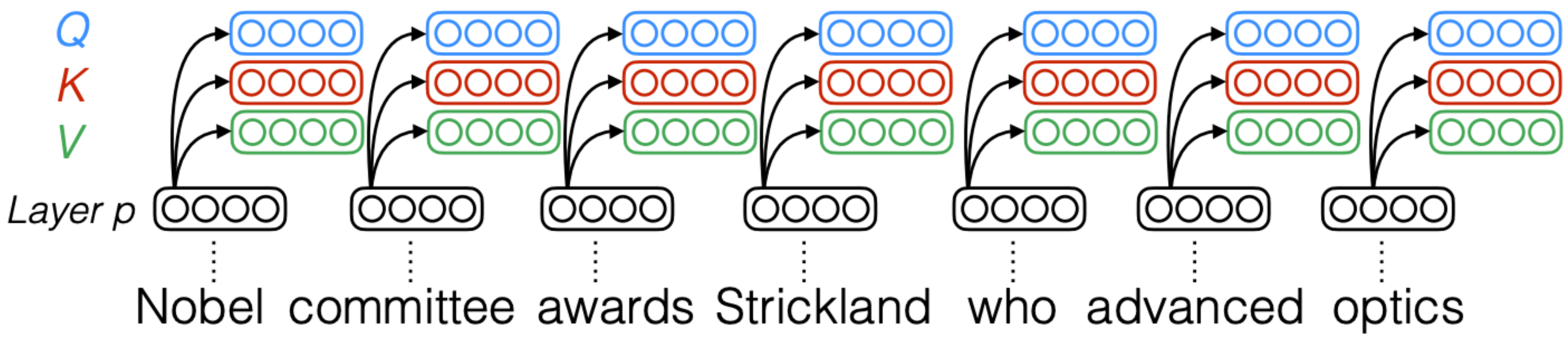
Introduction to Large Language Models

CONTENT

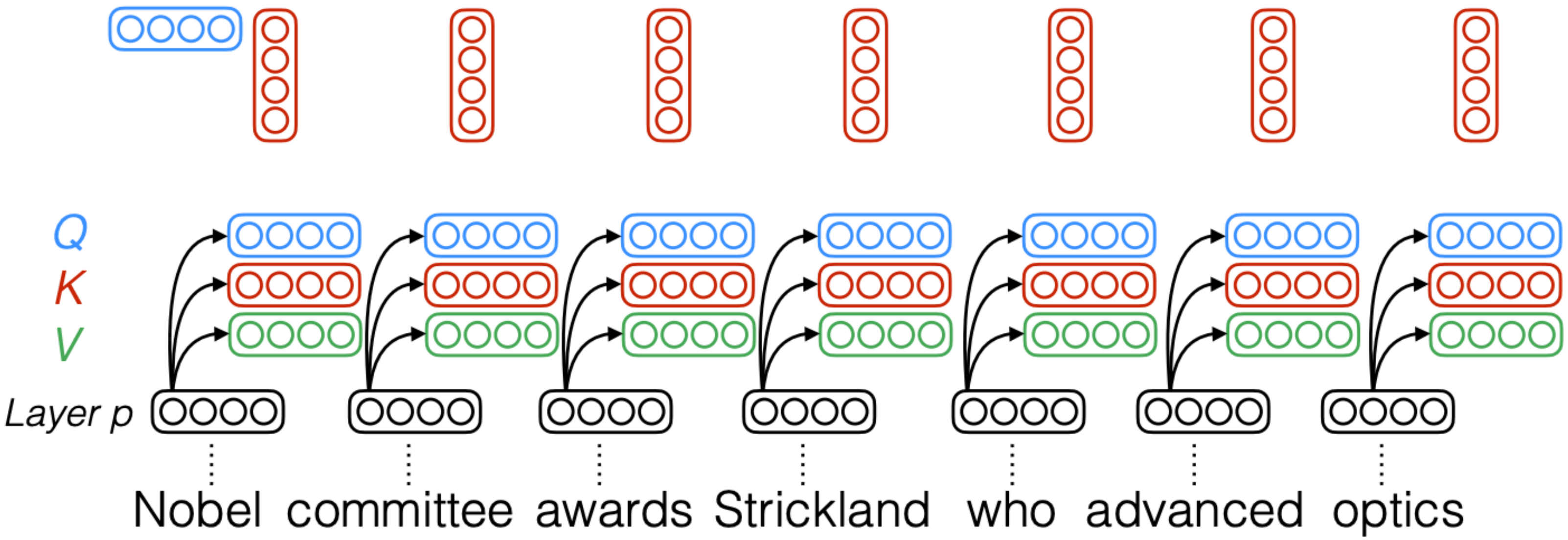
- Attention mechanism
- Model structure
- Pre-training & IT & RLHF
- Scaling
- Multi Modal LLM

