

B **AOLONG**

大模型

· 主讲人：王靖辉

Stable Diffusion

拳击手



悬崖瀑布
水彩画



麦田秋天
油画



树林溪流
水墨画



漫画脸
+微笑



改发色+
长发





目 录

0 1 | 大模型概述

0 2 | 大模型组成结构

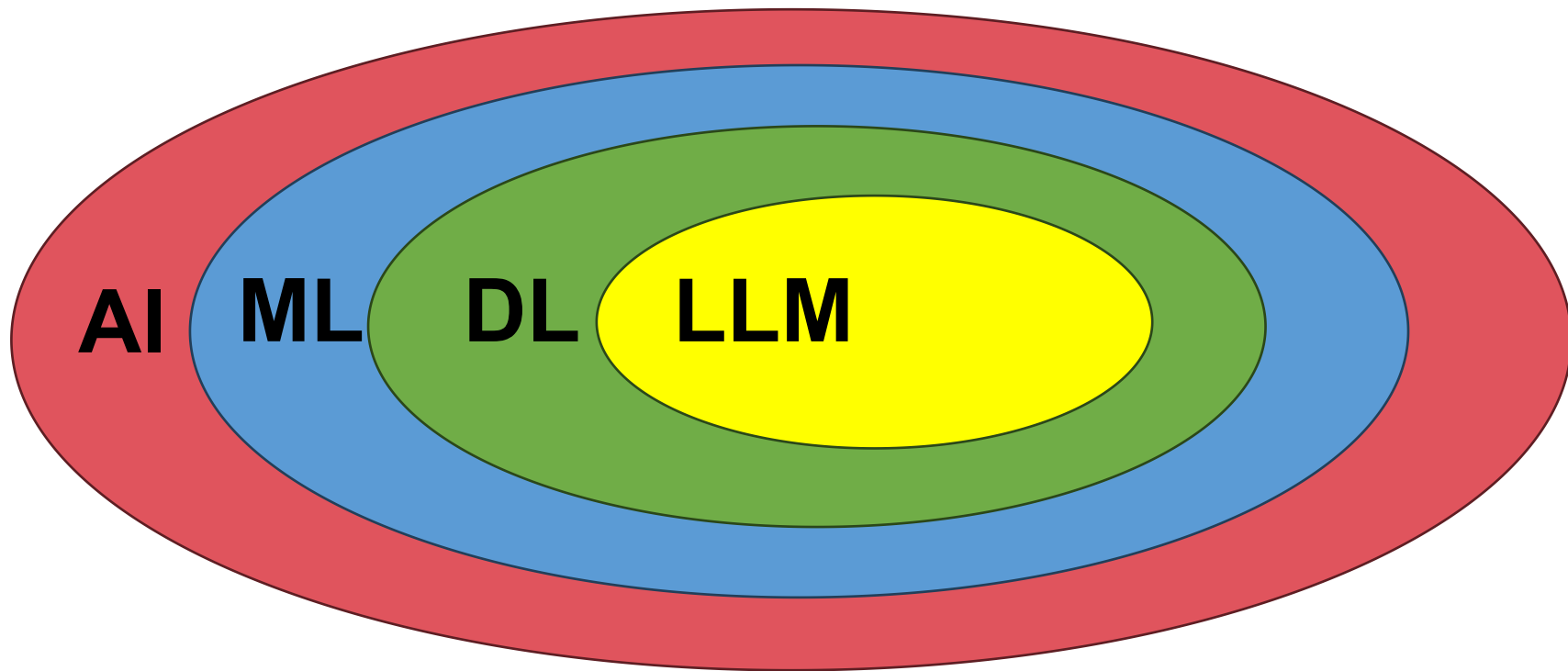
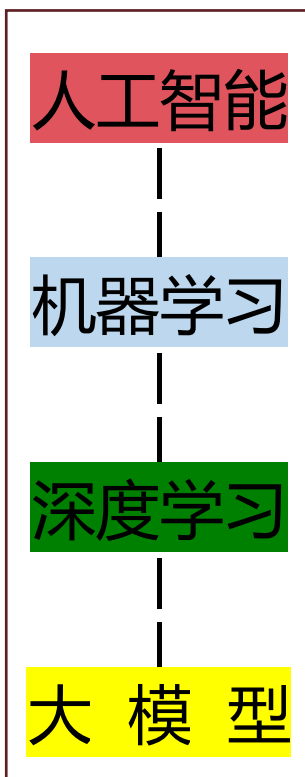
0 3 | 大模型应用场景

01 大模型概述

大模型概述

定义

大模型是指容量较大，用于深度学习任务的模型，通常具有海量的参数和复杂的架构。具有更好的通用性，可以通过预训练或其他方式在大型数据集上进行学习，再通过微调高效地处理计算机视觉、自然语言处理等复杂任务。



大模型概述-起源

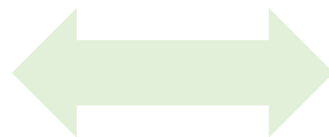
深度学习

过拟合问题

泛化能力有限

数据依赖性强

需要大量人工干预



大模型

泛化能力较强

仅需要微调

需要较少的人工干预

图像识别
手势估计
车道线检测
路径规划
目标检测
语言识别
文本翻译
等等

大模型概述-特点



大规模

- 1、拥有数以亿计的参数和权重
- 2、能够在海量数据上进行学习和建模



自监督学习

- 1、无标签数据开展预训练
- 2、使用迁移学习进行微调



复合性

- 1、多个不同的组件组成
- 2、根据不同任务和领域的需求进行调整和改进



通用性

- 1、适应不同的语言、领域和任务
- 2、应用于不同类型的问题

02 大模型组成结构

CNN

优点:

局部感知和权值共享
层次化特征提取

缺点:

缺乏全局感知
对输入尺寸敏感

应用场景:

图像分类
目标检测
语义分割

RNN

优点:

处理序列数据
短距离依赖

缺点:

计算效率低

应用场景:

自然语言处理
语音识别

LSTM

优点:

长距离依赖

缺点:

计算复杂性
参数数量较多

应用场景:

自然语言处理
时间序列预测

Transformer

优点:

长距离依赖
强大的语义理解

缺点:

计算资源消耗较大
训练难度较高

应用场景:

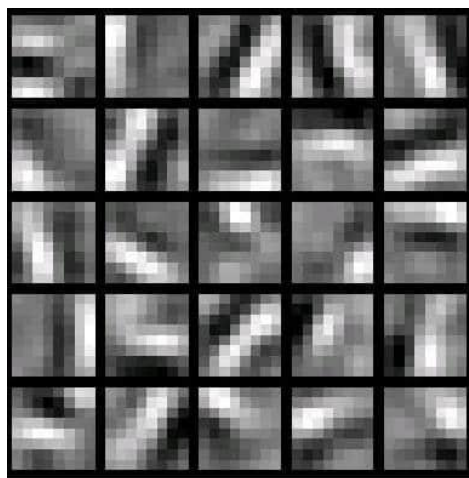
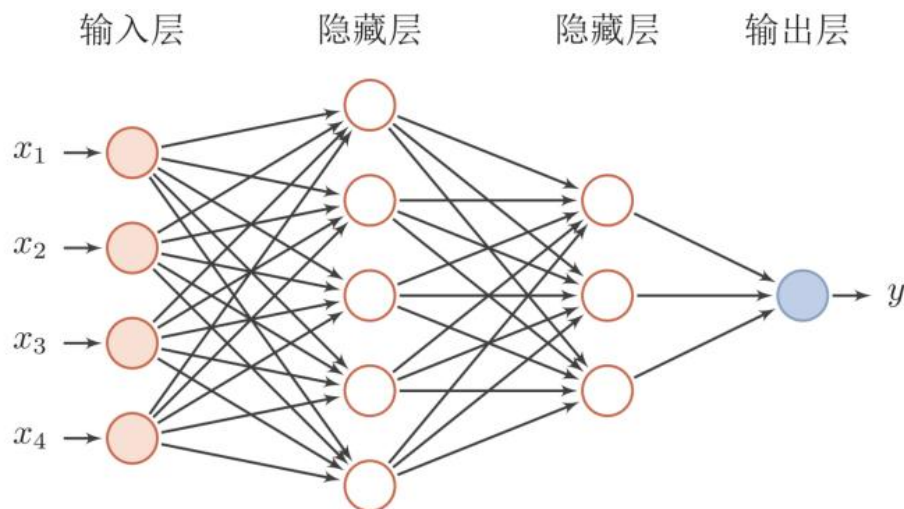
机器翻译
文本生成
文本分类和情感分析
序列到序列任务

大模型组成结构-CNN

卷积神经网络CNN

基本组成:

- 1) 卷积层 (带有激活函数) : 提取图像局部特征
- 2) 池化层: 数据降维
- 3) 全连接层: 特征表示, 信息整合

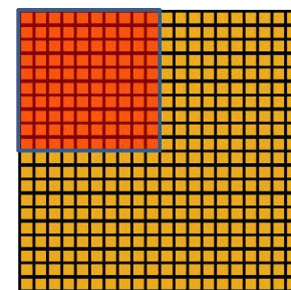


1 _{x1}	1 _{x0}	1 _{x1}	0	0
0 _{x0}	1 _{x1}	1 _{x0}	1	0
0 _{x1}	0 _{x0}	1 _{x1}	1	1
0	0	1	1	0
0	1	1	0	0

Image

4		

Convolved Feature



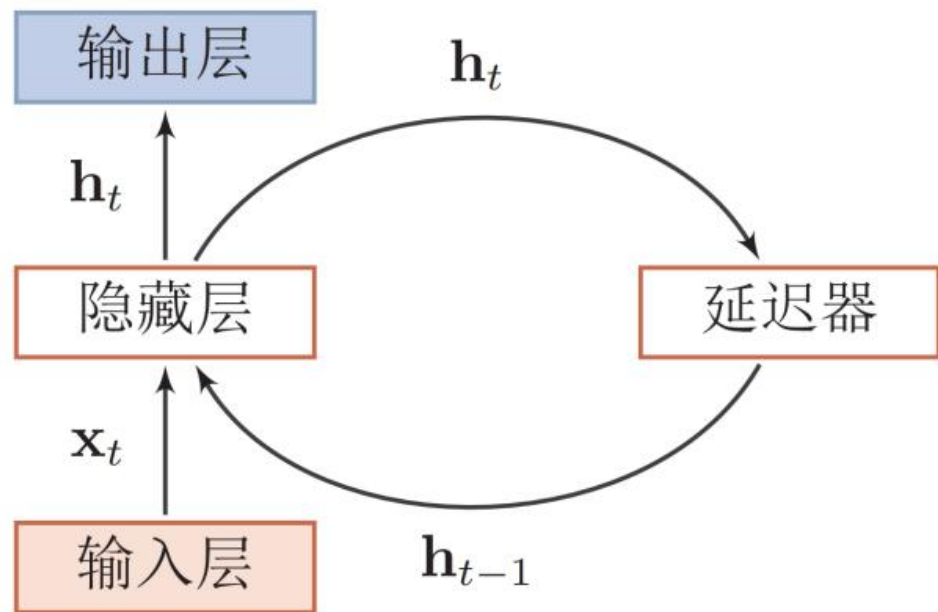
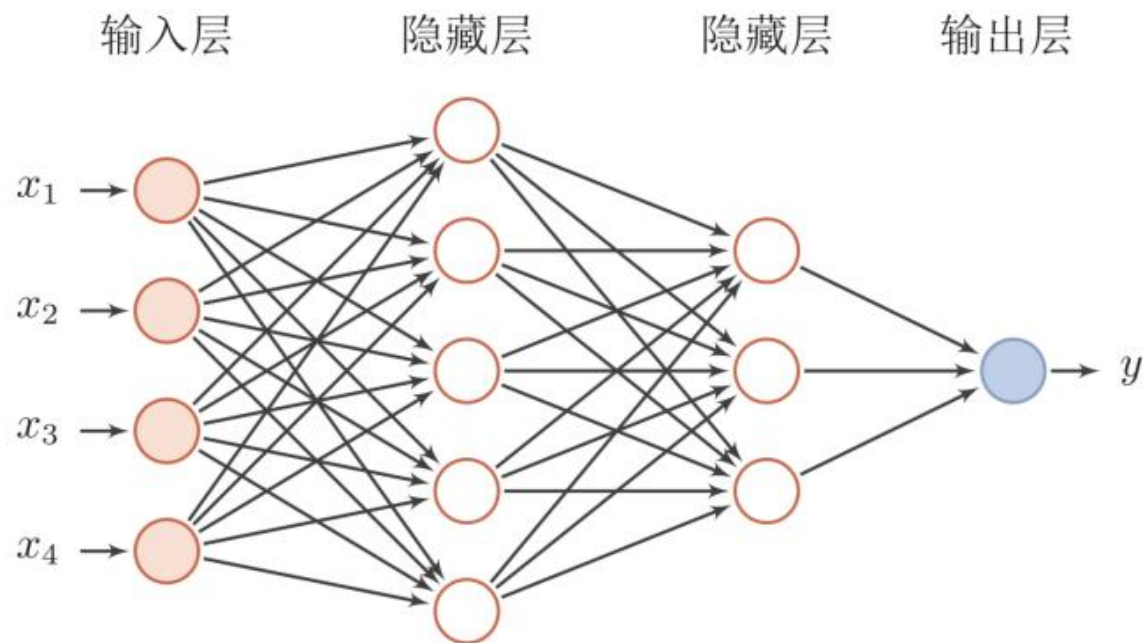
Convolved feature

