

教育部-华为智能基座课程

《人工智能基础与实践》

第13章：大模型原理与应用

授课教师：丛润民

山东大学
控制科学与工程学院

章节目录

CONTENTS

- 01 | 大模型发展与演练
- 02 | 大模型预训练范式
- 03 | 多模态大模型
- 04 | 基础大模型
- 05 | 大模型应用与范式扩展



章节目录

CONTENTS

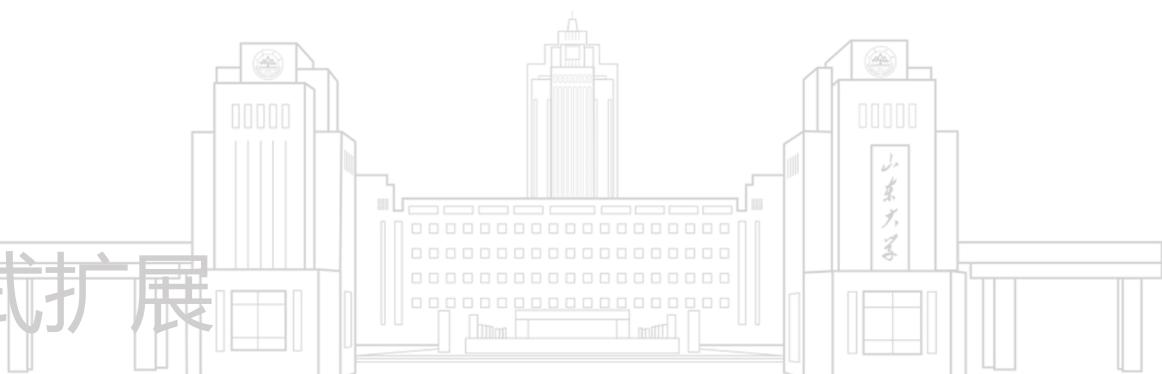
01 | 大模型发展与演练

02 | 大模型预训练范式

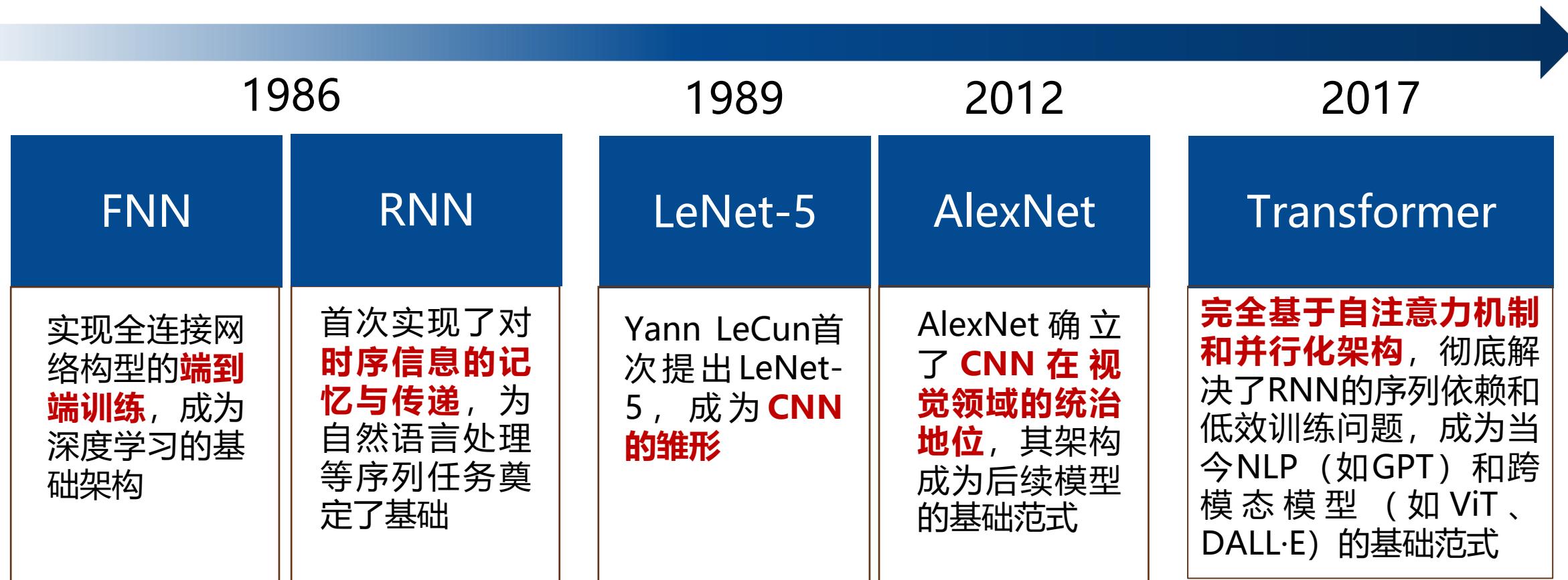
03 | 预训练、微调与对齐技术

04 | 基础大模型

05 | 大模型应用与范式扩展

