

Introduction to Large Language Models



AI Methods Lab Meeting

Ruurd Kuiper
Julius Centrum – Datascience and Biostatistics



Contents



What are LLMs?



How do they work?

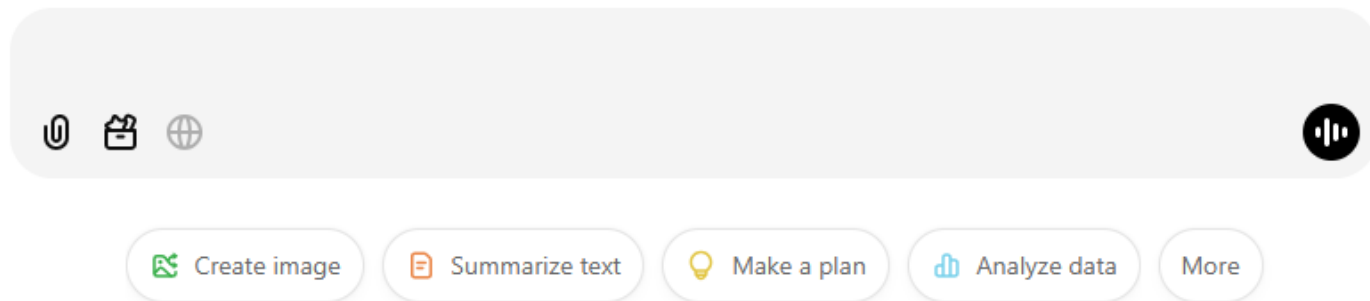


LLMs in
Healthcare

What are LLMs?

You have probably used them before..

What can I help with?



What are LLMs?

You have probably used them before..

What can I help with?

Message ChatGPT



 Create image

 Summarize text

 Make a plan

 Analyze data

More



GPT-4



UMC Utrecht

What are LLMs?

Different implementations



What are LLMs?

Different implementations



What are LLMs?

Different implementations



What are LLMs?

Different implementations

Proprietary



Open-source



What are LLMs

A definition

A **large language model (LLM)** is:

A **machine learning model** trained on **large amounts** of **(text) data** to process and generate **natural language**, using **contextual understanding** through its **large neural network architecture**.

What are LLMs?

What do they do?

LLMs estimate the conditional probability distribution $P(w_{t+1} | w_1, w_2, \dots, w_t)$, where $w_1, w_2, \dots, w_t$ are the preceding words in a sequence, and w_{t+1} represents the next word to be predicted.

What are LLMs?

What do they do?

LLMs estimate the conditional probability distribution $P(w_{t+1}|w_1, w_2, \dots, w_t)$, where $w_1, w_2, \dots, w_t$ are the preceding words in a sequence, and w_{t+1} represents the next word to be predicted.

w_1

w_2

w_3

w_4

$w_{t+1} = w_5$

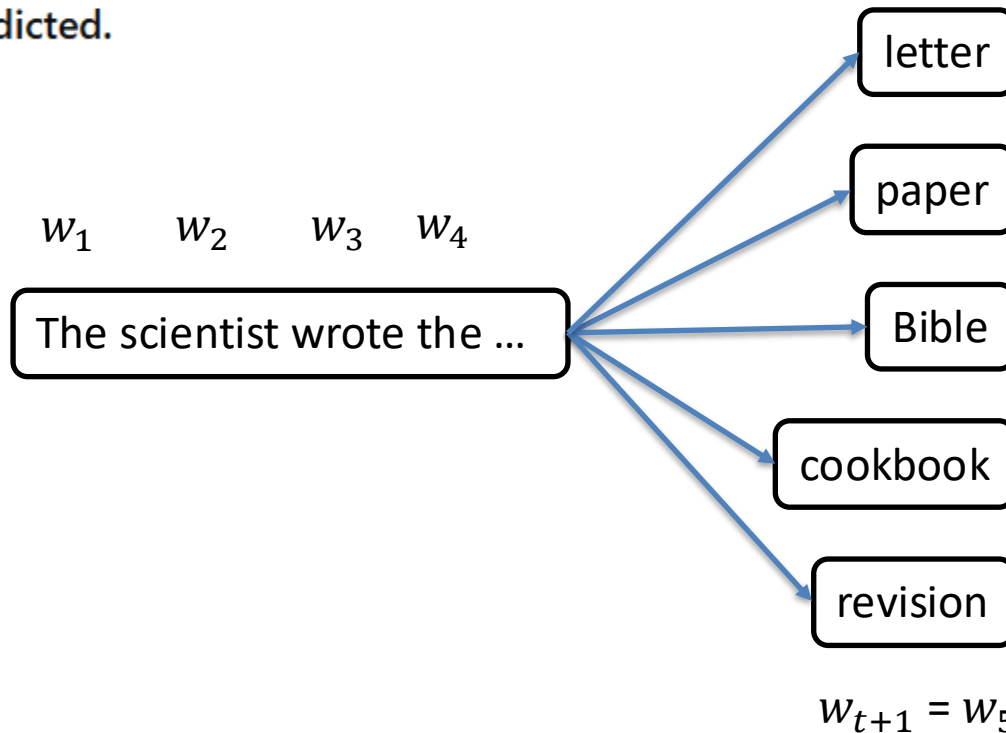
The scientist wrote the ...

?

What are LLMs?

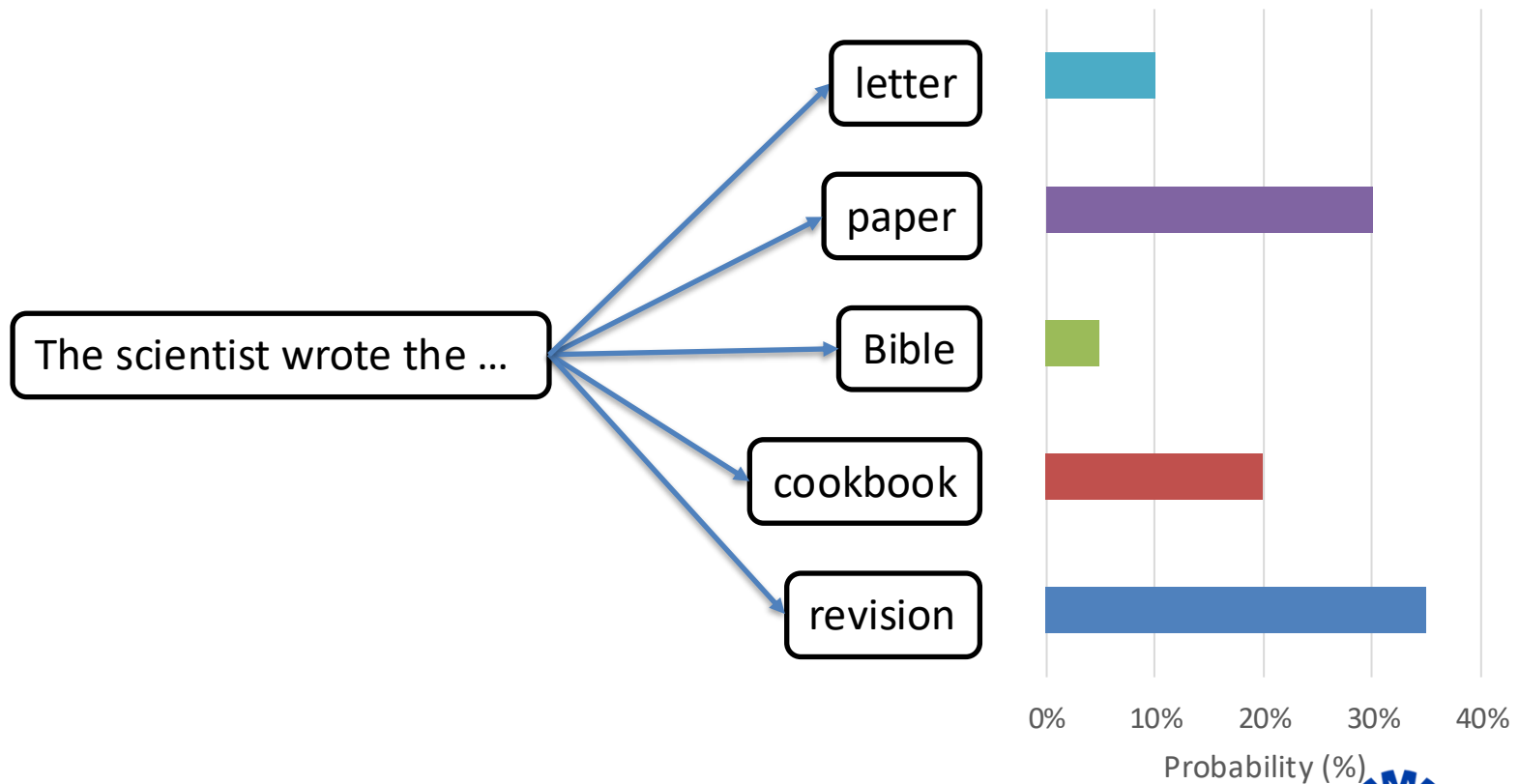
What do they do: **Next-word-prediction**

LLMs estimate the conditional probability distribution $P(w_{t+1}|w_1, w_2, \dots, w_t)$, where $w_1, w_2, \dots, w_t$ are the preceding words in a sequence, and w_{t+1} represents the next word to be predicted.