1	

Pooled feature

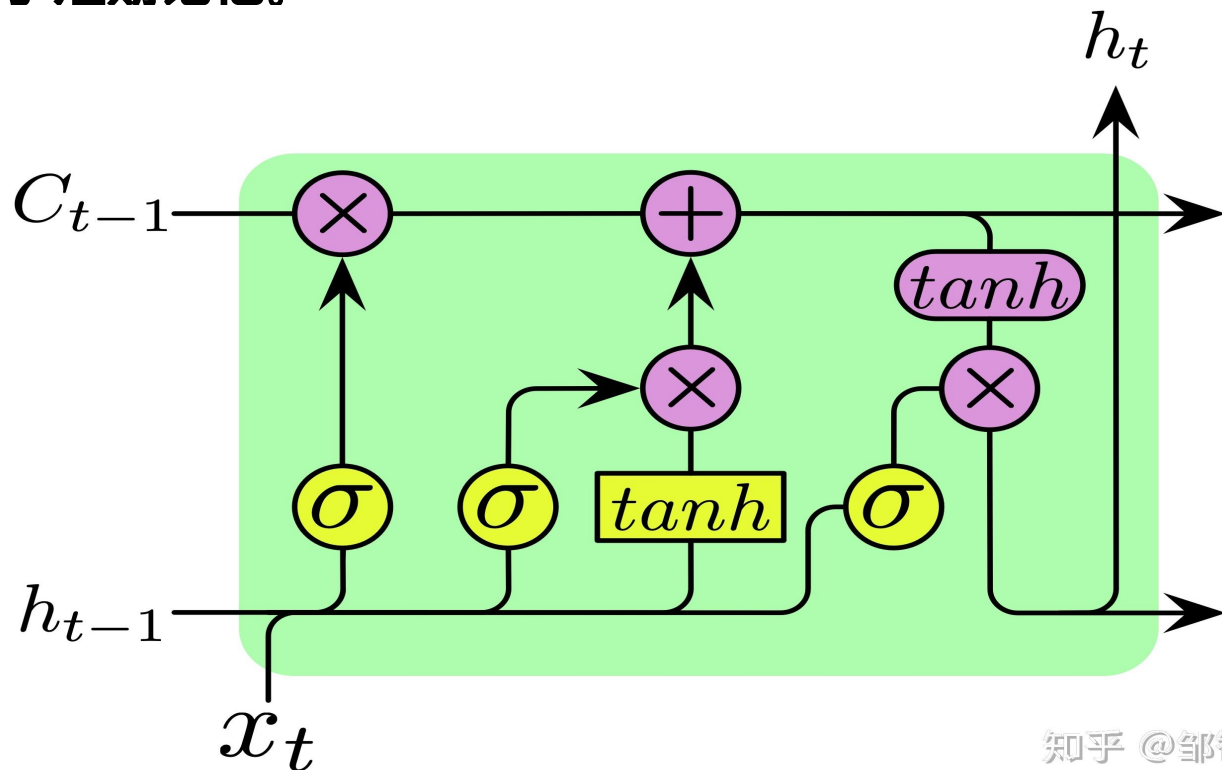
循环神经网络RNN

- 具有内部环状连接的人工神经网络，即一个序列当前的输出与前面的输出也有关，隐藏层之间的节点不再无连接而是有连接的。



长短期记忆网络LSTM

循环神经网络中的隐状态 h 存储了历史信息，可以看作是一种记忆，但这是一种短期记忆因为隐状态每个时刻都会被重写，而长期记忆可以看作是网络参数，隐含了从训练数据中学到的经验，其更新周期要远远慢于短期记忆。



