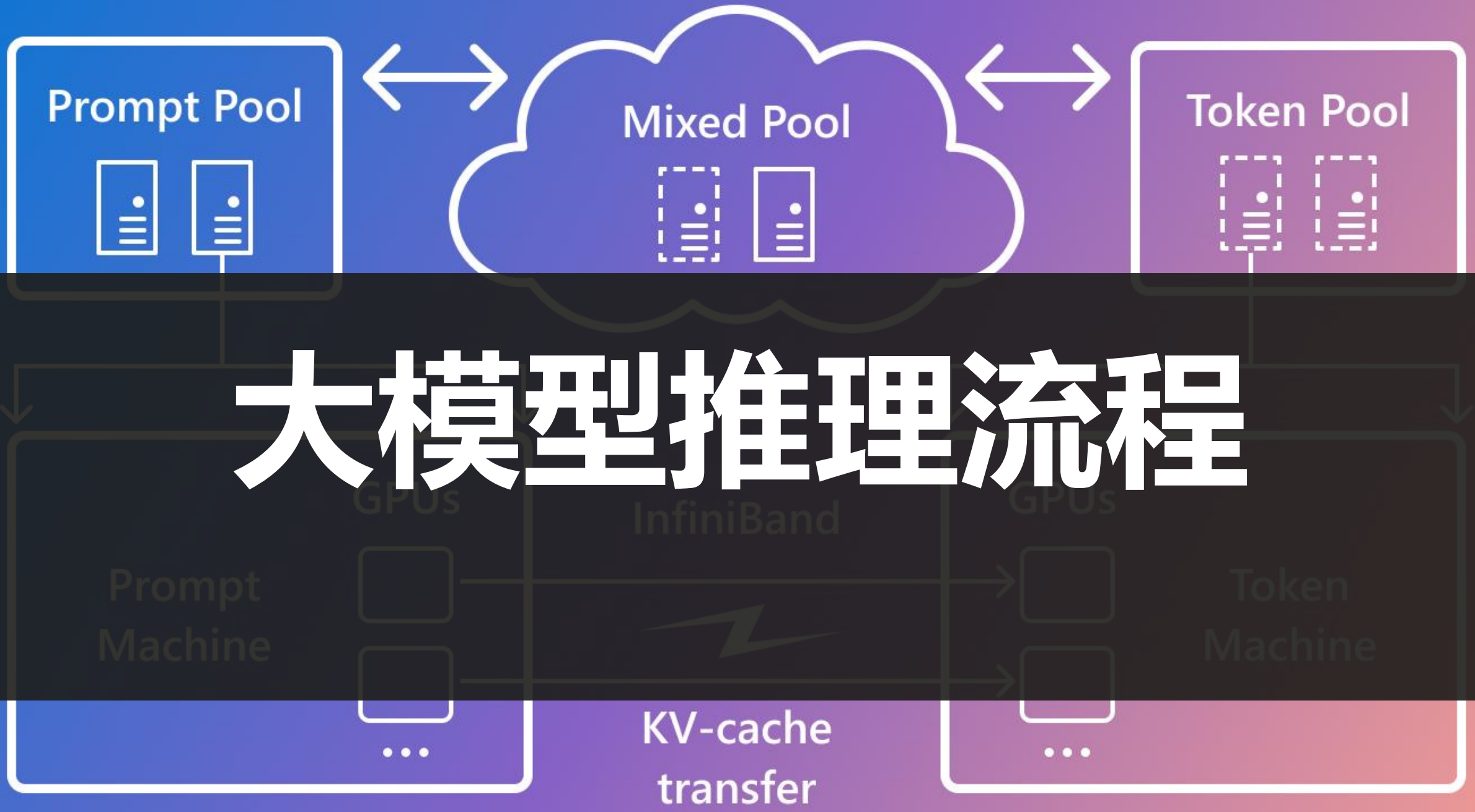
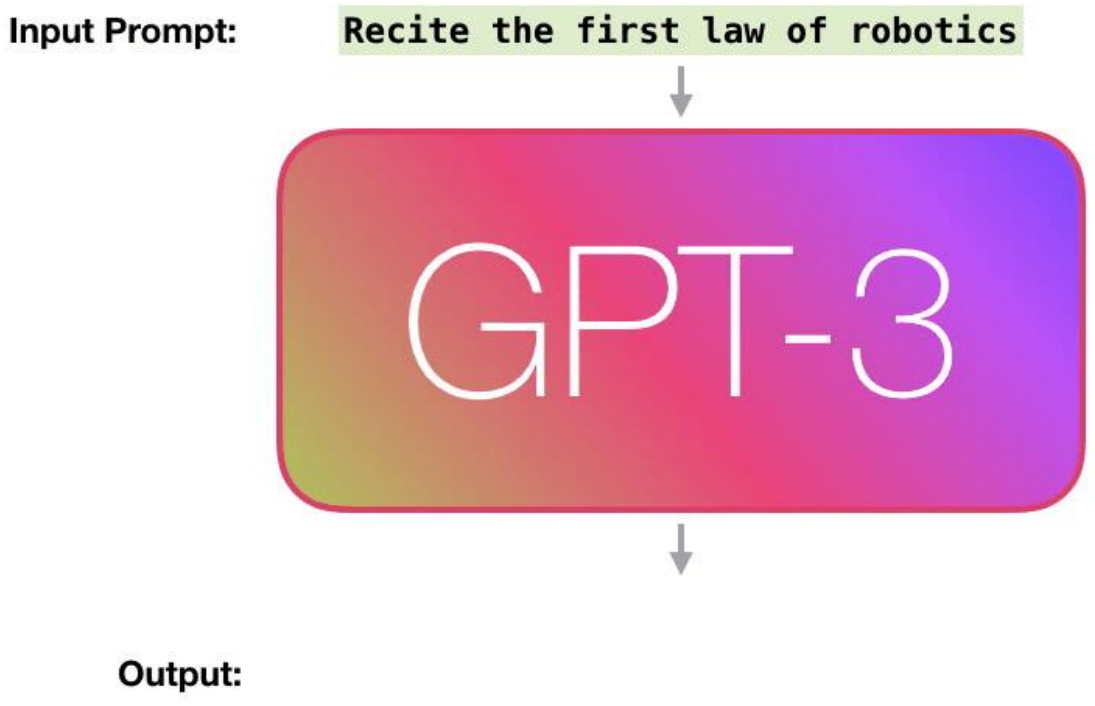
