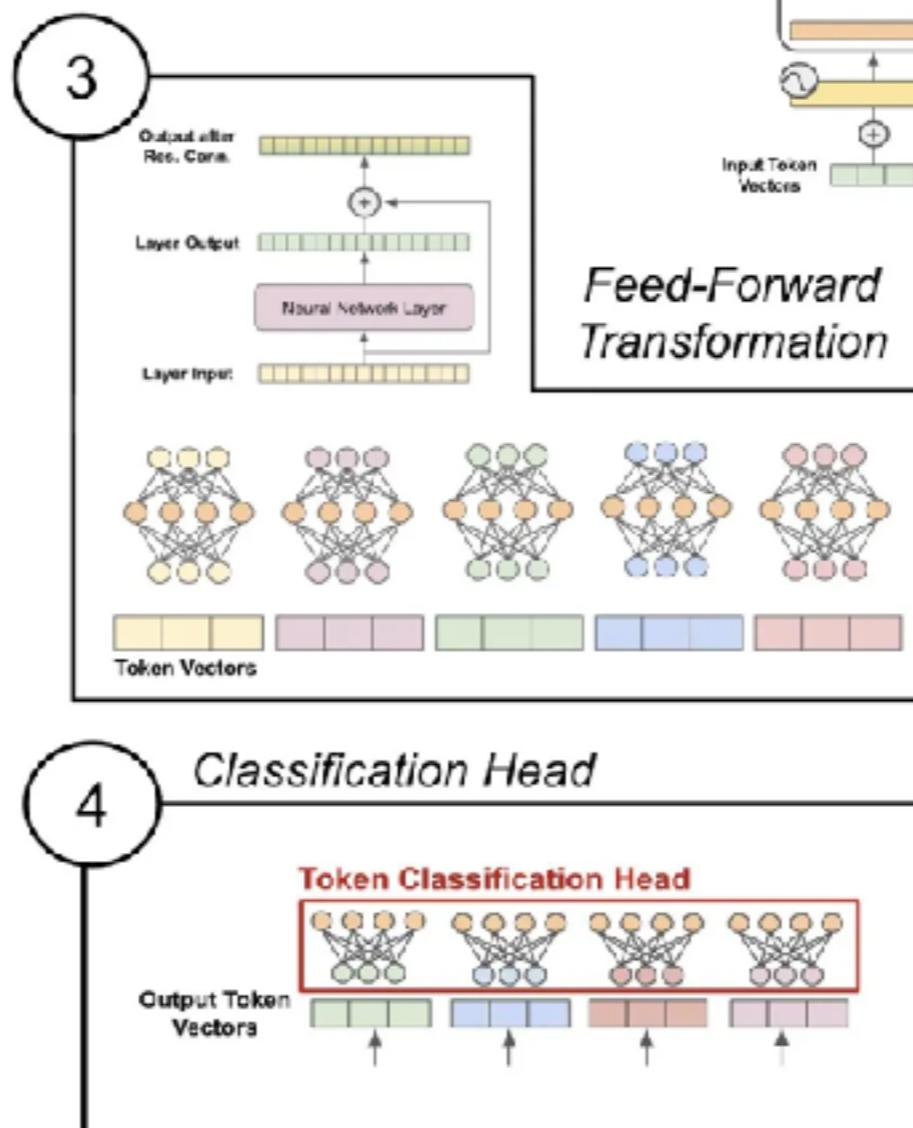
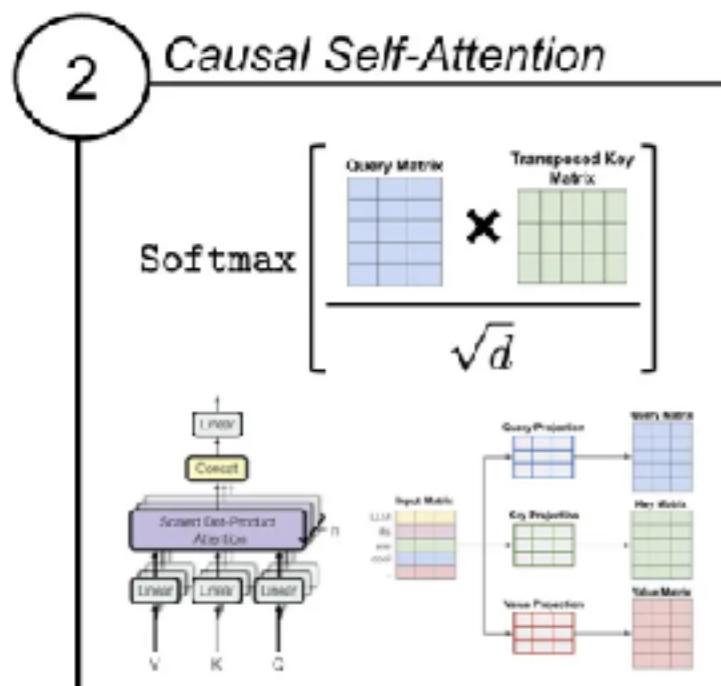
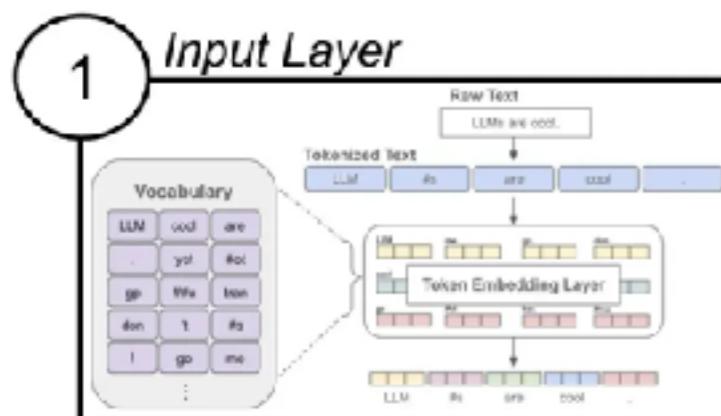


AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

LLM components !!

Components of the Decoder-only Transformer



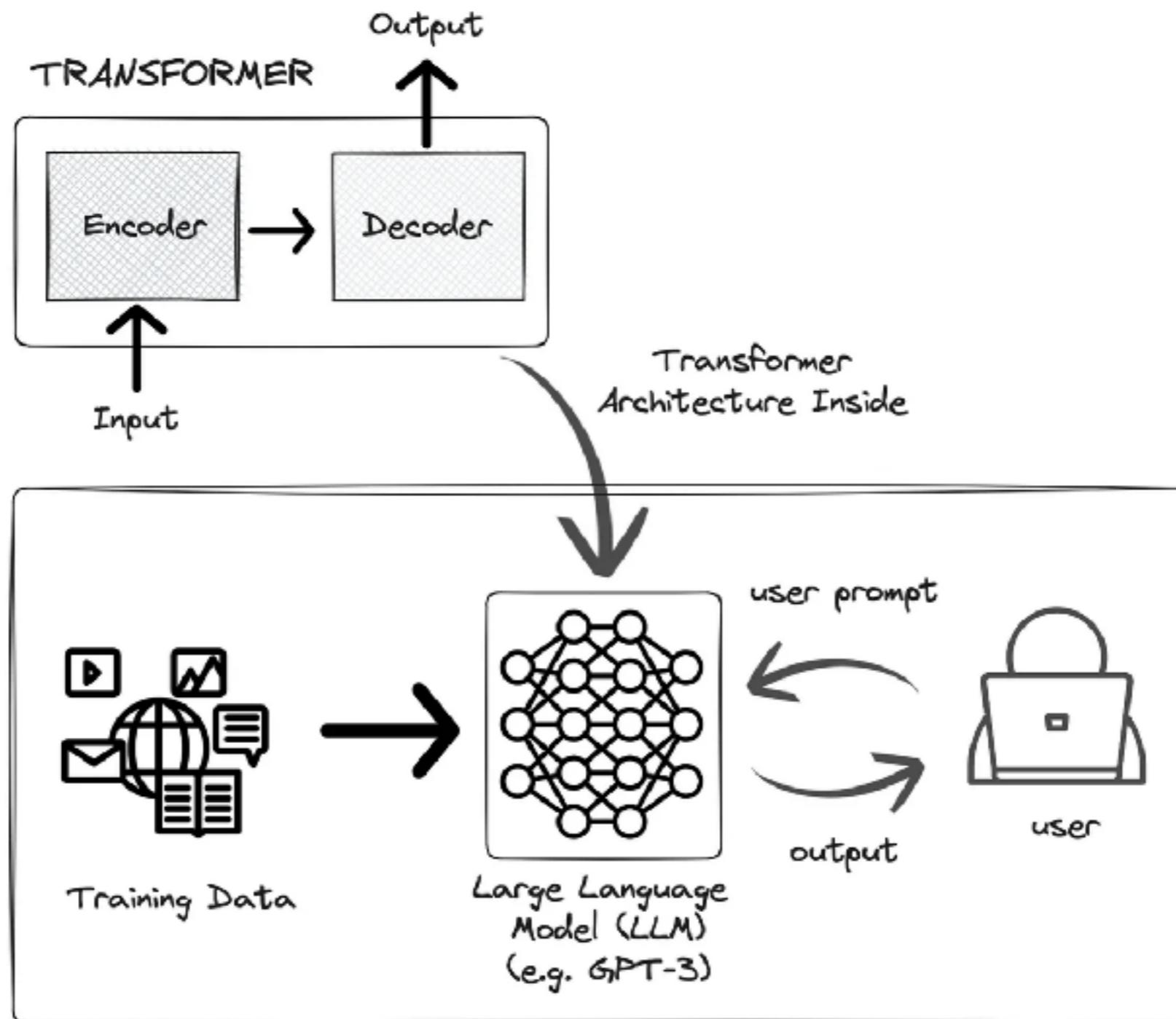
<https://stackoverflow.blog/2024/08/22/lms-evolve-quickly-their-underlying-architecture-not-so-much>



AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

Transformer inside



Transformer Models ?

BERT

Bidirectional
Encoder
Representations from
Transformers

GPT

Generative
Pre-trained
Transformer

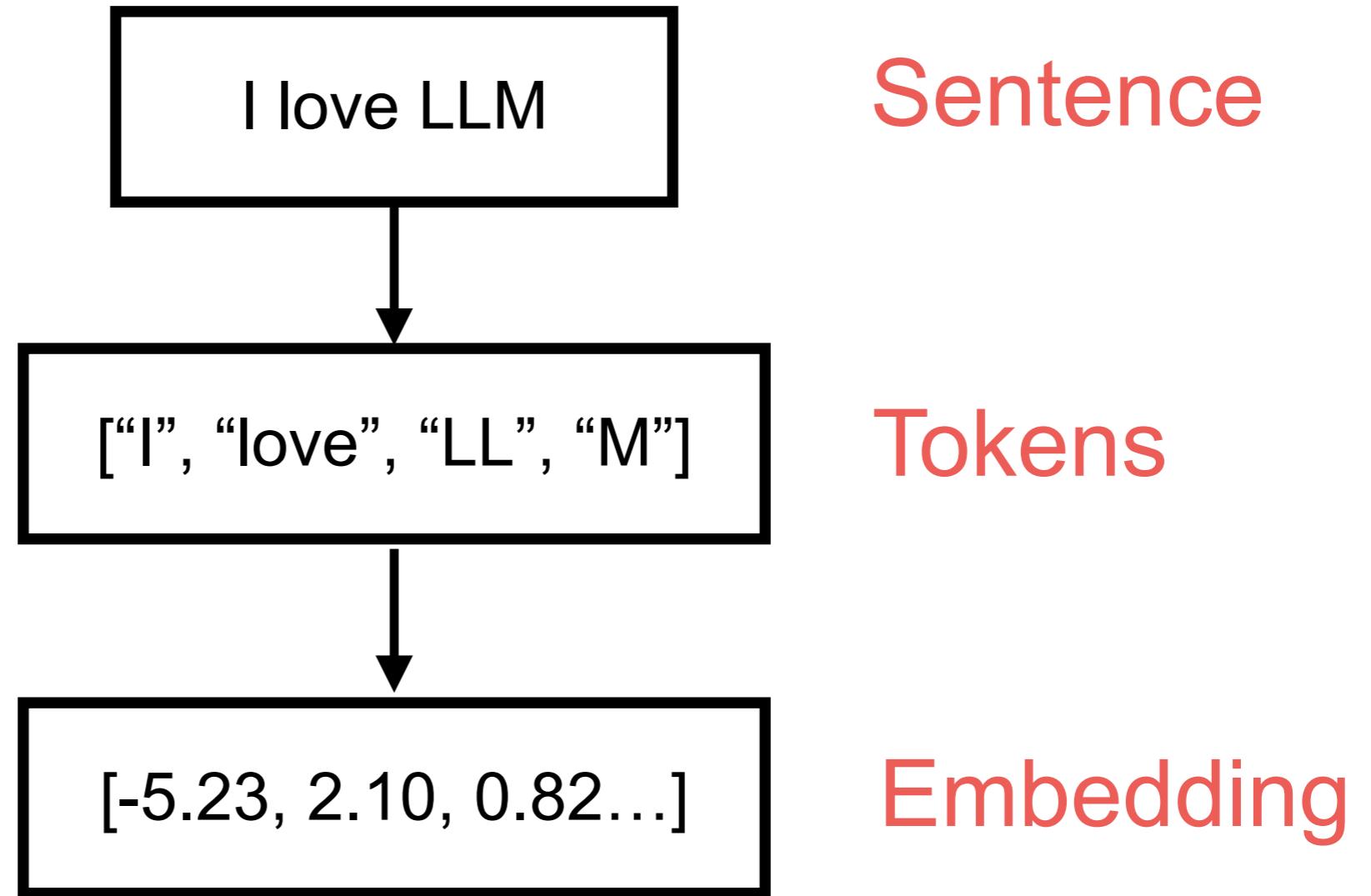


AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

34

Transformer process



OpenAI Tokenizer

GPT-4o & GPT-4o mini (coming soon) **GPT-3.5 & GPT-4** GPT-3 (Legacy)

ประเทศไทย

[Clear](#) [Show example](#)

Tokens **Characters**

10 9

ประเทศไทย

[Text](#) [Token IDs](#)

<https://platform.openai.com/tokenizer>

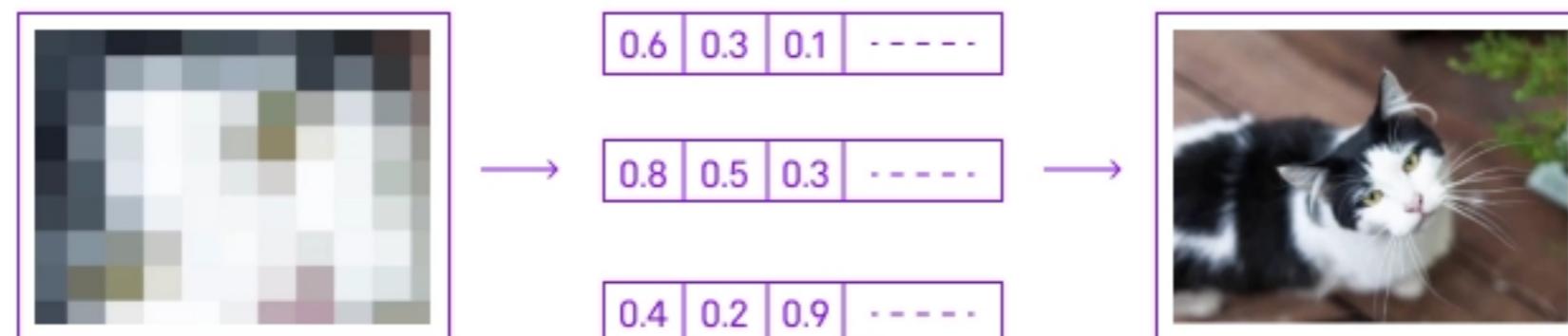


Vectorization and Embedding



Embedding ?

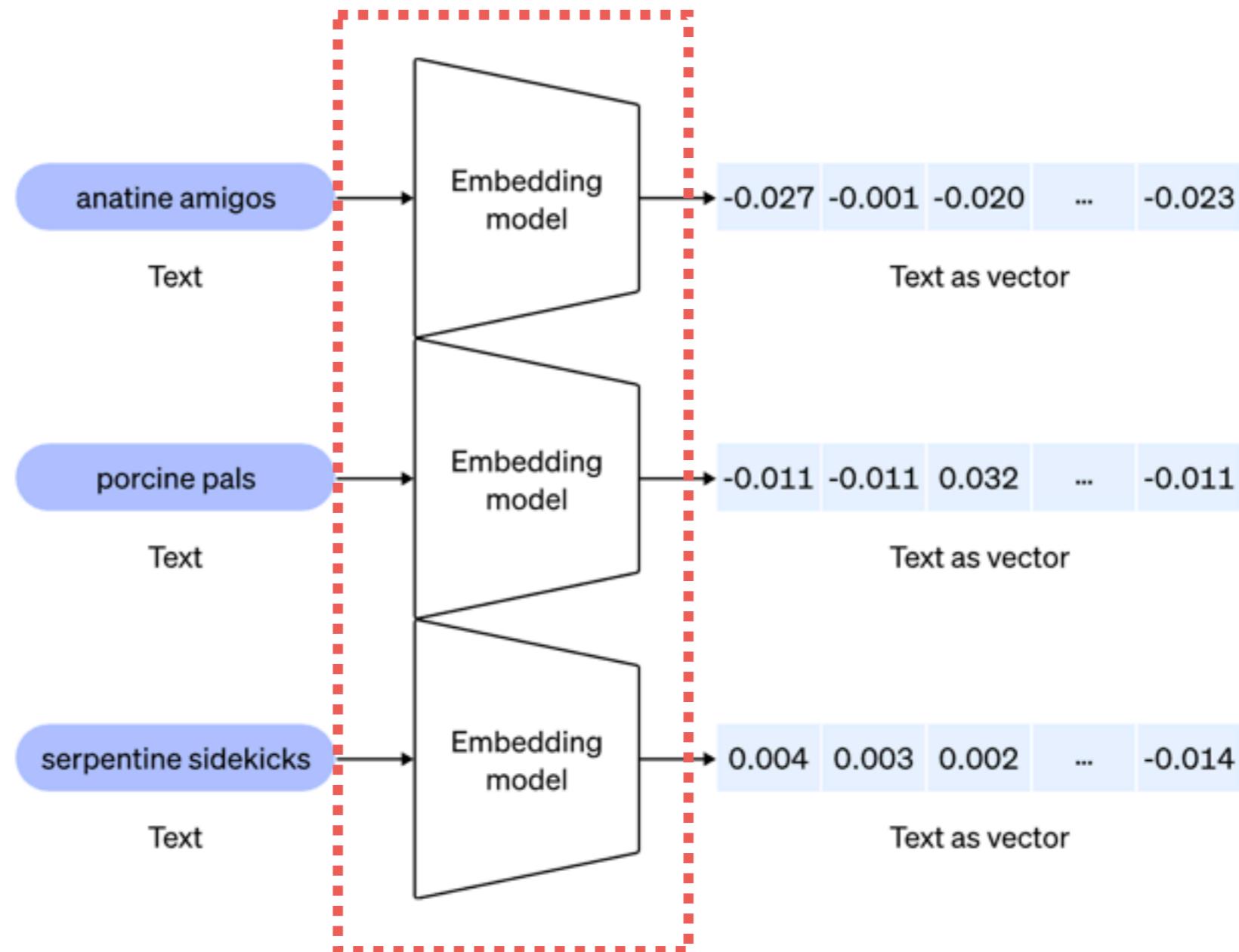
Map items of **unstructured data** to high-dimensional real vectors



<https://towardsdatascience.com transformers-in-depth-part-1-introduction-to-transformer-models-in-5-minutes-ad25da6d3cca>



Embedding models



<https://openai.com/index/new-embedding-models-and-api-updates/>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

Embedding models

Rank (Box..)	Model	Zero-shot	Memory U...	Number of P..	Embedding D..	Max Tokens	Mean (T..	Mean (TaskT..	Bitext ..	Classific
1	gemini-embedding-001	99%	Unknown	Unknown	3072	2048	68.37	59.59	79.28	71.82
2	Qwen3-Embedding-8B	99%	28966	7B	4096	32768	79.58	61.69	80.89	74.00
3	Qwen3-Embedding-4B	99%	15341	4B	2560	32768	69.45	60.86	79.36	72.33
4	Qwen3-Embedding-0.6B	99%	2272	595M	1024	32768	64.34	56.01	72.23	66.83
5	Ling-Embed-Mistral	99%	13563	7B	4096	32768	61.47	54.14	70.34	62.24
6	gte-Qwen2-7B-instruct	⚠ NA	29040	7B	3584	32768	62.51	55.93	73.92	61.55
7	multilingual-e5-large-instruct	99%	1058	560M	1024	514	63.22	55.08	80.13	64.94
8	SFR-Embedding-Mistral	95%	13563	7B	4096	32768	69.90	53.92	70.00	60.02
9	text-multilingual-embedding-002	99%	Unknown	Unknown	768	2048	62.16	54.25	70.73	64.64
9	GritLM-7B	99%	13813	7B	4096	4096	69.92	53.74	70.53	61.83
11	GritLM-8x7B	99%	89079	57B	4096	4096	69.49	53.31	68.17	61.55

Choose the right model for your use case !!

<https://huggingface.co/spaces/mteb/leaderboard>
<https://modal.com/blog/mteb-leaderboard-article>



AI

40

Types of embedding

**Word
embedding**

**Sentence
embedding**

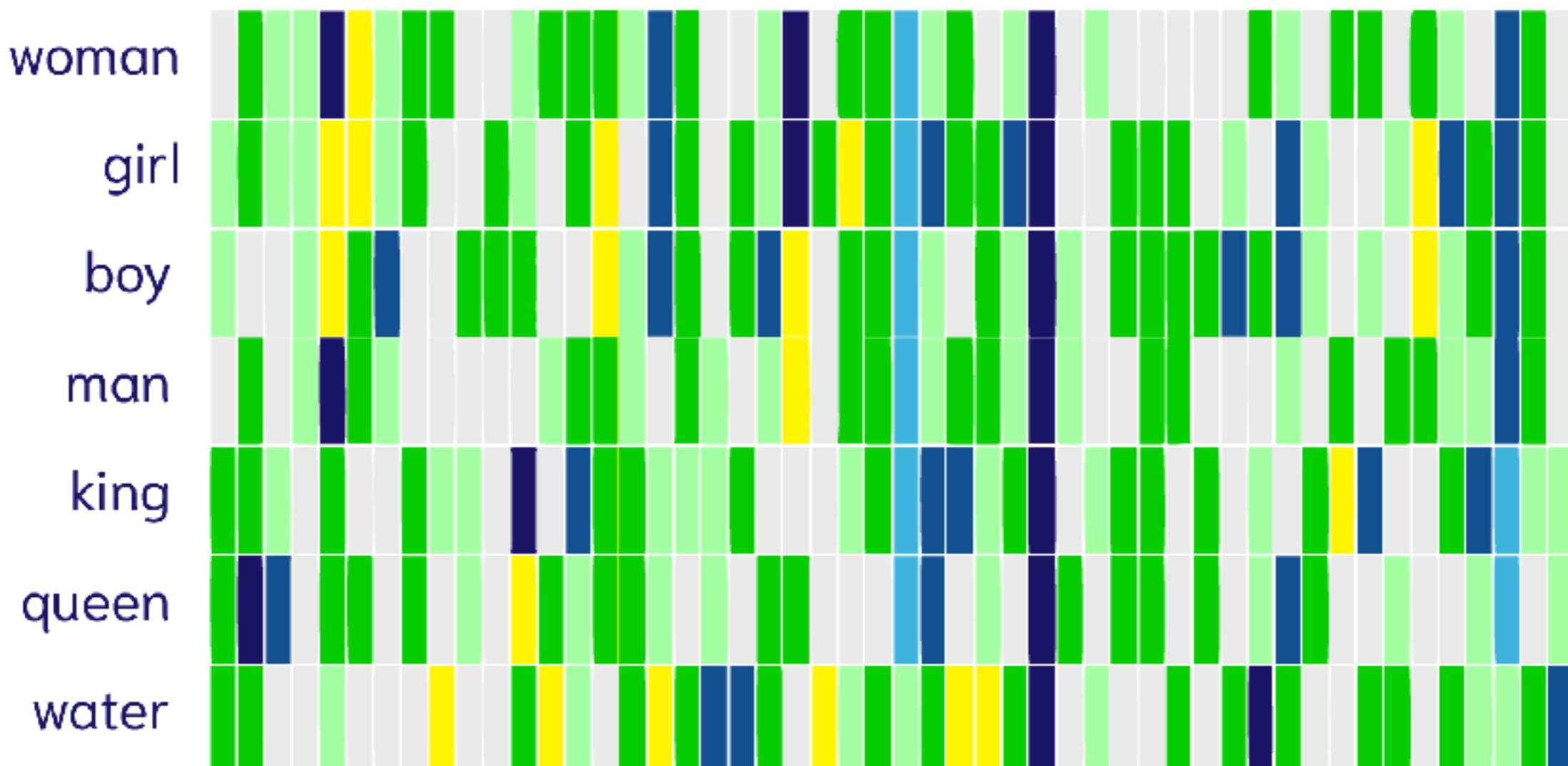
**Document
embedding**

**Multimodal
embedding**

<https://www.analyticsvidhya.com/blog/2024/09/vector-embeddings-with-cohere-and-huggingface/>