Attention mechanism

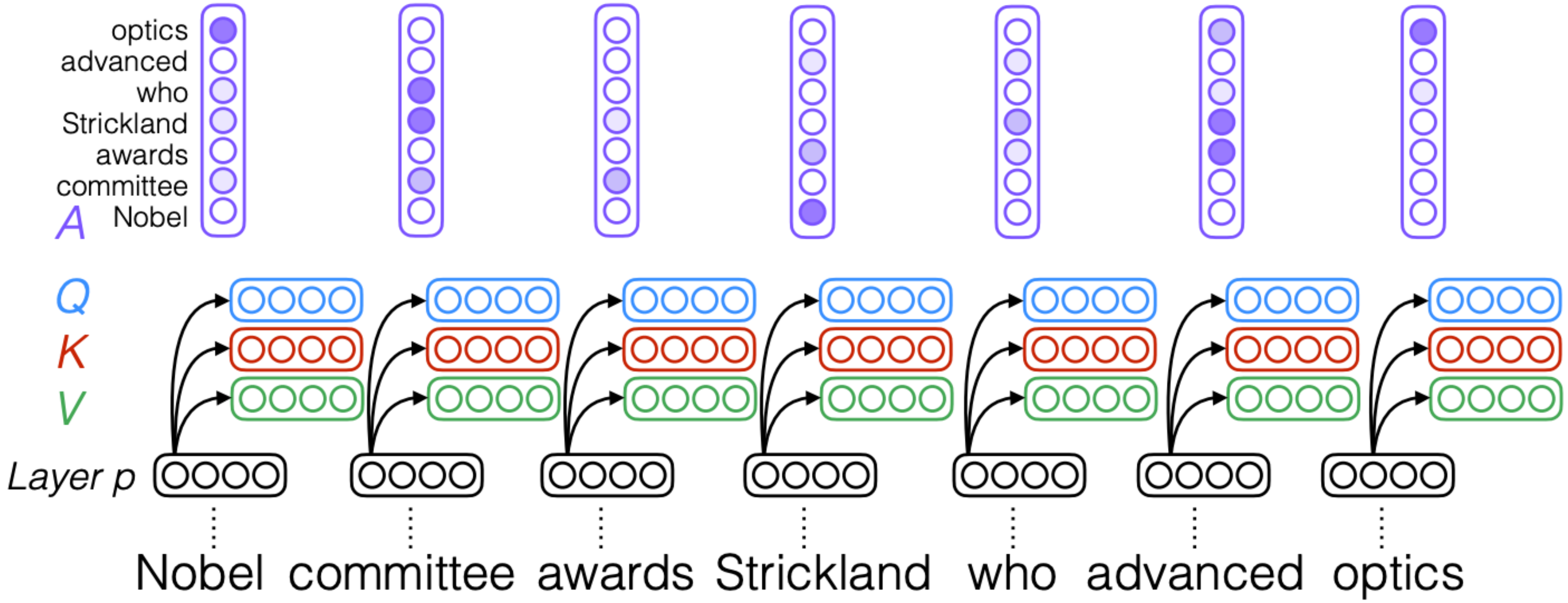
Attention mechanism



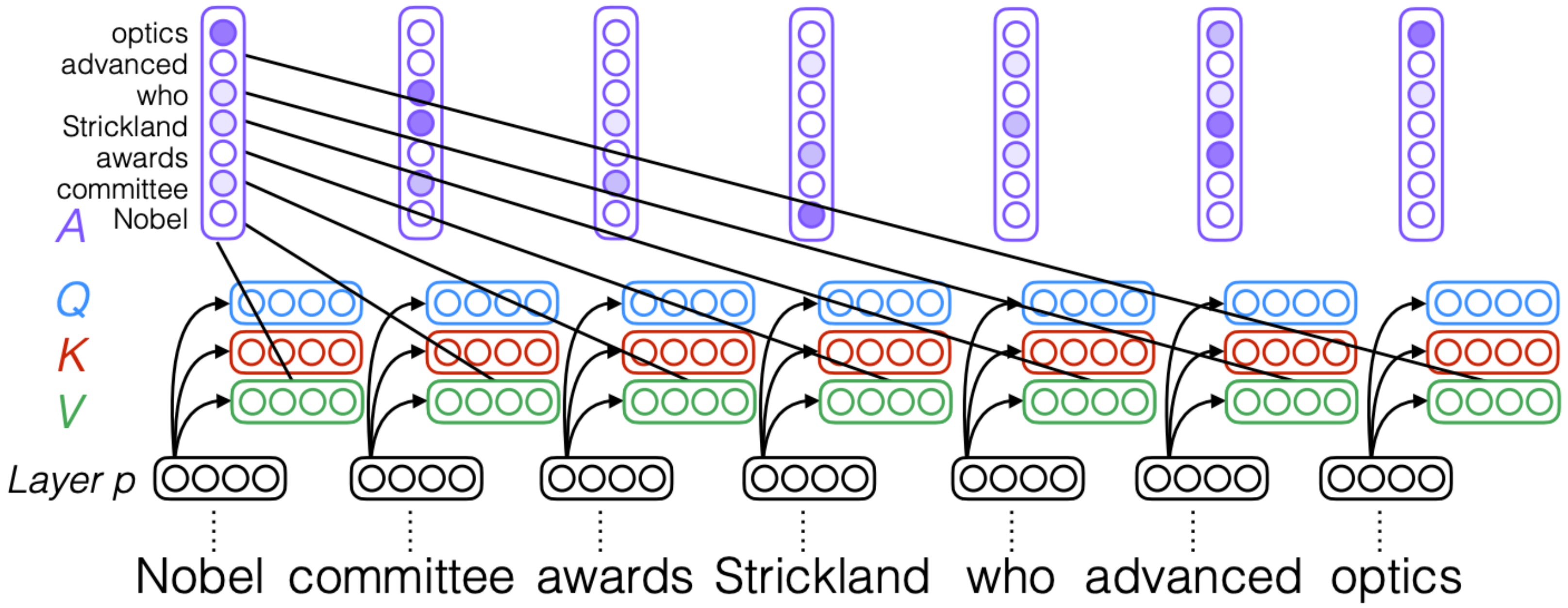
Attention mechanism



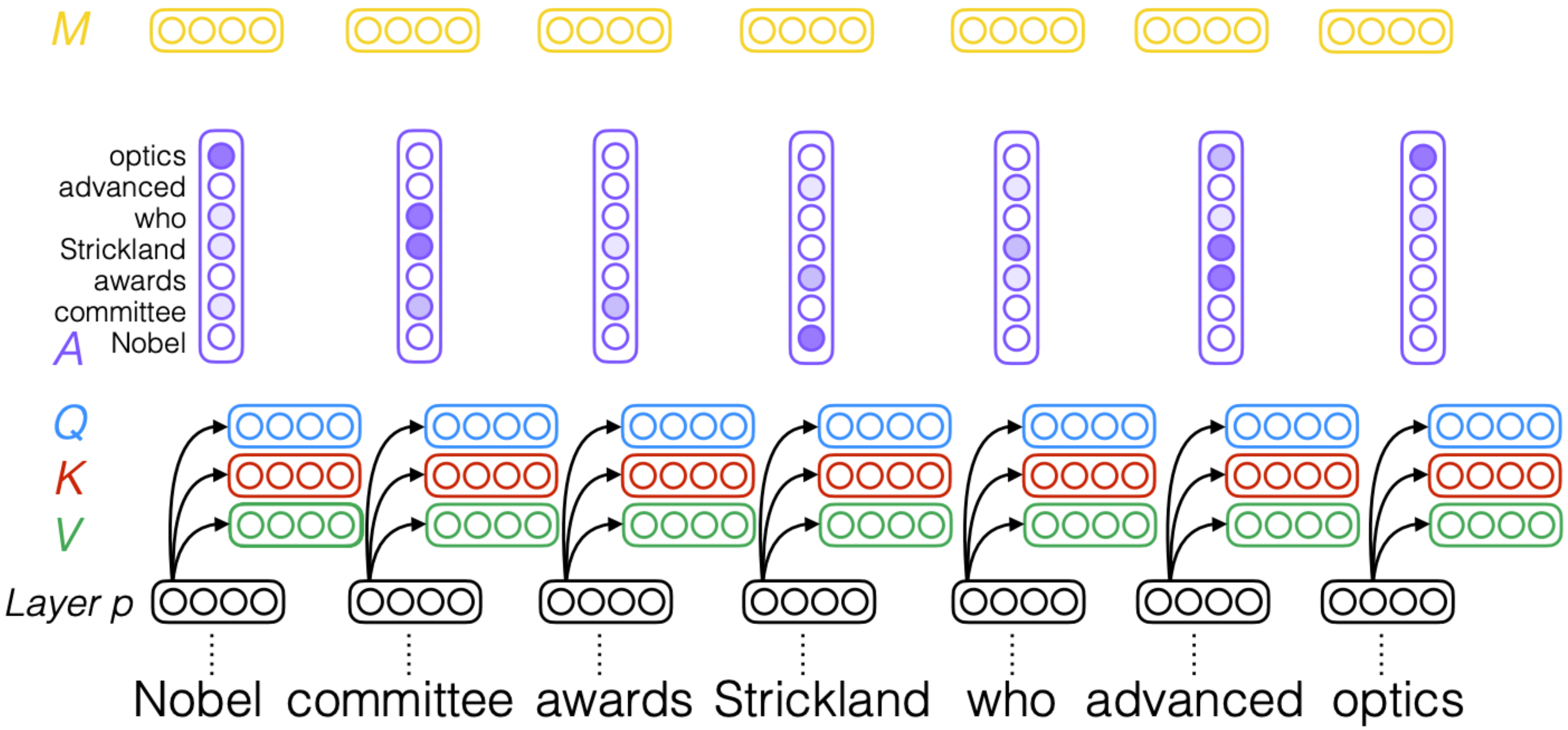
Attention mechanism



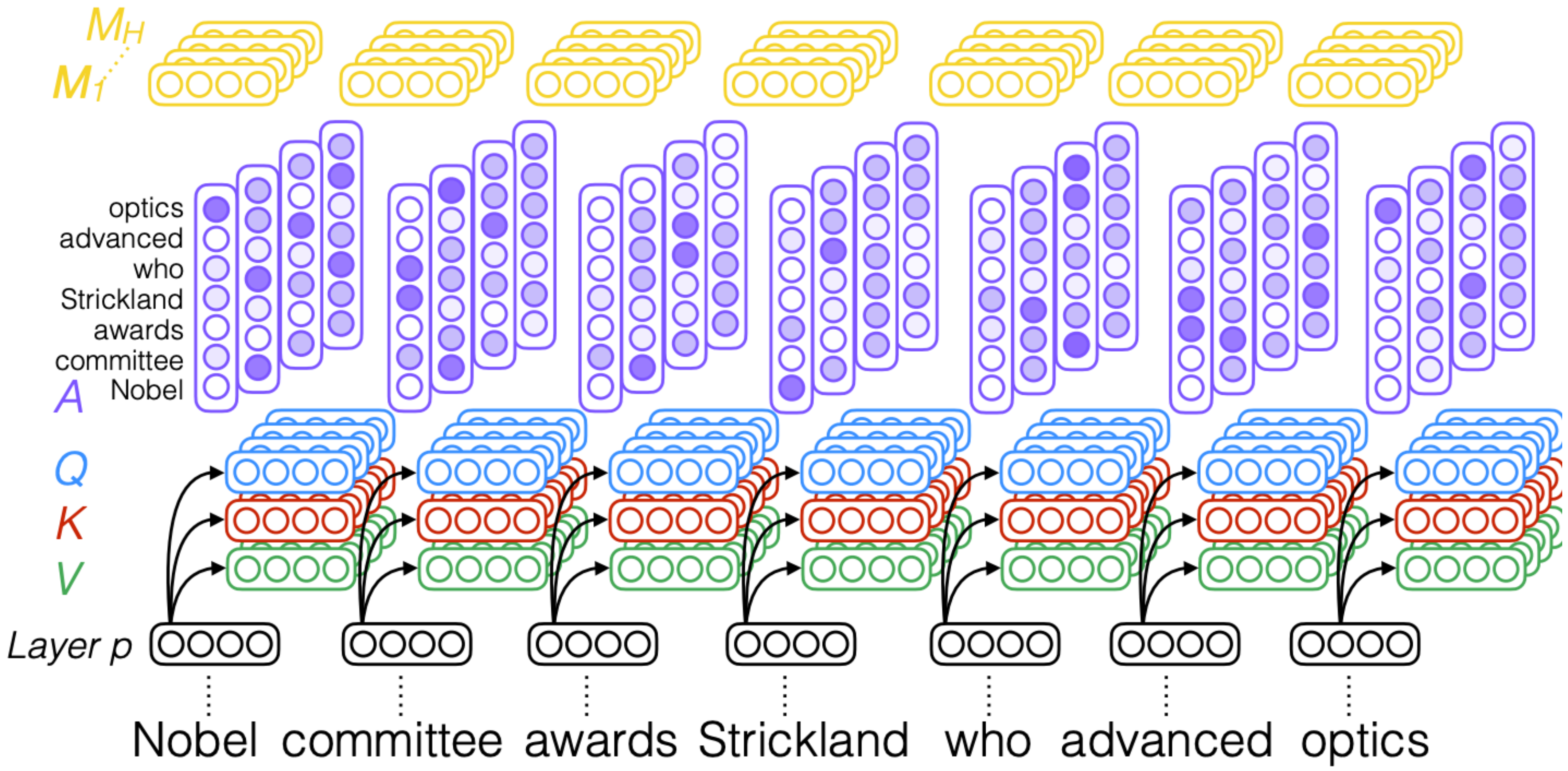
Attention mechanism



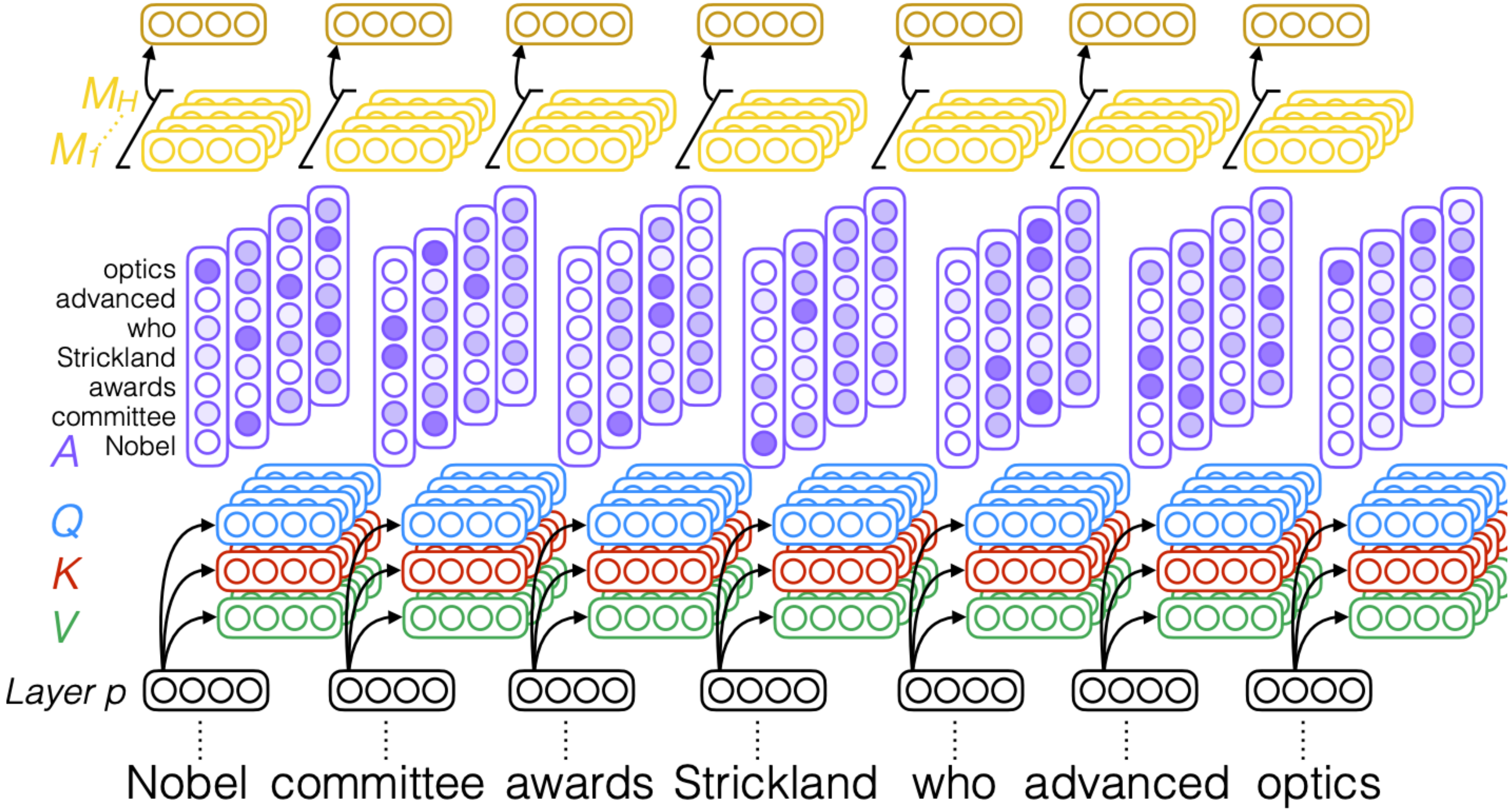
Attention mechanism



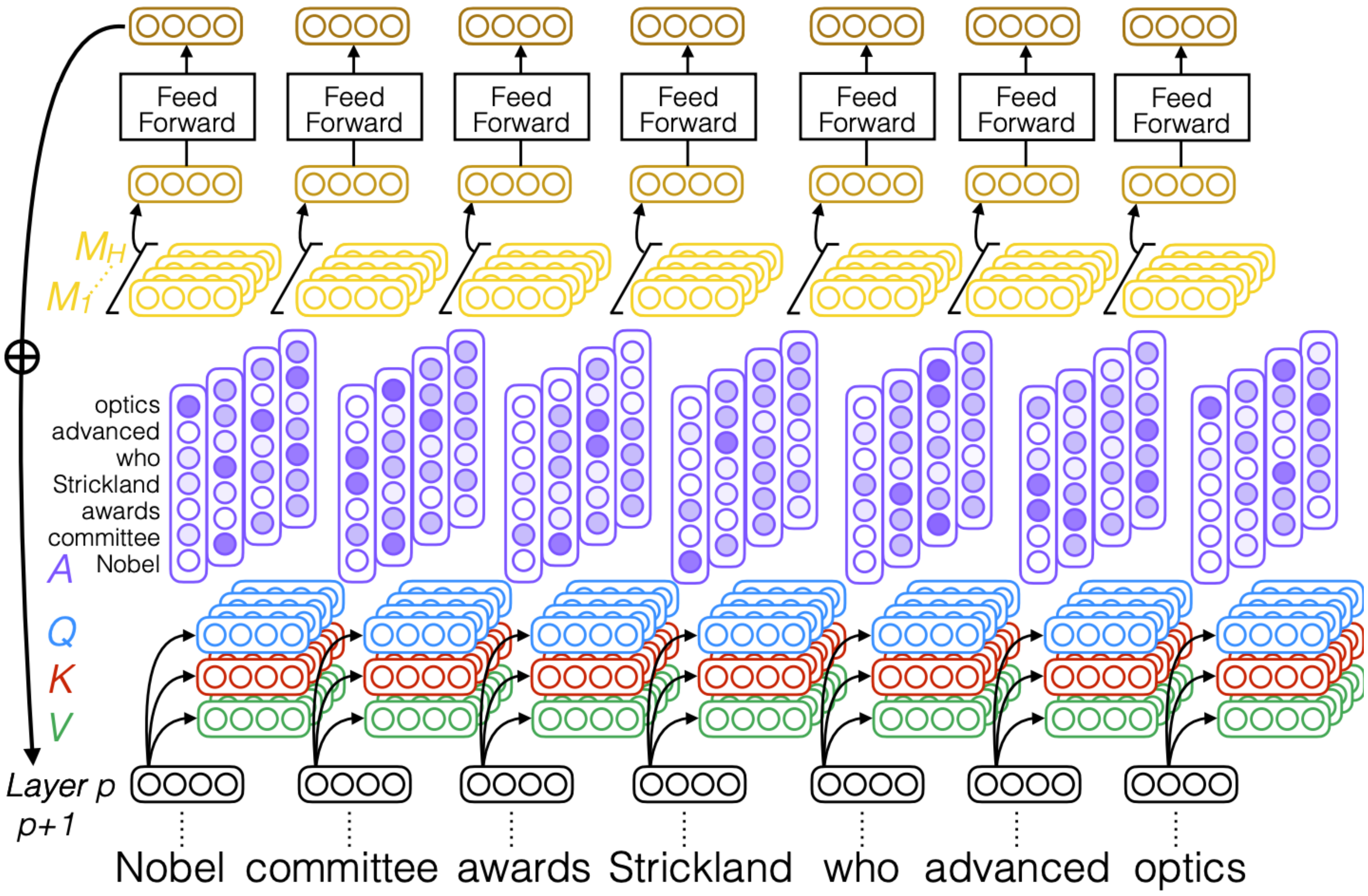
Attention mechanism



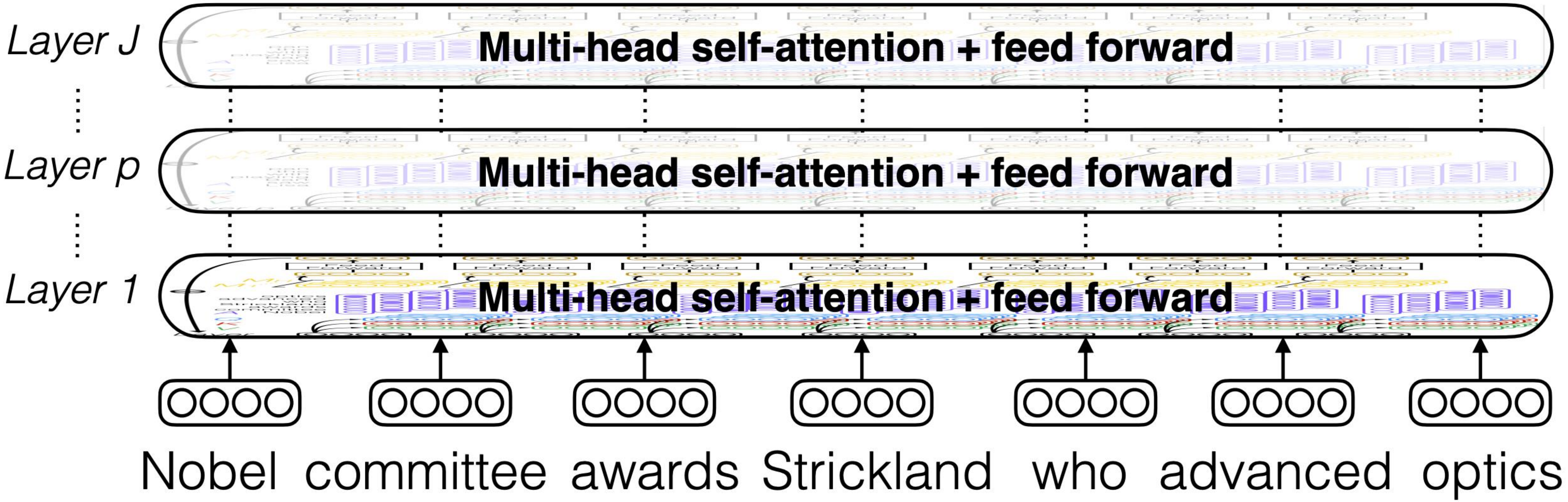
Attention mechanism



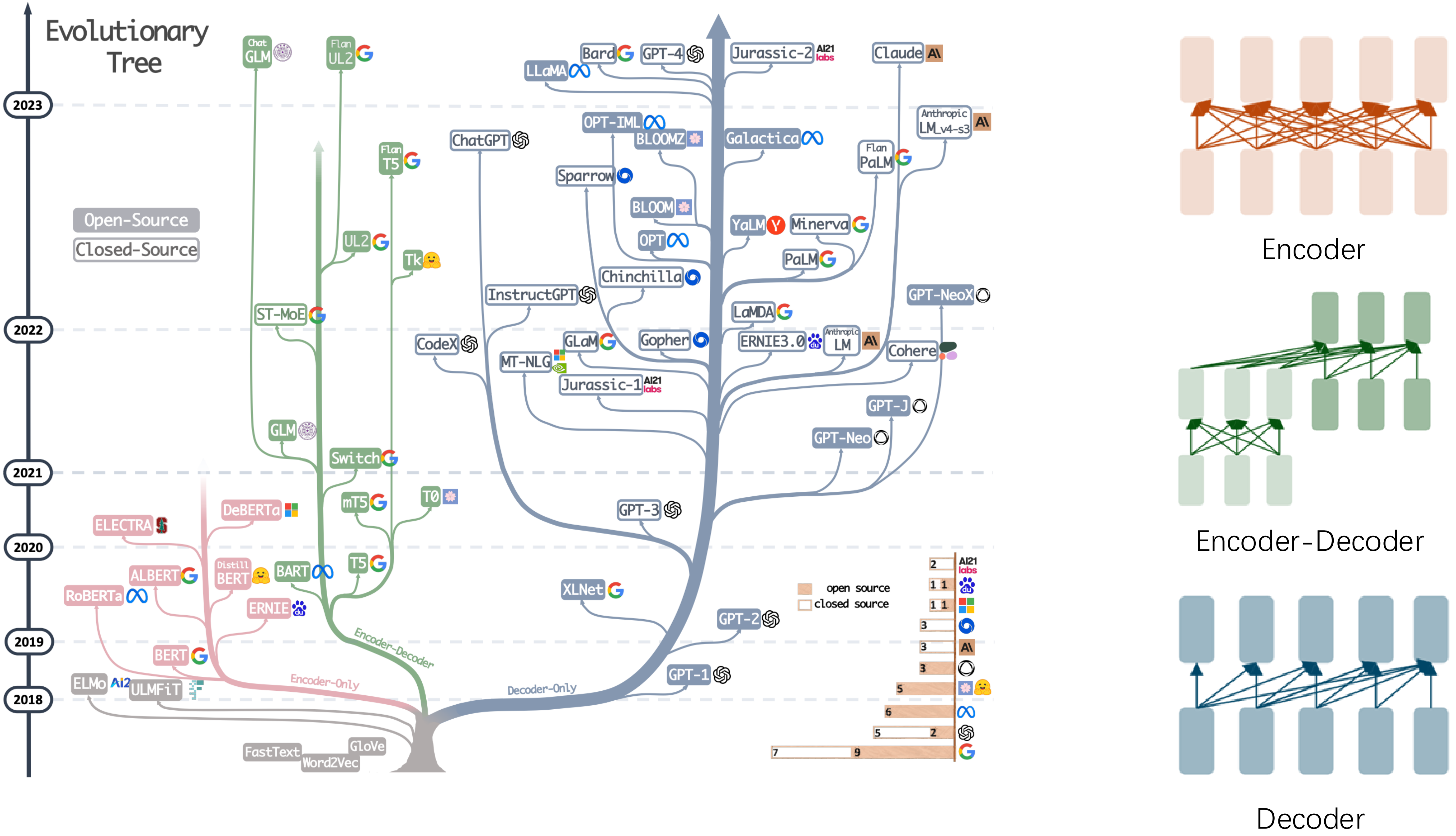
Attention mechanism



Attention mechanism

