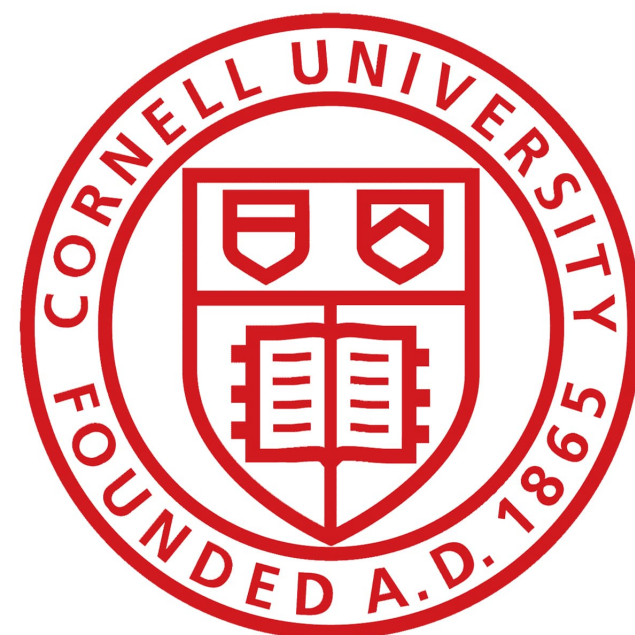


Beyond RNNs: An Introduction to Transformers and LLMs

Sharan Sahu | Stats and Data Sci. PhD | Cornell University



Cornell University®

LLMs and Transformers are everywhere now!

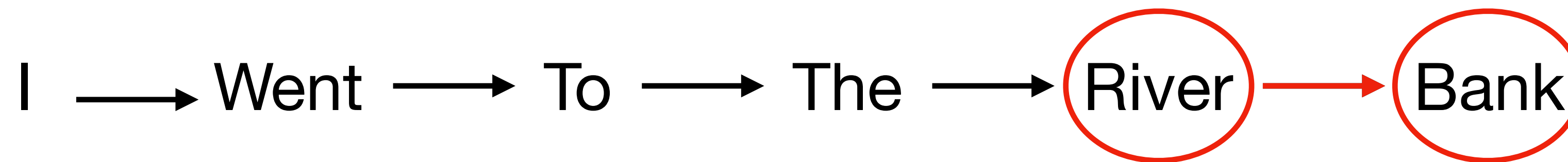


ChatGPT

The Gemini logo, featuring the word "Gemini" in a blue-to-purple gradient font. Above the letter 'i' is a small purple four-pointed star.The Claude logo, featuring an orange starburst icon followed by the word "Claude" in a black serif font.

Recurrent Neural Networks (RNNs)

Language is inherently sequential. The meaning of a word is dependent on the words that precede it

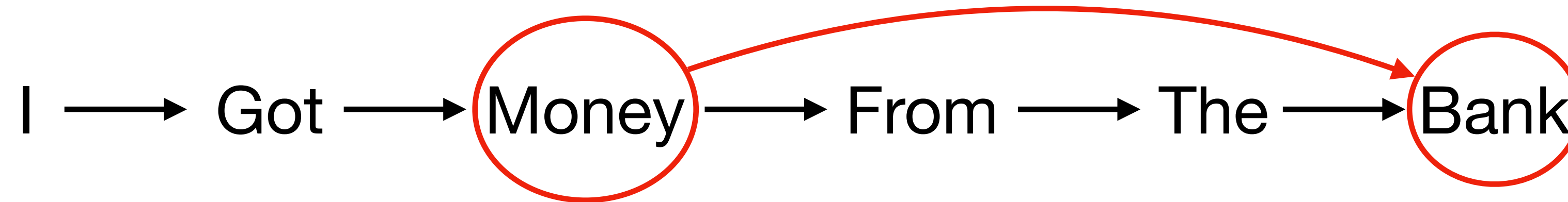


Bank (noun)

Definition: the rising ground bordering a lake, river, or sea or forming the edge of a cut or hollow

Recurrent Neural Networks (RNNs)

Language is inherently sequential. The meaning of a word is dependent on the words that precede it

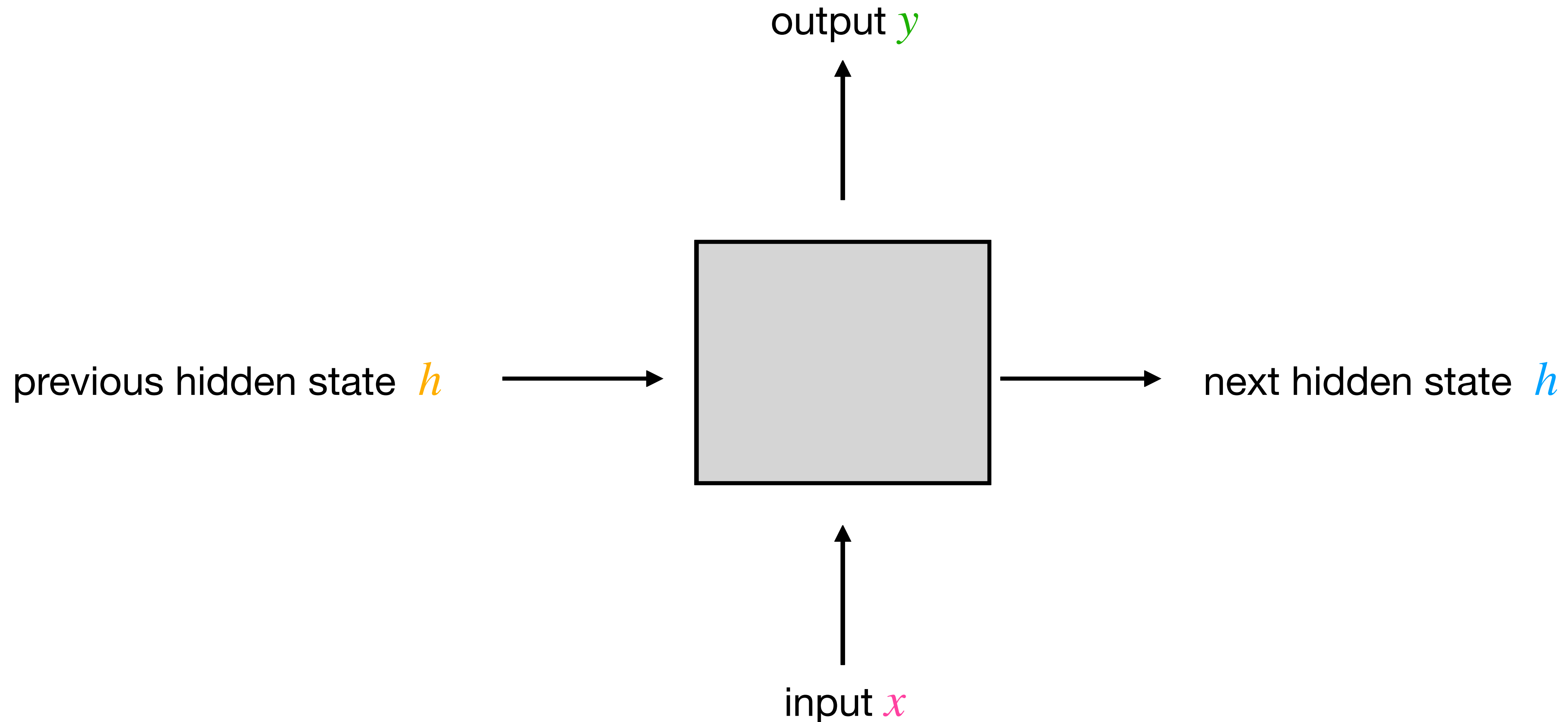


Bank (noun)

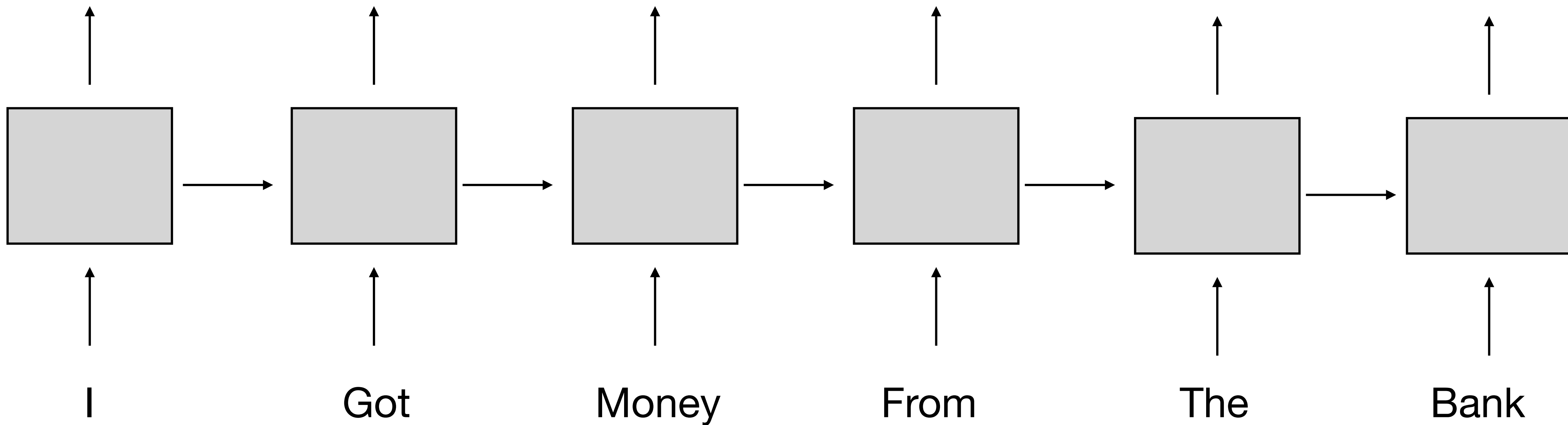
Definition: an establishment for the custody, loan, exchange, or issue of money, for the extension of credit, and for facilitating the transmission of funds

