Text to Intelligence

LLM Training Camp

Content prepared by Veeramanohar A

For Vidyut 2025, Amrita Vishwa Vidyapeetham.

Agenda

1

Section 1

What Are LLMs, Where We Use Them Daily, and How Businesses Leverage Them To Solve Problems

2

Section 2

Behind the Scenes of LLMs: Transformers, Tokenization, Embedding, GPT Architecture, Training Process, and Hands-on Text Generation using LLMs

3

Section 3

Domain-Specific Applications of LLMs, Fine-Tuning Techniques, and Data Preparation for Fine-Tuning

4

Section 4

Hands-on LLM Fine-Tuning, Evaluating Performance of LLM, and Best practices in Industry

Agenda

1

Section 1

What Are LLMs, Where We Use Them Daily, and How Businesses Leverage Them To Solve Problems

2

Section 2

Behind the Scenes of LLMs: Transformers, Tokenization, Embedding, GPT Architecture, Training Process, and Hands-on Text Generation using LLMs

3

Section 3

Domain-Specific Applications of LLMs, Fine-Tuning Techniques, and Data Preparation for Fine-Tuning

4

Section 4

Hands-on LLM Fine-Tuning, Evaluating Performance of LLM, and Best practices in Industry

What Are Large Language Models (LLMs)?

The model is simply (pre)trained to predict the next word

Next word prediction


Sample Text
LLMs predict one word at a time
Input to LLM
Target to predict
LLM can’t access words past the target

Iteration 1

LLMs predict one word at a time

Iteration 1

LLMs predict one word at a time

Iteration 2

LLMs predict one word at a time

Iteration 1 LLMs predict one word at a time

Iteration 2 LLMs predict one word at a time

Iteration 3 LLMs predict one word at a time

Iteration 1 LLMs predict one word at a time

Iteration 2 LLMs predict one word at a time

Iteration 3 LLMs predict one word at a time

Iteration 4 LLMs predict one word at a time

Iteration 1 LLMs predict one word at a time

Iteration 2 LLMs predict one word at a time

Iteration 3 LLMs predict one word at a time

Iteration 4 LLMs predict one word at a time

Iteration 5 LLMs predict one word at a time

Iteration 1 LLMs predict one word at a time

Iteration 2 LLMs predict one word at a time

Iteration 3 LLMs predict one word at a time

Iteration 4 LLMs predict one word at a time

Iteration 5 LLMs predict one word at a time

Iteration 6 LLMs predict one word at a time

Where do we use LLMs in daily life?

Content Generation

Virtual Assistants

Code Generation

Translation

LLMs

Education

Customer Support

Idea Brainstorming

Market Research

How we ask the model matters a lot!

How we ask the model matters a lot!

“Explain Gravity.”

Gravity is the force that pulls objects toward each other. It's the reason why things fall when you drop them and why planets orbit the Sun.

In simple terms:

Gravity pulls objects together. The more mass something has, the stronger its gravitational pull.

The Earth’s gravity pulls everything toward its center — that’s why you stay on the ground and why dropped objects fall downward.

In space, gravity keeps the Moon orbiting the Earth, and the Earth orbiting the Sun.

How we ask the model matters a lot!

“Explain Gravity.”

Gravity is the force that pulls objects toward each other. It's the reason why things fall when you drop them and why planets orbit the Sun.

In simple terms:

Gravity pulls objects together. The more mass something has, the stronger its gravitational pull.

The Earth’s gravity pulls everything toward its center — that’s why you stay on the ground and why dropped objects fall downward.

In space, gravity keeps the Moon orbiting the Earth, and the Earth orbiting the Sun.

“Explain Gravity to me like I’m five.”

Alright! Imagine you have a ball. When you throw it up, it always comes back down, right? That’s because of something called gravity.

Gravity is like an invisible hand the Earth has. It gently pulls everything toward it—balls, people, even you! That’s why you don’t float away like a balloon.

So, when you jump, gravity pulls you back down. It’s like Earth saying, “Hey! Come back here!” 🌍✨

That's gravity!

Prompting Methods

Zero-Shot Prompting

Few-Shot Prompting

Chain of Thought Prompting

Prompt Chaining

LLMs Don’t Know What They Don’t Know!

LLMs Don’t Know What They Don’t Know!

⚠️ Knowledge Cutoff

Hallucinations

LLMs Don’t Know What They Don’t Know!

⚠️ Knowledge Cutoff

Hallucinations

✅ **R**etrieval

Augmented

Generation

Question → Retrieve Relevant Data → Augment to LLM’s input → Generate the final response

Best Practices to follow when implementing RAG systems

Multi-Pipeline Framework → Router

Best Practices to follow when implementing RAG systems

Multi-Pipeline Framework → Router

Chunking Strategies

Best Practices to follow when implementing RAG systems

Multi-Pipeline Framework → Router

Chunking Strategies

Utilize Structured Data

Best Practices to follow when implementing RAG systems

Multi-Pipeline Framework → Router

Chunking Strategies

Utilize Structured Data

Chain of Verification Prompting

Solving Business Problems with LLMs

Solving Business Problems with LLMs

Product Recommendations

RAG on Product Database

Solving Business Problems with LLMs

Product Recommendations	→	RAG on Product Database
Candidate Screening & Onboarding	→	Document Summarization & QnA Chatbot

Solving Business Problems with LLMs

Product Recommendations	→	RAG on Product Database
Candidate Screening & Onboarding	→	Document Summarization & QnA Chatbot
Marketing Copy & Social Media Content	→	Text Generation & Feature Extraction

Solving Business Problems with LLMs

Product Recommendations	→	RAG on Product Database
Candidate Screening & Onboarding	→	Document Summarization & QnA Chatbot
Marketing Copy & Social Media Content	→	Text Generation & Feature Extraction
Data Analysis and Visualization	→	Code Generation & Chat over document

Agenda

1

Section 1

What Are LLMs, Where We Use Them Daily, and How Businesses Leverage Them To Solve Problems

2

Section 2

Behind the Scenes of LLMs: Transformers, Tokenization, Embedding, GPT Architecture, Training Process, and Hands-on Text Generation using LLMs

3

Section 3

Domain-Specific Applications of LLMs, Fine-Tuning Techniques, and Data Preparation for Fine-Tuning

4

Section 4

Hands-on LLM Fine-Tuning, Evaluating Performance of LLM, and Best practices in Industry

Under the hood of LLMs.

