EE P 596
LLMs: From Transformers to ChatGPT

Introduction | LLM Motivation | History of LLMs

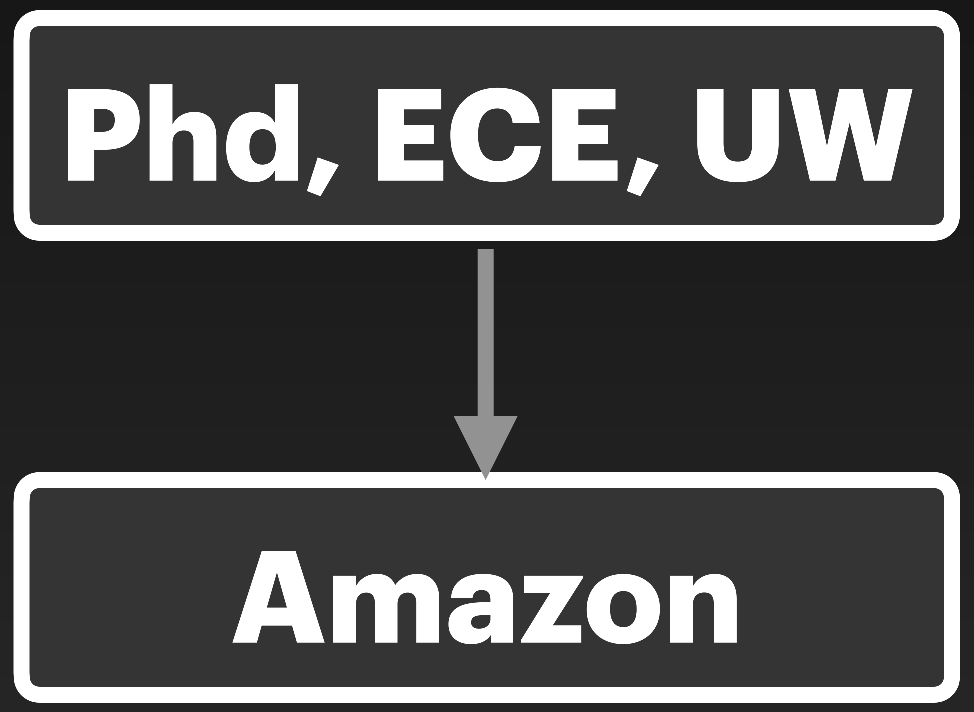
Dr. Karthik Mohan, Jan 4 2024 | Winter Quarter course | PMP, ECE, UW

Bit about Me

Bit about Me

Phd, ECE, UW

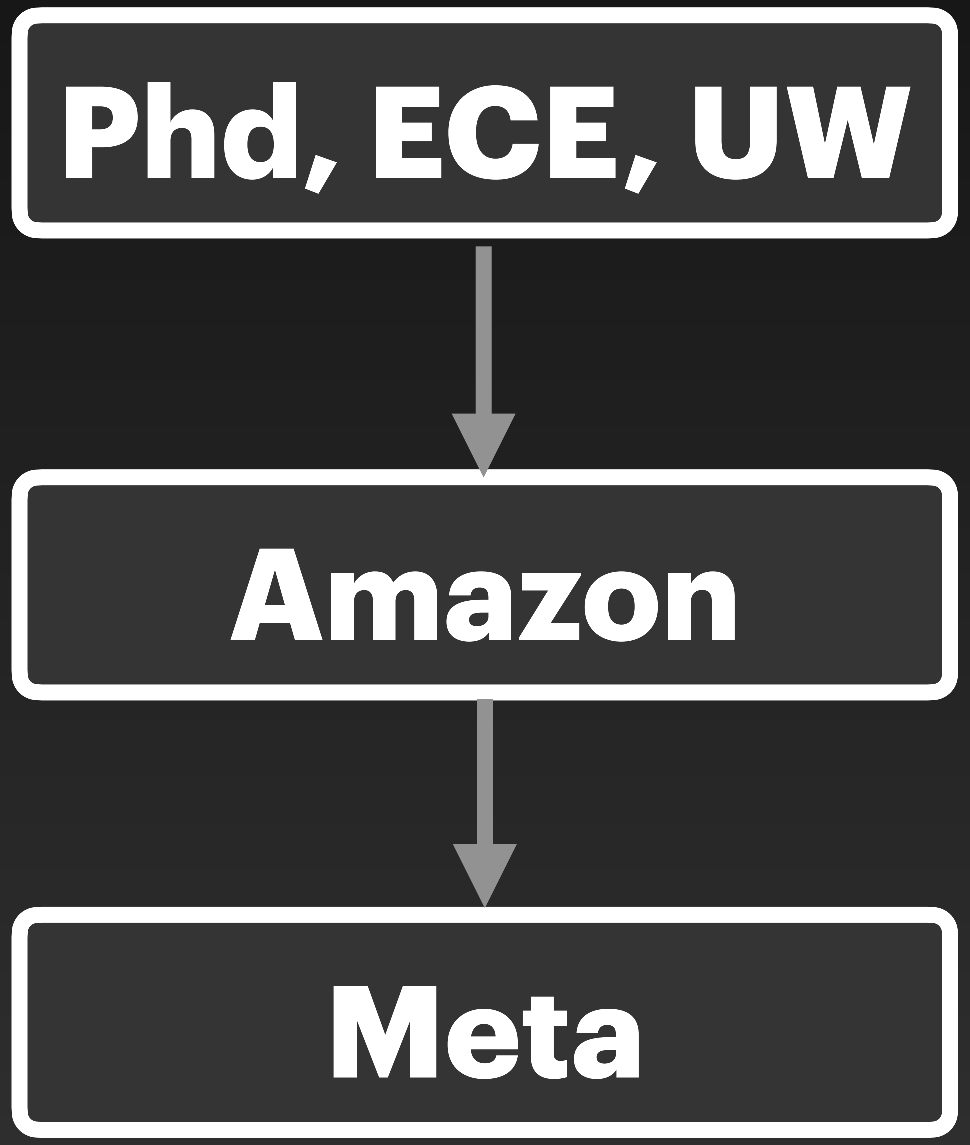
Bit about Me



Phd, ECE, UW

Amazon

Bit about Me

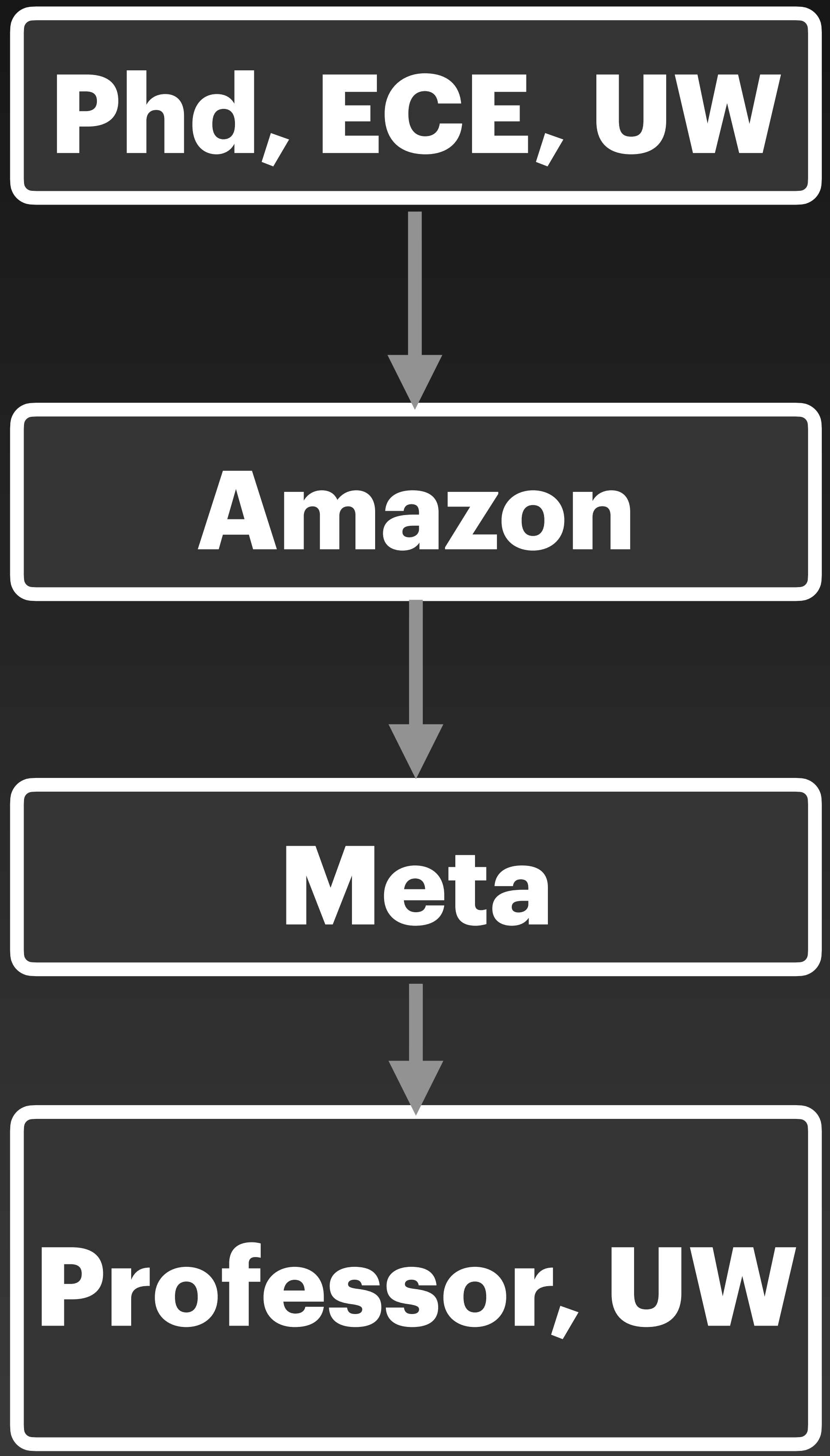


Phd, ECE, UW

Amazon

Meta

Bit about Me



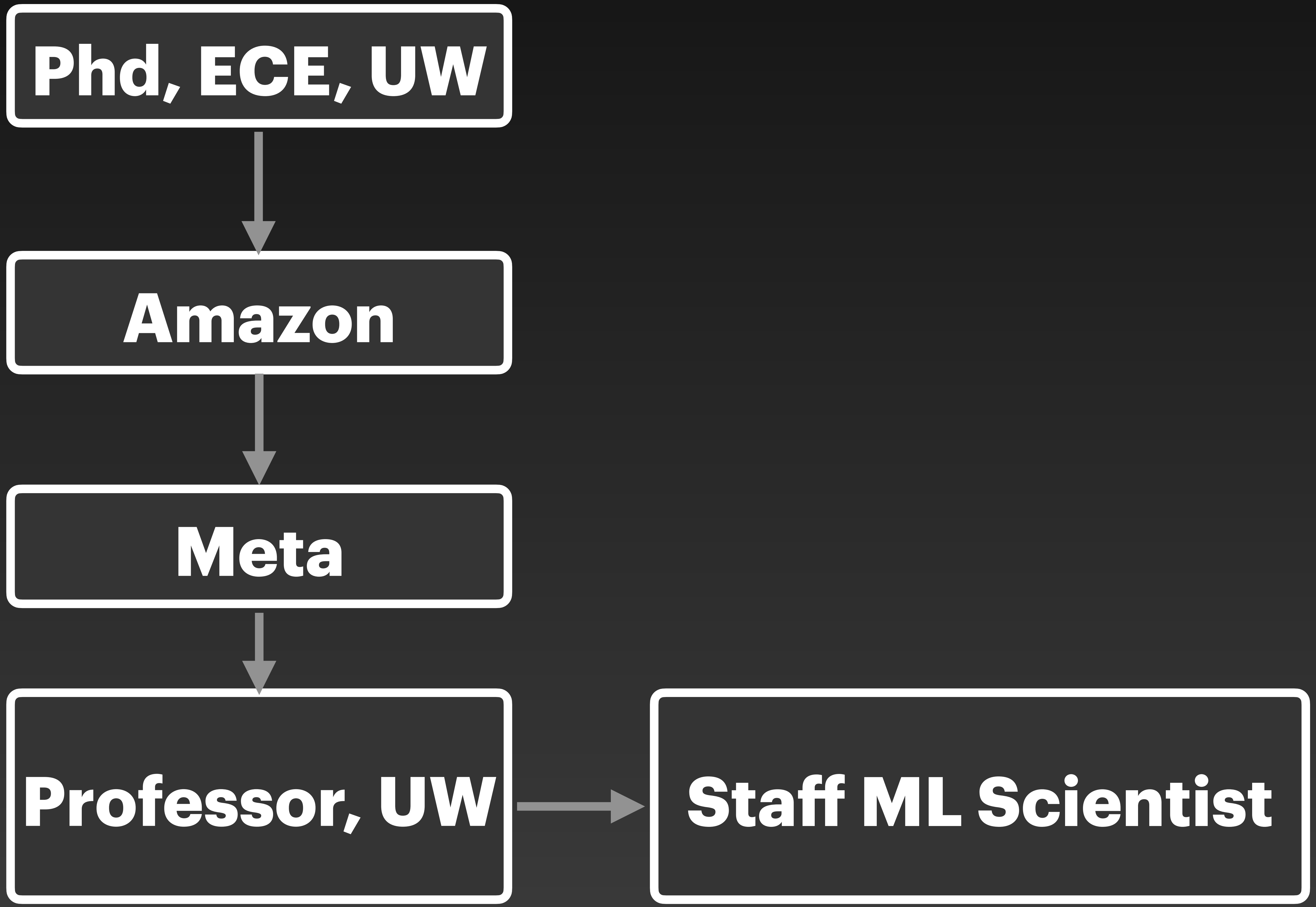
Phd, ECE, UW

Amazon

Meta

Professor, UW

Bit about Me



Phd, ECE, UW

Amazon

Meta

Professor, UW

Staff ML Scientist

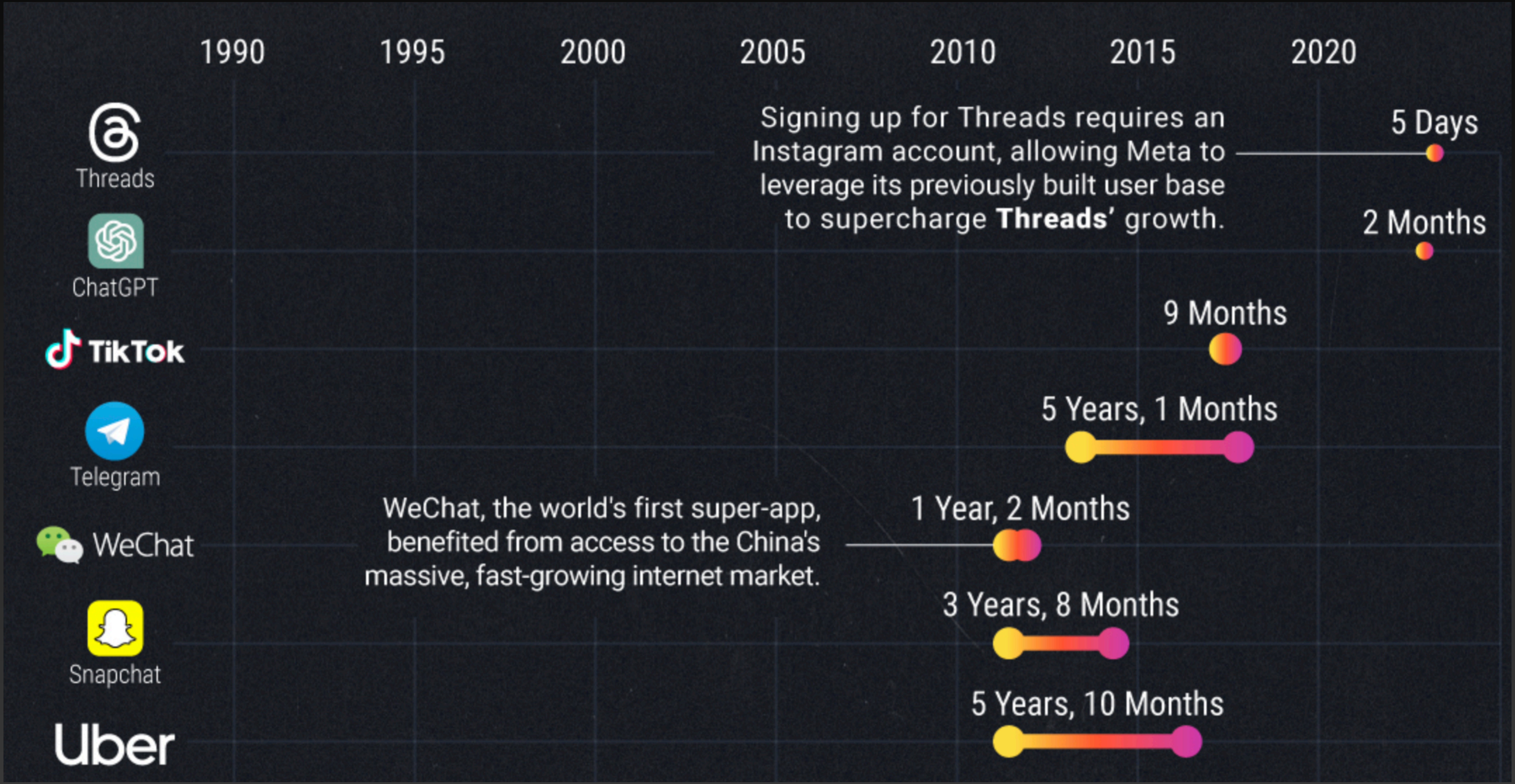
Teaching Support Team

Shreemit (TA)

Michael (TA)

Nikhil (Grader)