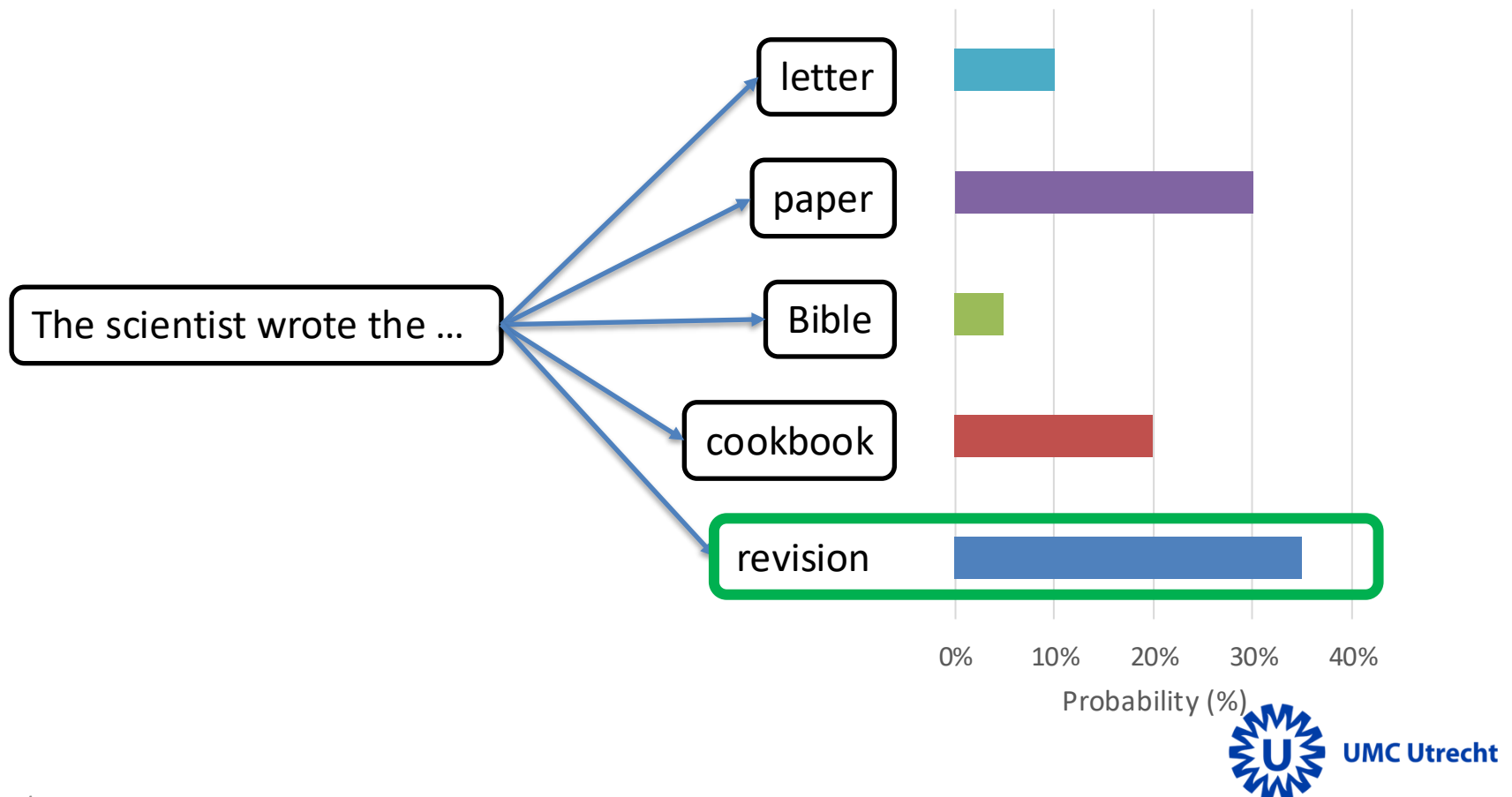
What are LLMs?

What do they do: **Next-word-prediction**



What are LLMs?

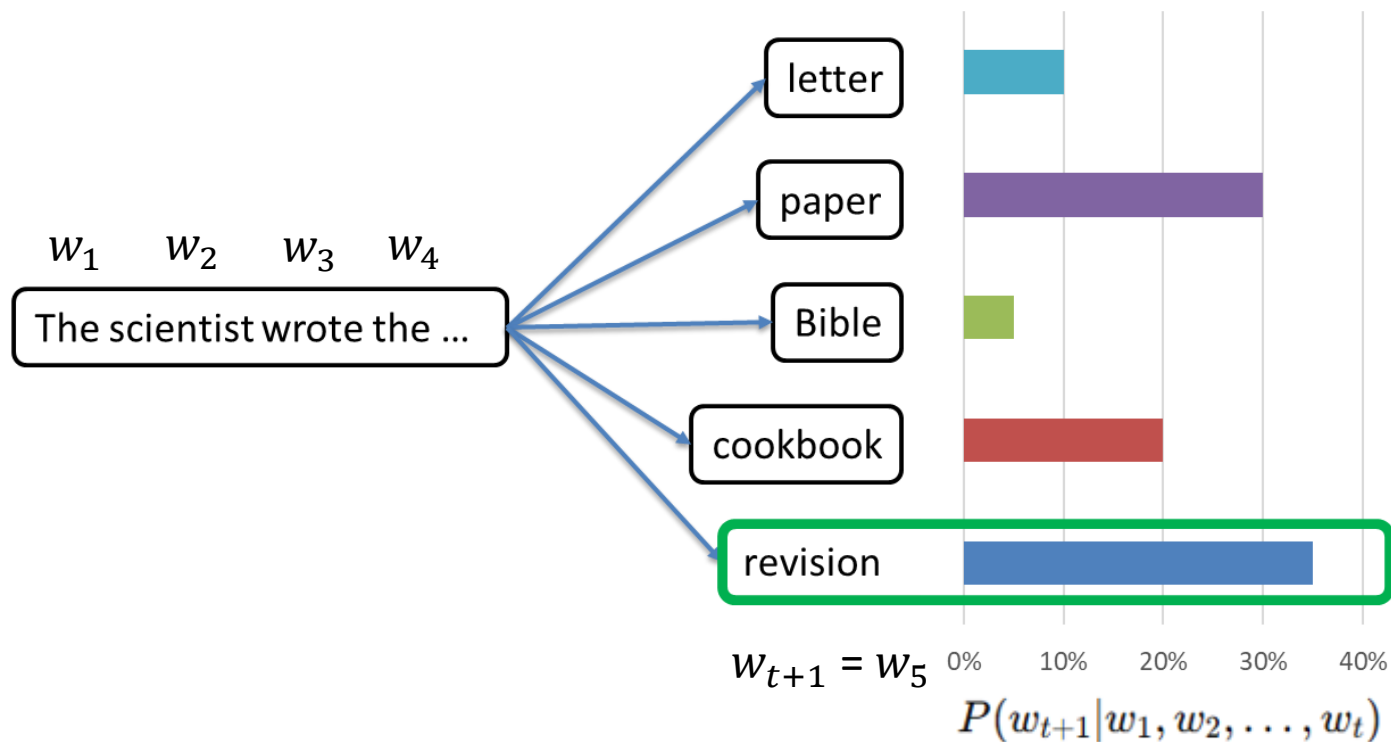
What do they do: **Next-word-prediction**



What are LLMs?

What do they do: **Next-word-prediction**

"The process of estimating the conditional probability distribution $P(w_{t+1}|w_1, w_2, \dots, w_t)$, where $w_1, w_2, \dots, w_t$ are the preceding words in a sequence, and w_{t+1} represents the next word to be predicted."



How Do LLMs Work

How Do LLMs Work



DATA:

LOTS OF TRAINING ON LARGE DATASETS.



MODEL:

TRANSFORMERS AND THE ATTENTION MECHANISM



PREDICTION:

PREDICTS THE NEXT TOKEN BASED ON CONTEXT.

How Do LLMs Work: Model



DATA:

LOTS OF TRAINING ON LARGE DATASETS.



MODEL:

TRANSFORMERS AND THE ATTENTION MECHANISM

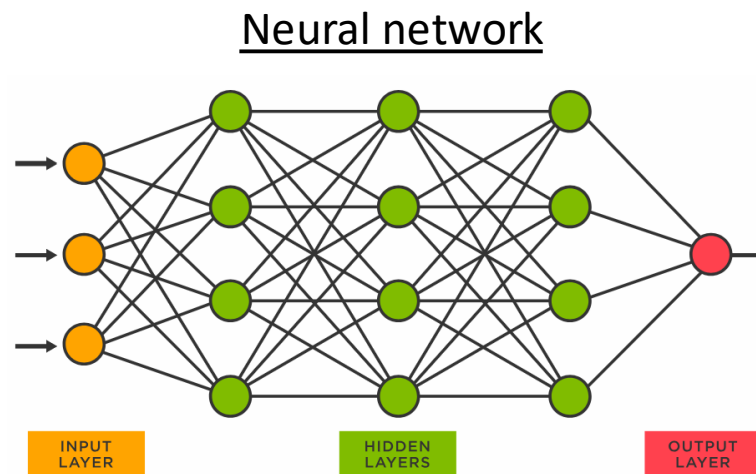


PREDICTION:

PREDICTS THE NEXT TOKEN BASED ON CONTEXT.