Recurrent Neural Networks (RNNs)

Formally, this is called recurrence. This allows us to build a recurrent neural network



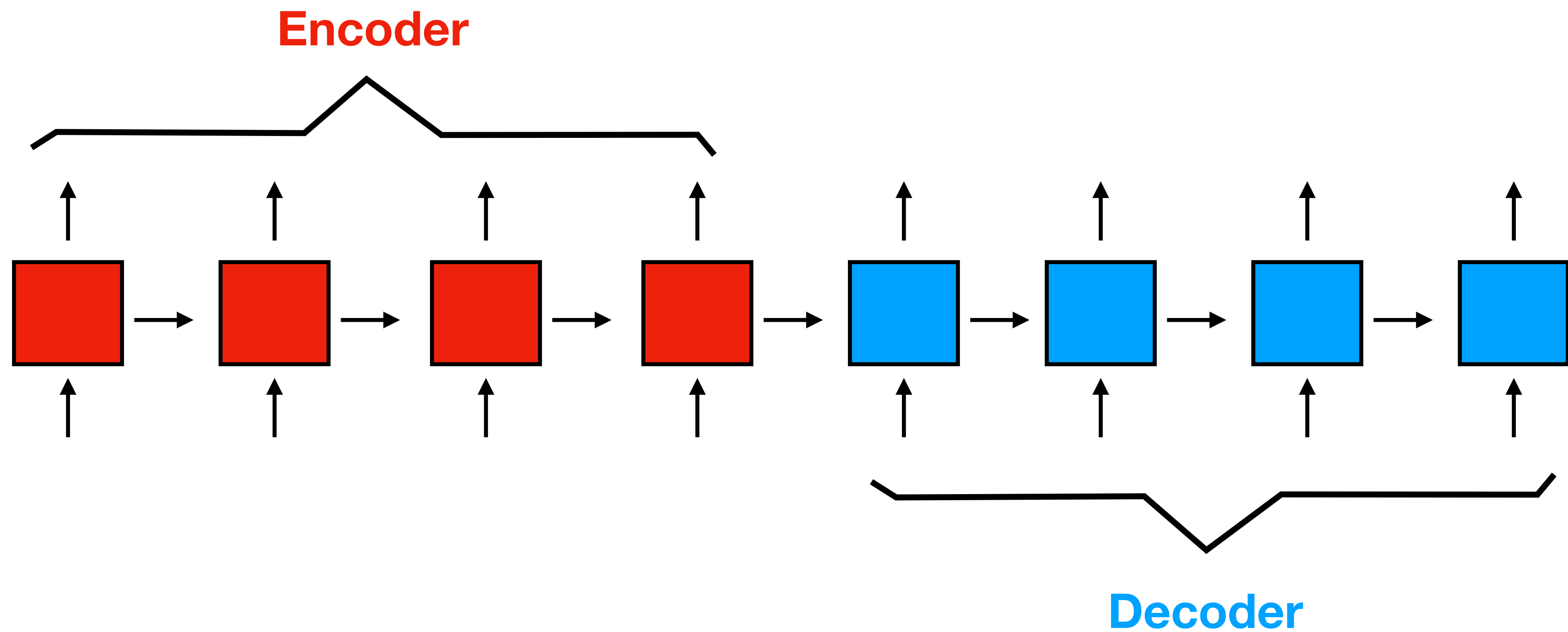
Recurrent Neural Networks (RNNs)