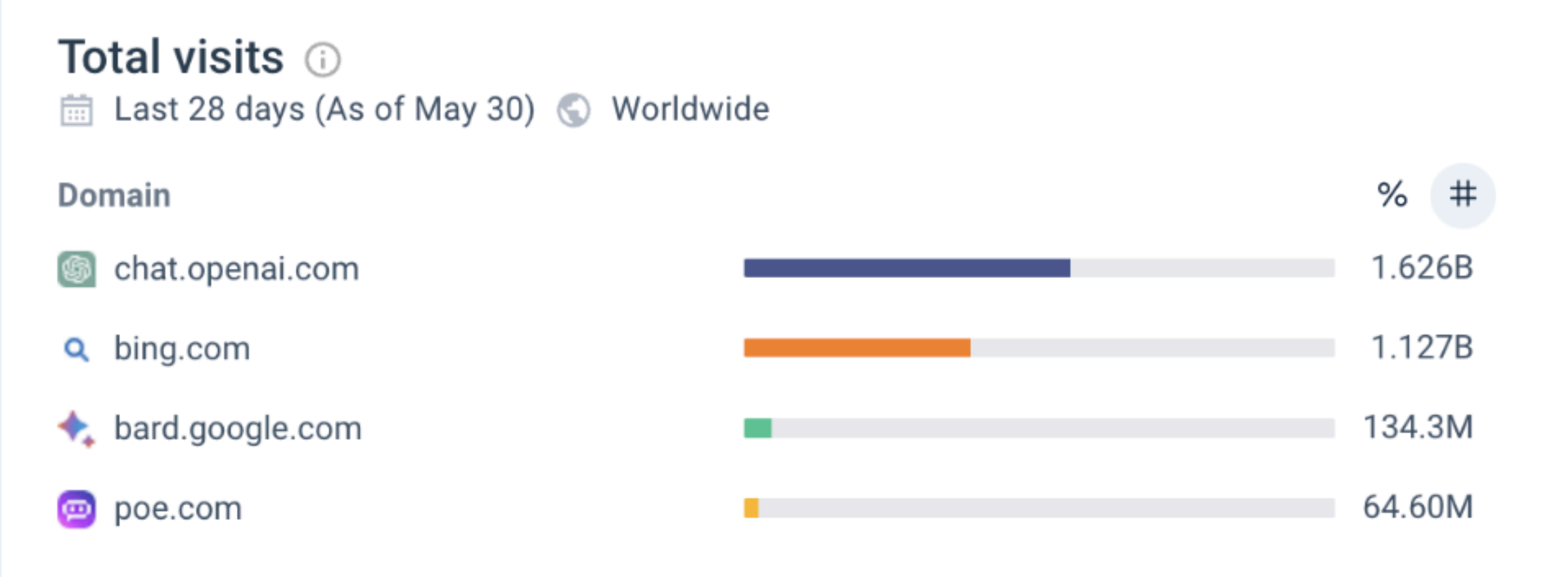
ChatGPT and LLMs are everywhere!



Signing up for Threads requires an Instagram account, allowing Meta to leverage its previously built user base to supercharge **Threads'** growth.

WeChat, the world's first super-app, benefited from access to the China's massive, fast-growing internet market.

ChatGPT and LLMs are everywhere!

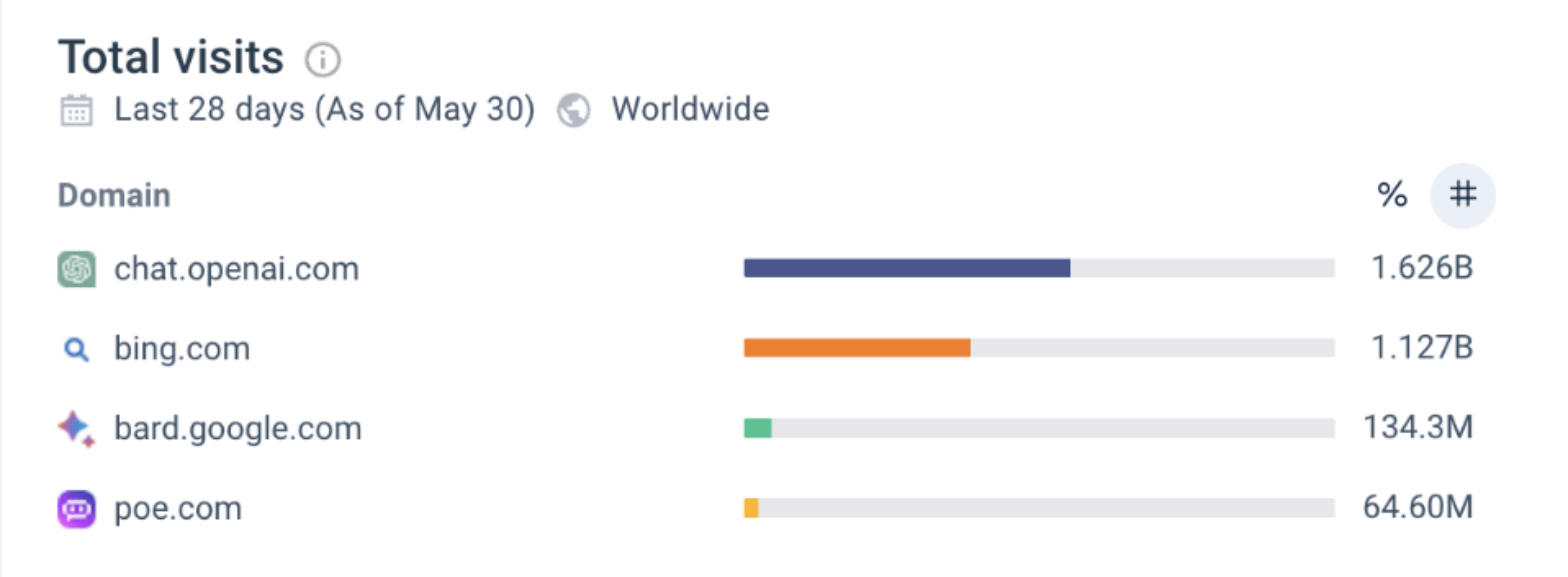


Total visits ⓘ

Last 28 days (As of May 30) Worldwide

Domain	% #
chat.openai.com	1.626B
bing.com	1.127B
bard.google.com	134.3M
poe.com	64.60M

ChatGPT and LLMs are everywhere!



Total visits

Last 28 days (As of May 30) Worldwide

Domain	
chat.openai.com	1.626B
bing.com	1.127B
bard.google.com	134.3M
poe.com	64.60M

ChatGPT and LLMs are everywhere!

Let’s look at some examples!

ChatGPT and LLMs are everywhere!

Paraphrasing

ChatGPT and LLMs are everywhere!

Paraphrasing

Math

ChatGPT and LLMs are everywhere!

Paraphrasing

Math

Coding

ChatGPT and LLMs are everywhere!

Let’s go checkout ChatGPT live!

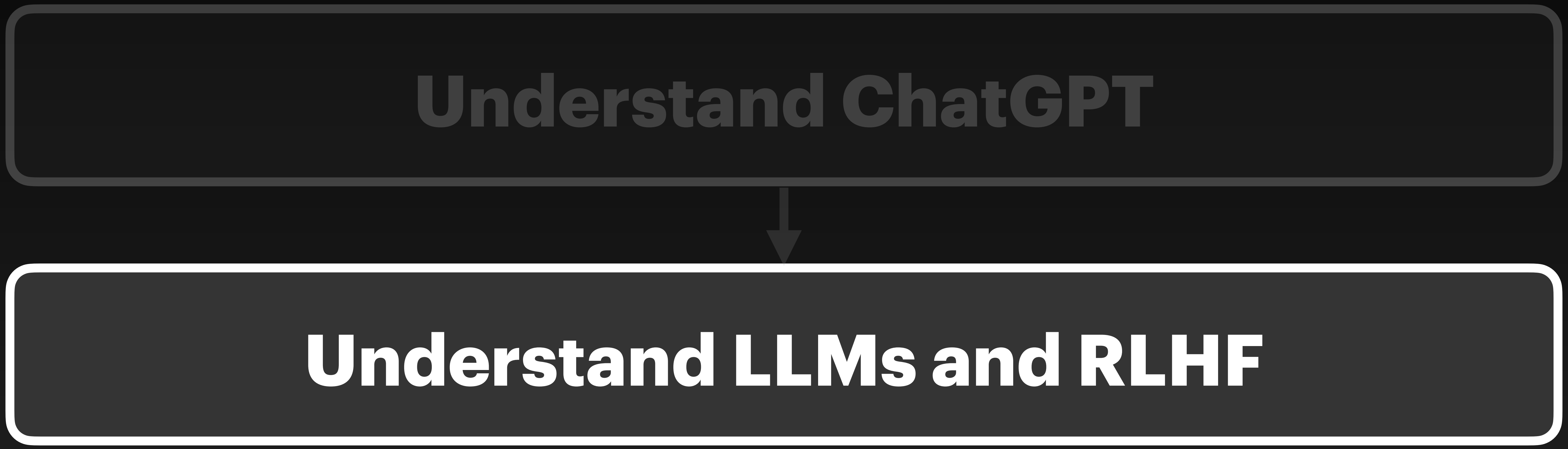
Engine behind ChatGPT

ChatGPT heavily relies on Large Language Models to power its responses to users!

How do you understand ChatGPT?

To Understand ChatGPT?

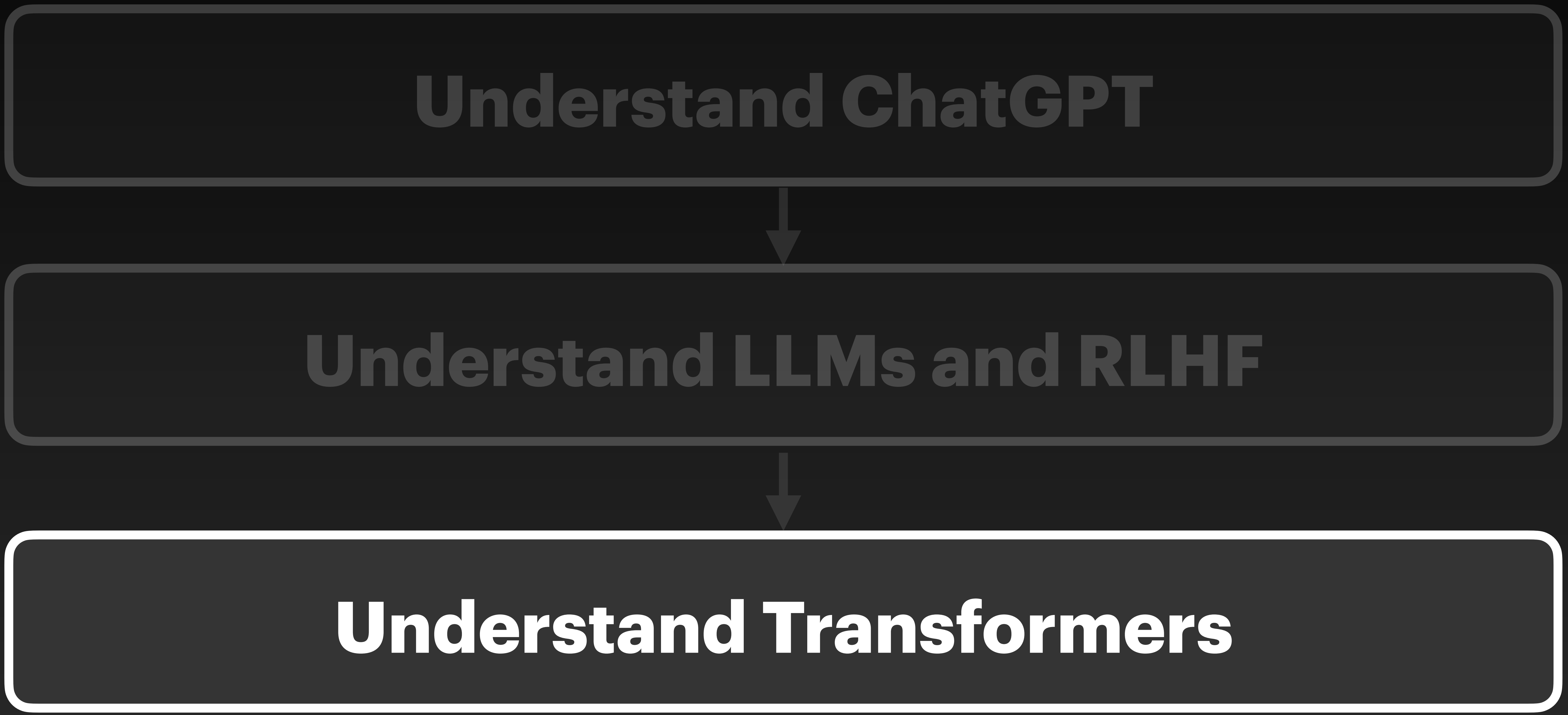
How do you understand ChatGPT?



Understand ChatGPT

Understand LLMs and RLHF

How do you understand ChatGPT?

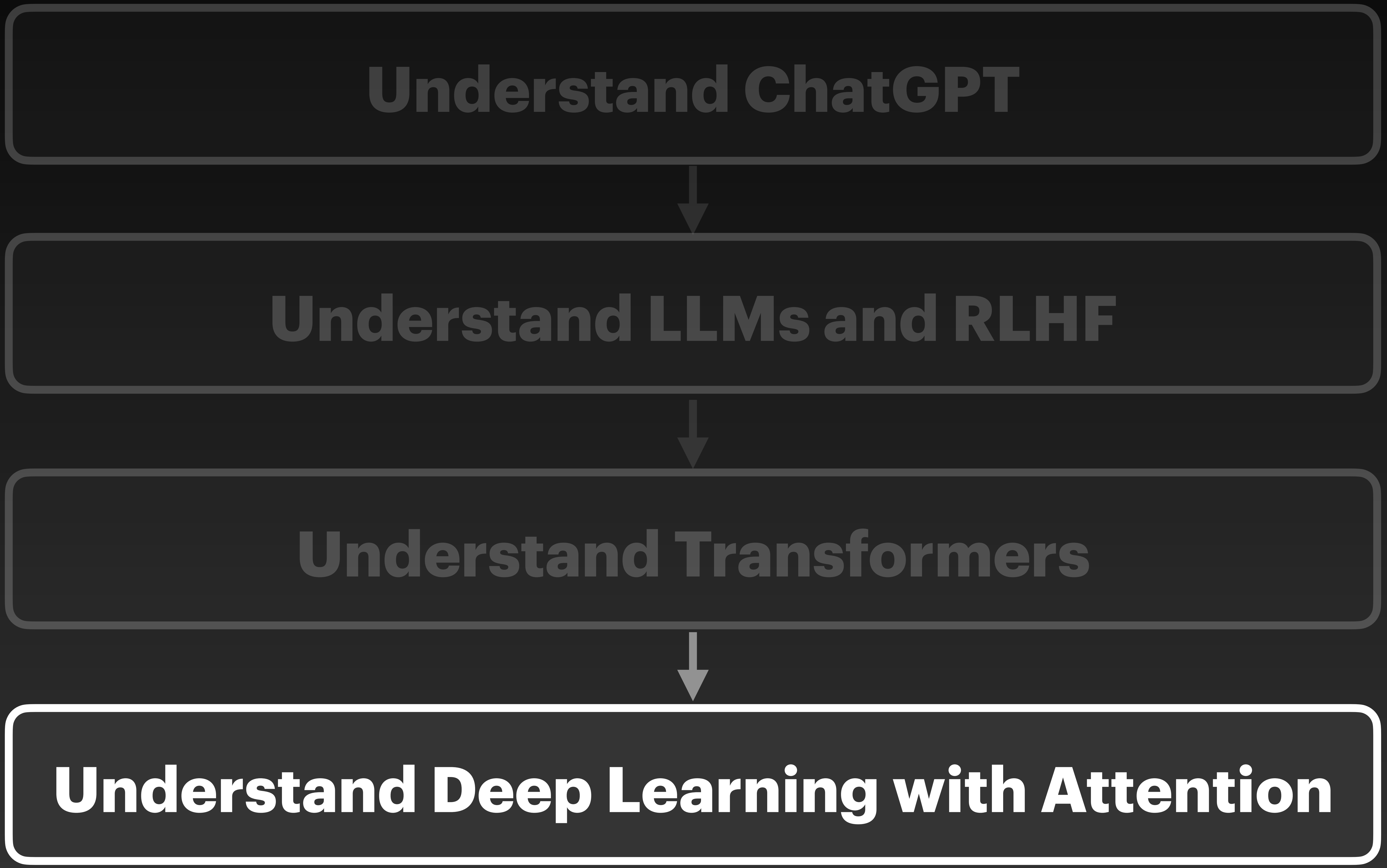


Understand ChatGPT

Understand LLMs and RLHF

Understand Transformers

How do you understand ChatGPT?