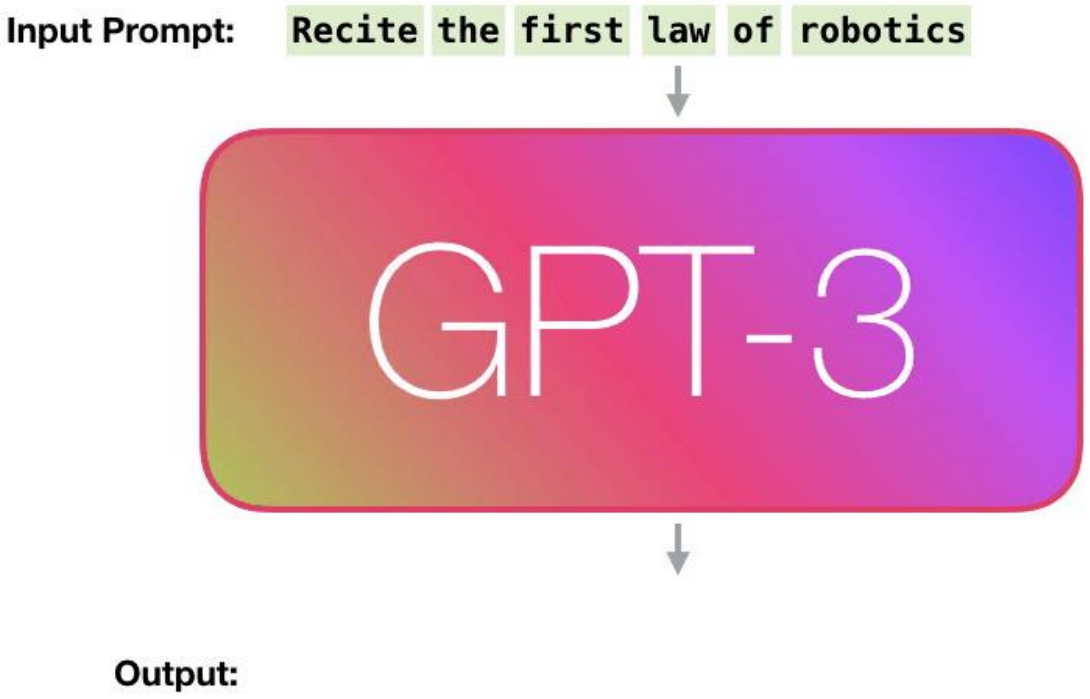