Word embedding with GloVe

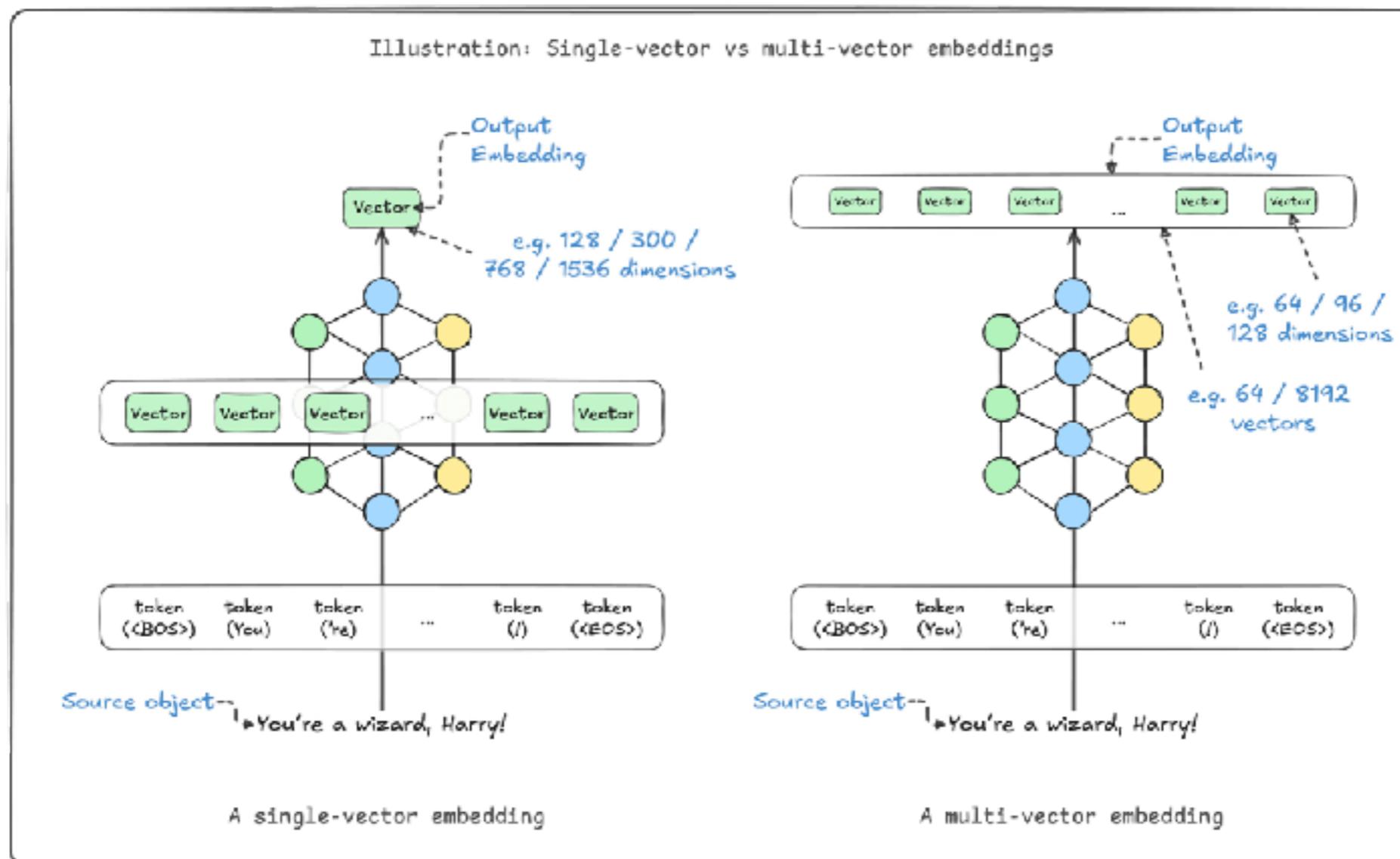


<https://en.wikipedia.org/wiki/GloVe>



Single vs Multi-vector Embedding

Improve retrieval and similar objects



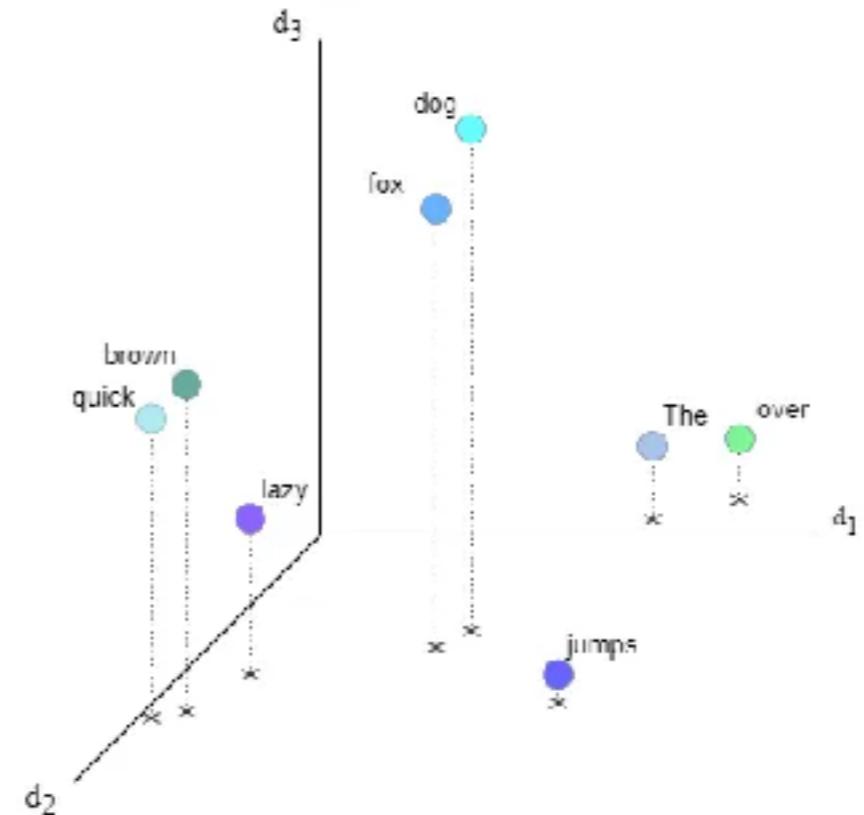
<https://docs.weaviate.io/weaviate/tutorials/multi-vector-embeddings>



Embedding ?

Map items of **unstructured data** to high-dimensional real vectors

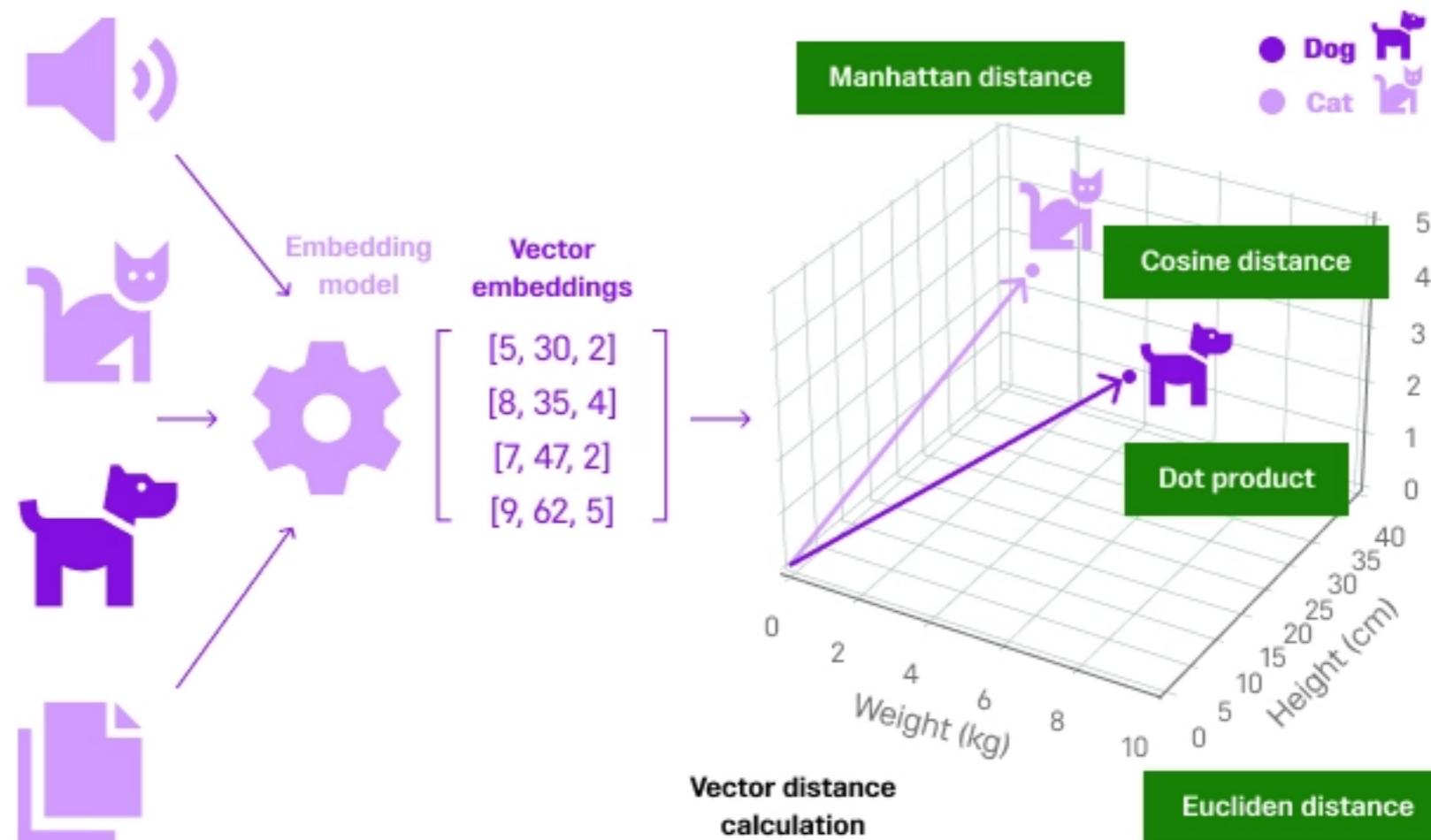
	0	1	2	d_{model}
The	0.64	-0.09	0.23	0.005
quick	0.05	0.79	0.47	-0.54
brown	0.12	0.71	0.51	-0.3
fox	0.12	0.52	0.83	0.01
jumps	0.88	0.69	0.02	0.27
over	0.84	0.15	0.13	0.05
the	0.64	0.00	0.23	0.005
lazy	0.1	0.65	0.28	0.19
dog	0.54	0.49	0.90	0.000



<https://towardsdatascience.com transformers-in-depth-part-1-introduction-to-transformer-models-in-5-minutes-ad25da6d3cca>



Similarity search with Vecrtor



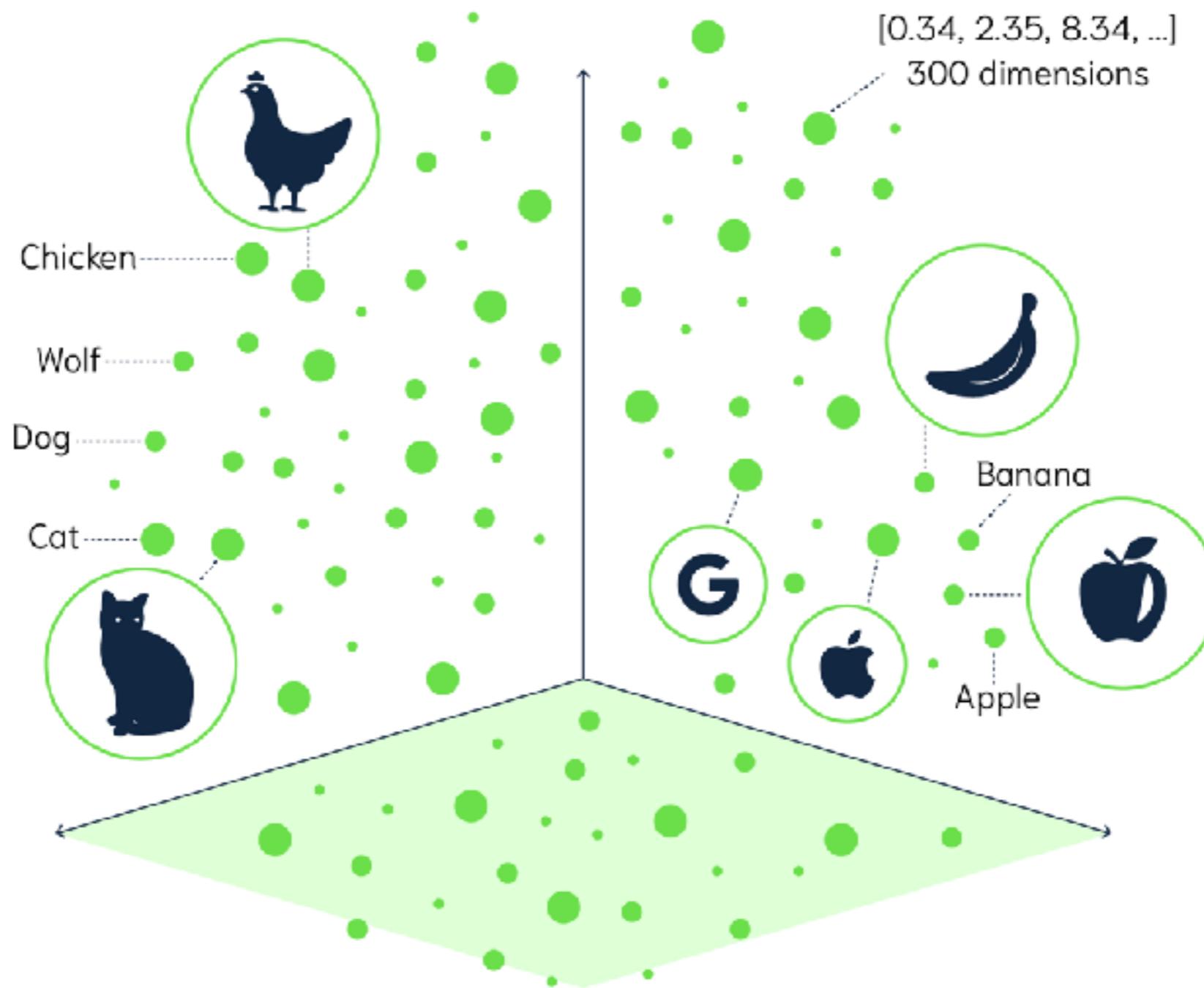
<https://www.singlestore.com/blog/distance-metrics-in-machine-learning-simplified/>



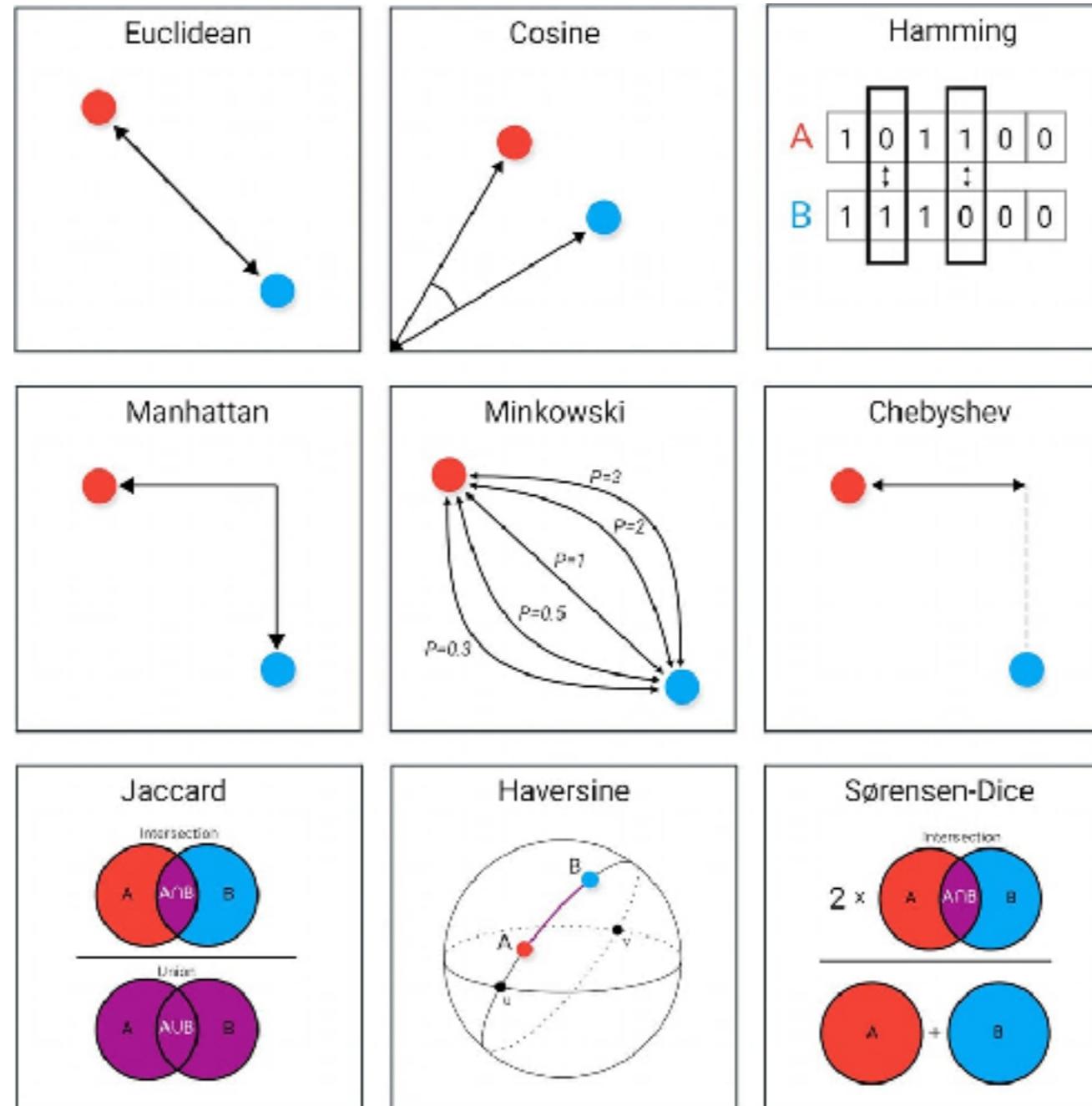
AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

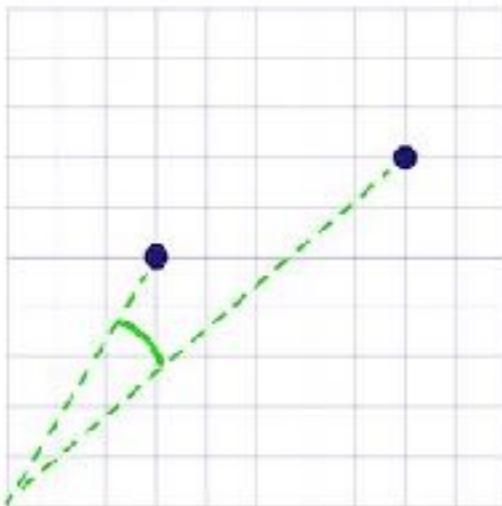
Visual of Vector space



Distance measure !!

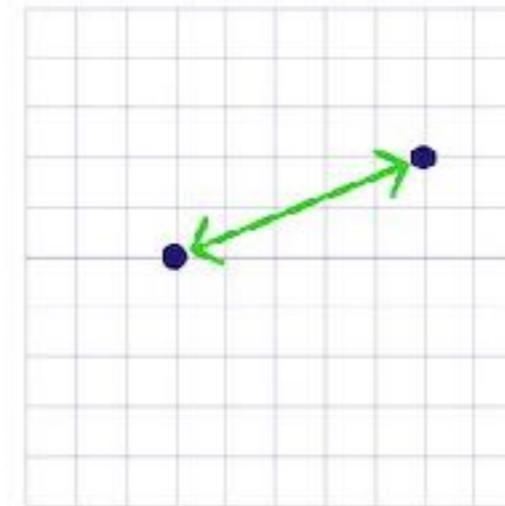


Distance Metrics in Vector Search



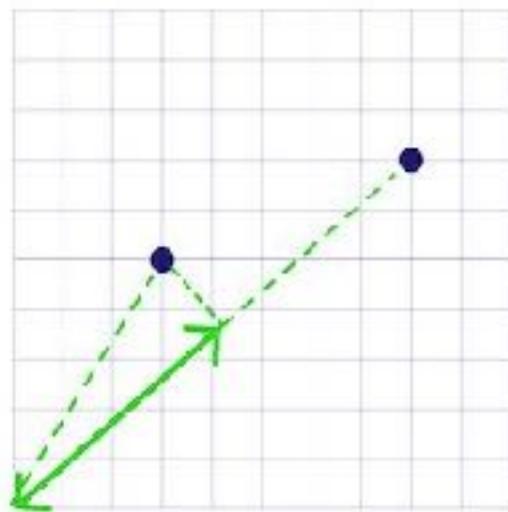
Cosine Distance

$$1 - \frac{A \cdot B}{\|A\| \|B\|}$$



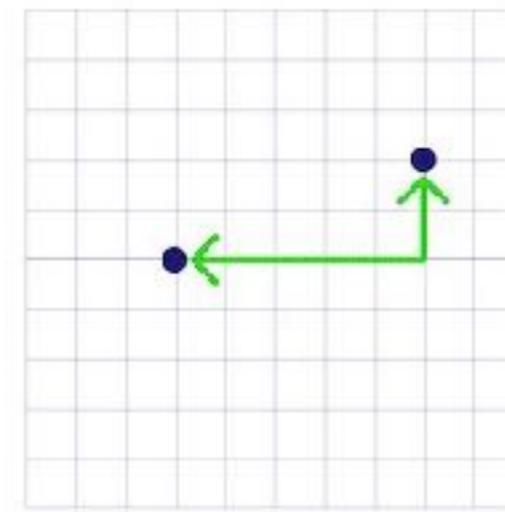
Squared Euclidean
(L2 Squared)

$$\sum_{i=1}^n (x_i - y_i)^2$$



Dot Product

$$A \cdot B = \sum_{i=1}^n A_i B_i$$

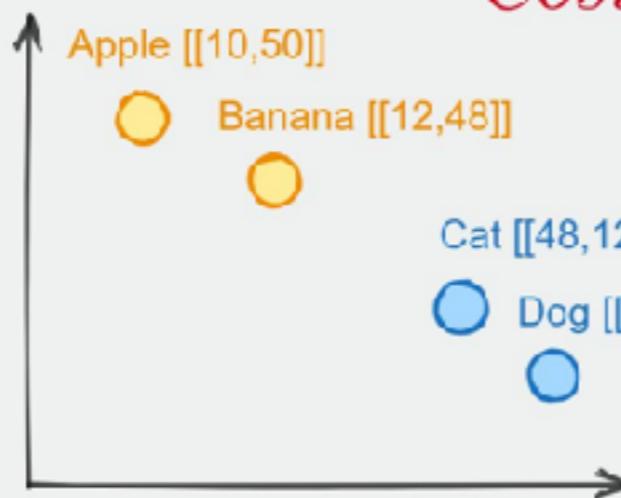


Manhattan (L1)

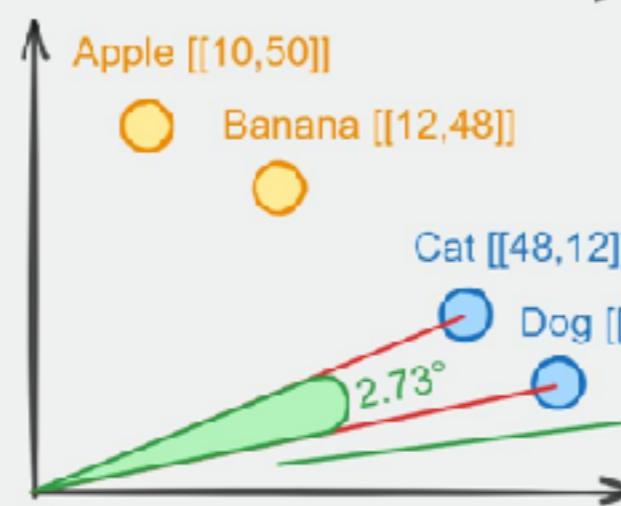
$$\sum_{i=1}^n |x_i - y_i|$$



Cosine in LLM



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



$$\mathbf{A} \cdot \mathbf{B} = 50 \times 48 + 10 \times 12 = 2400 + 120 = 2520$$

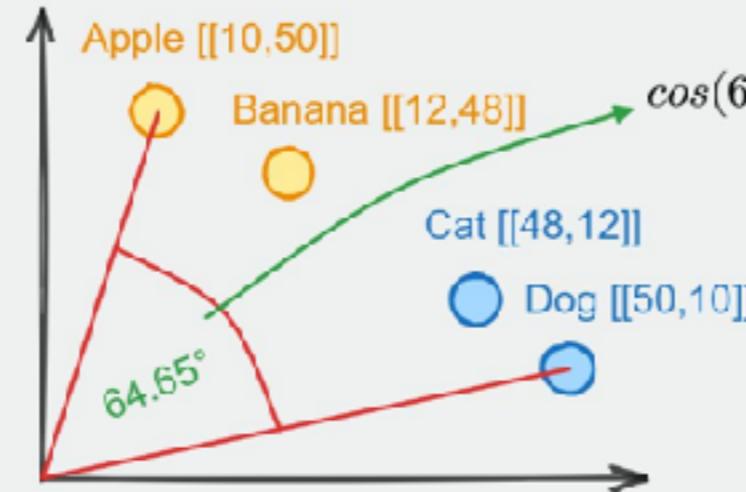
$$\|\mathbf{B}\| = \sqrt{48^2 + 12^2} = \sqrt{2304 + 144} = \sqrt{2448} \approx 49.5$$

$$\|\mathbf{A}\| = \sqrt{50^2 + 10^2} = \sqrt{2500 + 100} = \sqrt{2600} \approx 51.0$$

$$\text{cosine similarity} = \frac{2520}{51.0 \times 49.5} \approx 0.998$$

$\cos(2.73^\circ) \approx 0.998$

Cat & Dog are similar!



$$\cos(64.65^\circ) \approx 0.4284$$

Apple & Dog are not similar!



<https://x.com/levikul09/status/1771843190745948233/photo/1>



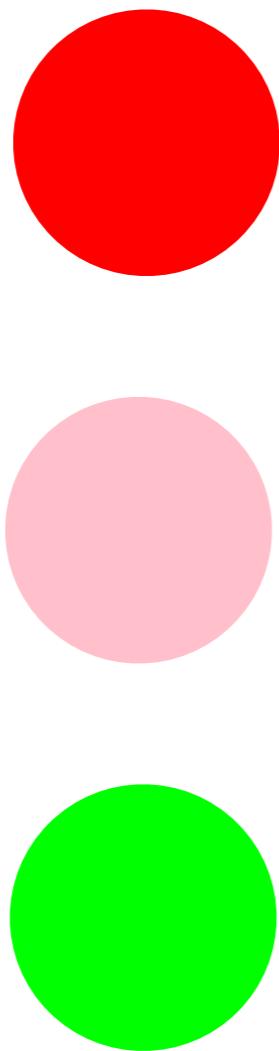
AI

49

Example with Vector embedding ?



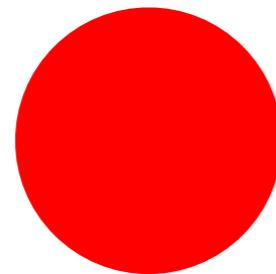
RGB ?



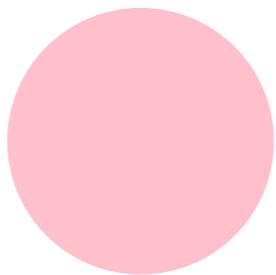
<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/demo-rgb>



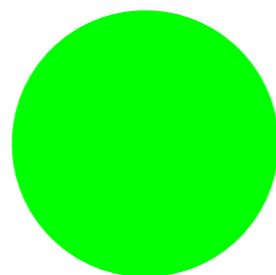
RGB ?



