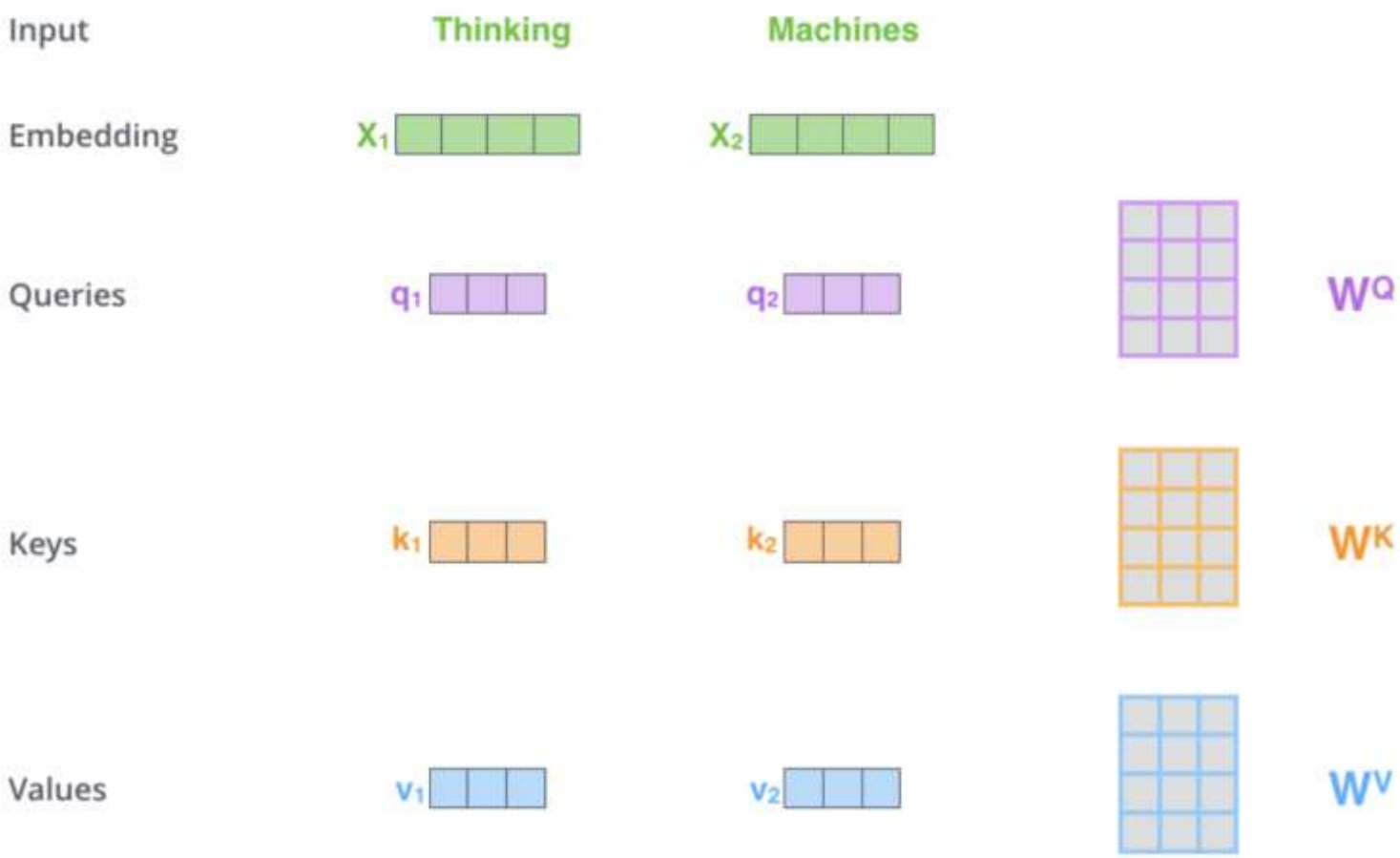
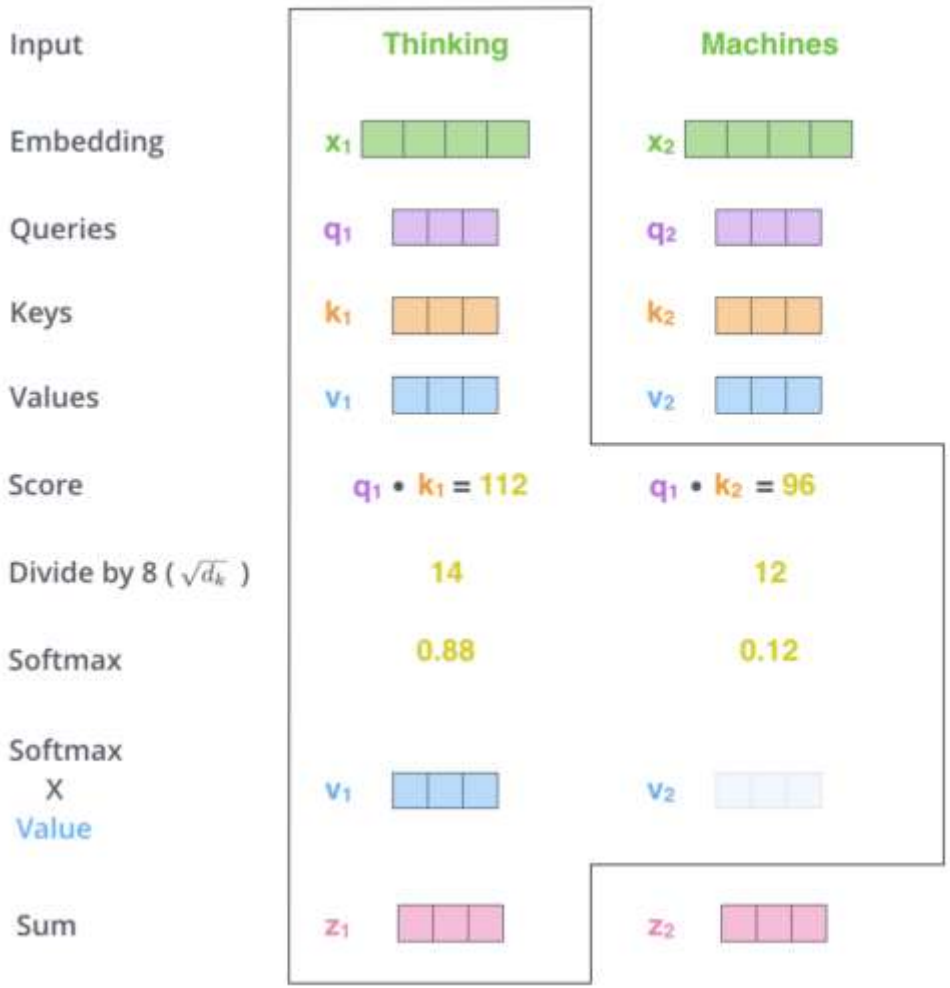
Introduction to LLMs

Attention

Attention

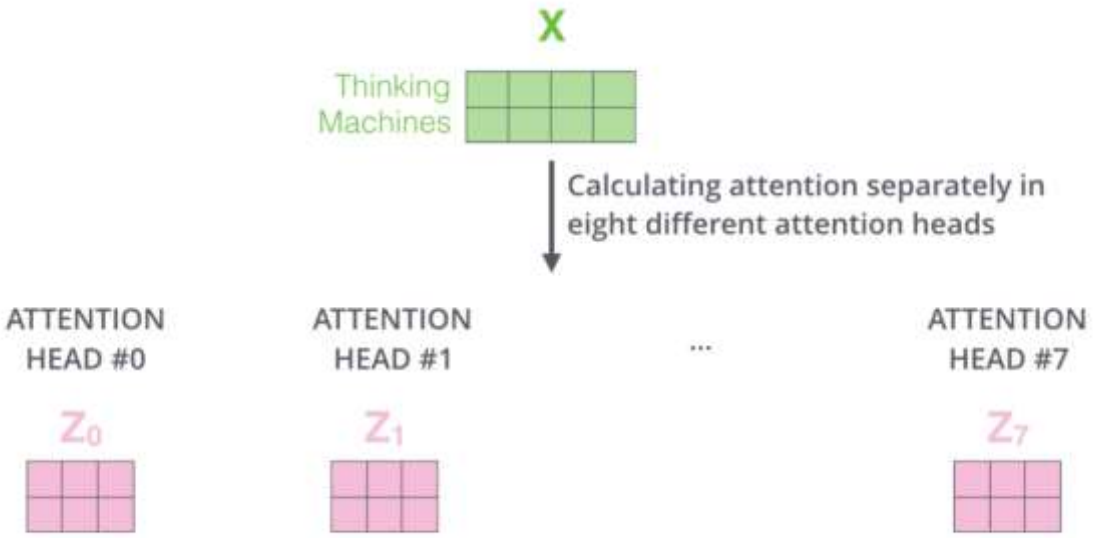

Input
Thinking
Machines
Embedding
X1
X2
Queries
q1
q2
WQ
Keys
k1
k2
WK
Values
v1
v2
WV

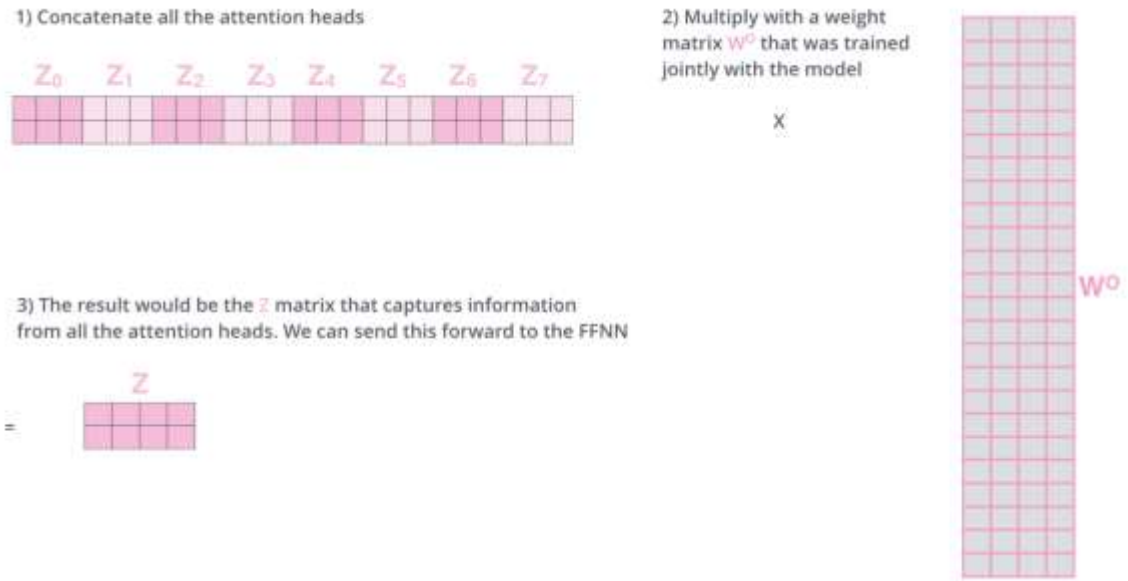
Attention



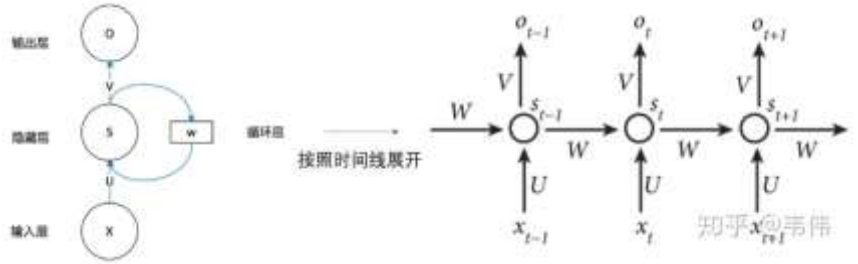
$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