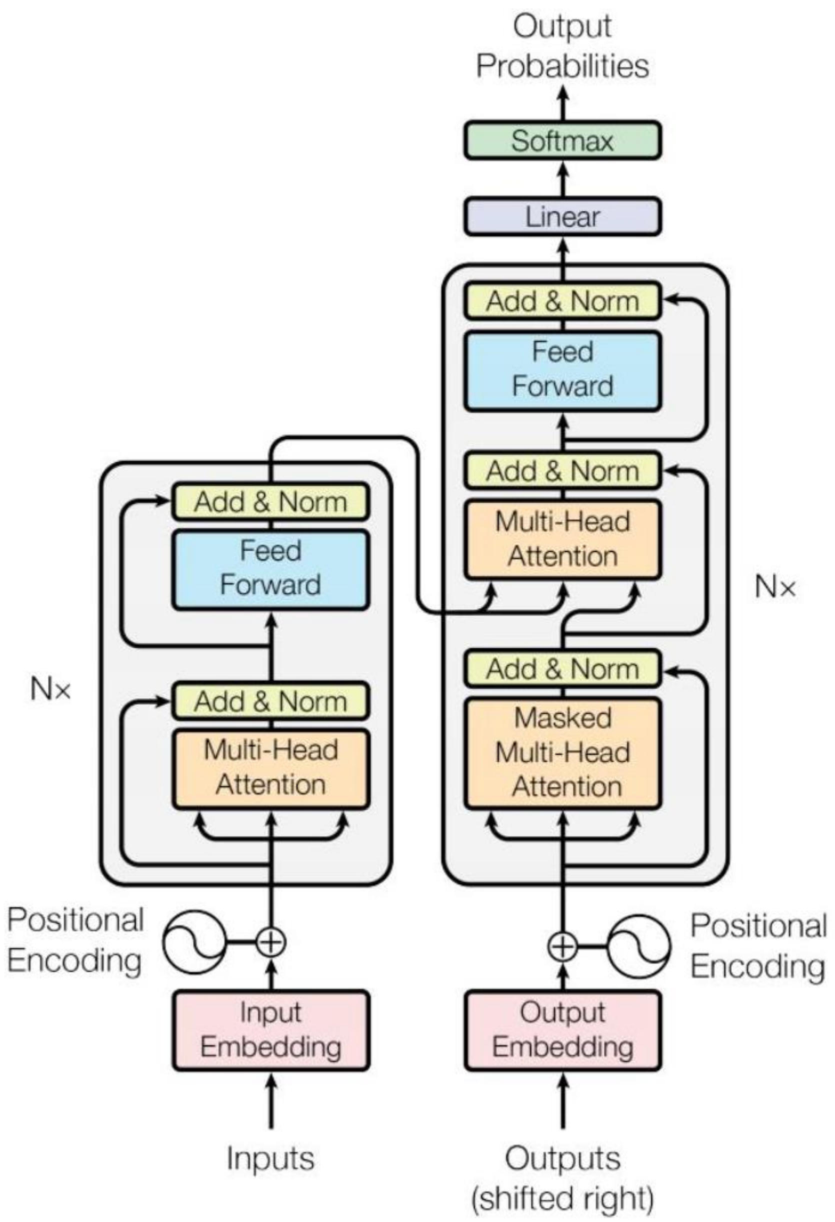


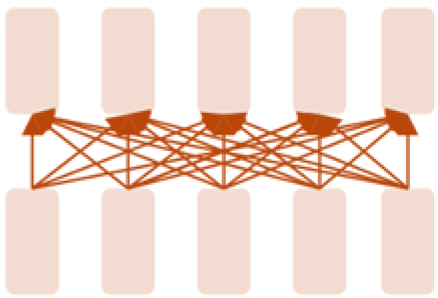
Model structure



Evolutionary Tree
2023
2022
2021
2020
2019
2018
Open-Source
Closed-Source
Encoder-Only
Encoder-Decoder
Decoder-Only
ChatGLM
Flan UL2
LLaMA
Bard
GPT-4
Jurassic-2
Claude
OPT-IML
BLOOMZ
Galactica
ChatGPT
Anthropic LM_v4-s3
Flan T5
Flan PaLM
Sparrow
BLOOM
YaLM
Minerva
UL2
Tk
OPT
PaLM
Chinchilla
ST-MoE
InstructGPT
GPT-NeoX
LaMDA
CodeX
GLaM
Gopher
ERNIE3.0
Anthropic LM
Cohere
MT-NLG
Jurassic-1
GPT-J
GLM
GPT-Neo
Switch
mT5
T0
GPT-3
DeBERTa
ELECTRA
T5
ALBERT