大模型组成结构-Transformer

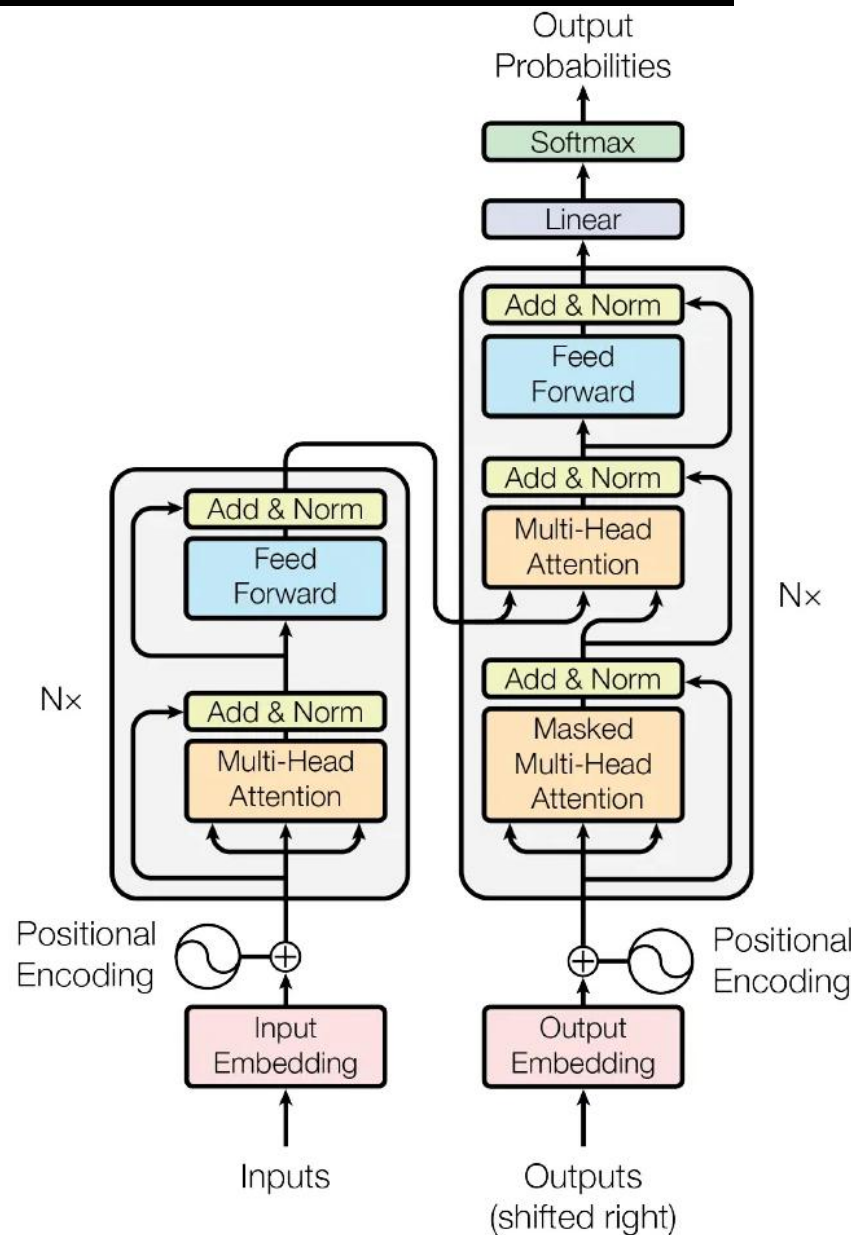
Transformer

- 模型架构

抛弃了传统的CNN和RNN，整个网络结构完全是由Attention机制组成，通过搭建编码器和解码器，在NLP任务中取得优异成绩。

- 基本原理

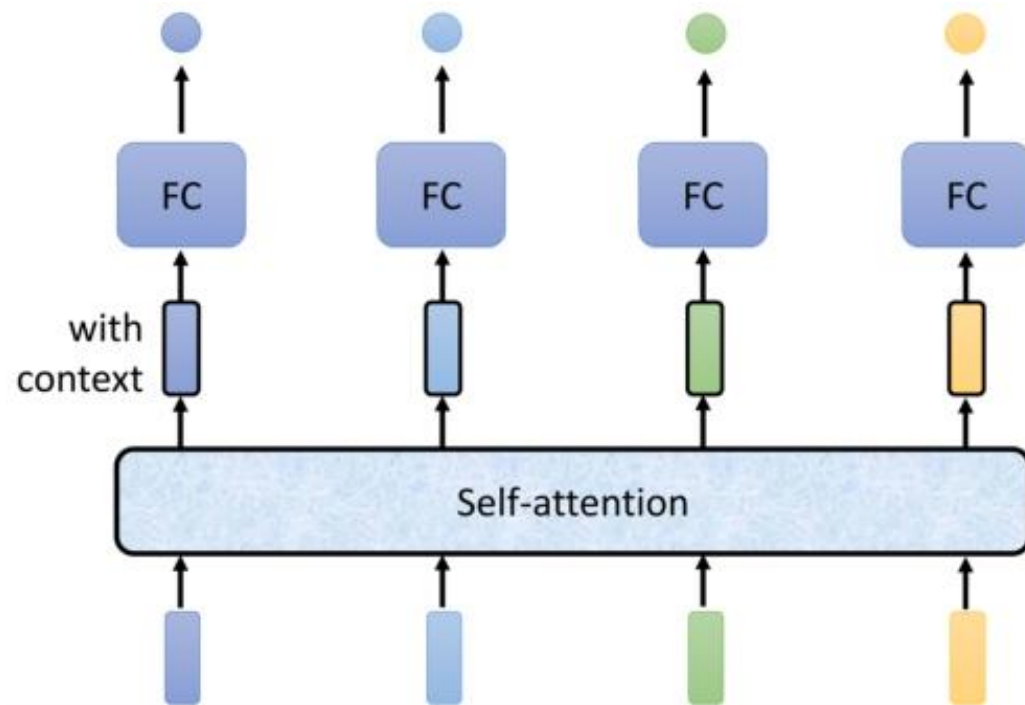
Attention（注意力）机制从关注全部到关注重点，将有限的注意力集中在重点信息上，从而节省资源，快速获得最有效的信息。