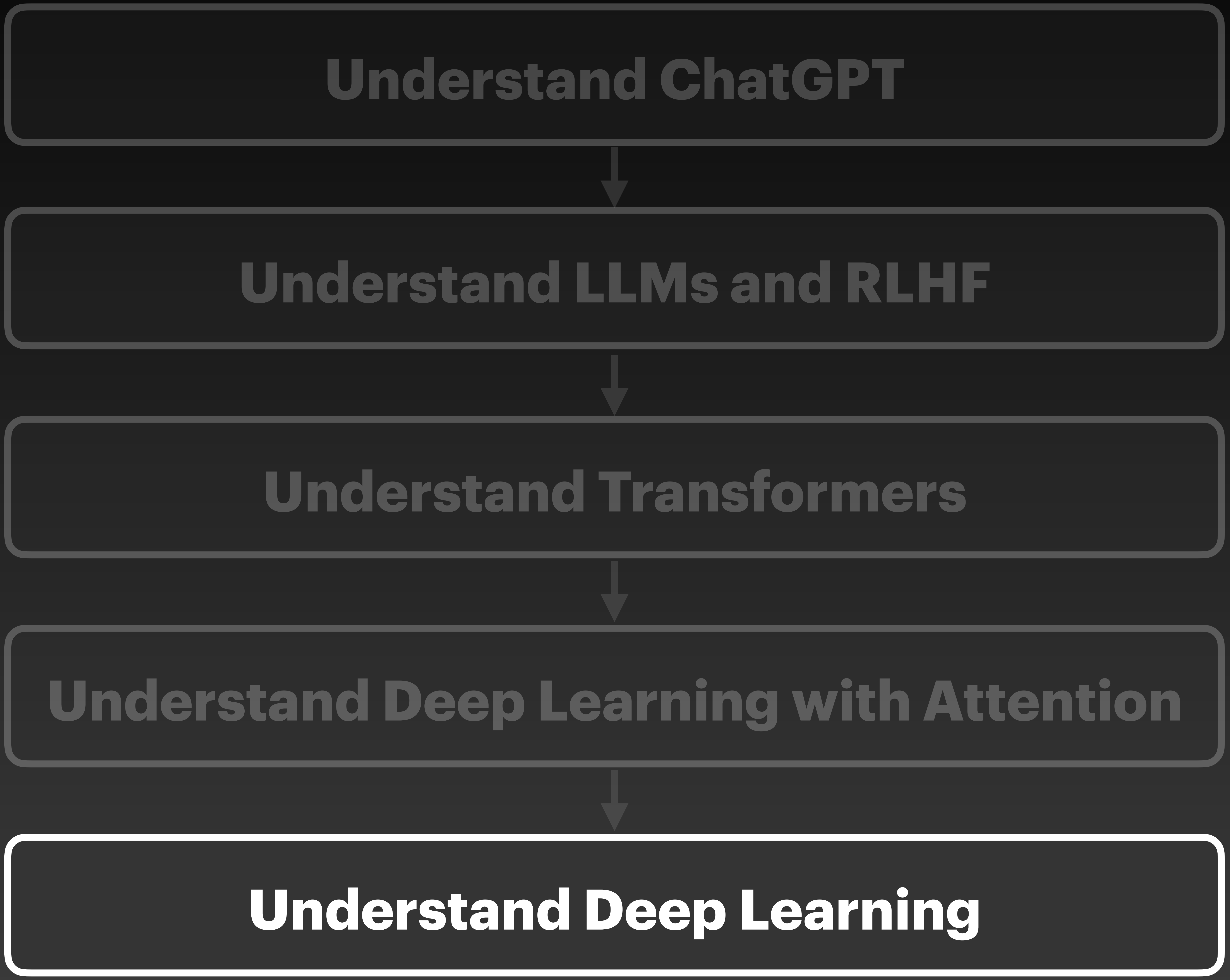
Understand ChatGPT

Understand LLMs and RLHF

Understand Transformers

Understand Deep Learning with Attention

How do you understand ChatGPT?



Understand ChatGPT

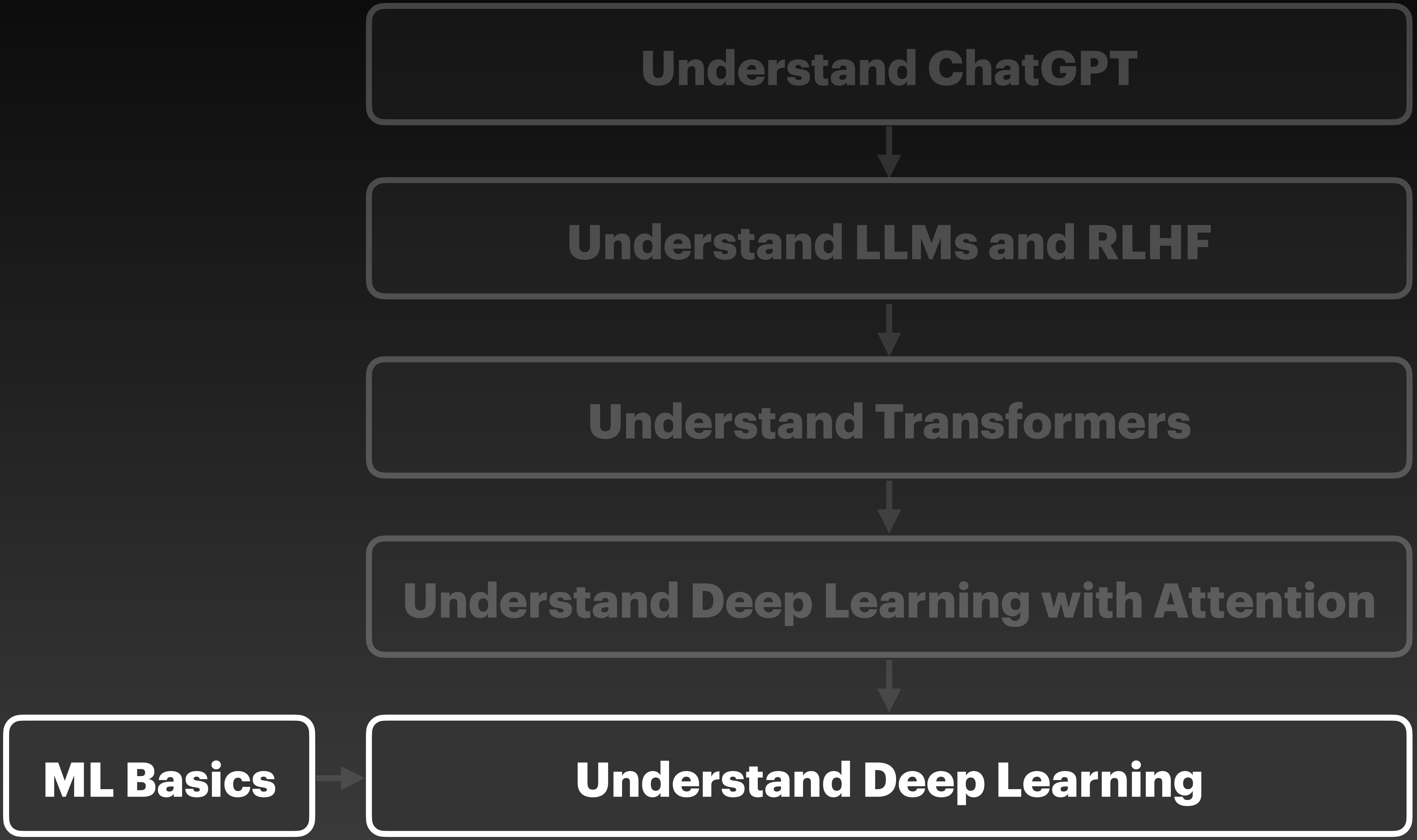
Understand LLMs and RLHF

Understand Transformers

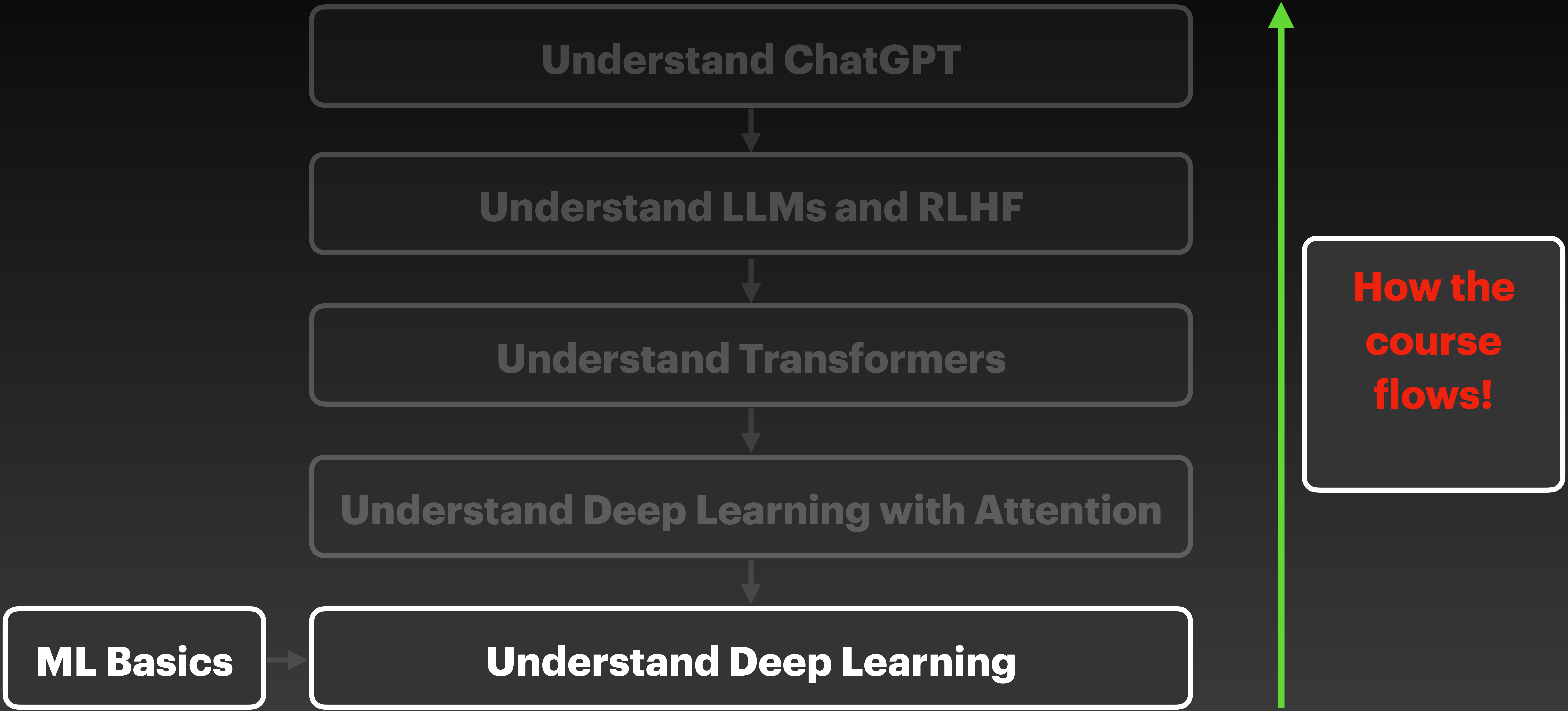
Understand Deep Learning with Attention

Understand Deep Learning

How do you understand ChatGPT?



How do you understand ChatGPT?



Understand ChatGPT

Understand LLMs and RLHF

Understand Transformers

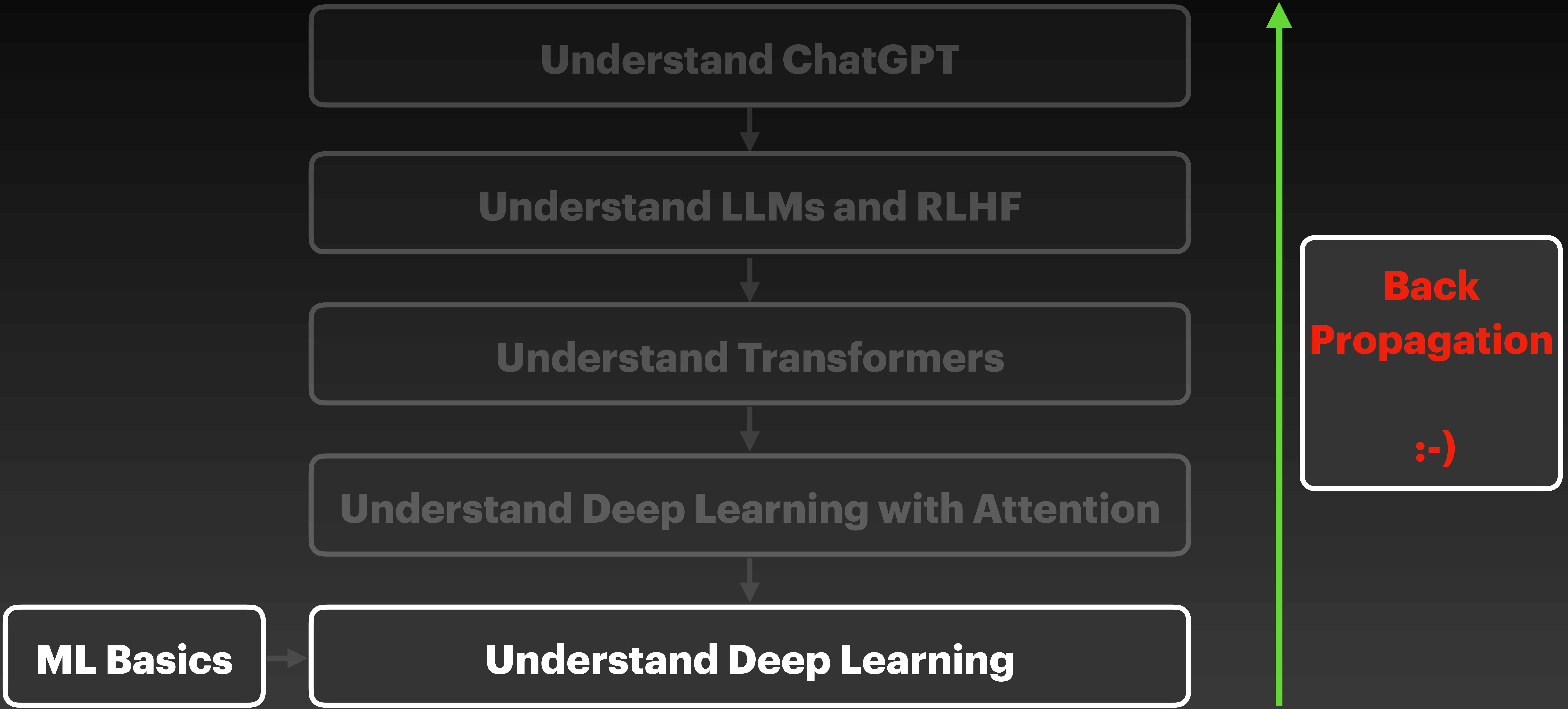
Understand Deep Learning with Attention

ML Basics

Understand Deep Learning

How the course flows!

How do you understand ChatGPT?



Understand ChatGPT

Understand LLMs and RLHF

Understand Transformers

Understand Deep Learning with Attention

ML Basics

Understand Deep Learning

Back Propagation

:-)

Course Outline

1. Building the foundations

- Logistics and Motivation
- ML fundamentals
- Logistic Regression
- Deep Learning

2. Transformers

- Transformers
- Discriminative and Generative
- Embeddings
- Applications
-

3. Generative AI

- LLMs
- GPT, GPT-2,GPT-3
- GPT 3.5, GPT 4
- Prompt Engineering
- Fine-tuning and Evaluating LLMs
- Open source vs closed LLMs

4. LVMs

- Auto Encoders
- Stable Diffusion
- Text to Image models
- Applications

Course Outline

1. Building the foundations

- Logistics and Motivation
- ML fundamentals
- Logistic Regression
- Deep Learning

2. Transformers

- Transformers
- Discriminative and Generative
- Embeddings
- Applications

3. Generative AI

- LLMs
- GPT, GPT-2,GPT-3
- GPT 3.5, GPT 4
- Prompt Engineering
- Fine-tuning and Evaluating LLMs
- Open source vs closed LLMs

4. LVMs

- Auto Encoders
- Stable Diffusion
- Text to Image models
- Applications

Course Outline

1. Building the foundations

- Logistics and Motivation
- ML fundamentals
- Logistic Regression
- Deep Learning

2. Transformers

- Transformers
- Discriminative and Generative
- Embeddings
- Applications

3. Generative AI

- LLMs
- GPT, GPT-2,GPT-3
- GPT 3.5, GPT 4
- Prompt Engineering
- Fine-tuning and Evaluating LLMs
- Open source vs closed LLMs

4. LVMs

- Auto Encoders
- Stable Diffusion
- Text to Image models
- Applications

Course Outline

1. Building the foundations

- Logistics and Motivation
- ML fundamentals
- Logistic Regression
- Deep Learning

2. Transformers

- Transformers
- Discriminative and Generative
- Embeddings
- Applications

3. Generative AI

- LLMs
- GPT, GPT-2,GPT-3
- GPT 3.5, GPT 4
- Prompt Engineering
- Fine-tuning and Evaluating LLMs
- Open source vs closed LLMs

4. LVMs

- Auto Encoders
- Stable Diffusion
- Text to Image models
- Applications

Course Webpage and Resources

https://bytesizeml.github.io/llm2024

(Almost) Every Class

First 60 Minutes

- Theory
- Code samples

Next 15 minutes

- In-Class Exercise

Next 35 minutes

- Theory
- In-class Coding Exercise

(Almost) Every Class

First 60 Minutes

- Theory
- Code samples

Next 15 minutes

- In-Class Exercise

Next 35 minutes

- Theory
- In-class Coding Exercise

(Almost) Every Class

First 60 Minutes

- Theory
- Code samples

Next 15 minutes

- In-Class Exercise

Next 35 minutes

- Theory
- In-class Coding Exercise

What I would like you to take away!

Conceptually

- Better understanding of LLMs
- Of LLM application areas
- Of APIs
- Intuition behind LLM models
- Theory behind LLMs

Implementation

- Coding up baselines in Colab
- Comfort with APIs
- Use of Hugging Face models
- Showcasing your work on webpage
- Fine-tuning LLM models

Ideas

- Where can you apply LLMs next?
- How can you chain LLMs
- to solve a problem?

What I would like you to take away!

Conceptually

- Better understanding of LLMs
- Of LLM application areas
- Of APIs
- Intuition behind LLM models

Implementation

- Coding up baselines in Colab
- Comfort with APIs
- Use of Hugging Face models
- Showcasing your work on webpage
- Fine-tuning LLM models

Ideas

- Where can you apply LLMs next?
- How can you chain LLMs
- to solve a problem?

What I would like you to take away!

Conceptually

- Better understanding of LLMs
- Of LLM application areas
- Of APIs
- Intuition behind LLM models

Implementation

- Coding up baselines in Colab
- Comfort with APIs
- Use of Hugging Face models
- Showcasing your work on webpage
- Fine-tuning LLM models

Ideas

- Where can you apply LLMs next?
- How can you chain LLMs
- to solve a problem?