[255, 0, 0]



[255, 192, 203]



[0, 255, 0]

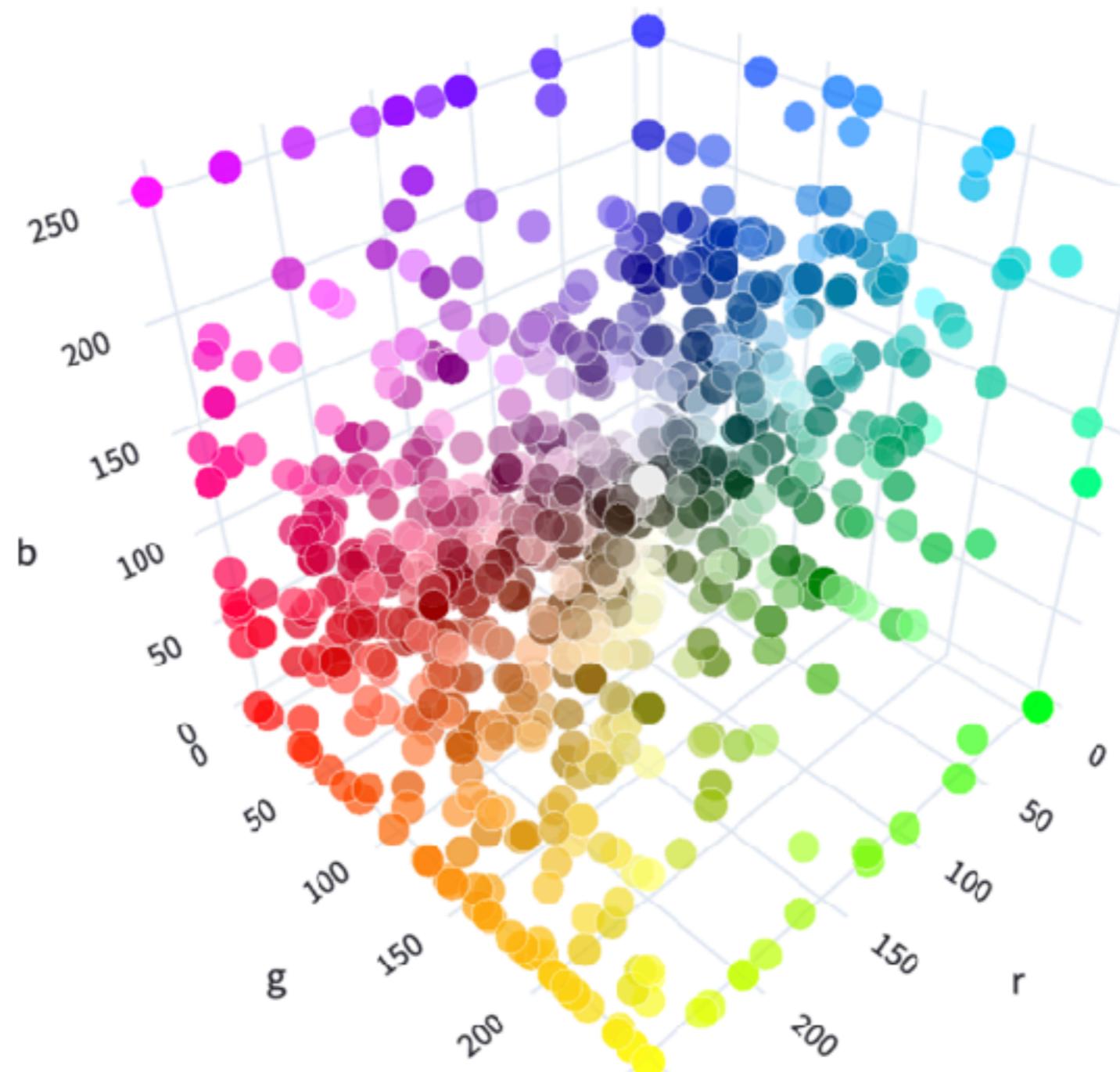
<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/demo-rgb>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

Visualize RGB



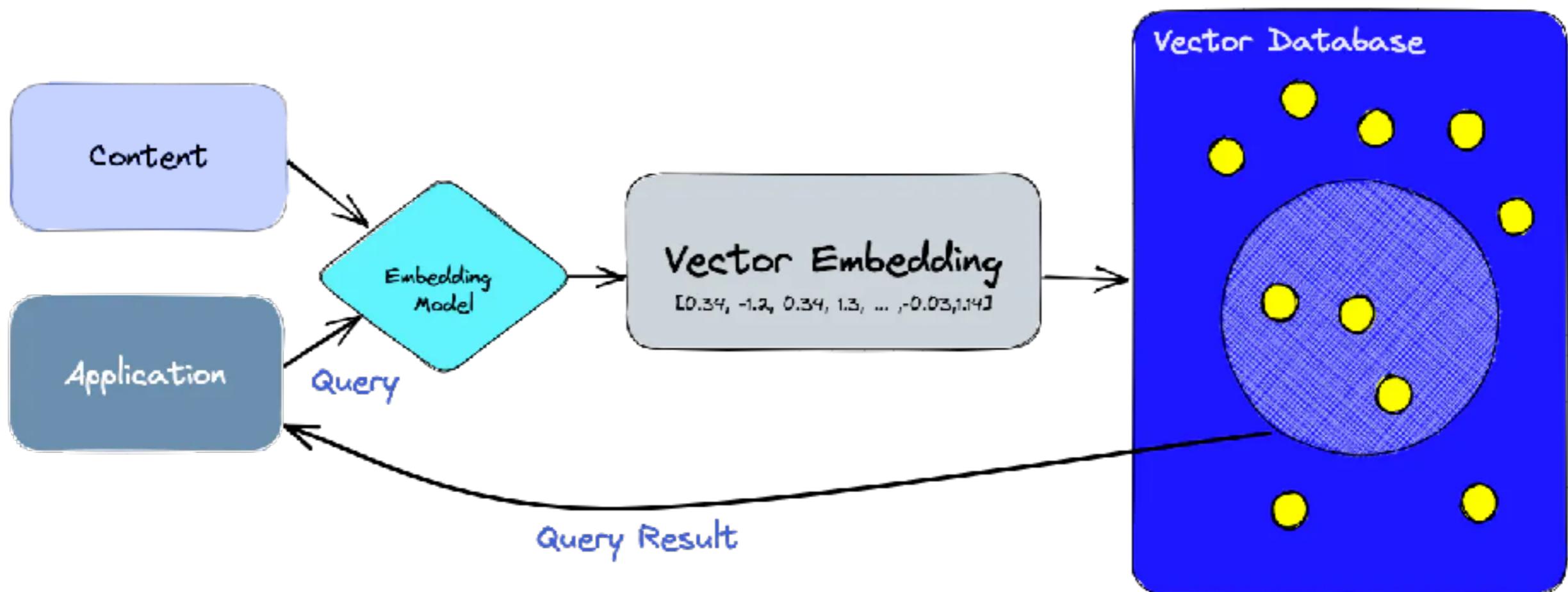
https://huggingface.co/spaces/jphwang/colorful_vectors



Store data in Vector Database



Store data in Vector Database



<https://www.pinecone.io/learn/vector-database/>



Vector Database ?

Index and Store vector embedding
Fast retrieval and similar search

Keyword search

Similarity search

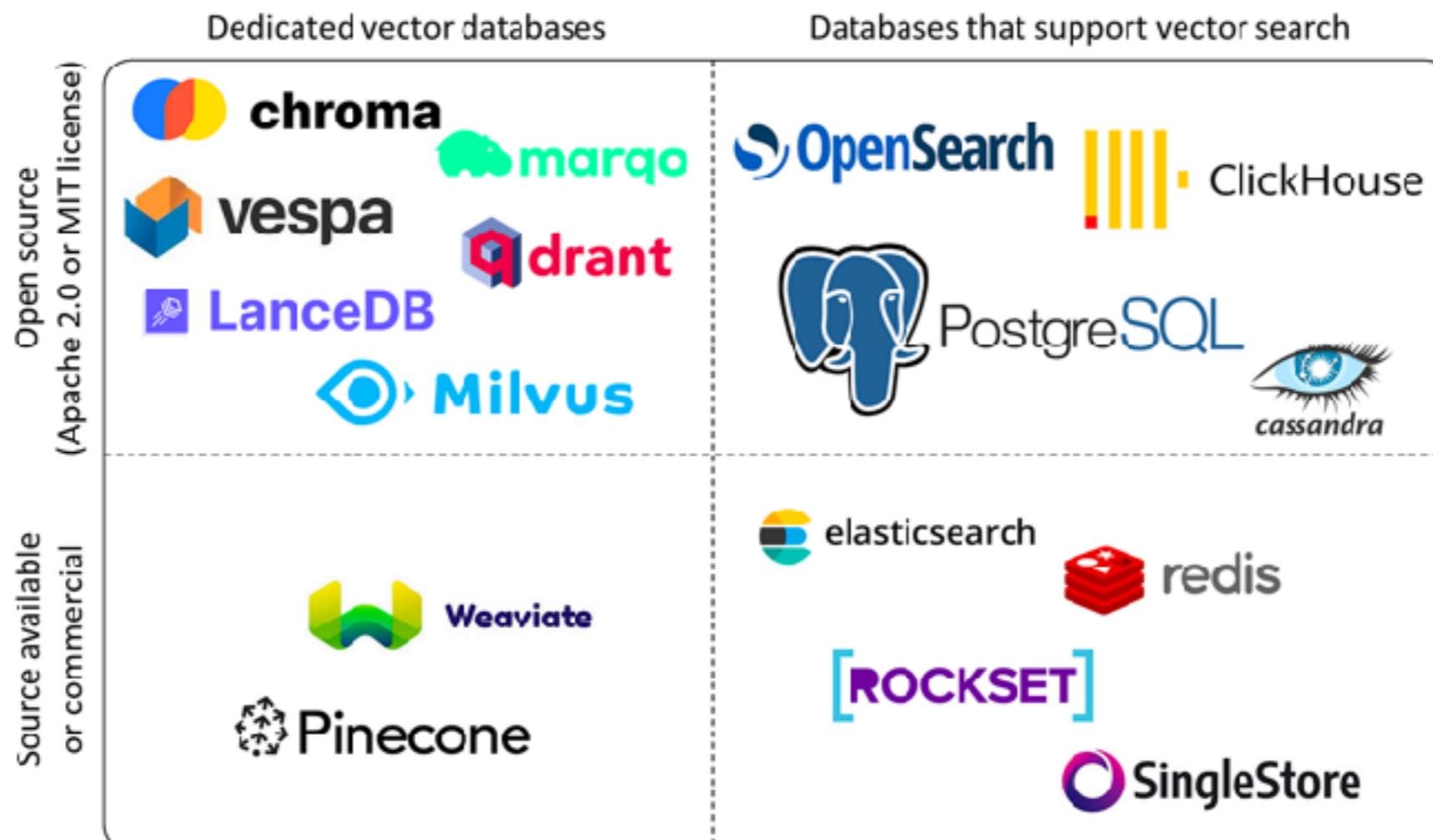
Full-text search

Semantic search



Vector Database ?

Map items of unstructured data to high-dimensional real vectors



<https://towardsdatascience.com/transformers-in-depth-part-1-introduction-to-transformer-models-in-5-minutes-ad25da6d3cca>



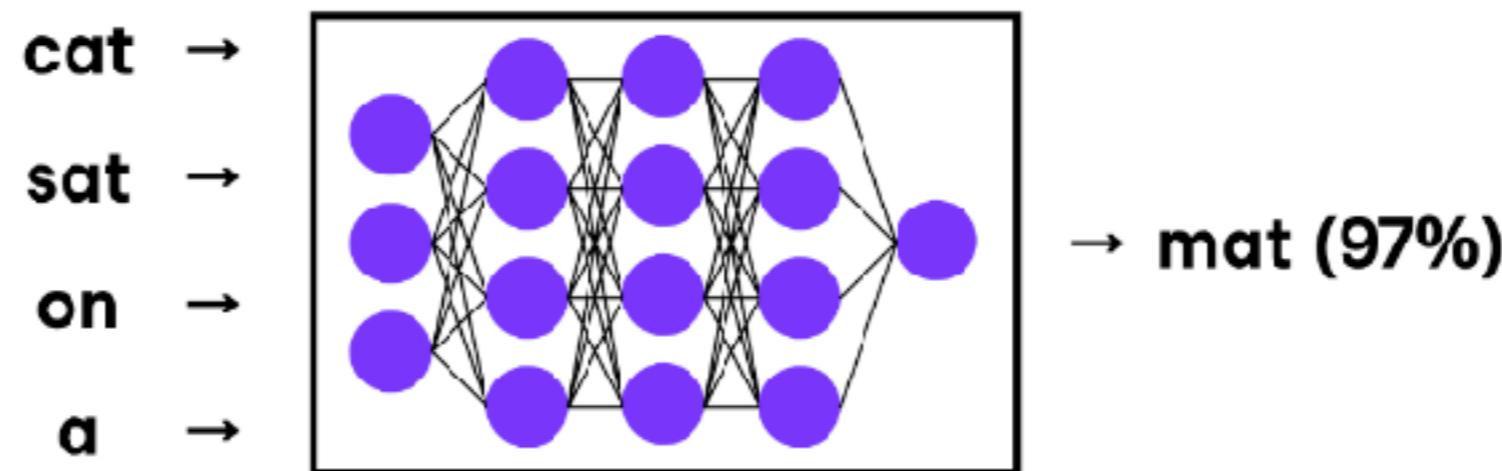
Let's go !!



Large Language Model (LLM)

Neural network

Predicts the next word in a sequence



LLM Development timeline

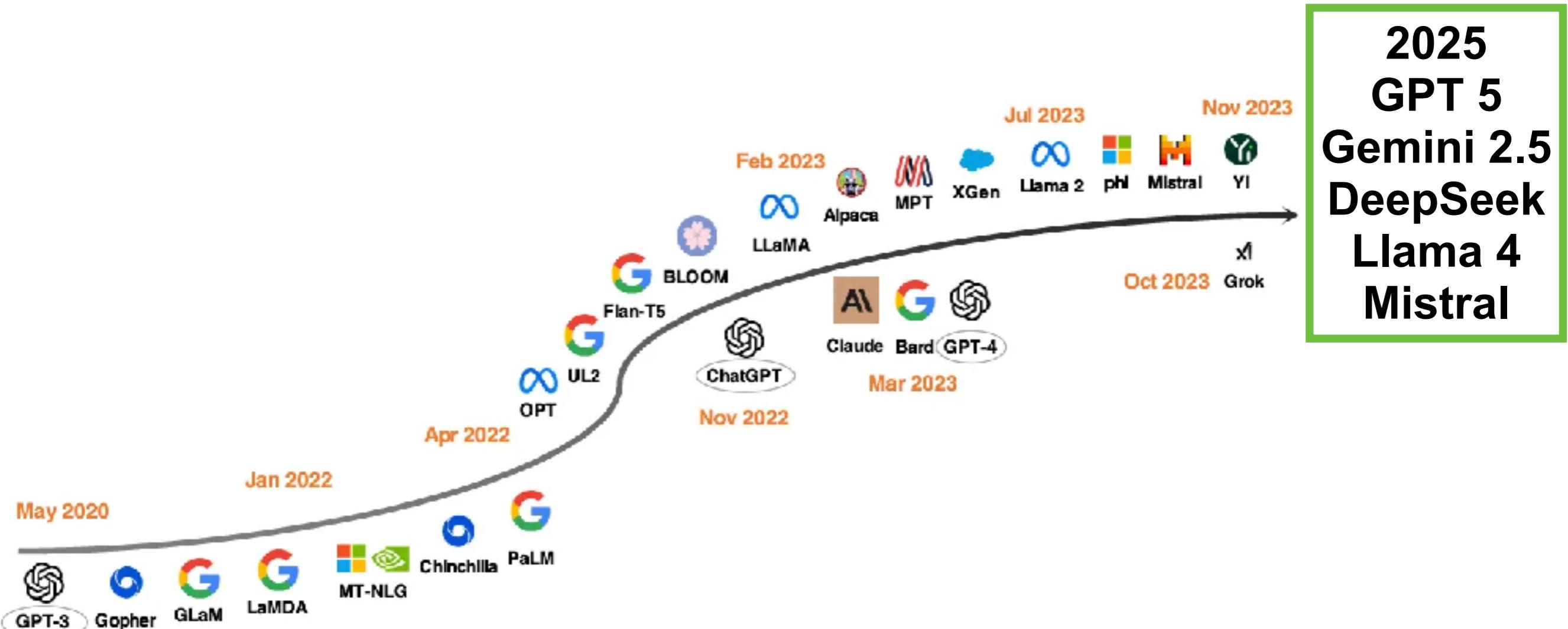


Figure 3: LLM development timeline. The models below the arrow are closed-source while those above the arrow are open-source.

<https://arxiv.org/abs/2311.16989>



LLMs in industry

Model name	Company
GPT 4	OpenAI
Gemini	Google AI
BLOOM	NVIDIA AI
Llama	Facebook/Meta
Claude Sonnet, Haiku, Opus	Anthropic
Phi	Microsoft
DeepSeek	DeepSeek





Module 2

Introduction to Claude AI



Claude AI

The screenshot shows the Claude AI web interface. At the top, there is a "Free plan" button and an "Upgrade" link. Below that, a greeting says "Hi Somkiat, how are you?". A large input field asks "How can I help you today?". Below the input field are three buttons: a plus sign, a document icon, and a double equals sign. To the right of the input field, a dropdown menu is open, showing the current selection "Claude Sonnet 4" with a red up arrow button. The dropdown also lists other models: "Claude Opus 4.1" (PRO), "Claude Sonnet 4" (selected, marked with a checkmark), "More models" (with a right arrow), "Claude Opus 4" (PRO), "Claude Sonnet 3.7", "Claude Opus 3", and "Claude Haiku 3.5" (PRO). The "Claude Opus 4.1" entry includes the description "Powerful, large model for complex challenges". The "Claude Opus 4" entry includes the description "Fastest model for daily tasks".

<https://claude.ai/>



Anthropic

ANTHROPIC

Claude ▾ API ▾ Solutions ▾ Research ▾ Commitments ▾ Learn ▾ News

Try Claude

AI research and products that put safety at the frontier

CLAUDE.AI

Meet Claude Opus 4.1

Claude Opus 4.1, our most intelligent AI model, is now available.

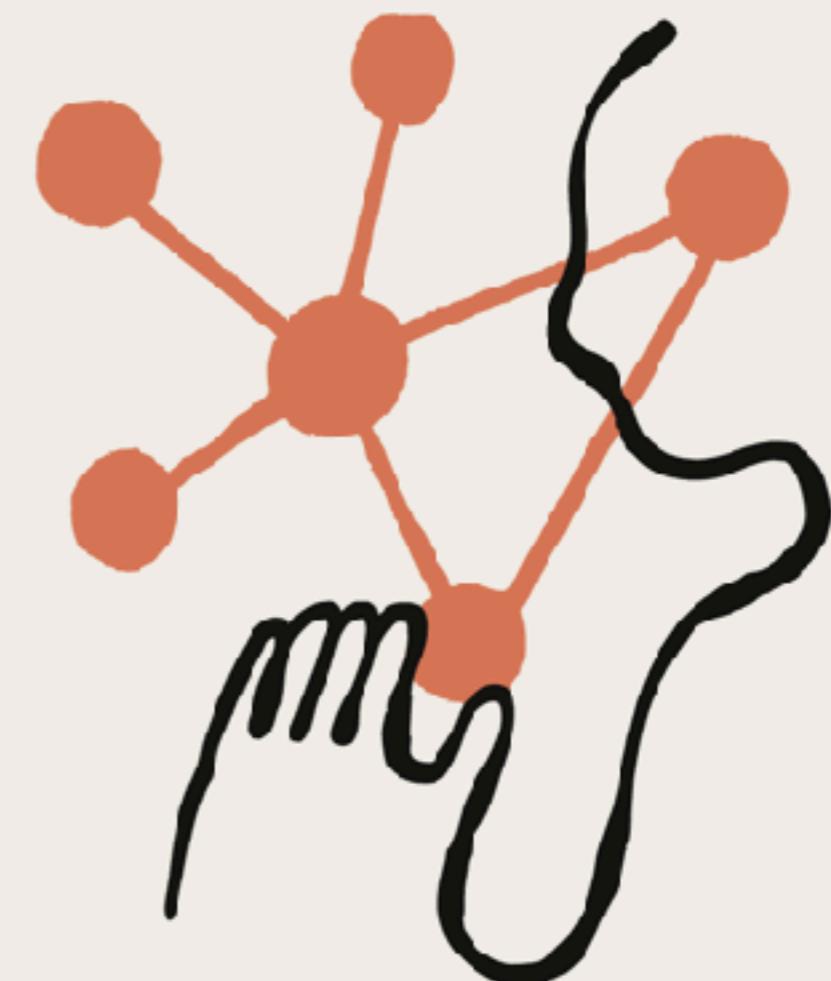
Talk to Claude

API

Build with Claude

Create AI-powered applications and custom experiences using Claude.

Learn more



<https://www.anthropic.com/>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

65

Models ?

Sonnet

Opus

Haiku

Balance speed and
Intelligent
Analysis

Reasoning and creative
Complex and long task

Fast
Simple tasks

<https://docs.anthropic.com/en/docs/about-claude/models/overview>



Model names

Model names

Model	Anthropic API	AWS Bedrock	GCP Vertex AI
Claude Opus 4.1	claude-opus-4-1-20250805	anthropic.claude-opus-4-1- 20250805-v1:0	claude-opus-4- 1@20250805
Claude Opus 4	claude-opus-4-20250514	anthropic.claude-opus-4- 20250514-v1:0	claude-opus- 4@20250514
Claude Sonnet 4	claude-sonnet-4-20250514	anthropic.claude-sonnet-4- 20250514-v1:0	claude-sonnet- 4@20250514
Claude Sonnet 3.7	claude-3-7-sonnet-20250219 (claude-3-7-sonnet-latest)	anthropic.claude-3-7-sonnet- 20250219-v1:0	claude-3-7- sonnet@20250219
Claude Haiku 3.5	claude-3-5-haiku-20241022 (claude-3-5-haiku-latest)	anthropic.claude-3-5-haiku- 20241022-v1:0	claude-3-5- haiku@20241022
Claude Haiku 3	claude-3-haiku-20240307	anthropic.claude-3-haiku- 20240307-v1:0	claude-3- haiku@20240307

<https://docs.anthropic.com/en/docs/about-claude/models/overview>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

Comparison models ?

Model name	Capabilities	Limitation	Use cases
Claude Haiku	Fast, low latency short task	Reasoning depth Complex task Creative	Real-time chatbot Quick customer support
Claude Sonnet	Balance between speed and intelligent	Slow than Haiku Less powerful than Opus	Business report Code assistance Research and writing task
Claude Opus	Most powerful Reasoning and creative Complex problem solving	Slow, more response time Expensive	Deep research Legal, medical analysis Complex sw development

<https://docs.anthropic.com/en/docs/about-claude/models/overview#model-comparison-table>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

Comparison models ?

	Claude Opus 4.1	Claude Opus 4	Claude Sonnet 4	OpenAI o3	Gemini 2.5 Pro
Agentic coding <i>SWE-bench Verified</i> ¹	74.5%	72.5%	72.7%	69.1%	67.2%
Agentic terminal coding <i>Terminal-Bench</i> ²	43.3%	39.2%	35.5%	30.2%	25.3%
Graduate-level reasoning <i>GPQA Diamond</i>	80.9%	79.6%	75.4%	83.3%	86.4%
Agentic tool use <i>TAU-bench</i>	Retail 82.4%	Retail 81.4%	Retail 80.5%	Retail 70.4%	—
	Airline 56.0%	Airline 59.6%	Airline 60.0%	Airline 52.0%	—
Multilingual Q&A <i>MMMLU</i> ³	89.5%	88.8%	86.5%	88.8%	—
Visual reasoning <i>MMMU (validation)</i>	77.1%	76.5%	74.4%	82.9%	82%
High school math competition <i>AMC 2025</i> ⁴	78.0%	75.5%	70.5%	88.9%	88%

<https://www.anthropic.com/clause/opus>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

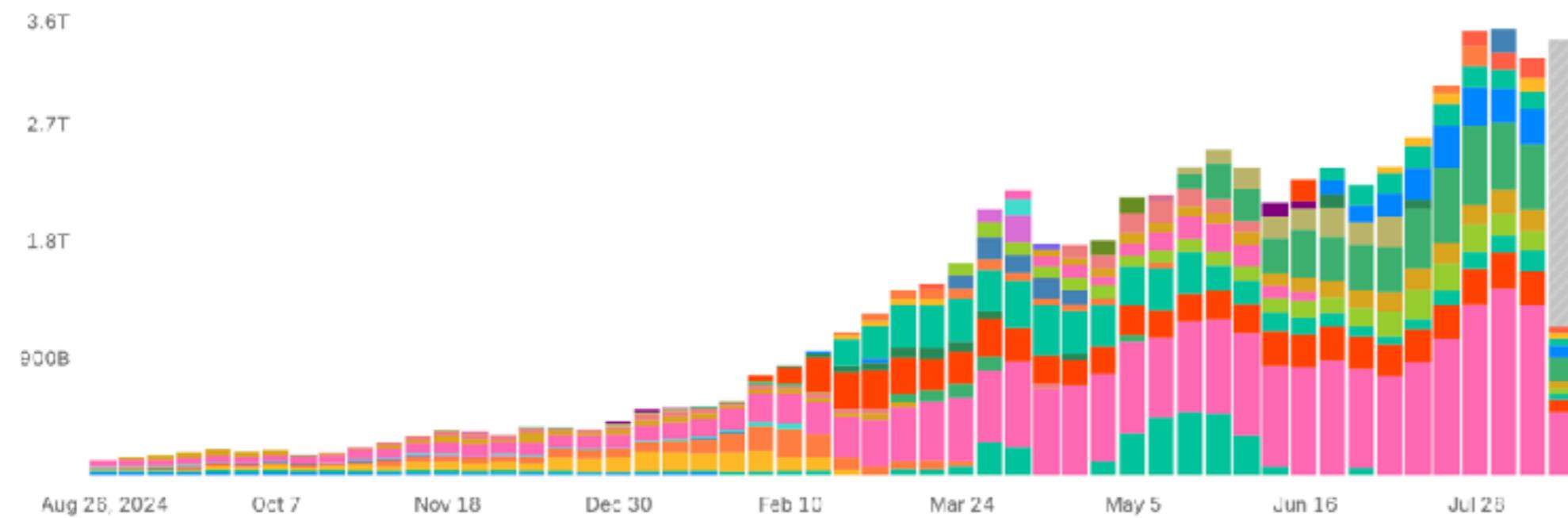
69

Token usage from OpenRouter

Leaderboard

Top this week 3

Token usage across models on OpenRouter ⓘ



1.	Claude Sonnet 4 by anthropic	515B tokens ↑2%	6.	DeepSeek V3 0324 (free) by deepseek	141B tokens ↓14%
2.	Gemini 2.0 Flash by google	274B tokens ↓0%	7.	Gemini 2.5 Pro by google	135B tokens ↓9%
3.	Gemini 2.5 Flash by google	267B tokens ↓0%	8.	Claude 3.7 Sonnet by anthropic	129B tokens ↓8%
4.	DeepSeek V3 0324 by deepseek	159B tokens ↓11%	9.	R1 0528 (free) by deepseek	109B tokens ↑5%

<https://openrouter.ai/rankings>



AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

70

SWE-bench



SWE-bench

Leaderboards

BENCHMARKS

SWE-bench

SWE-bench Lite

SWE-bench Multilingual

SWE-bench Multimodal

SWE-bench Bash Only

SWE-bench Verified

ABOUT

Paper

Docs

Contact

Citations

Press



Leaderboards

There's an all-new, challenging SWE-bench Multimodal, containing software issues described with images. [Learn more here.](#)

Bash Only Verified Lite Full Multimodal

Bash Only evaluates all LMs with a [minimal agent](#) on SWE-bench Verified ([details](#))

Filters: [Open Scaffold](#) [All Tags](#)

Model	% Resolved	Org	Date	Logs	Trajs	Site	Release
Claude 4 Opus (20250514)	67.60		2025-08-02	✓	✓		1.0.0
GPT-5 (2025-08-07) (medium reasoning)	65.00		2025-08-07	✓	✓		1.7.0
Claude 4 Sonnet (20250514)	64.93		2025-05-21	✓	✓		1.0.0
GPT-5 mini (2025-08-07) (medium reasoning)	59.80		2025-08-07	✓	✓		1.7.0
o3 (2025-04-16)	58.40		2025-05-21	✓	✓		1.0.0
Qwen3-Coder 480B/A35B Instruct	55.40		2025-08-02	✓	✓		1.0.0
Gemini 2.5 Pro (2025-05-06)	53.60		2025-05-21	✓	✓		1.0.0
Claude 3.7 Sonnet (20250219)	52.80		2025-05-21	✓	✓		0.0.0
o4-mini (2025-04-16)	45.00		2025-05-21	✓	✓		1.0.0