Distill BERT
BART
XLNet
RoBERTa
ERNIE
GPT-2
BERT
ELMo
ULMFiT
GPT-1
FastText
GloVe
Word2Vec
open source
closed source
Encoder
Encoder-Decoder
Decoder

Eecoder & encoder-decoder





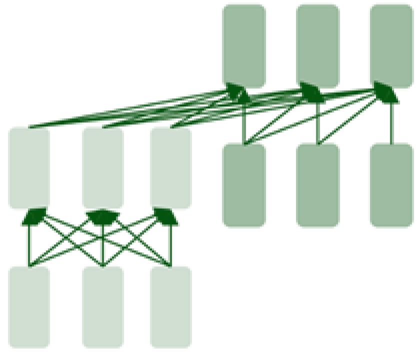
encoder

Trained by predicting words from surrounding words on both sides.

good: Strong comprehension ability.

bad: Limited generation ability.

Application: Discrimination task



encoder-decoder

Trained to map from one sequence to another

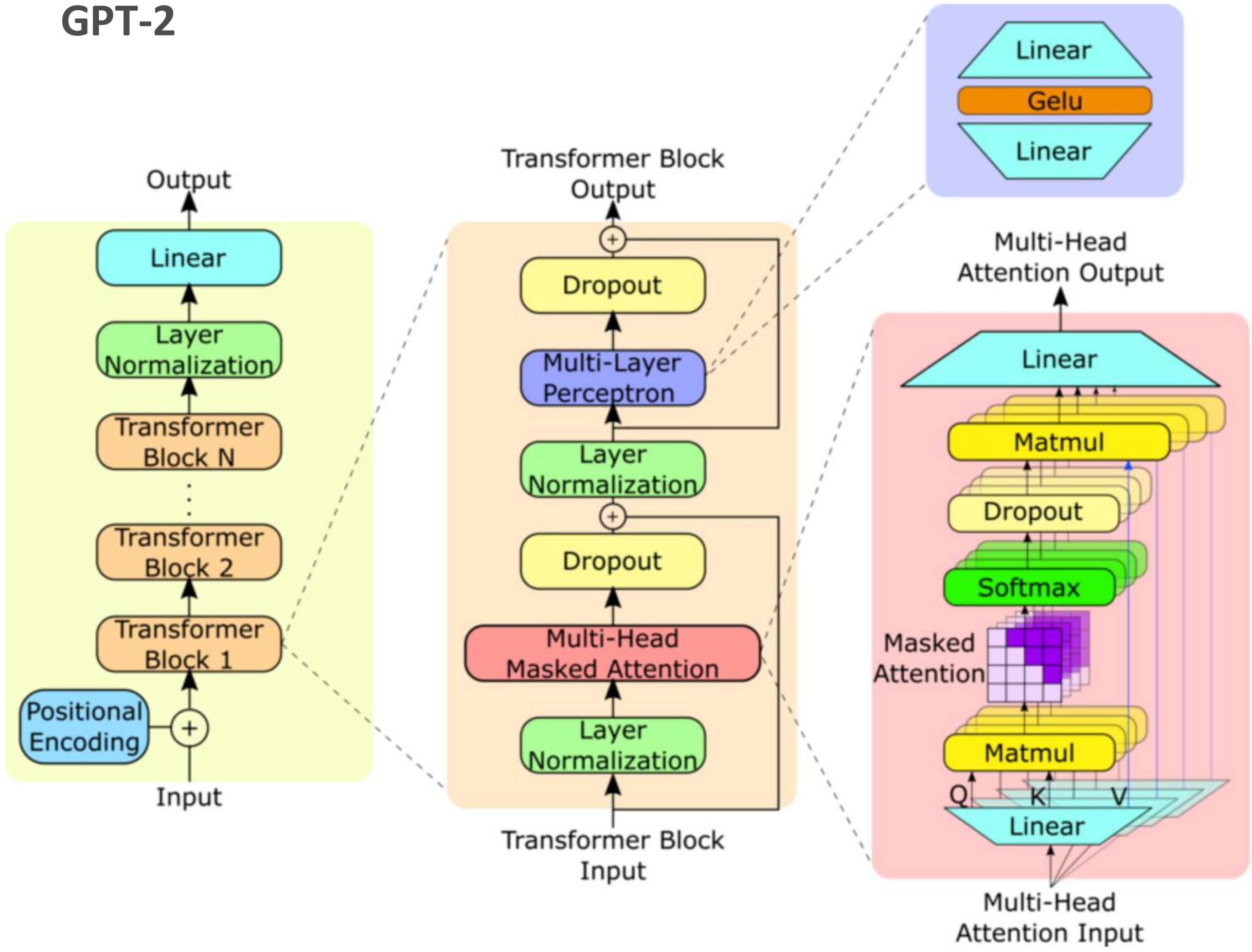
good: good at both comprehension and generation