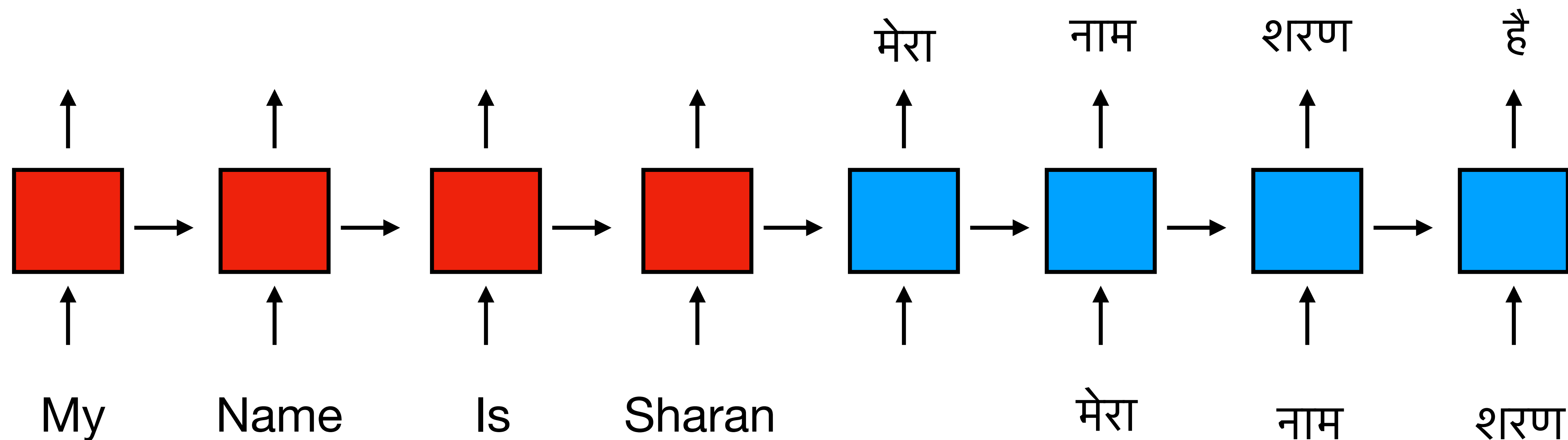
Machine Translation



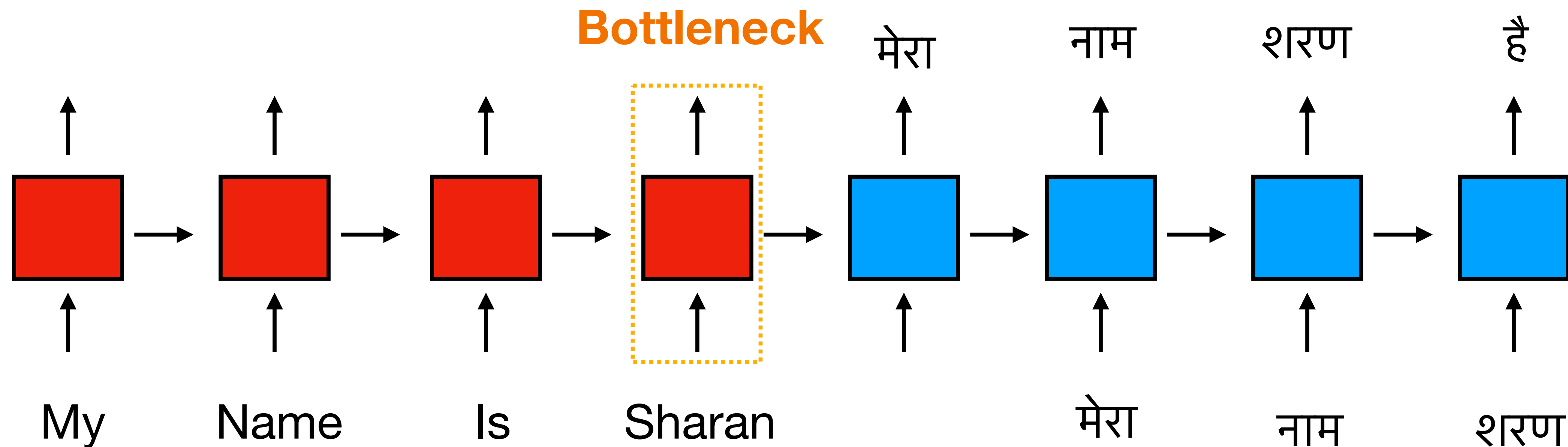
Machine Translation



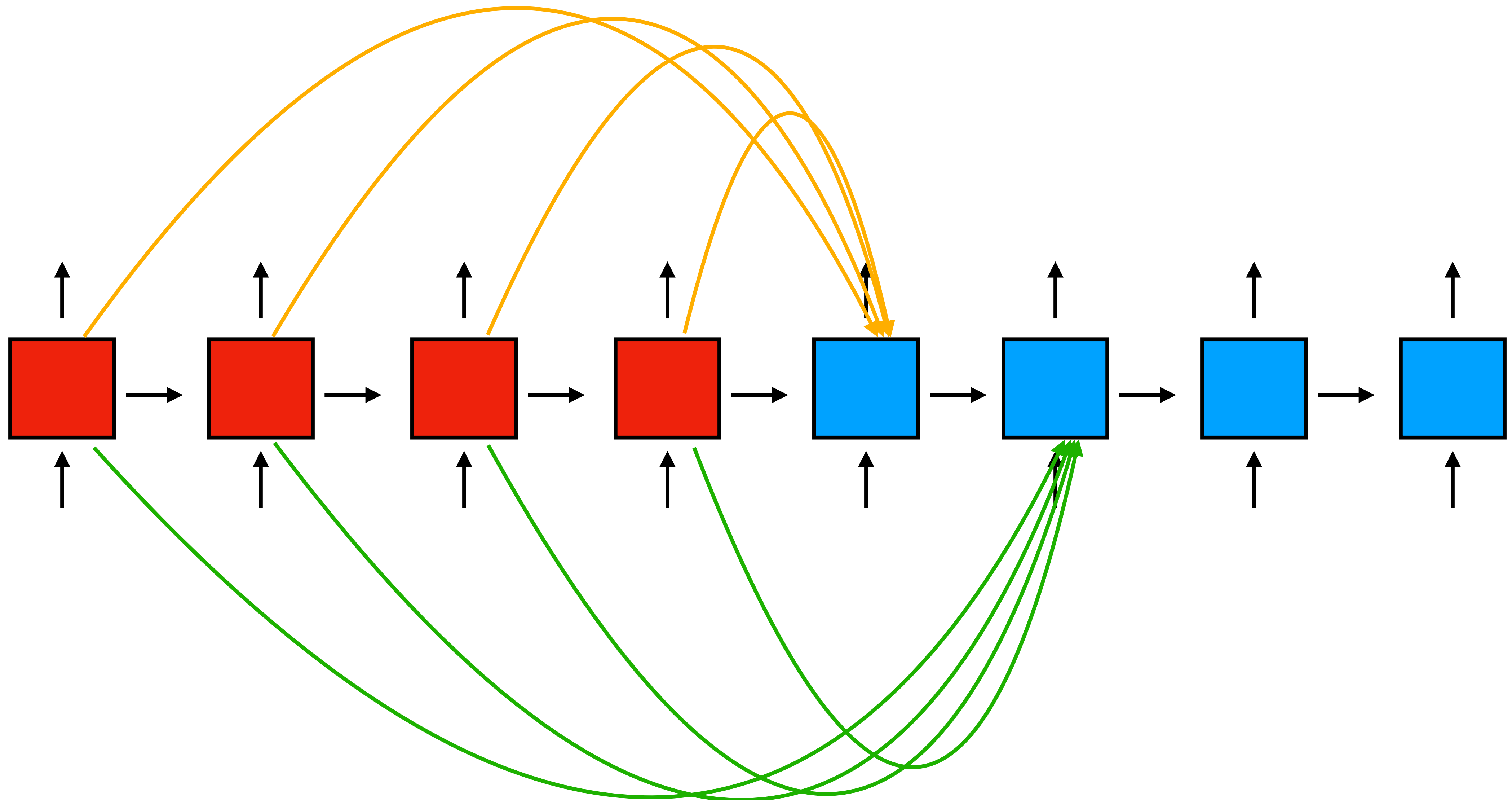
Machine Translation



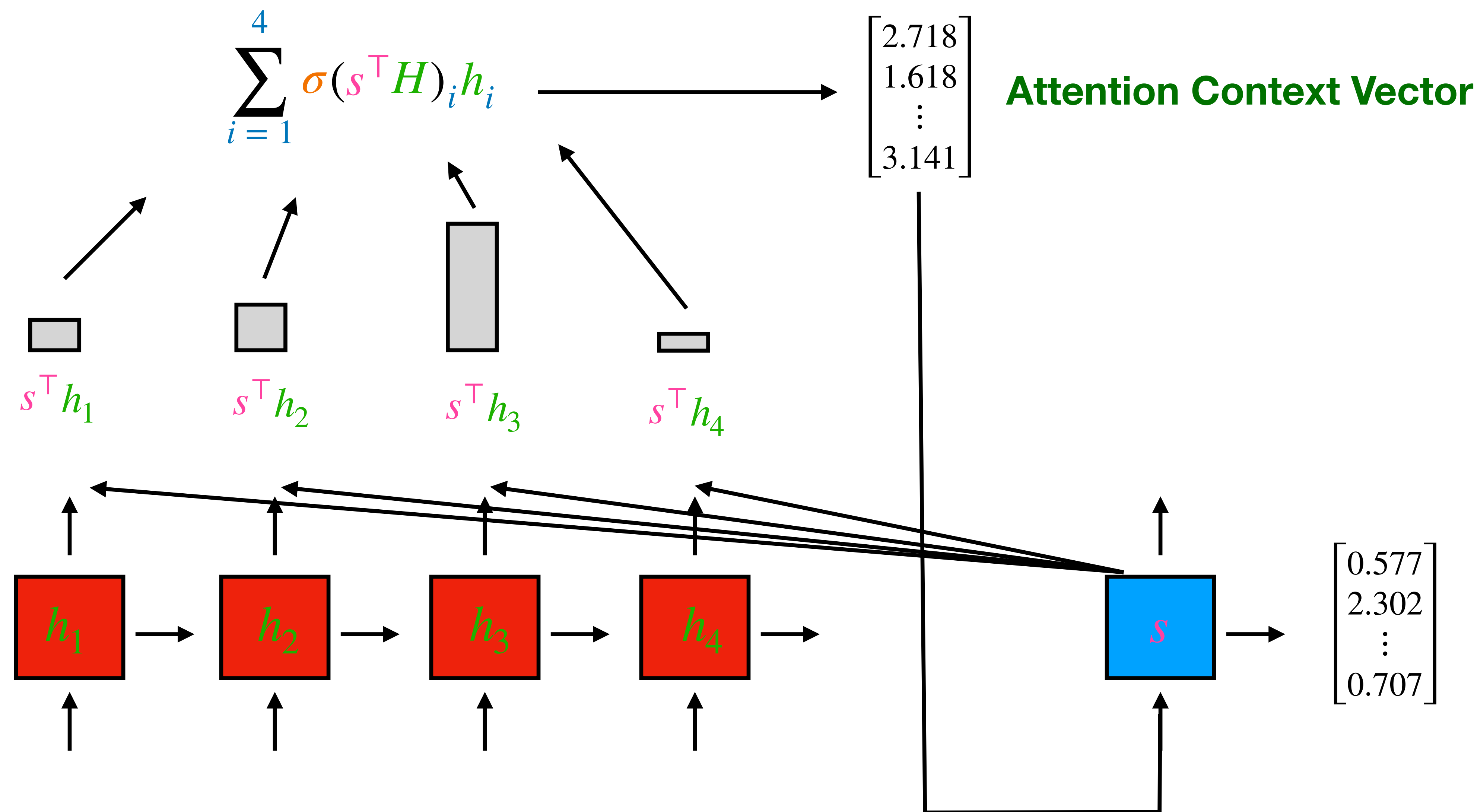
Machine Translation

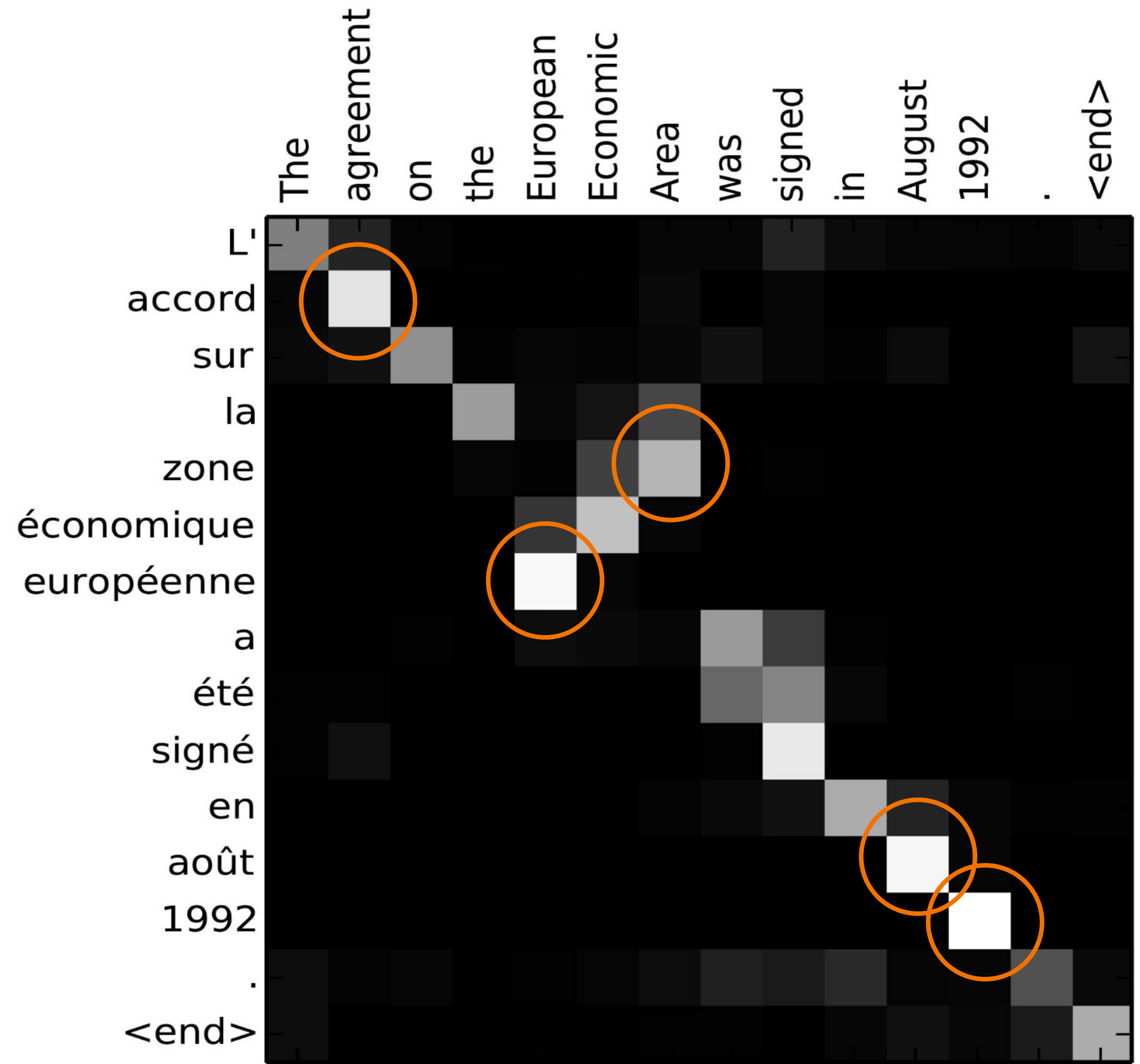