Survey Results

What are you looking to learn/work on ?

Discuss in groups of 3 or 4

Assignments

1. Conceptual Assignment (20%)

- Typically once a week
- Tests your understanding of concepts
- Typically multiple choice questions
- Assigned on Thu, due next Sat
- Portion of this grade from
- In-class exercises

2. Coding Assignments (35%)

- Typically once a week
- Google colab based assignments
- Working with pytorch, LLM apis, etc
- Assigned on Thu, due next Sat

3. Mini-projects (30%)

- 2 or 3 for this class
- Get 2 weeks to work on it
- More involved than a coding assignment
- Could include a Kaggle Contest
- Could include a web demo

4. Project Presentation (15%)

- Present on one of the mini-projects
- Presentation on Tu or Th of finals week
- 7 minutes per team + 3 minute questions
- Methodology + working demo and results

Assignments

1. Conceptual Assignment (20%)

- Typically once a week
- Tests your understanding of concepts
- Typically multiple choice questions
- Assigned on Thu, due next Sat
- Portion of this grade from
- In-class exercises

2. Coding Assignments (35%)

- Typically once a week
- Google colab based assignments
- Working with pytorch, LLM apis, etc
- Assigned on Thu, due next Sat

3. Mini-projects (30%)

- 2 or 3 for this class
- Get 2 weeks to work on it
- More involved than a coding assignment
- Could include a Kaggle Contest
- Could include a web demo

4. Project Presentation (15%)

- Present on one of the mini-projects
- Presentation on Tu or Th of finals week
- 7 minutes per team + 3 minute questions
- Methodology + working demo and results

Assignments

1. Conceptual Assignment (20%)

- Typically once a week
- Tests your understanding of concepts
- Typically multiple choice questions
- Assigned on Thu, due next Sat
- Portion of this grade from
- In-class exercises

2. Coding Assignments (35%)

- Typically once a week
- Google colab based assignments
- Working with pytorch, LLM apis, etc
- Assigned on Thu, due next Sat

3. Mini-projects (30%)

- 2 or 3 for this class
- Get 2 weeks to work on it
- More involved than a coding assignment
- Could include a Kaggle Contest
- Could include a web demo

4. Project Presentation (15%)

- Present on one of the mini-projects
- Presentation on Tu or Th of finals week
- 7 minutes per team + 3 minute questions
- Methodology + working demo and results

Assignments

1. Conceptual Assignment (20%)

- Typically once a week
- Tests your understanding of concepts
- Typically multiple choice questions
- Assigned on Thu, due next Sat
- Portion of this grade from
- In-class exercises

2. Coding Assignments (35%)

- Typically once a week
- Google colab based assignments
- Working with pytorch, LLM apis, etc
- Assigned on Thu, due next Sat

3. Mini-projects (30%)

- 2 or 3 for this class
- Get 2 weeks to work on it
- More involved than a coding assignment
- Could include a Kaggle Contest
- Could include a web demo

4. Project Presentation (15%)

- Present on one of the mini-projects
- Presentation on Tu or Th of finals week
- 7 minutes per team + 3 minute questions
- Methodology + working demo and results

ChatGPT and LLMs are everywhere!

Engine vs API

Engines are different from APIs and we shouldn’t confuse the two.

Engine vs API

Engines are different from APIs and we shouldn’t confuse the two.

BERT and Llama are Engines/Foundation Models whereas ChatGPT 3.5 is an API

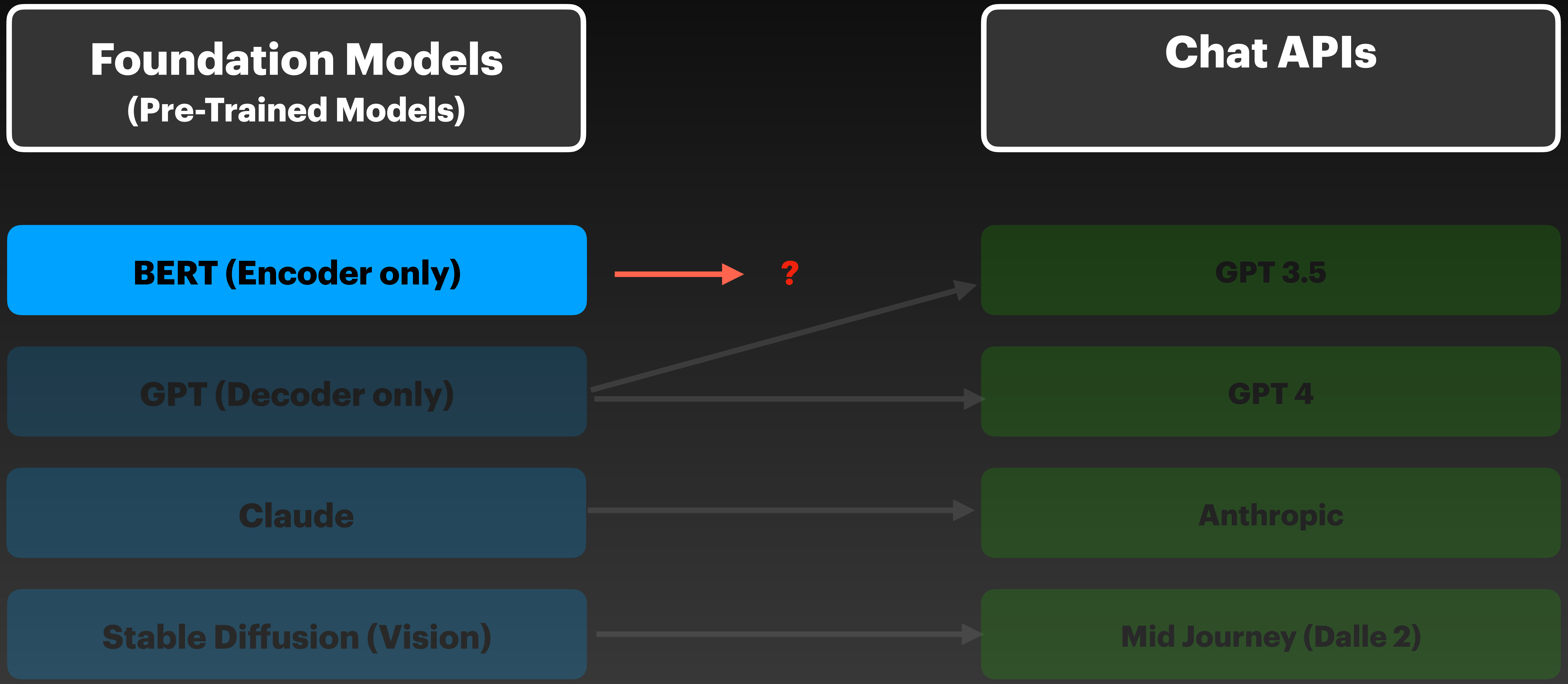
Engine vs API

Foundation Models
(Pre-Trained Models)

Chat APIs

- **BERT (Encoder only)**
- **GPT (Decoder only)**
- **Claude**
- **Stable Diffusion (Vision)**

Engine vs API



Foundation Models
(Pre-Trained Models)

Chat APIs

BERT (Encoder only)

?

GPT (Decoder only)

Claude

Stable Diffusion (Vision)

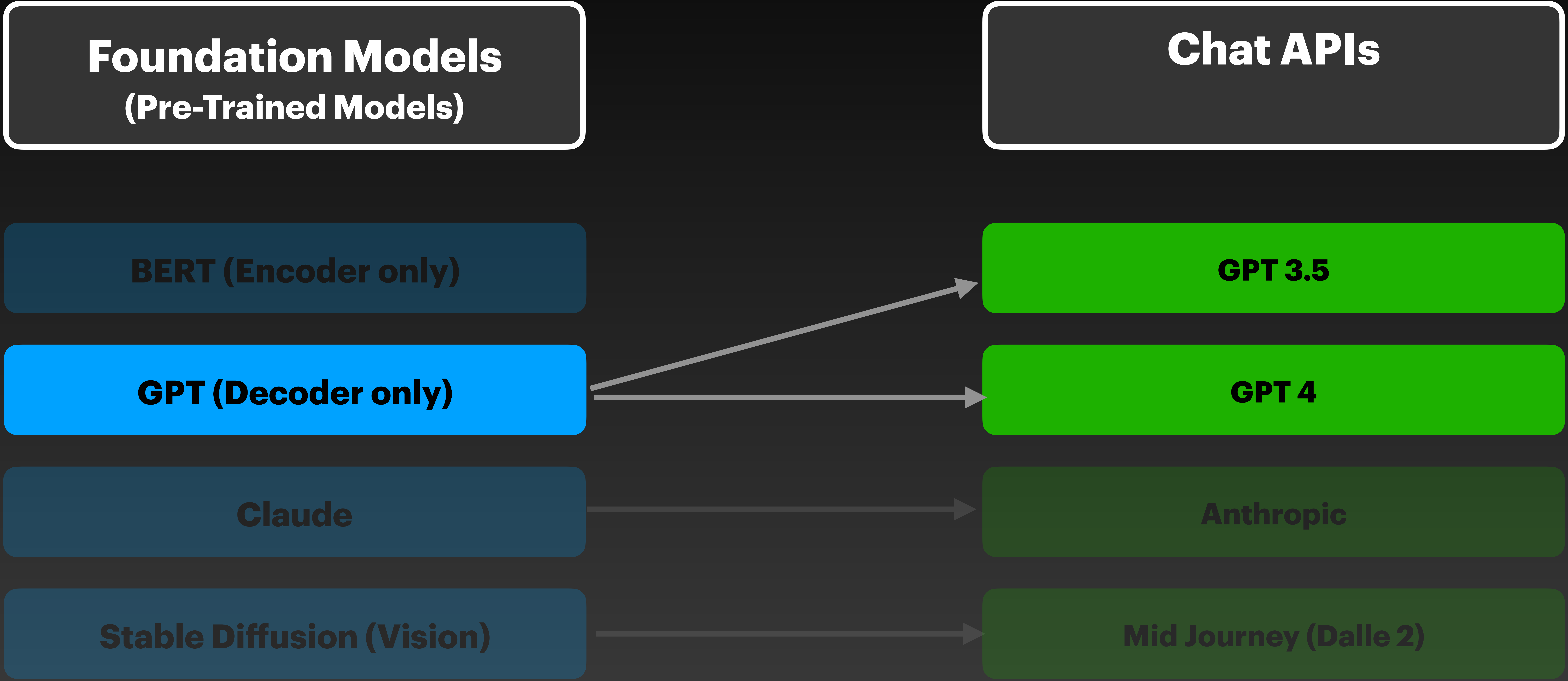
GPT 3.5

GPT 4

Anthropic

Mid Journey (Dalle 2)

Engine vs API



Foundation Models
(Pre-Trained Models)

Chat APIs

BERT (Encoder only)

GPT (Decoder only)

Claude

Stable Diffusion (Vision)

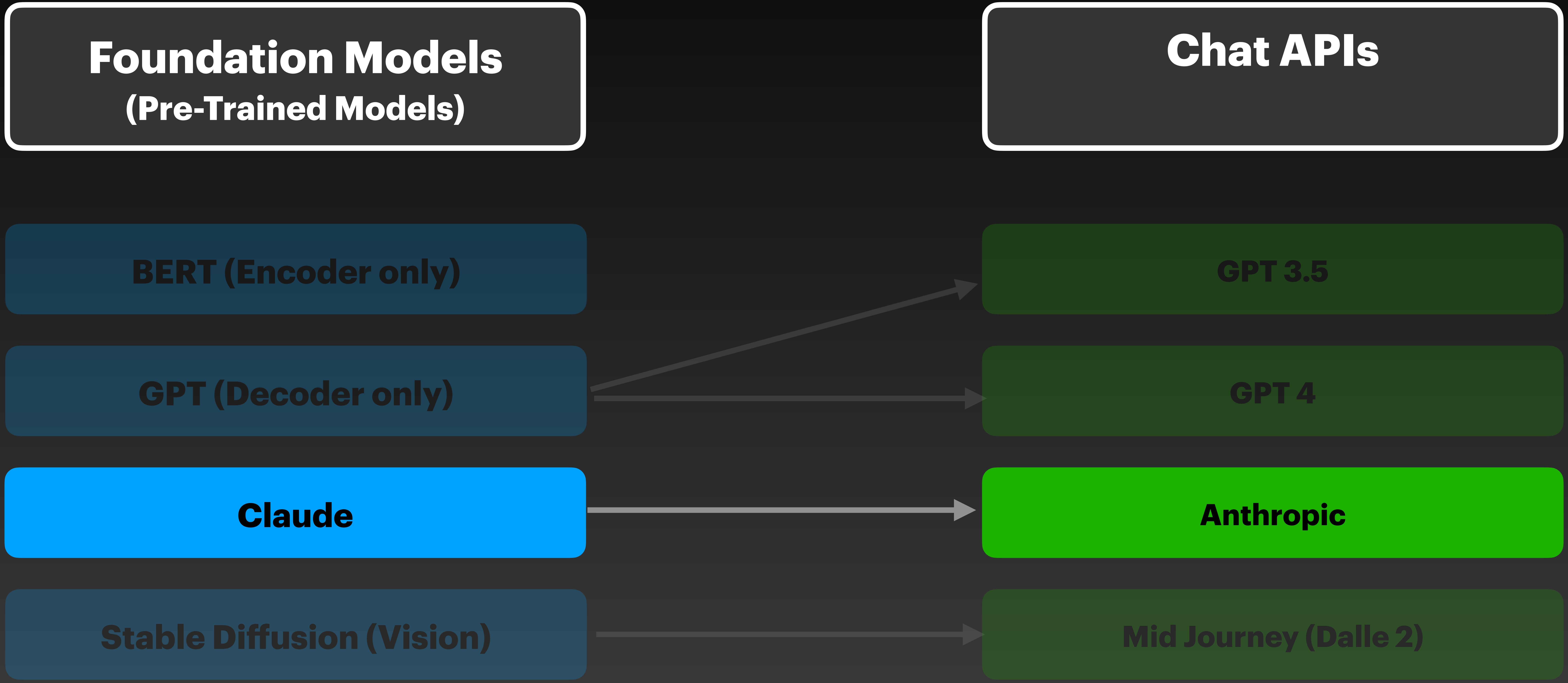
GPT 3.5

GPT 4

Anthropic

Mid Journey (Dalle 2)

Engine vs API



Foundation Models
(Pre-Trained Models)

Chat APIs

BERT (Encoder only)

GPT (Decoder only)

Claude

Stable Diffusion (Vision)

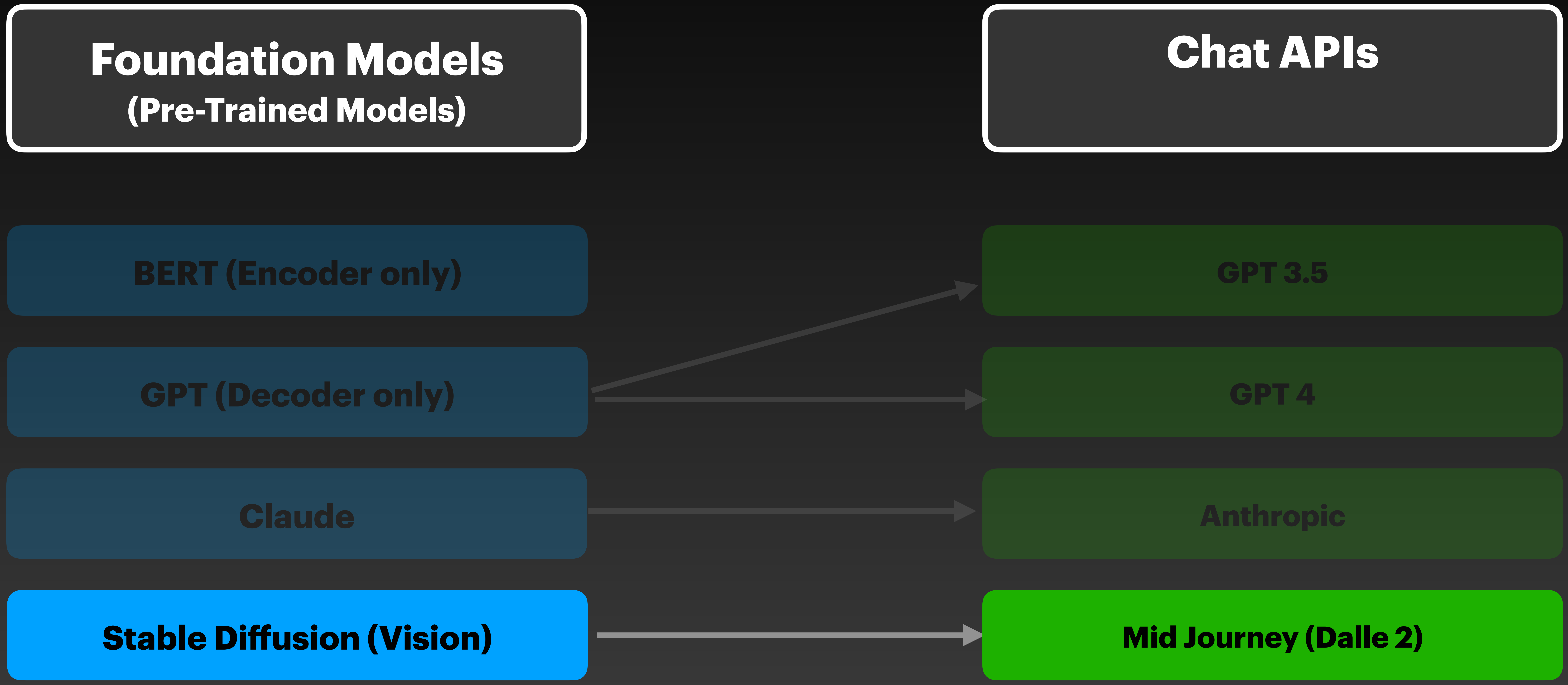
GPT 3.5

GPT 4

Anthropic

Mid Journey (Dalle 2)