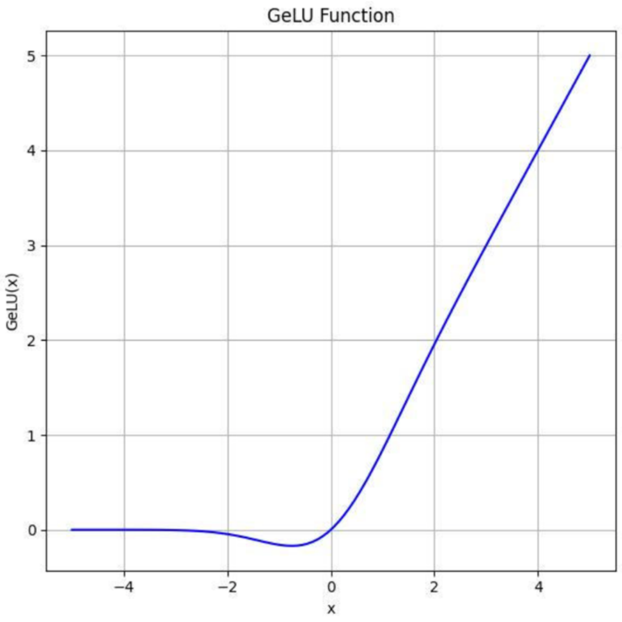
bad: Hard and expensive to train

Application: Translation task

Decoder

GPT-2





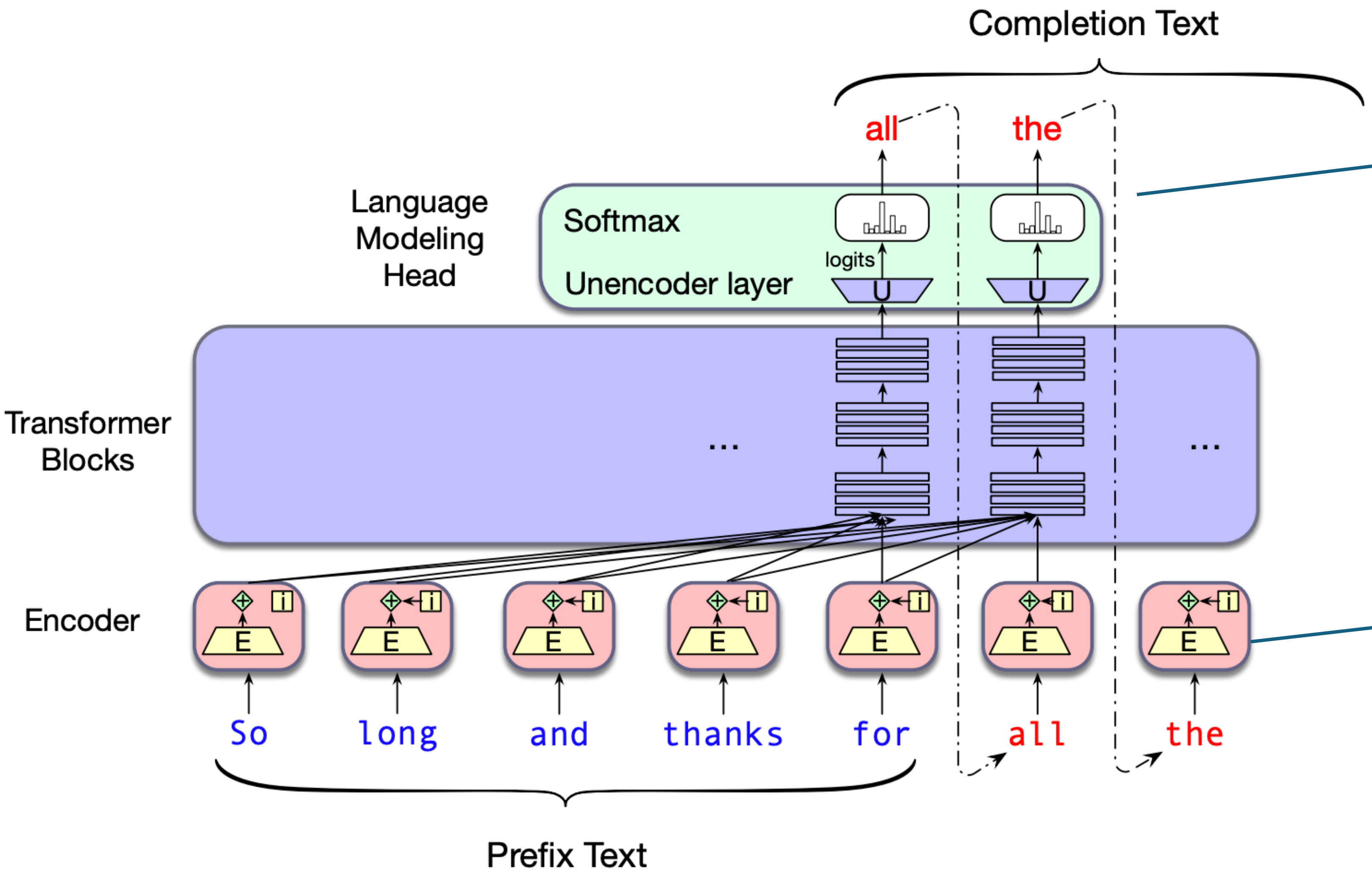
$$\mathrm{GELU}(\mathrm{x}) = \mathrm{x} * \Phi(\mathrm{x})$$

Predict from left to right

Good at generation

Easy for scaling up !

Decoder



Sampling:

- Top k
- Top p
- Softmax with temperature

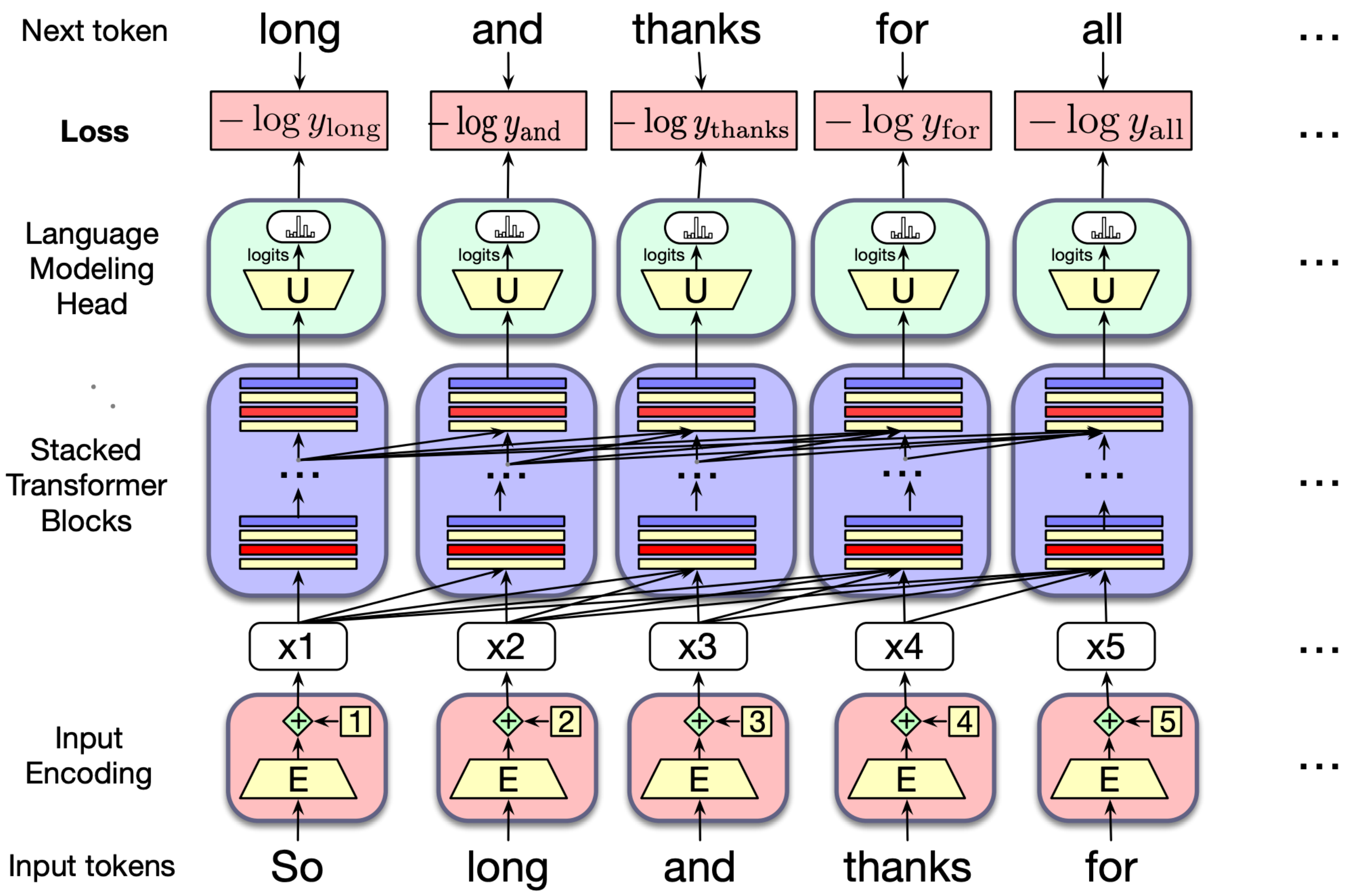
Tokenization:

Granularity: Subword

Subword -> index -> embedding

Pre-training & IT & RLHF

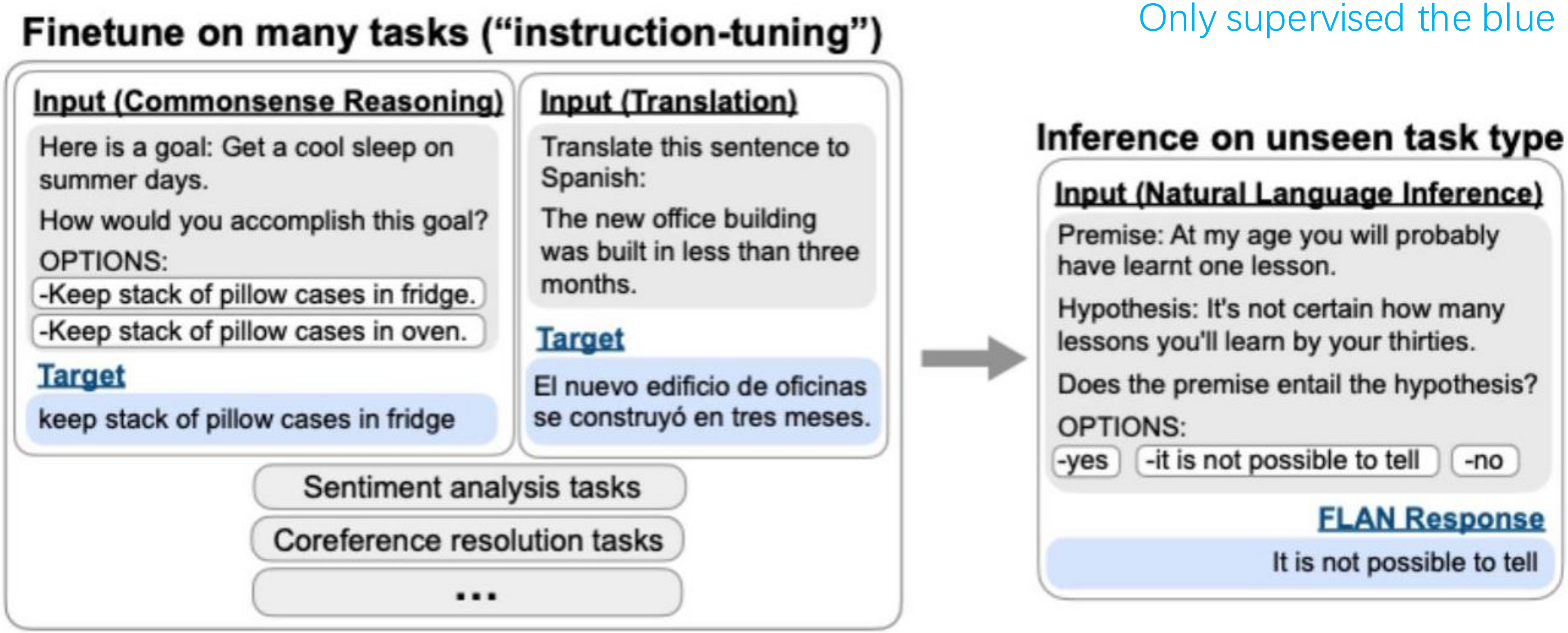
Pre-training



- Train to predict the next token (self-supervised training)
- Large-scale corpora from the internet
- Cross-entropy loss
- Good generalist auto-completes