Attention Mechanism



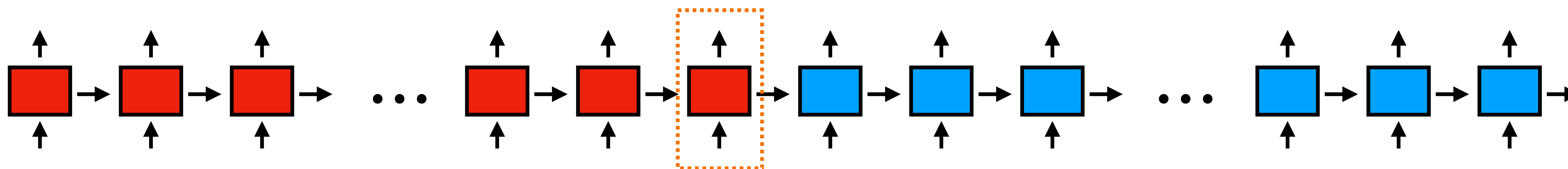
Attention Mechanism





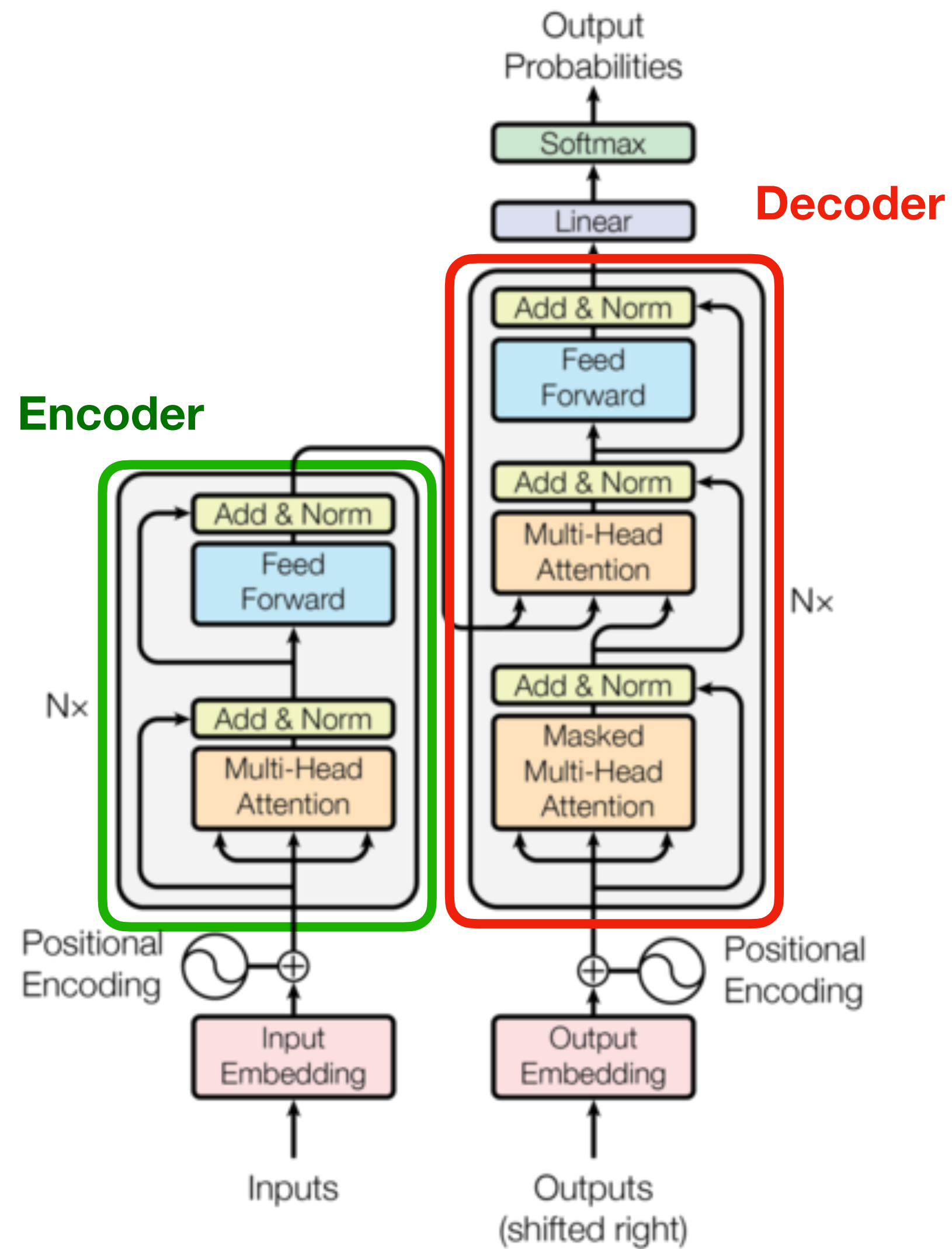
RNNs are bad at handling long range sequences!

It becomes very hard to compress a long sequence of text into a single vector!