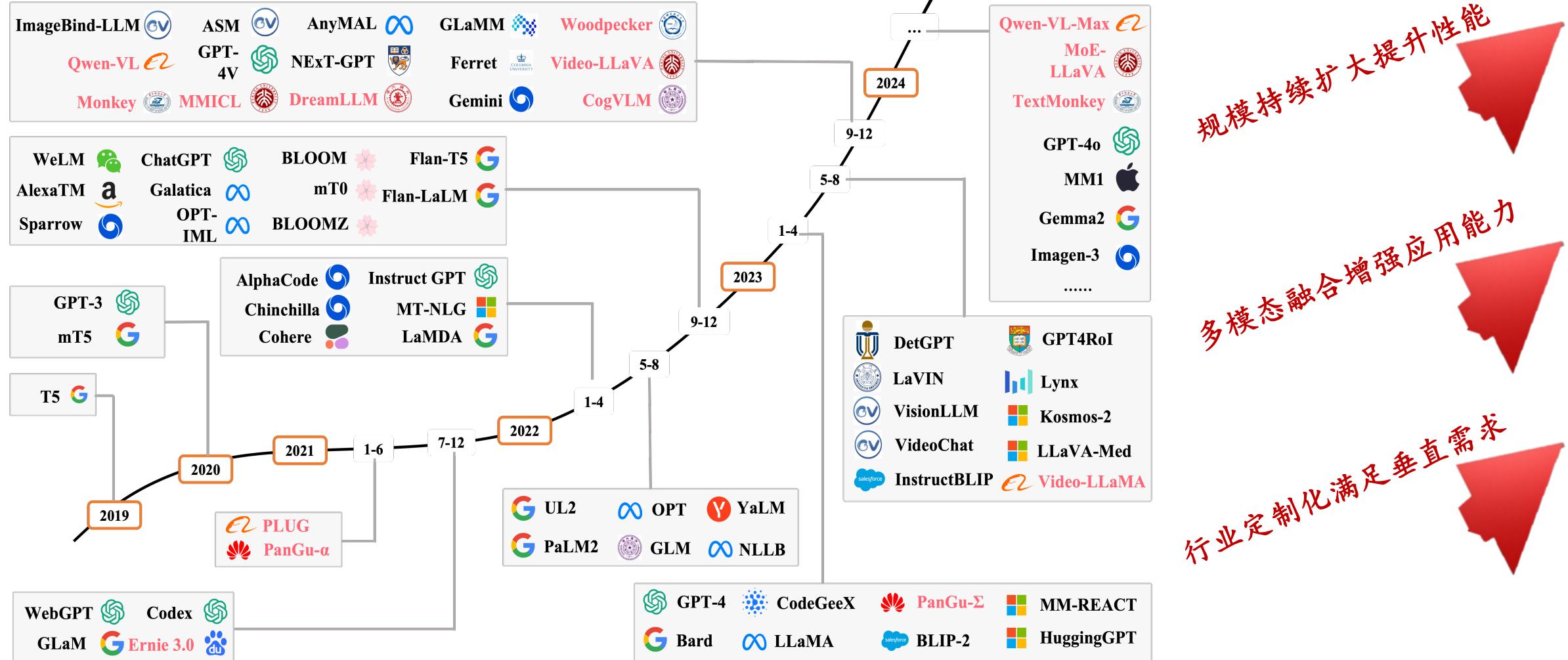


- 人工智能作为目前的热门研究方向，发展历程也是极为迅猛。
- 从“结构简单、能力专用”向“结构统一、能力通用”的范式转变

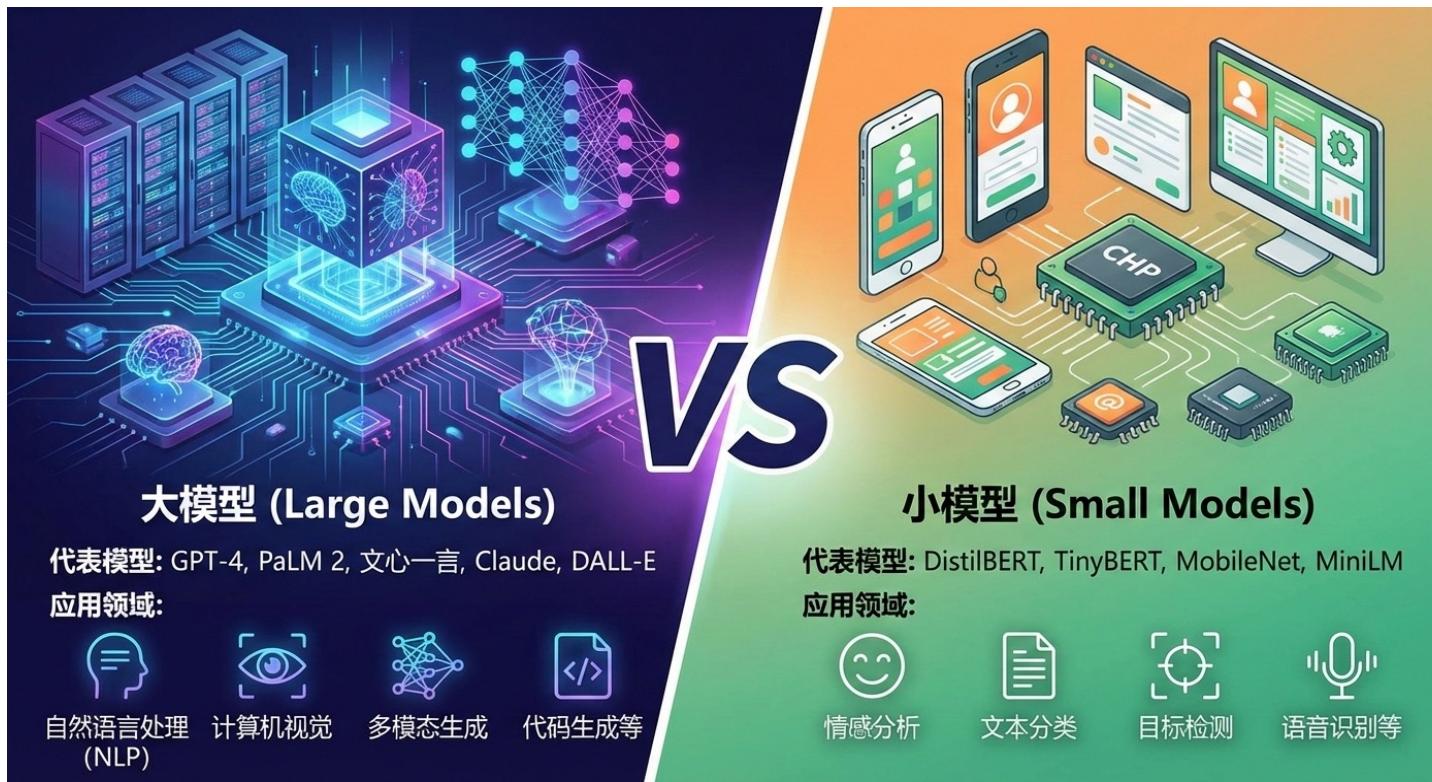


- 每次范式变革，核心驱动力都来自**计算能力与数据规模**

■ 大模型，简单来说，是指那些参数规模巨大、计算量复杂的深度学习模型



- 大模型是指具备超**大参数量和海量训练数据的AI模型**，通常应用与深度学习、大模型预训练和多模态任务。可以在没有微调或少量调整的情况下，执行多种复杂任务。
- 小模型是指**参数规模相对较小**，专注于特定任务或垂直领域的模型。小模型训练成本较低，运行速度快，适合边缘计算和移动设备部署。