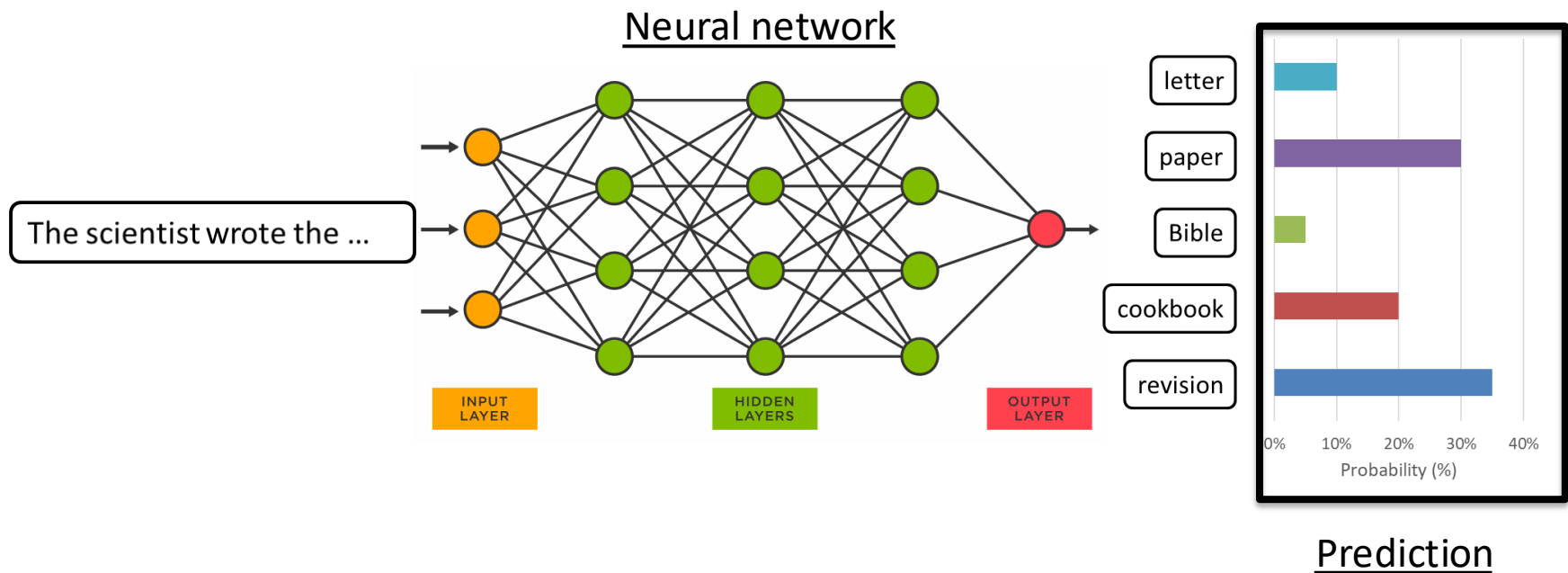
How Do LLMs Work: Model

Neural networks



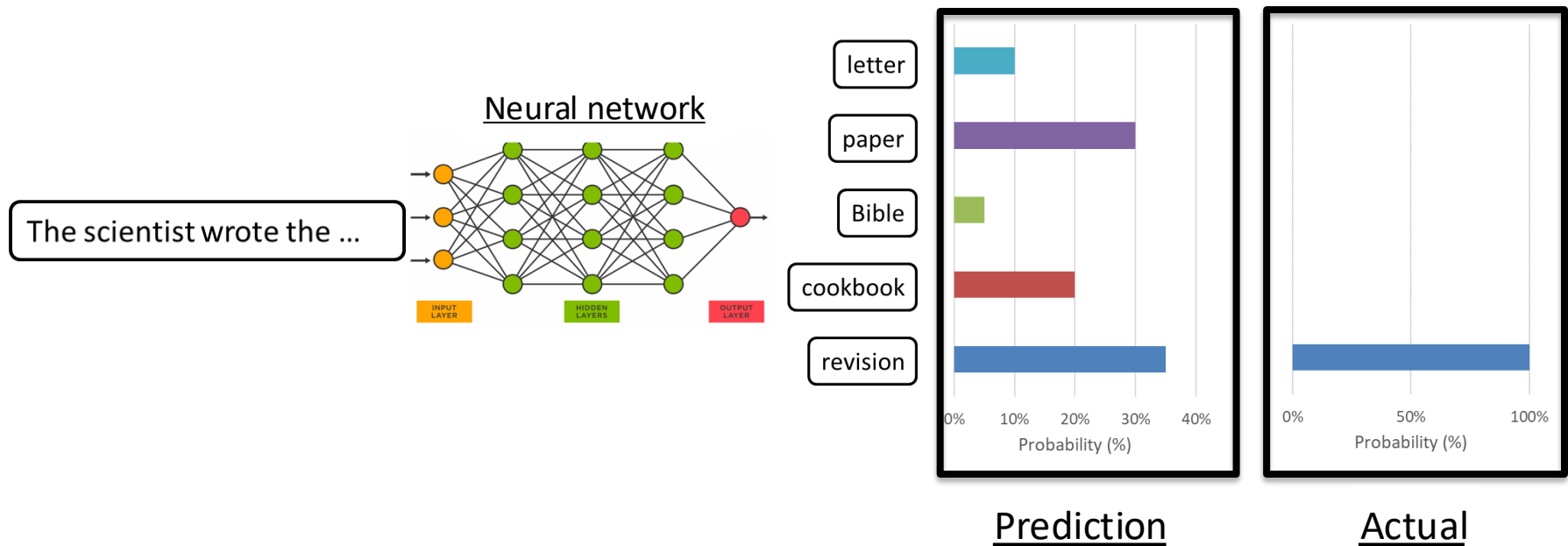
How Do LLMs Work: Model

Neural networks



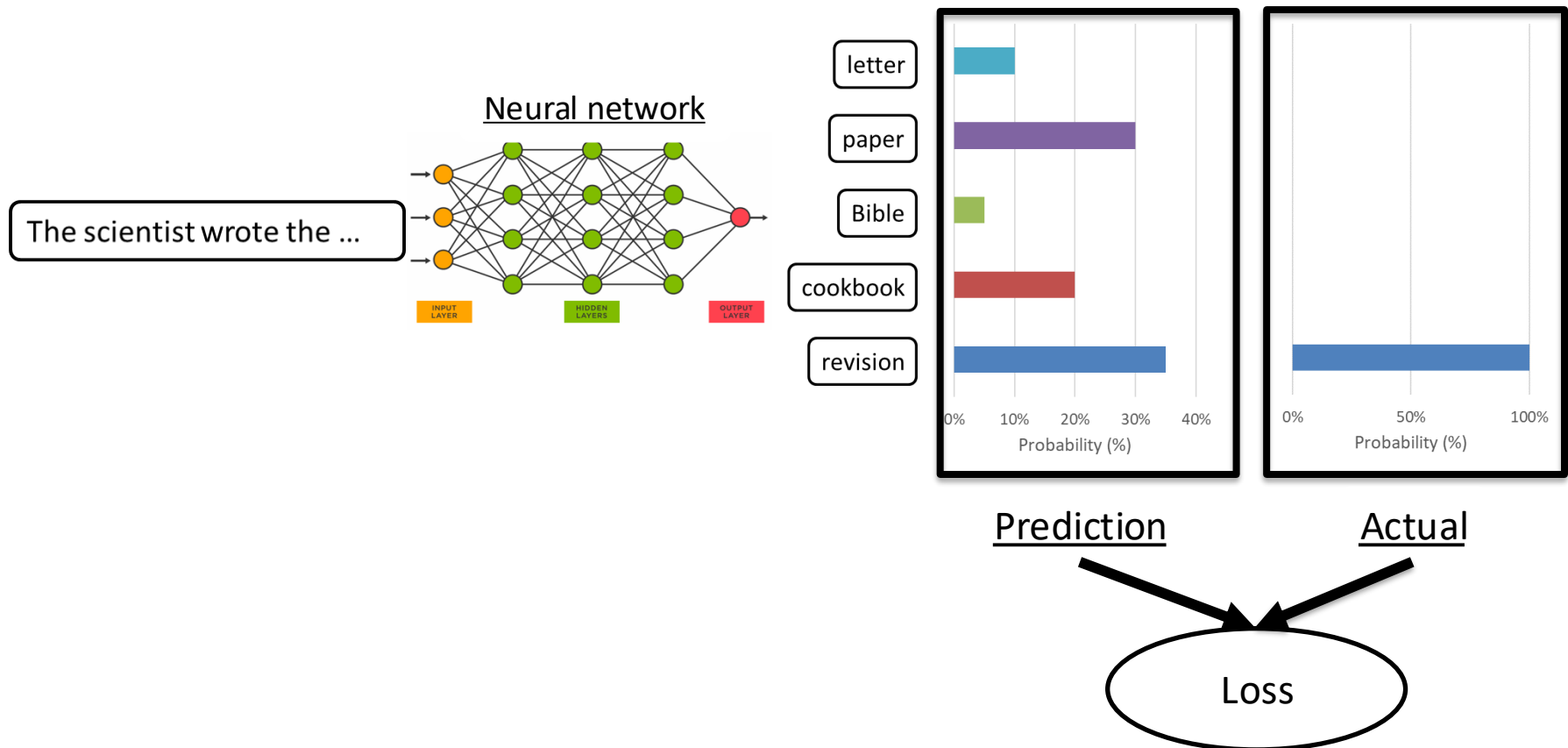
How Do LLMs Work: Model

Neural networks



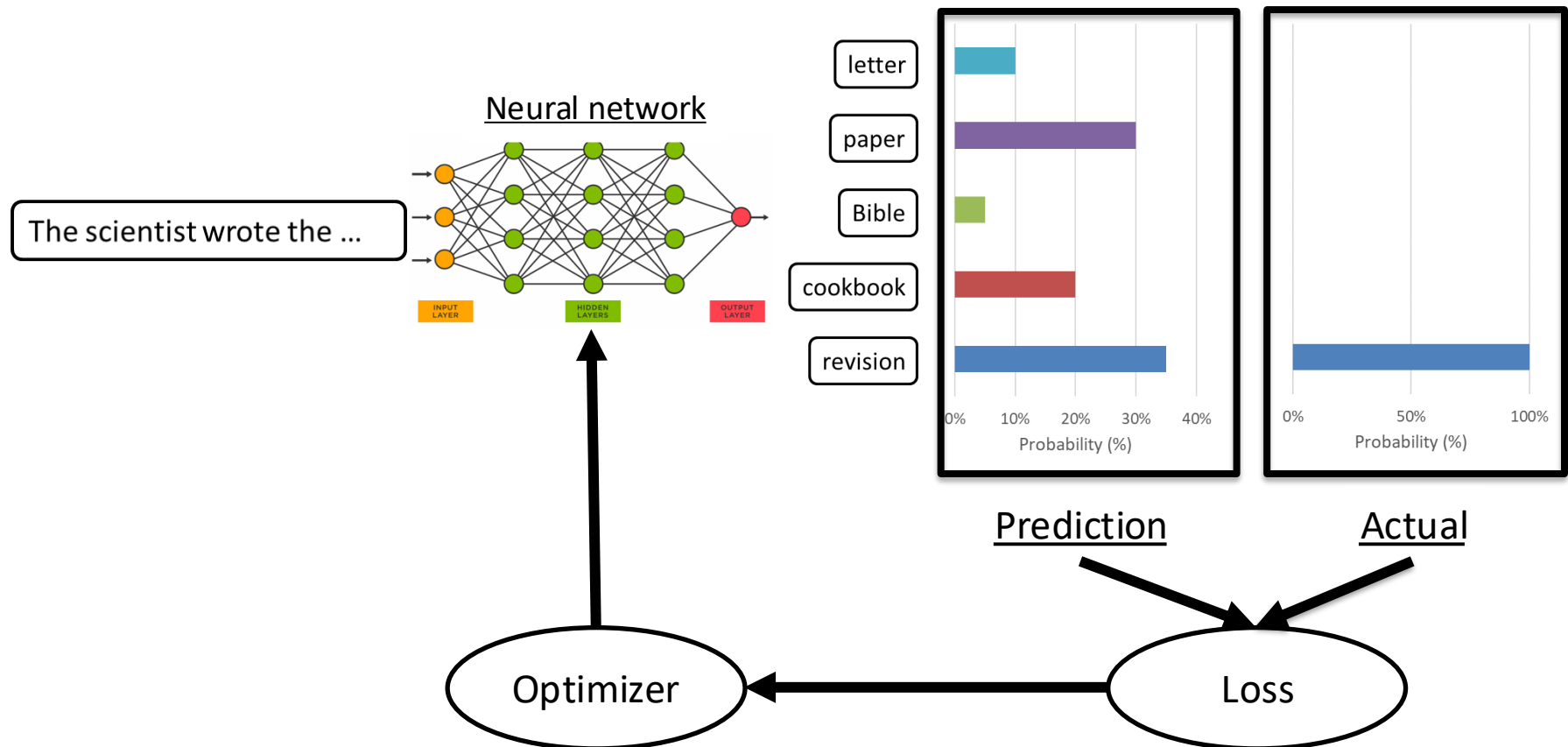
How Do LLMs Work: Model

Neural networks



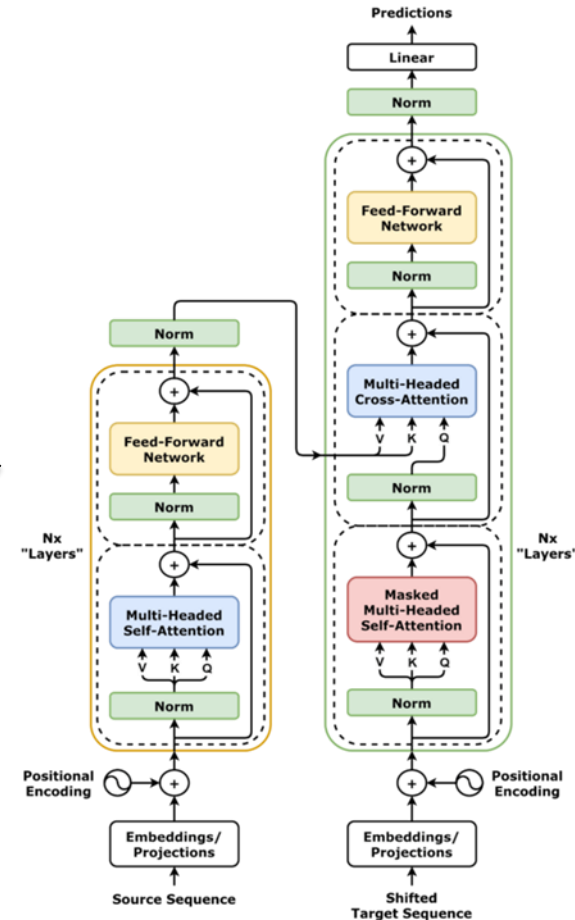
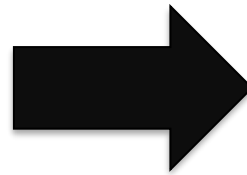
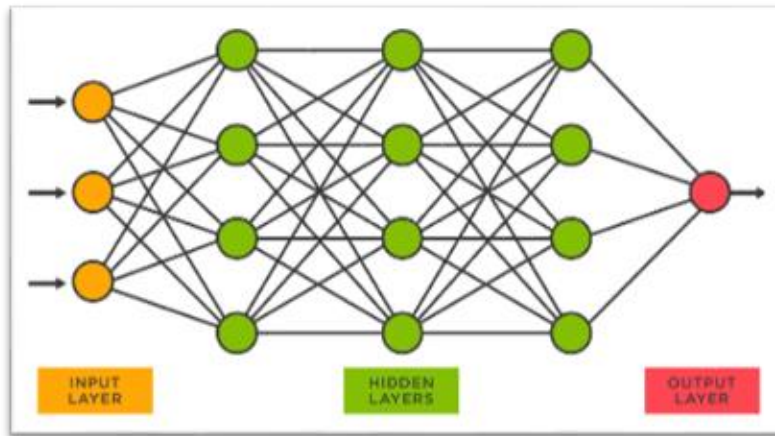
How Do LLMs Work: Model

Neural networks



How Do LLMs Work: Model

Transformer architecture

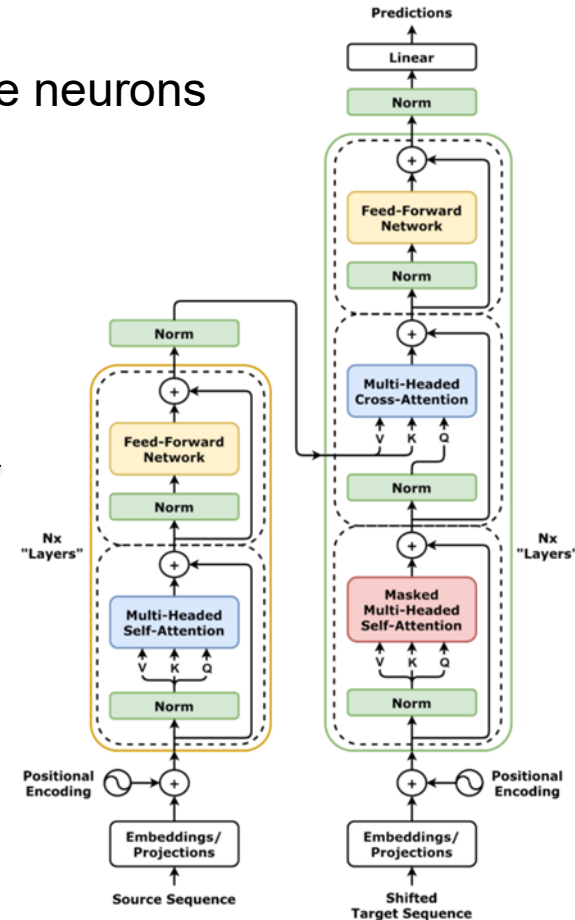
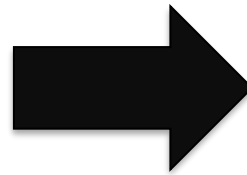