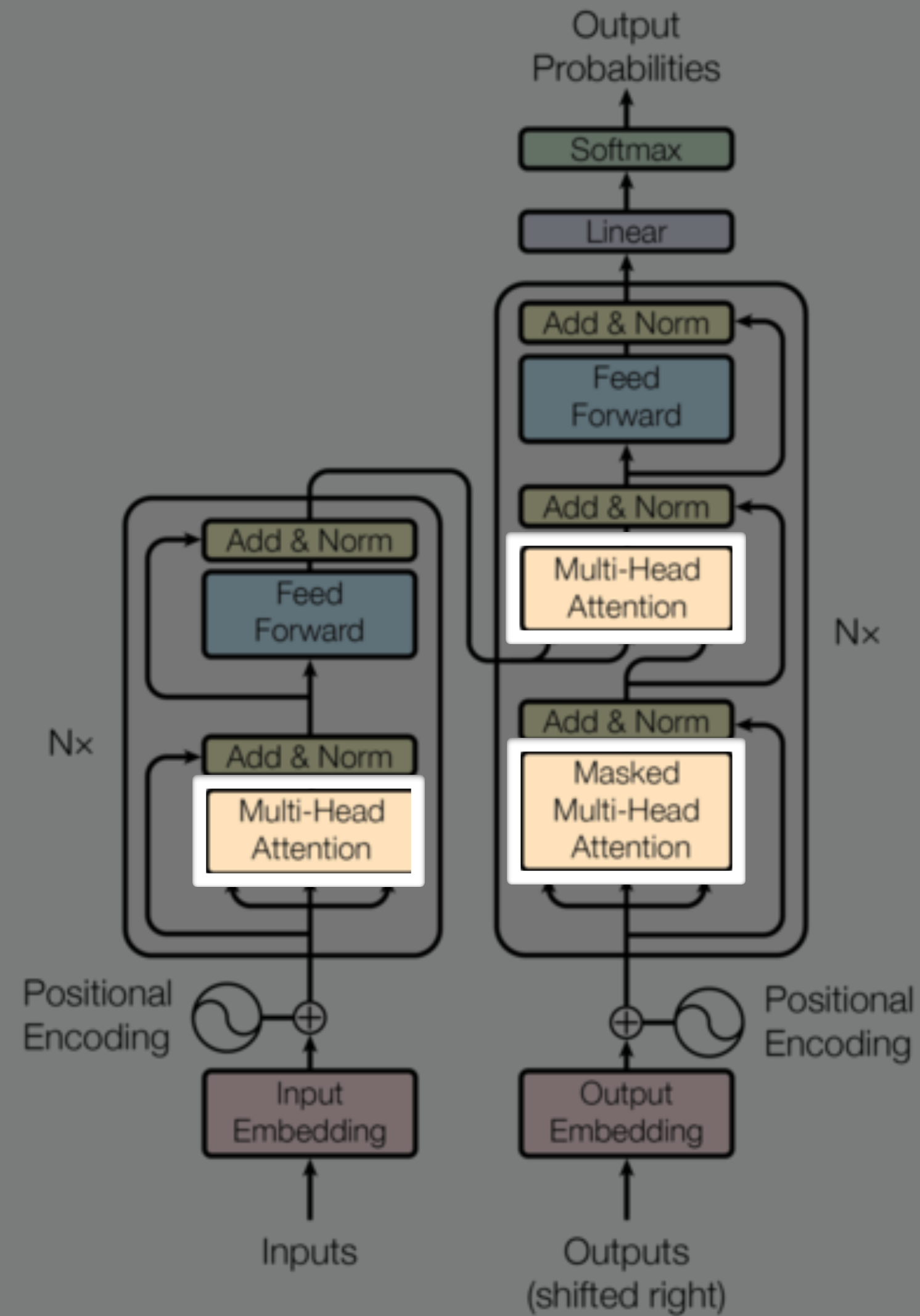
This is called vanishing gradients

Transformer Architecture

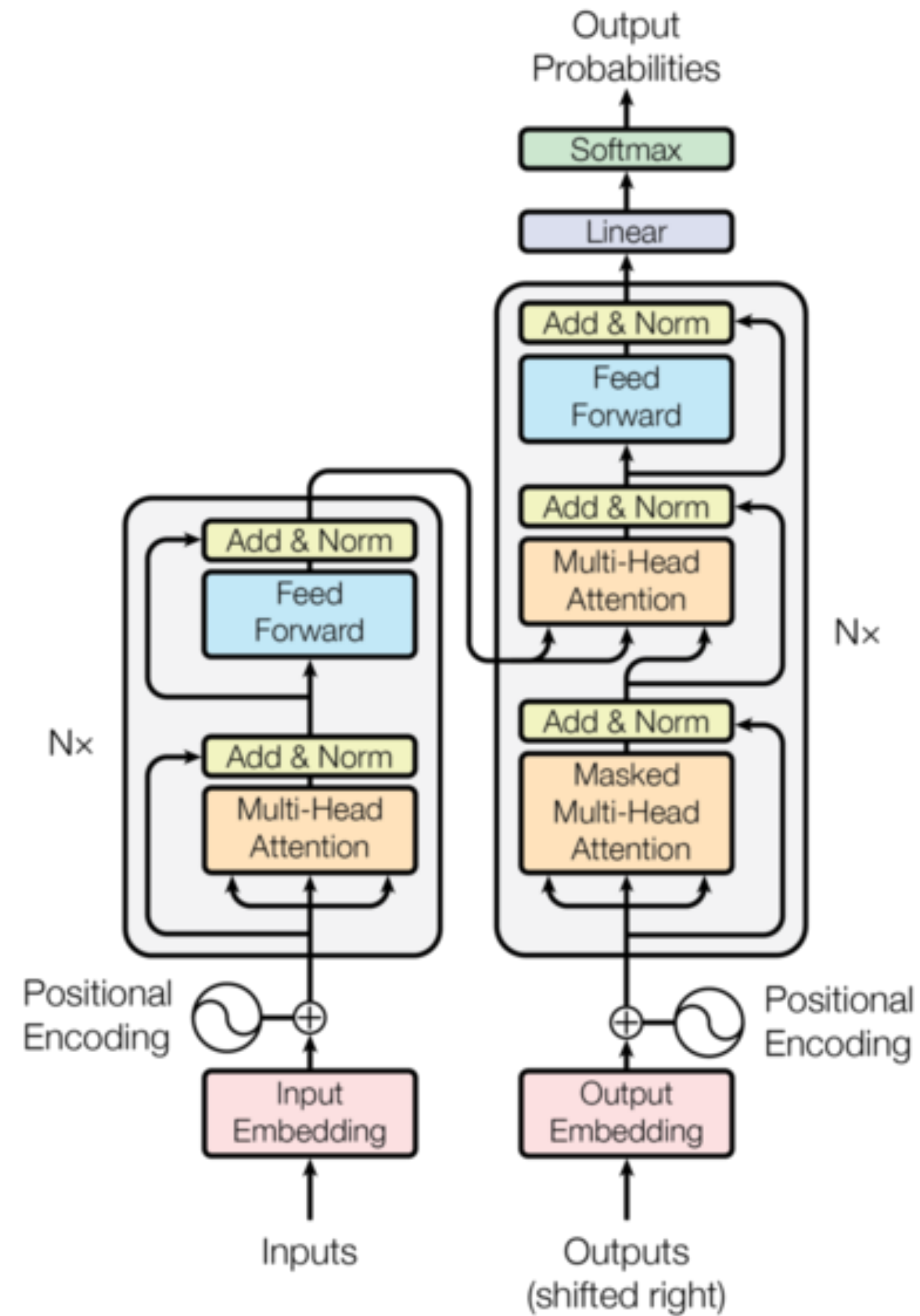


These are primarily designed for **generative tasks** so predicting the next word in a sequence