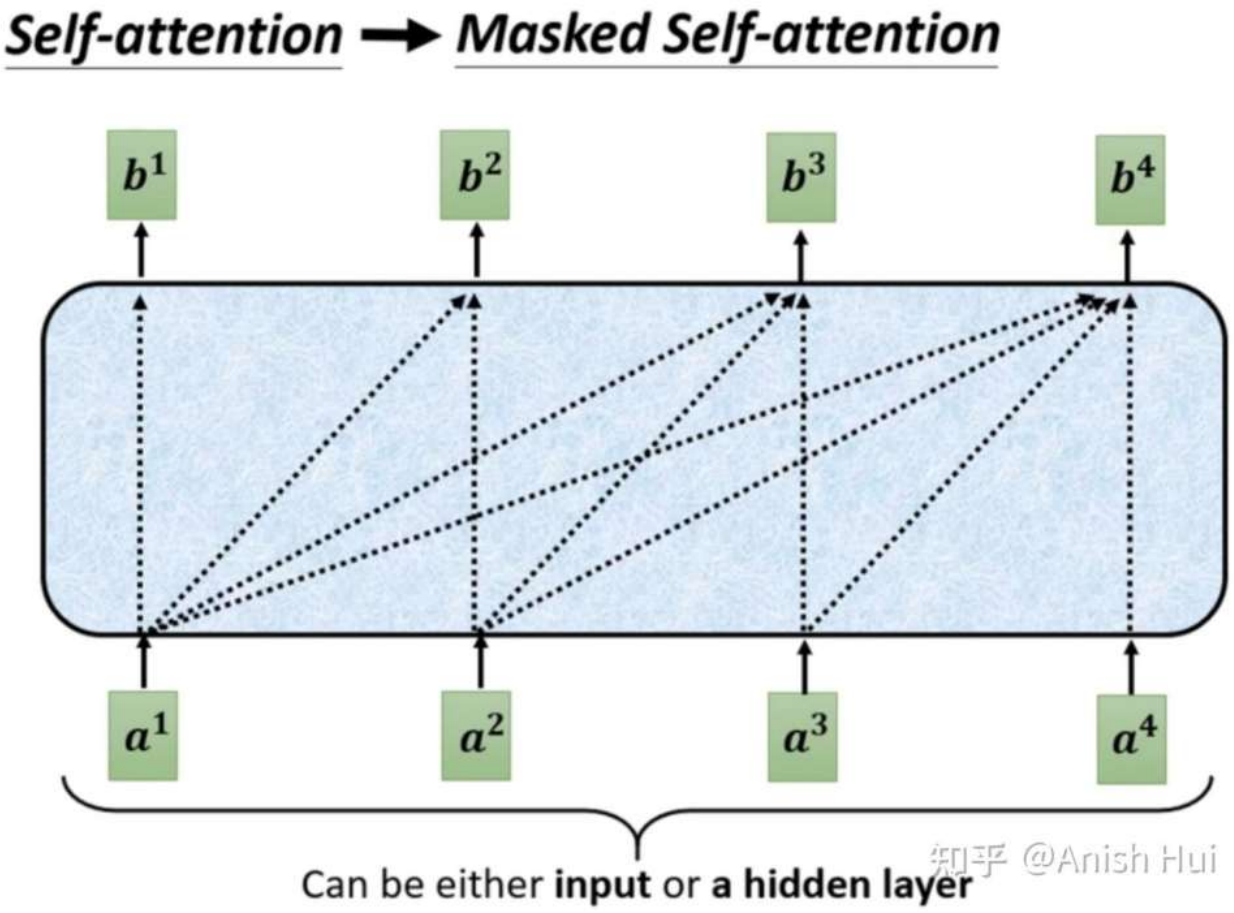
Multi-head attention





Masked self-attention

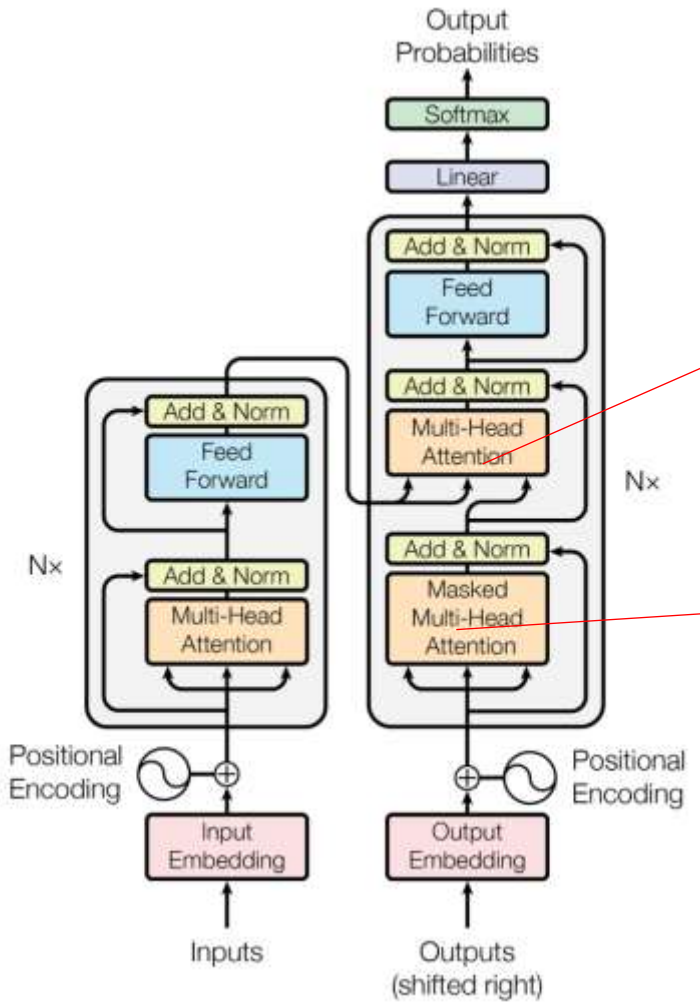




Ensures causal dependency

Transformer Models

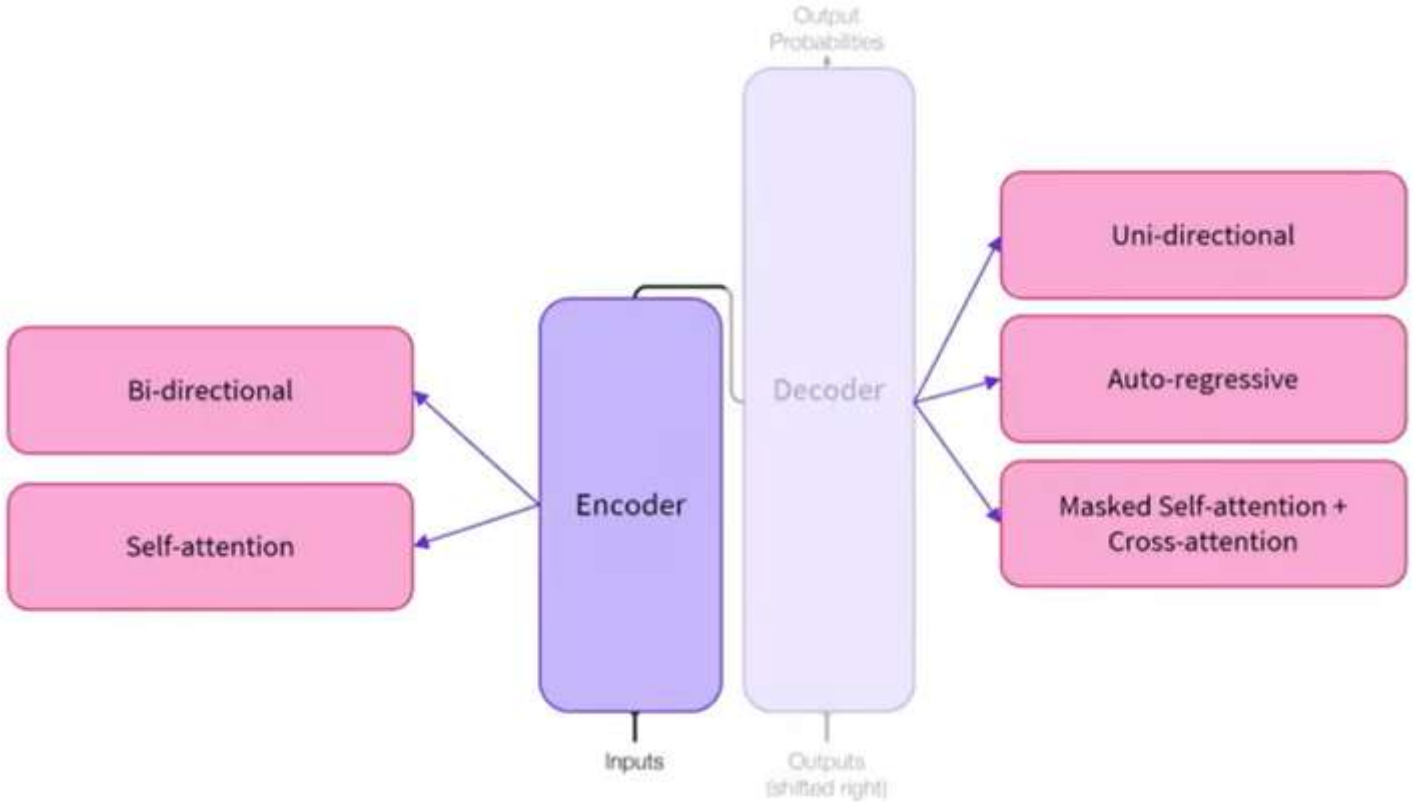
Encoder-Decoder



Cross-attention

Masked self-attention

Encoder-only



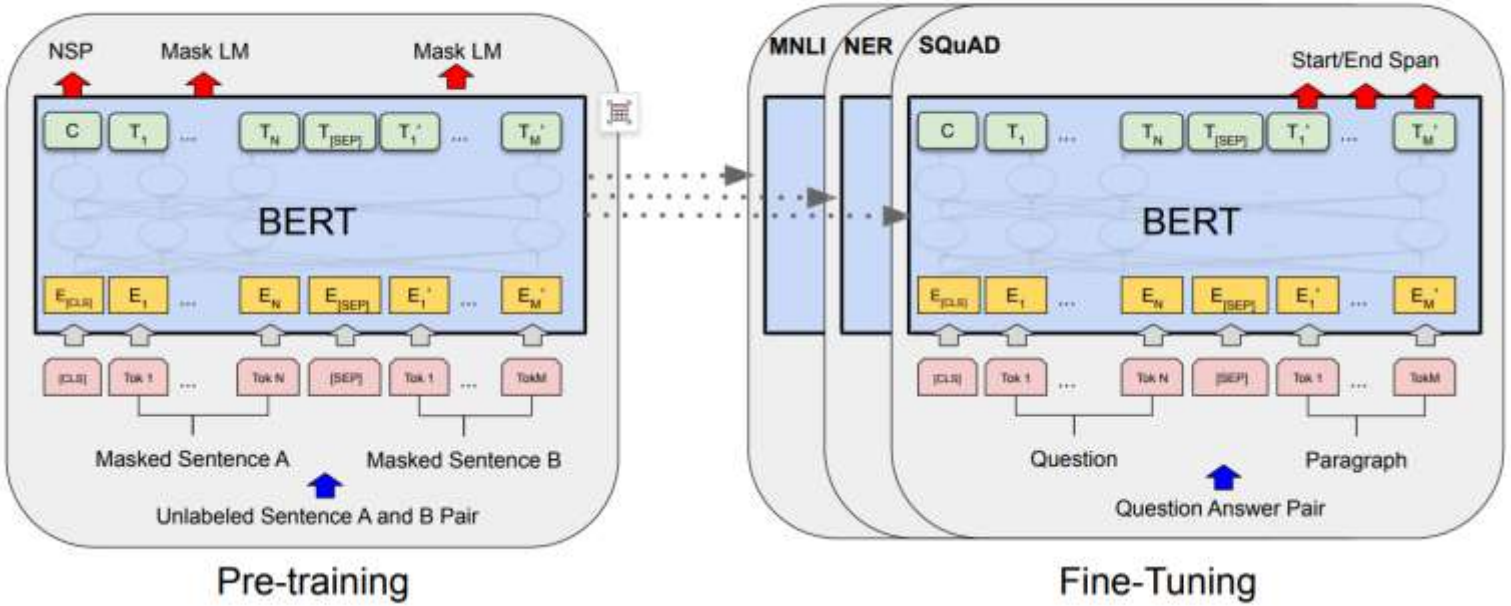
Training method:
predict masked words

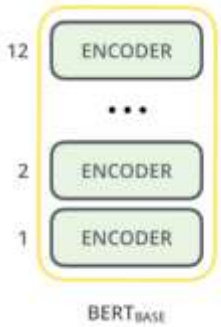
Advantages:
comprehension

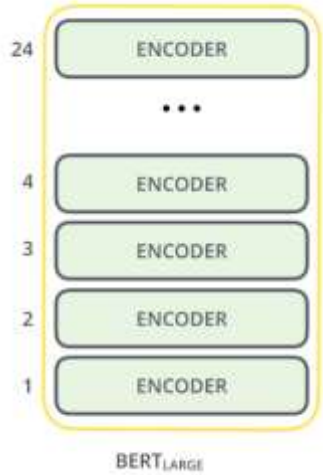
Disadvantages:
generation

Bert

BERT: Bidirectional Encoder Representations from Transformers







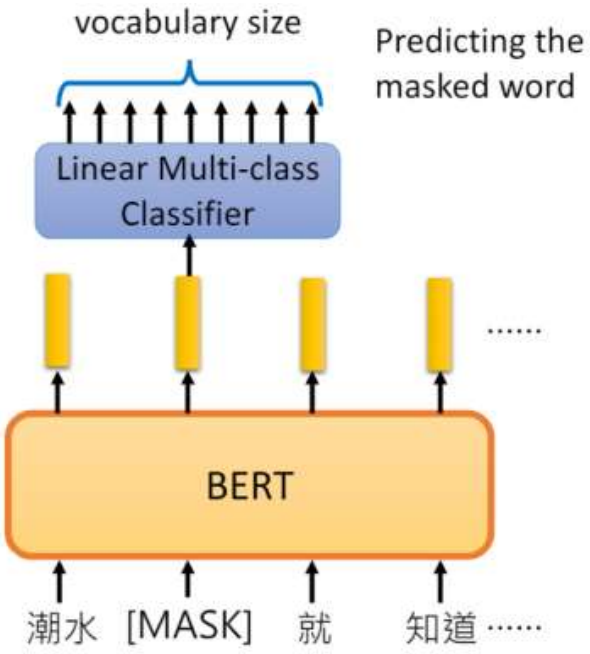
Encoder-only structure

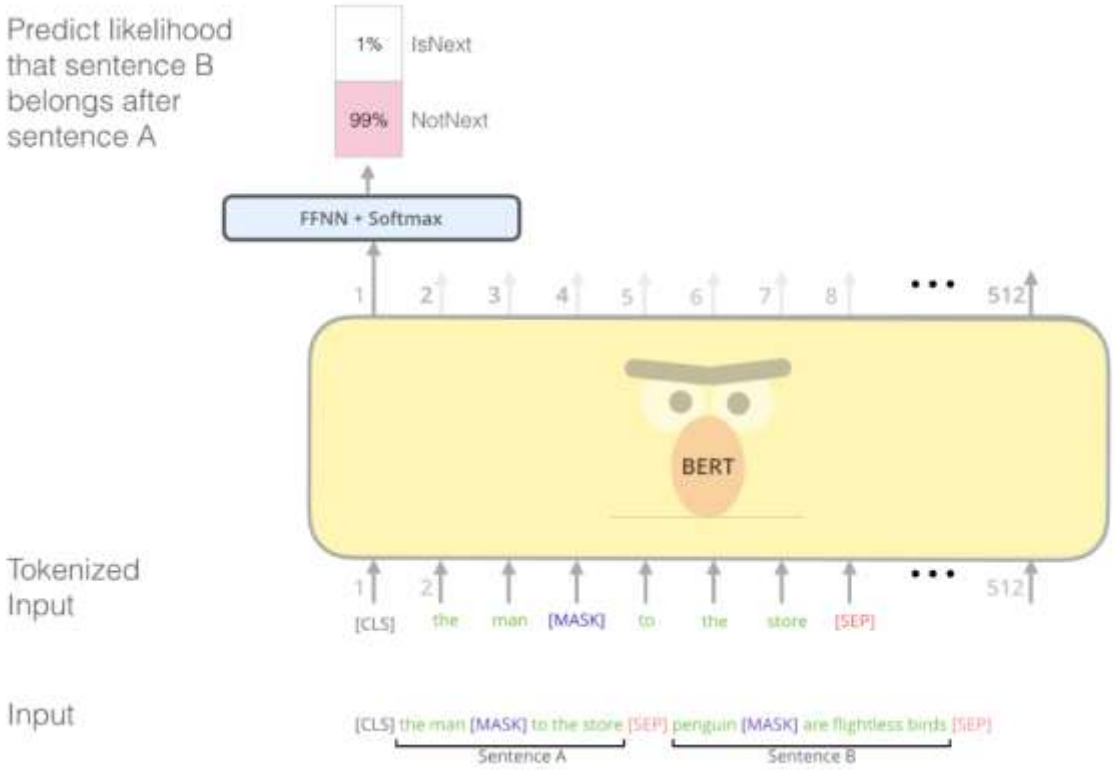
Bert

Pretrain:

1、Masked LM

2、Next sentence prediction





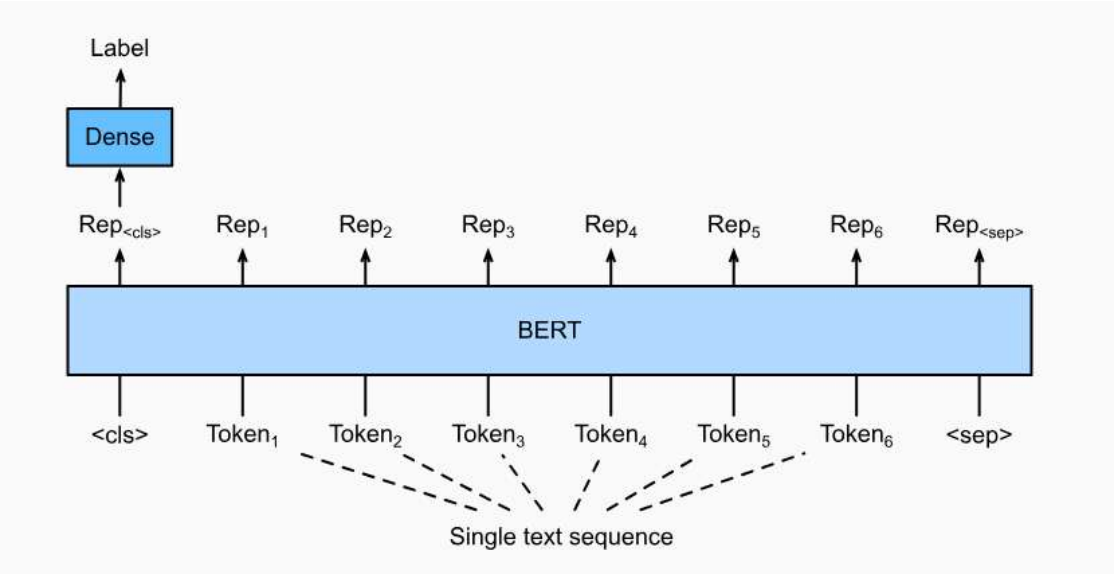
Bert

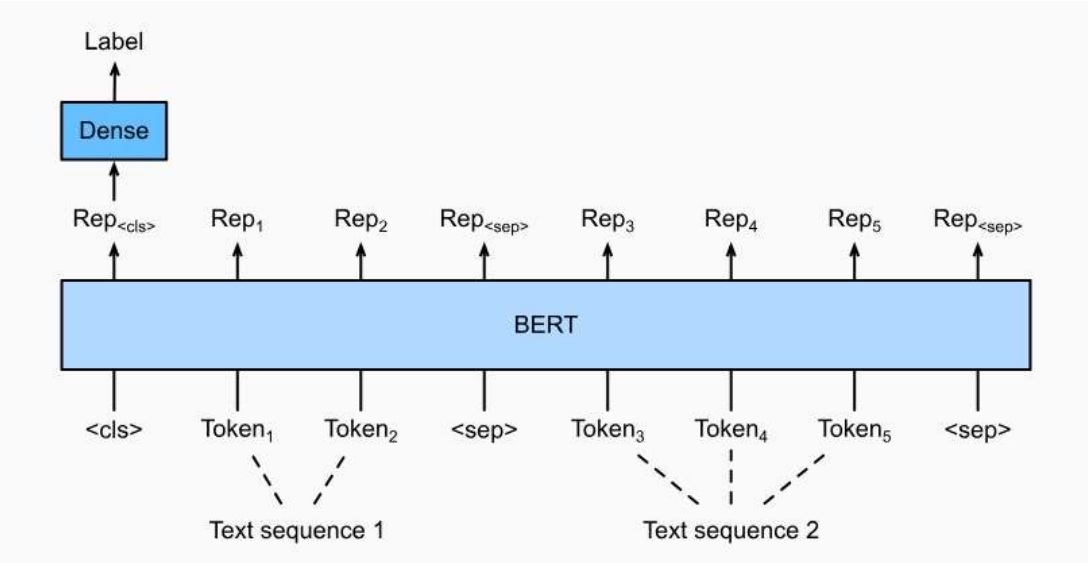
Fine-tuning：

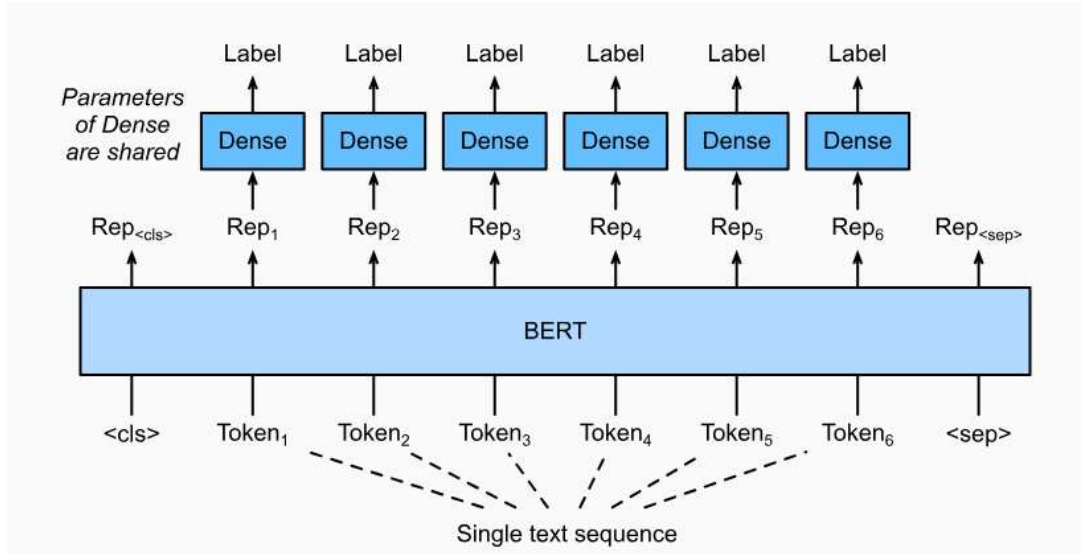
1、Single text classification

2、Text Pair Classification or Regression

3、Text Tagging



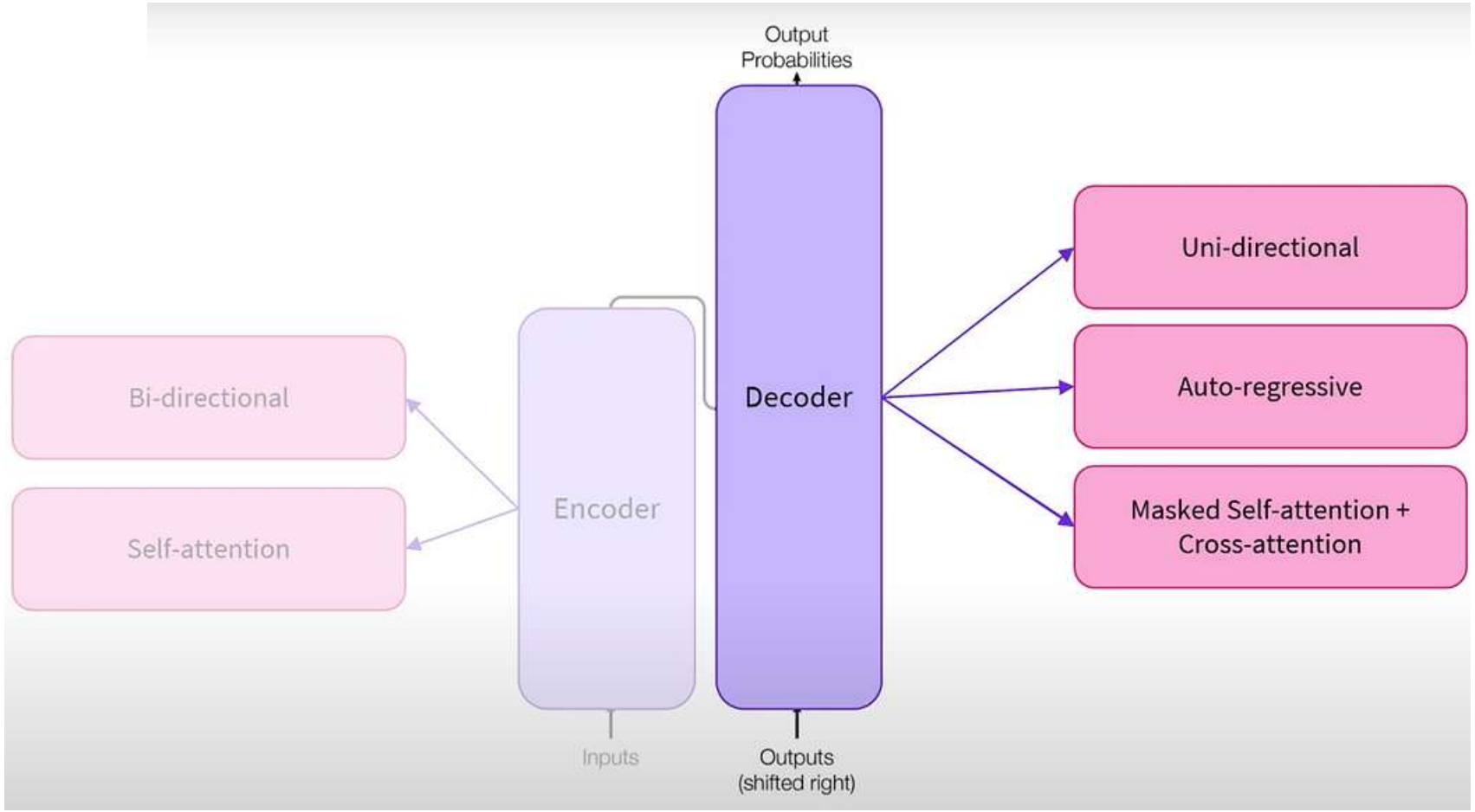




Highly dependent on fine-tuning

❌ Generation/Zero-shot/Few-shot

Decoder only

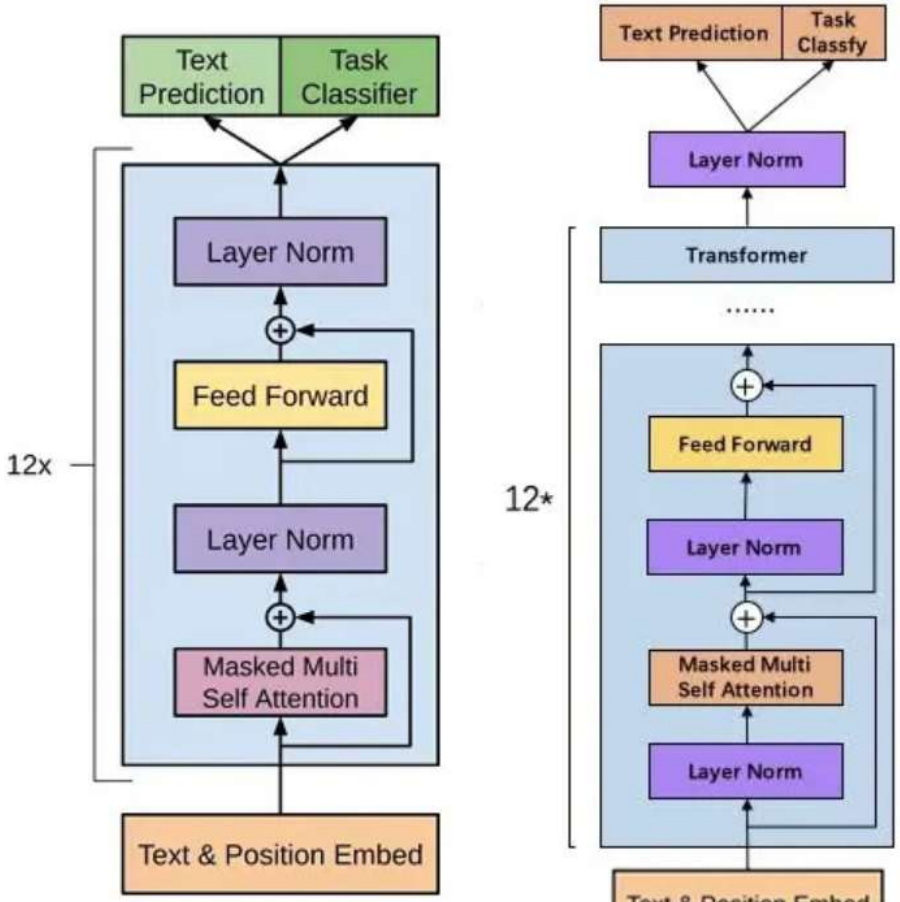


Training method:
autoregressively predict the next token

Advantages:
generation

Disadvantages:
suboptimal semantic comprehension

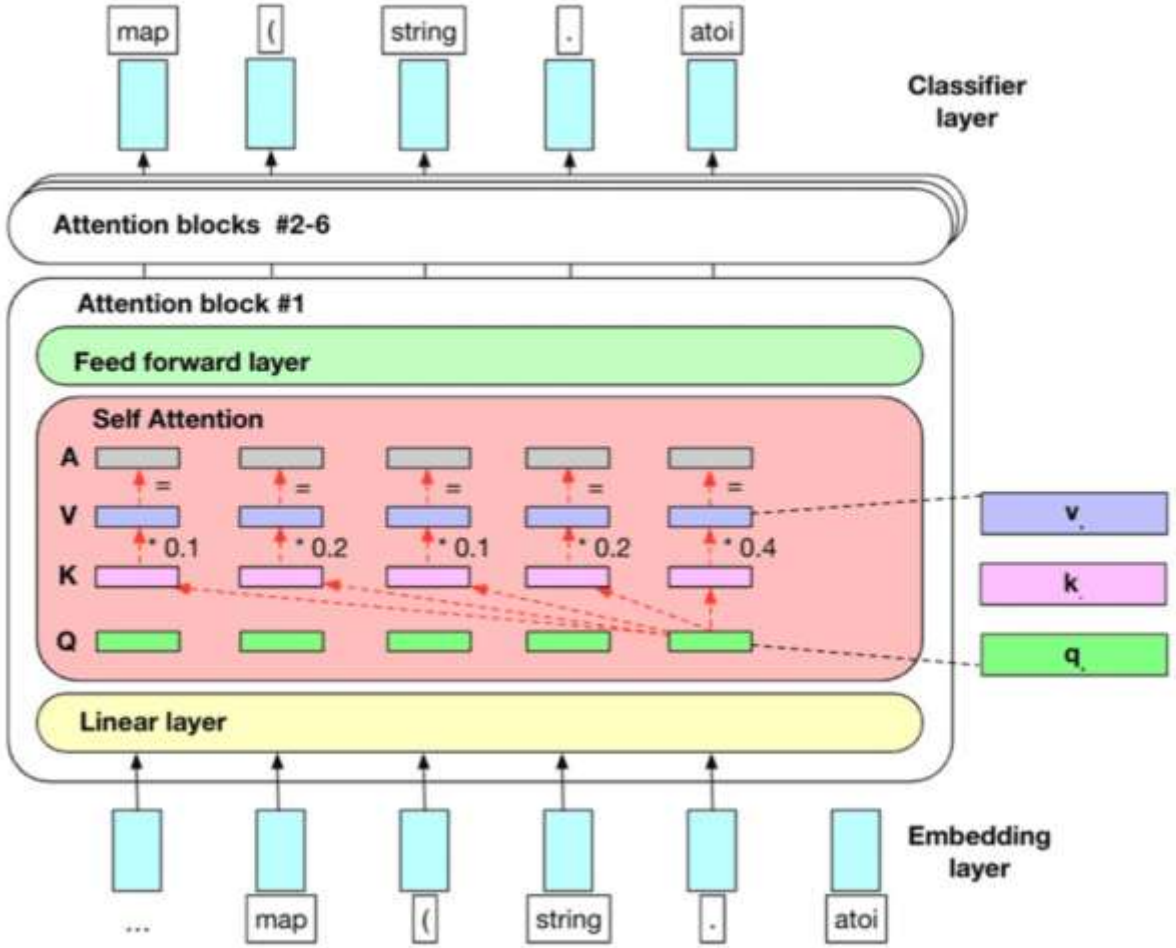
GPT-1/2



Gpt1:
Training: unsupervised pretrain
+ supervised fine-tuning

Gpt2:
Training: unsupervised pretrain

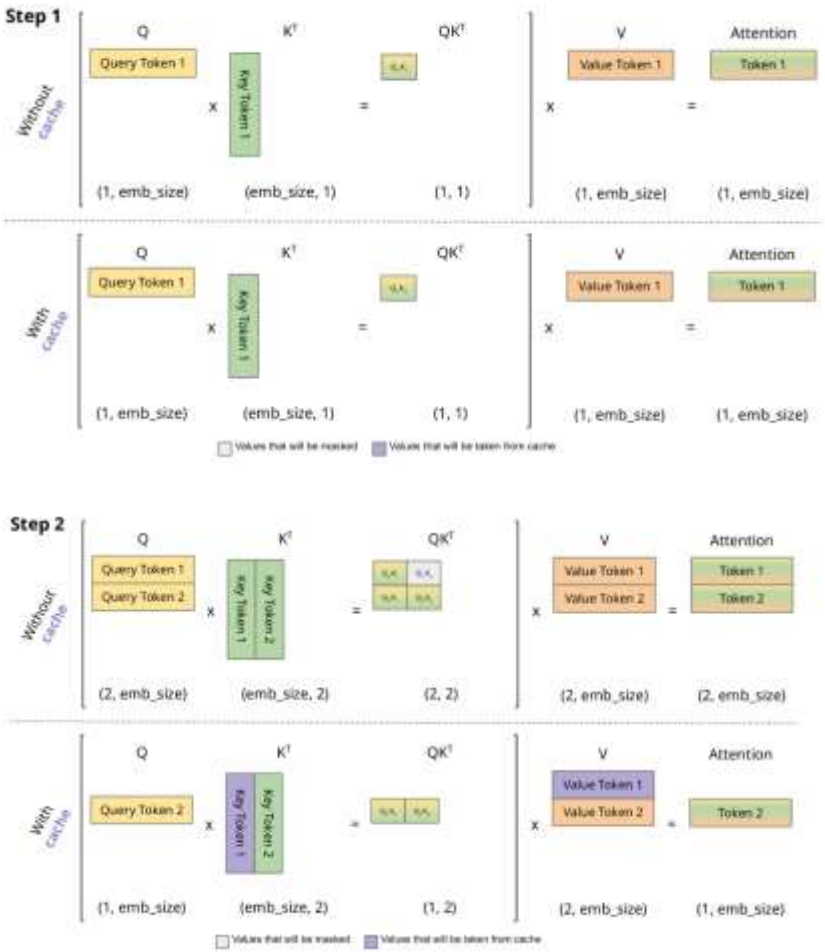
Pretrain

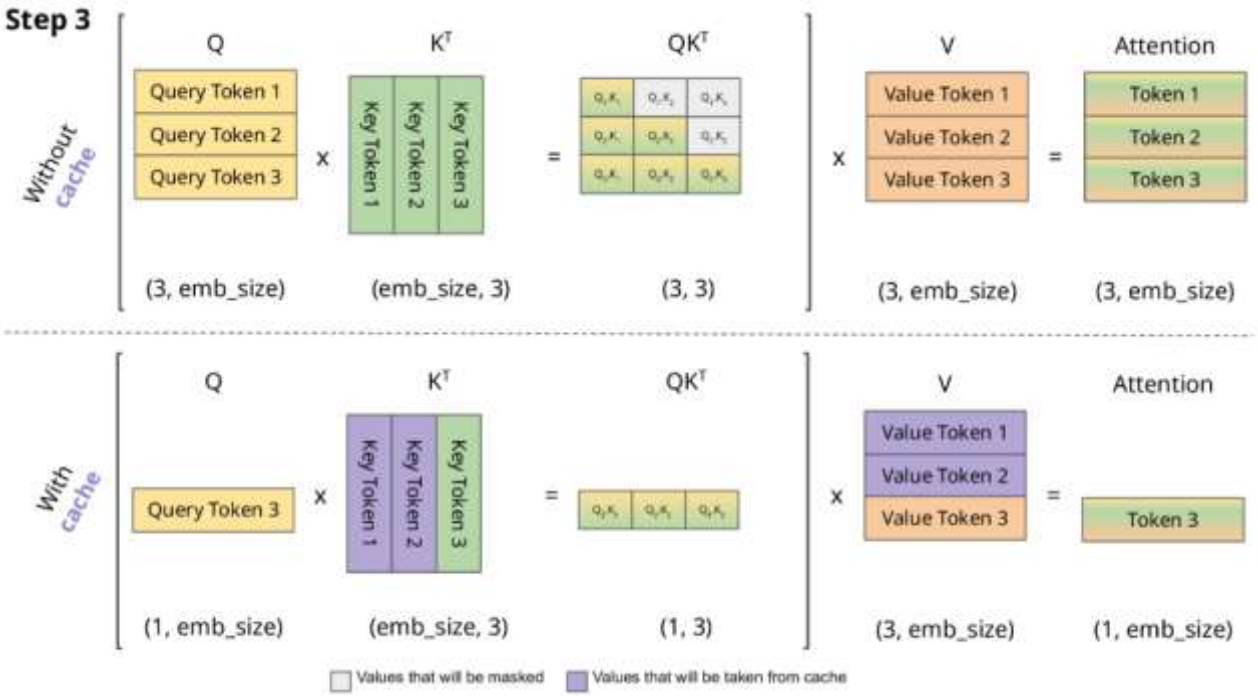


Cross-entropy loss

$$L_1(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$$

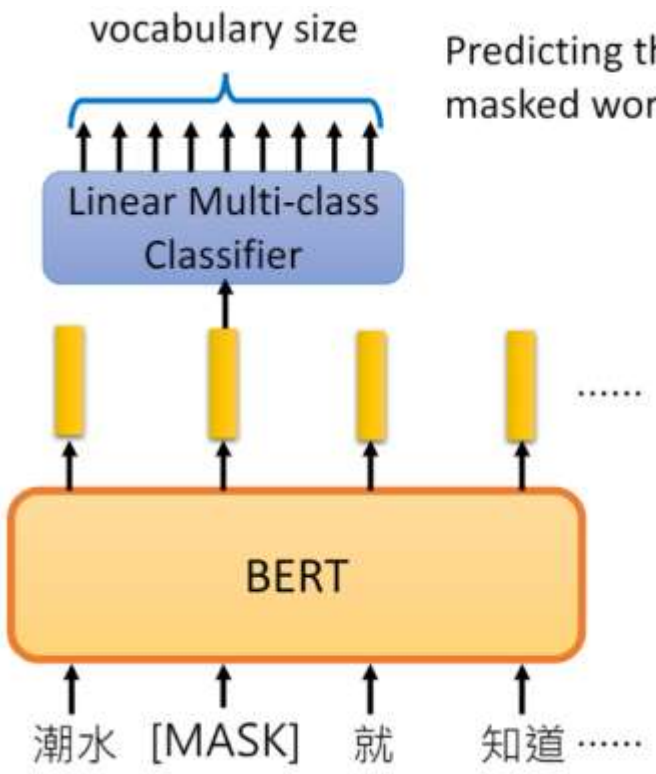
KV cache

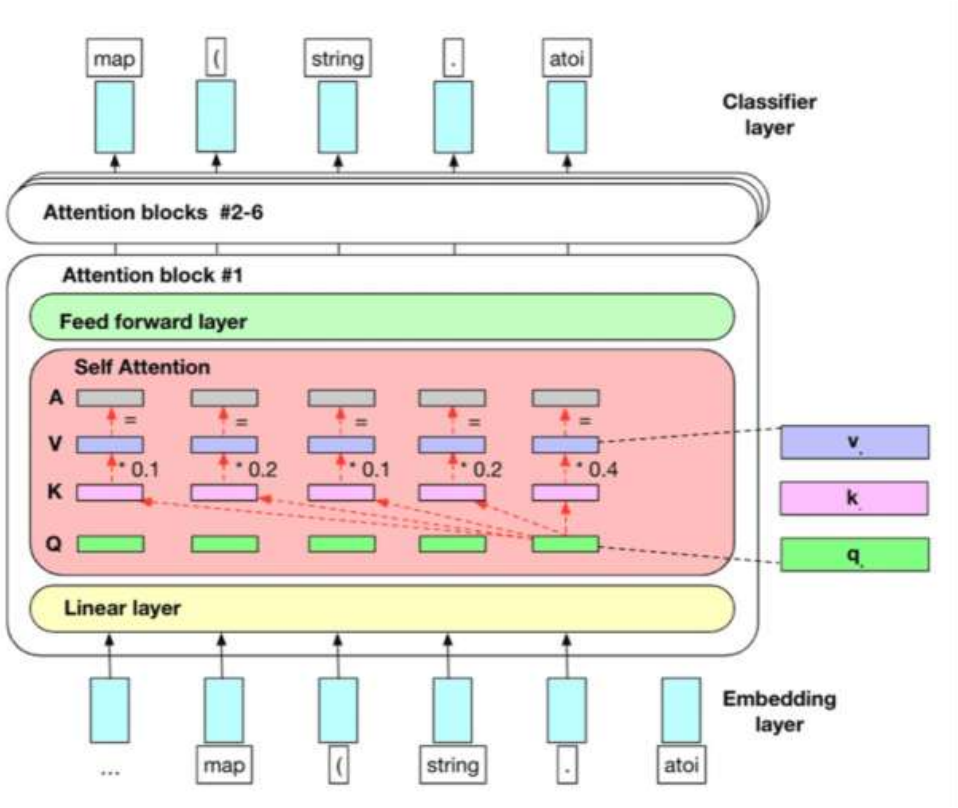




$O(n^2)$ -> O(n)