Engine vs API



Foundation Models

(Pre-Trained Models)

Chat APIs

BERT (Encoder only)

GPT (Decoder only)

Claude

Stable Diffusion (Vision)

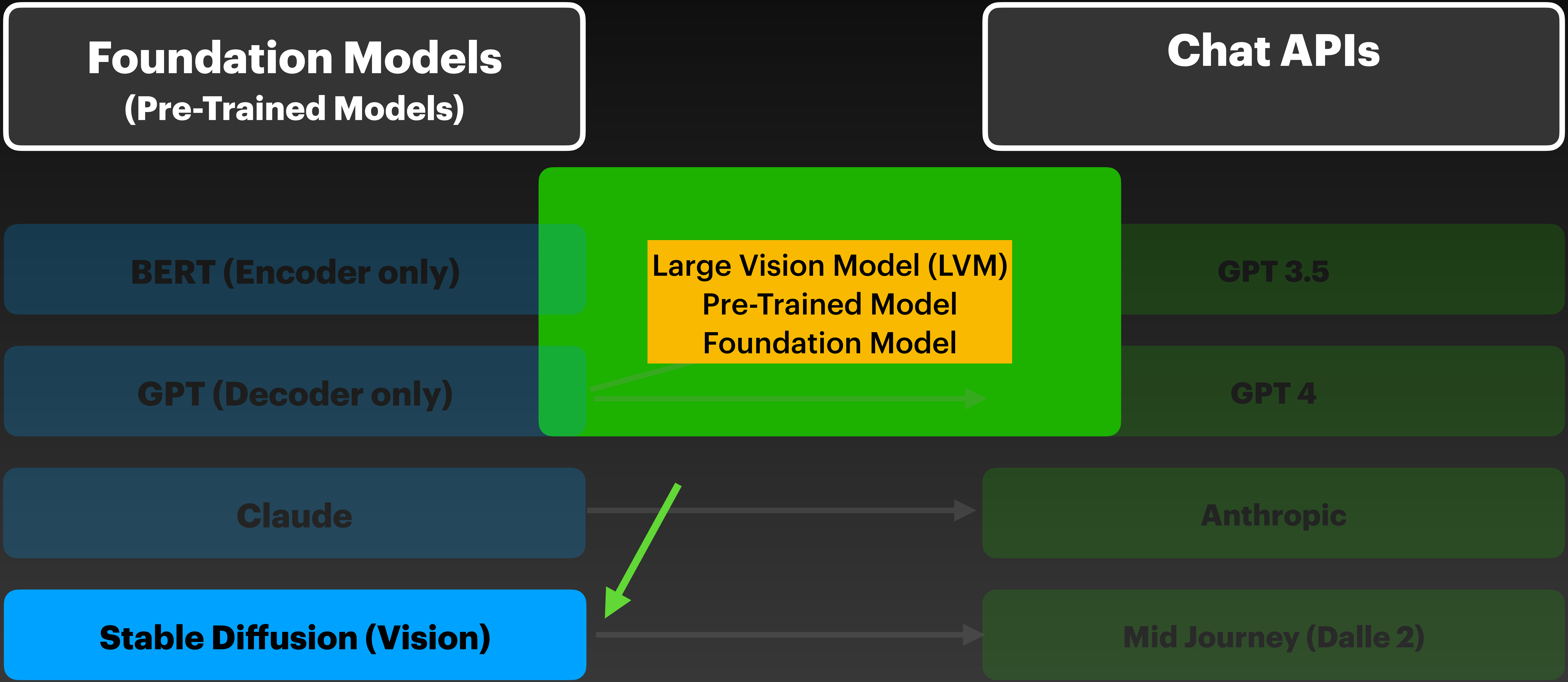
GPT 3.5

GPT 4

Anthropic

Mid Journey (Dalle 2)

Engine vs API



Foundation Models
(Pre-Trained Models)

Chat APIs

Large Vision Model (LVM)
Pre-Trained Model
Foundation Model

BERT (Encoder only)

GPT 3.5

GPT (Decoder only)

GPT 4

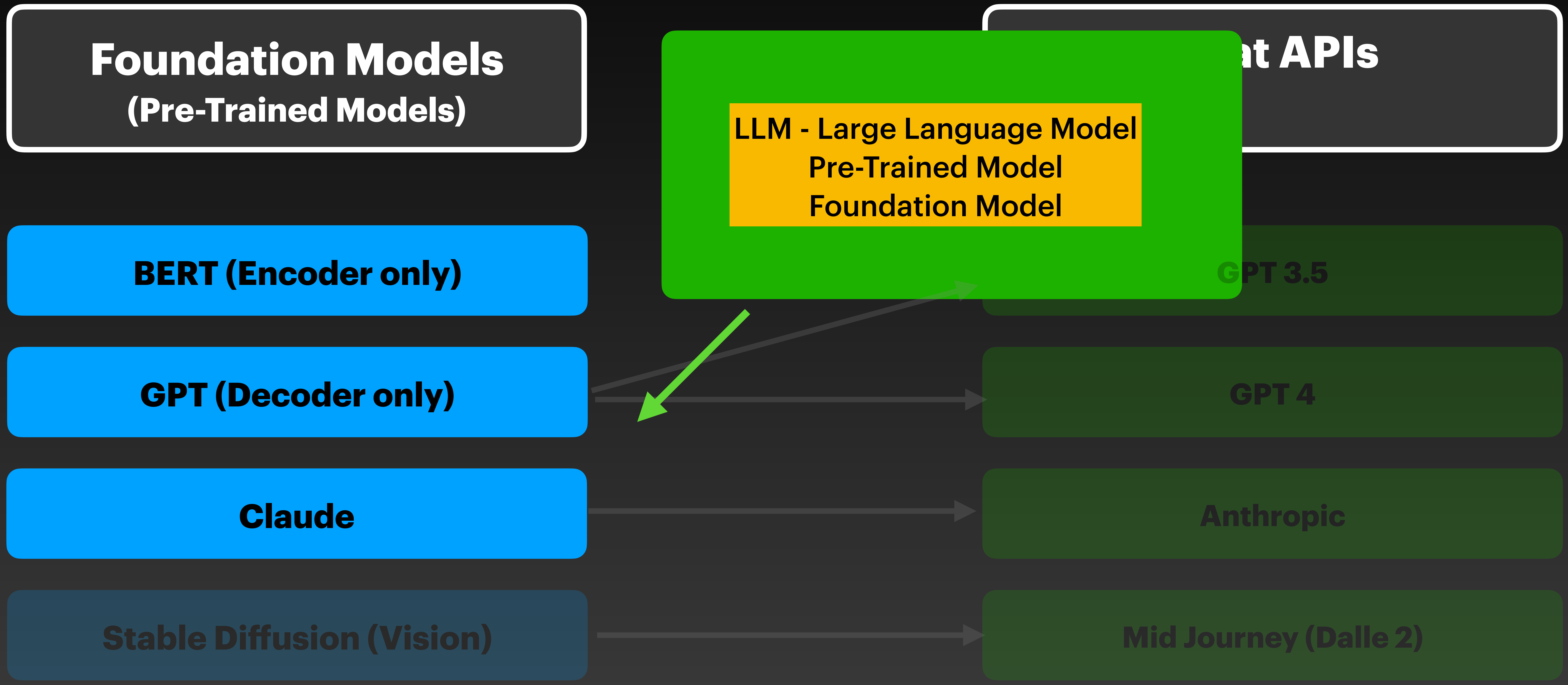
Claude

Anthropic

Stable Diffusion (Vision)

Mid Journey (Dalle 2)

Engine vs API



Foundation Models
(Pre-Trained Models)

at APIs

LLM - Large Language Model
Pre-Trained Model
Foundation Model

BERT (Encoder only)

GPT (Decoder only)

Claude

Stable Diffusion (Vision)

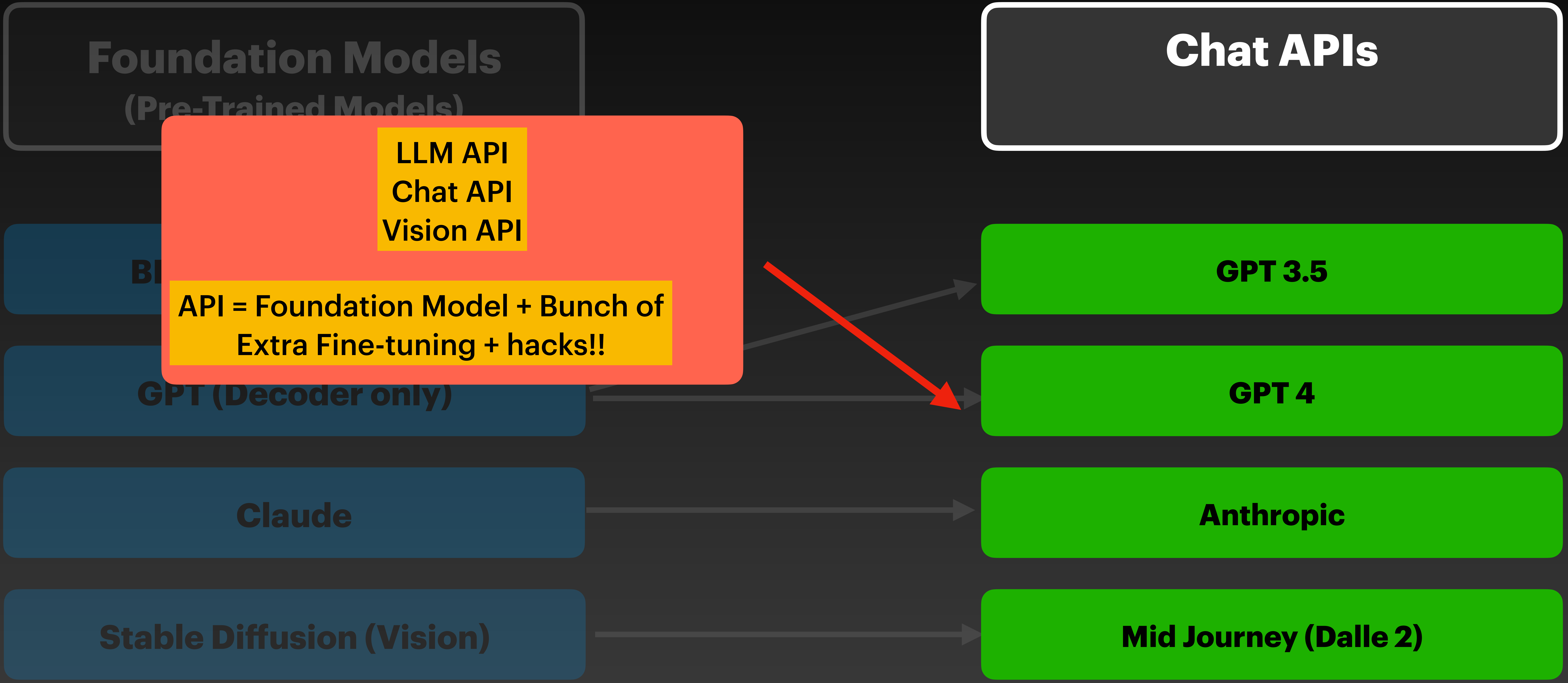
GPT 3.5

GPT 4

Anthropic

Mid Journey (Dalle 2)

Engine vs API



Foundation Models
(Pre-Trained Models)

BI

GPT (Decoder only)

Claude

Stable Diffusion (Vision)

Chat APIs

GPT 3.5

GPT 4

Anthropic

Mid Journey (Dalle 2)

LLM API
Chat API
Vision API

API = Foundation Model + Bunch of Extra Fine-tuning + hacks!!

What is a Language Model?

Scientific **Data-driven Model** that helps machines understand language and patterns in sentence construction

What is a Language Model?

Example: I just got promoted. I am feeling so ———

What is a Language Model?

Example: I just got promoted. I am feeling so happy

What is a Language Model?

Example: I just checked my application status and it got ———. It's frustrating!

What is a Language Model?

Example: I just checked my application status and it got rejected. It’s frustrating!

What is a Large Language Model (LLM)?

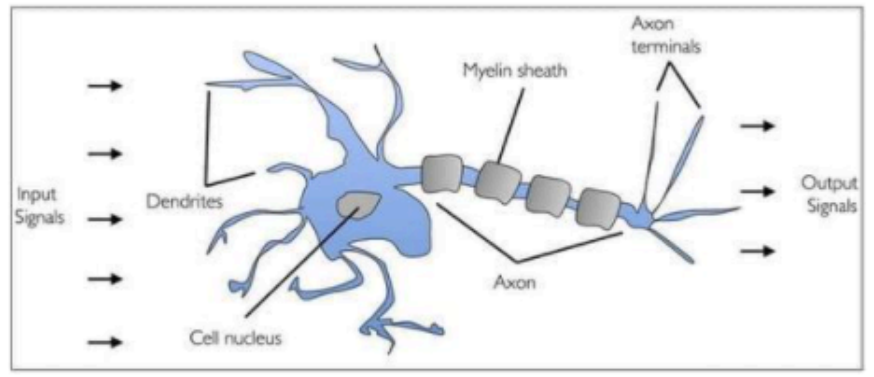
LLMs are language models that are learned from massive corpuses of text, that are mined from the web. They are known to be sophisticated in understanding language and can be **generative** in nature.

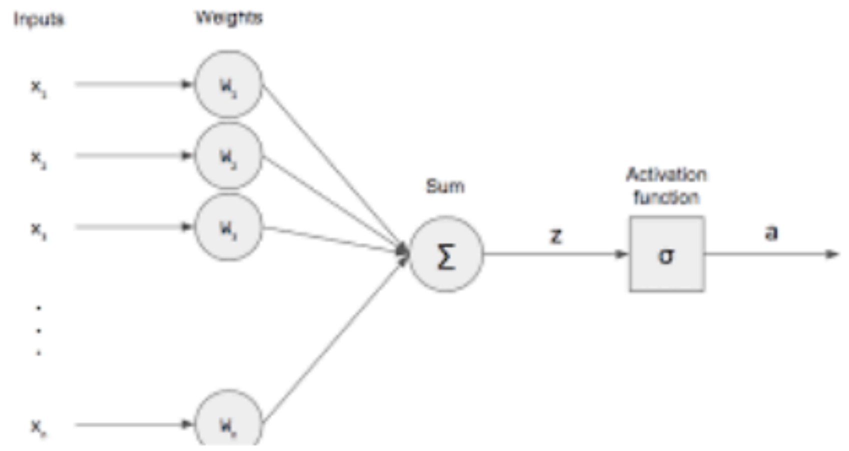
History of (Large) Language Models

How did machines work with language before and how we do it now?