*ChatGPT Nano-Banana-Pro生成

➤ 自然语言处理: ChatGPT - 文本问答示例



如何看待世界互联网大会这个平台?



世界互联网大会是一个重要的国际平台，专门讨论互联网技术及其对社会、经济和文化的影响。这个平台汇聚了来自世界各地的互联网专家、技术人员、政府官员和商界领袖，共同探讨互联网发展的未来趋势和挑战。

世界互联网大会的重要性在于，它促进了全球互联网技术和应用的发展，促进了国际间的技术交流和合作。它还为政府、企业和其他组织提供了一个展示最新互联网技术和创新成果的平台，



你能帮我写个类似于“羊了个羊”的小程序代码嘛



当然可以！这是一份用 Python 实现的“羊了个羊”小程序代码：

```
python
import random

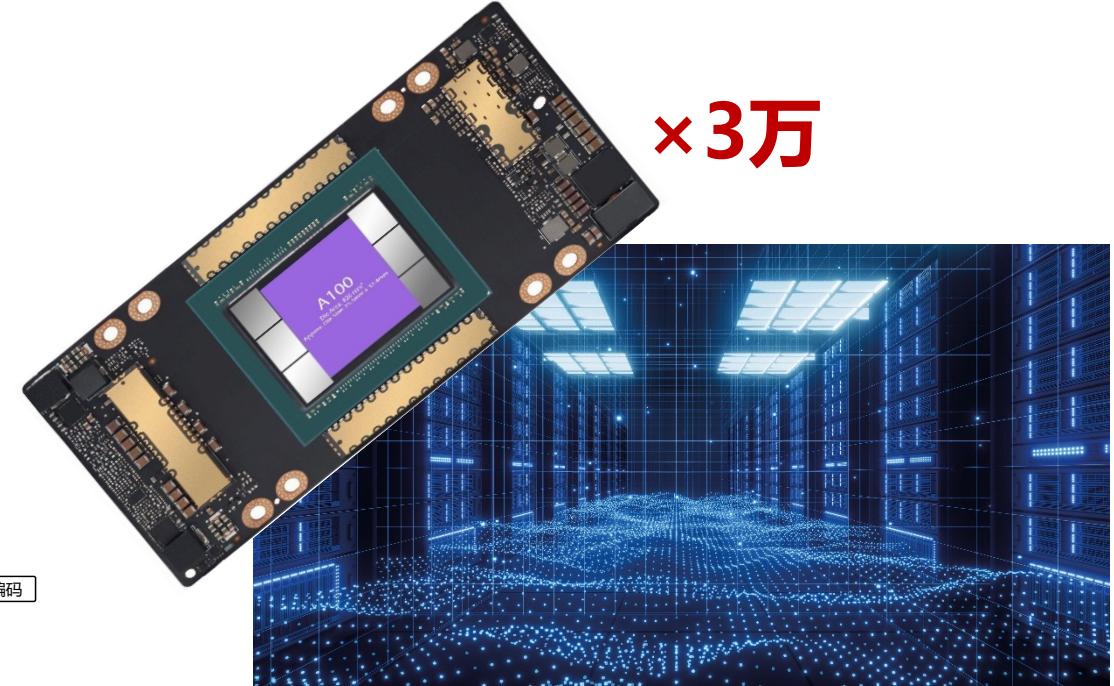
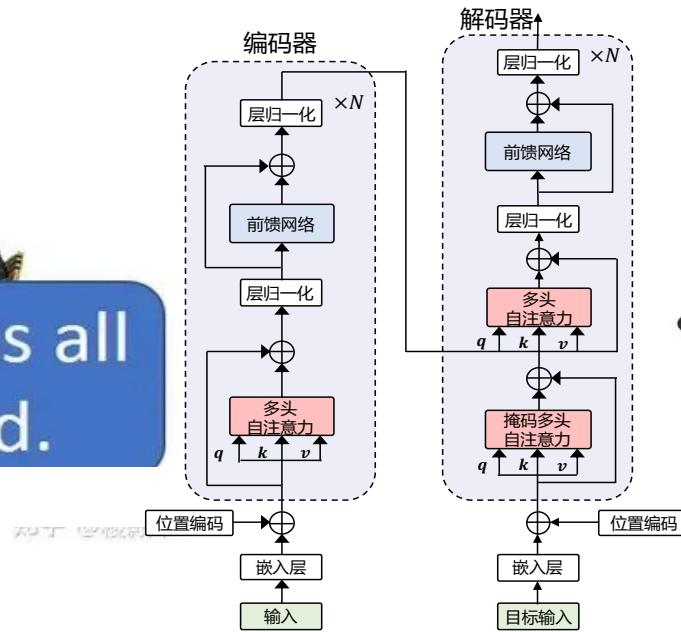
def sheep_game():
    # 设置羊数量
    sheep_count = random.randint(10, 100)

    # 设置循环次数
    loop_count = 0

    while sheep_count > 0:
        print("羊数量: ", sheep_count)
        sheep_count -= 1
        loop_count += 1

    print("经过了", loop_count, "轮羊了个羊，所有羊都被羊完了。")
```