Iteration 1

“This is”

↑

Output Layers

↑

LLM

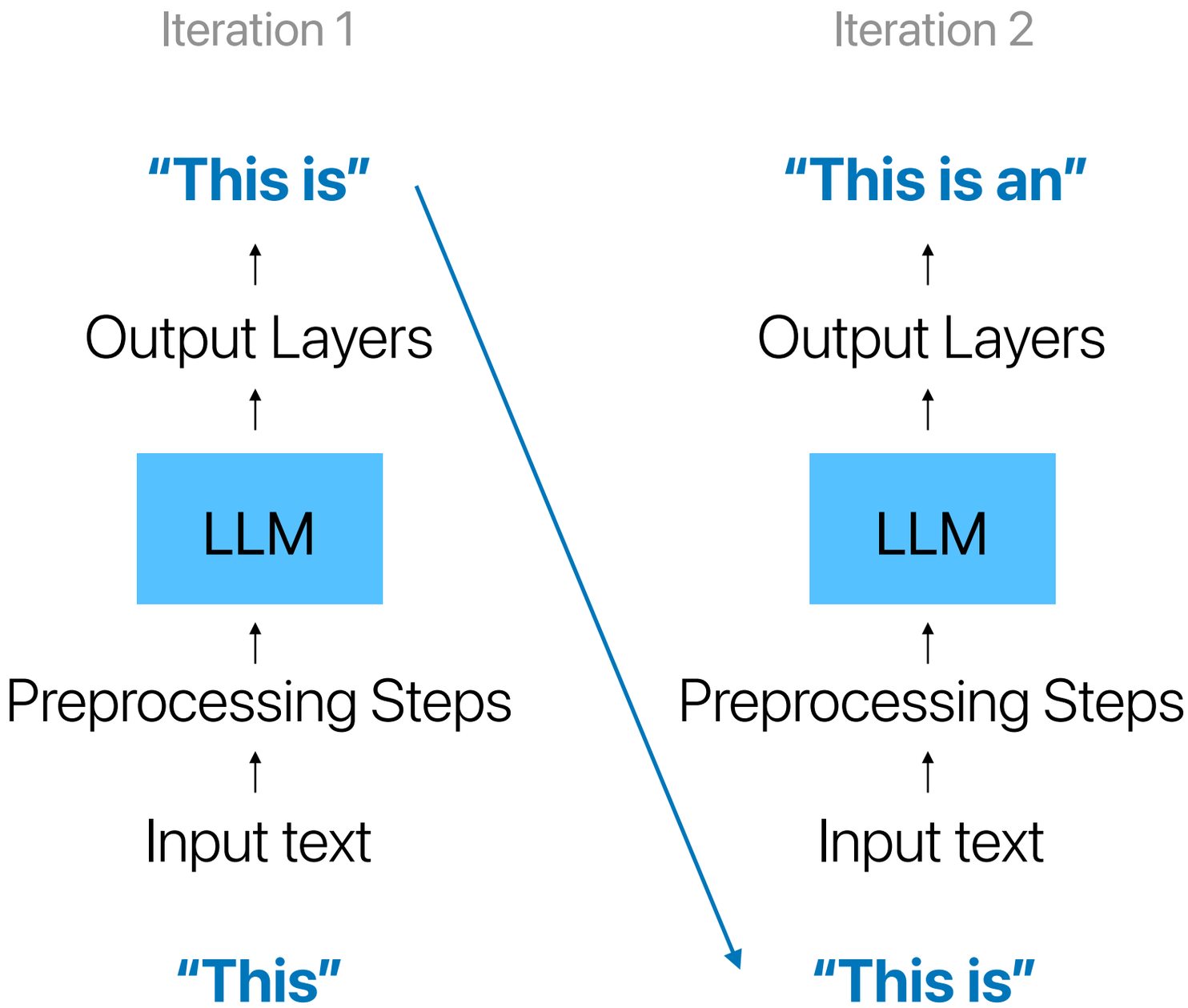
↑

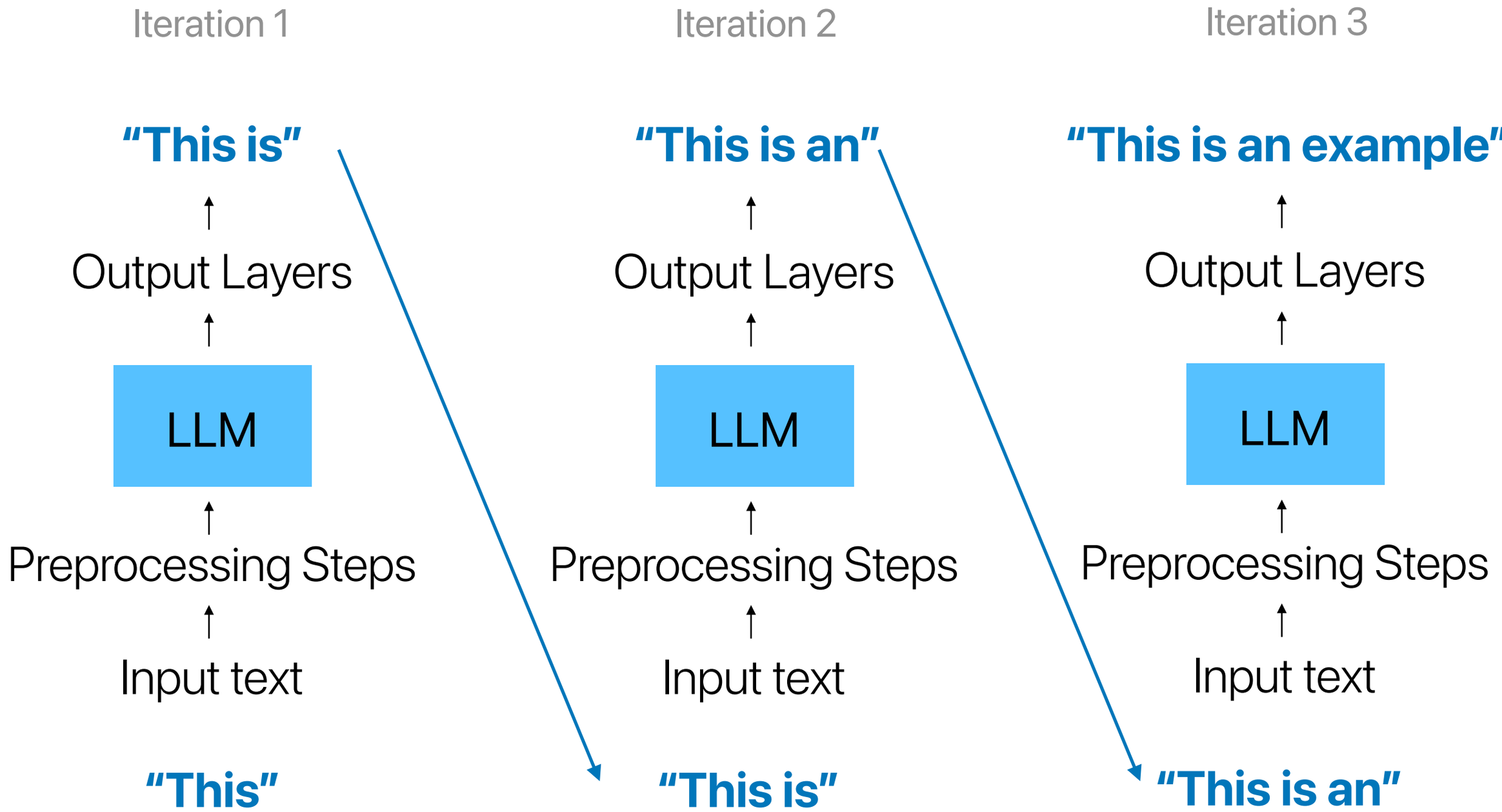
Preprocessing Steps

↑

Input text

“This”