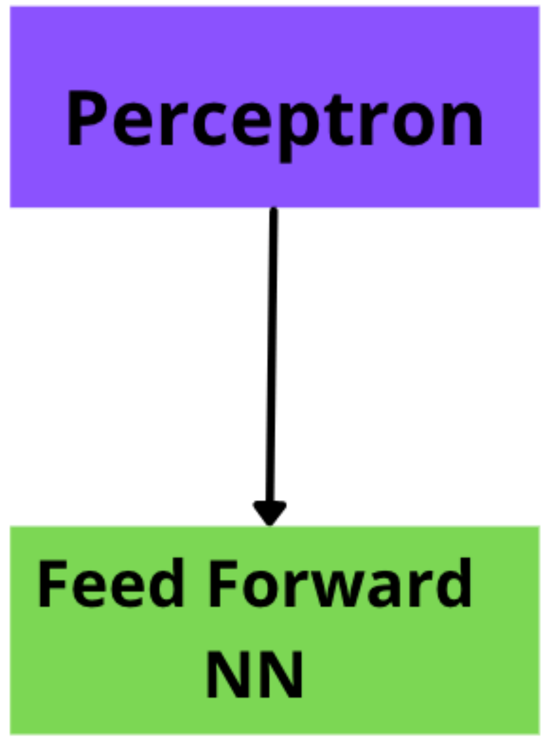
History of (Large) Language Models

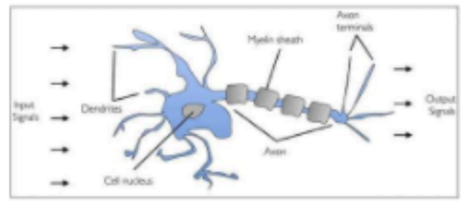
Perceptron

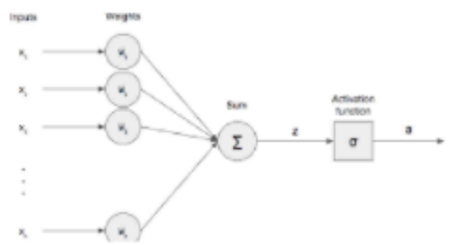


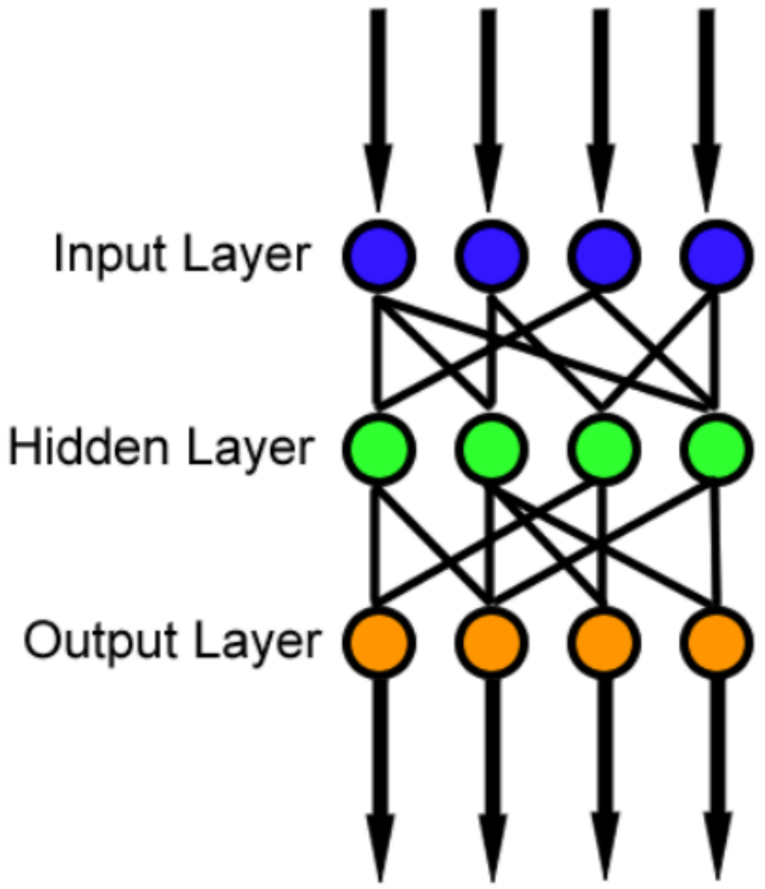


History of (Large) Language Models

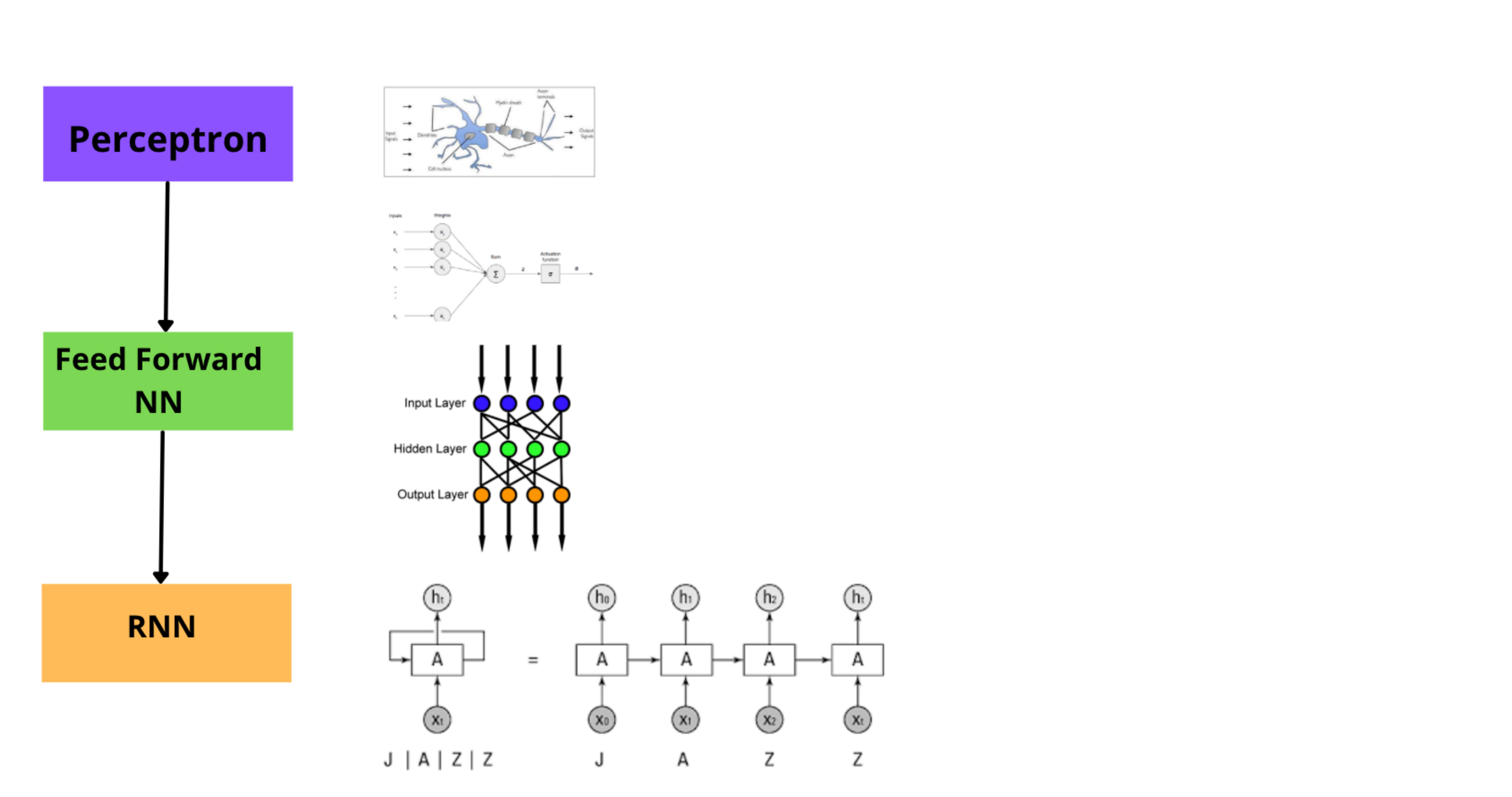








History of (Large) Language Models



Perceptron

Feed Forward NN

RNN

History of (Large) Language Models

RNN Issue:

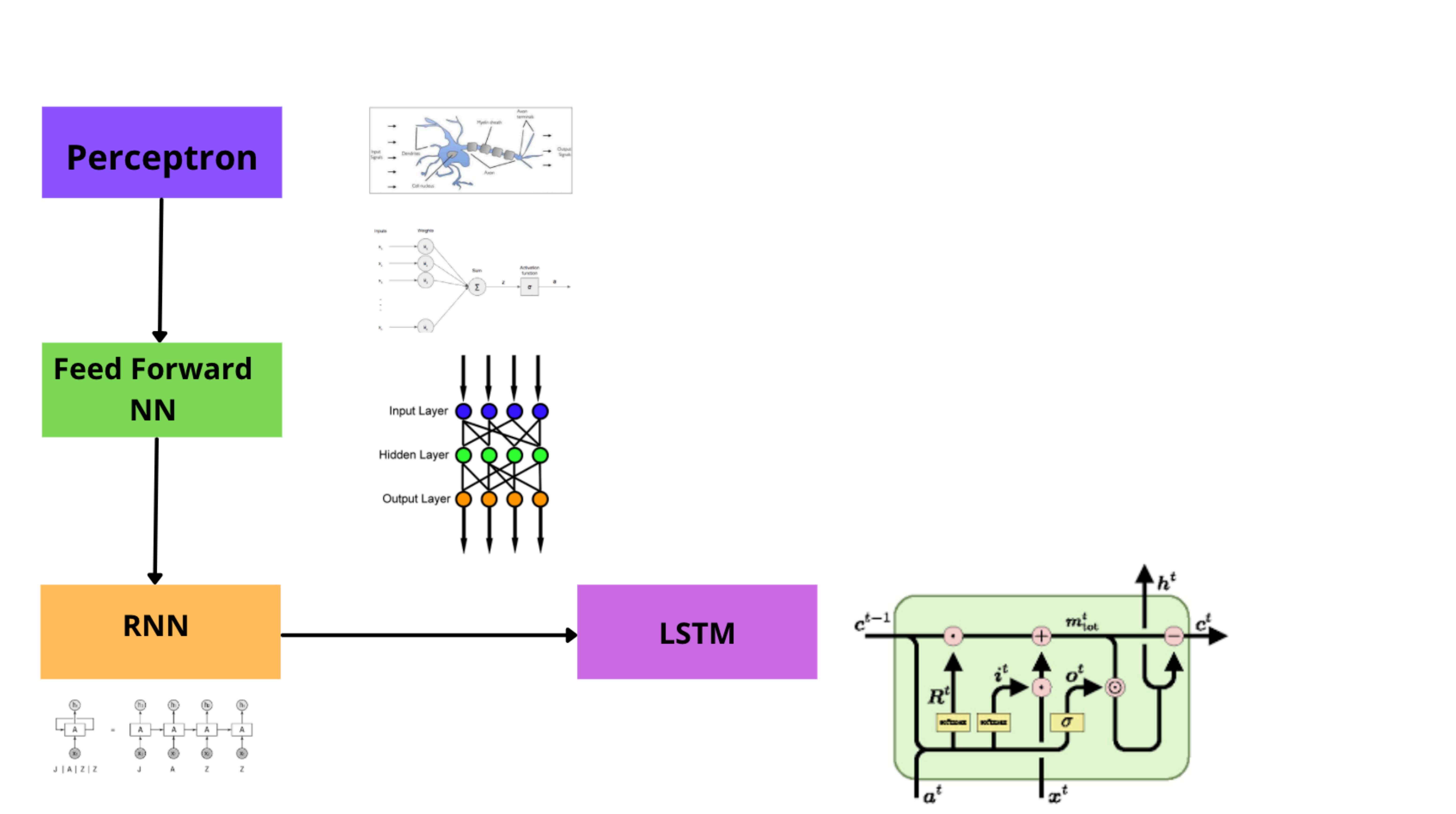
I just arrived in NY. In a few days, I would like to visit the city, ———

History of (Large) Language Models

RNN Issue:

I just arrived in NY. In a few days, I would like to visit the city, NY

History of (Large) Language Models



Perceptron

Feed Forward NN

RNN

LSTM

History of (Large) Language Models

LSTM

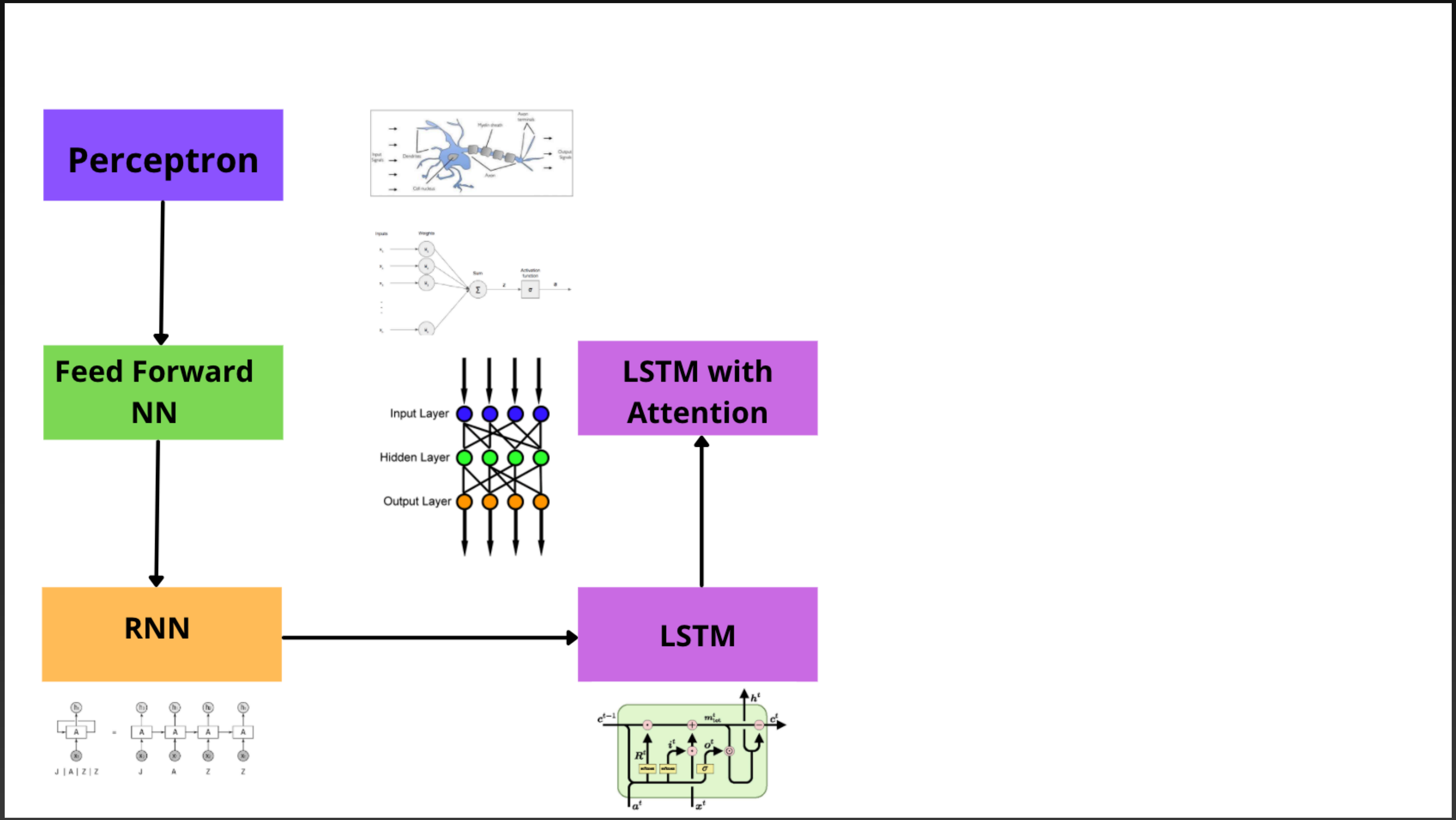
I just arrived in NY. In a few days, I would like to visit the city, ———