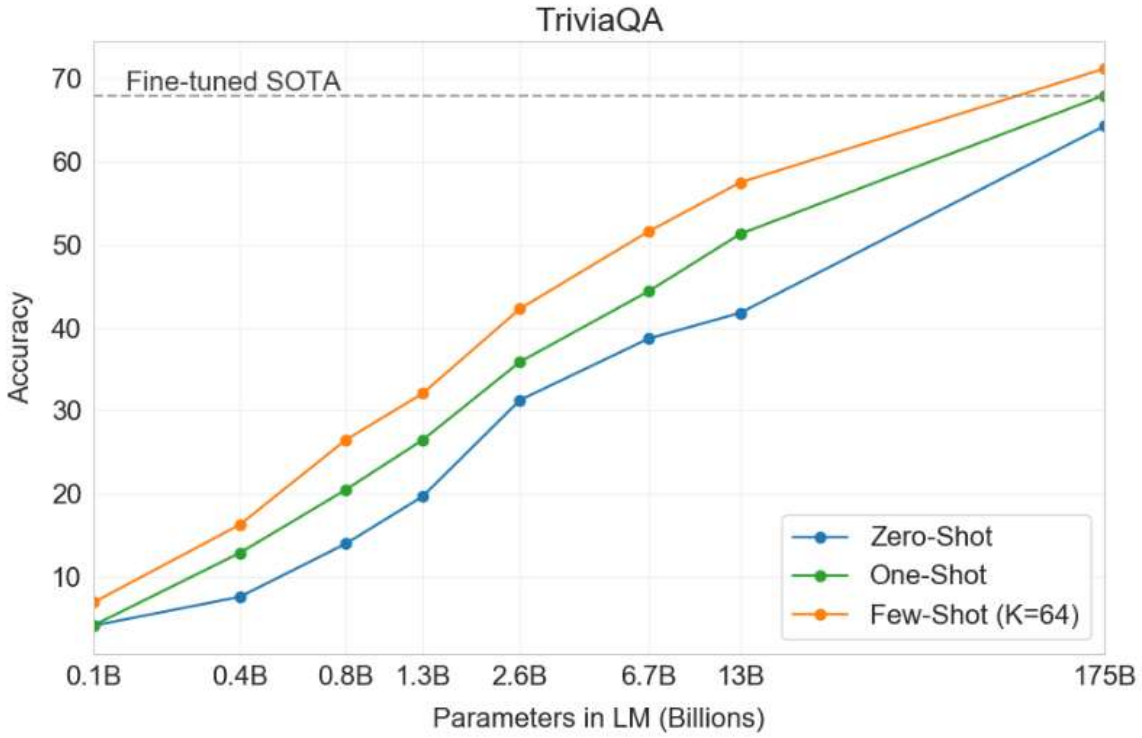
Scale up——Encoder or Decoder?



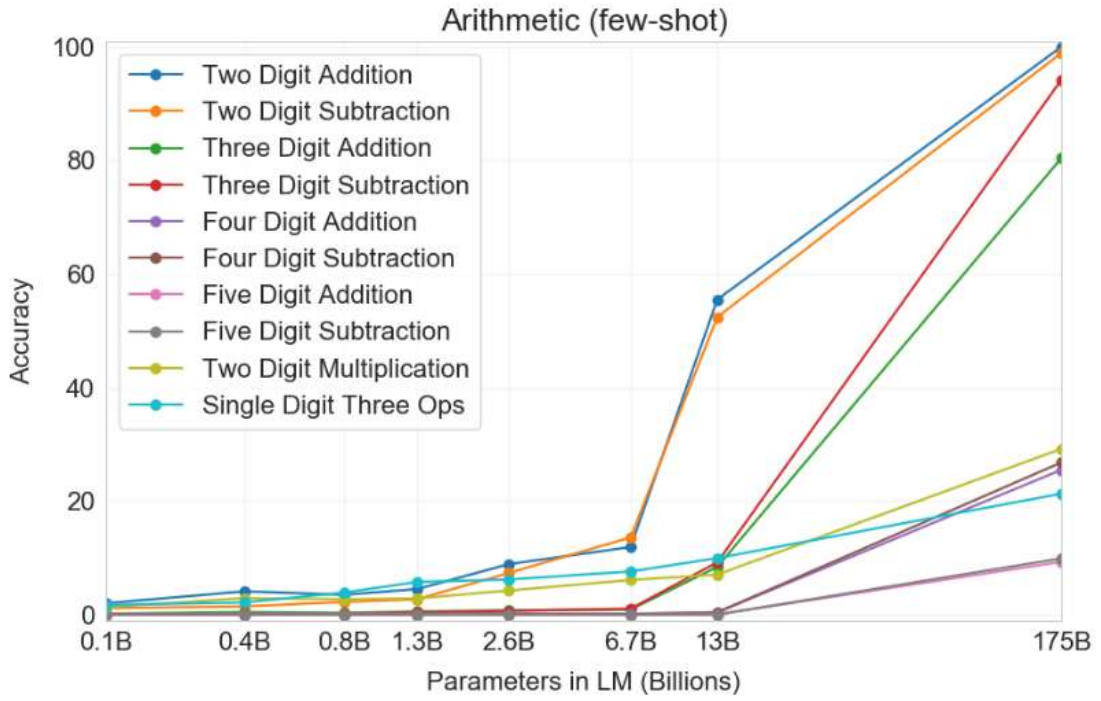


1、Auto-regression training makes full use of data
2、Causal attention enables KV cache

GPT-3——Scaling up



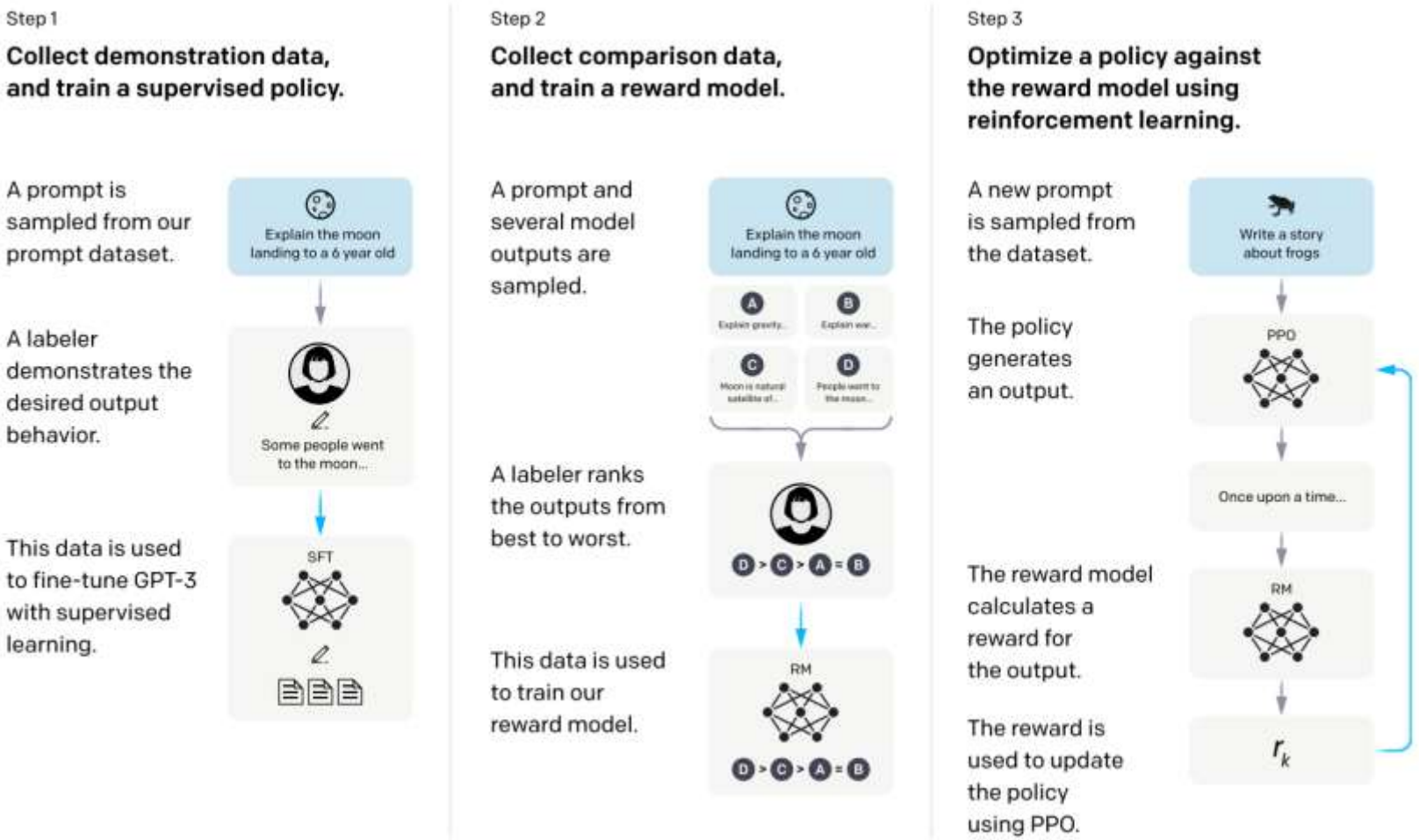
The larger GPT-3's parameter count, the higher its zero-shot/few-shot QA accuracy



arithmetic abilities of GPT-3

Instruct-GPT(RLHF)

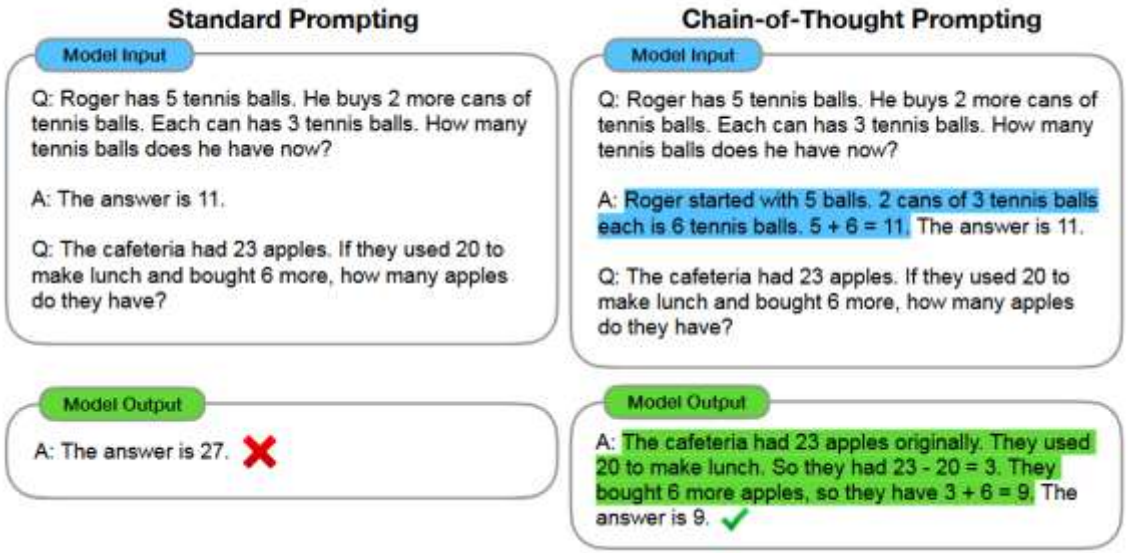
trained to produce outputs that align with human preferences