<https://www.swebench.com/>

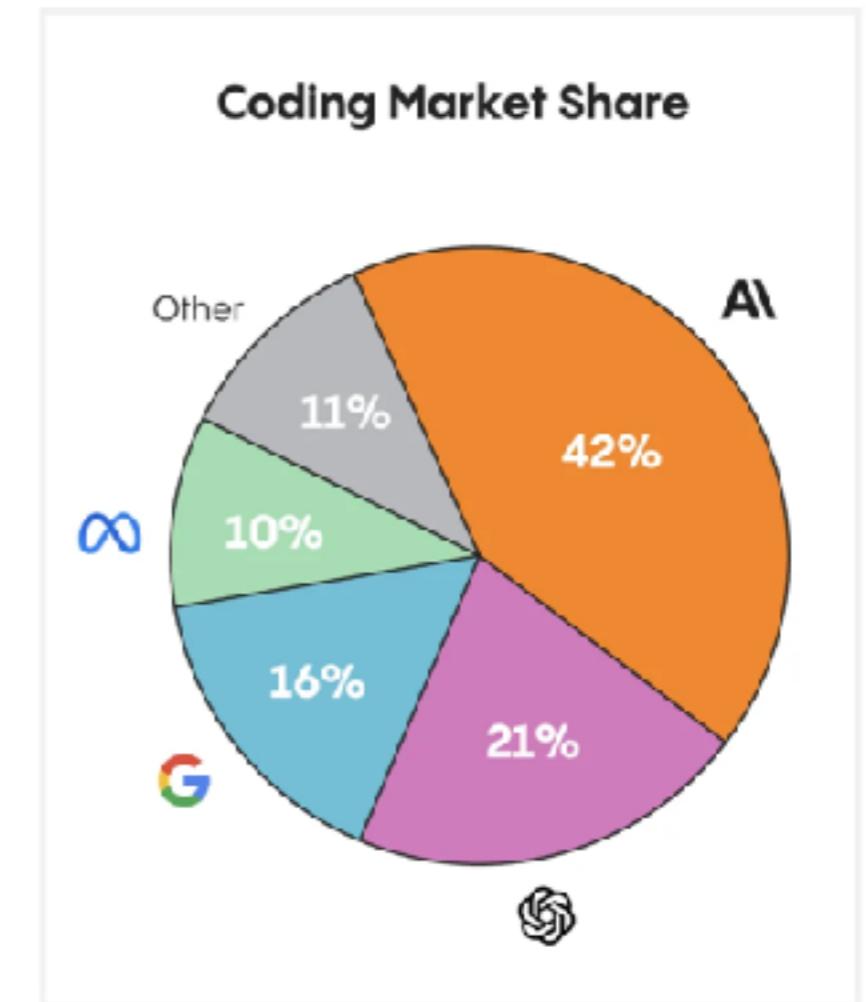
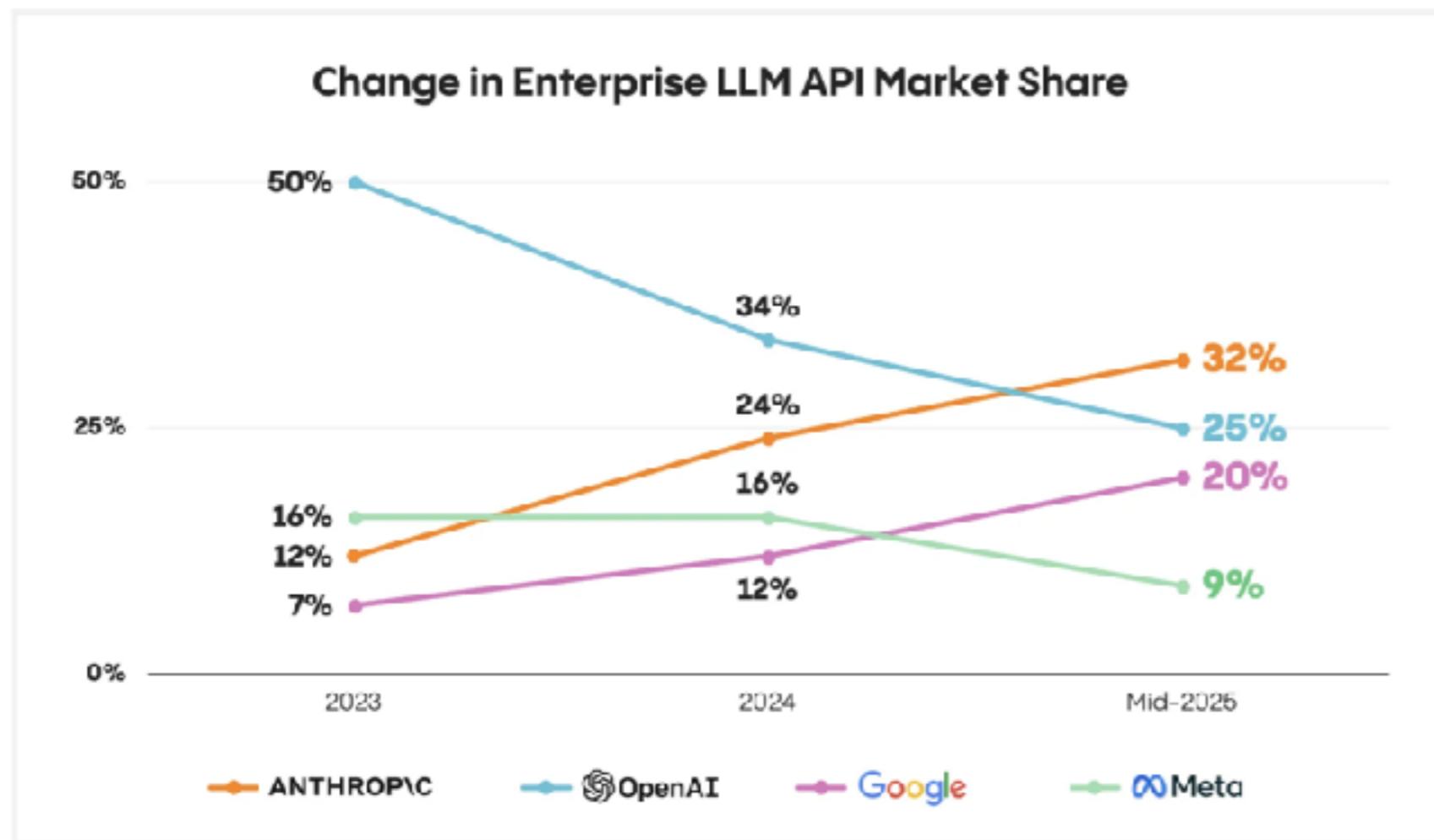
AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

71

Market Share !!

Enterprise LLM API Market Share by Usage



© 2025 Menlo Ventures

<https://menlovc.com/perspective/2025-mid-year-llm-market-update/>



Anthropic Products ?

Chat-based

API-based

Claude Code

Web
Mobile app

Integrate with external tools

CLI tools
Code Agent

MCP
Model Context Protocol

<https://www.anthropic.com/>



ສືເໜືອງ



มีด



Prompt Engineering

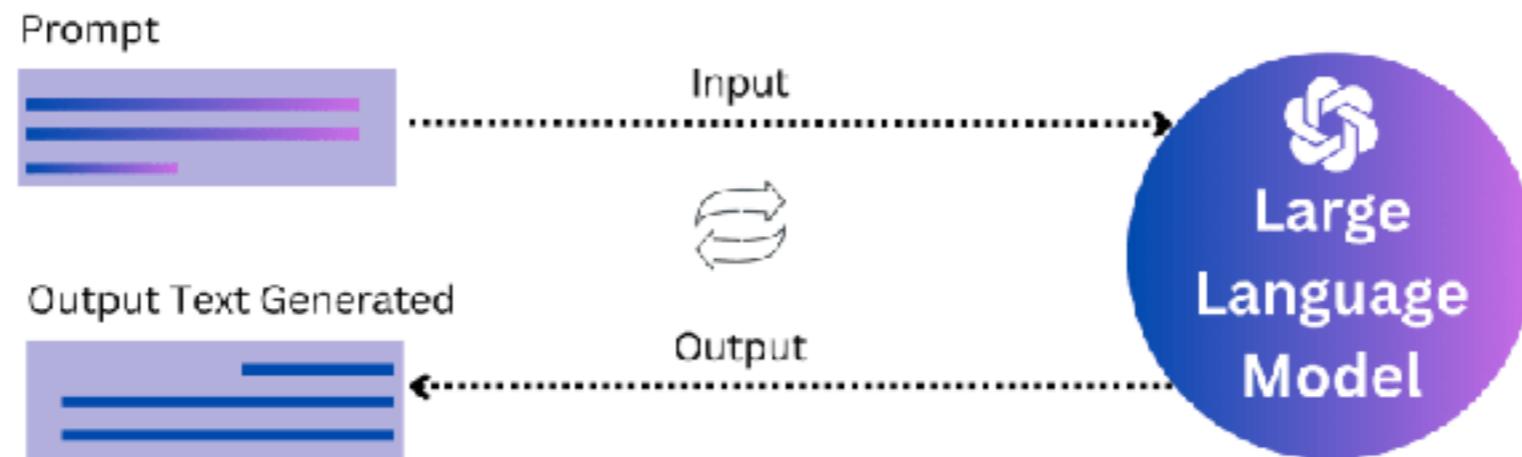
<https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/overview>



Prompt composition

Prompt engineering

Compose prompts from user inputs and context



<https://platform.openai.com/docs/guides/prompt-engineering>



Better Prompt

Write clear instructions

Provide reference text

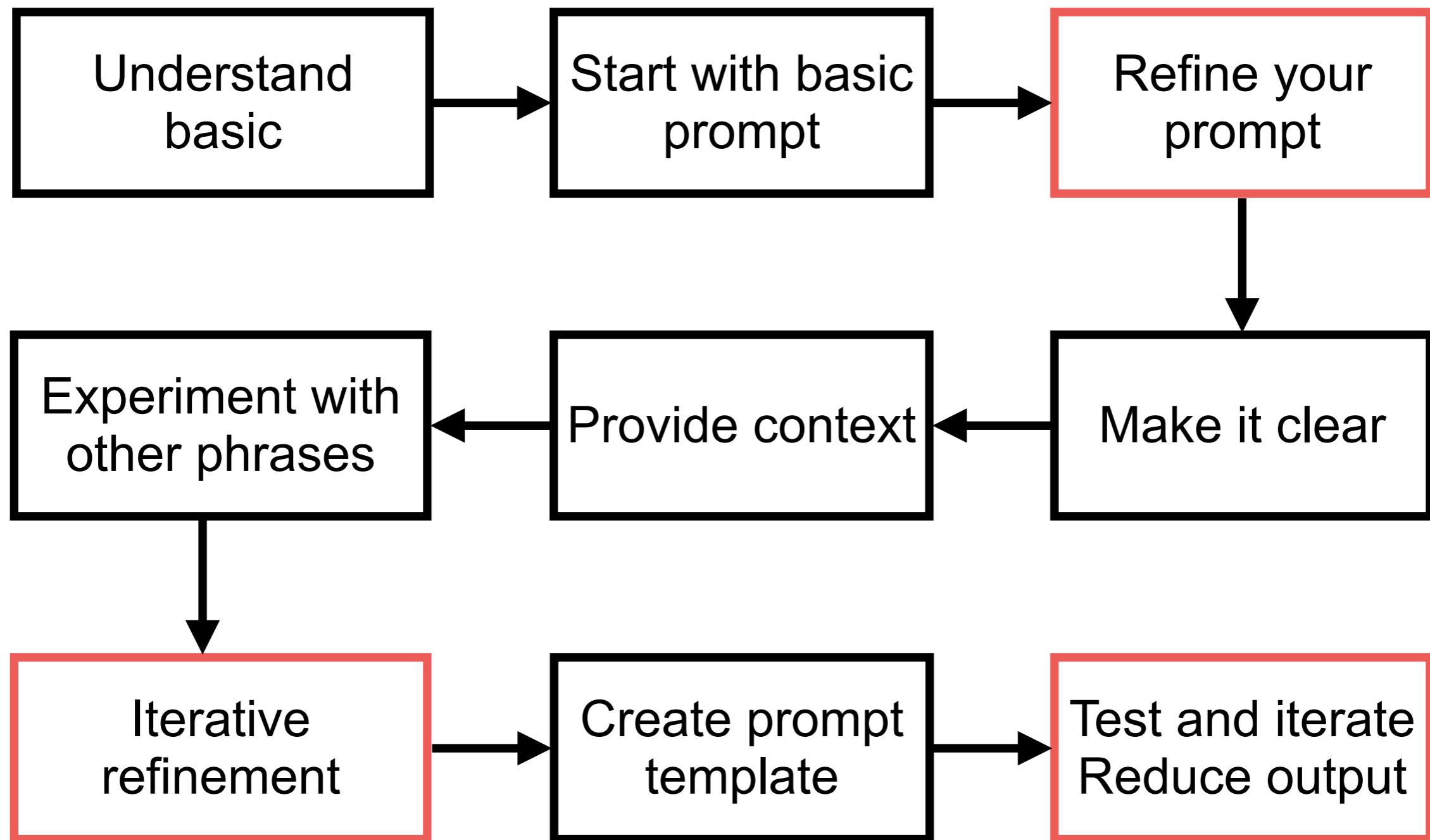
Split complex tasks into simpler subtasks

Give the model time to think

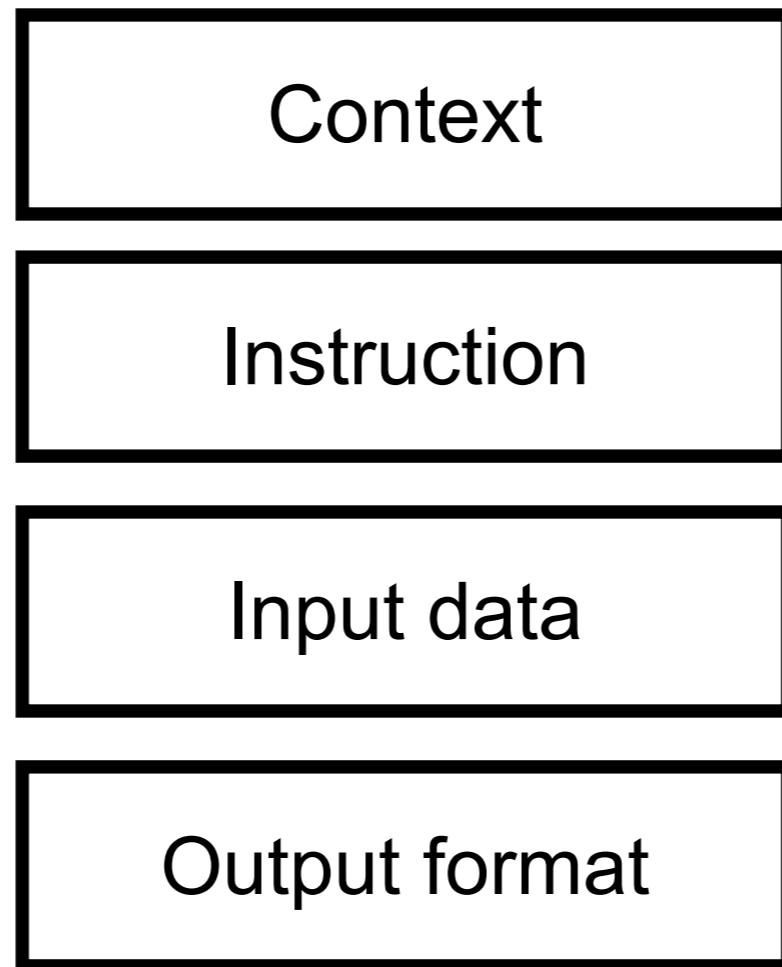
Testing and improve ...



Basic of Prompt Engineer



Structure of Prompt



Prompting Guide

Prompt Engineering

Prompt Engineering Guide

Prompt engineering is a relatively new discipline for developing and optimizing prompts to efficiently use language models (LMs) for a wide variety of applications and research topics. Prompt engineering skills help to better understand the capabilities and limitations of large language models (LLMs).

Researchers use prompt engineering to improve the capacity of LLMs on a wide range of common and complex tasks such as question answering and arithmetic reasoning. Developers use prompt engineering to design robust and effective prompting techniques that interface with LLMs and other tools.

Prompt engineering is not just about designing and developing prompts. It encompasses a wide range of skills and techniques that are useful for interacting and developing with LLMs. It's an important skill to interface, build with, and understand capabilities of LLMs. You can use prompt engineering to improve safety of LLMs and build new capabilities like augmenting LLMs with domain knowledge and external tools.

<https://www.promptingguide.ai/>



Prompt Techniques

Zero-shot

Chain-of
Thought (CoT)

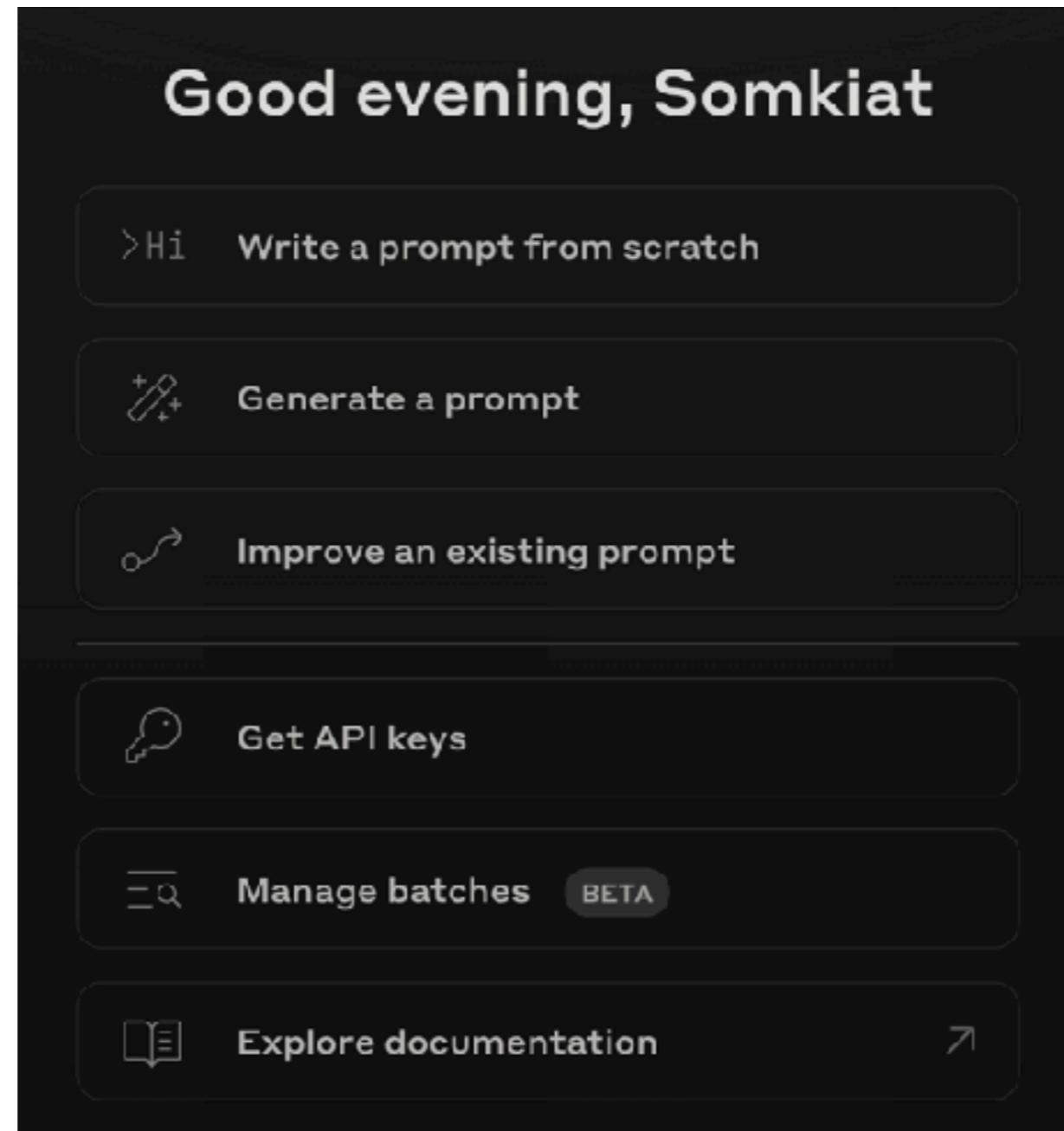
Few-shot

Meta or structure

<https://www.promptingguide.ai/techniques>



Anthropic Dashboard



<https://console.anthropic.com/dashboard>



AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

Chain of Thought Prompting (CoT)

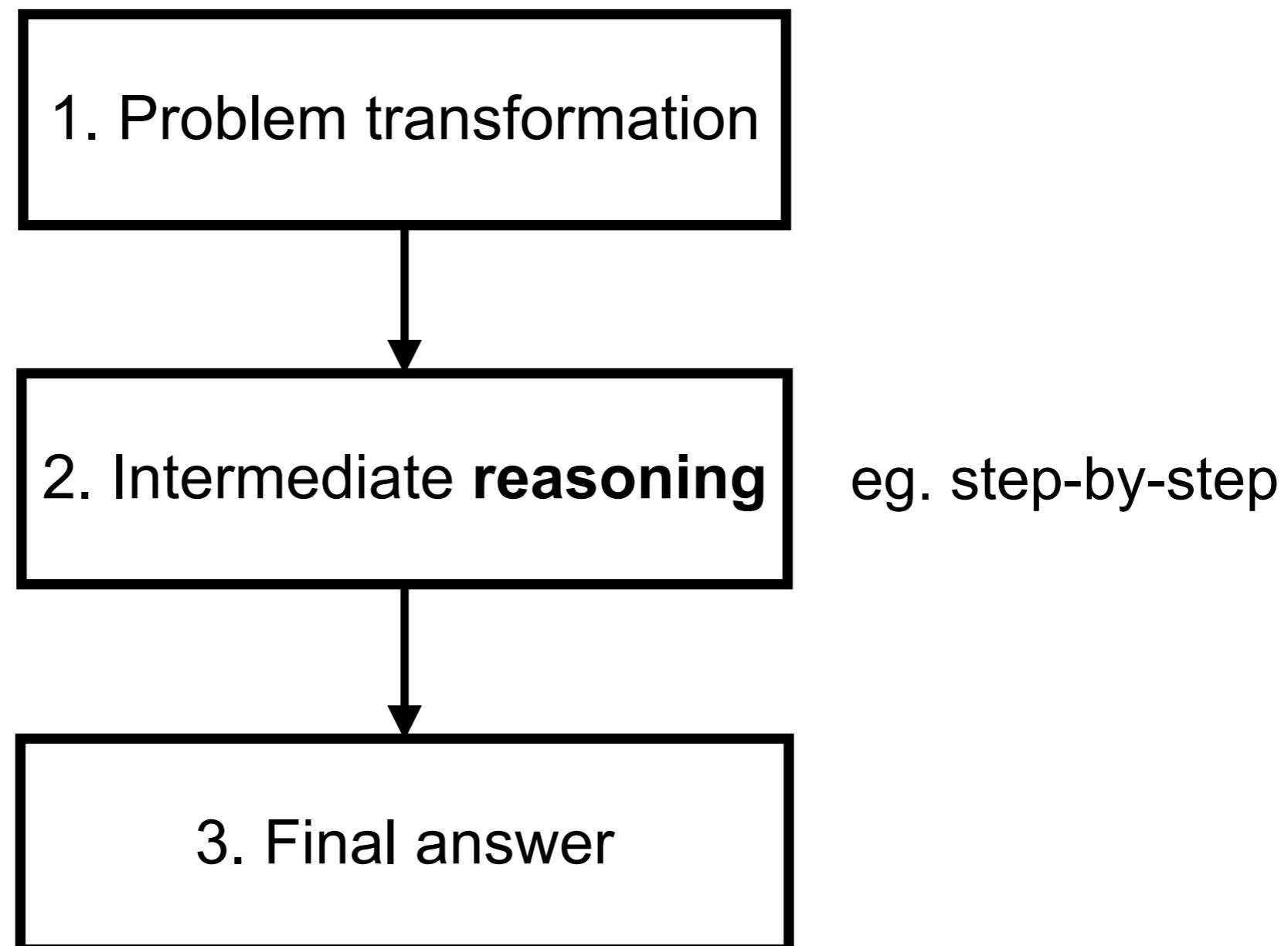
Technique used to improve the reasoning ability of LLM

Try to break down a complex problem into smaller, More manageable steps, lead to final answer

Reasoning model !!



Chain of Thought Prompting (CoT)



Chain of Thought Prompting (CoT)

(a) Few-shot

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The answer is 8. ✗

(b) Few-shot-CoT

Q: Roger has 5 tennis balls. He buys 2 more cans of tennis balls. Each can has 3 tennis balls. How many tennis balls does he have now?

A: Roger started with 5 balls. 2 cans of 3 tennis balls each is 6 tennis balls. $5 + 6 = 11$. The answer is 11.

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A:

(Output) The juggler can juggle 16 balls. Half of the balls are golf balls. So there are $16 / 2 = 8$ golf balls. Half of the golf balls are blue. So there are $8 / 2 = 4$ blue golf balls. The answer is 4. ✓

(c) Zero-shot

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 ✗

(d) Zero-shot-CoT (Ours)

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: Let's think step by step.

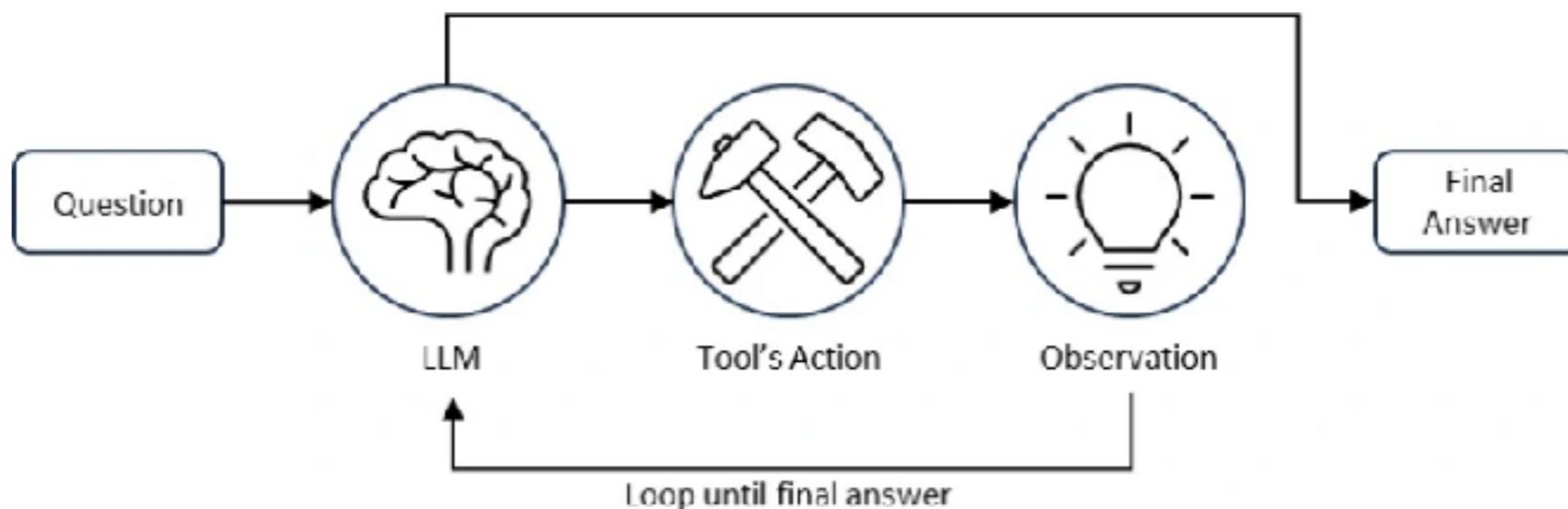
(Output) There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓

<https://www.promptingguide.ai/techniques/cot>



ReAct Prompt

LLM reasoning and additional tools (expert)
Improve better answer



<https://www.promptingguide.ai/techniques/react>



Hallucinations in LLM !!