History of (Large) Language Models

LSTM

I just arrived in NY. In a few days, I would like to visit the city, Seattle

History of (Large) Language Models



Perceptron

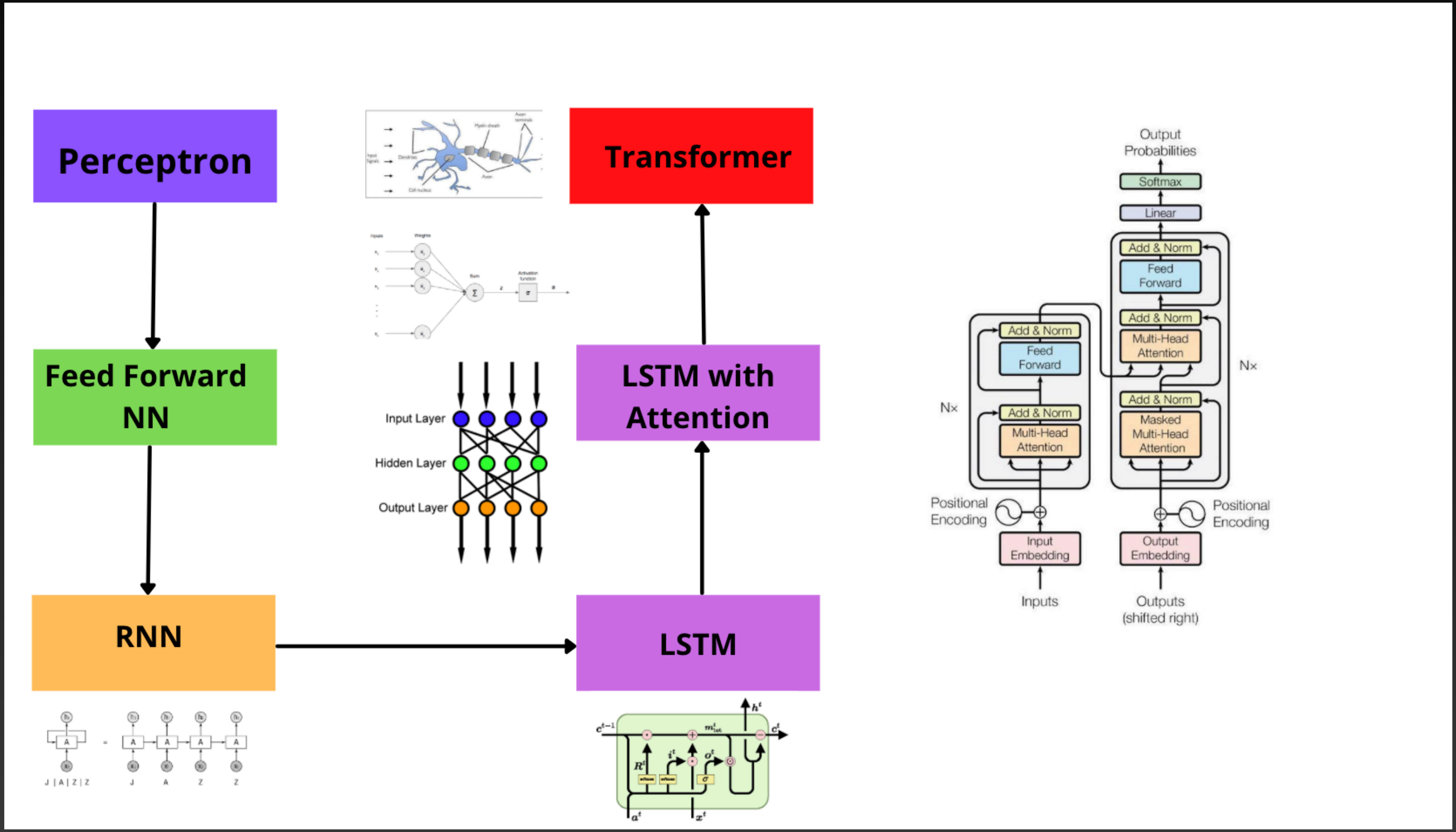
Feed Forward NN

RNN

LSTM

LSTM with Attention

History of (Large) Language Models



Perceptron

Feed Forward NN

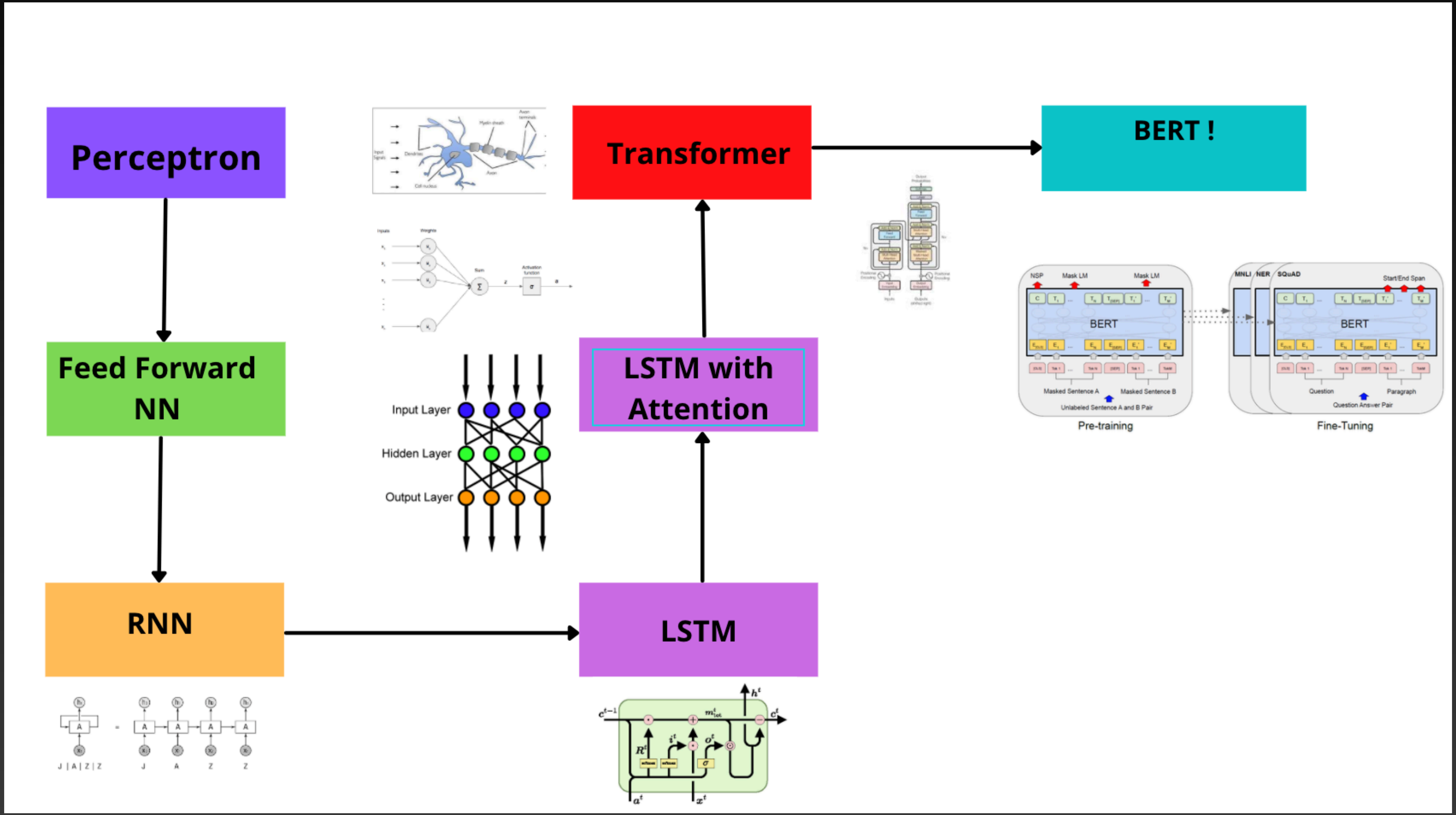
RNN

LSTM

LSTM with Attention

Transformer

History of (Large) Language Models



Perceptron

Feed Forward NN

RNN

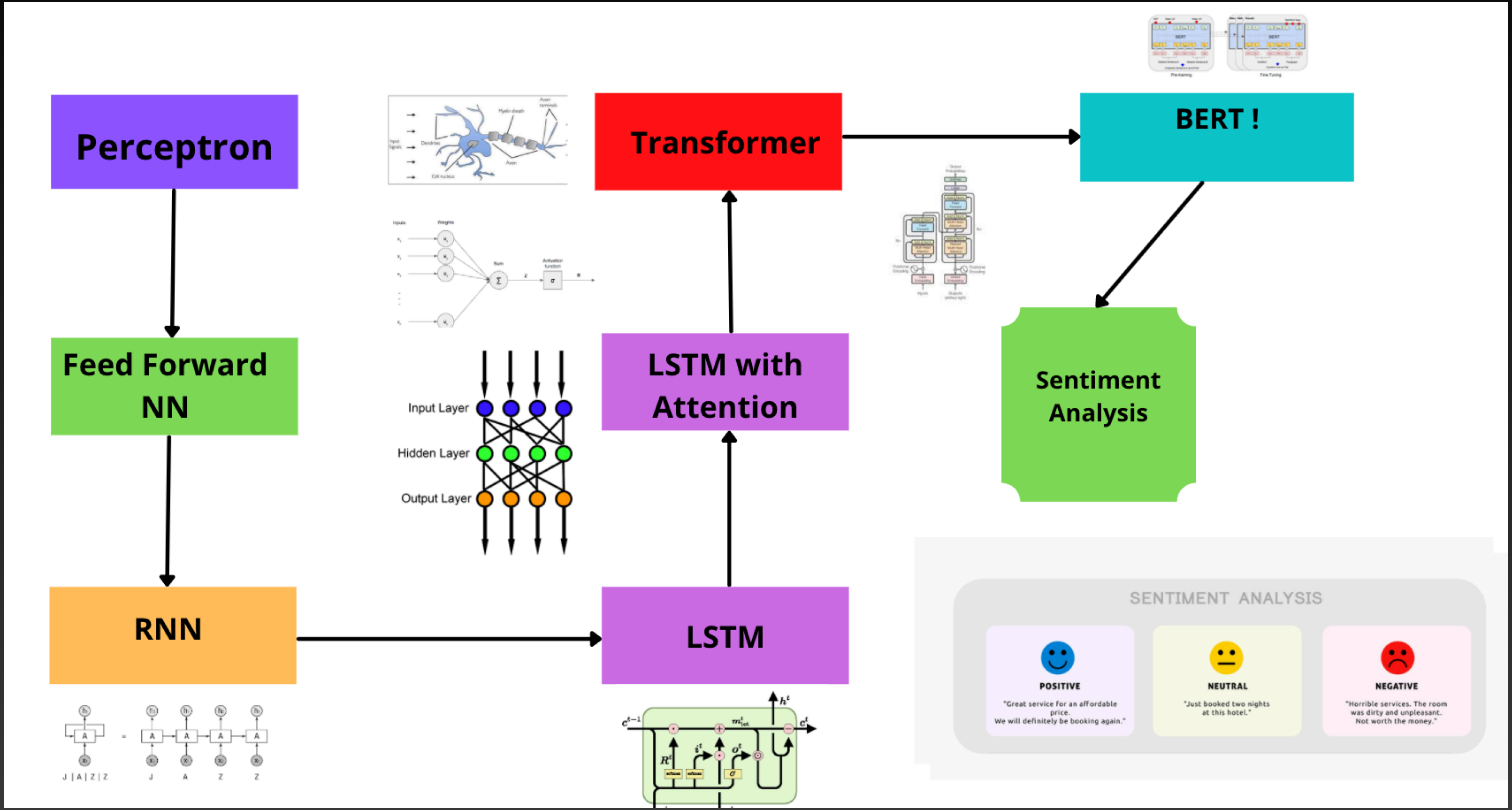
LSTM

LSTM with Attention

Transformer

BERT !

History of (Large) Language Models



Perceptron

Feed Forward NN

RNN

LSTM

LSTM with Attention

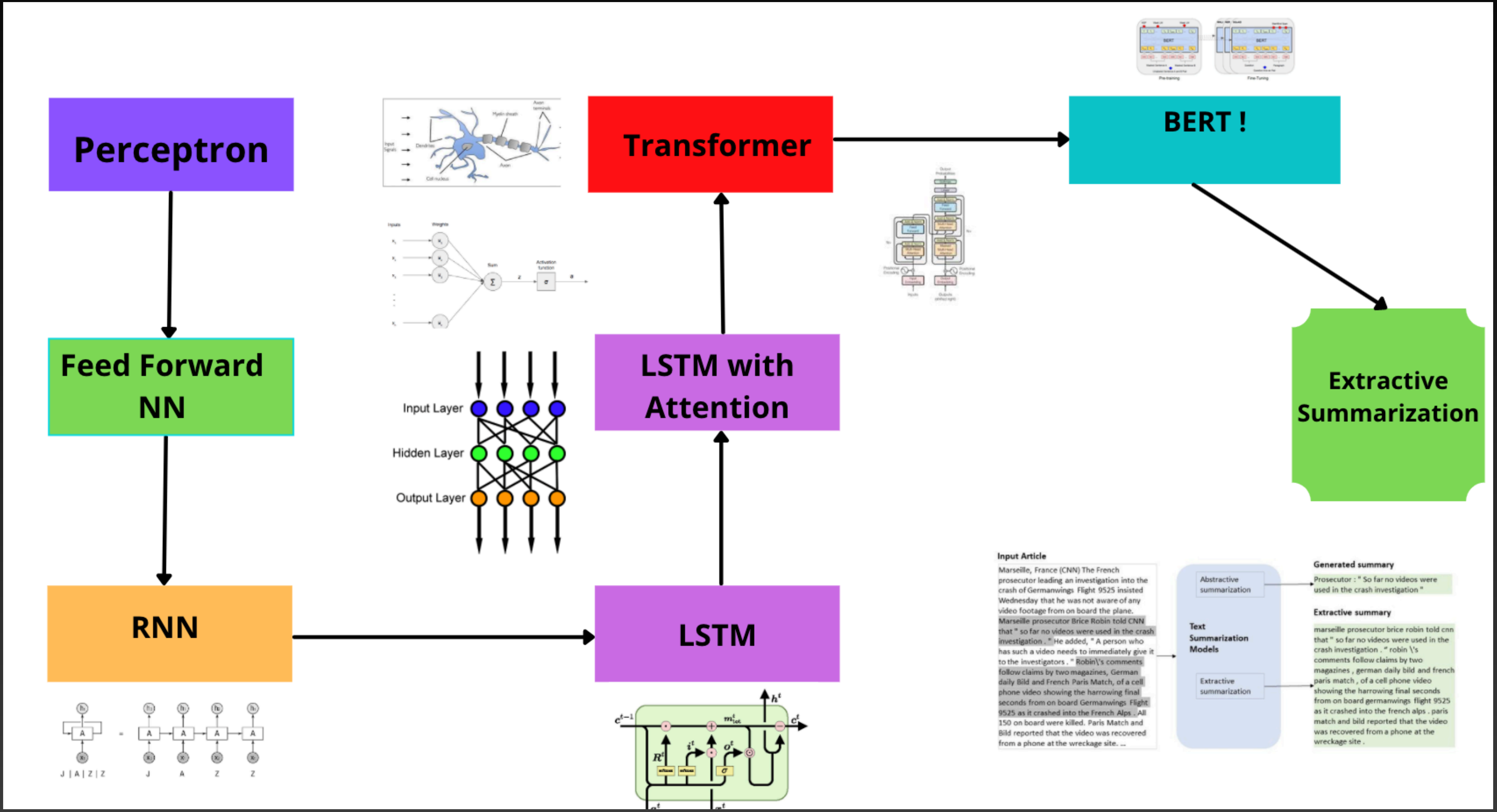
Transformer

BERT !

Sentiment Analysis

Input Layer
Hidden Layer
Output Layer
SENTIMENT ANALYSIS
POSITIVE
"Great service for an affordable price.
We will definitely be booking again."
NEUTRAL
"Just booked two nights at this hotel."
NEGATIVE
"Horrible services. The room was dirty and unpleasant.
Not worth the money."

History of (Large) Language Models



Perceptron

Feed Forward NN

RNN

LSTM

LSTM with Attention

Transformer

BERT !

Extractive Summarization

Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators . " Robin\'s comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...

Text Summarization Models

Abstractive summarization

Extractive summarization

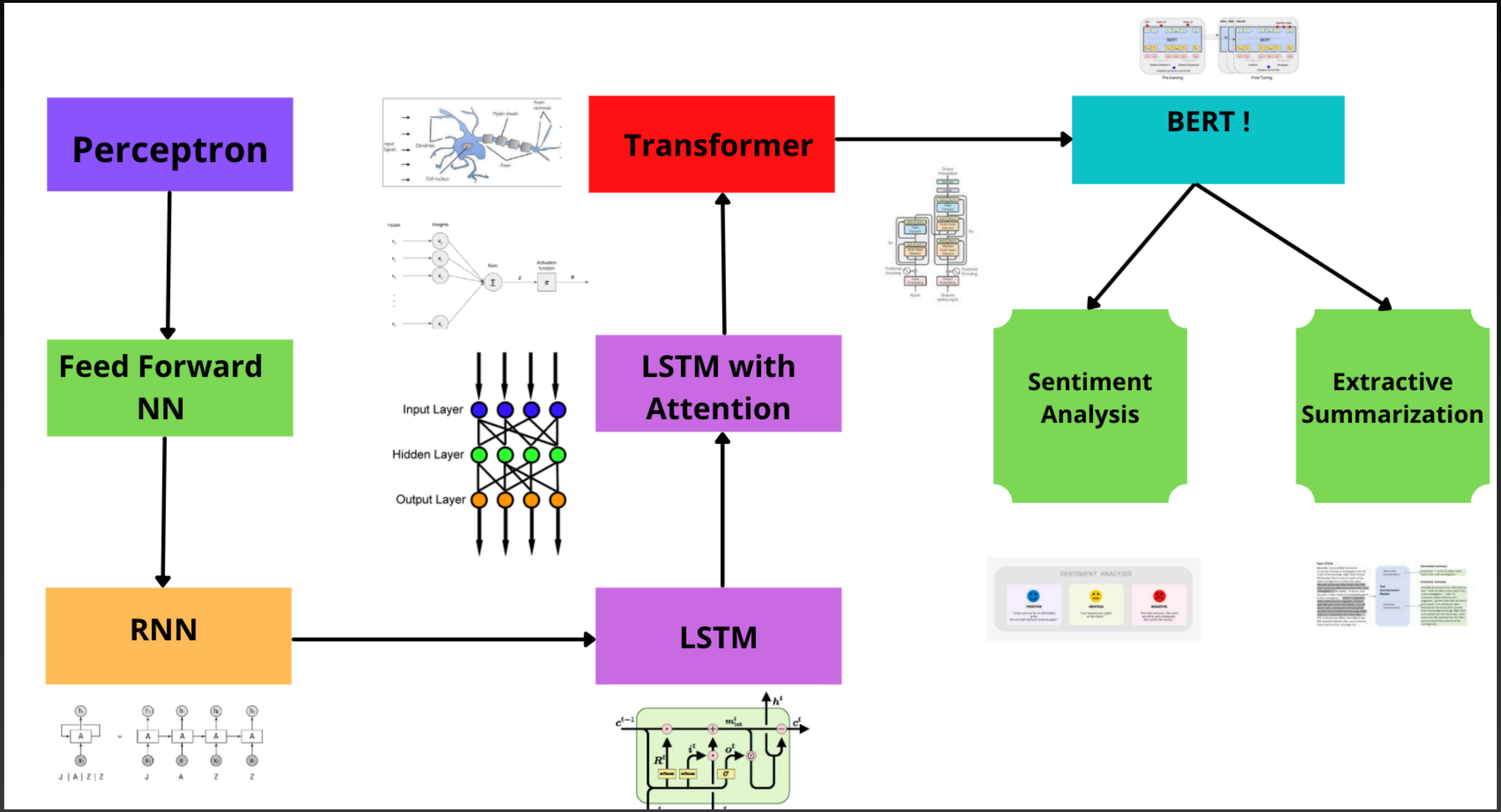
Generated summary

Prosecutor : " So far no videos were used in the crash investigation "

Extractive summary

marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . " robin \'s comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

History of (Large) Language Models



Perceptron

Feed Forward NN

RNN

LSTM

LSTM with Attention

Transformer

BERT !

Sentiment Analysis

Extractive Summarization

History of (Large) Language Models

GPT vs BERT

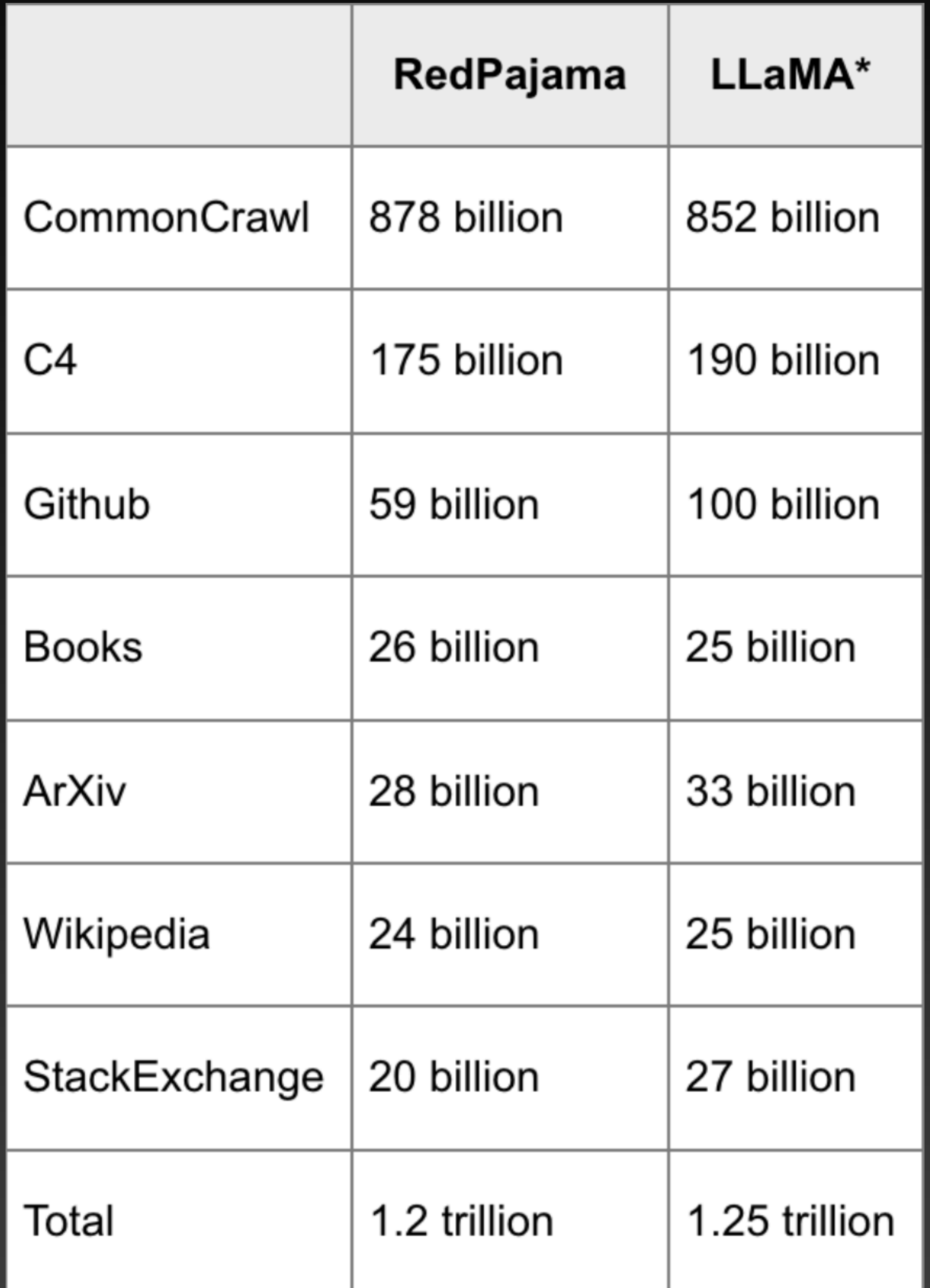
While BERT is purely about encoding and is called an encoding Transformer. GPT is purely a decoder and is called a decoding transformer.