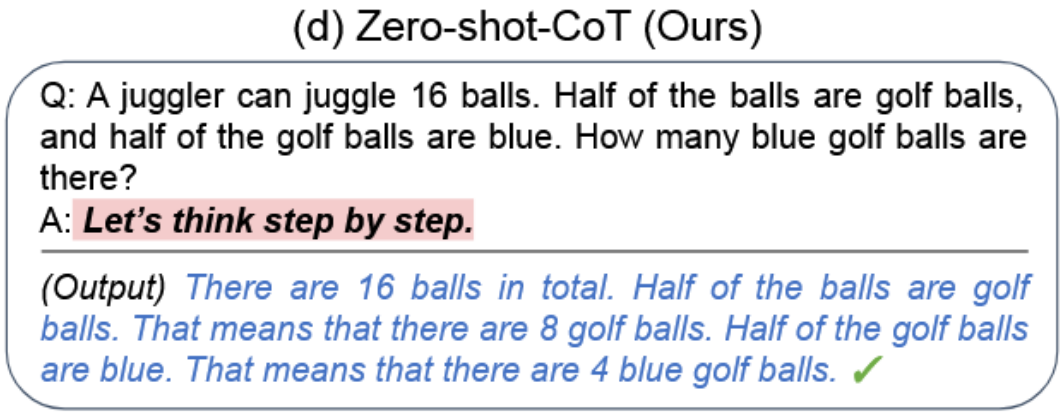
Reasoning

Cot

The reasoning ability of large language models can be unlocked by simple methods

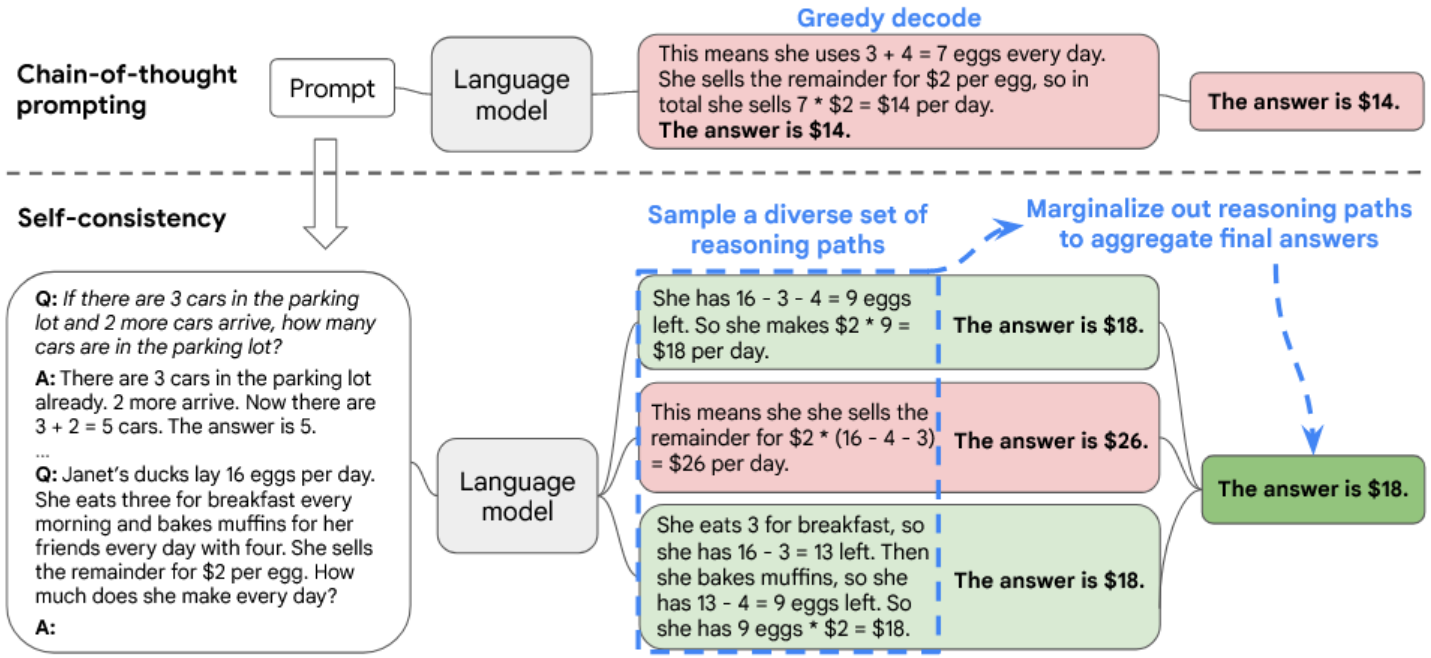


Few-shot cot



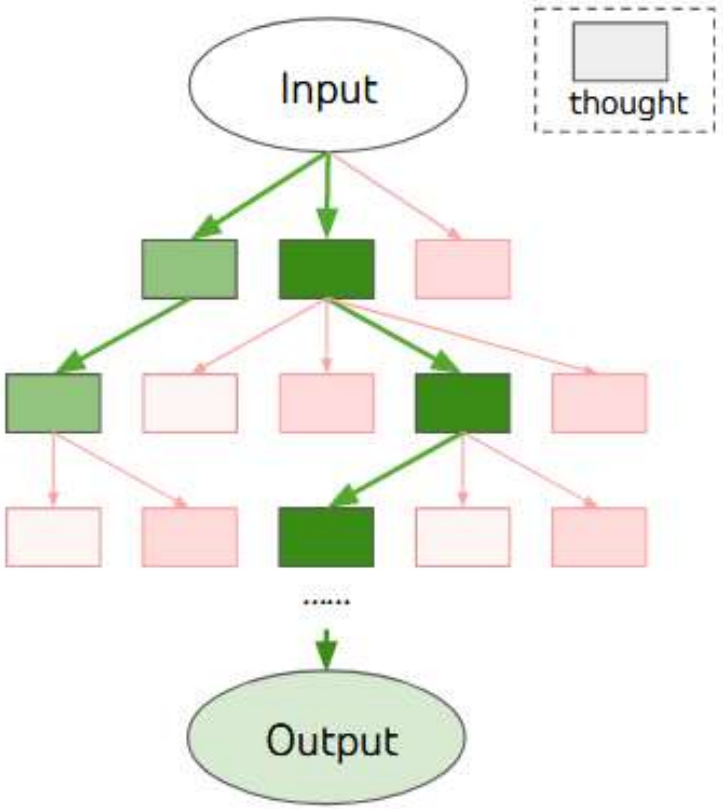
Zero-shot cot

Self-Consistency Cot



Self-consistency leverages the intuition that complex reasoning tasks typically admit multiple reasoning paths that reach a correct answer.

Tot



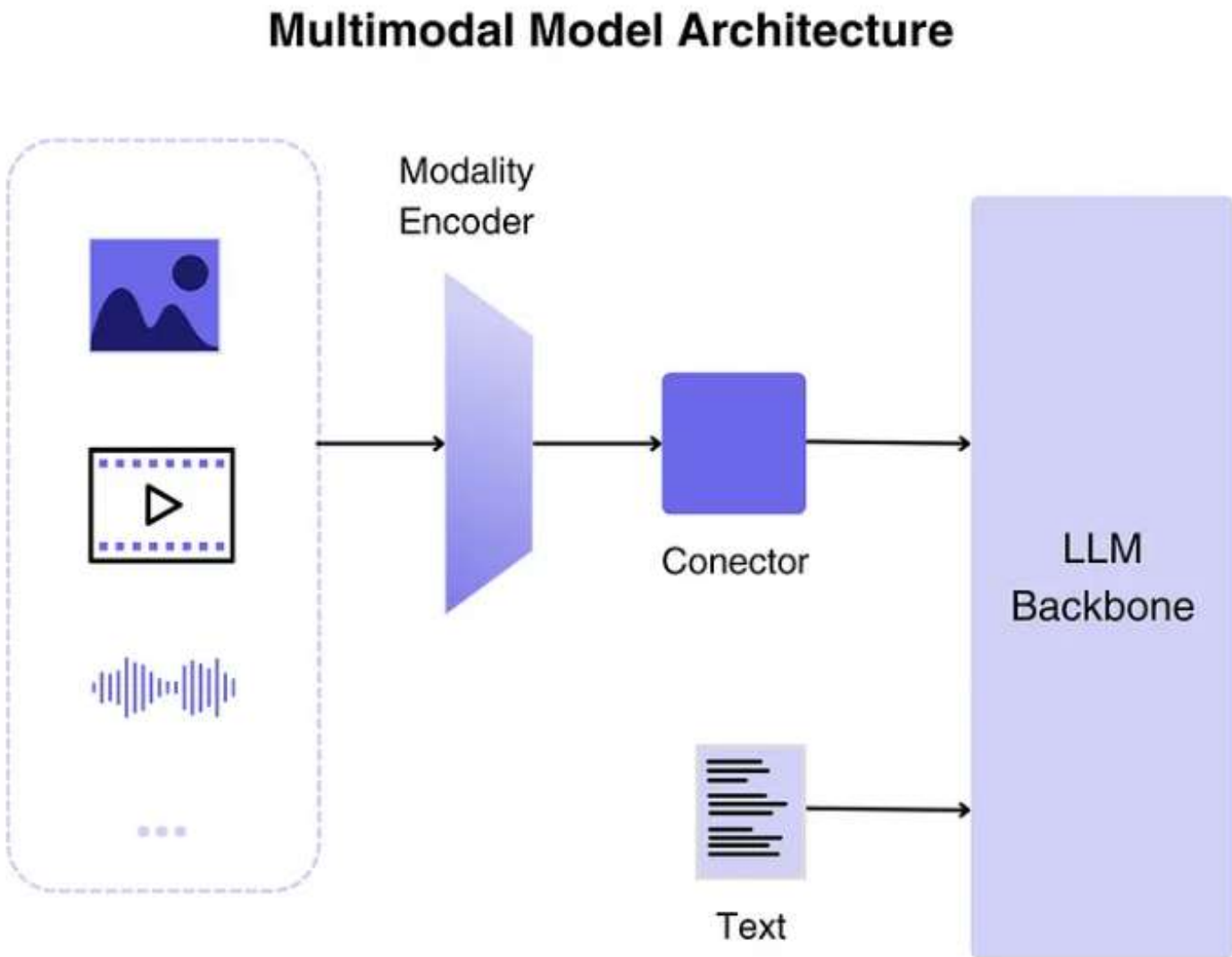
(d) Tree of Thoughts (ToT)

Thought decomposition: Decompose intermediate thinking into coherent based on task characteristics, providing a clear carrier for reasoning.

Thought generator: Generate k diverse candidate thoughts via independent sampling or sequential proposal based on the current reasoning state.