Generated content that irrelevant or inconsistent

Incorrect
information

Trust

Training data quality

<https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models>



Consequences of LLM Hallucination

Privacy Issues



6x increase in toxicity:
ChatGPT persona assignment perpetuates harmful stereotypes

Misinformation and Disinformation



74% of IT decision-makers
concerned about cybersecurity risks with GPT-4

Discriminating and Toxic Content



49% of individuals foresee
GPT-4 as a tool for spreading misinformation

How to mitigate LLM hallucinations?

1

Pre-processing and Input Control

Limiting response length

Controlled input

2

Model Configuration & Behavior

Adjusting model parameters

Using a moderation layer

3

Learning and Improvement

Feedback, Monitoring & Improvement Mechanisms

Domain Adaptation and Augmentation

4

Context and Data Enhancement

Incorporating an external database

Contextual prompt engineering



Structured Prompt



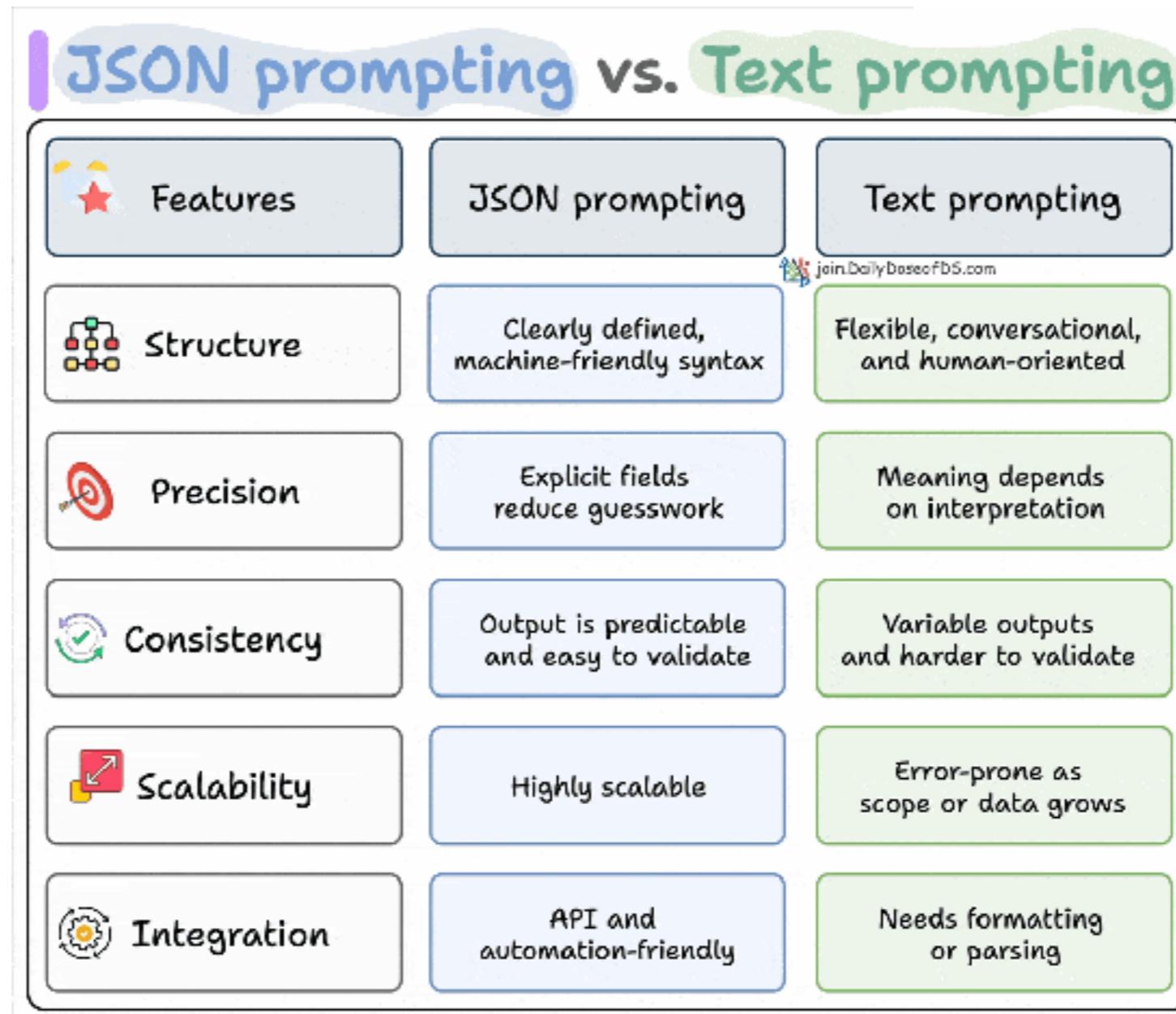
Structured Prompt

Text, Markdown, XML, JSON



Text vs JSON Prompt !!

Structured prompt, reduce hallucinations

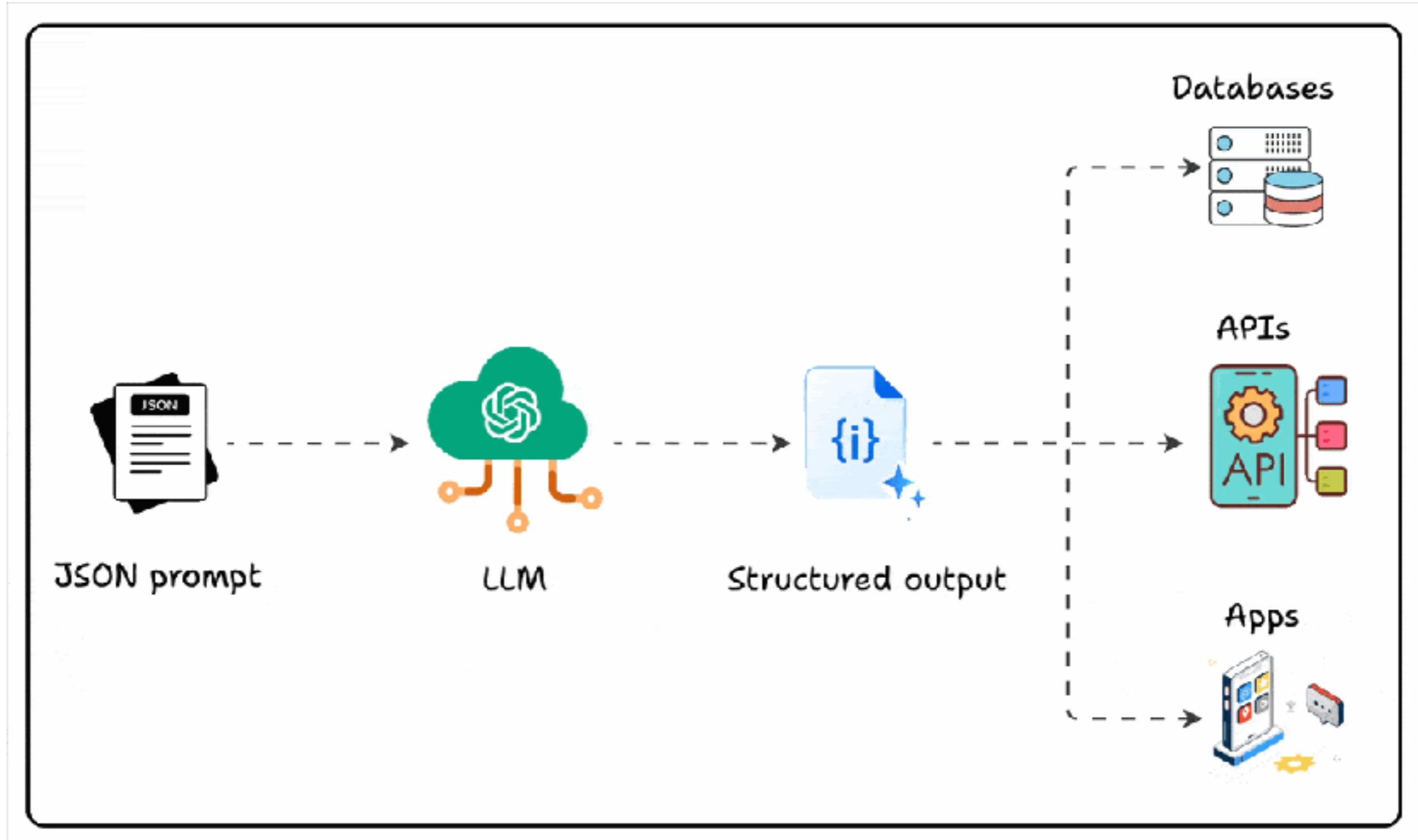


JSON Prompt

```
{  
  "task": "Summarize",  
  "format": "bullet points",  
  "tone": "professional",  
  "length": "3 key takeaways"  
}
```



JSON Prompt



Many models excel at other formats !!

Claude handles **XML**

Markdown provide structure without overhead



JSON and Markdown

The image shows two side-by-side terminal windows comparing JSON and Markdown prompts for an AI model.

JSON Prompt

```
{  
  "task": "Provide details for each movie",  
  "movies": ["Inception", "The Matrix", "Interstellar"],  
  "output_format": {  
    "title": "",  
    "director": "",  
    "year": "",  
    "imdb_rating": ""  
  }  
}
```

Braces, commas,
colons etc. are
a token overhead

Tokens
59
Characters
205

Markdown Prompt (Structured)

```
# Task  
Provide details for each movie  
  
## Movies  
- Inception  
- The Matrix  
- Interstellar  
  
## Output Format  
- Title:  
- Director:  
- Year:  
- IMDb Rating:
```

Less tokens.
Saves money 💰

Tokens
41
Characters
151



Use XML in Claude AI

Clarify

Accuracy

Flexibility

Parse-ability

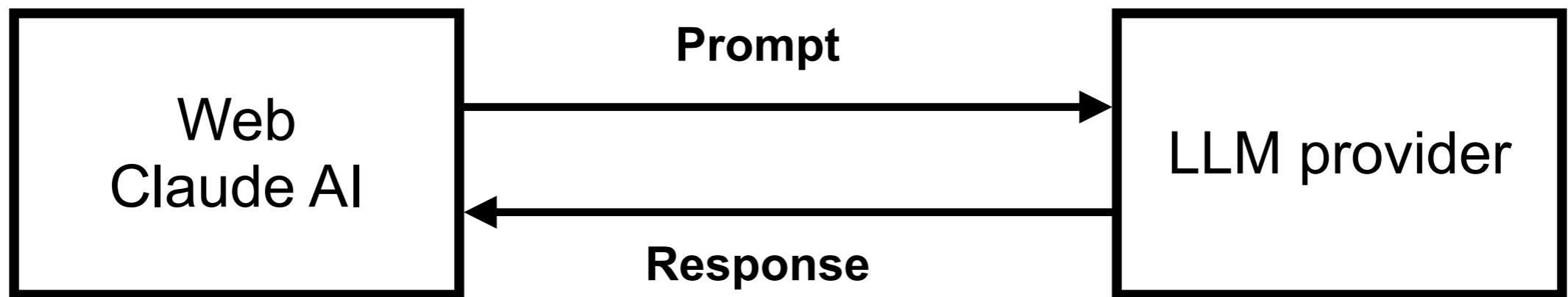
<https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/use-xml-tags>



Workshop



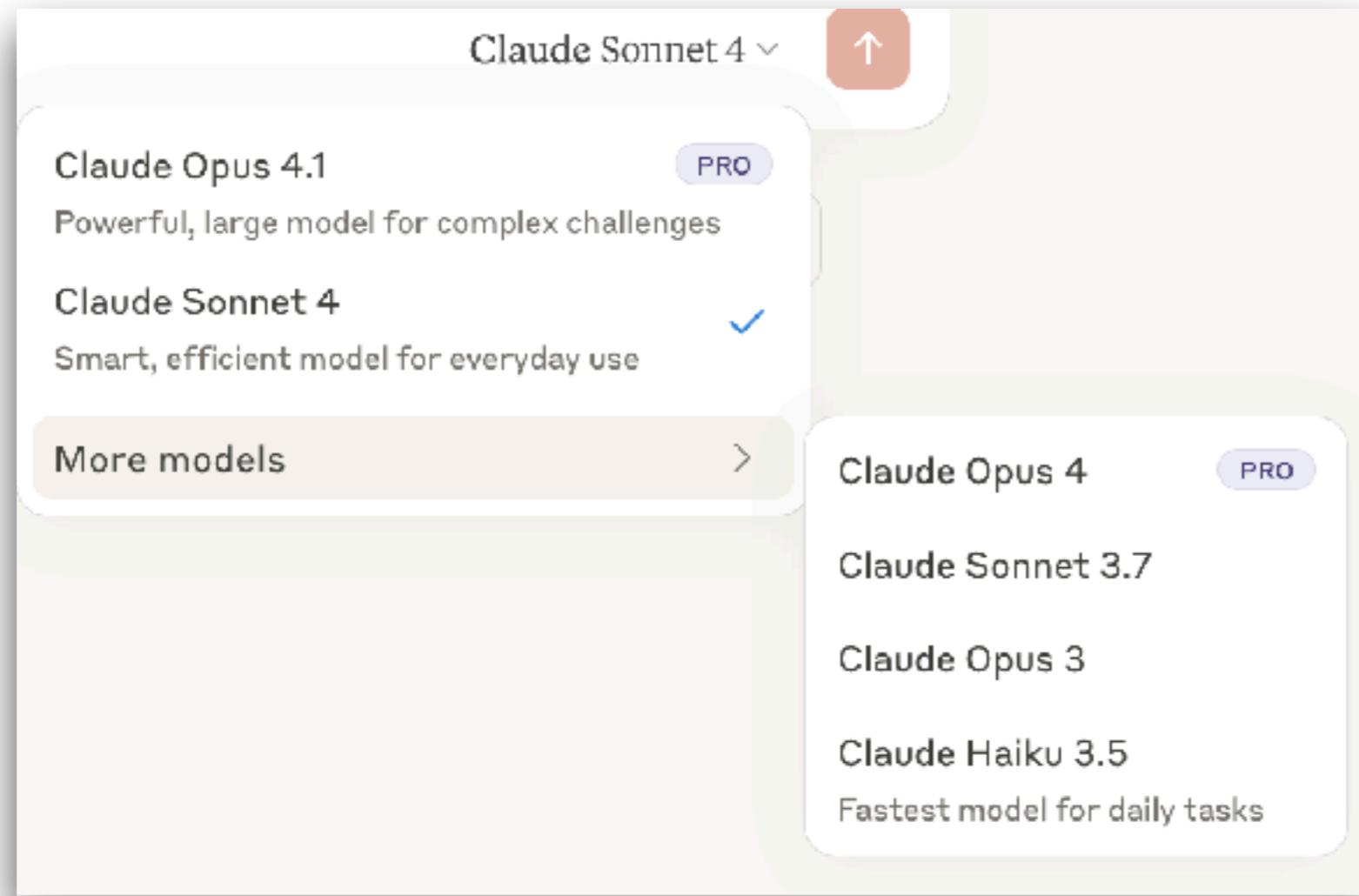
Chat with Claude AI



<https://claude.ai/new>



Choose model !!

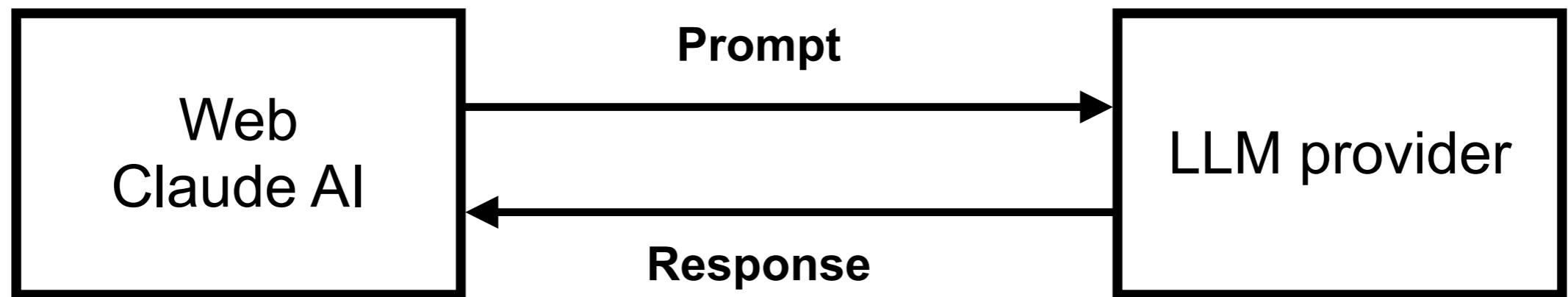


<https://docs.anthropic.com/en/docs/about-claude/models/overview>



Chat with Claude AI

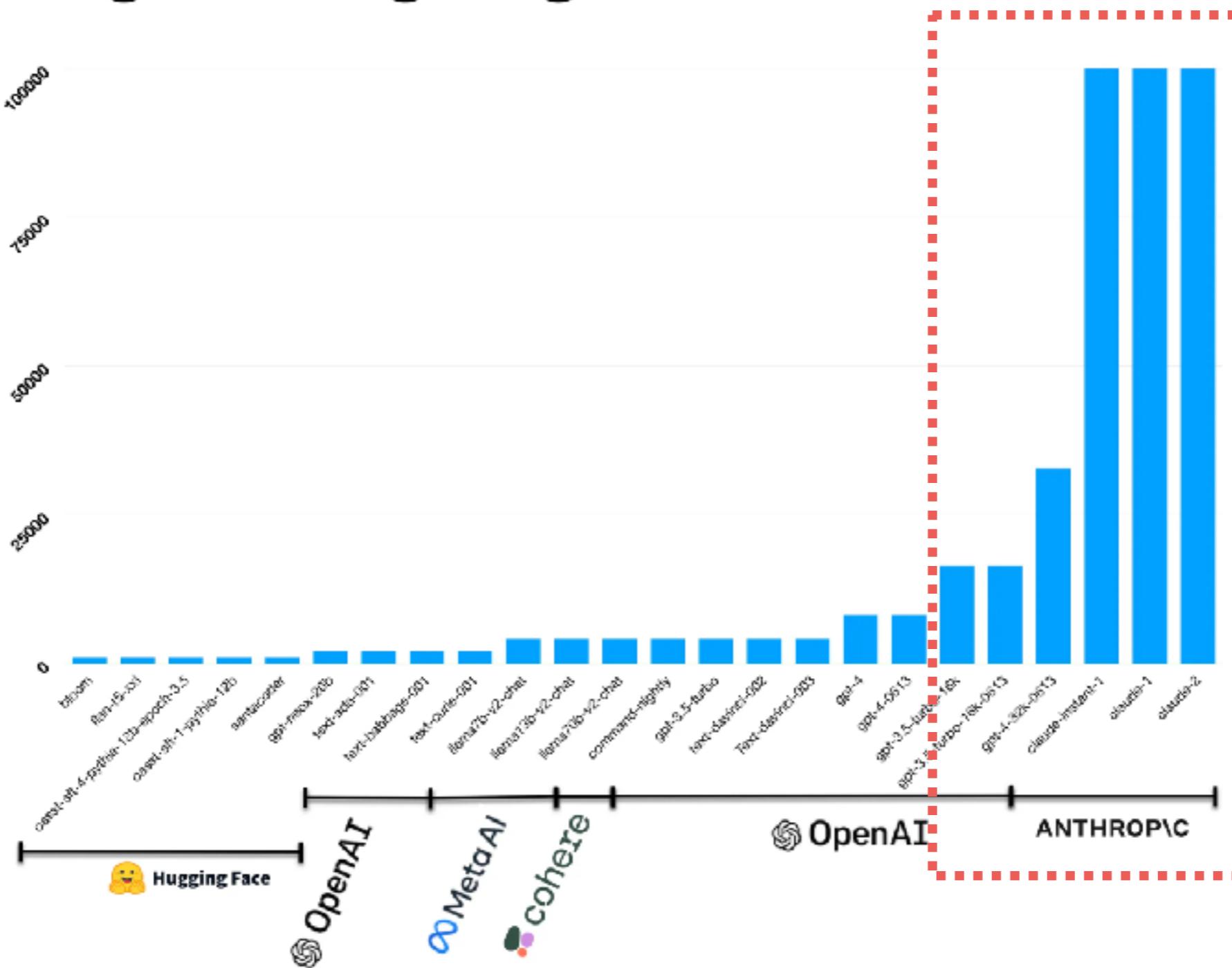
Context window
200,000 **tokens**



<https://claude.ai/new>



Large Language Model Context Size



www.cobusgreyling.com

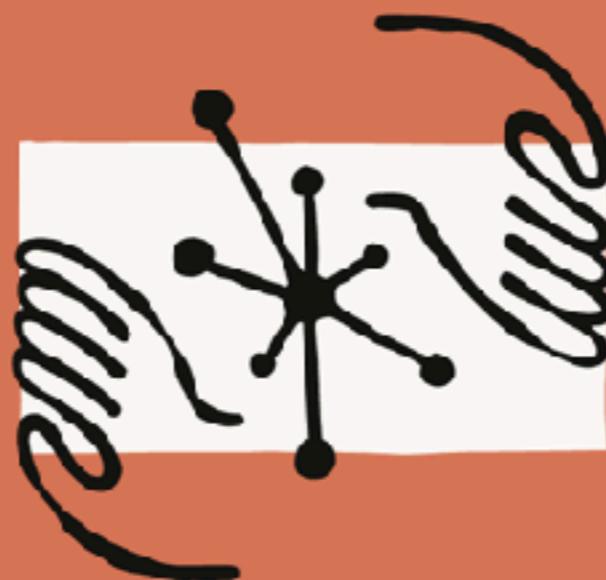
<https://www.humanfirst.ai/blog/how-does-large-language-models-use-long-contexts>



Claude Sonnet 4 => 1M

Claude Sonnet 4 now supports 1M tokens of context

Aug 12, 2025 • 2 min read



<https://www.anthropic.com/news/1m-context>



Llama 4 from Meta (10M)

Llama 4: Leading Multimodal Intelligence

Newest model suite offering unrivaled speed and efficiency

Llama 4 Behemoth

288B active parameter, 16 experts

2T total parameters

The most intelligent teacher model for distillation

Preview

Llama 4 Maverick

17B active parameters, 128 experts

400B total parameters

Native multimodal with 1M context length

Available

Llama 4 Scout

17B active parameters, 16 experts

109B total parameters

Industry leading 10M context length

Optimized inference

Available

<https://ai.meta.com/blog/llama-4-multimodal-intelligence/>



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

105

Long context, more use cases

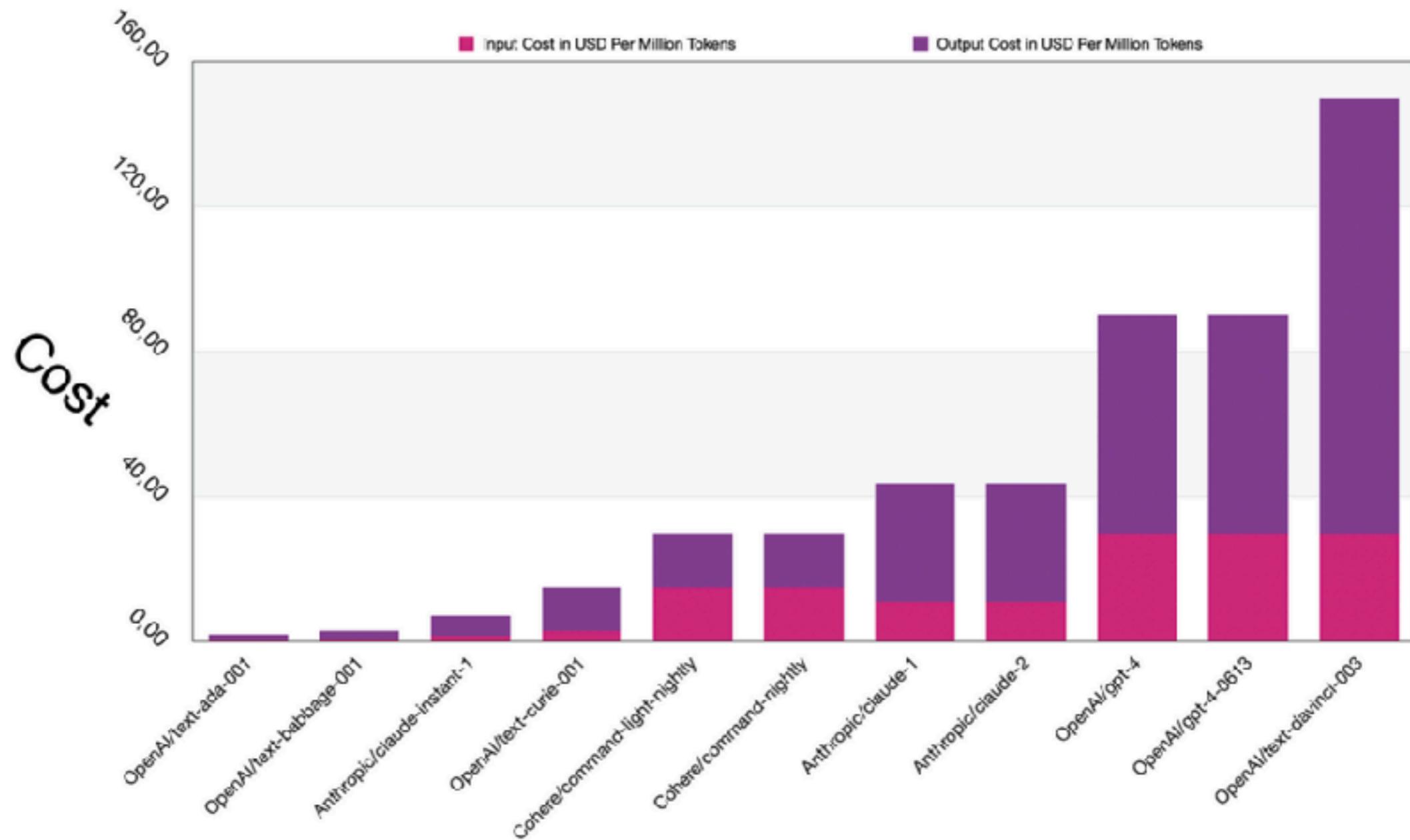
Large scale code analysis
Document synthesis
Context-aware agents

Pricing !!

	Input	Output
Prompts ≤ 200K	\$3 / MTok	\$15 / MTok
Prompts > 200K	\$6 / MTok	\$22.50 / MTok



LLM Models Cost !!