Iteration 1
“This is”
Output Layers
LLM
Preprocessing Steps
Input text
“This”
Iteration 2
“This is an”
Output Layers
LLM
Preprocessing Steps
Input text
“This is”


Iteration 1
Iteration 2
Iteration 3
“This is”
“This is an”
“This is an example”
Output Layers
Output Layers
Output Layers
LLM
LLM
LLM
Preprocessing Steps
Preprocessing Steps
Preprocessing Steps
Input text
Input text
Input text
“This”
“This is”
“This is an”

Tokenization

Input Text

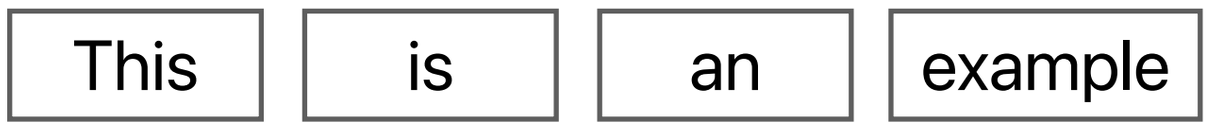
This is an example

Tokenization

Input Text

This is an example

Tokenized text



Tokenization

Input Text

This is an example

Tokenized text

This	is	an	example

Token IDs

4013	2052	133	390

Embeddings

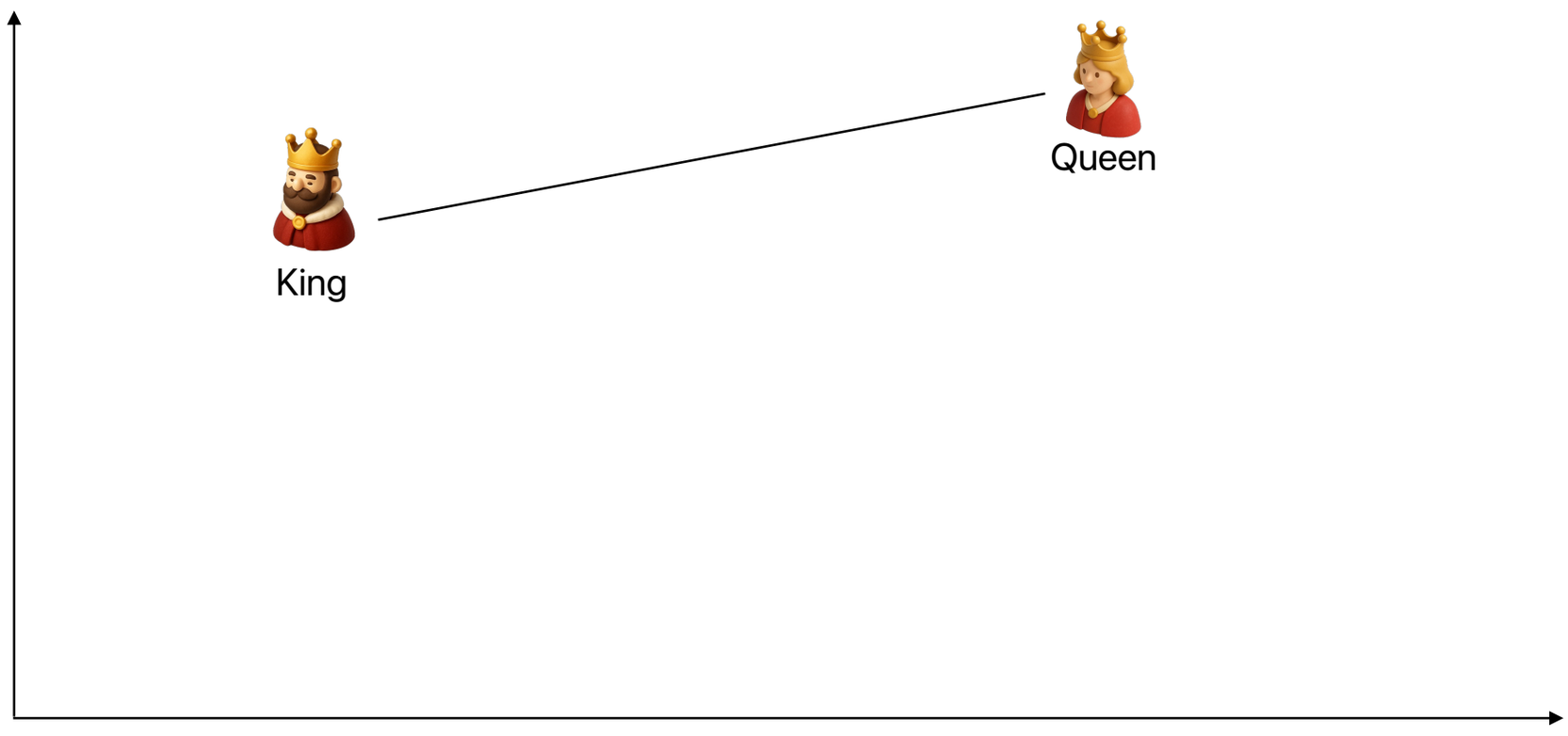
Capture Semantic Relationship

Learned during the training process

Embeddings

Capture Semantic Relationship

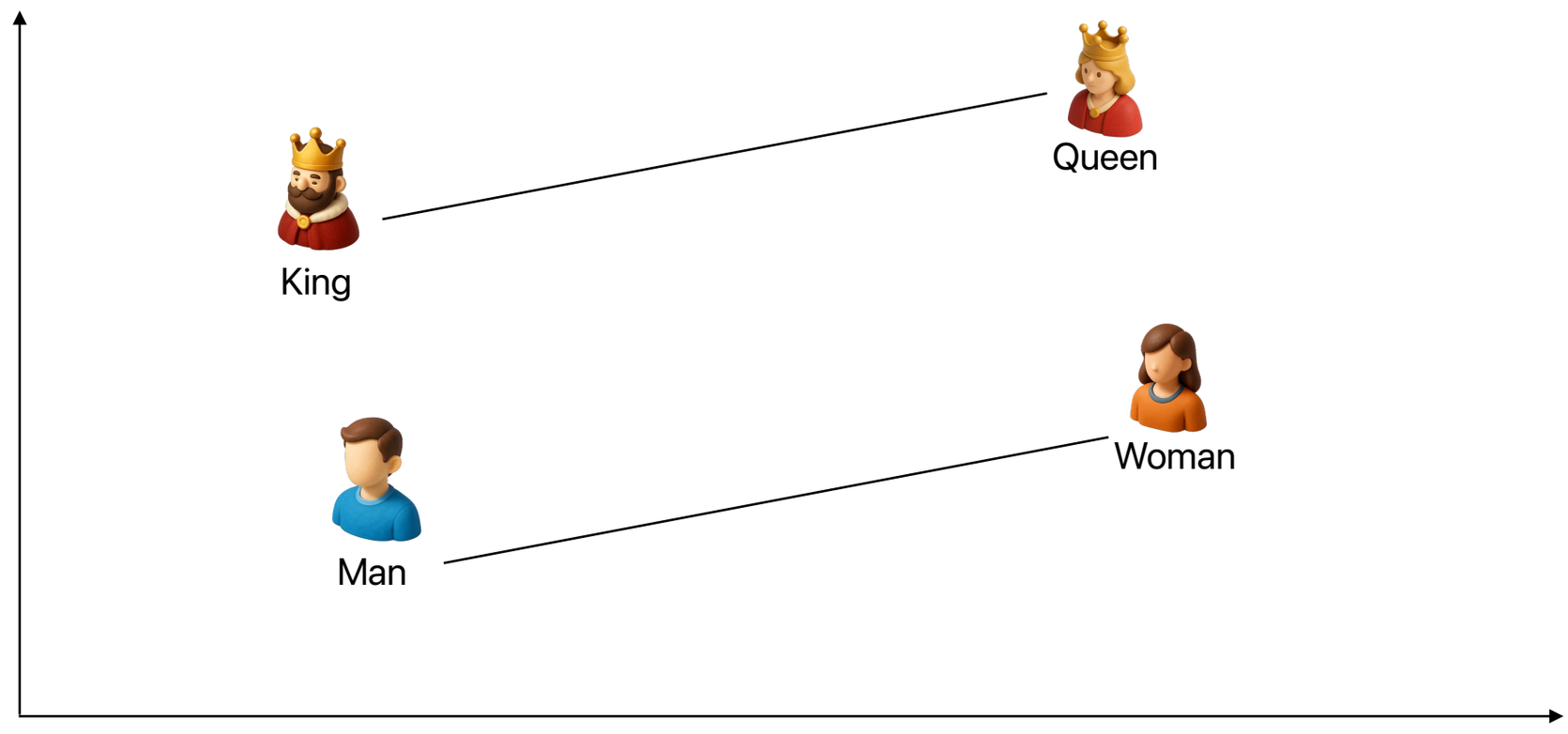
Learned during the training process



Embeddings

Capture Semantic Relationship

Learned during the training process



Attention is all you need!

Attention is all you need!

Positional Encoding

Attention is all you need!

Positional Encoding

Attention

Attention is all you need!

Positional Encoding

Attention

Self-Attention

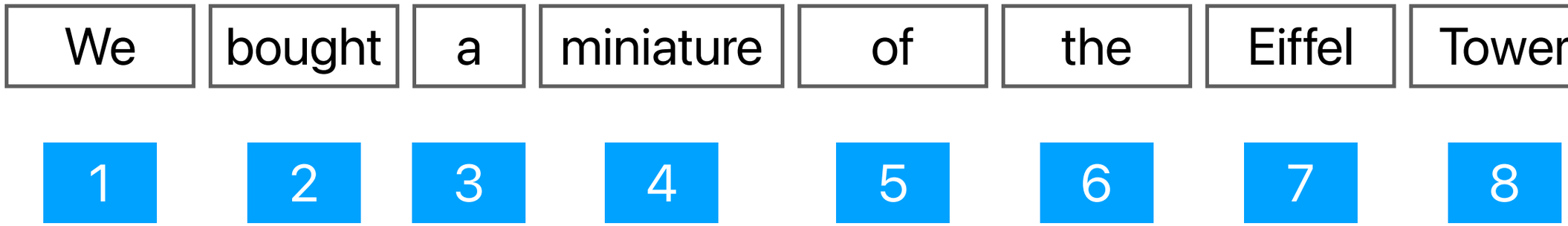
Positional Encoding

We bought a miniature of the Eiffel Tower

We	bought	a	miniature	of	the	Eiffel	Tower

Positional Encoding

We bought a miniature of the Eiffel Tower



Learn the importance of word order form the data.

Attention & Self-Attention

We bought a miniature of the Eiffel Tower

We	bought	a	miniature	of	the	Eiffel	Tower

What words does the model attends to,
when it is making the prediction for the output sentence.

Attention & Self-Attention

What was bought?

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought								
a								
miniature								
of								
the								
Eiffel								
Tower								

Attention & Self-Attention

What was bought?

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought				●			●	●
a								
miniature		●						
of								
the								
Eiffel		●						
Tower		●						