Attention

Attention为一种注意力机制，它将一个序列的不同位置联系起来，以计算序列的表示。作用为**全局关联权重**，然后做输入的加权和。

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



大模型组成结构-Transformer

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

Input

Embedding

Queries

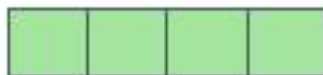
Keys

Values

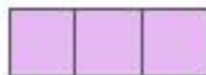
Score

Thinking

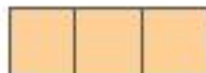
x_1



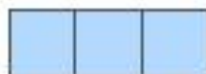
q_1



k_1



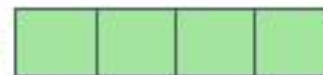
v_1



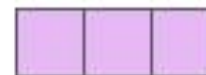
$$q_1 \cdot k_1 = 112$$

Machines

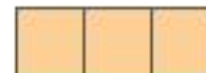
x_2



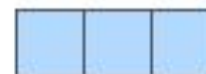
q_2



k_2



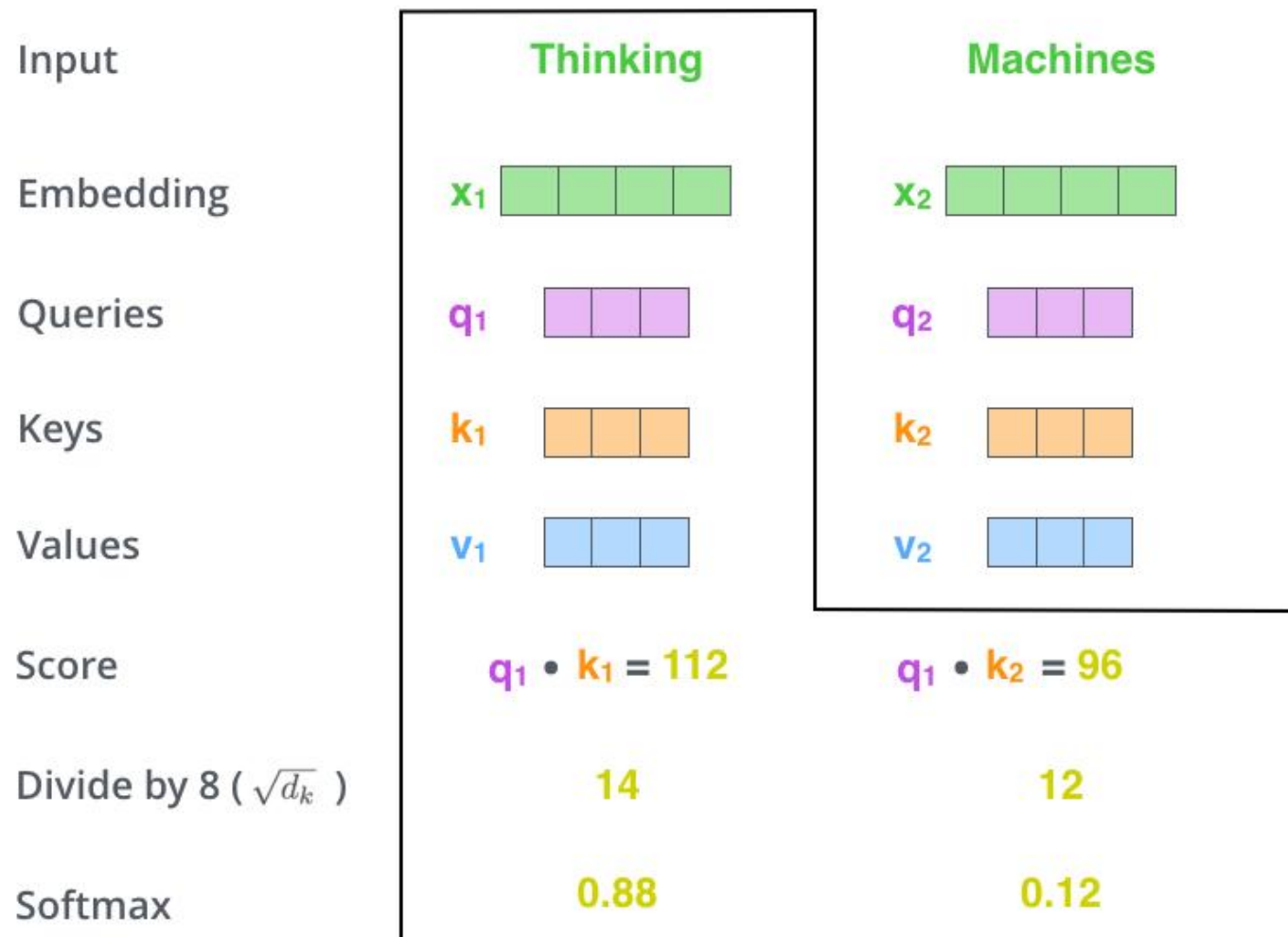
v_2



$$q_1 \cdot k_2 = 96$$

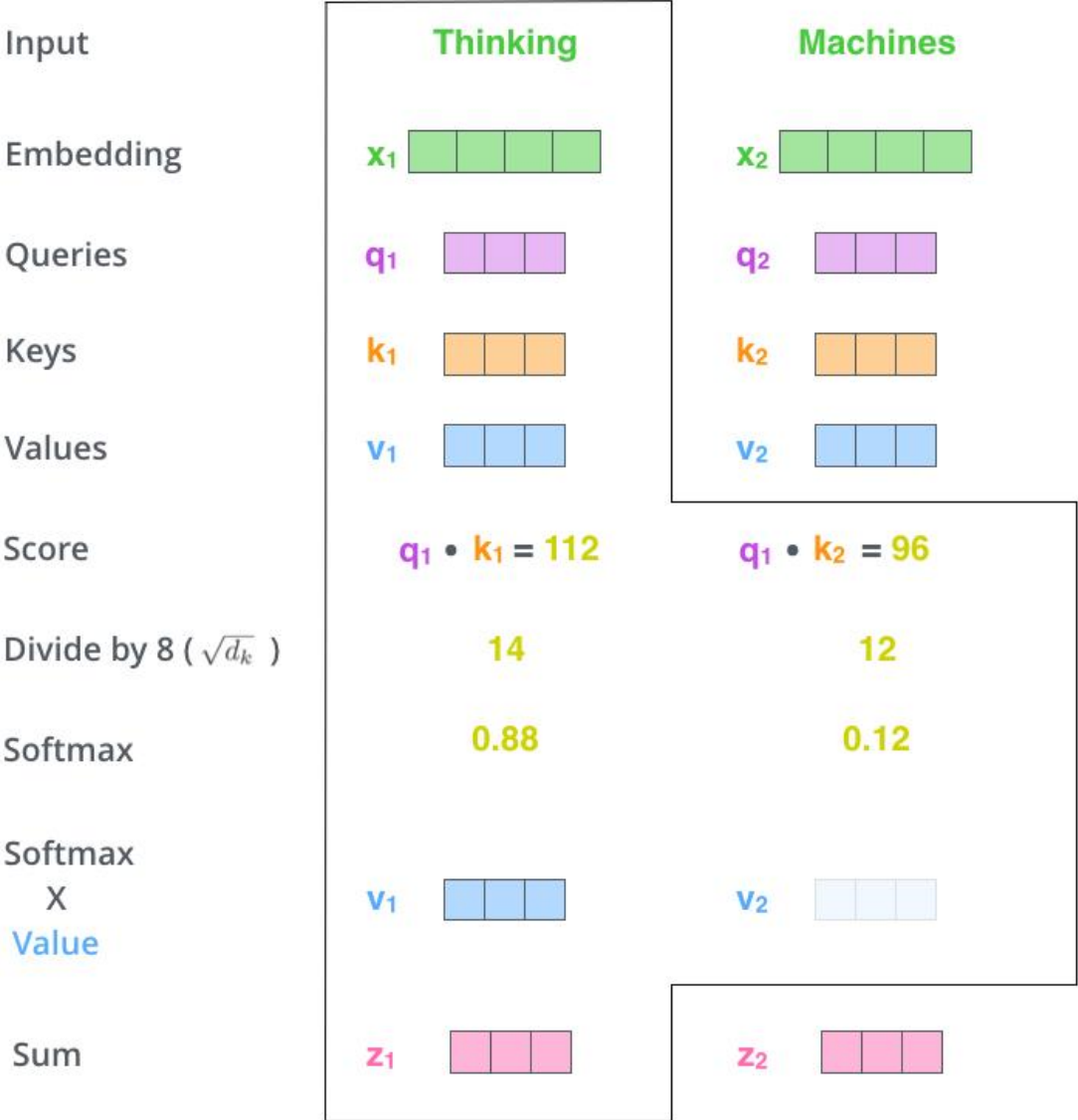
大模型组成结构-Transformer

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