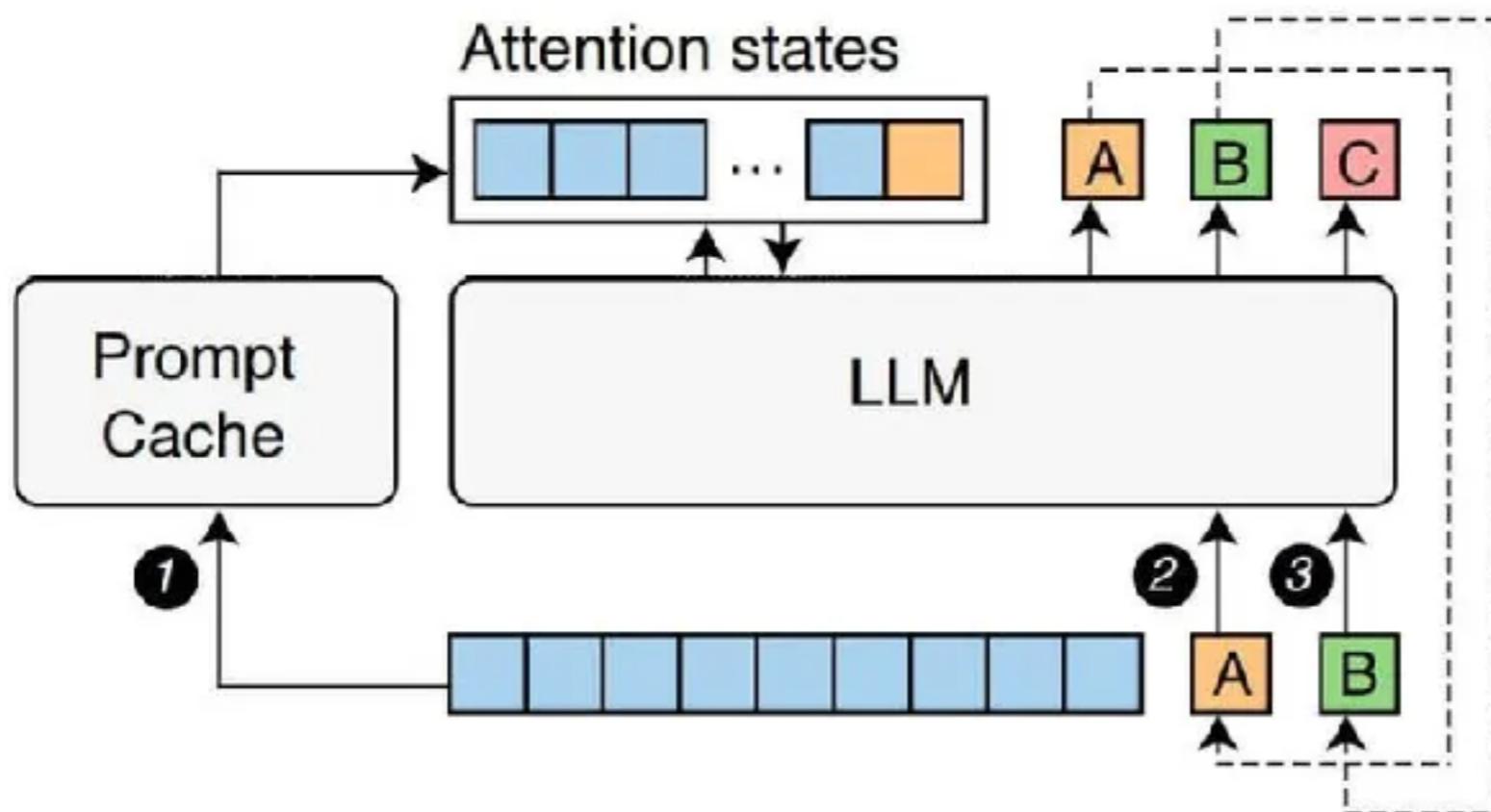
<https://www.humanfirst.ai/blog/how-does-large-language-models-use-long-contexts>



Context Caching



Prompt caching



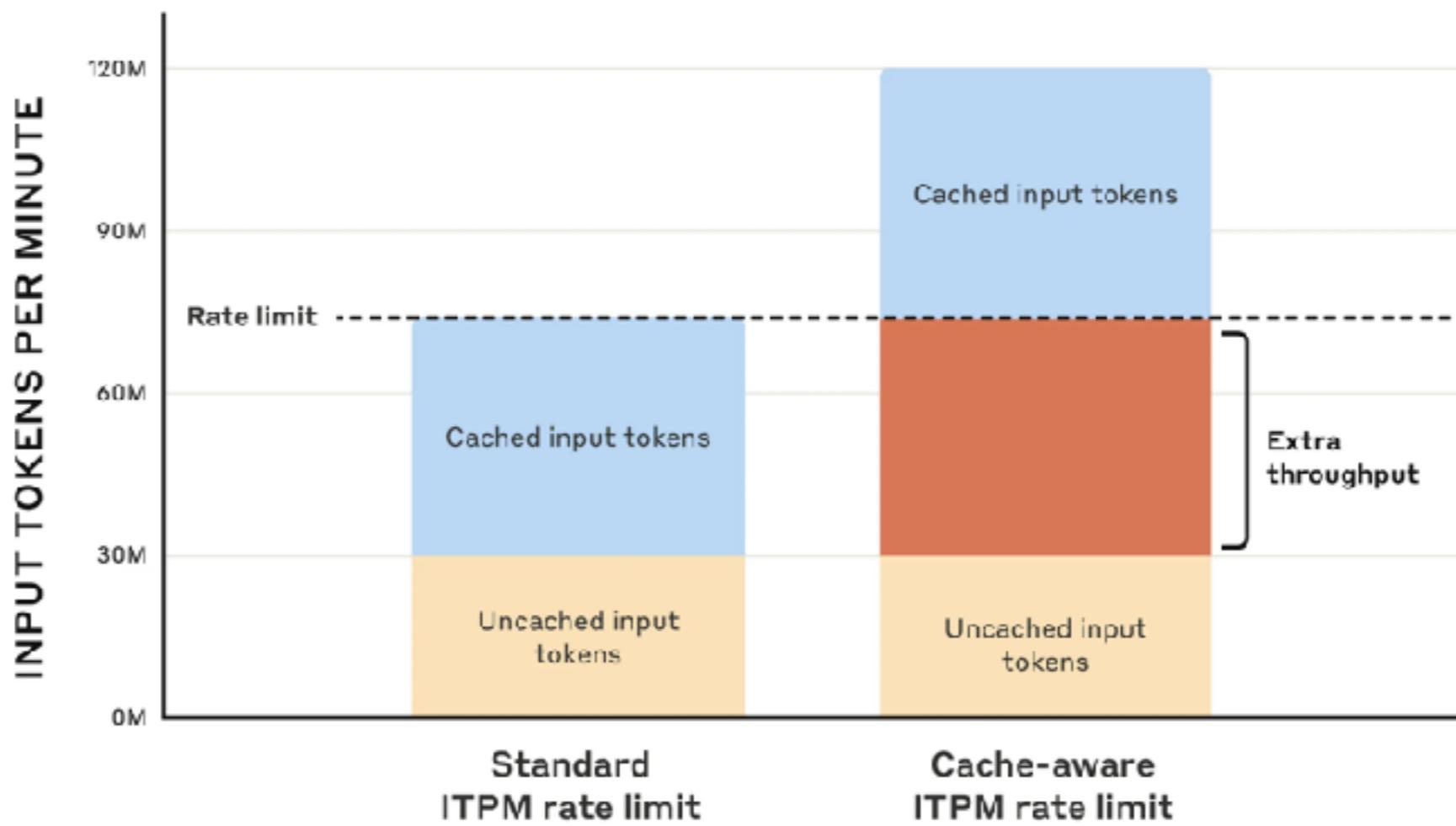
(c) Generation with Prompt Cache

https://github.com/anthropics/anthropic-cookbook/blob/main/misc/prompt_caching.ipynb



Context caching with Anthropic API

Optimize throughput with prompt caching



<https://www.anthropic.com/news/token-saving-updates>



Prompt caching

Before (manual cache control on each element)

System Prompt:

```
... long system prompt cache_control: {type: "ephemeral"}
```

User Message:

```
Hello, can you tell me more about the solar system?  
cache_control: {type: "ephemeral"}
```

Assistant Reply:

```
Certainly! The solar system is the collection of  
celestial bodies ...
```

New User Message:

```
Tell me more about Mars. cache_control: {type:  
"ephemeral"}
```

After (automatic use of longest cached prefix)

System Prompt:

```
... long system prompt
```

User Message:

```
Hello, can you tell me more about the solar system?
```

Assistant Reply:

```
Certainly! The solar system is the collection of  
celestial bodies ...
```

New User Message:

```
Tell me more about Mars. cache_control: {type:  
"ephemeral"}
```

<https://docs.anthropic.com/en/docs/build-with-claude/prompt-caching>



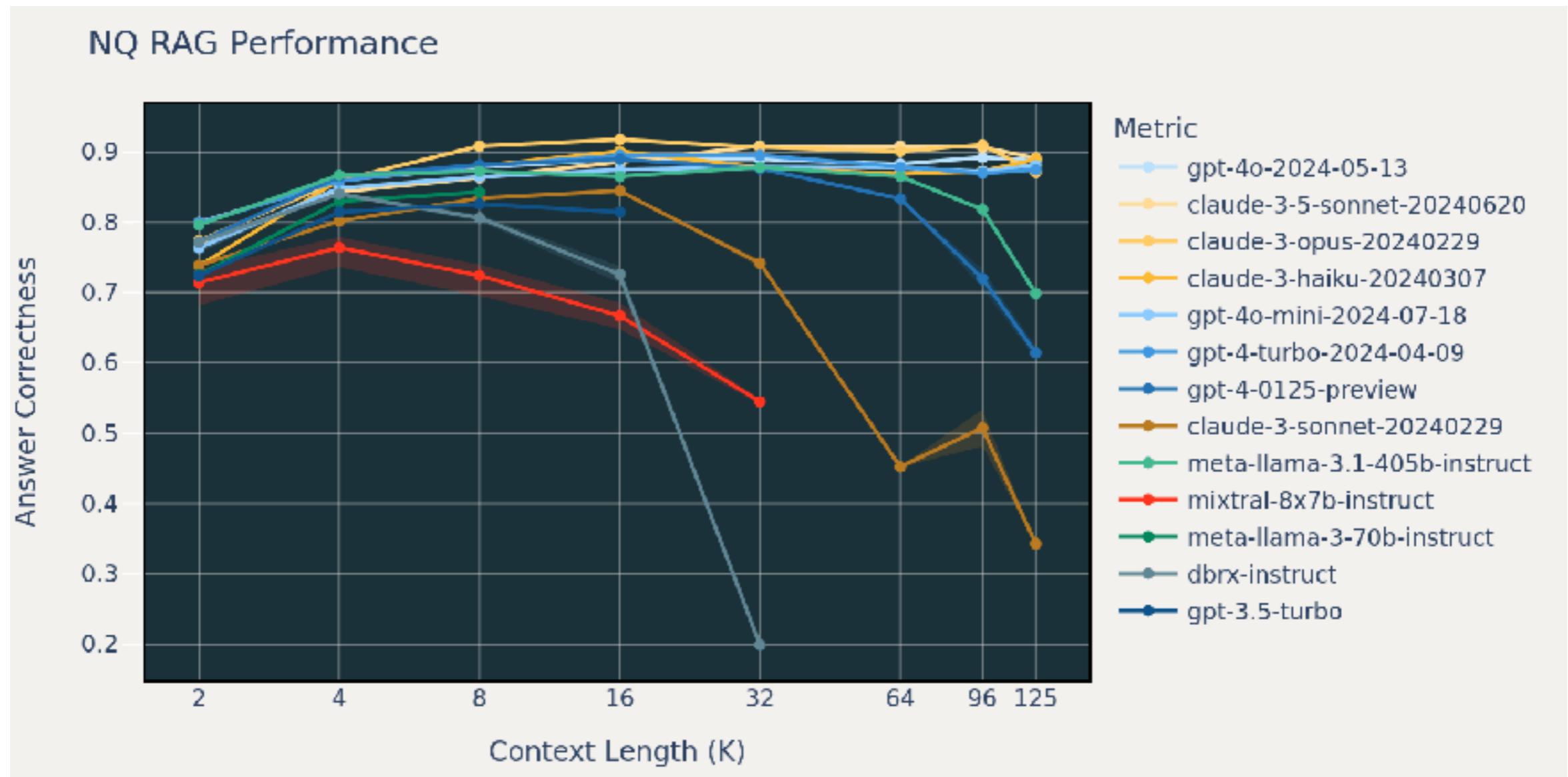
AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

Long context with correctness !!



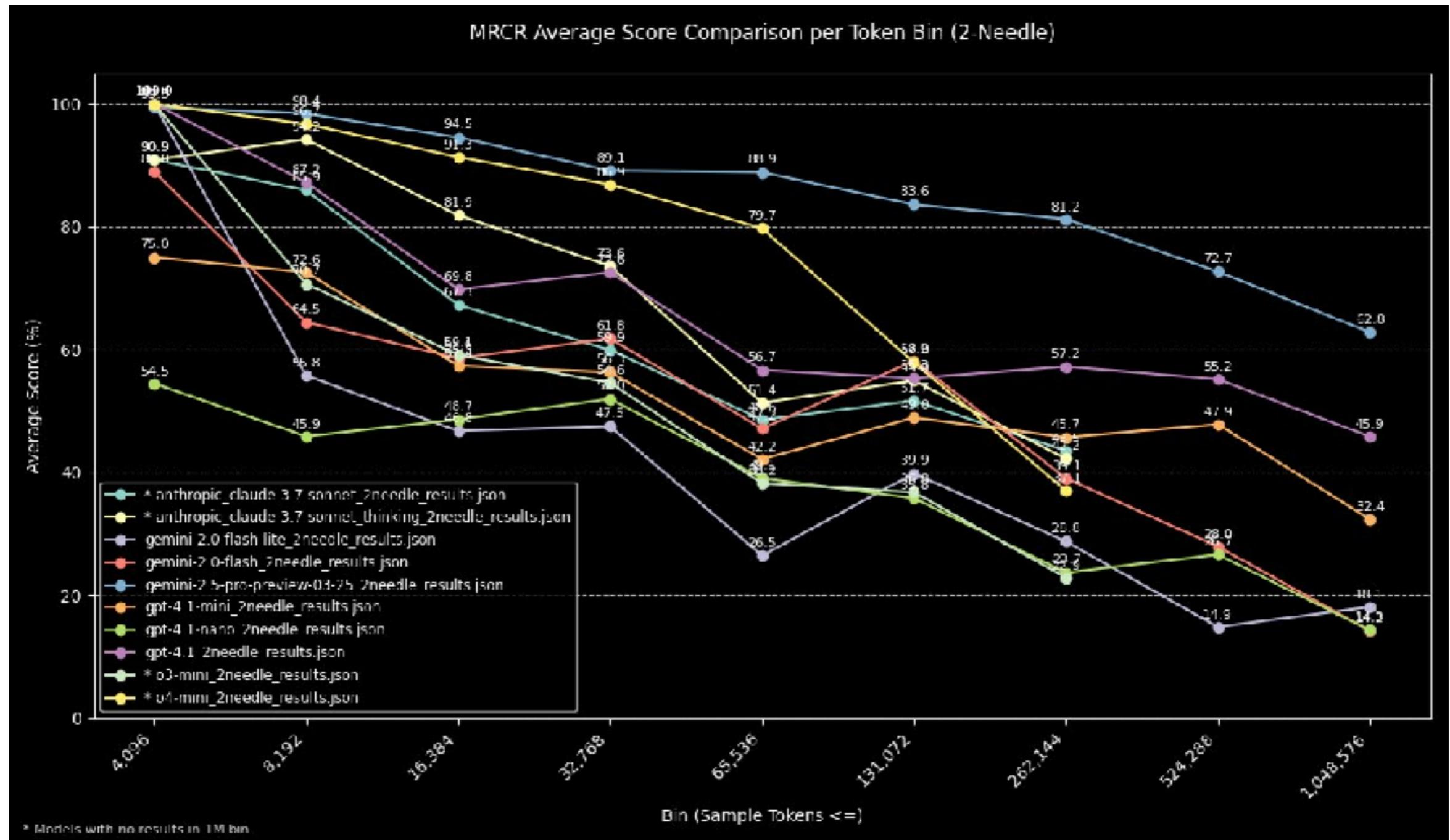
Long context with correctness !



<https://www.databricks.com/blog/long-context-rag-performance-lms>



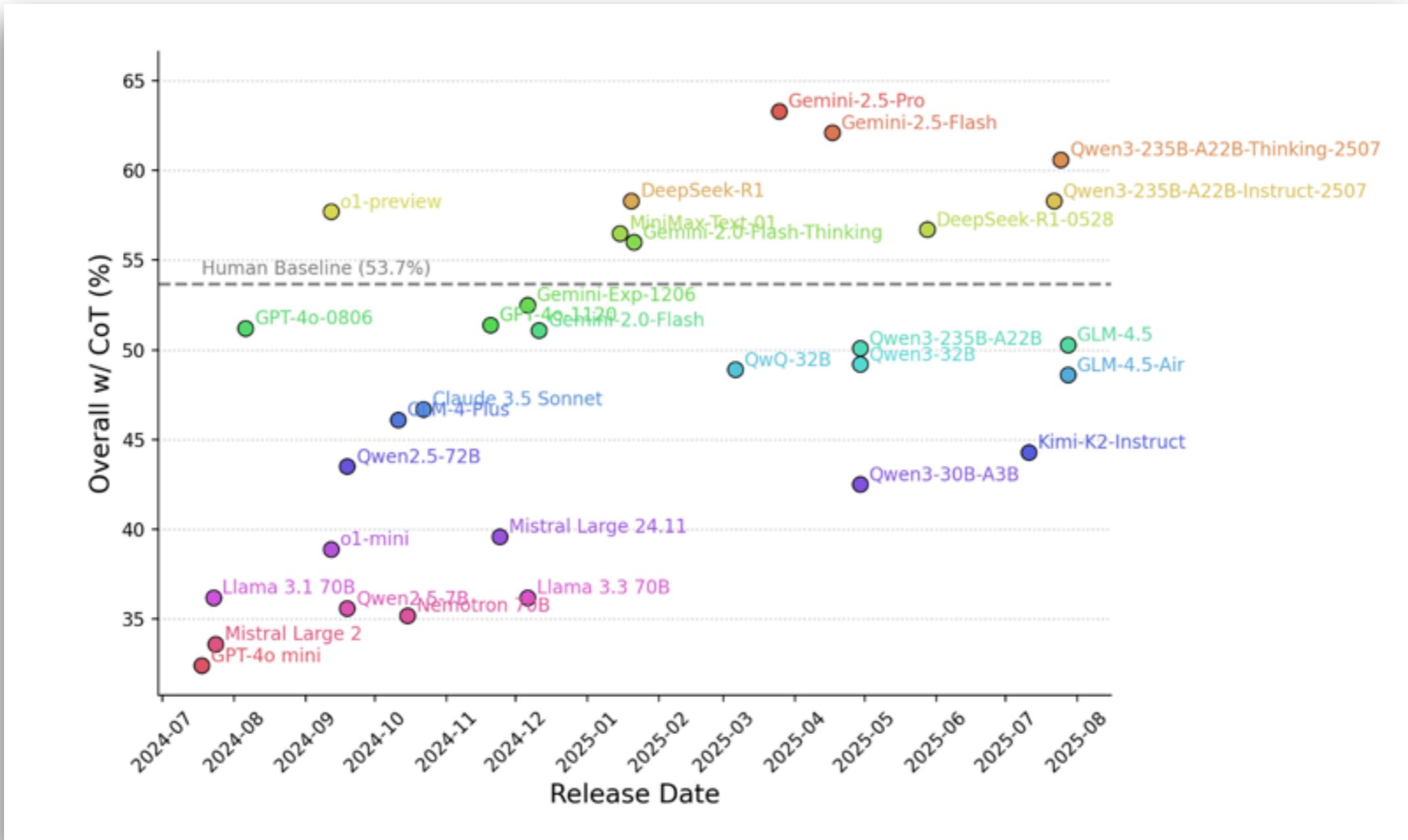
Long context with correctness !



<https://x.com/wintermoat/status/1913001771506561369/photo/1>



LongBench v2



<https://longbench2.github.io/>

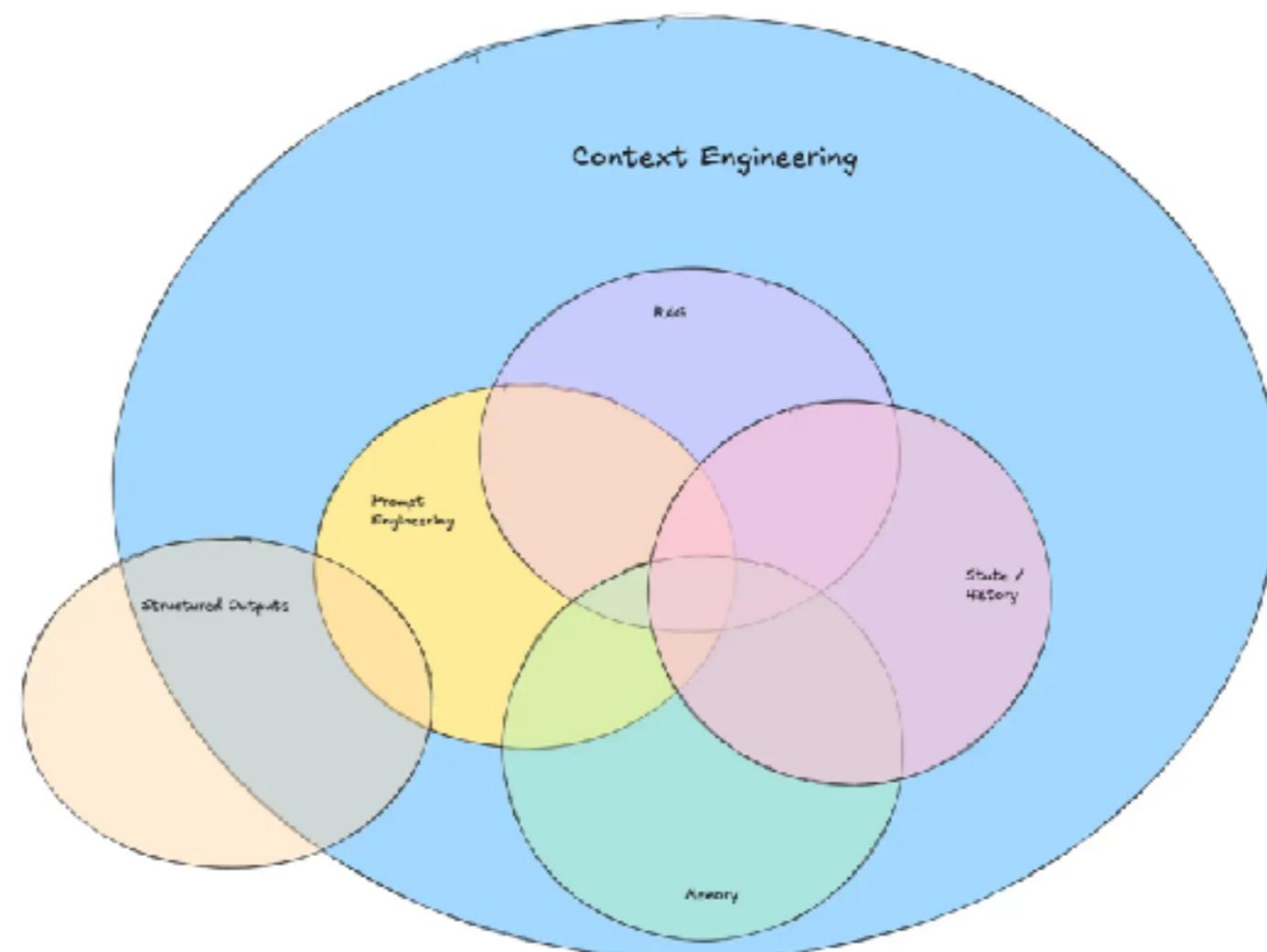


Context Engineering !!