Prompt Pool
Mixed Pool
Token Pool
GPUs
InfiniBand
GPUs
Prompt
Machine
Token
Machine
KV-cache
transfer

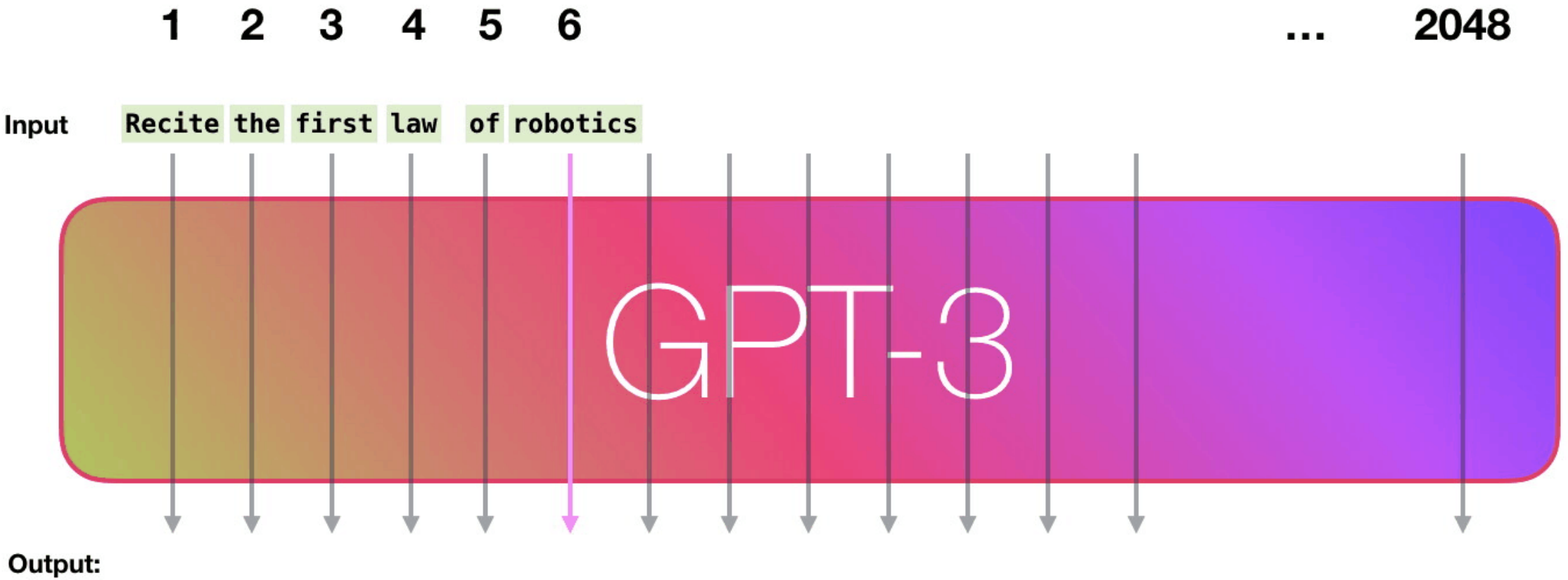
大模型推理流程

LLM 推理过程


Input Prompt:
Recite the first law of robotics
GPT-3
Output:

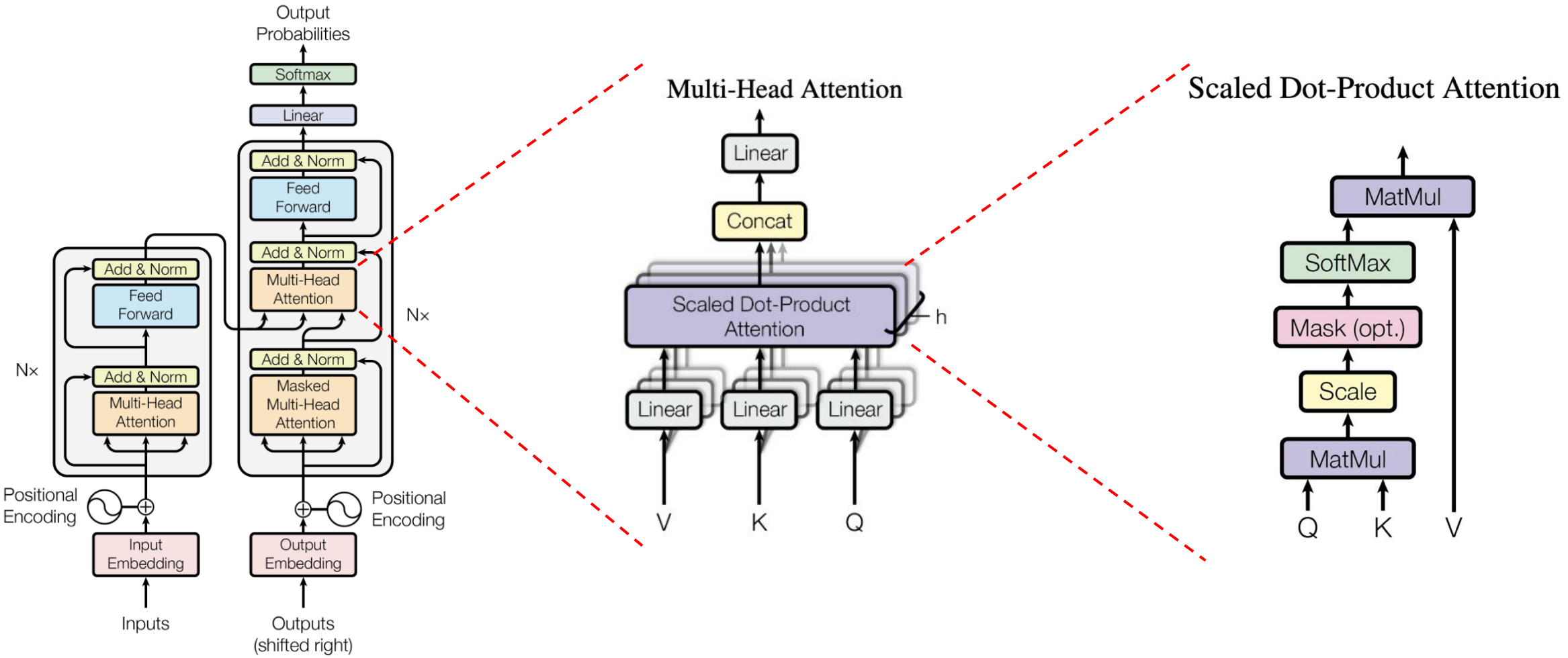

Input Prompt:
Recite the first law of robotics
GPT-3
Output:

LLM 推理过程


1 2 3 4 5 6 ... 2048
Input
Recite the first law of robotics
GPT-3
Output:

Attention

- 在 LLM 推理中最关键的就是下图中的 Attention：



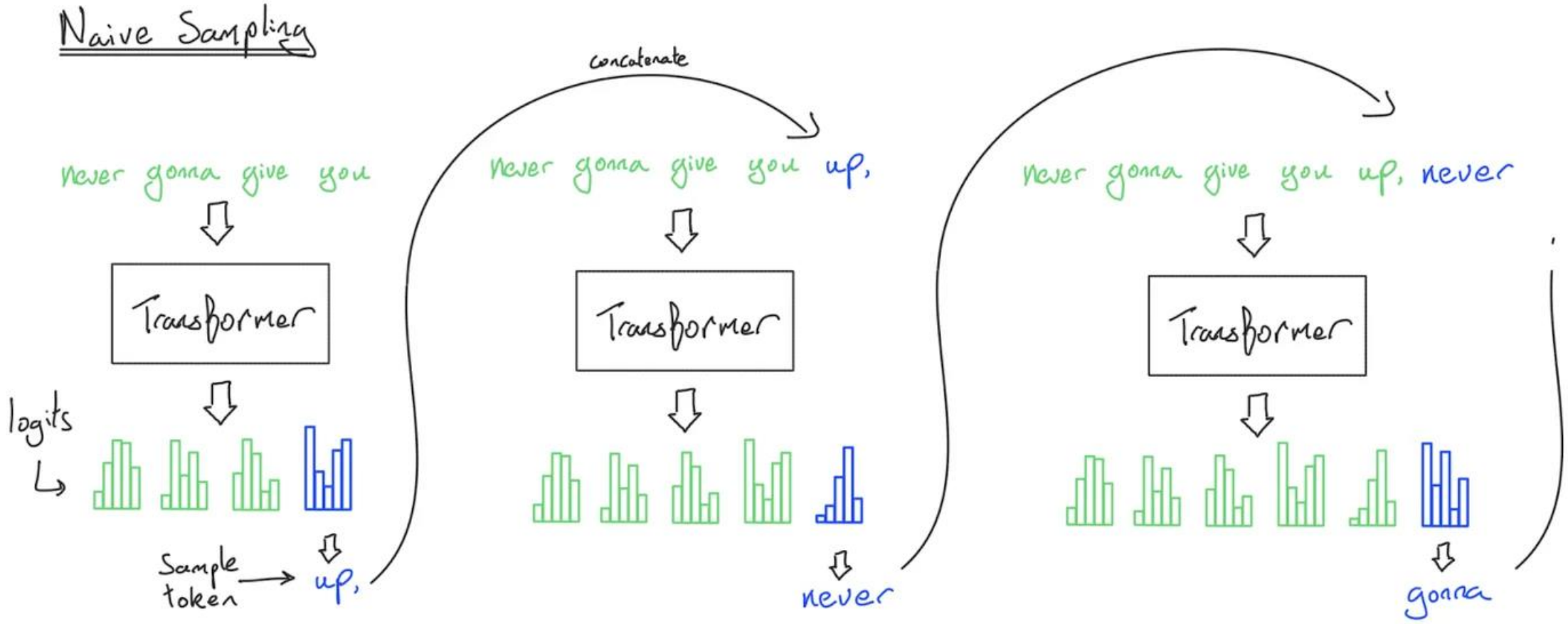
Attention

- 展开 Attention 的计算：

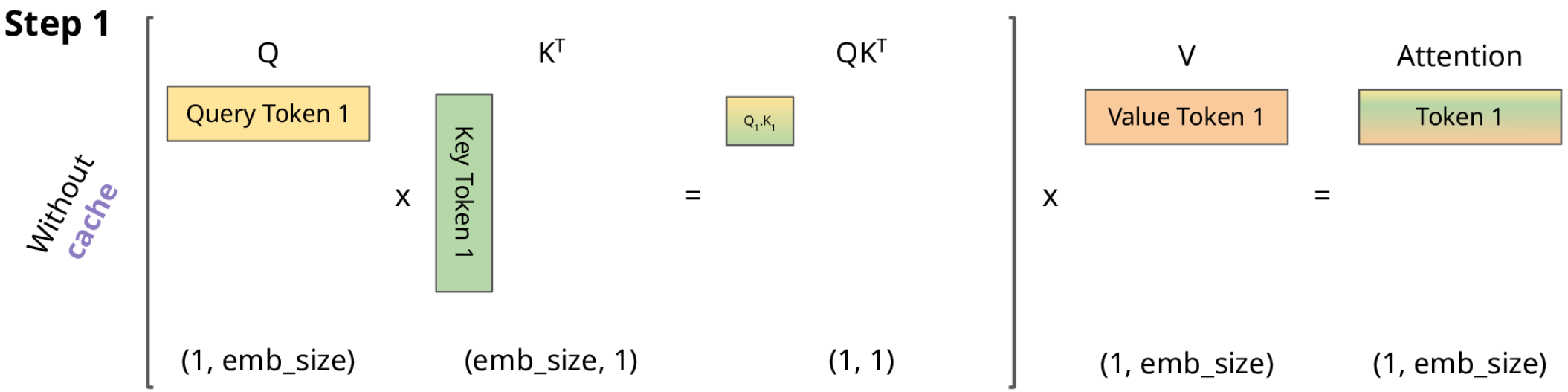
$$Attention = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- 每次生成一个 token 之后，会将该 token 拼接到上一次输入的 tokens 之后，进行下一次的 Attention 计算以用于预测下一个输出的 token，这个过程称为自回归推理过程