Task specifical ?
In context learning ~

Instruction-tuning



Only supervised the blue part (answer)

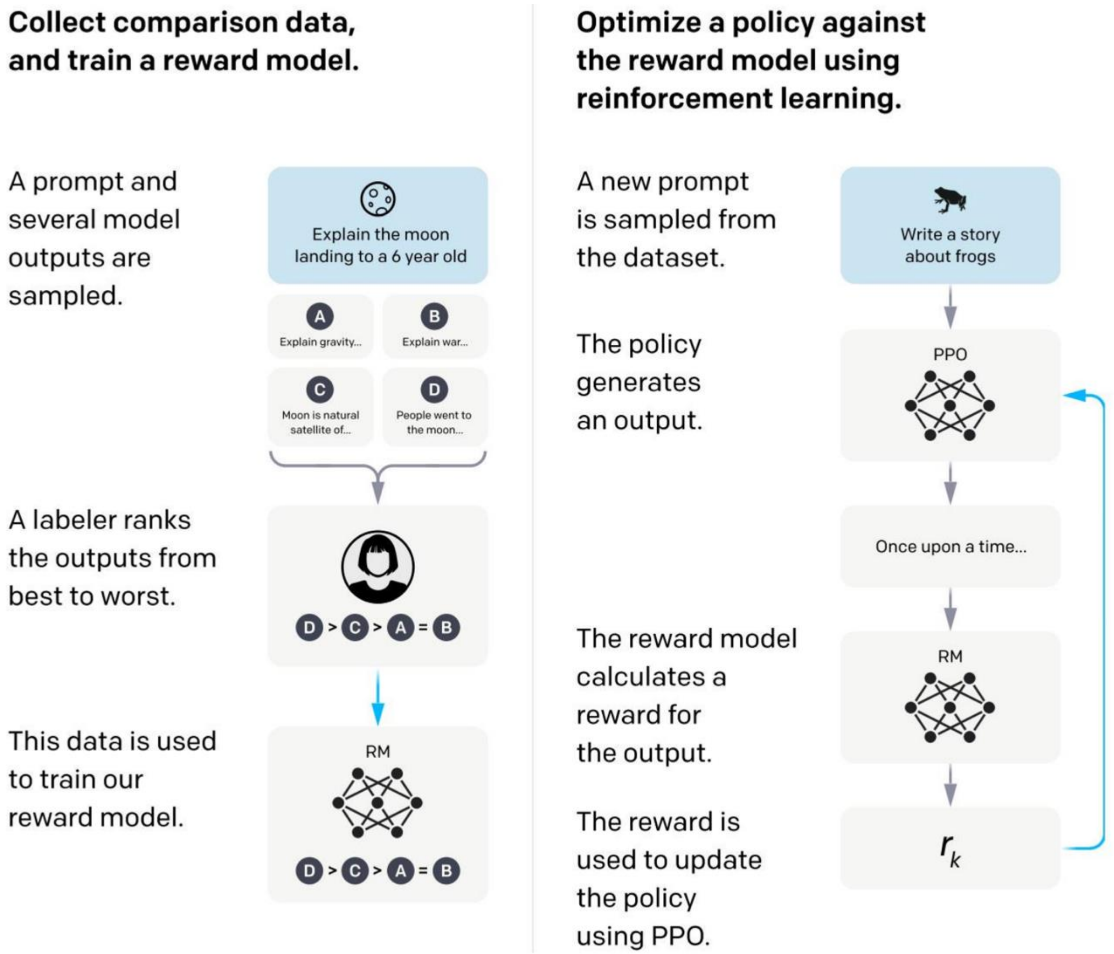
- Input and Target : Instruction + input as input with the target in SFT
- Objective function : Loss computed only for target tokens in SFT
- Purpose : good SFT builds models that can do many unseen tasks

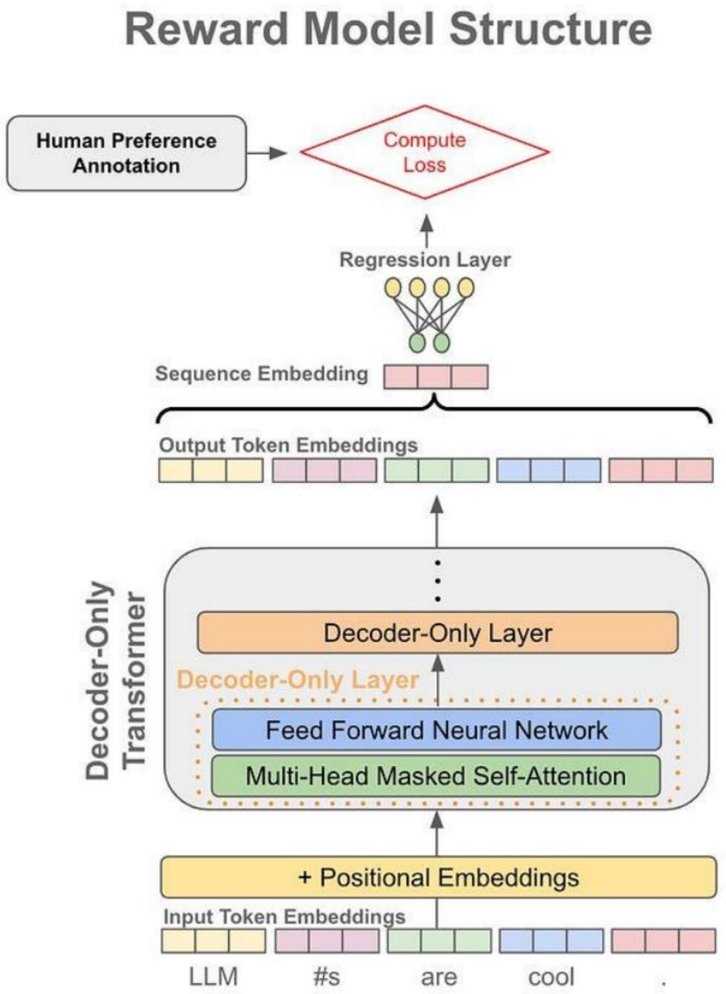
Expensive data labeling···

Not optimal answer for human ···

Reinforce learning with human feedback

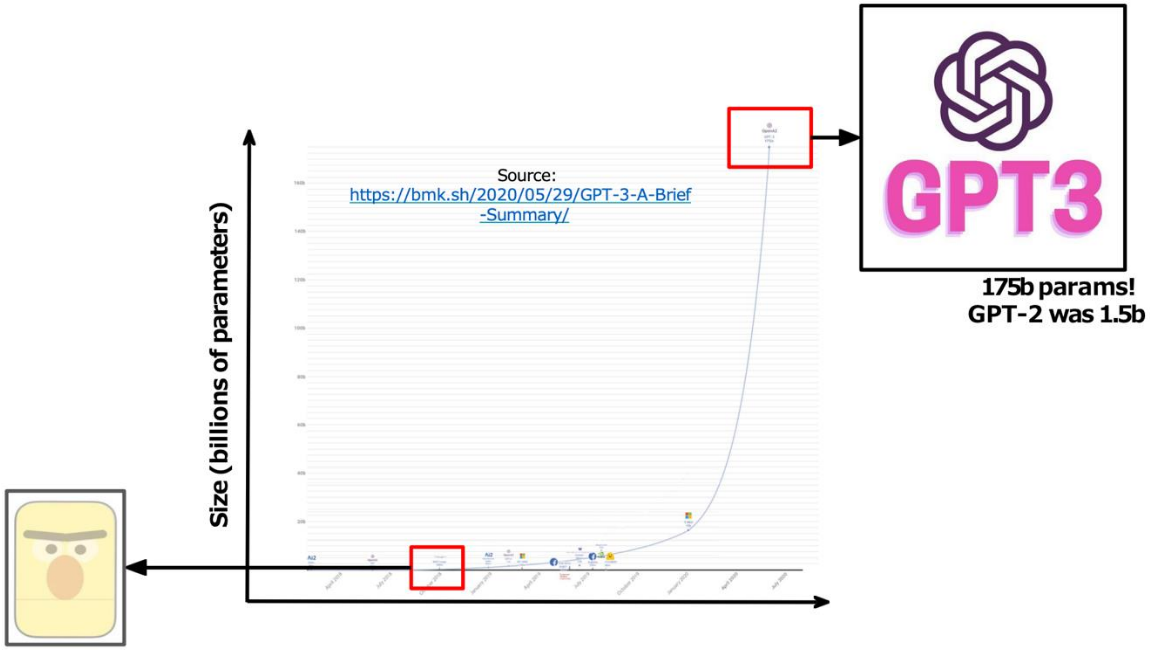
LLMs may produce text that can cause direct harm – allowing easy access to dangerous information. Therefore, LLMs should be trained to produce outputs that align with human preferences and values.