Context Engineering

Tuning instructions and relevant context that LLM need



<https://www.promptingguide.ai/guides/context-engineering-guide>



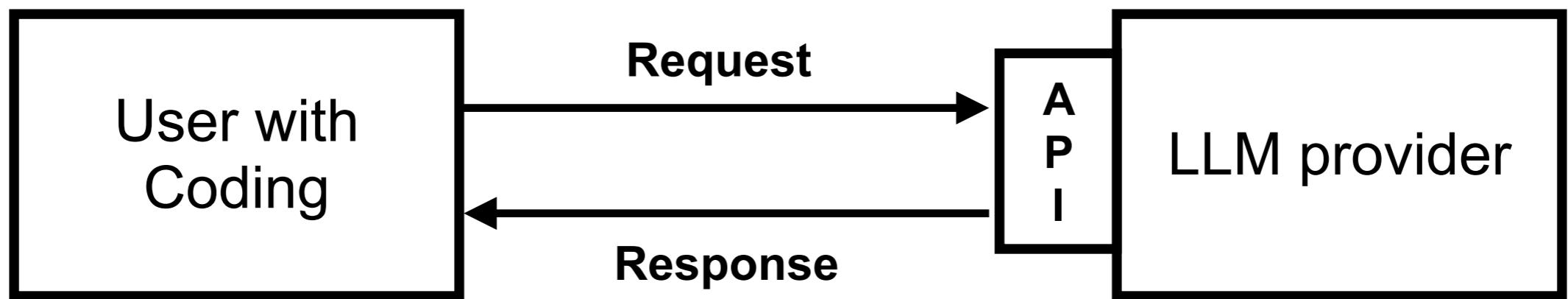
AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

Working with API



Working with APIs



<https://console.anthropic.com/dashboard>



Anthropic and OpenAI SDK

```
from openai import OpenAI

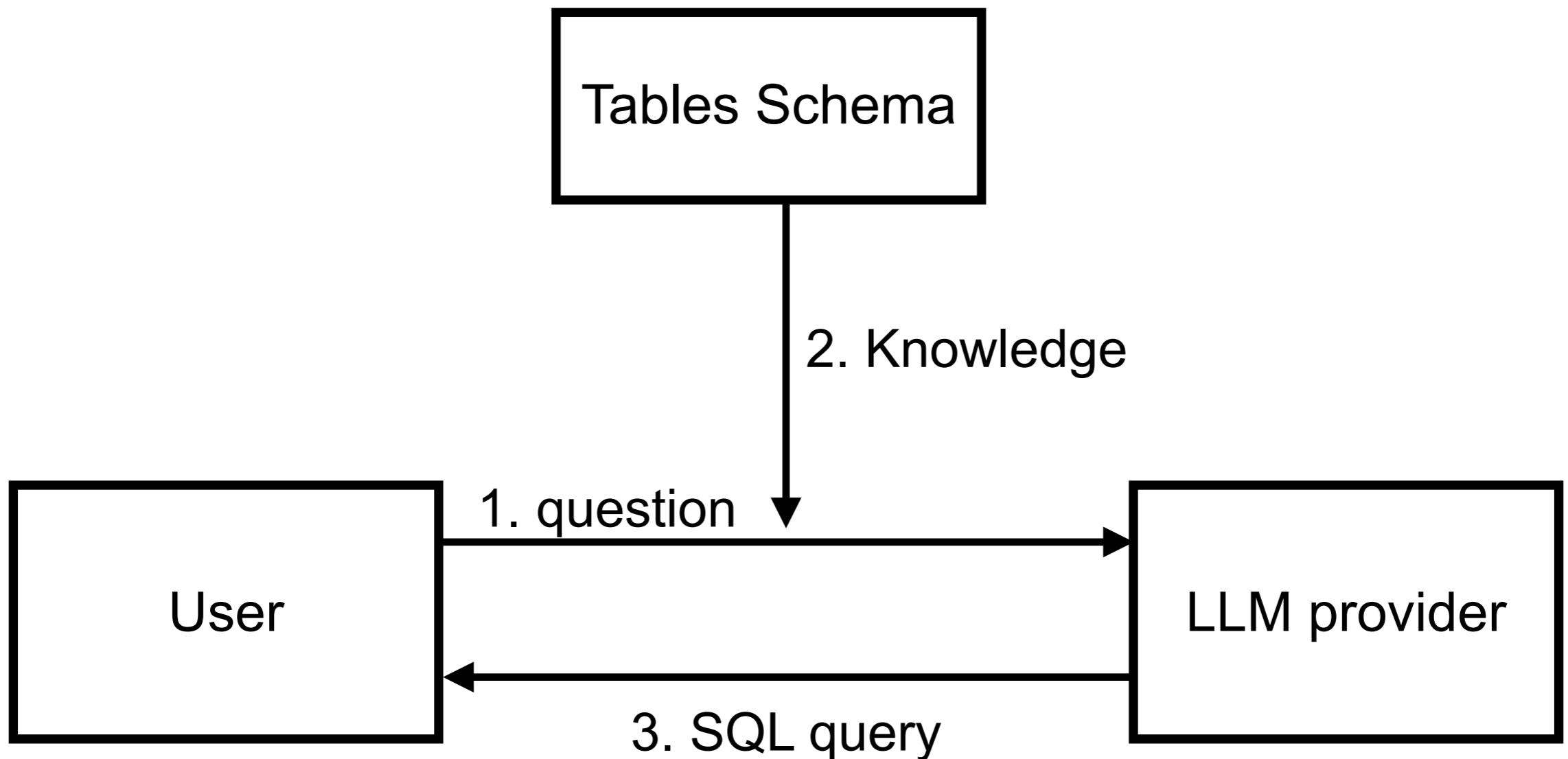
client = OpenAI(
    api_key="ANTHROPIC_API_KEY", # Your Anthropic API key
    base_url="https://api.anthropic.com/v1/" # Anthropic's API endpoint
)

response = client.chat.completions.create(
    model="claude-opus-4-1-20250805", # Anthropic model name
    messages=[
        {"role": "system", "content": "You are a helpful assistant."},
        {"role": "user", "content": "Who are you?"}
    ],
)
print(response.choices[0].message.content)
```

<https://docs.anthropic.com/en/api/openai-sdk>



Text-to-SQL

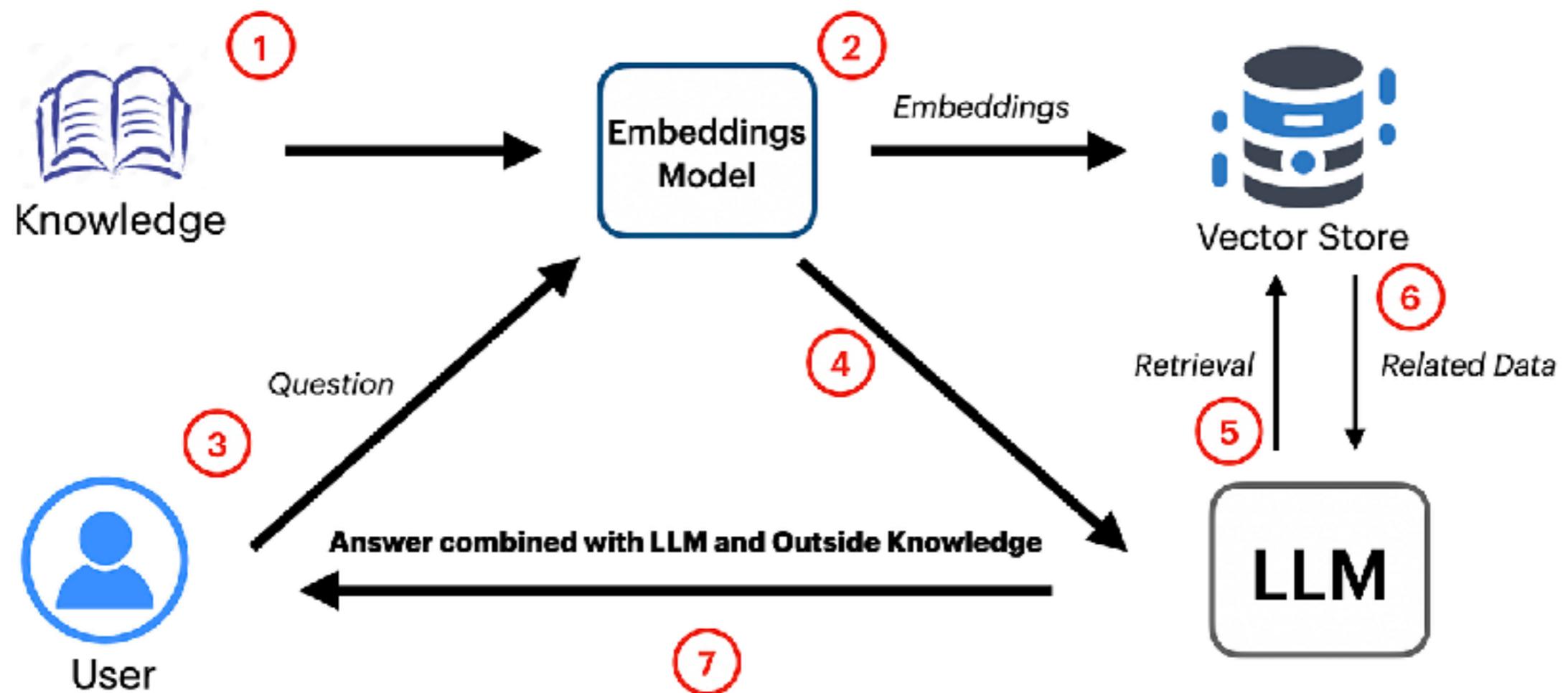


<https://github.com/up1/workshop-ai-with-technical-team/tree/main/workshop/demo-sql>



ChatBot !!

RAG Enhanced Chatbot



https://github.com/up1/workshop-basic-llm/tree/main/workshop/customer_support

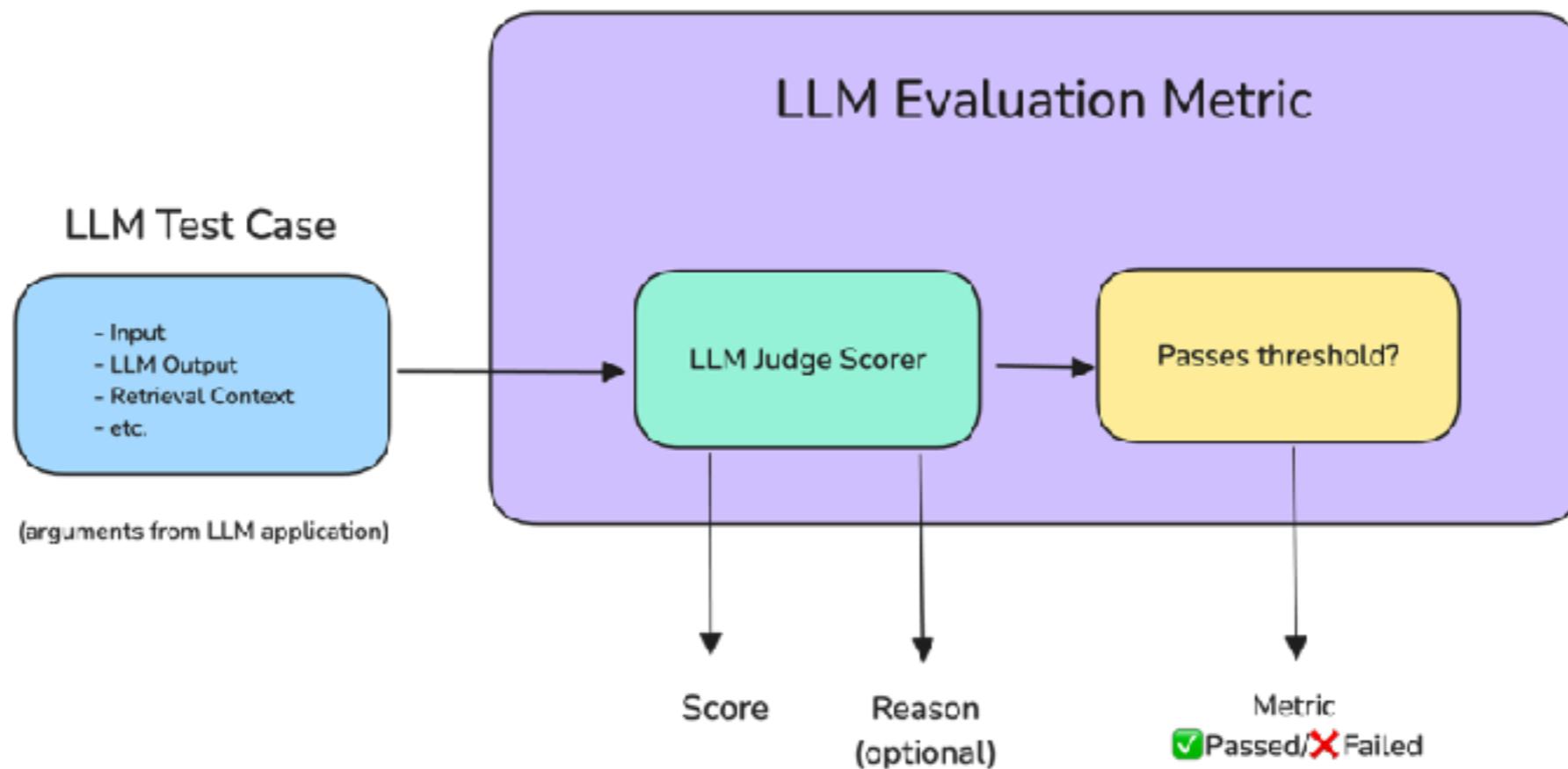


Evaluation in LLM



Evaluation in LLM

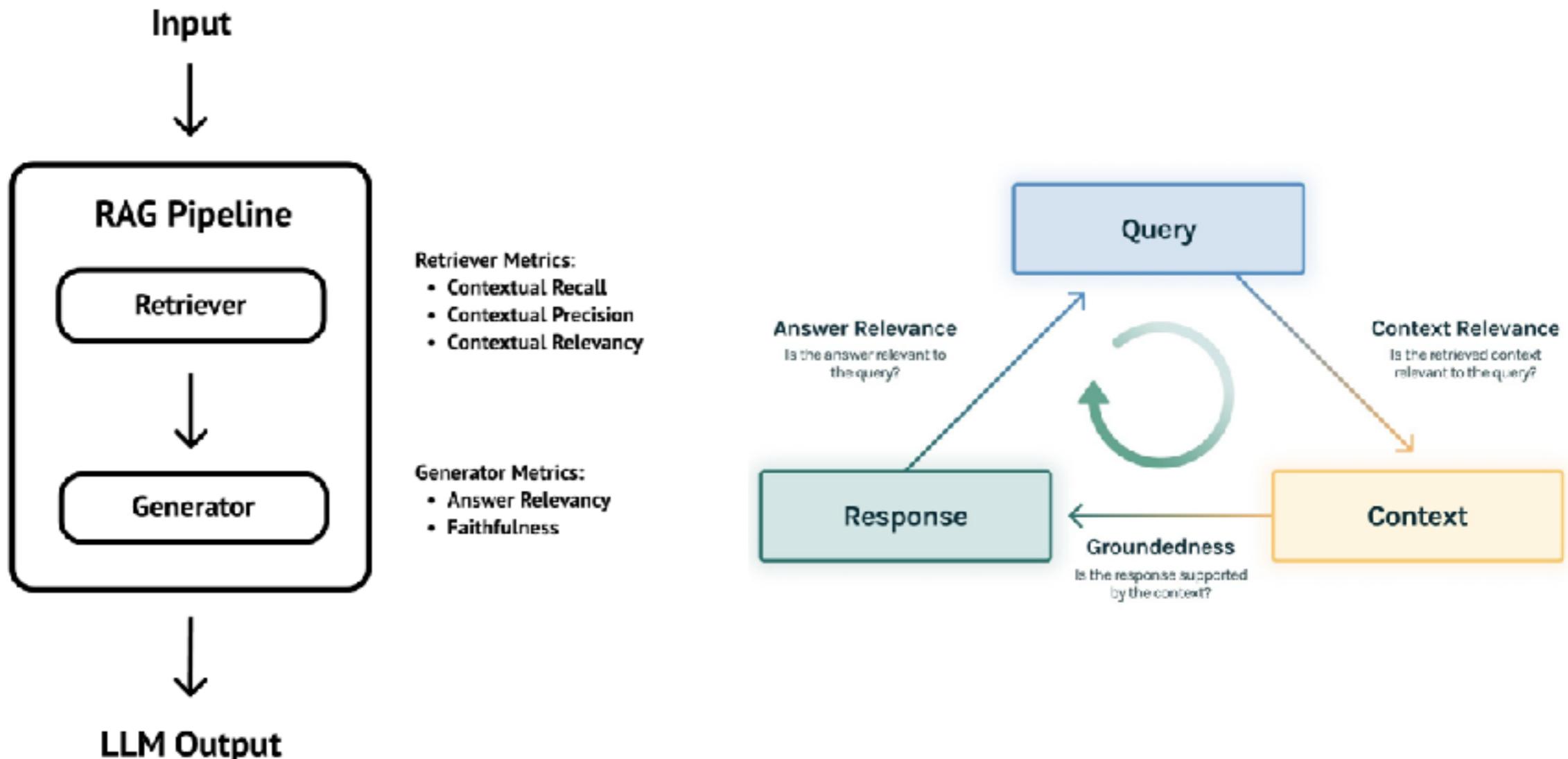
Answer correctness, context relevant
Semantic similarity
Hallucinations



<https://www.confident-ai.com/blog/llm-evaluation-metrics-everything-you-need-for-llm-evaluation>



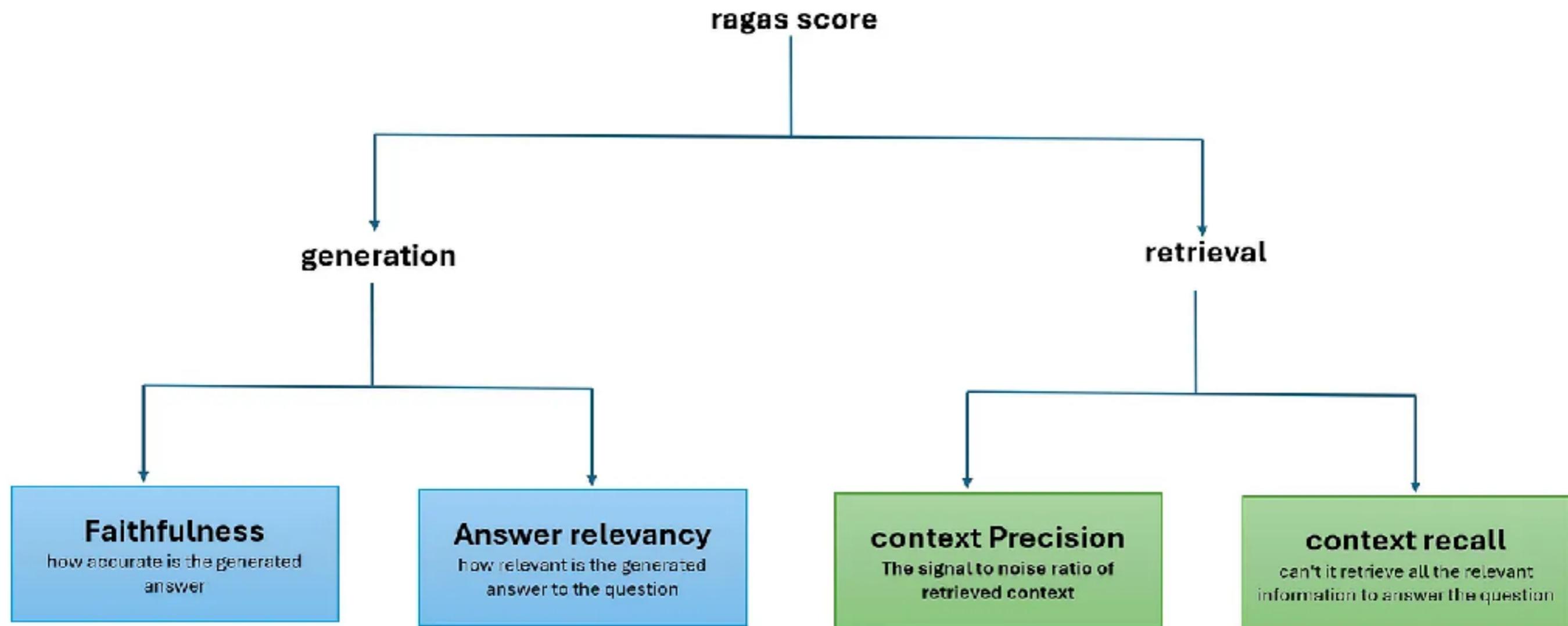
Evaluation in LLM



https://atamel.dev/posts/2025/01-14_rag_evaluation_deepeval/



Evaluation in LLM



<https://dkaarthick.medium.com/ragas-for-rag-in-llms-a-comprehensive-guide-to-evaluation-metrics-3aca142d6e38>



3 Evaluation Framework Integrations

Haystack Version: All ▾

Search integrations



Reset filters

Evaluation Framework ▾

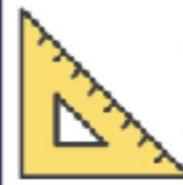
All Maintainers ▾

DeepEval.

DeepEval

Use the DeepEval evaluation framework to calculate model-based metrics

☆ Haystack 2.0 ☆ Evaluation Framework



Ragас

Ragас

Use the Ragас evaluation framework to calculate model-based metrics

☆ Haystack 2.0 ☆ Evaluation Framework



UpTrain

UpTrain

Use the UpTrain evaluation framework to calculate model-based metrics

☆ Haystack 2.0 ☆ Evaluation Framework

https://x.com/Haystack_AI/status/1760726816526925839/photo/1



AI

© 2020 - 2025 Siam Chamnankit Company Limited. All rights reserved.

127

Working with Claude Code



Claude Code

AI coding assistant by Anthropic
Supervised coding agent

Your code's new collaborator

Unleash Claude's raw power directly in your terminal. Search million-line codebases instantly. Turn hours-long workflows into a single command. Your tools. Your workflow. Your codebase, evolving at thought speed.

[Try Claude Code on Max](#)

[See our pricing options](#)



<https://www.anthropic.com/clause-code>



Claude Code

Terminal-based, not IDE

Work with all tools

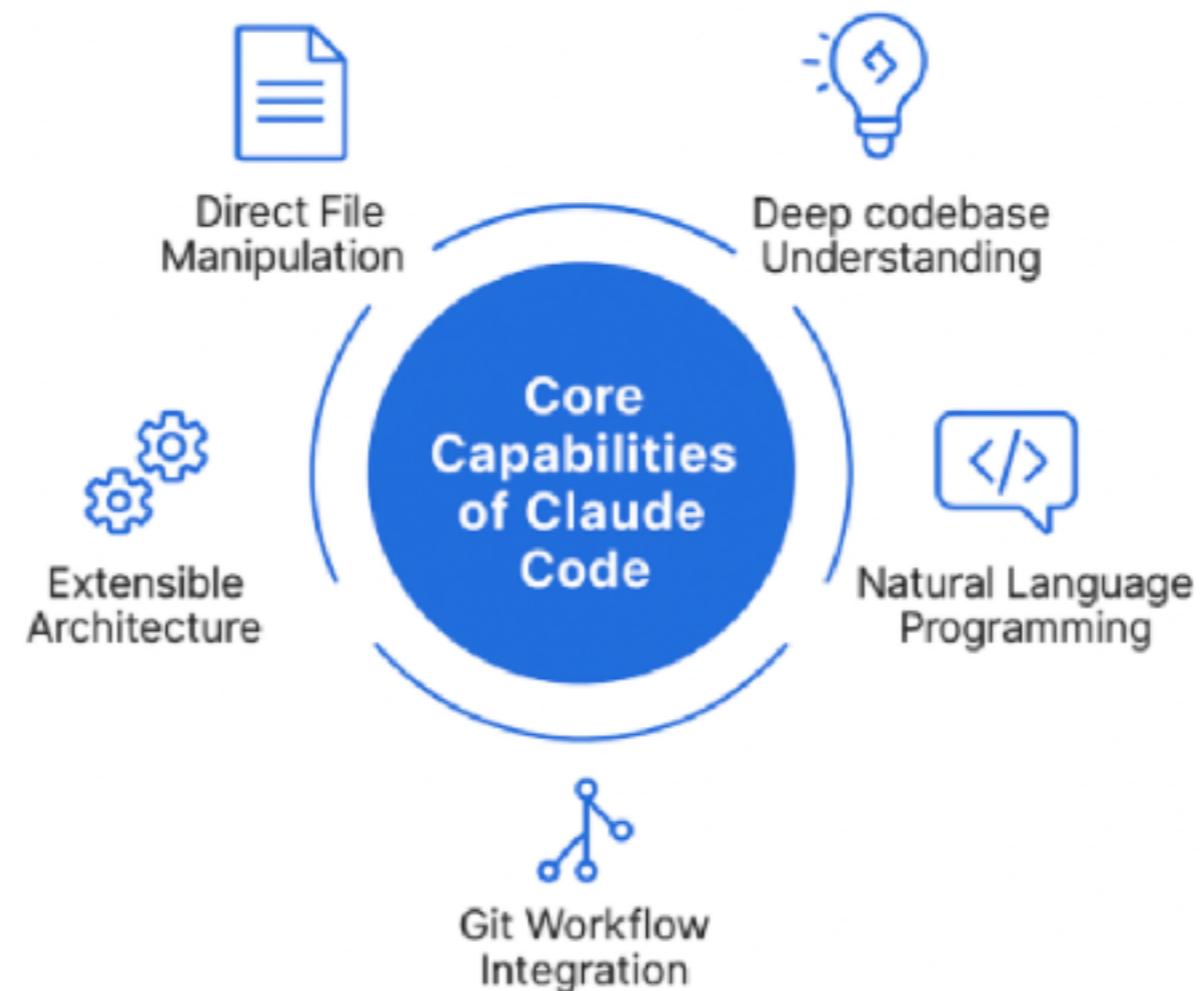
Fit into existing workflows

General purpose

Infinitely hackable



Capability of Claude Code



<https://www.analyticsvidhya.com/blog/2025/07/what-is-claude-code/>



Build-in tools in Claude Code

Bash, cmd

File search

File listing

File read and write

Web fetch and search

TODOs

Sub-agents

Working with MCP

<https://www.youtube.com/watch?v=6eBSHbLKuN0>



More context = better performance

CLAUDE.md
#create memory

MCP resources
(Memory)

Slash commands

Mention with
@filename

Take time to tune context for your team



Let's start

```
$npm install -g @anthropic-ai/clause-code  
$clause
```

<https://docs.anthropic.com/en/docs/clause-code/common-workflows>



Common workflows ?

Explore > plan > confirm > code > commit

Write tests > commit > code > iterate > commit

Write code > screenshot result > iterate



Example

Implement [google.png] then screenshot it with Puppeteer and iterate until it look like the google

Plan and create TODO list before coding !!

Update Todos

- └ Create HTML structure for Google homepage
- Implement CSS styling to match Google's design
- Add Google logo and search functionality
- Set up Puppeteer for screenshots
- Take initial screenshot and compare
- Iterate and refine styling based on comparison



AI

© 2020 - 2025 Siam Chamnkit Company Limited. All rights reserved.

Check API Usage !!!

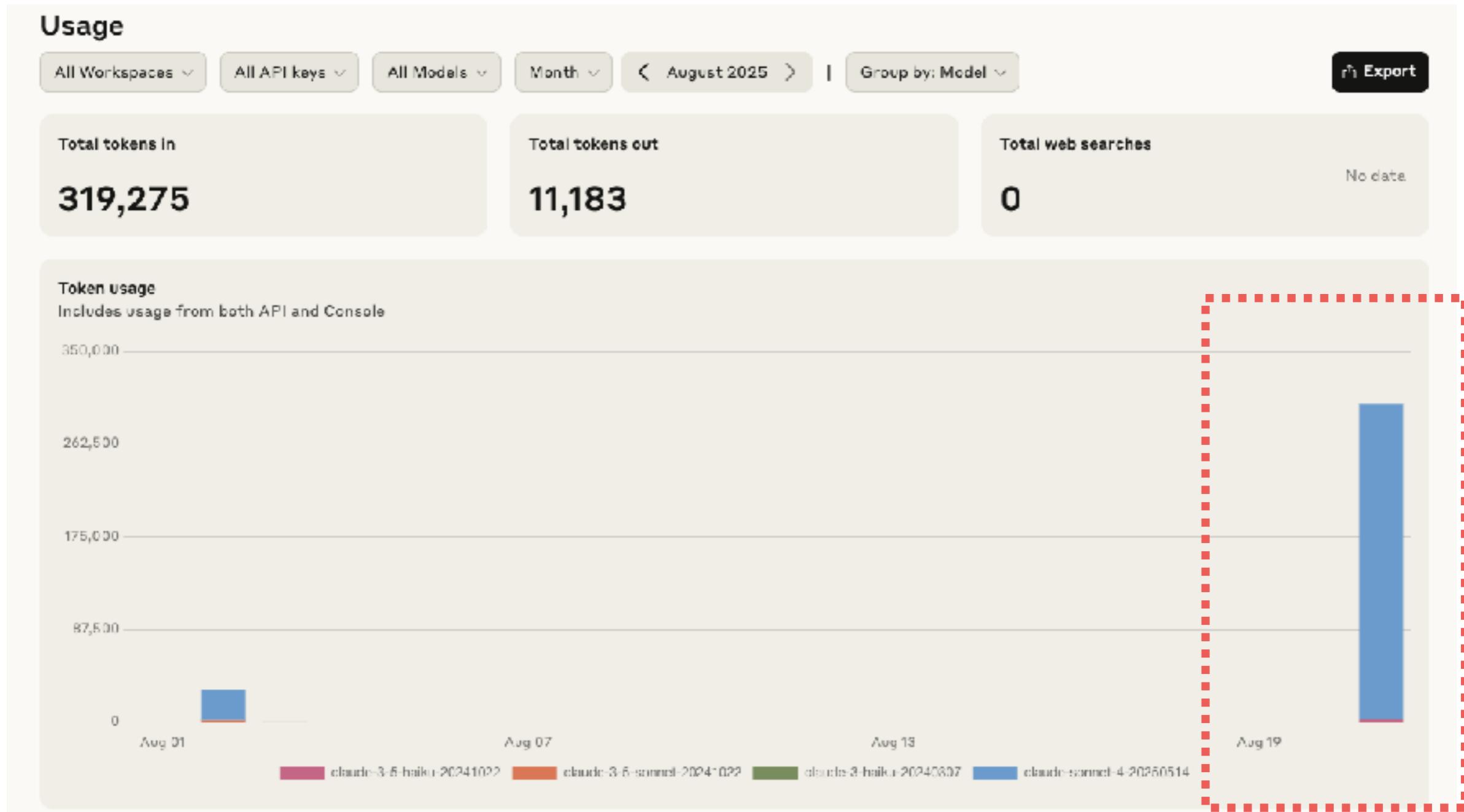
/cost

```
/cost
└ Total cost:          $0.3486
    Total duration (API): 2m 26.7s
    Total duration (wall): 20m 0.4s
    Total code changes:   290 lines added, 0 lines removed
    Usage by model:
        claude-3-5-haiku: 6.2k input, 110 output, 0 cache read, 0 cache
        write
        claude-sonnet:   84 input, 5.6k output, 558.4k cache read, 24.4k
        cache write
```

<https://docs.anthropic.com/en/docs/clause-code/costs>



Check API Usage !!!