Naïve Attention



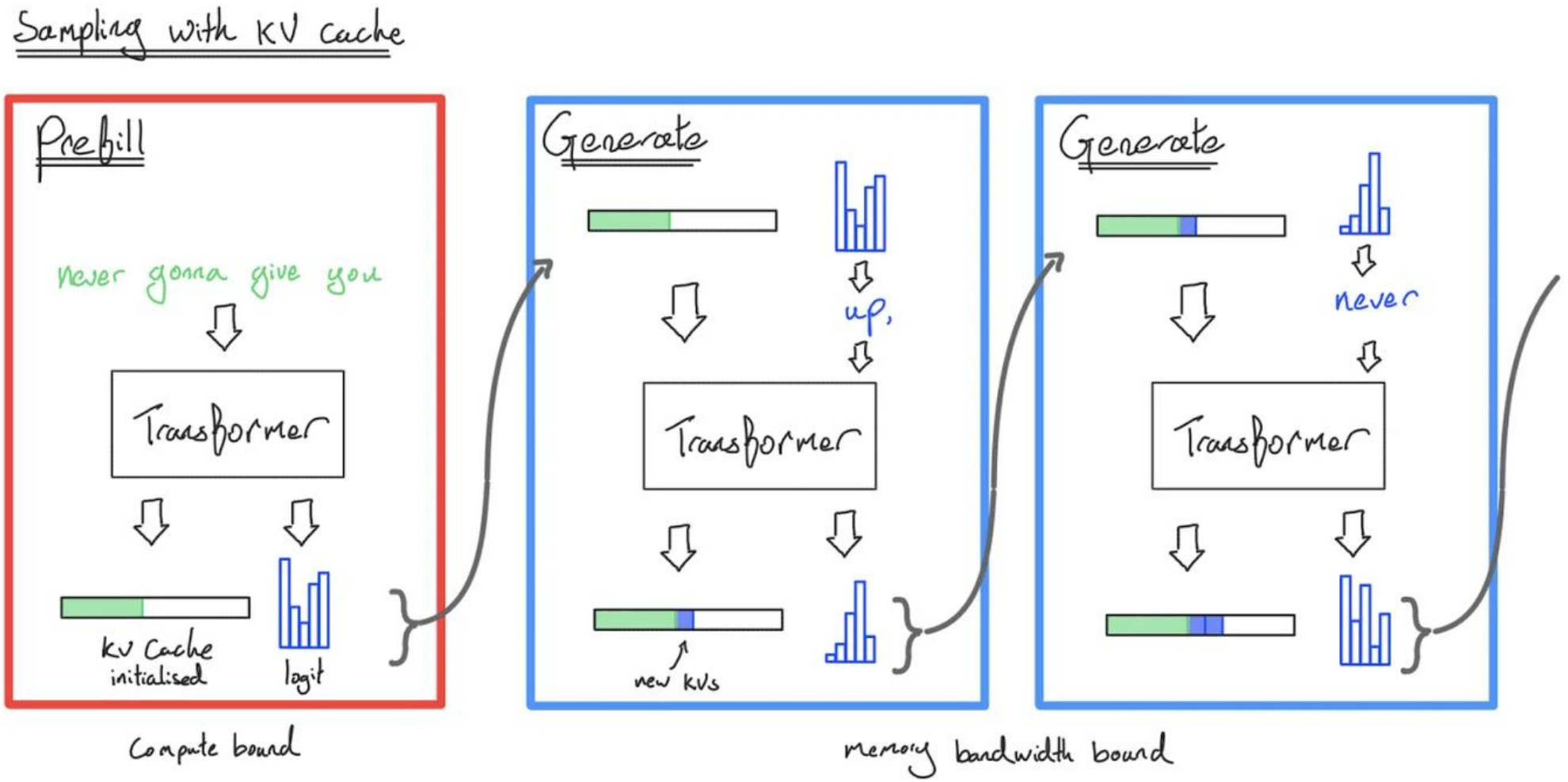
Attention


Step 1
Without cache
Q
Query Token 1
(1, emb_size)
x
K^T
Key Token 1
(emb_size, 1)
=
QK^T
Q1.K1
(1, 1)
x
V
Value Token 1
(1, emb_size)
=
Attention
Token 1
(1, emb_size)