Attention & Self-Attention

What is the miniature?

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought				●			●	●
a								
miniature		●						
of								
the								
Eiffel		●						
Tower		●						

Attention & Self-Attention

What is the miniature?

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought				●			●	●
a								
miniature		●					●	●
of								
the								
Eiffel		●		●				
Tower		●		●				

Attention & Self-Attention

What tower is it?

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought				●			●	●
a								
miniature		●					●	●
of								
the								
Eiffel		●		●				
Tower		●		●				

Attention & Self-Attention

What tower is it?

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought				●			●	●
a								
miniature		●					●	●
of								
the								
Eiffel		●		●				●
Tower		●		●			●	

Attention & Self-Attention

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought								
a								
miniature								
of								
the								
Eiffel								
Tower								

Queries, Keys and Values

Attention & Self-Attention

	We	bought	a	miniature	of	the	Eiffel	Tower
We								
bought								
a								
miniature								
of								
the								
Eiffel								
Tower								

Queries, Keys and Values

Attention & Self-Attention

	We	bought	a	miniature	of	the	Eiffel	Tower
We	1.2	-inf	-inf	-inf	-inf	-inf	-inf	-inf
bought	1	1.3	-inf	-inf	-inf	-inf	-inf	-inf
a	0.6	1.1	1.5	-inf	-inf	-inf	-inf	-inf
miniature	0.5	1	1.2	1.6	-inf	-inf	-inf	-inf
of	0.4	0.9	1.1	1.3	1.7	-inf	-inf	-inf
the	0.3	0.7	1	1.2	1.4	1.8	-inf	-inf
Eiffel	0.2	0.5	0.9	1.1	1.3	1.5	2	-inf
Tower	0.1	0.4	0.8	1	1.2	1.4	1.8	2.2

Queries, Keys and Values

Query, Key and Values

Query Vector - how much focus to place on other parts of the input

"What is this miniature of?"

Query, Key and Values

Query Vector - how much focus to place on other parts of the input

"What is this miniature of?"

Key Vectors - represent each word in the sentence

"Eiffel" and "Tower"

Query, Key and Values

Query Vector - how much focus to place on other parts of the input

"What is this miniature of?"