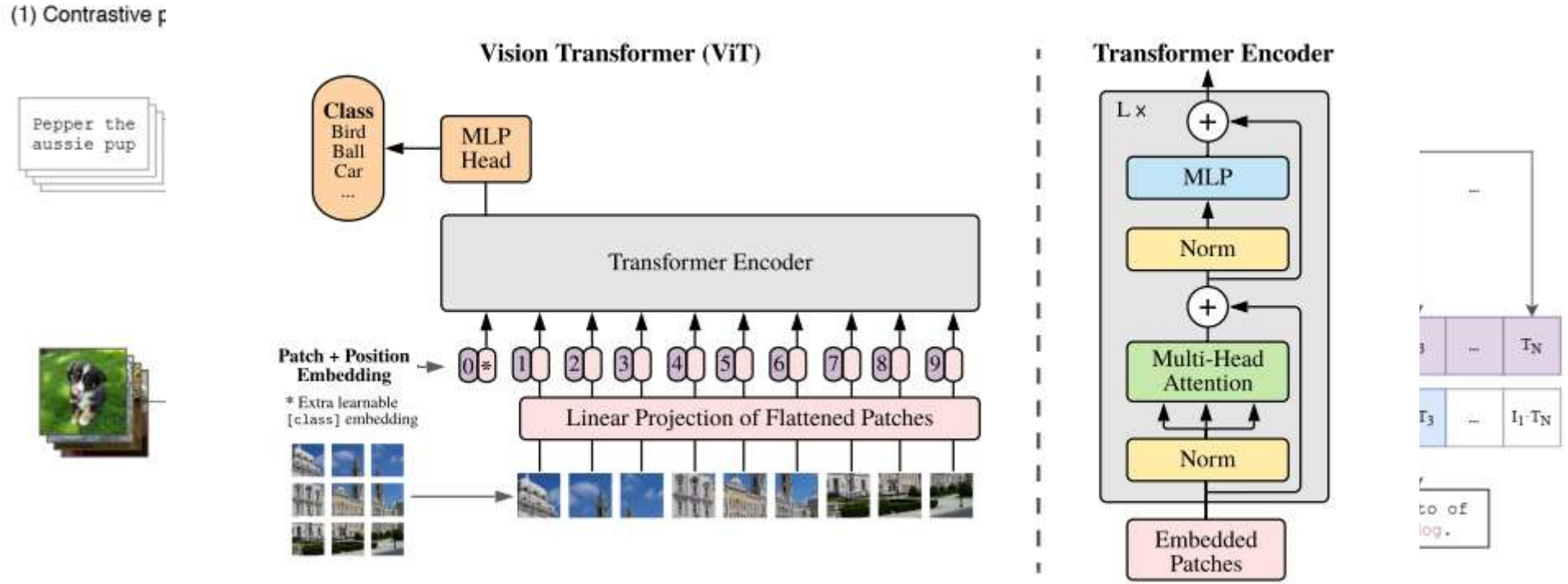
State evaluator: Leverage LLM‘s autonomous reasoning to assess candidate states.

MLLM



1、Modality encoder
2、Connector
3、LLM

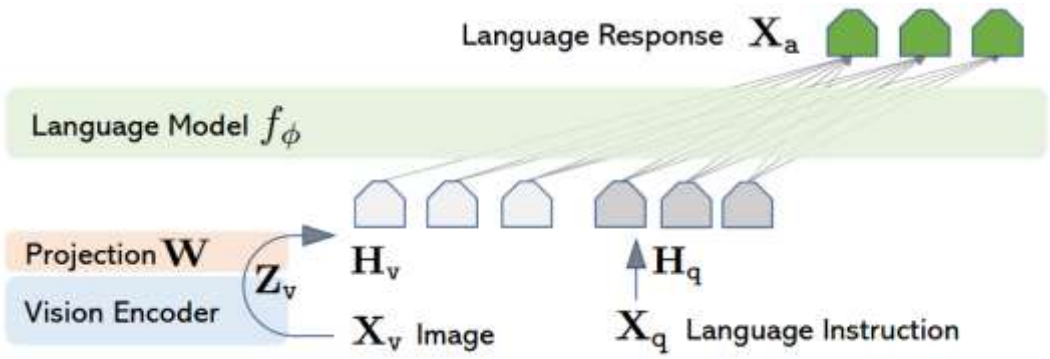
Modality encoder



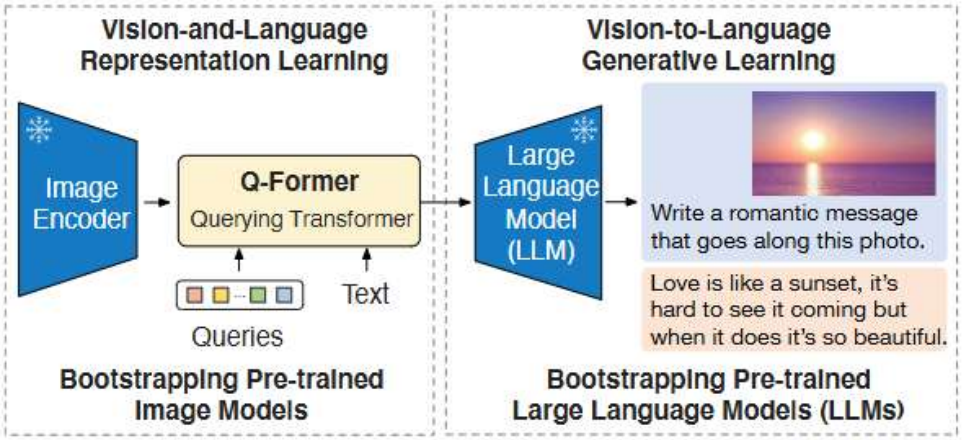
Text-encoder：transformer language model(decoder-only)
Image-encoder：VIT/ResNet

Connector

LLaVA connector: Linear layer

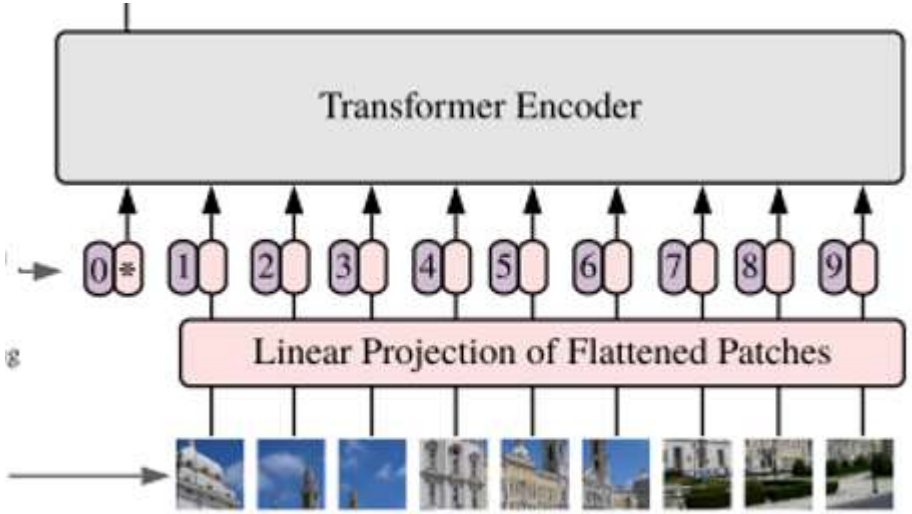


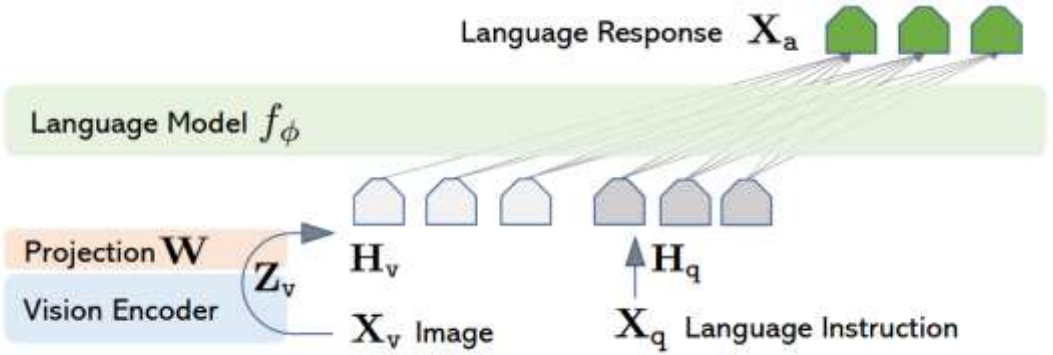
Blip-2 connector: Q-former



Connector

LLaVA connector: Linear layer

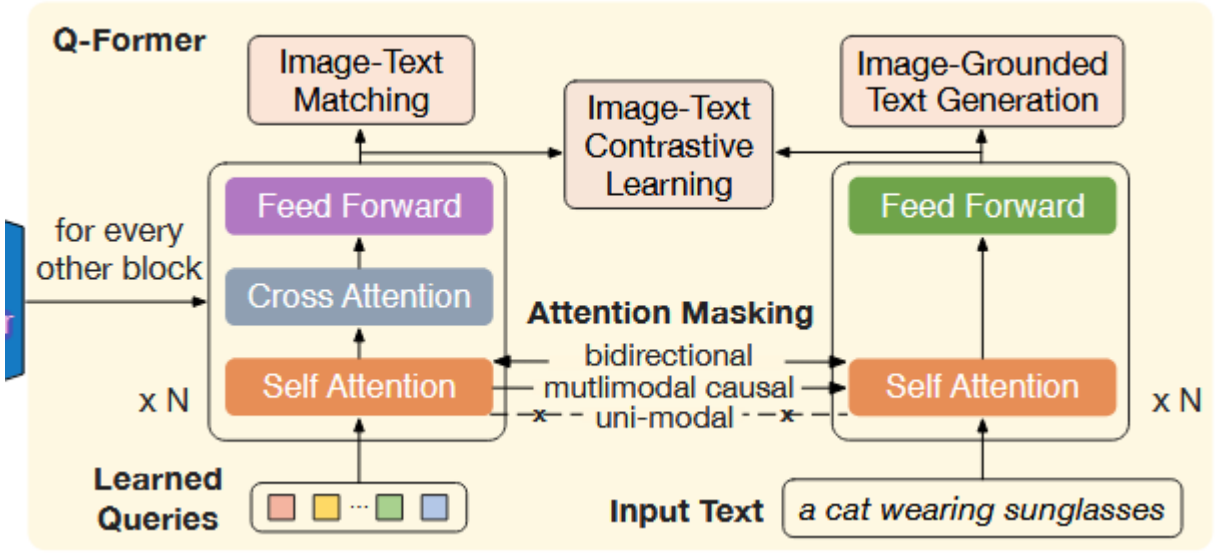




Connector (Projector): Aligns ViT Feature Dimension with LLaMA Word Embedding Dimension

Connector

Blip-2 connector: Q-former



1、self-attention: Bert base(pretrained)
2、cross-attention: randomly initialized

ITC: Image-text Contrastive
ITG: Image-text Generation
ITM: Image-Text Matching