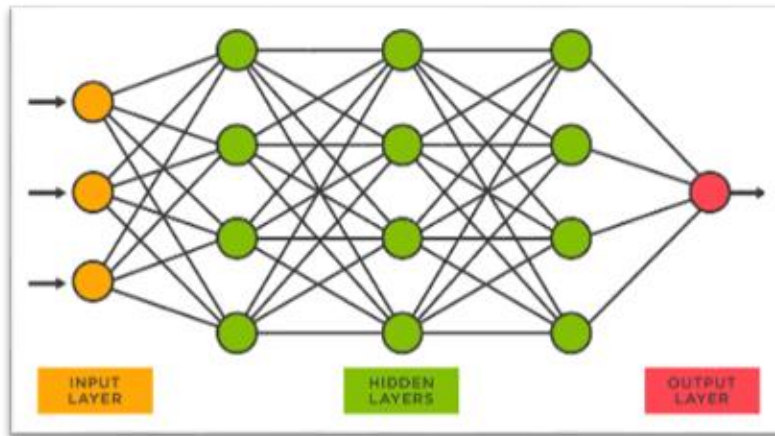


How Do LLMs Work: Model

Transformer architecture



The 'architecture' is the way we connect the neurons

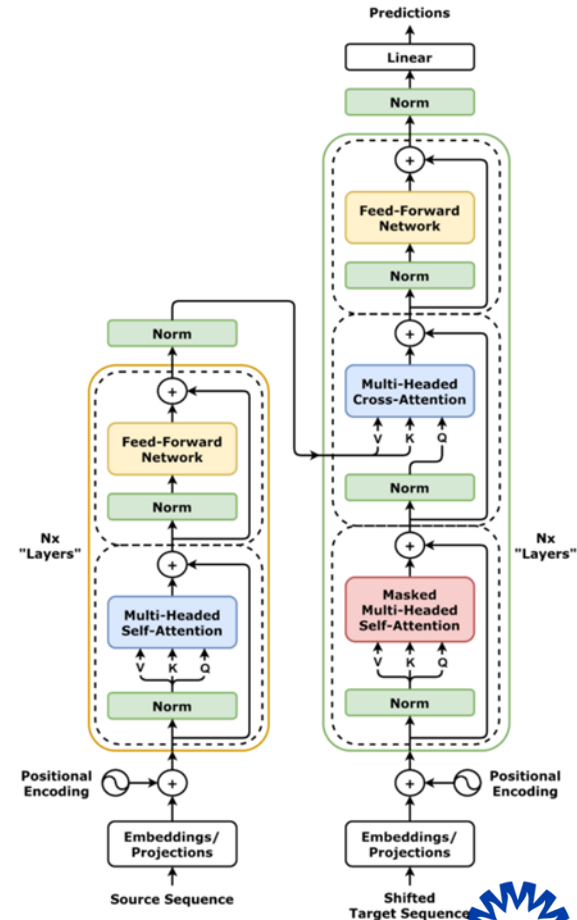


How Do LLMs Work: Model

Transformer architecture



- Let's walk through the architecture step-by-step...



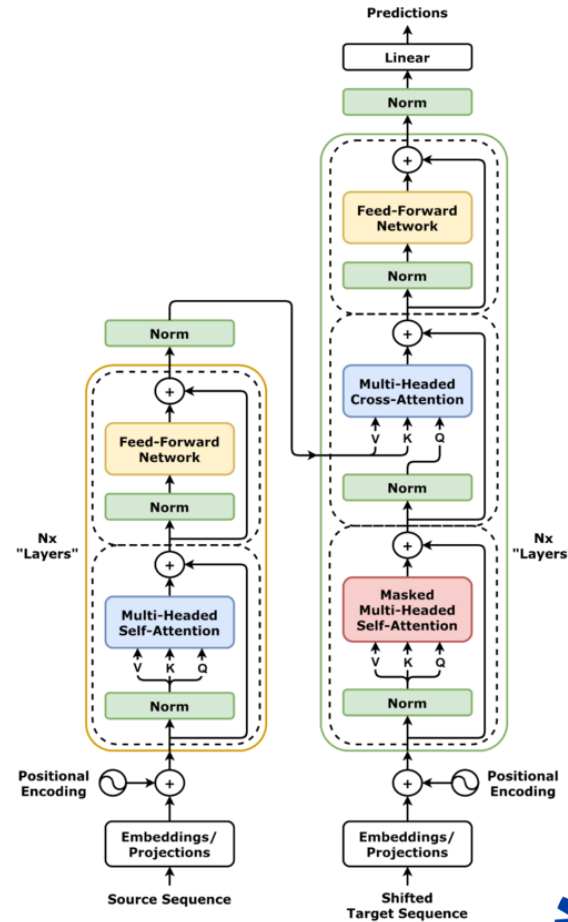
How Do LLMs Work: Model

Transformer architecture



- Let's walk through the architecture step-by-step...

Just kidding!



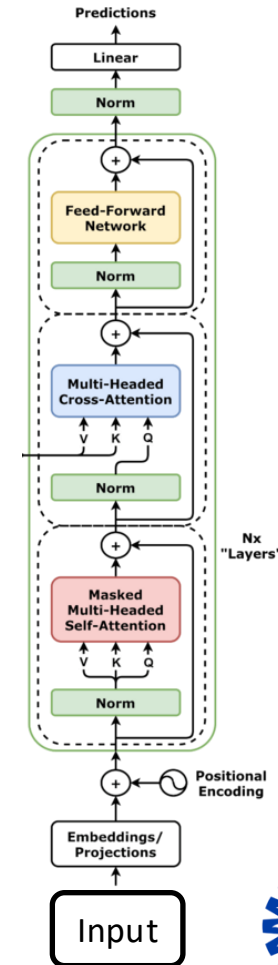
How Do LLMs Work: Model

Transformer architecture – decoder only



- Let's walk through the architecture step-by-step...

Just kidding!