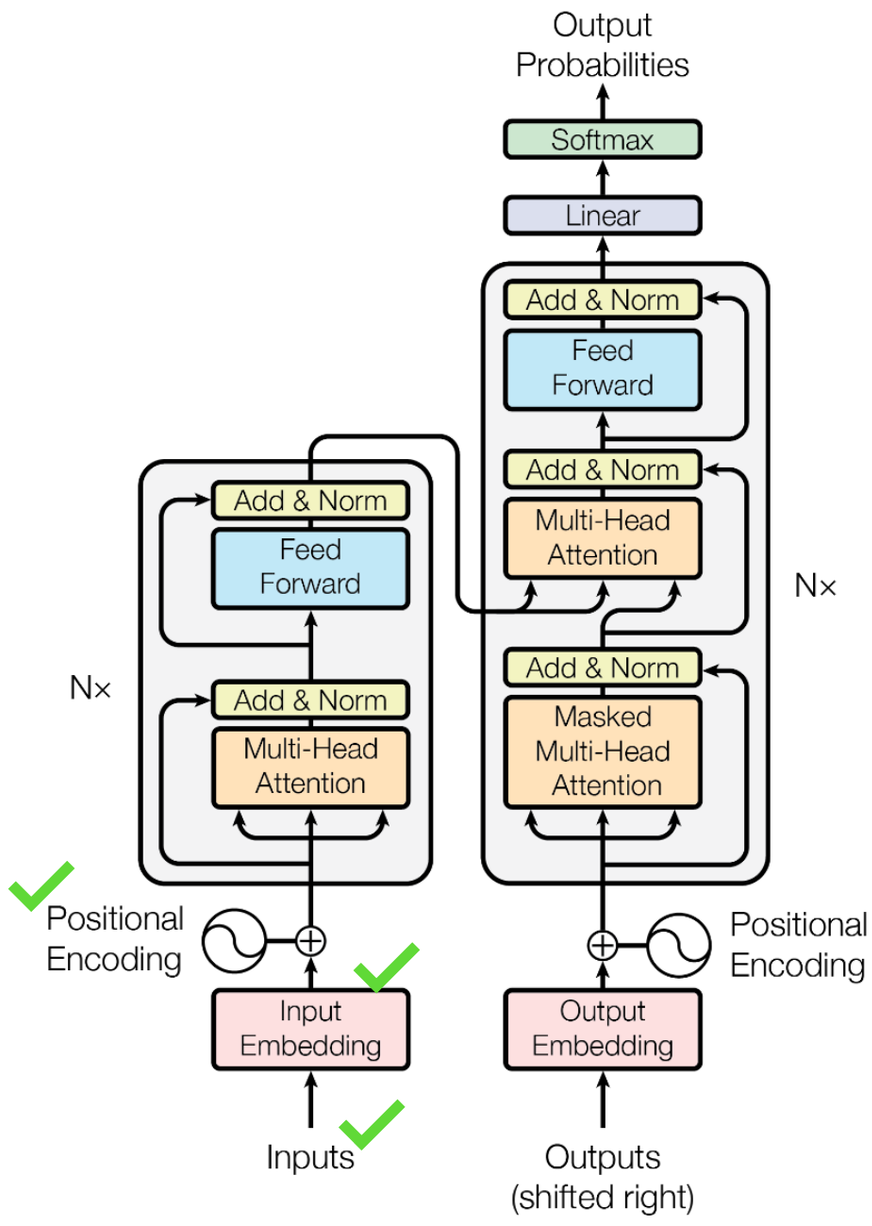
Key Vectors - represent each word in the sentence

"Eiffel" and "Tower"

Value Vectors - contextual meaning of the word

Recognizing that "miniature" is related to "Eiffel Tower" in this context.

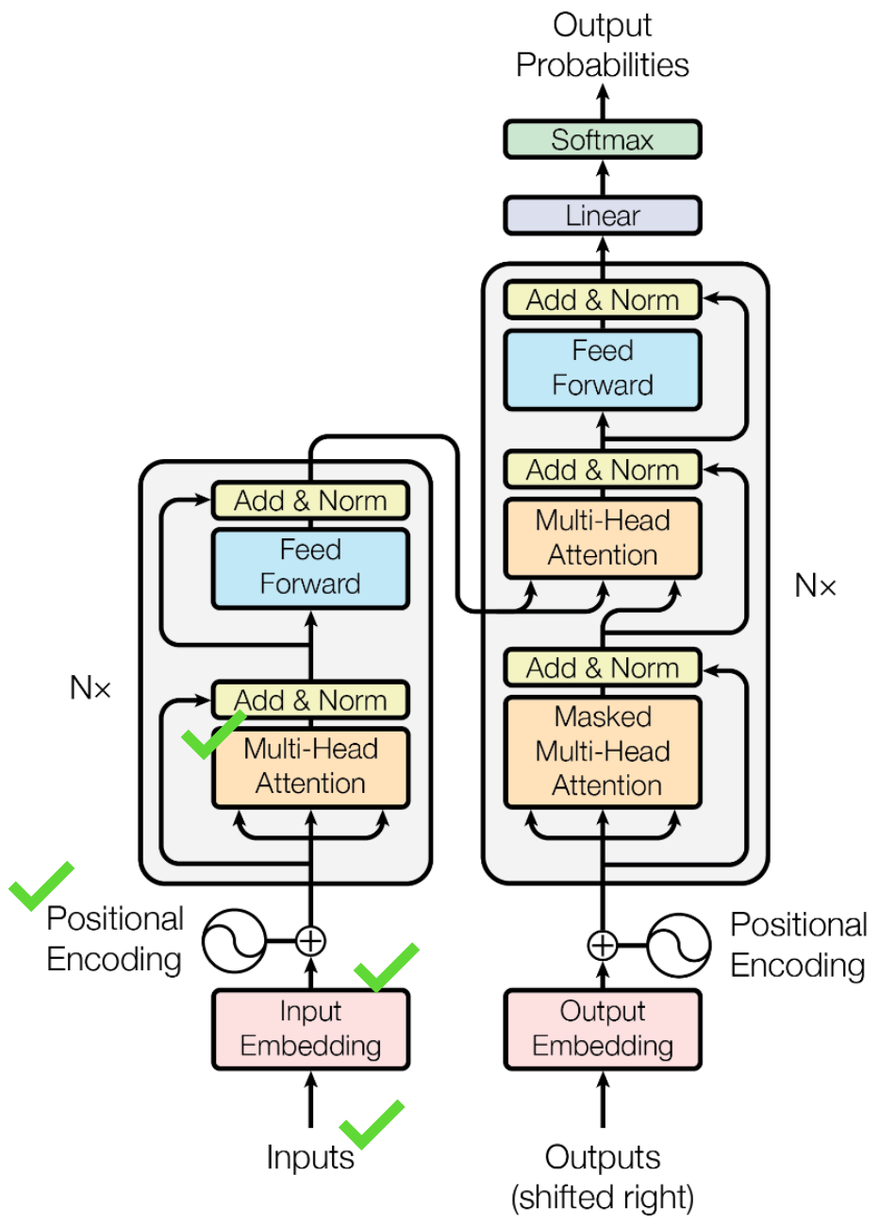
Attention is all you need!



Multi-Headed Attention

Learn from different angles at the same time

Attention is all you need!