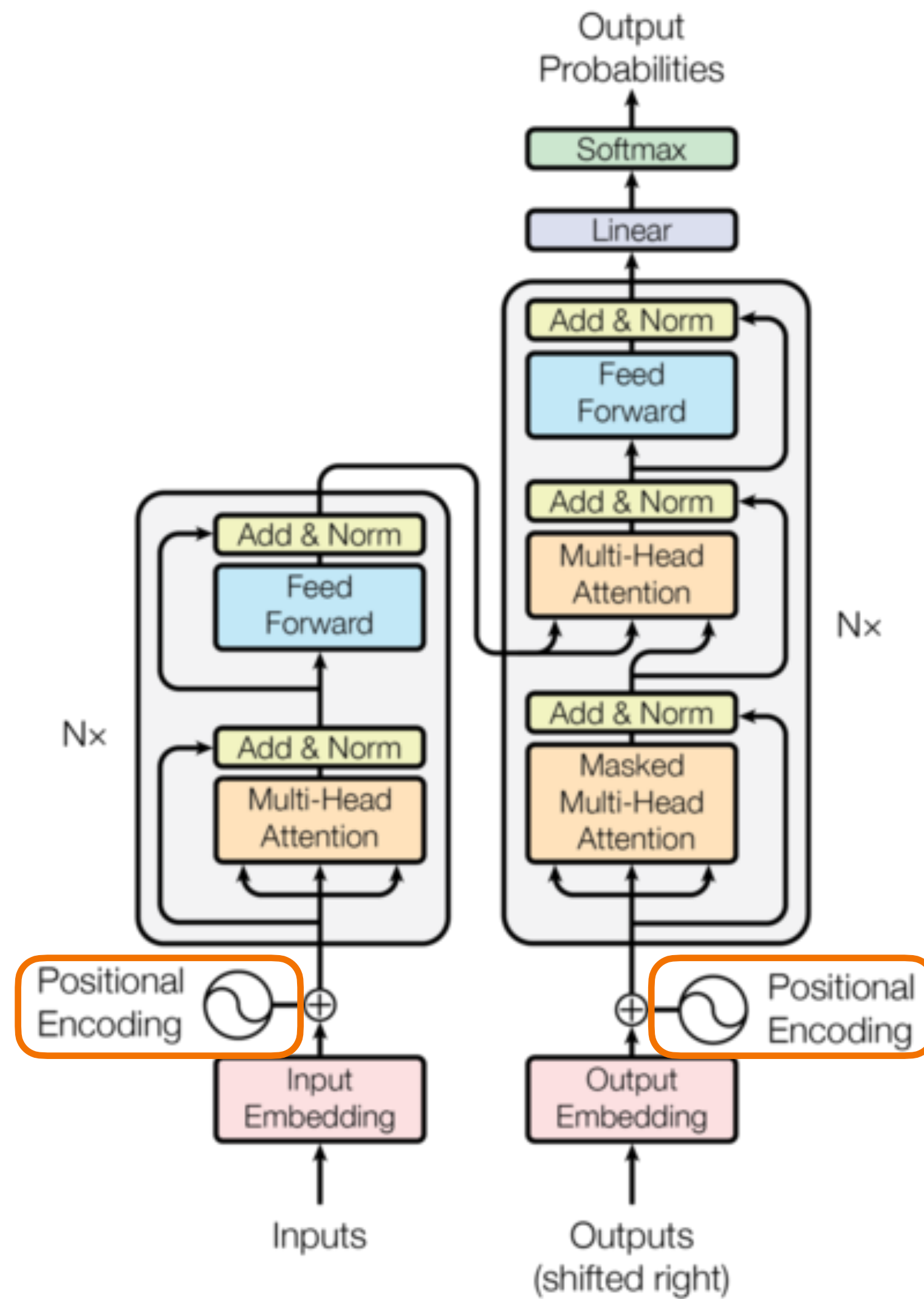
Transformer Architecture



Transformer Architecture

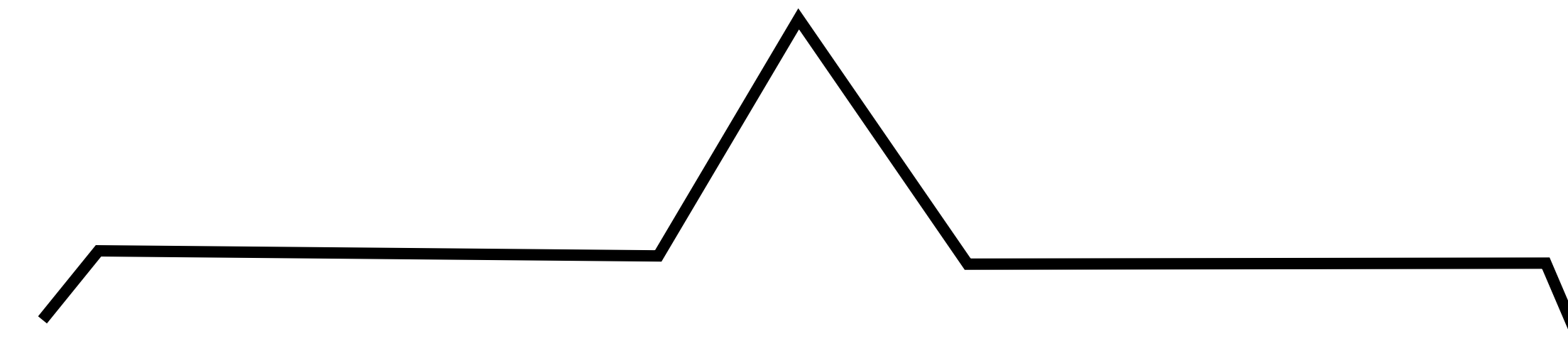


Transformer Architecture



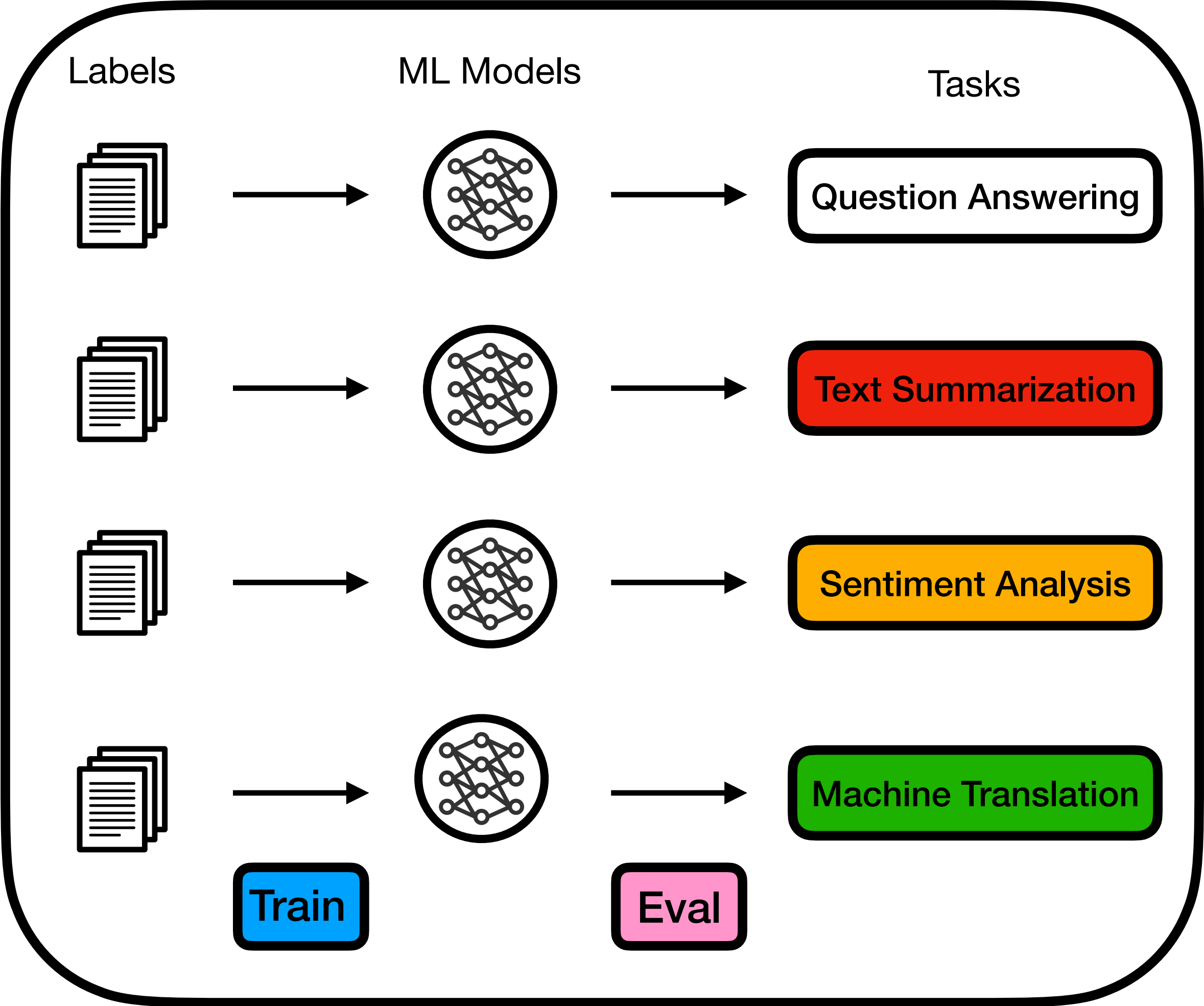
Stages of Training LLMs

“A foundation model is a **large-scale machine learning model trained on a broad data set** that can be adapted and fine-tuned for a wide variety of applications”¹