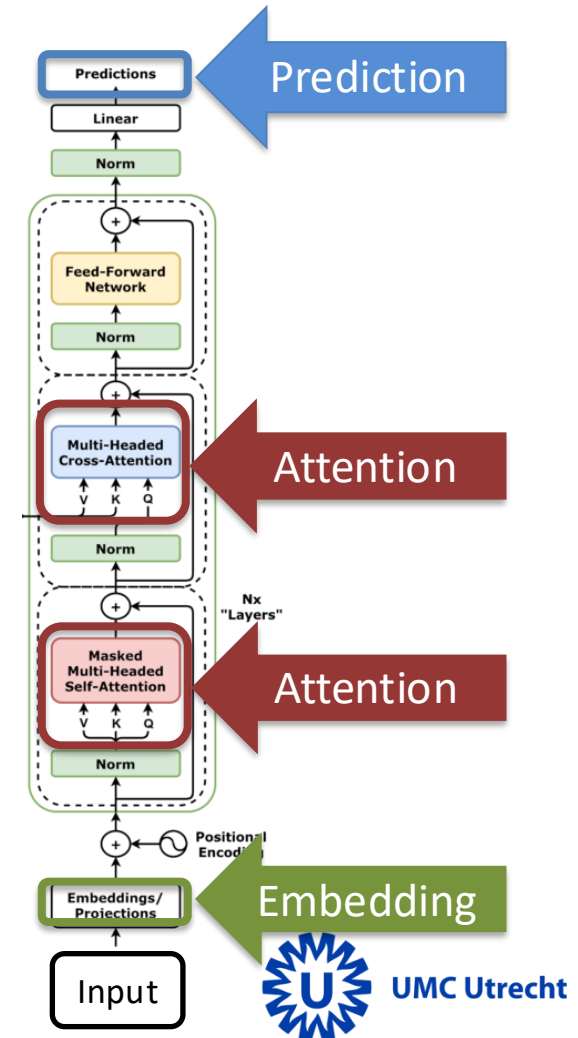


How Do LLMs Work: Model

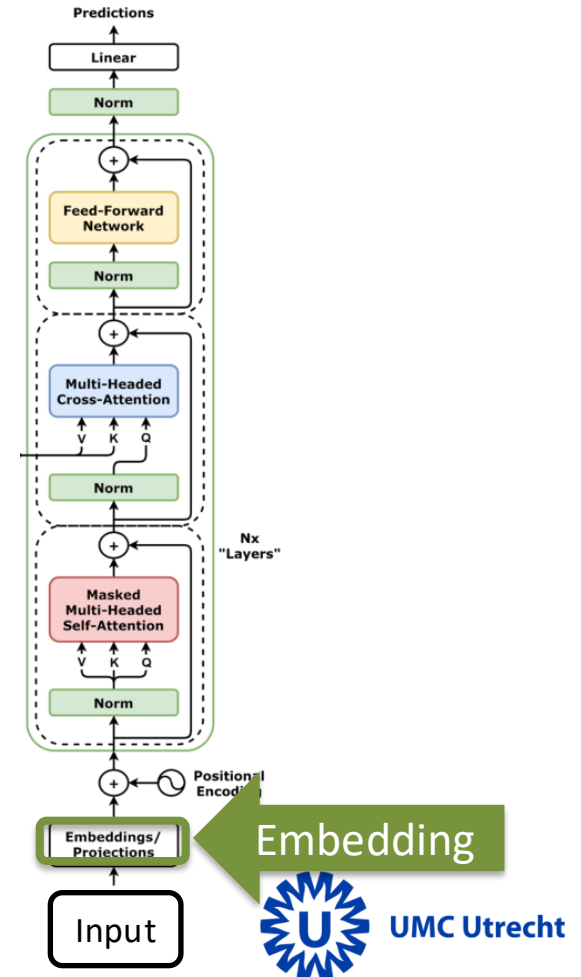
Transformer architecture



- **Embedding**
- **Attention**
- **Prediction**



Embedding

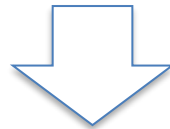


How Do LLMs Work: Model

Embedding: Words to tokens



Short words each get their own token, but exquisite, convoluted, or infrequent words are partitioned into diminutive segments.



Tokenizer

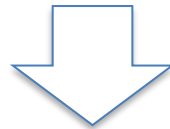
Short words each get their own token, but exquisite, convoluted, or infrequent words are partitioned into diminutive segments.

How Do LLMs Work: Model

Embedding: Words to token IDs

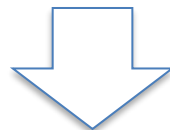


Short words each get their own token, but exquisite, convoluted, or infrequent words are partitioned into diminutive segments.



Tokenizer

Short words each get their own token, but exquisite, convoluted, or infrequent words are partitioned into diminutive segments.



Map to ID

[15900, 6391, 2454, 717, 1043, 2316, 6602, 11, 889, 67557, 11, 191866, 26369, 11, 503, 5603, 191346, 6391, 553, 31512, 295, 1511, 33870, 143209, 33749, 13]

How Do LLMs Work: **Model**

Embedding: **Token IDs to embeddings**



The king loves the queen

The king loves the queen

[976, 13793, 19620, 290, 42300]

How Do LLMs Work: Model

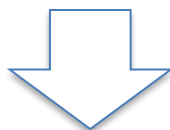
Embedding: **Token IDs to embeddings**



The king loves the queen

The king loves the queen

[976, 13793, 19620, 290, 42300]



Vector embedding

1	0.02	0.93	0.04	0.02	0.94
2	0.01	0.80	0.40	0.01	0.21
3	0.01	0.21	0.60	0.01	0.83
N	...	...	...	...	...

For example: N = 4096 for Llama 3



UMC Utrecht

How Do LLMs Work: Model

Embedding: **Embeddings are vectors**



	The	King	Loves	The	Queen
...	...	...	...	...	...
Royalness	0.02	0.93	0.04	0.02	0.94
Menliness	0.01	0.80	0.40	0.01	0.21
Womanliness	0.01	0.21	0.60	0.01	0.83
...	...	...	...	...	...

How Do LLMs Work: Model

Embedding: **Embeddings are vectors**

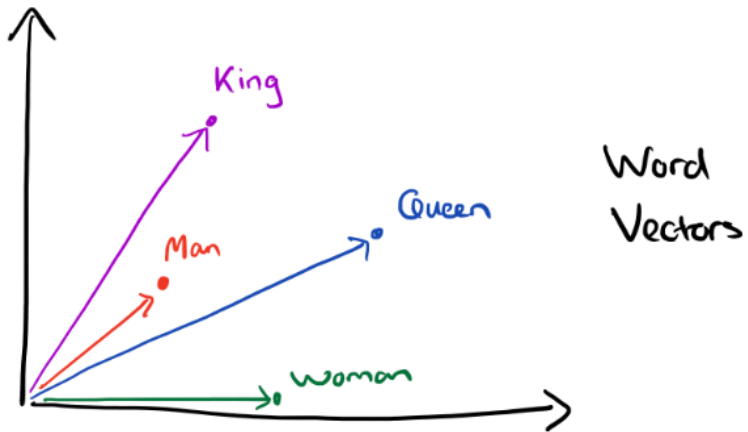


	The	King	Loves	The	Queen
...	...	...	...	...	...
Royalness	0.02	0.93	0.04	0.02	0.94
Menliness	0.01	0.80	0.40	0.01	0.21
Womanliness	0.01	0.21	0.60	0.01	0.83
...	...	...	...	...	...

Source of bias!