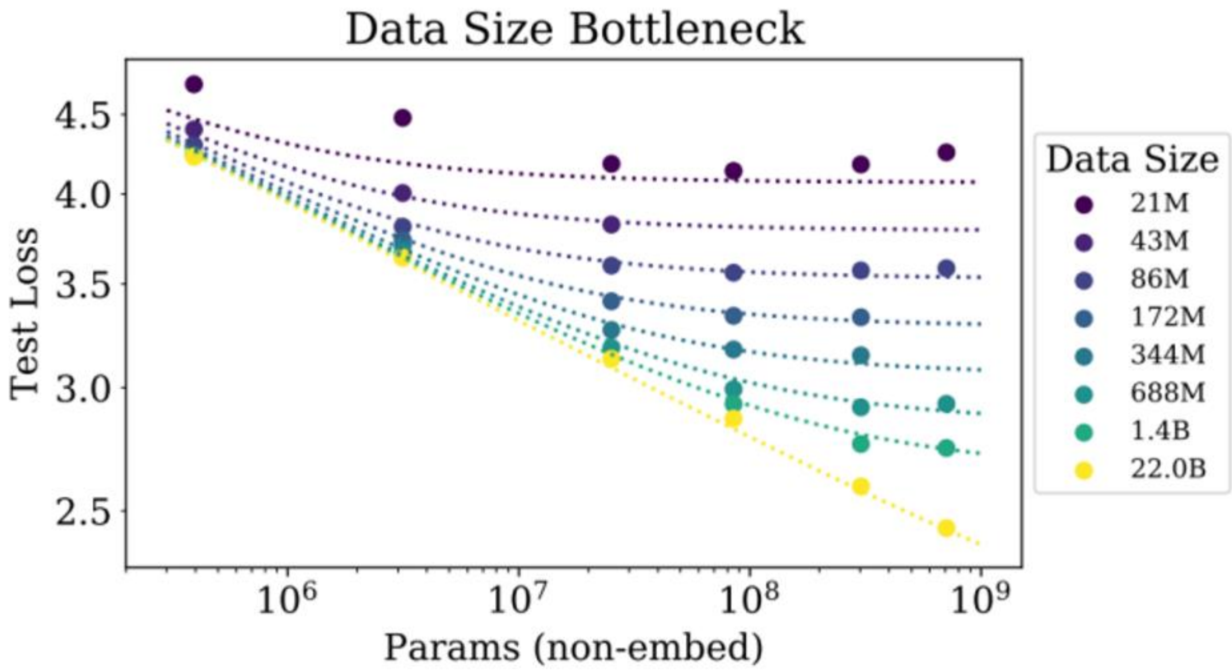




Scaling

Scaling up





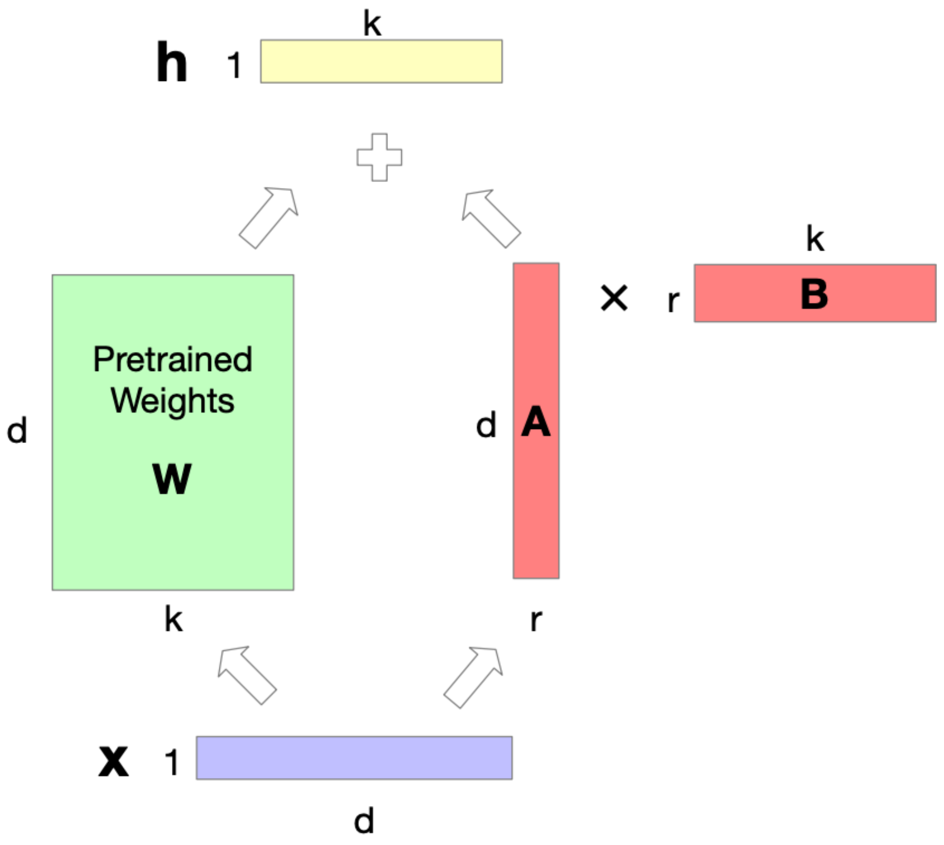
○ GPT-3 trained on text can do arithmetic problems like addition and subtraction

Larger means stronger

○ Different abilities “emerge” at different scales

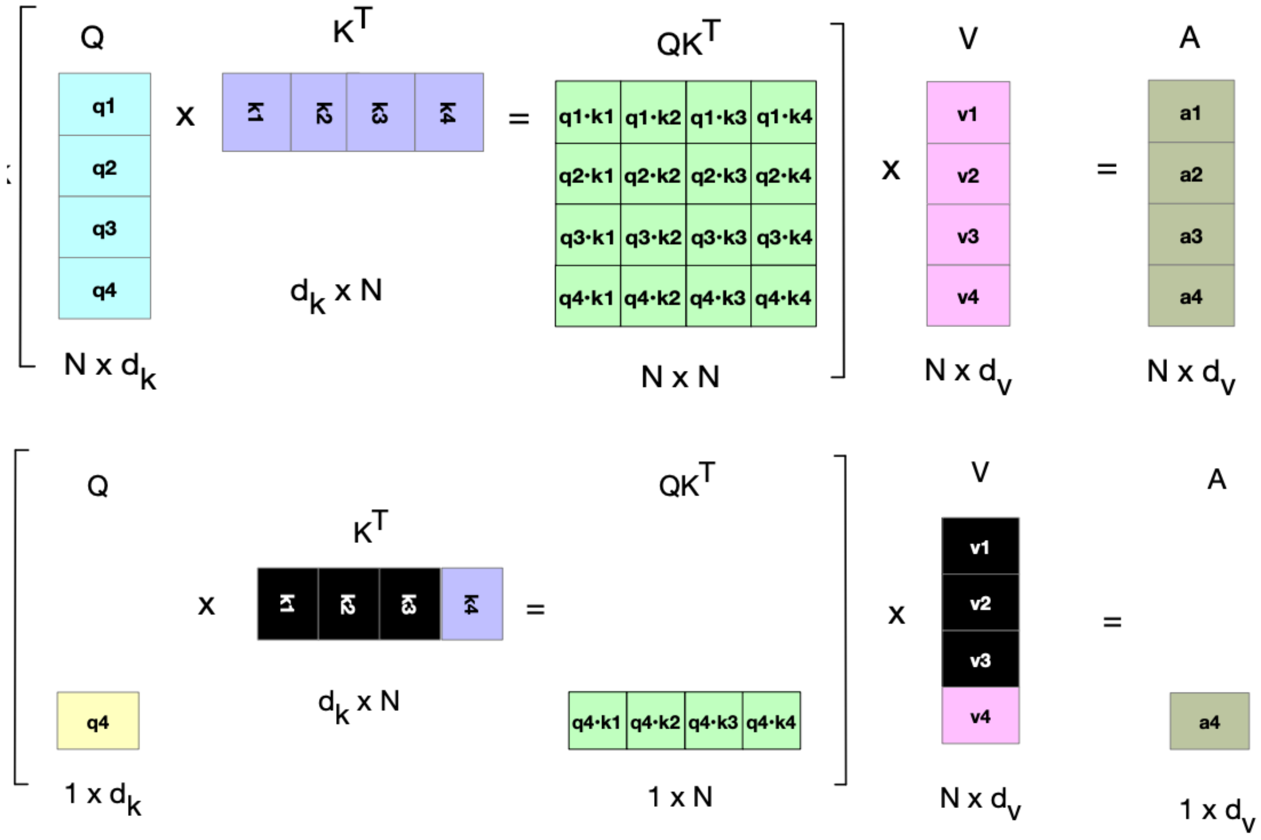
Efficiency …

Parameter-Efficient Finetuning (PEFT)