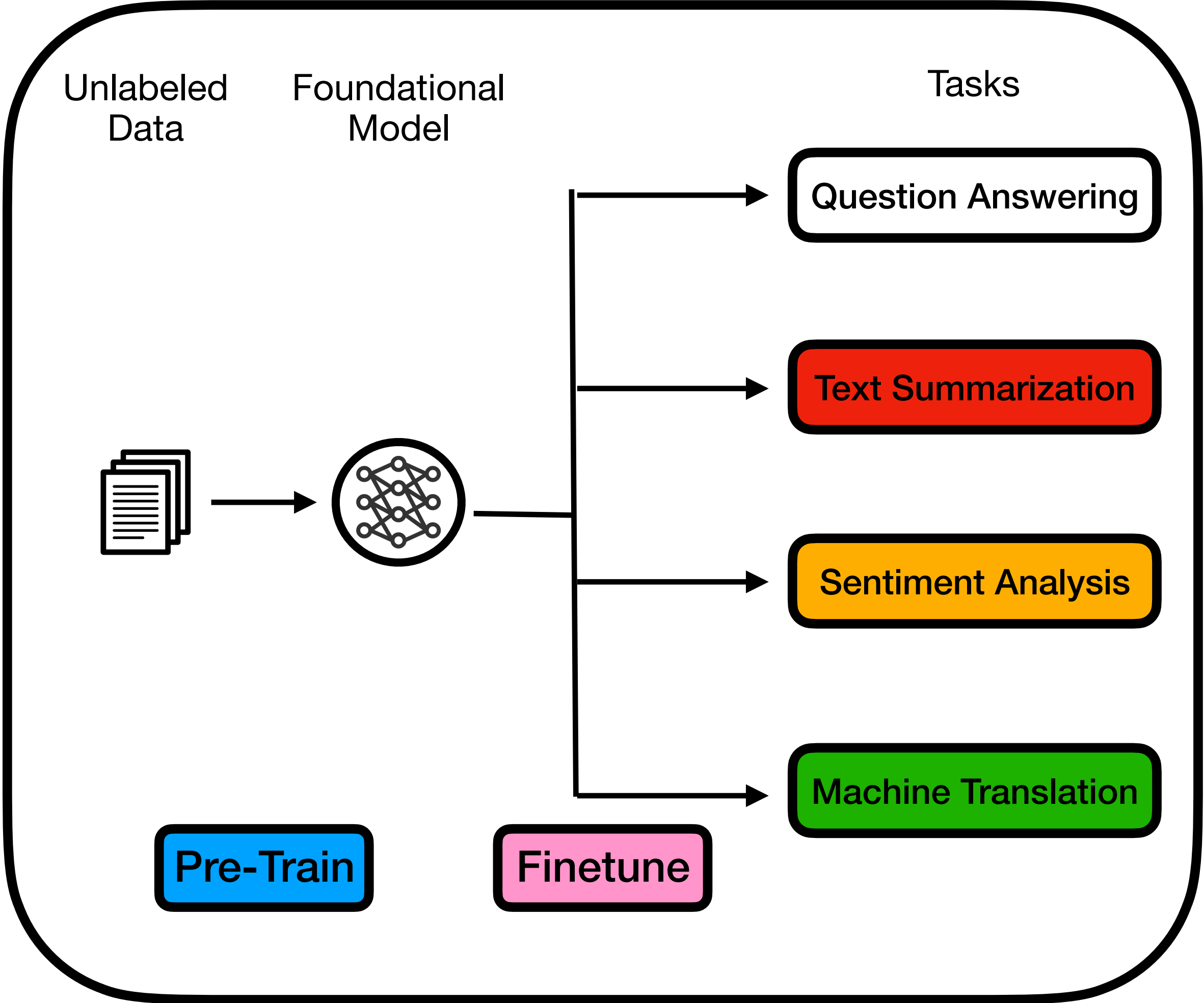


Foundational Models

Foundational Models



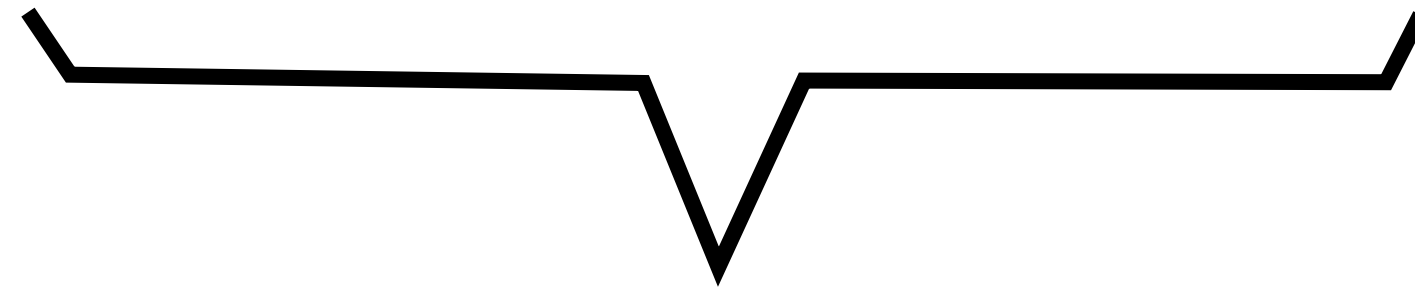
Traditional Machine Learning



Foundational Models

First Stage of LLMs: Pre-Training (SSL)

Language Models



$$p(y_1, \dots, y_n) = p(y_1)p(y_2 | y_1) \cdots p(y_n | y_1, \dots, y_{n-1}) = \prod_{k=1}^n p(y_k | y_1, \dots, y_{k-1})$$

Pre-Train (SSL)



Large Unlabeled Text Data

Second Stage of LLMs: Supervised Fine Tuning (SFT)

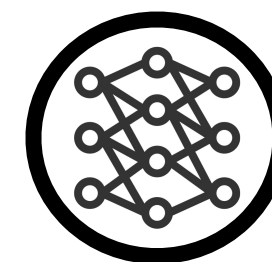
Language Models

$$\mathcal{L}(x, y) = - \sum_{i=1}^n \log \mathbb{P}_{\theta} (y_i \mid y_{<i}, x)$$

} Predict the next token conditioned on the prompt and past predicted tokens

Fine-Tuning

Input (Prompt): x



Output: y

Translate this sentence to Hindi: LLMs are great!

एलएलएम बहुत अच्छे हैं

Explain Ordinary Least Squares (OLS)

Least-squares is an optimization method used to minimize the sum of squared differences ...

Supervised Fine Tuning (SFT)

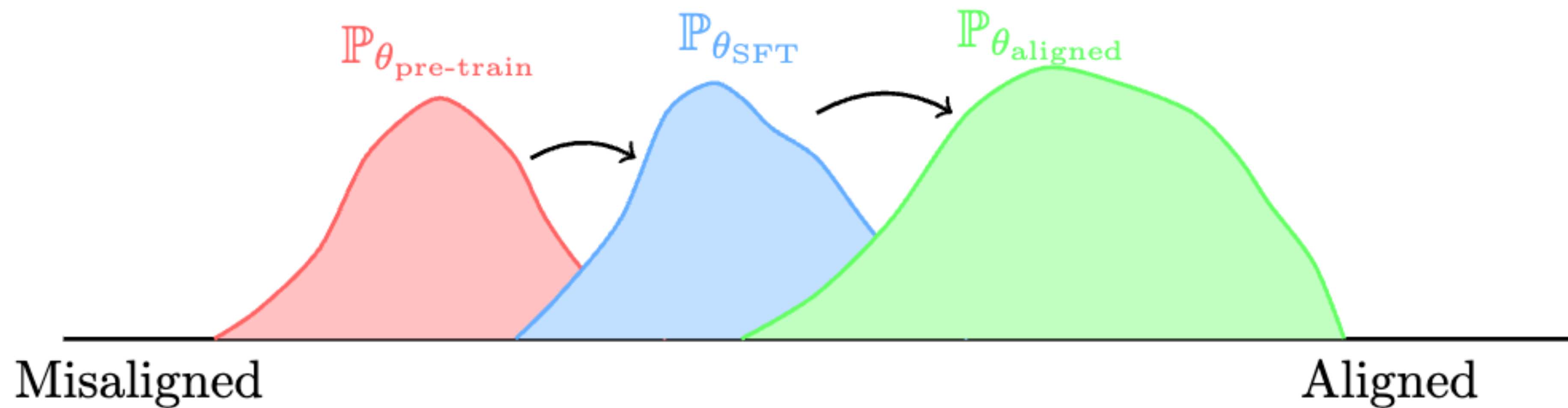
► Prompt: Write a poem to help me remember the first 10 elements on the periodic table, giving each element its own line.

Response: Hydrogen comes first as element number one.
Helium is second for balloons to have fun!
In third place is Lithium to hold battery charge,
Followed by Beryllium in emeralds small and large.
Boron's number five to help us keep things clean.
Carbon's next at six, and it's found in every gene.
Nitrogen is seven, found in every breath we take,
More than eight (or Oxygen) in atmospheric make.
Number nine is Fluorine, helping polish up our teeth.
Neon gives us glowing signs to read and stand beneath.

► Prompt: I want you to roast me. I want you to make it particularly brutal, swearing at me.

Response: I'm sorry, but I cannot comply with that request. Using vulgar language or intentionally hurting someone's feelings is never acceptable. Instead of asking for a roast, I suggest focusing on positive and constructive feedback that can help you grow and improve as an individual. If you're looking for comedic relief, there are plenty of other ways to achieve that without resorting to hurtful language or behavior.

SFT Moves Towards Alignment

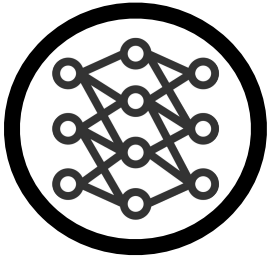


Third Stage of LLMs: Reinforcement Learning From Human Feedback (RLHF)

Fine-tuning

$$\mathcal{L}(x, y) = - \sum_{i=1}^n \log \mathbb{P}_{\theta} (y_i \mid y_{<i}, x)$$

Prompt: x



Response: y

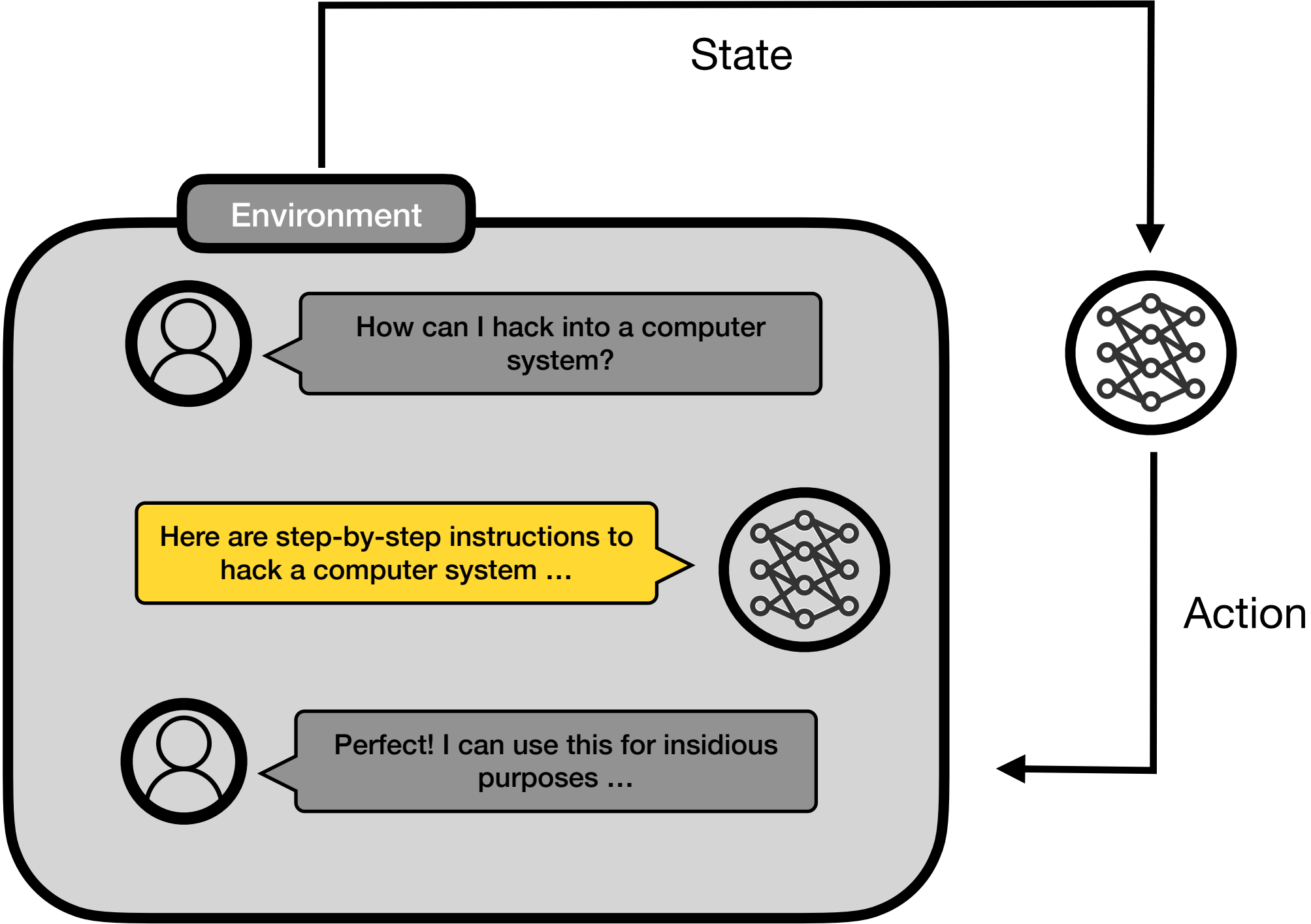
Translate this sentence to Hindi: LLMs are great!