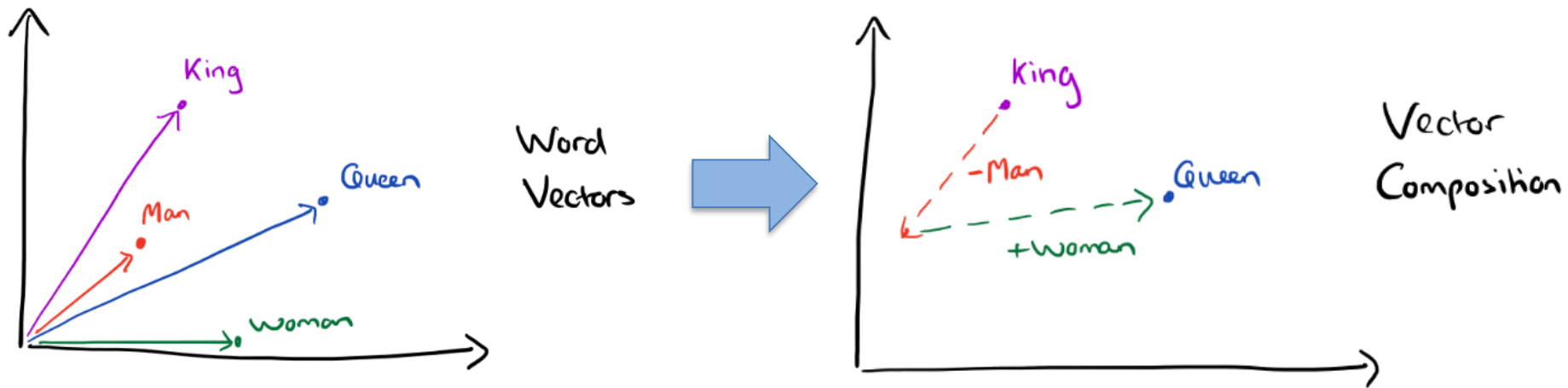
How Do LLMs Work: Model

Embeddings



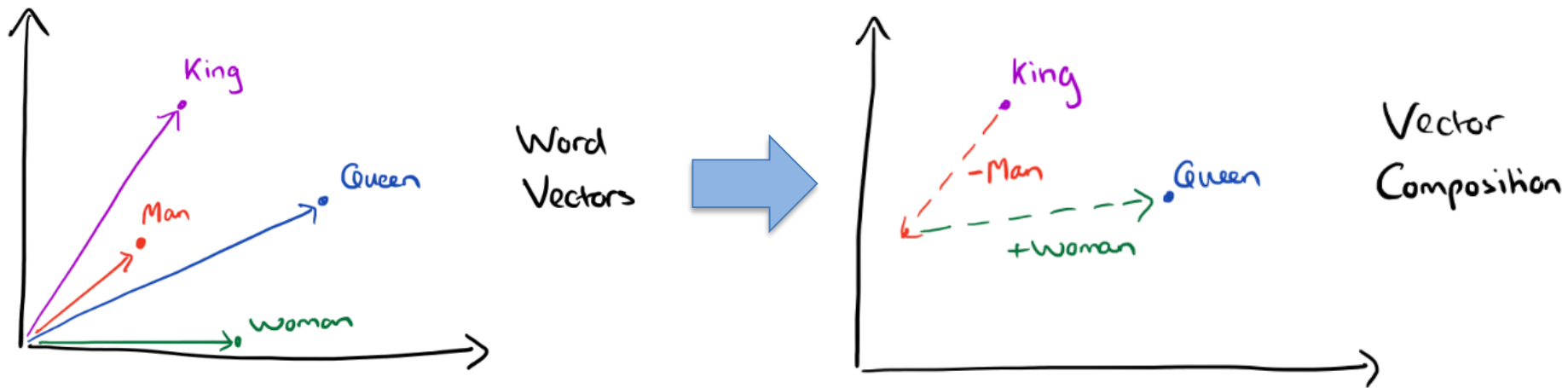
How Do LLMs Work: Model

Embeddings



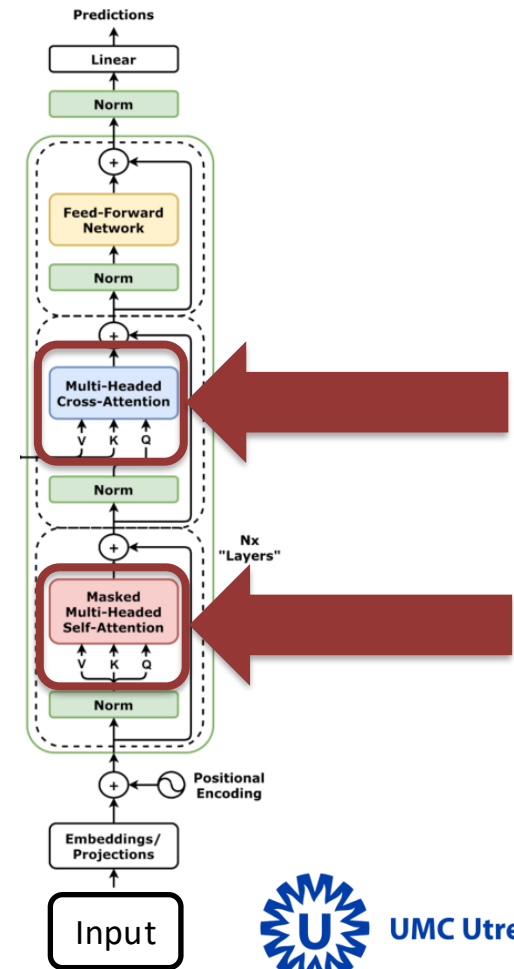
How Do LLMs Work: Model

What does that get us? **Meaning!**



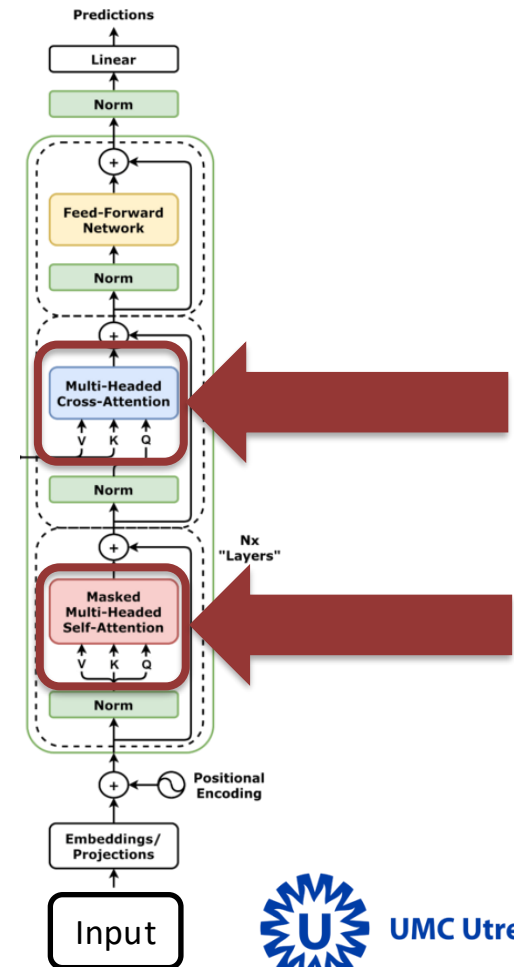
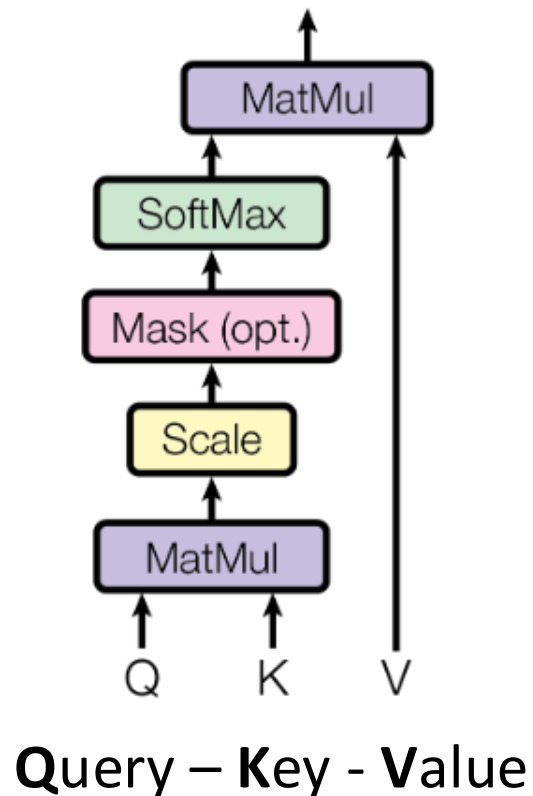
How Do LLMs Work: Model

Attention mechanism



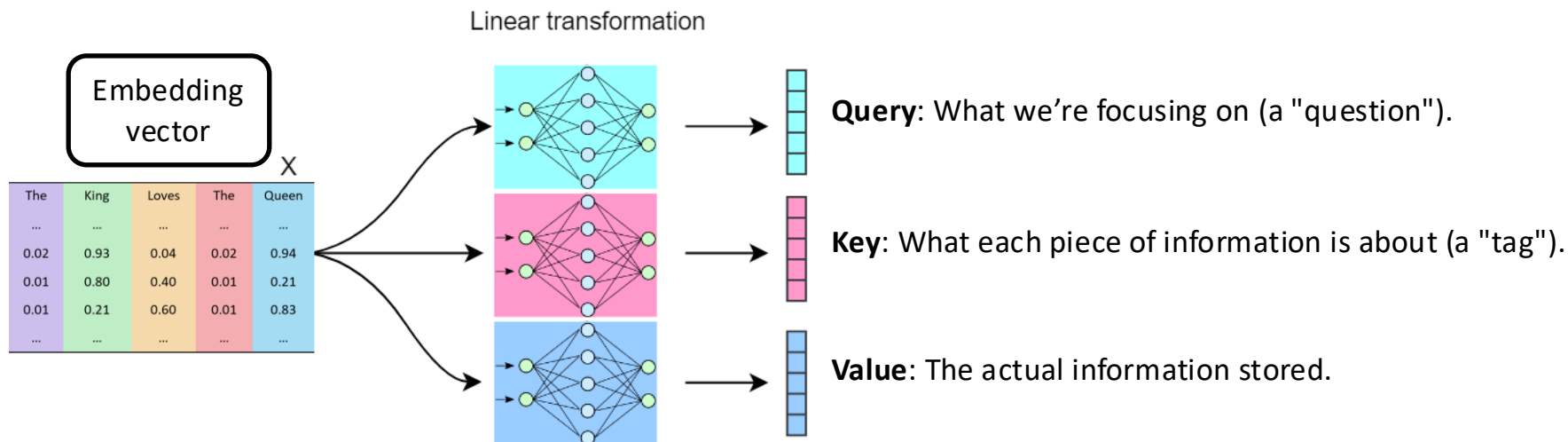
How Do LLMs Work: Model

Query – Key - Value



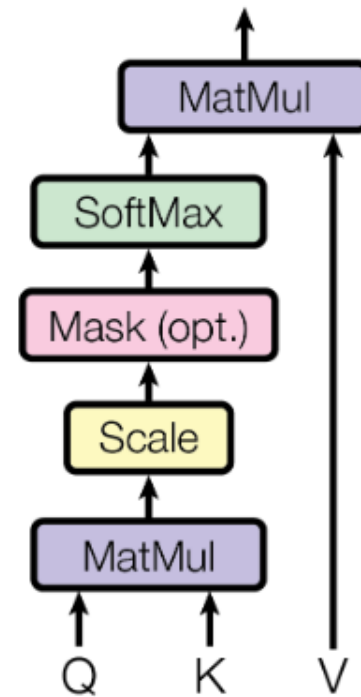
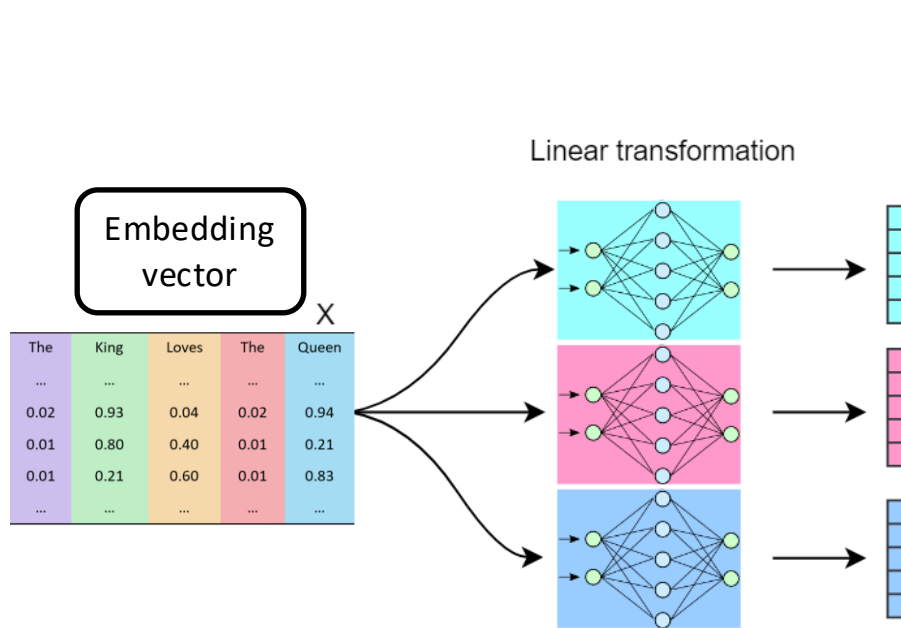
How Do LLMs Work: Model