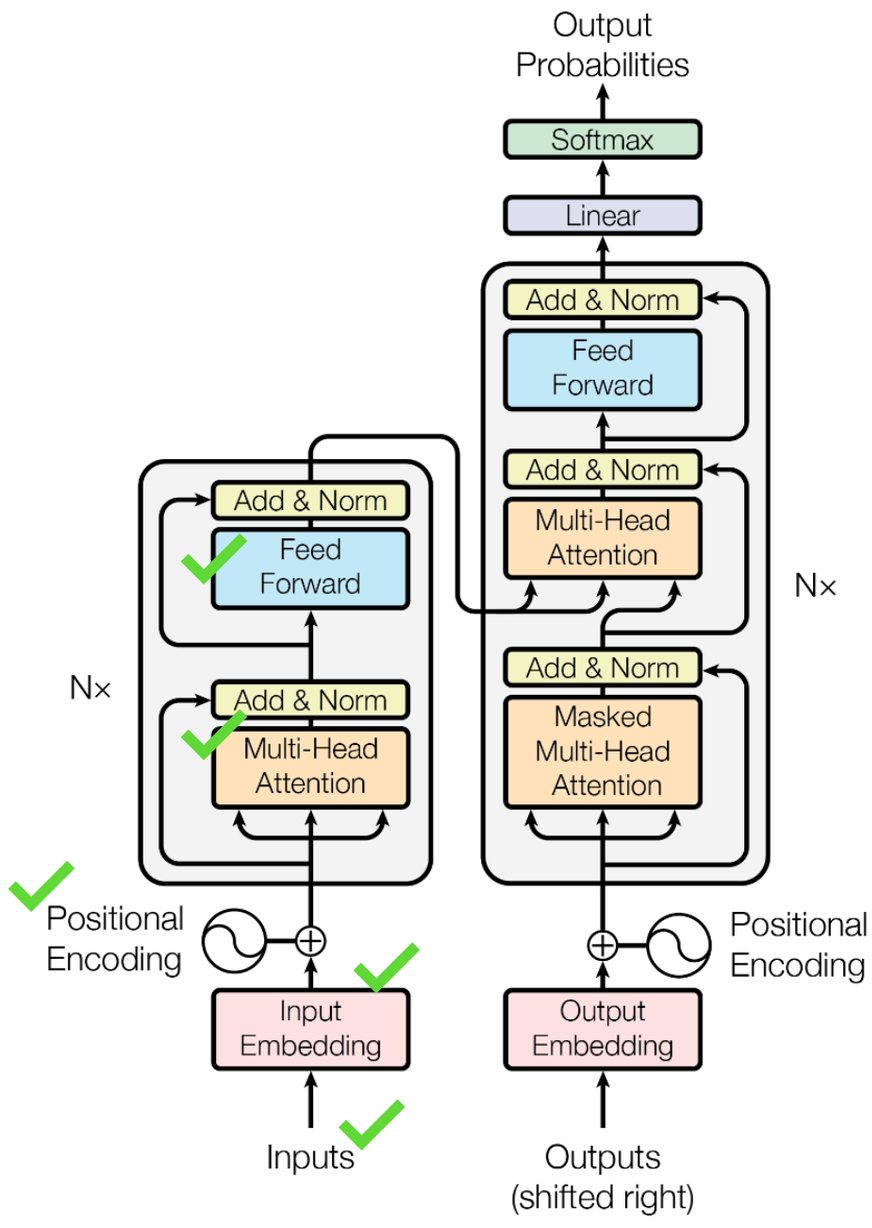
Multi-Headed Attention

Learn from different angles at the same time

Feed Forward

Thinks about the role of each word

Attention is all you need!



Multi-Headed Attention

Learn from different angles at the same time

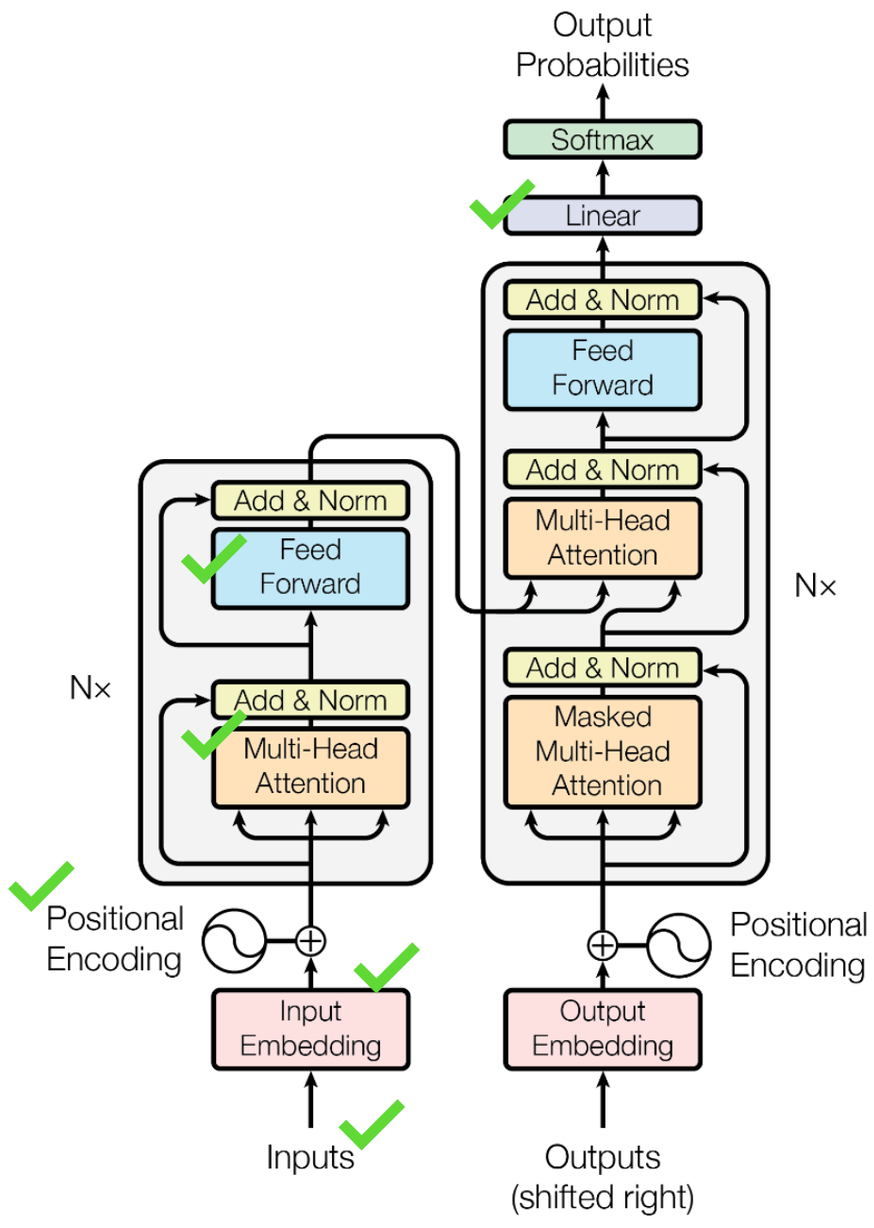
Feed Forward

Thinks about the role of each word

Add & Norm

Add your old thoughts back

Attention is all you need!