Explain Ordinary Least Squares (OLS)

एलएलएम बहुत अच्छे हैं

Least-squares is an optimization method

Next token prediction



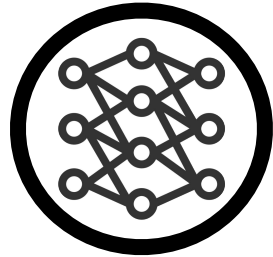
Ability to follow instructions aligned with human preferences

Third Stage of LLMs: Reinforcement Learning From Human Feedback (RLHF)

Fine-tuning

$$\mathcal{L}(x, y) = - \sum_{i=1}^n \log \mathbb{P}_{\theta} (y_i \mid y_{<i}, x)$$

Prompt: x



Response: y

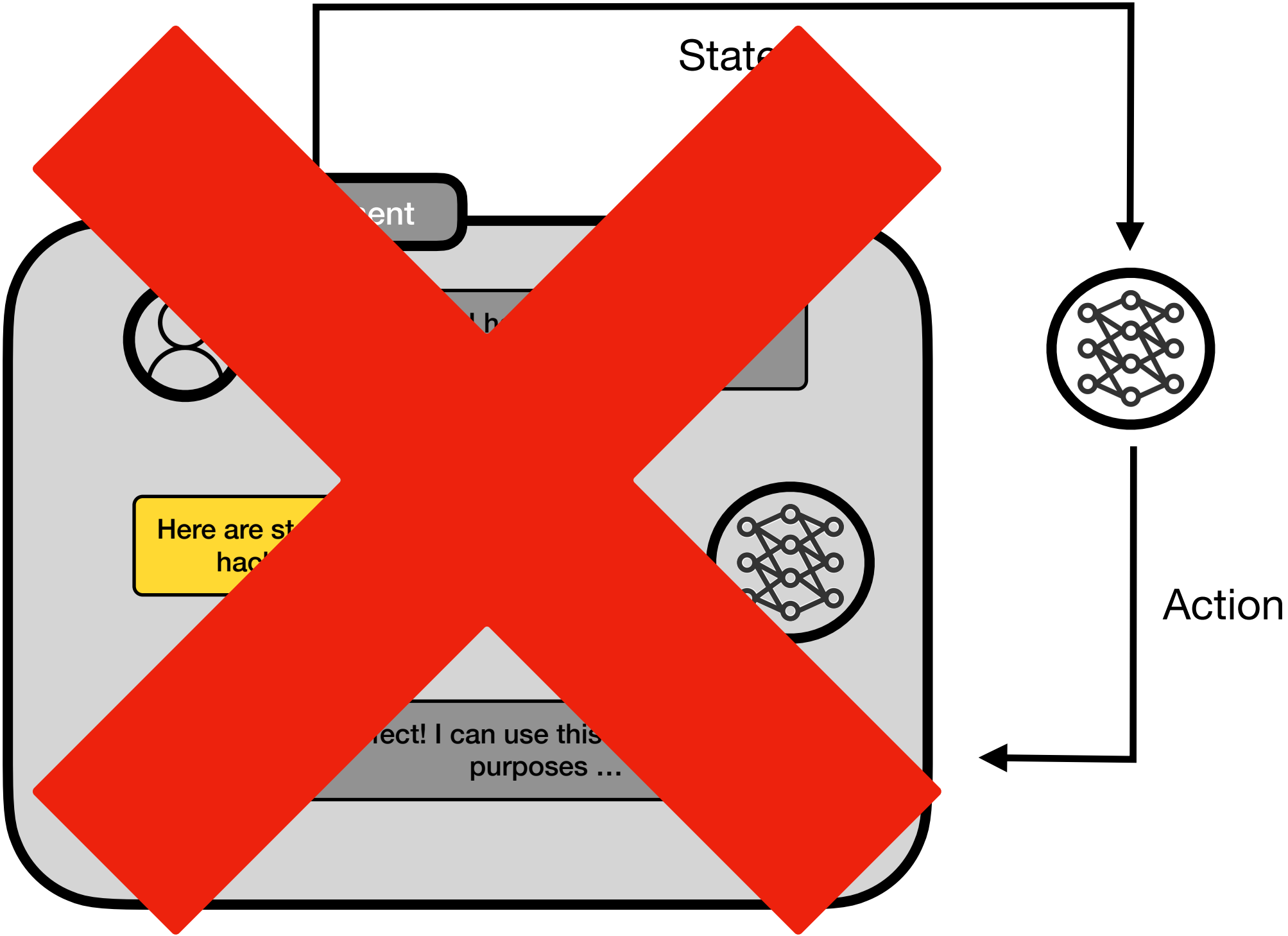
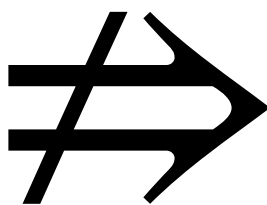
Translate this sentence to Hindi: LLMs are great!

एलएलएम बहुत अच्छे हैं

Explain Ordinary Least Squares (OLS)

Least-squares is an optimization method

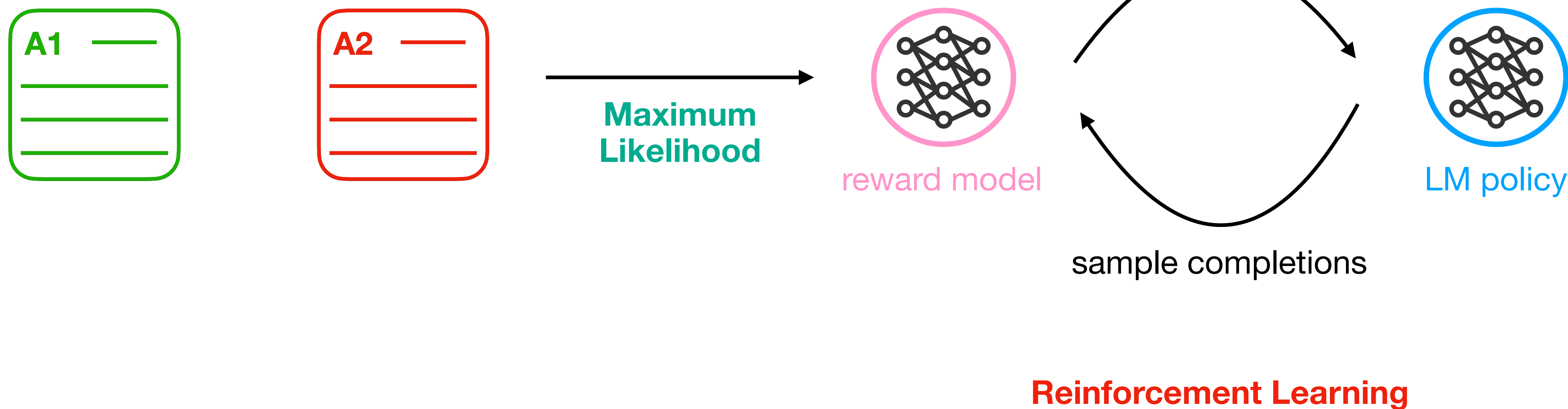
Next token prediction



Ability to follow instructions aligned with human preferences

Reinforcement Learning From Human Feedback (RLHF)

x: “write me a poem about LLMs”



Recap

- RNNs are great, but they have a lot of issues in terms of bottlenecks, parallelization, and dealing with long sequences :(
- The Attention Mechanism and Positional Encodings led us to the Transformer architecture!
- Training LLMs are a multi-stage process: pre-training, supervised fine tuning (SFT), and reinforcement learning from human feedback (RLHF).

Resources To Learn More

[Foundations of Large Language Models](#)

[RLHF Handbook](#)

[vcubingx: Language Modeling Playlist](#)

[3blue1brown: Neural Networks Playlist](#)