Query – Key - Value



How Do LLMs Work: Model

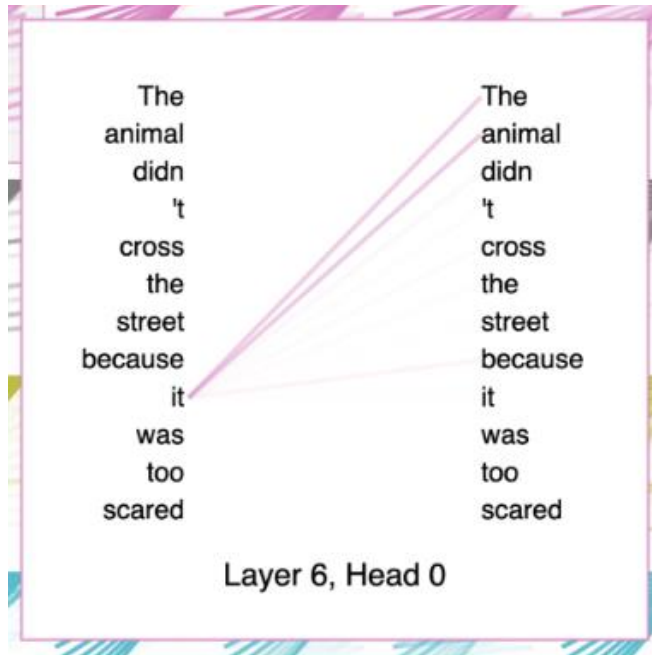
Query – Key - Value



Query – Key - Value

How Do LLMs Work: Model

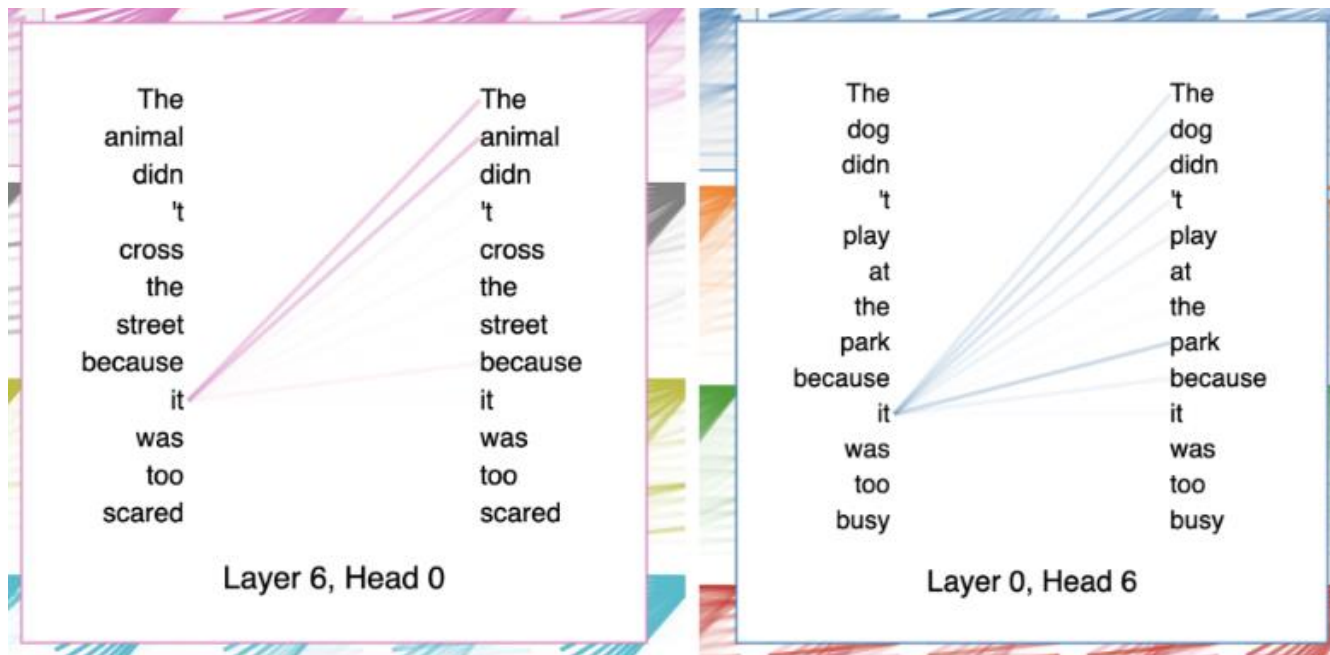
What does that get us? **Attention!**



Source: <https://www.comet.com/site/blog/explainable-ai-for-transformers/>

How Do LLMs Work: Model

What does that get us? **Attention!**



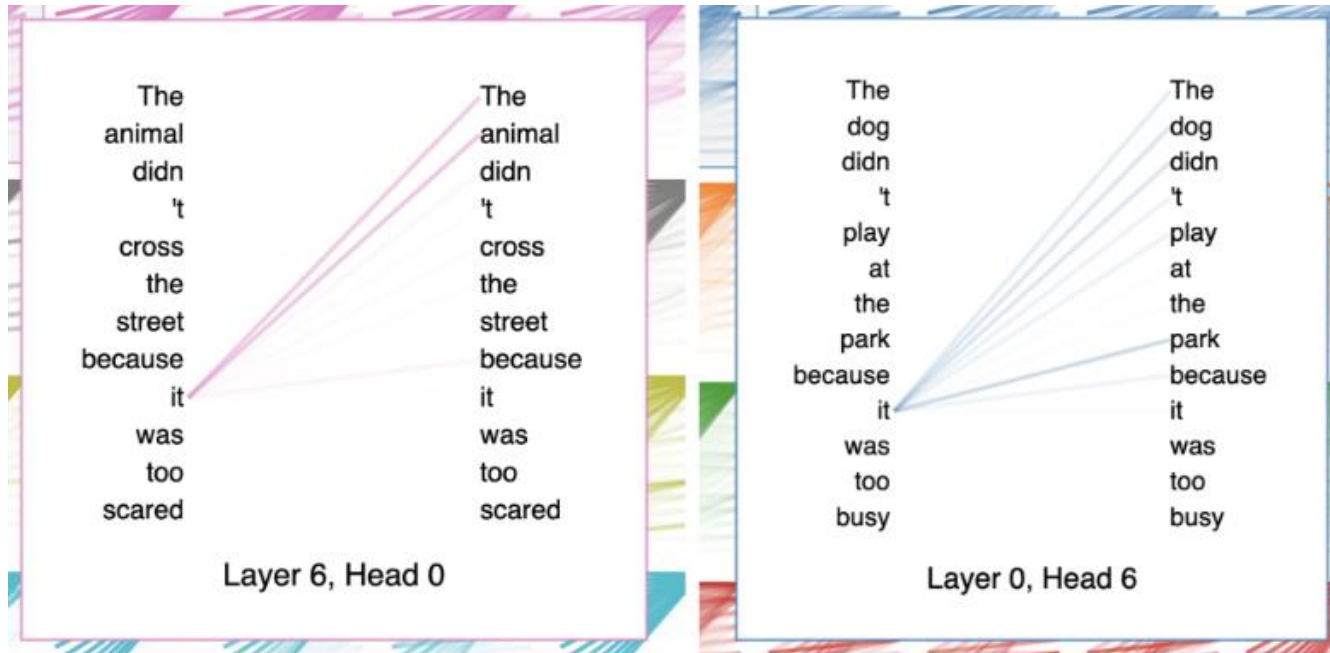
Source: <https://www.comet.com/site/blog/explainable-ai-for-transformers/>



UMC Utrecht

How Do LLMs Work: Model

What does that get us? **Attention!**



Source: <https://www.comet.com/site/blog/explainable-ai-for-transformers/>

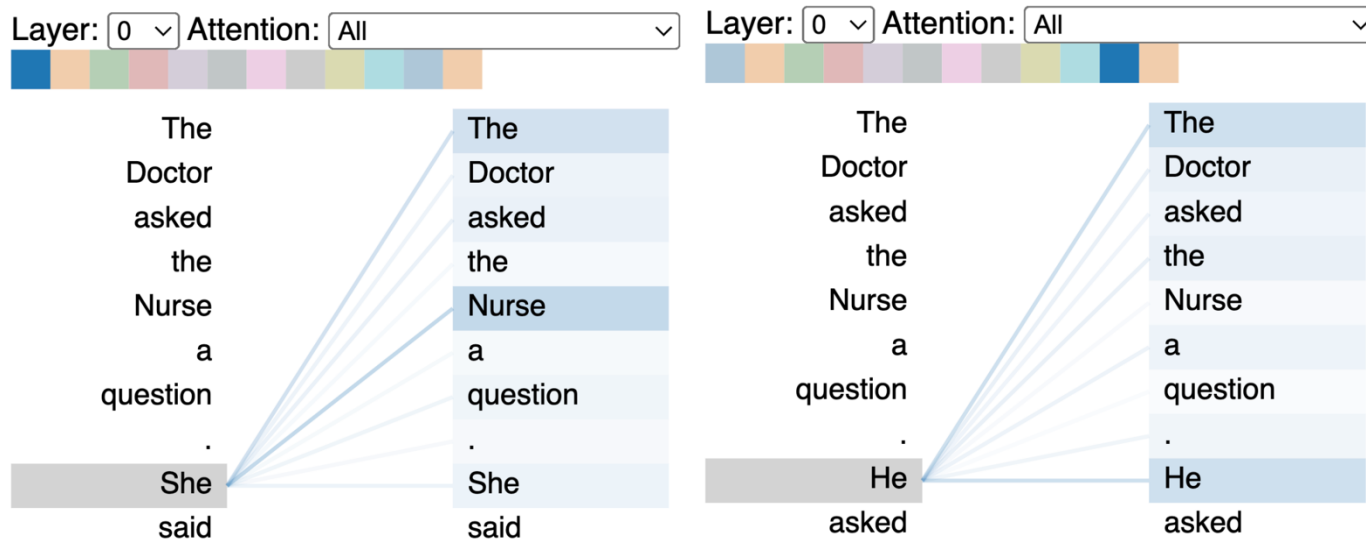
Meaning + Attention = Understanding?

How Do LLMs Work: Model

Risks



Why is this important to know?



Source: <https://www.comet.com/site/blog/explainable-ai-for-transformers/>

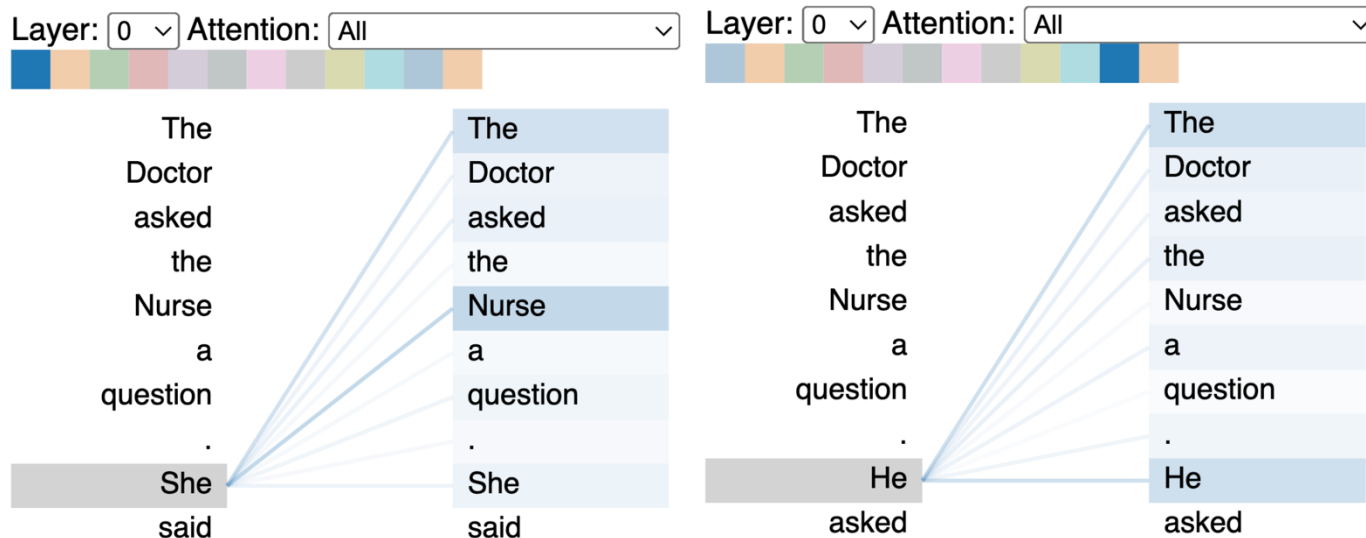
How Do LLMs Work: Model

Risks



Why is this important to know?