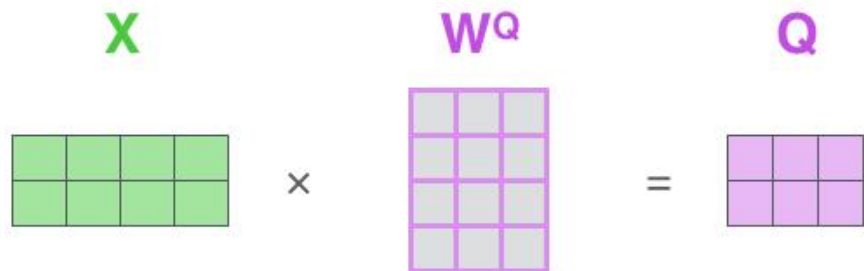


大模型组成结构-Transformer

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

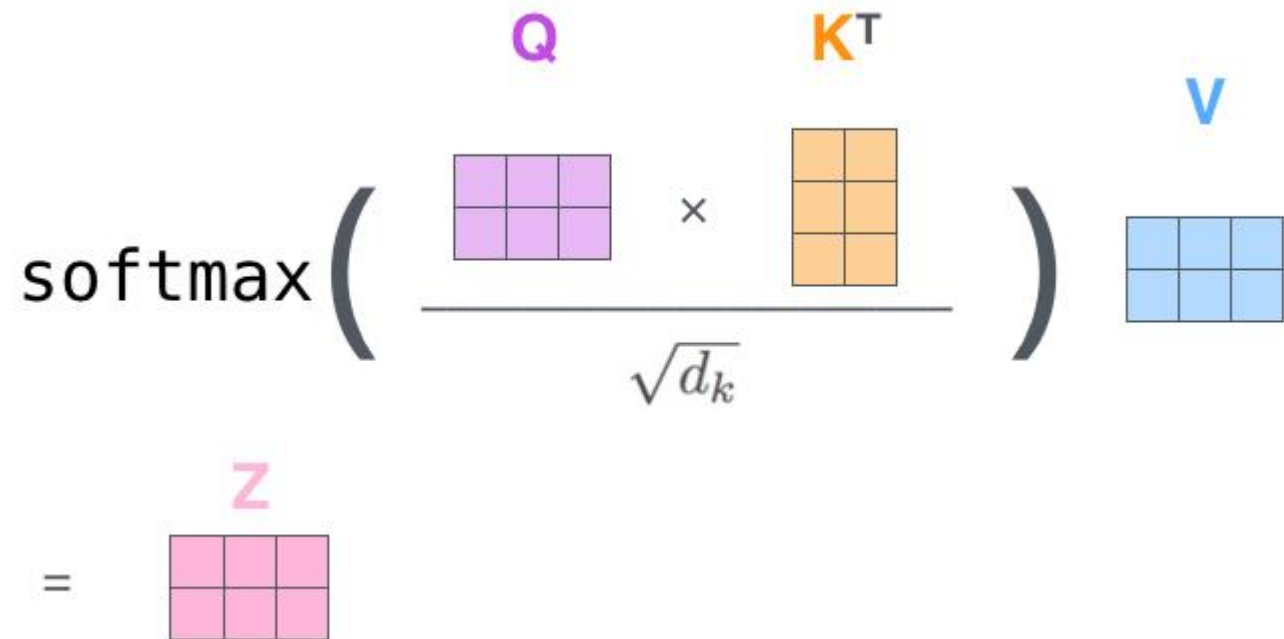


大模型组成结构-Transformer

$$X \times W^Q = Q$$


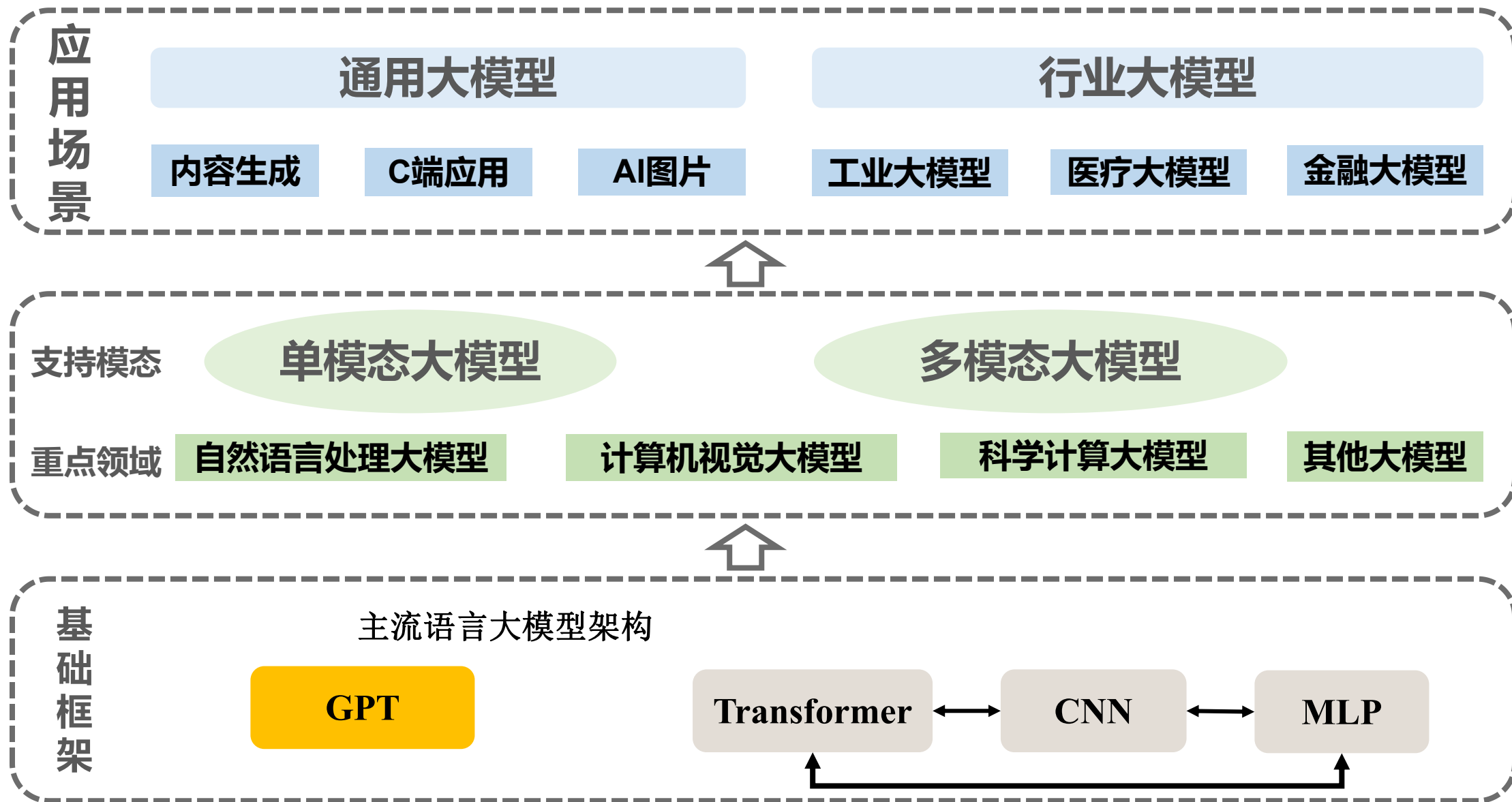
$$X \times W^K = K$$


$$X \times W^V = V$$

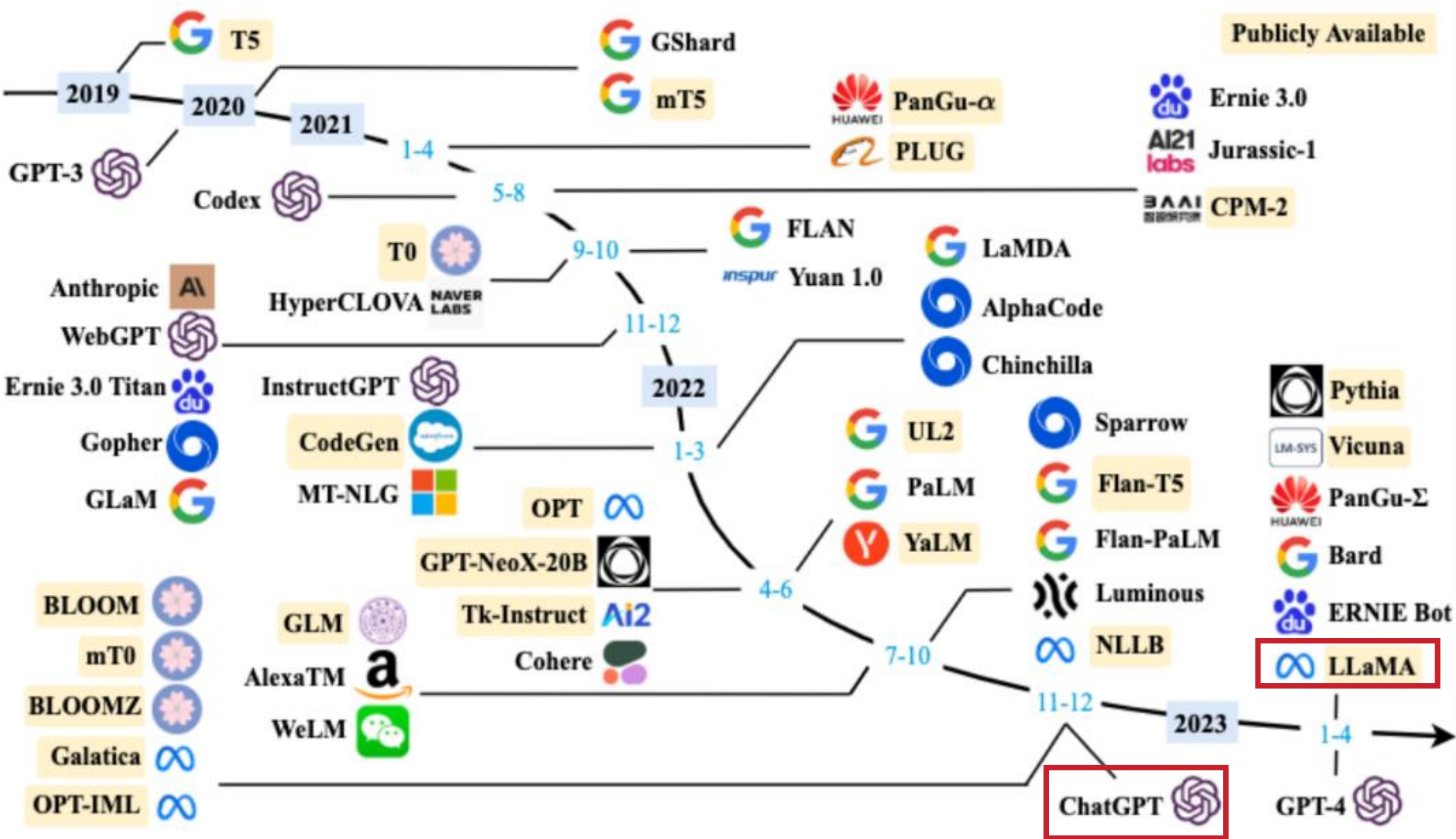

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d_k}}\right) \times V = Z$$


03 大模型的应用场景

大模型的应用场景