History of (Large) Language Models

GPT-x

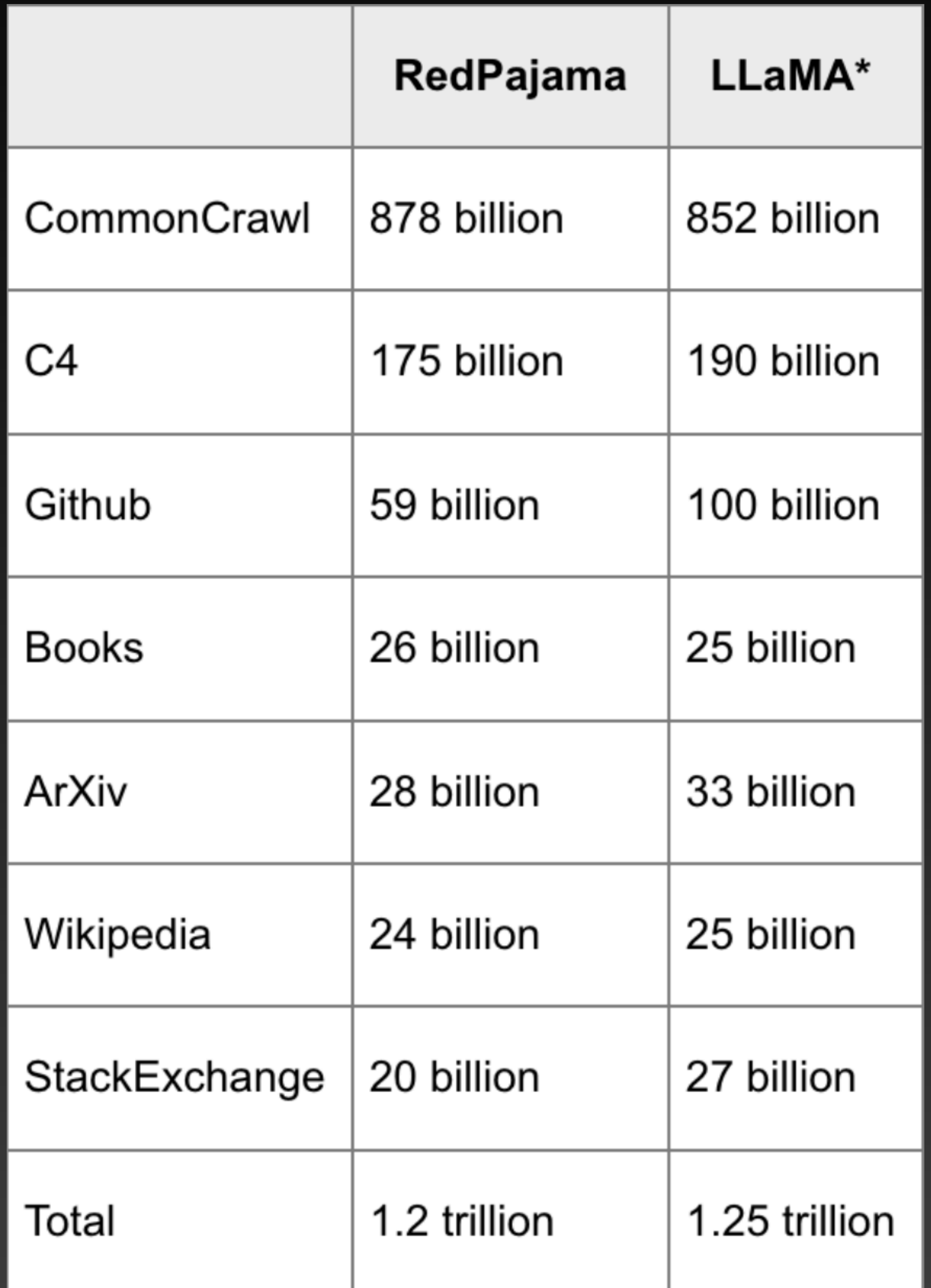
GPT-x (GPT, GPT-2, GPT-2.5, etc) are decoding transformers that are trained to predict the next token given the past and do a very good job at it! That’s how they can generate entire paragraphs that look logical, grammatical and structured.

1 Trillion Tokens!



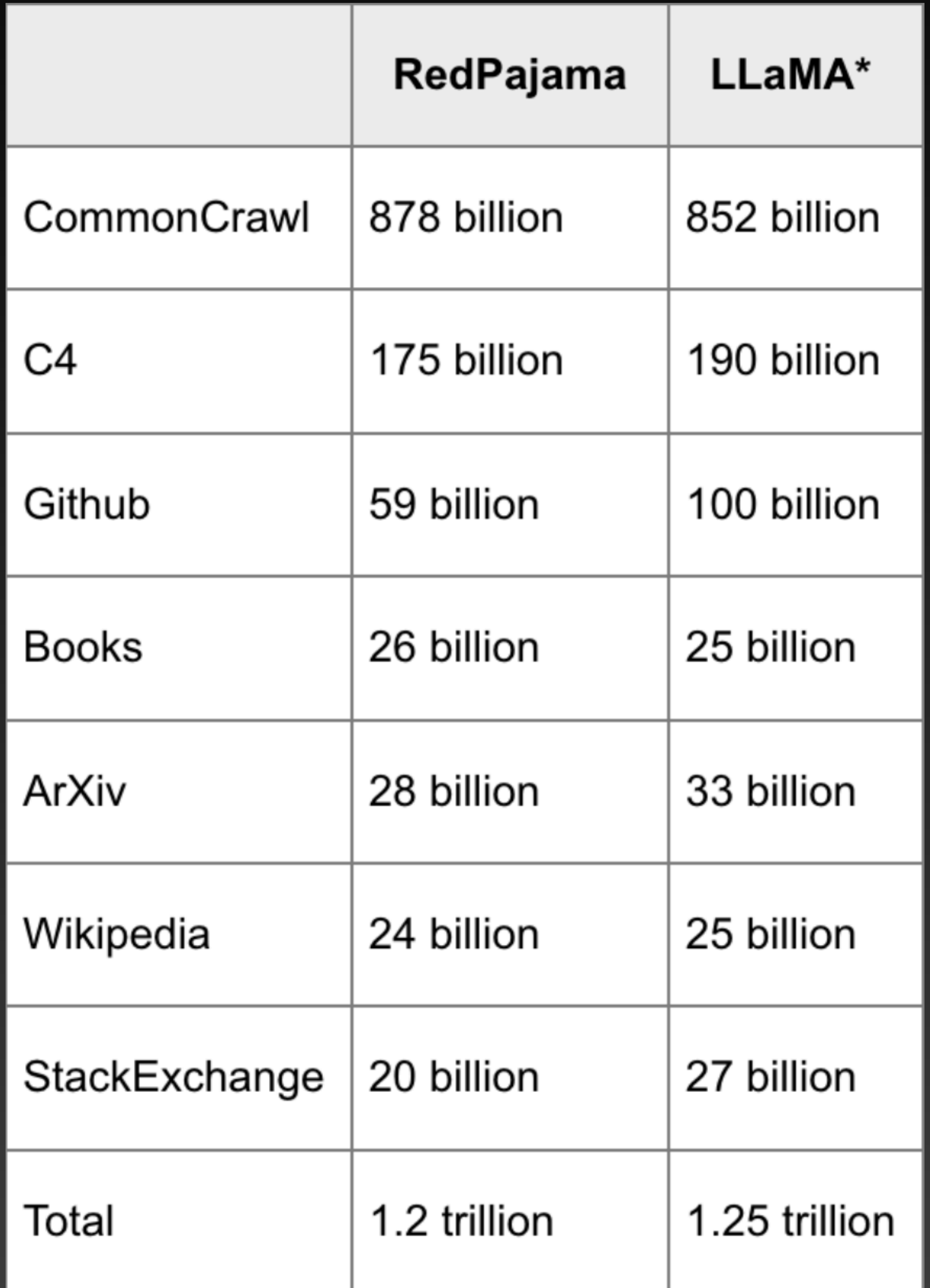
	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

1 Trillion Tokens requires how many books?



	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

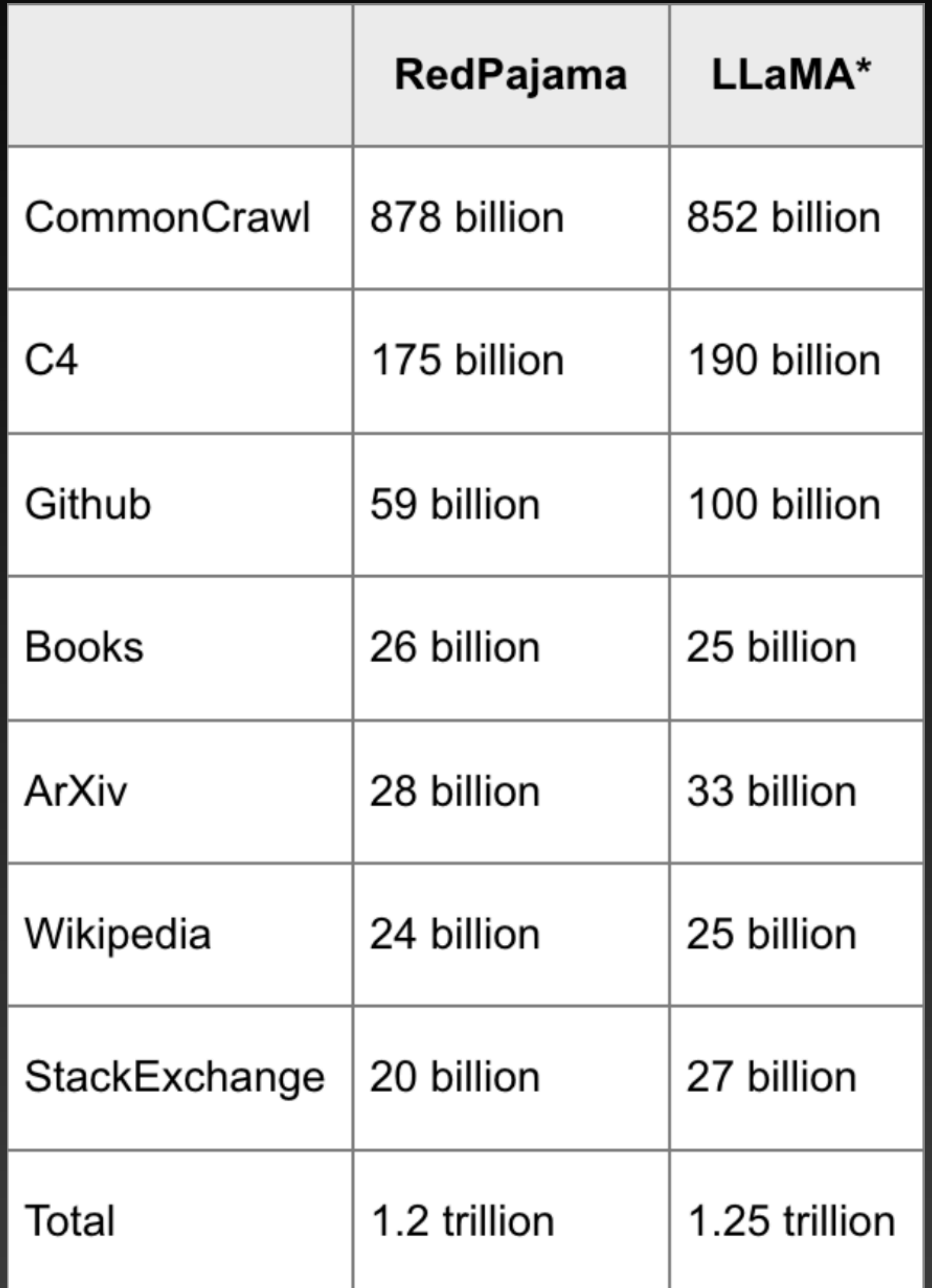
1 Trillion Tokens requires how many books?



	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

1 Book ~ 50k Tokens

1 Trillion Tokens requires how many books?



	RedPajama	LLaMA*
CommonCrawl	878 billion	852 billion
C4	175 billion	190 billion
Github	59 billion	100 billion
Books	26 billion	25 billion
ArXiv	28 billion	33 billion
Wikipedia	24 billion	25 billion
StackExchange	20 billion	27 billion
Total	1.2 trillion	1.25 trillion

1 Book ~ 50k Tokens

15 Million Books ~ 1 Trillion Tokens

ChatGPT use cases for NLP

ChatGPT use cases for NLP

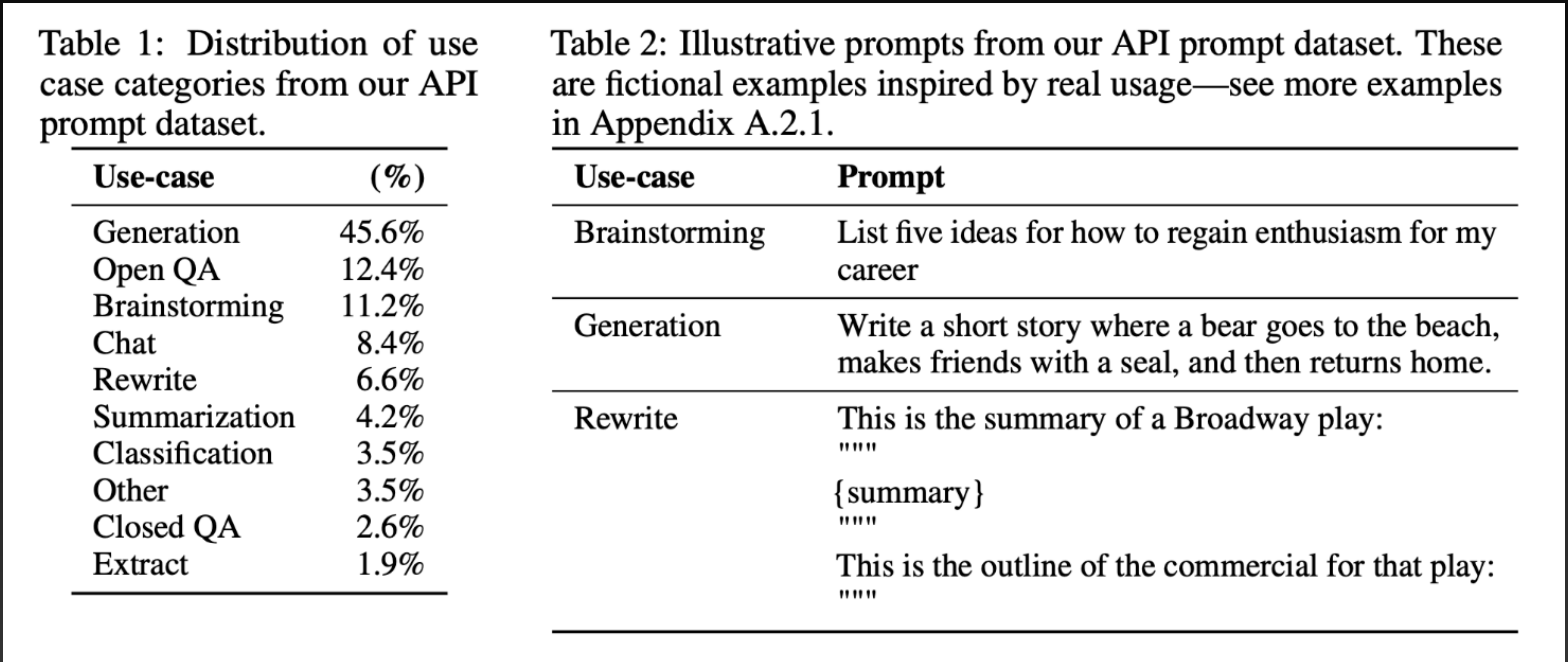


Table 1: Distribution of use case categories from our API prompt dataset.

Use-case	(%)
Generation	45.6%
Open QA	12.4%
Brainstorming	11.2%
Chat	8.4%
Rewrite	6.6%
Summarization	4.2%
Classification	3.5%
Other	3.5%
Closed QA	2.6%
Extract	1.9%

Table 2: Illustrative prompts from our API prompt dataset. These are fictional examples inspired by real usage—see more examples in Appendix A.2.1.

Use-case	Prompt
Brainstorming	List five ideas for how to regain enthusiasm for my career
Generation	Write a short story where a bear goes to the beach, makes friends with a seal, and then returns home.
Rewrite	This is the summary of a Broadway play: """ {summary} """ This is the outline of the commercial for that play: """

The distribution of prompts used to finetune InstructGPT

Dialing it back a bit...

Deep Learning Foundations