<https://console.anthropic.com/usage>



Claude Code Modes ?

Default mode

Auto mode

Plan mode

Strategic thinking first

<https://www.siddharthbharath.com/clause-code-the-complete-guide/>



Workshop

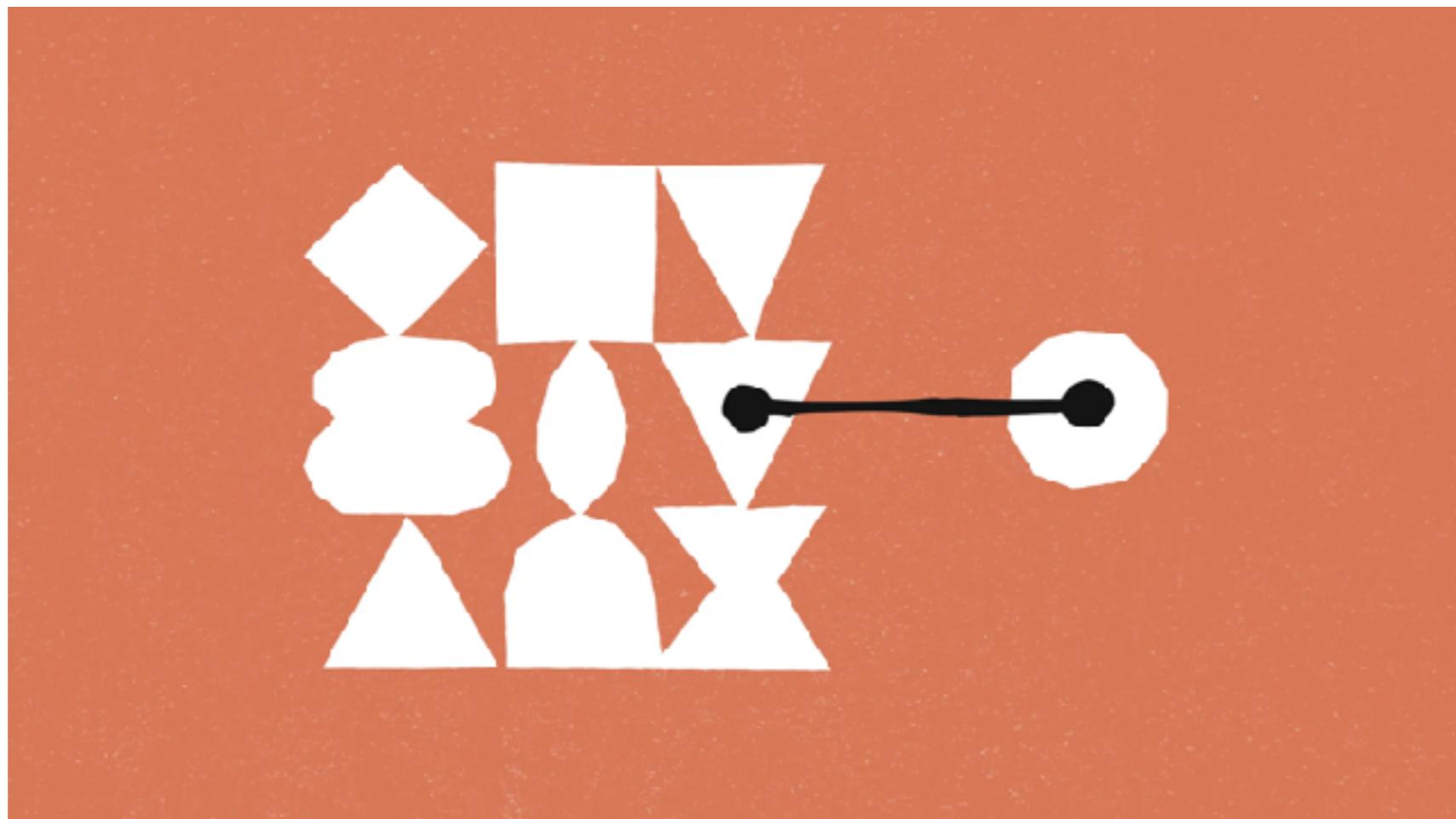


Model Context Protocol (MCP)



Model Context Protocol (MCP)

New standard for connecting AI assistants to system



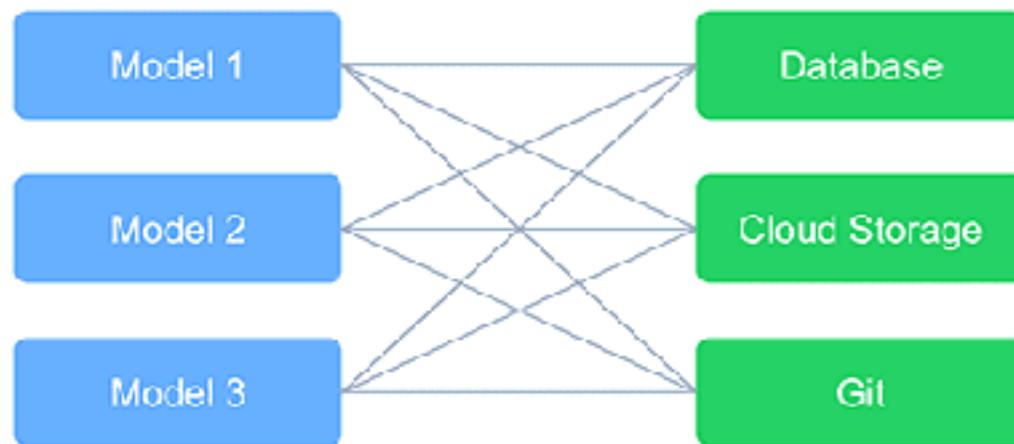
<https://www.anthropic.com/news/model-context-protocol>



Model Context Protocol (MCP)

Traditional Integration vs MCP Approach

Traditional: N×M Connections



Each model needs custom integration
with each data source

9 Total Connections

MCP: N+M Connections



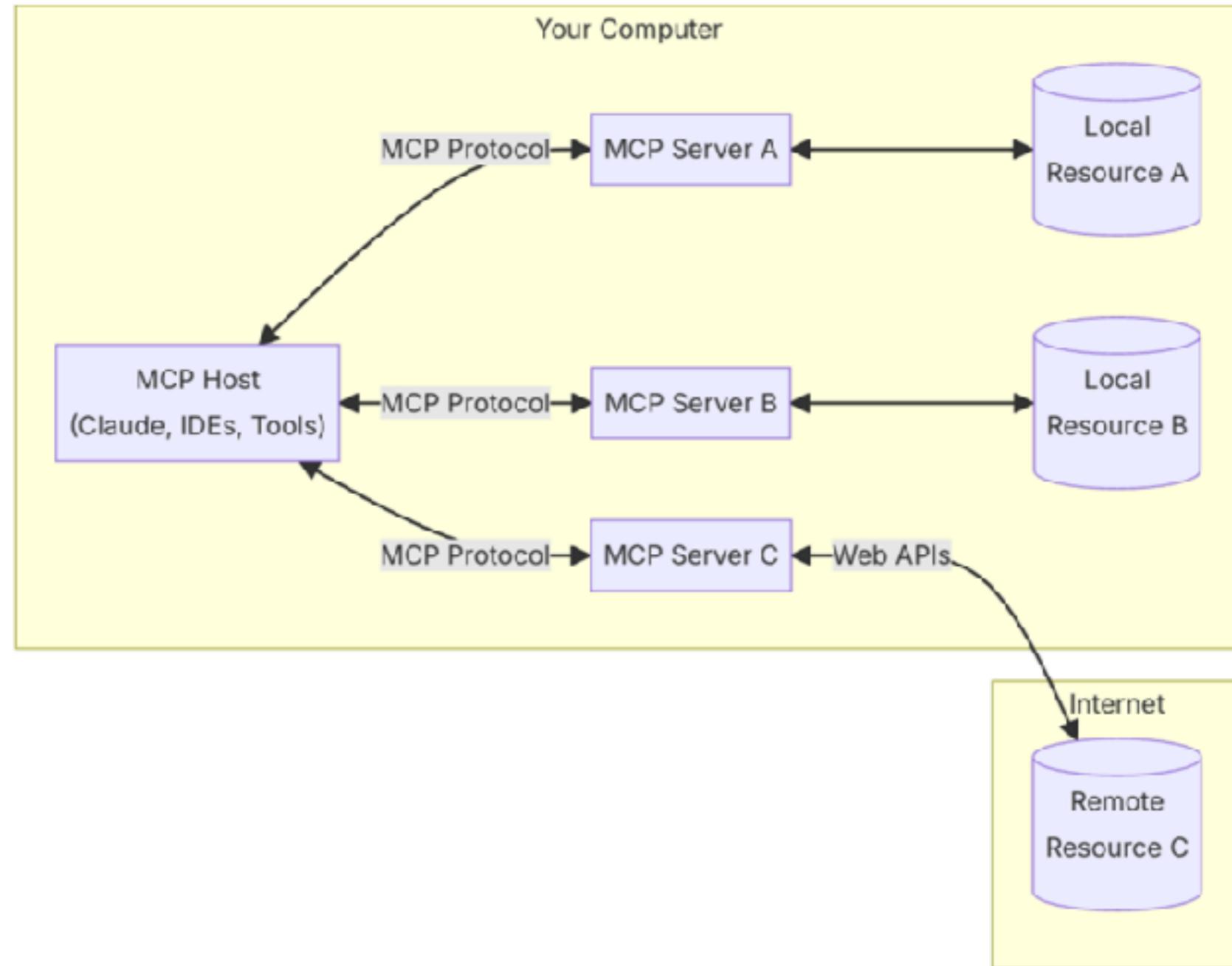
Models and data sources only need
to integrate once with MCP

6 Total Connections

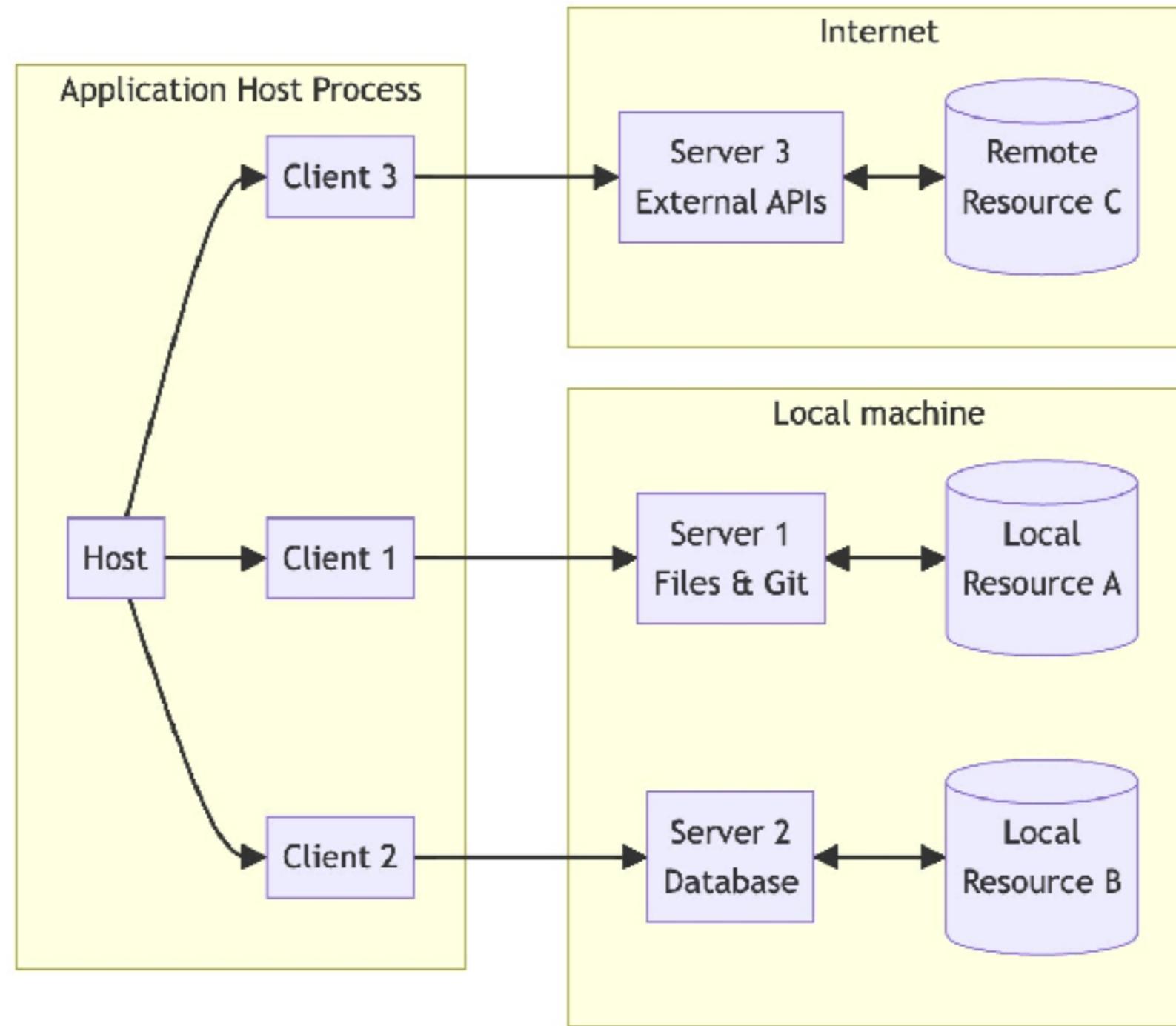
<https://salesforcedevops.net/index.php/2024/11/29/anthropics-model-context-protocol/>



Model Context Protocol (MCP)



Model Context Protocol (MCP)



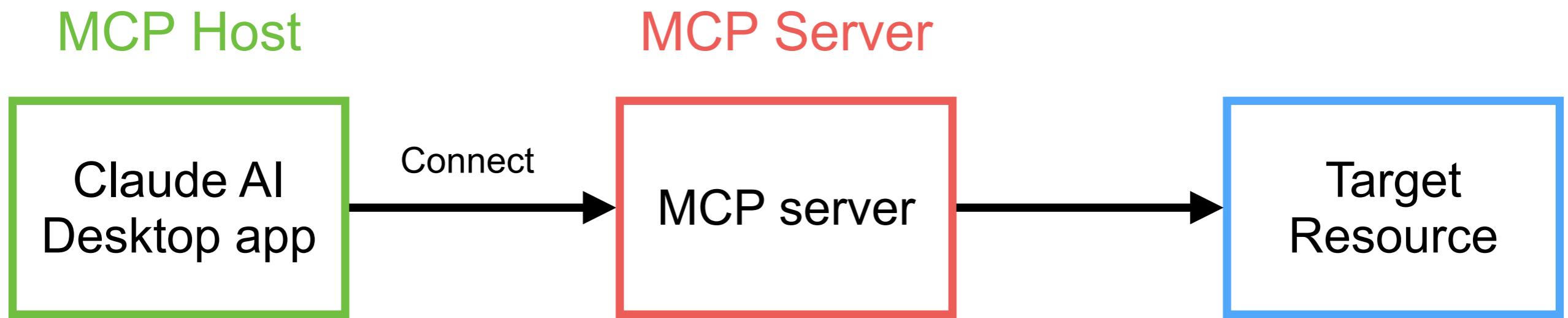
List of MCP Servers

The screenshot shows the MCP.so website interface. At the top left is the logo 'MCP.so'. To its right are navigation links for 'Servers' and 'Clients'. On the far right are a user icon and a 'Sign In' button. Below the header, a message box displays '1577 MCP Servers in list'. The main content area features a large heading 'Find Awesome MCP Servers and Clients' in bold black and red text. Below it is a subtitle 'The largest collection of MCP Servers.' A search bar at the bottom left contains the placeholder 'Search with keywords'.

<https://mcp.so/>



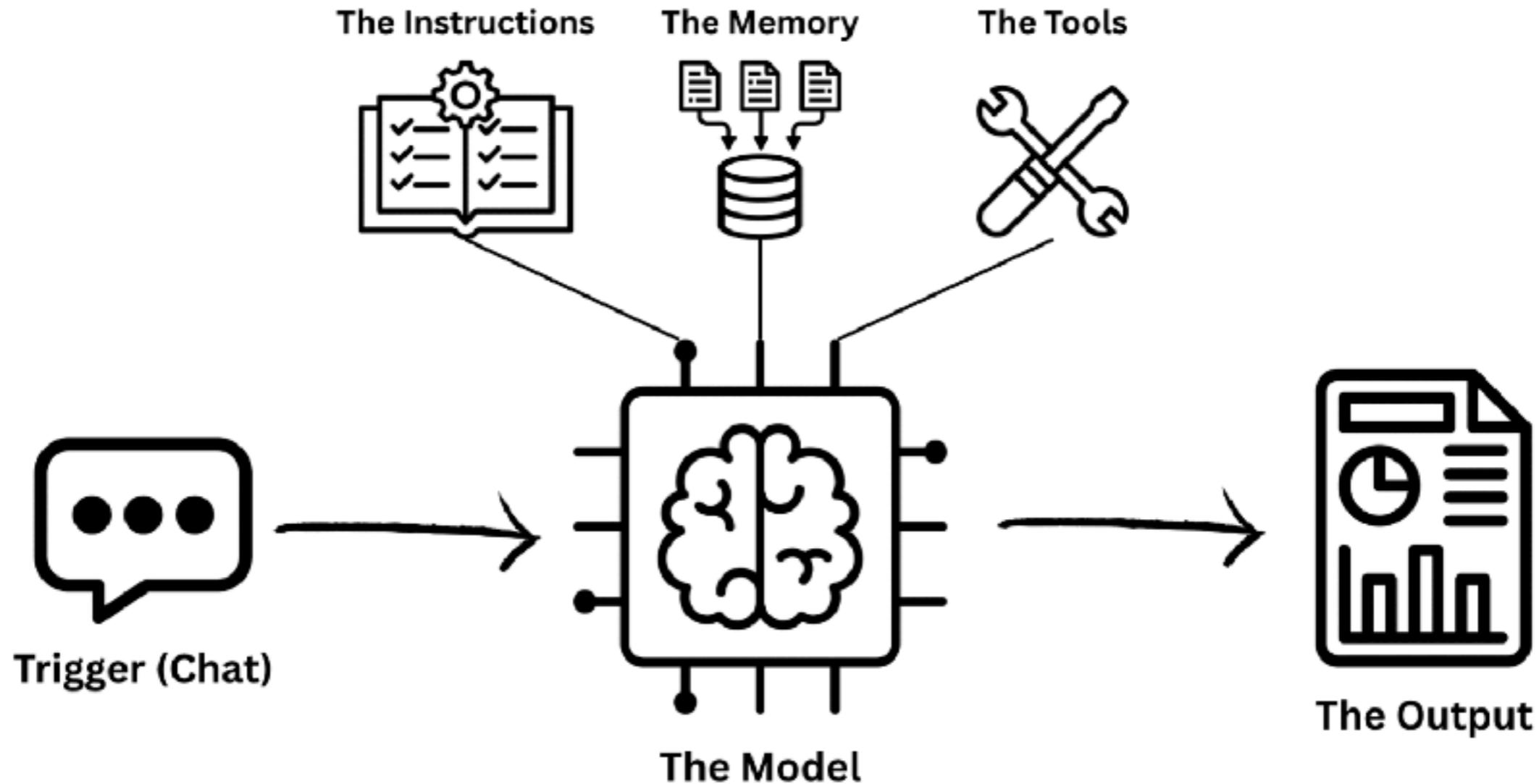
Example MCP by Anthropic



<https://www.somkiat.cc/model-context-protocol/>



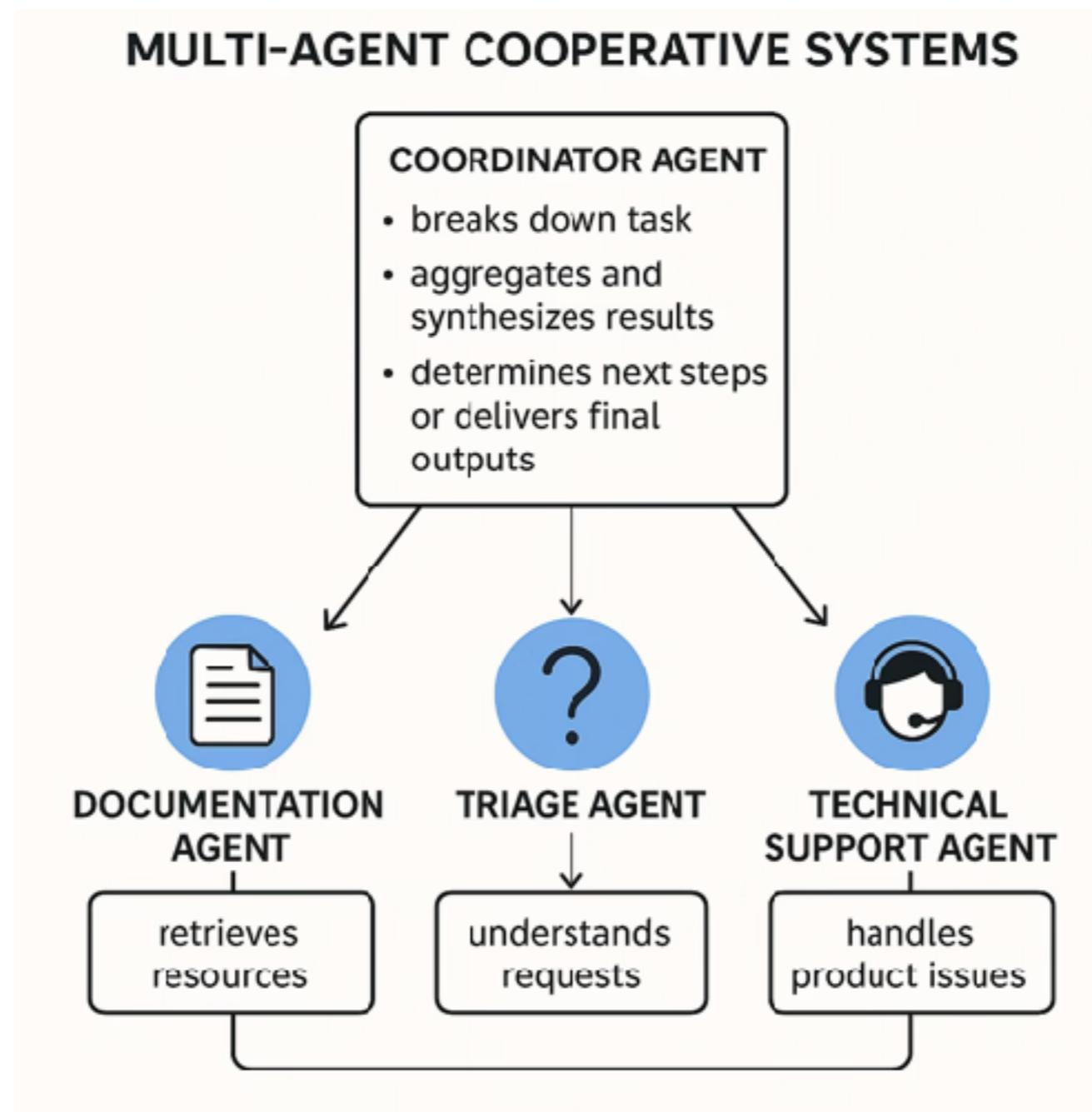
AI Agent



<https://www.siddharthbharath.com/ultimate-guide-ai-agents/>



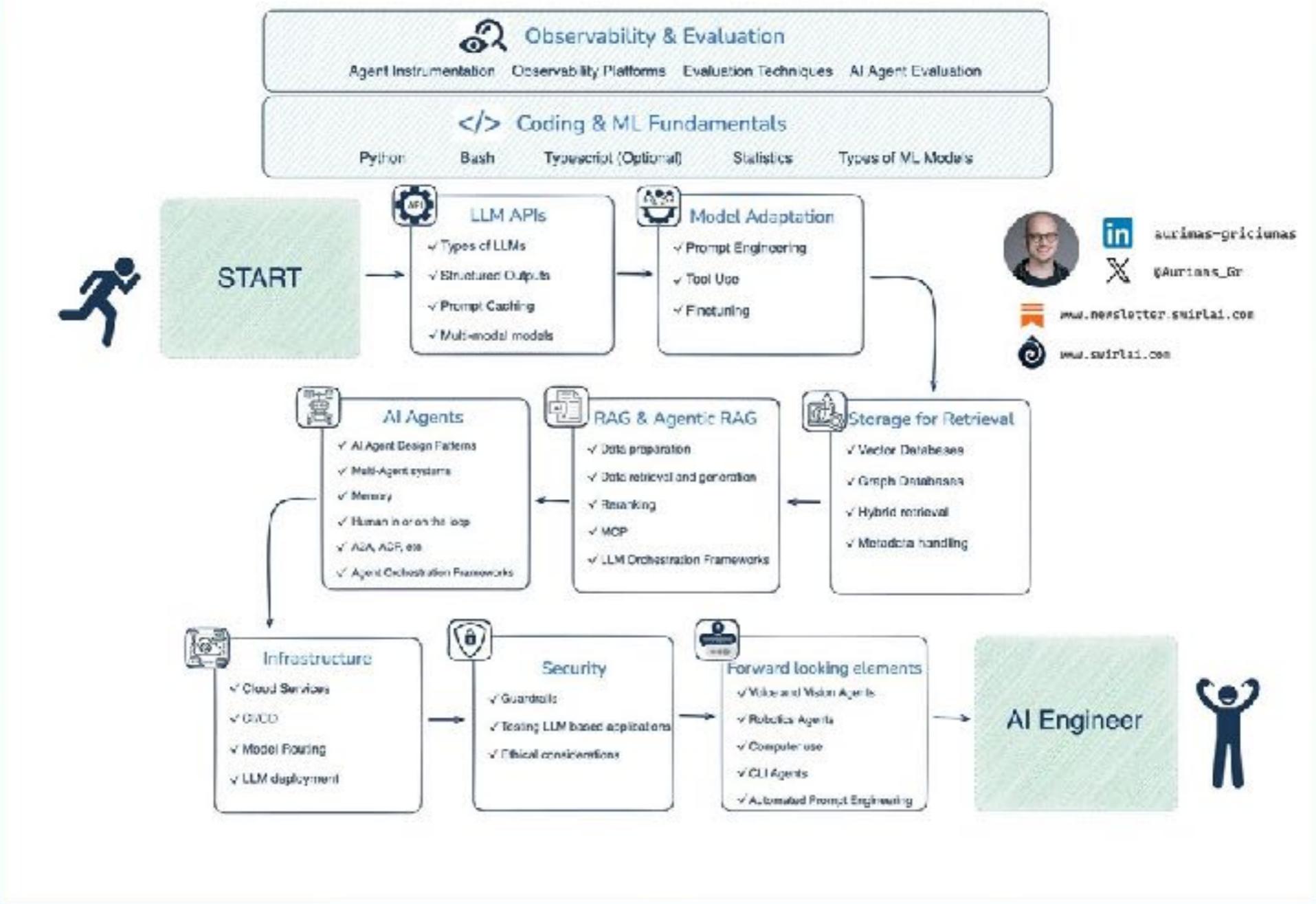
Multiple AI Agent



<https://www.siddharthbharath.com/ultimate-guide-ai-agents/>



AI Engineering Learning Roadmap



Q/A