Multi-Headed Attention

Learn from different angles at the same time

Feed Forward

Thinks about the role of each word

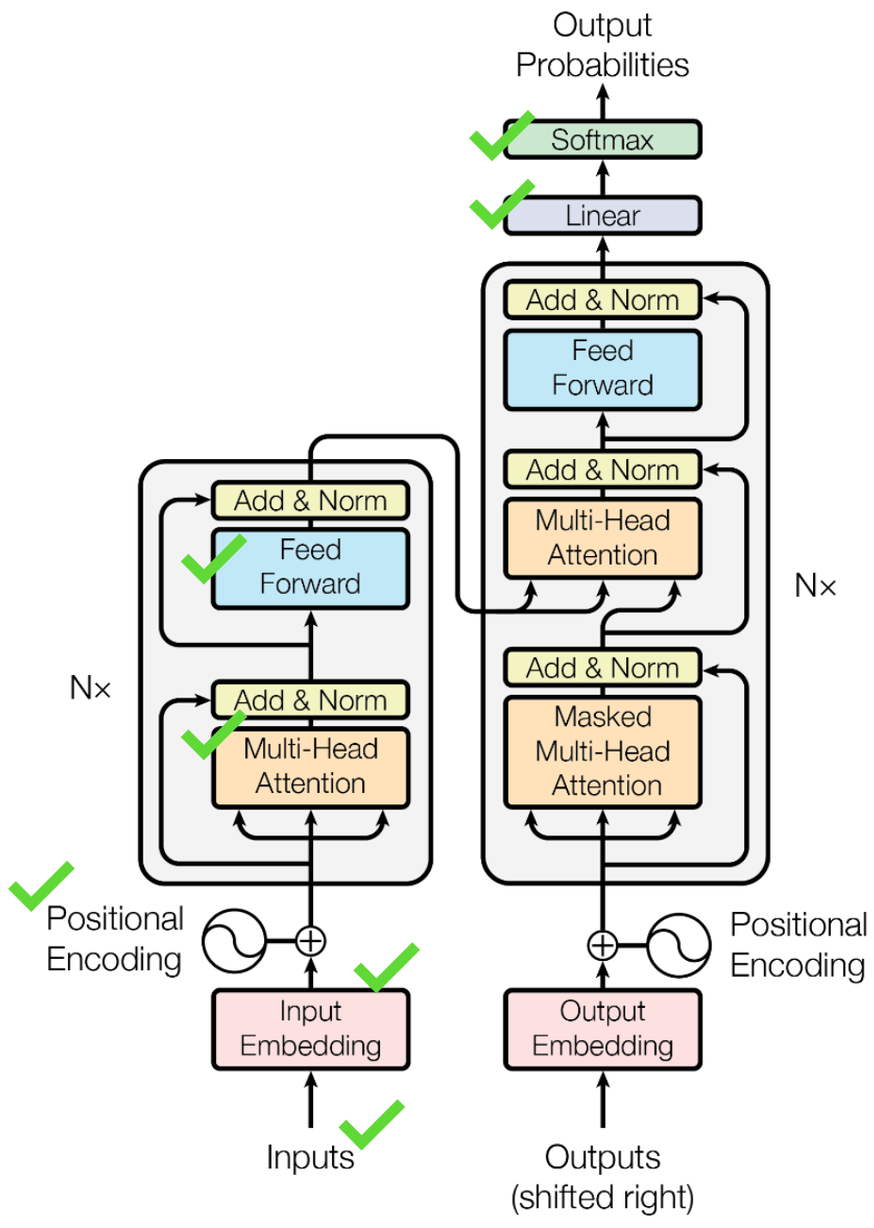
Add & Norm

Add your old thoughts back

Linear

Thoughts into words

Attention is all you need!



Multi-Headed Attention

Learn from different angles at the same time

Feed Forward

Thinks about the role of each word

Add & Norm

Add your old thoughts back

Linear

Thoughts into words

Softmax

Add a probability to each word

LLM Pre-training

LLM Pre-training

Parameters

Continuous values

LLM Pre-training

Parameters

Continuous values

Initial: Gibberish

LLM Pre-training

Parameters

Continuous values

Initial: Gibberish

Learns through examples

LLM Pre-training

Parameters

Continuous values

Initial: Gibberish

Learns through examples

Backpropagation

LLM Pre-training

Parameters

Continuous values

Initial: Gibberish

Learns through examples

Backpropagation

Trillions of Data

LLM Pre-training

Parameters

Continuous values

Initial: Gibberish

Learns through examples

Backpropagation

Trillions of Data

RLHF

Context Window

Maximum number of words to consider at once

Capacity to remember

Context Window & Parameters

Model Name	Parameters	Context Window (In tokens)
GPT-4o	~200B	128k
Gemini 2.5 Pro	Not disclosed	1M
Llama 4 Scout	17B (active)	10M
Llama 3.2	1B, 3B, 11B, 90B	128k
DeepSeek-V3	37B (active)	128k
OpenAI o1	200B	200k
Grok-1	314B	8k
Grok-3	2.7T	128k
GPT-4	1.76T	128k
Phi-3	7B	128k

Control the output

Temperature

Controls the randomness of the model's output.

↓ Lower temperature
More deterministic

↑ Higher temperature
More creative

Control the output

Temperature

Controls the randomness of the model's output.

↓ Lower temperature
More deterministic

↑ Higher temperature
More creative

Top-k Sampling

Limits the model’s token selection

↓ Lower k
More predictable

↑ Higher k
More diverse

Control the output

Temperature

Controls the randomness of the model's output.

Lower temperature
More deterministic

Higher temperature
More creative

Top-k Sampling

Limits the model’s token selection

Lower k
More predictable

Higher k
More diverse

Top-p Sampling

Adjusts # of tokens based on cumulative probability

Lower p
More focused

Higher p
More elaborate

Control the output

Temperature

Controls the randomness of the model's output.

↓ Lower temperature
More deterministic

↑ Higher temperature
More creative

Top-k Sampling

Limits the model's token selection

↓ Lower k
More predictable

↑ Higher k
More diverse

Top-p Sampling

Adjusts # of tokens based on cumulative probability

↓ Lower p
More focused

↑ Higher p
More elaborate

Frequency penalty, Presence penalty, Max tokens, Stop sequences, Repetition penalty

Hands-on

Hands-on

Agenda

1

Section 1

What Are LLMs, Where We Use Them Daily, and How Businesses Leverage Them To Solve Problems

2

Section 2

Behind the Scenes of LLMs: Transformers, Tokenization, Embedding, GPT Architecture, Training Process, and Hands-on Text Generation using LLMs

3

Section 3

Domain-Specific Applications of LLMs, Fine-Tuning Techniques, and Data Preparation for Fine-Tuning

4

Section 4

Hands-on LLM Fine-Tuning, Evaluating Performance of LLM, and Best practices in Industry

From General to Genius

From General to Genius

🏥 Healthcare & Biomedical Research

From General to Genius

🏥 Healthcare & Biomedical Research

⚖️ Legal & Regulatory Compliance

From General to Genius

🏥 Healthcare & Biomedical Research

⚖️ Legal & Regulatory Compliance

💰 Finance & Investment

From General to Genius

🏥 Healthcare & Biomedical Research

⚖️ Legal & Regulatory Compliance

💰 Finance & Investment

📐 Mathematics

From General to Genius

🏥 Healthcare & Biomedical Research

⚖️ Legal & Regulatory Compliance

💰 Finance & Investment

📐 Mathematics

🛠️ Enterprise & Customer Service

Why a genius is required?

Task-Specific Experts

Why a genius is required?

Task-Specific Experts

Brand Voice and Tone Customization

Why a genius is required?

Task-Specific Experts

Brand Voice and Tone Customization

Data Privacy and Compliance

Why a genius is required?

Task-Specific Experts

Brand Voice and Tone Customization

Data Privacy and Compliance

Multilingual and Cultural Adaptation

Why a genius is required?