To understand the risks you need to understand the sources of error



Source: <https://www.comet.com/site/blog/explainable-ai-for-transformers/>

How Do LLMs Work

The whole cycle



DATA:

LOTS OF TRAINING ON LARGE DATASETS.



MODEL:

TRANSFORMERS AND THE ATTENTION MECHANISM



PREDICTION:

PREDICTS THE NEXT TOKEN BASED ON CONTEXT.

How Do LLMs Work

The whole cycle



DATA:

LOTS OF TRAINING ON LARGE DATASETS.



MODEL:

TRANSFORMERS AND THE ATTENTION MECHANISM



PREDICTION:

PREDICTS THE NEXT TOKEN BASED ON CONTEXT.

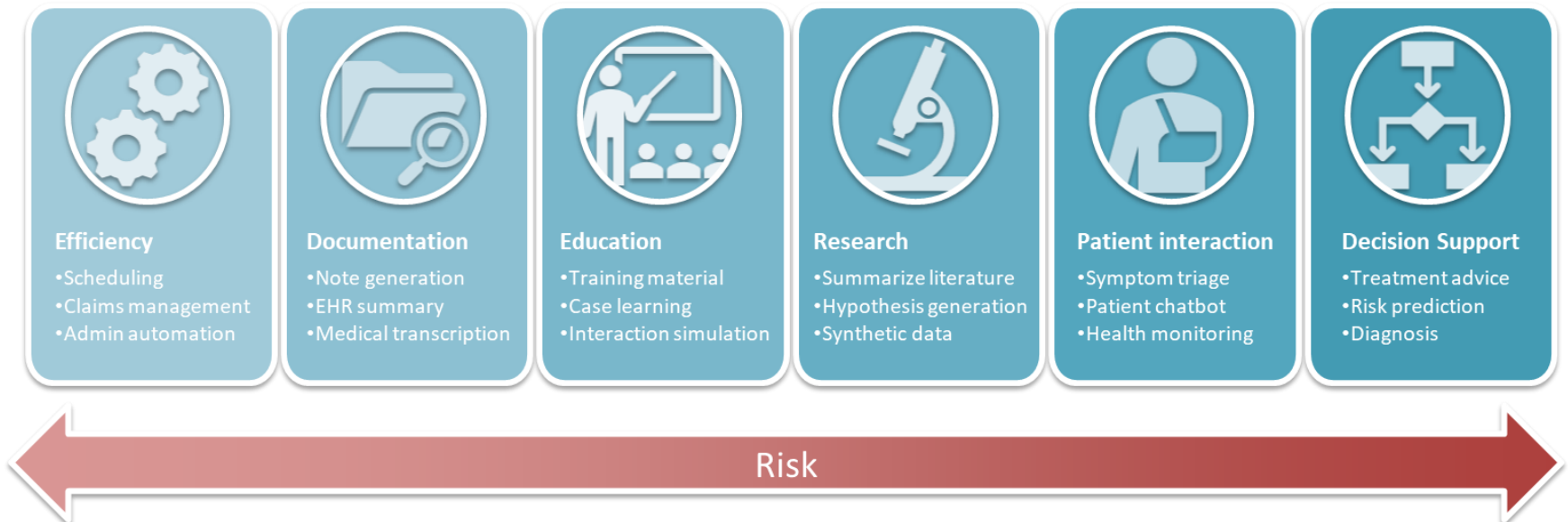


VALIDATION:

DETECT AND MITIGATE ERRORS AND BIAS

LLMs in Healthcare

Applications (sorted by risk)



LLMs in Healthcare

Not everything is text!

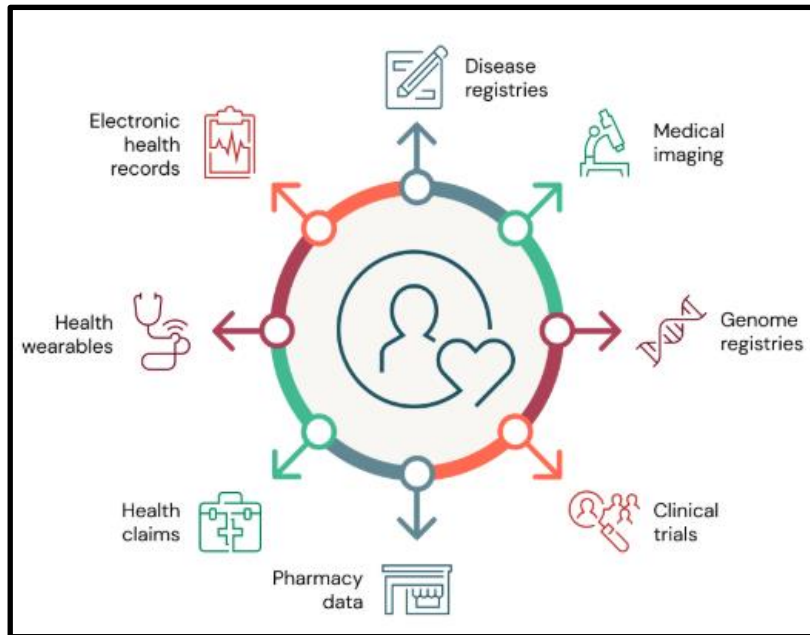


Image from: <https://www.databricks.com/blog/2021/07/19/unlocking-the-power-of-health-data-with-a-modern-data-lakehouse.html>

LLMs in Healthcare

Multi-modal models

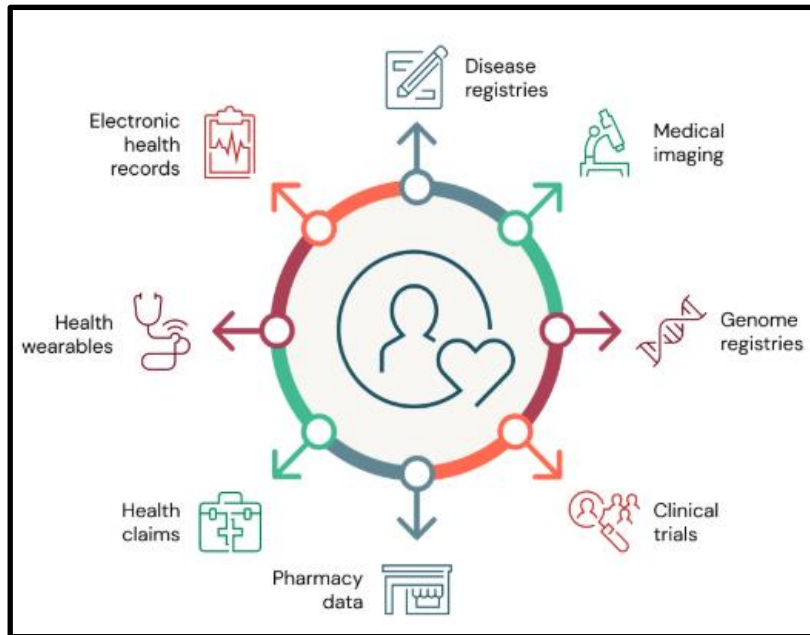


Image from: <https://www.databricks.com/blog/2021/07/19/unlocking-the-power-of-health-data-with-a-modern-data-lakehouse.html>

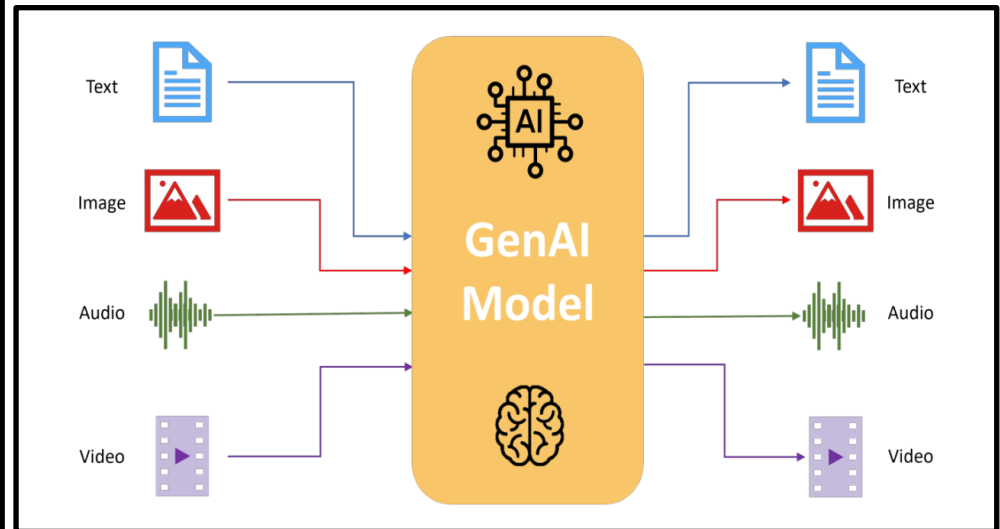


Image from: <https://www.linkedin.com/pulse/multimodal-generative-ai-tarun-sharma-zzf9c/>

LLMs in Healthcare

Big promises, small steps?

Can you write me a report analyzing this chest X-ray?



Findings:

- Devices: None.
- Lungs: No pneumothorax. No substantial pleural effusion. Lungs appear clear.
- Cardiomedial: Normal heart size. Mediastinal contours within normal limits.
- Other: No acute skeletal abnormality.

Impression:

No active disease seen in chest.

Enter a question here

Med-Palm 2: Clinical Multi-modal LLM

Singhal, K., Azizi, S., Tu, T. *et al.* Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023). <https://sites.research.google/med-palm/>

An Introduction to LLMs



Any questions?