Keep dense projection matrix frozen, update the low rank matrix

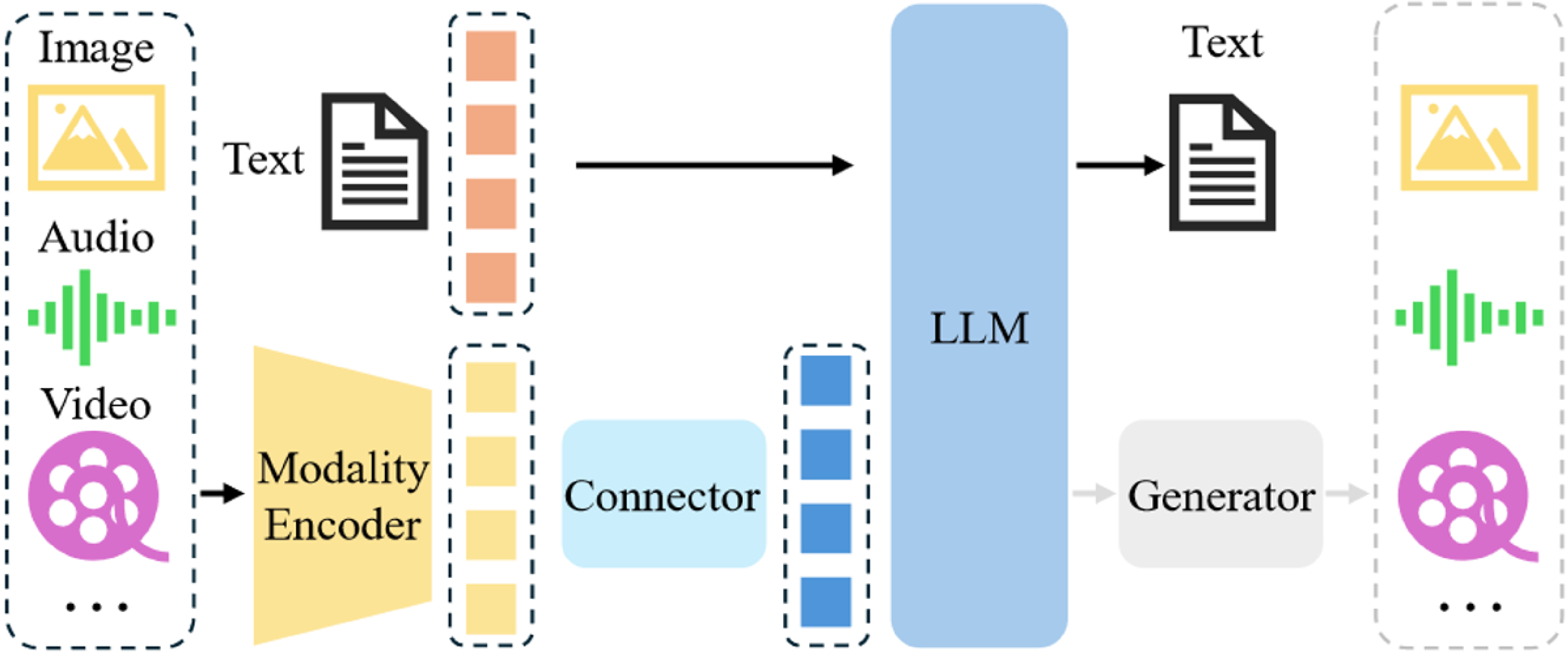
KV Cache



Avoid recompute the past keys and values during inference

Multi Modal LLM

Multi Modal LLM

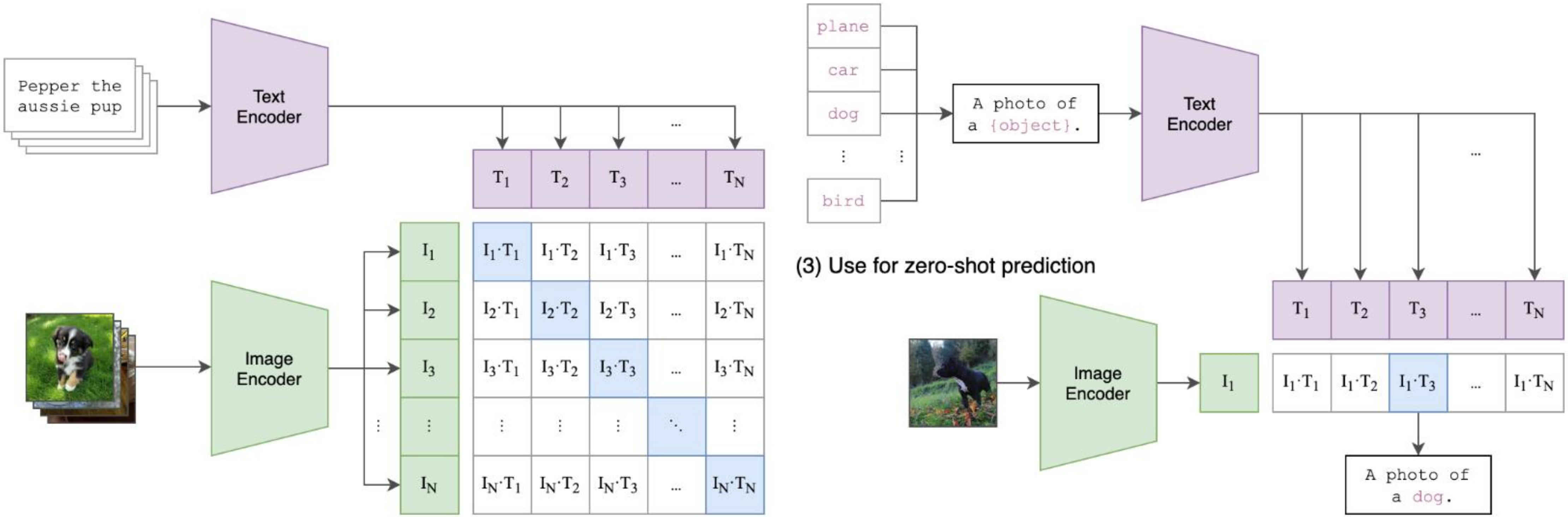


Four main components

(1) multimodal encoder (2) connector

(3) large language model (4) multimodal generator

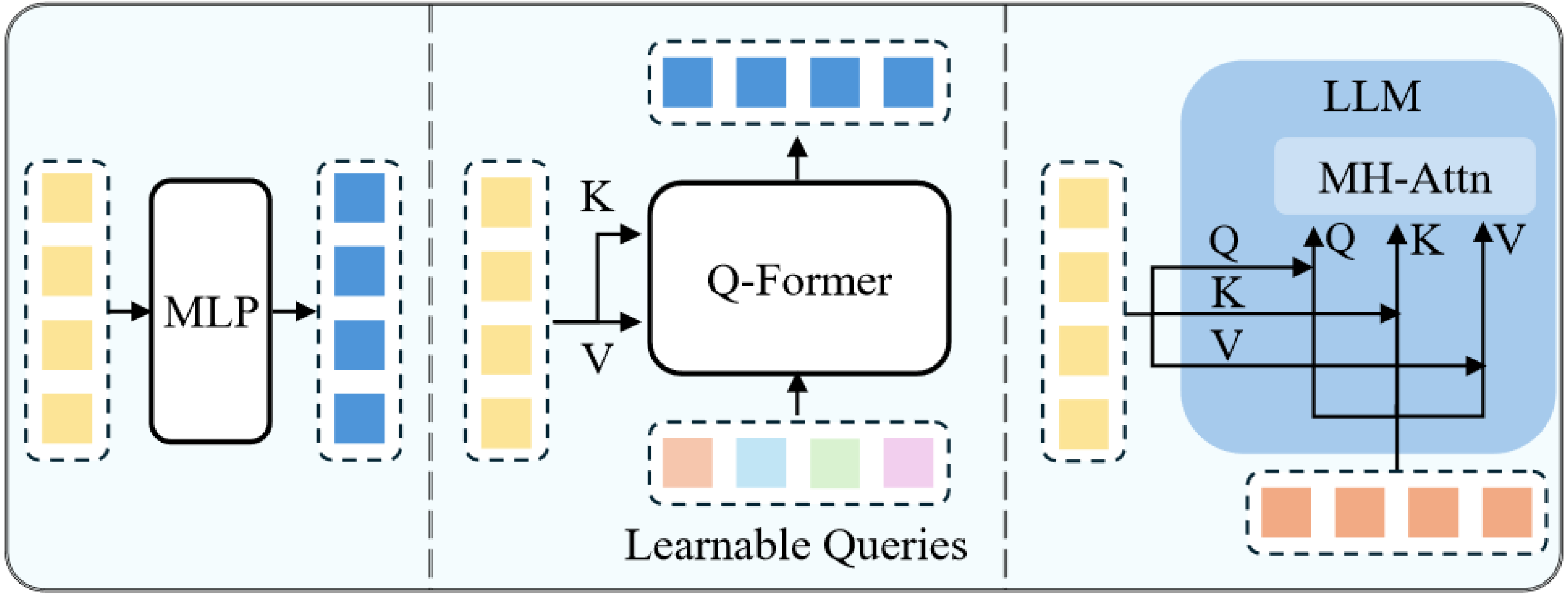
Multimodal Encoder



Key factor:

1. Encoder parameter number
2. Pretrained dataset size
3. Resolution ratio

Connector

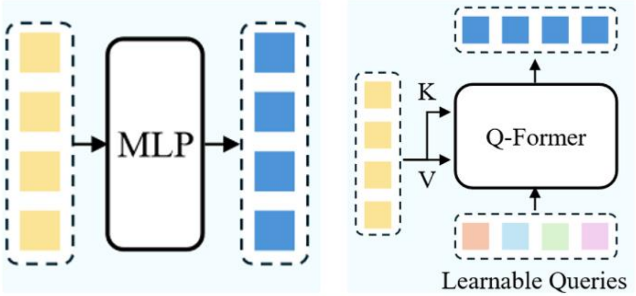


The role of connectors is to integrate multimodal information, which can be divided into token-level and feature-level based on the fusion hierarchy.

Three mainstream connectors are:
MLP (token-level), Q-Former (token-level), and multi-head attention (feature-level).

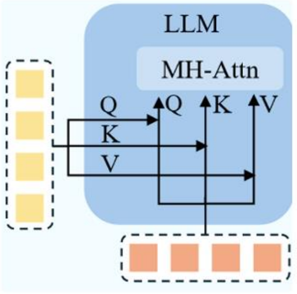
Connector

Token Level

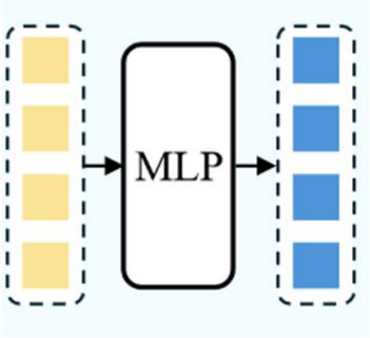


VS

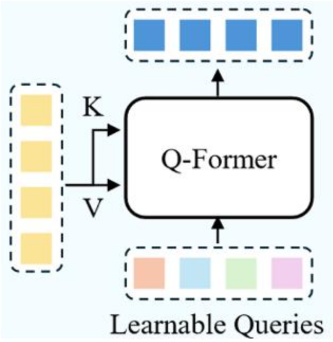
Feature Level



Token-level performs better in VQA benchmark tests (VQA: Visual Question Answering)



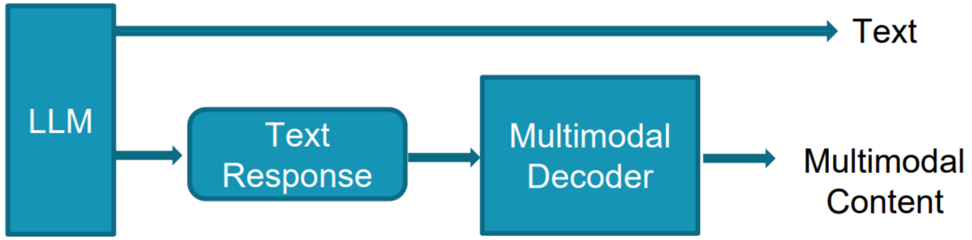
VS



The type of the connector is far less important than the number of visual tokens and the input resolution.

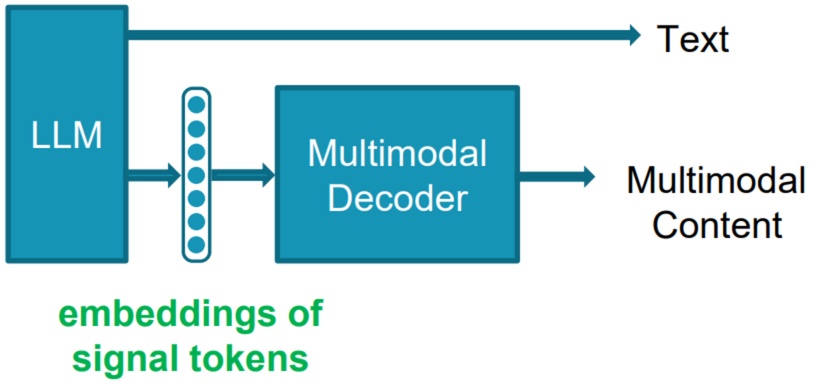
Multimodal Generator

There are two implementation methods for the multimodal decoder:
(1) Using the text output as the input.
(2) Using the embeddings corresponding to specific tokens as the input.



Advantages:
•Efficient, no need to fine-tune the Large Language Model (LLM).
•High lower limit of performance.

Disadvantages:
•Lack the ability of end-to-end fine-tuning.
•Low upper limit of performance, and some multimodal tasks cannot be translated into text.



Characteristics:
It can be fine-tuned in an end-to-end manner.
It has a high upper limit of performance and can convey information that cannot be carried by text, such as: visual spatial relationships.

Further⋯

LLM generate image?