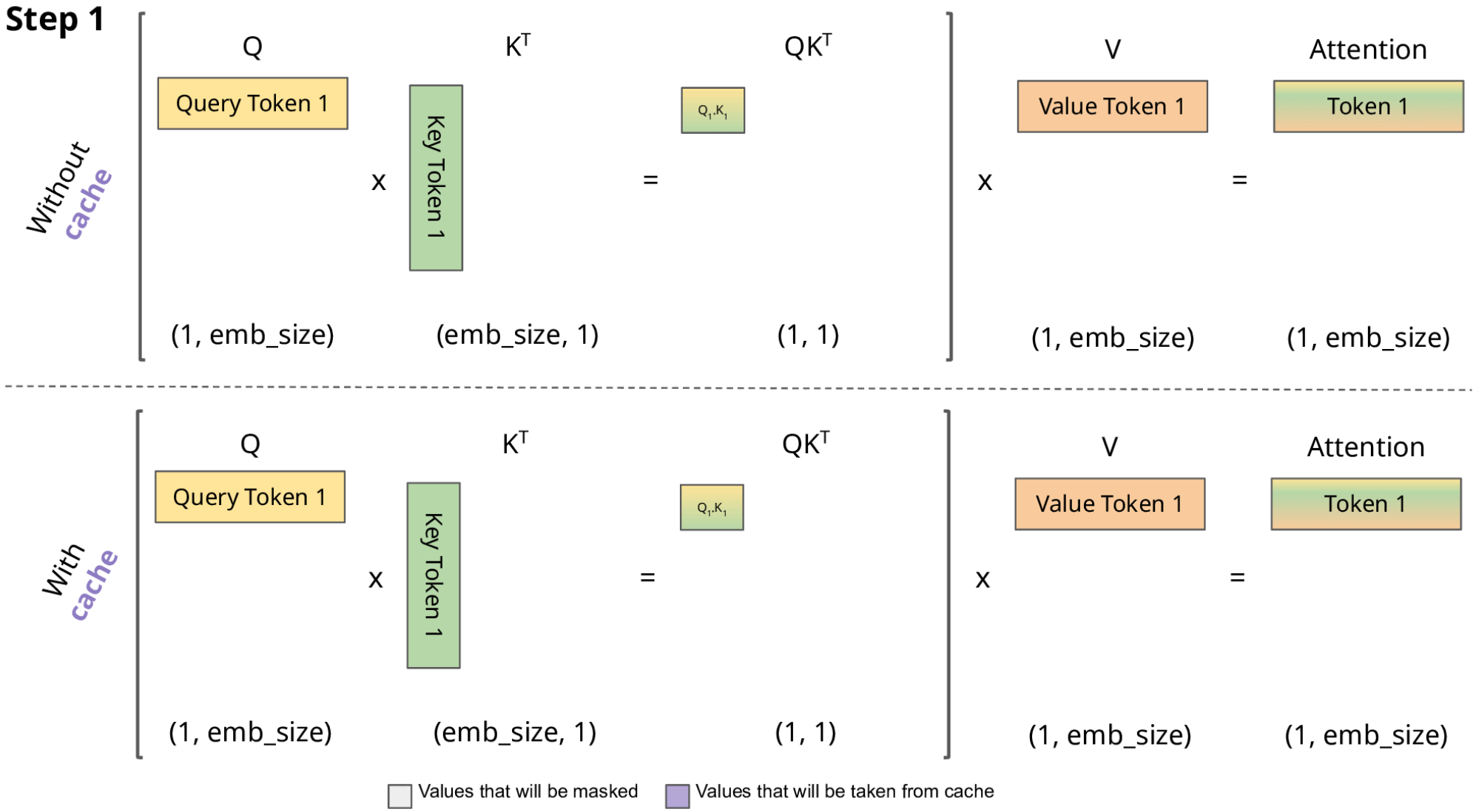
KV Cache

- 把每个 token 在过 Transformer 时乘以 W_k, W_v 这俩参数矩阵的结果缓存下来
- 推理解码生成时采用自回归 auto-regressive 方式，即每次生成一个 token，都要依赖之前 token 的结果
- 如果每生成一个 token 时候乘以 W_k, W_v 这俩参数矩阵要对所有 token 都算一遍，代价非常大，所以缓存起来就叫 KV Cache

KV Cache


Sampling with KV cache
Prefill
never gonna give you
Transformer
KV cache initialised
logit
Compute bound
Generate
up,
Transformer
new KVs
Generate
never
Transformer
memory bandwidth bound