Task-Specific Experts

Brand Voice and Tone Customization

Data Privacy and Compliance

Multilingual and Cultural Adaptation

Enhanced Safety and Alignment

Methods of Fine-Tuning

Full fine-tuning

Updating all the parameters

Methods of Fine-Tuning

Full fine-tuning	Updating all the parameters
Feature Extraction	Only the final layers are trained

Methods of Fine-Tuning

Full fine-tuning	Updating all the parameters
Feature Extraction	Only the final layers are trained
Parameter-Efficient Fine-Tuning (PEFT)	Trainable adapter layers

Parameter-Efficient Fine-Tuning (PEFT)

Low-Rank Adaptation (LoRA)

Parameter-Efficient Fine-Tuning (PEFT)

Low-Rank Adaptation (LoRA)

Doesn’t Update all model parameters

Approximate the necessary weight updates

Reducing computational resources and memory requirements.

Parameter-Efficient Fine-Tuning (PEFT)

Low-Rank Adaptation (LoRA)

Doesn’t Update all model parameters

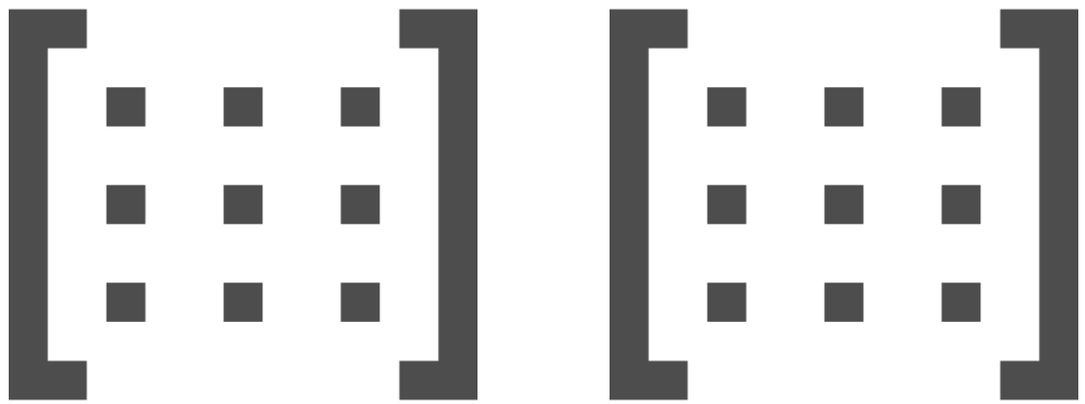
Approximate the necessary weight updates

Reducing computational resources and memory requirements.

Track change in weights

Low-Rank Adaptation (LoRA)

Track change in weights



A

B

Low Rank Matrices

$A \in R^{r \times k}$

$B \in R^{d \times r}$

Reduces the number of trainable parameters from,

$d \times k$ to $r \times (d+k)$

Low-Rank Adaptation (LoRA)

Rank	7B	13B	70B	180B
1	167k	228k	529k	849k
2	334k	456k	1M	2M
8	1M	2M	4M	7M
16	3M	4M	8M	14M
512	86M	117M	270M	434M
1024	171M	233M	542M	869M
8192	1.4B	1.8B	4.3B	7B

Low-Rank Adaptation (LoRA)

Rank of LoRA

Controls the number of trainable parameters

Target Modules

Controls the layers in which LoRA is applied

Alpha

Adjusts the weight updates during training

Dropout

To prevent overfitting

(↑-Small data, ↓-Large data)

Bias

Determines whether to modify the bias or not

Random state

To ensure reproducability

Gradient Checkpointing

To save memory during training

Loft Quantization

Applies quantization strategies during fine-tuning

Quantization

Compression Technique

Reduces Precision

Decreases Model's Computational Requirements

Quantization

Compression Technique

Reduces Precision

Decreases Model's Computational Requirements

Post-Training Quantization

Quantization Aware Training

Mixed-Precision Quantization

Fine-Tuning Parameters

Batch Size

Affects memory usage
and training speed

Warmup Steps

of steps for LR to increase
From 0 to initial value

Optimizer

Updates the model weight
to minimize the loss

LR Scheduler

Adjusts LR during training
to improve model convergence

Weight Decay

Prevents Overfitting by
Regularization

LR - Learning Rate

Reinforcement Learning from Human Feedback (RLHF)

Human-in-the-loop

Reinforcement Learning from Human Feedback (RLHF)

Human-in-the-loop

Train a reward model

Reinforcement Learning from Human Feedback (RLHF)

Human-in-the-loop

Train a reward model

Proximal Policy Optimization

Reinforcement Learning from Human Feedback (RLHF)

Human-in-the-loop

Train a reward model

Proximal Policy Optimization

Direct Preference Optimization

Reinforcement Learning from Human Feedback (RLHF)

Prompt

Explain the theory of relativity in simple terms

Reinforcement Learning from Human Feedback (RLHF)

Prompt	Explain the theory of relativity in simple terms
Response A	A concise explanation suitable for laypersons.
Response B	An overly technical description filled with jargon.
Response C	An incorrect or misleading explanation.

Reinforcement Learning from Human Feedback (RLHF)

Prompt	Explain the theory of relativity in simple terms
Response A	A concise explanation suitable for laypersons.
Response B	An overly technical description filled with jargon.
Response C	An incorrect or misleading explanation.

$$A > B > C$$

Agenda

1

Section 1

What Are LLMs, Where We Use Them Daily, and How Businesses Leverage Them To Solve Problems

2

Section 2

Behind the Scenes of LLMs: Transformers, Tokenization, Embedding, GPT Architecture, Training Process, and Hands-on Text Generation using LLMs

3

Section 3

Domain-Specific Applications of LLMs, Fine-Tuning Techniques, and Data Preparation for Fine-Tuning

4

Section 4

Hands-on LLM Fine-Tuning, Evaluating Performance of LLM, and Best practices in Industry

How well does an LLM perform?

How well does an LLM perform?

🧠 Intelligence Benchmarking

Evaluation	Field	Questions	Response Type
MMLU-Pro	Reasoning & Knowledge	12,032	Multiple-Choice
Humanity's Last Exam	Reasoning & Knowledge	2,684	Open Answer
GPQA Diamond	Scientific Reasoning	198	Multiple-Choice
MATH-500	Quantitative Reasoning	500	Open Answer
LiveCodeBench	Code Generation	315	Python
Global MMLU Lite	Multilingual Reasoning	~6000	Multiple-Choice
MGSM	Multilingual Mathematic	~2000	Open Answer

How well does an LLM perform?

⚡ Performance Benchmarking

Tests

Different Workloads
varied input token size

Load Scenarios
Single vs Parallel prompts

Frequency Testing

How well does an LLM perform?

⚡ Performance Benchmarking

Tests

Different Workloads

varied input token size

Load Scenarios

Single vs Parallel prompts

Frequency Testing

Metrics

Latency

- Time to first token
- Time to first answer token

Output Speed (token/s)

End-to-End

Response time

Structured Outputs

Are LLMs the catalysts for the next industrial revolution?

Hands-on

For any queries, contact

Veeramanohar A

veeramanohar.a@latentview.com | veeramanohar@icloud.com

LinkedIn

Workshop Materials

Thank You!