大模型的应用场景

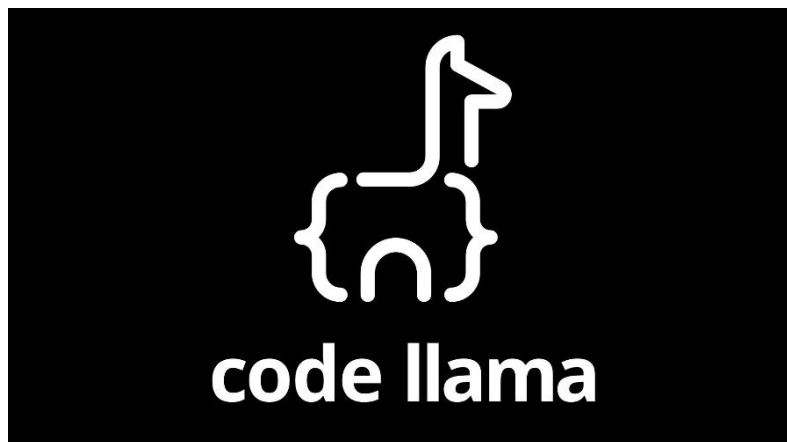
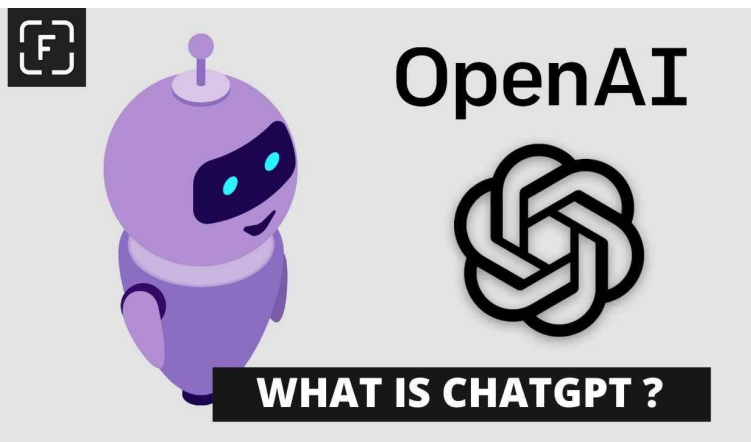


大模型的应用场景

ChatGPT

Codellama2

Stable diffusion



Codellama2生成代码



ChatGPT生成大纲



Autolabeling自动标注



Stable Diffusion

拳击手



悬崖瀑布
水彩画



麦田秋天
油画



树林溪流
水墨画



漫画脸
+微笑



改发色+
长发



大模型的应用场景



5月，百度Apollo汽车智能化业务展示了以大模型为基础的新一代AI智舱探索成果



大模型的应用场景-争议

01

偏见和不公平性：

03

资源消耗和环境成本

05

对人类工作的影响

02

缺乏透明度和可解释性

04

滥用和恶意用途



THANKS