KV Cache

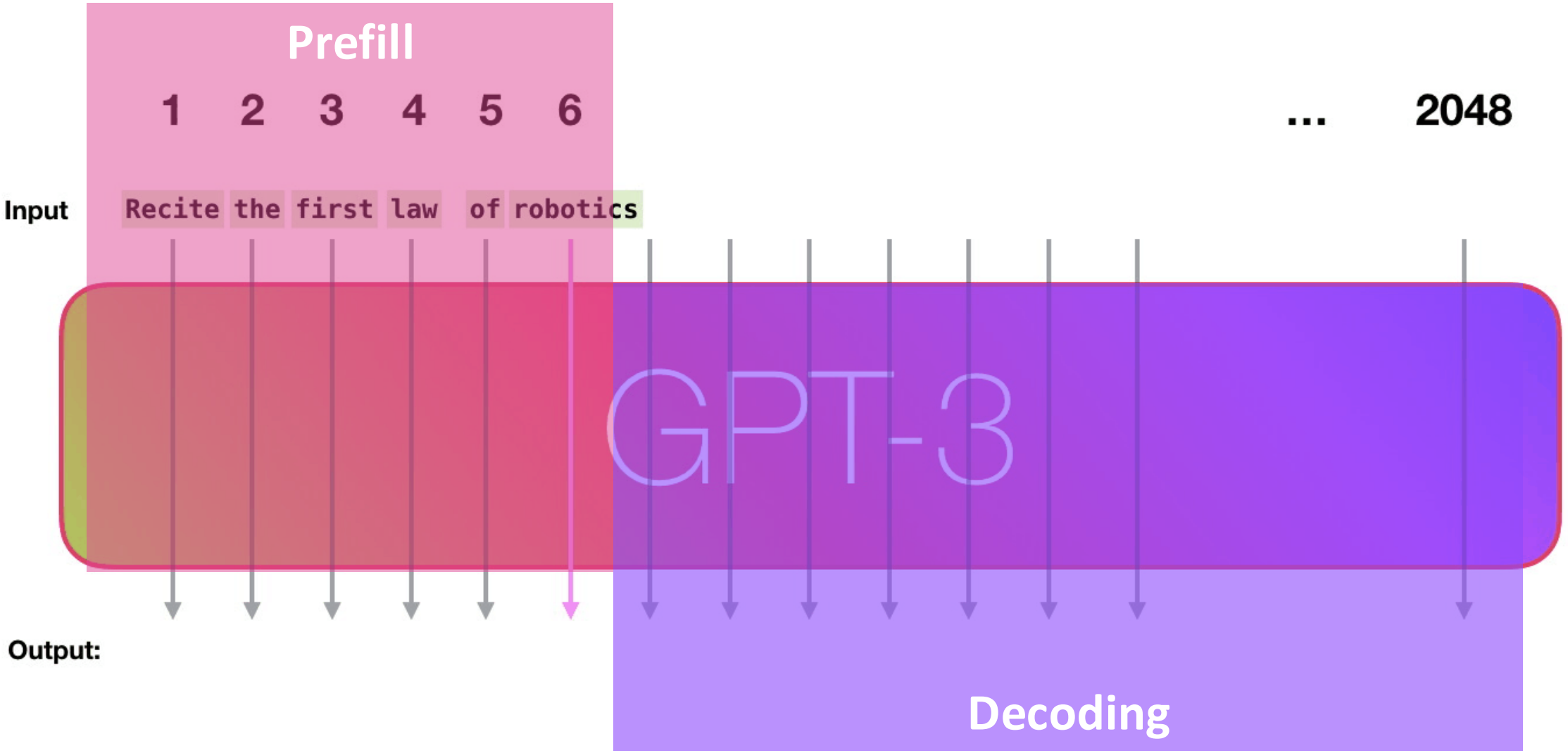

Step 1
Without cache
Q
K^T
QK^T
Query Token 1
Key Token 1
Q1.K1
x
=
V
Attention
Value Token 1
Token 1
x
=
(1, emb_size)
(emb_size, 1)
(1, 1)
(1, emb_size)
(1, emb_size)
With cache
Q
K^T
QK^T
Query Token 1
Key Token 1
Q1.K1
x
=
V
Attention
Value Token 1
Token 1
x
=
(1, emb_size)
(emb_size, 1)
(1, 1)
(1, emb_size)
(1, emb_size)
Values that will be masked
Values that will be taken from cache

LLM 推理过程

- **Prefill：**
  - 根据输入的 prompt 生成第一个输出 Token，通过一次 Forward 就可以完成
  - 在 Forward 中，输入 Tokens 间可以并行计算，因此执行效率很高，为计算瓶颈

- **Decoding：**
  - 从生成第一个 Token 后，采用自回归一次生成一个 Token，直到生成 Stop Token 或达到预设的输出长度上限结束
  - 设输出共 N x Token，Decoding 阶段需要执行 N-1 次 Forward，只能串行执行，效率很低
  - 在生成过程中，需要计算 Token 越来越多，计算量也会适当增大

LLM 推理过程


Prefill
1 2 3 4 5 6
... 2048
Input
Recite the first law of robotics
GPT-3
Output:
Decoding

Q&A