- 谷歌提出了Transformer架构，利用自注意力机制构建长程序关系的同时，实现了对模型的并行化训练，这为研发大模型提供了并行优化的基础模型结构。



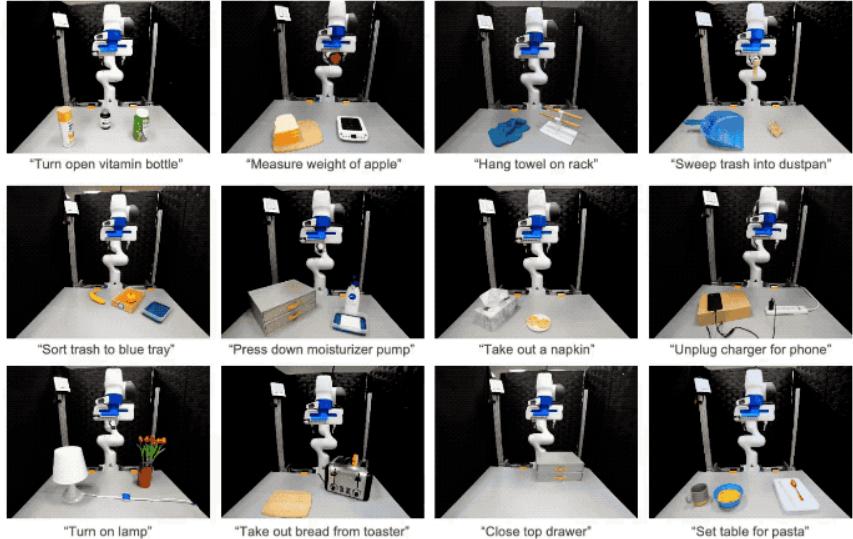
ChatGPT使用Transformer为核心组件

ChatGPT的初始版本参数量就达到了1750亿！！

训练一个ChatGPT的模
型需要上万张A100！！！



■ 大模型的优势与挑战皆根植于其“**大**”这一本质特征之中。



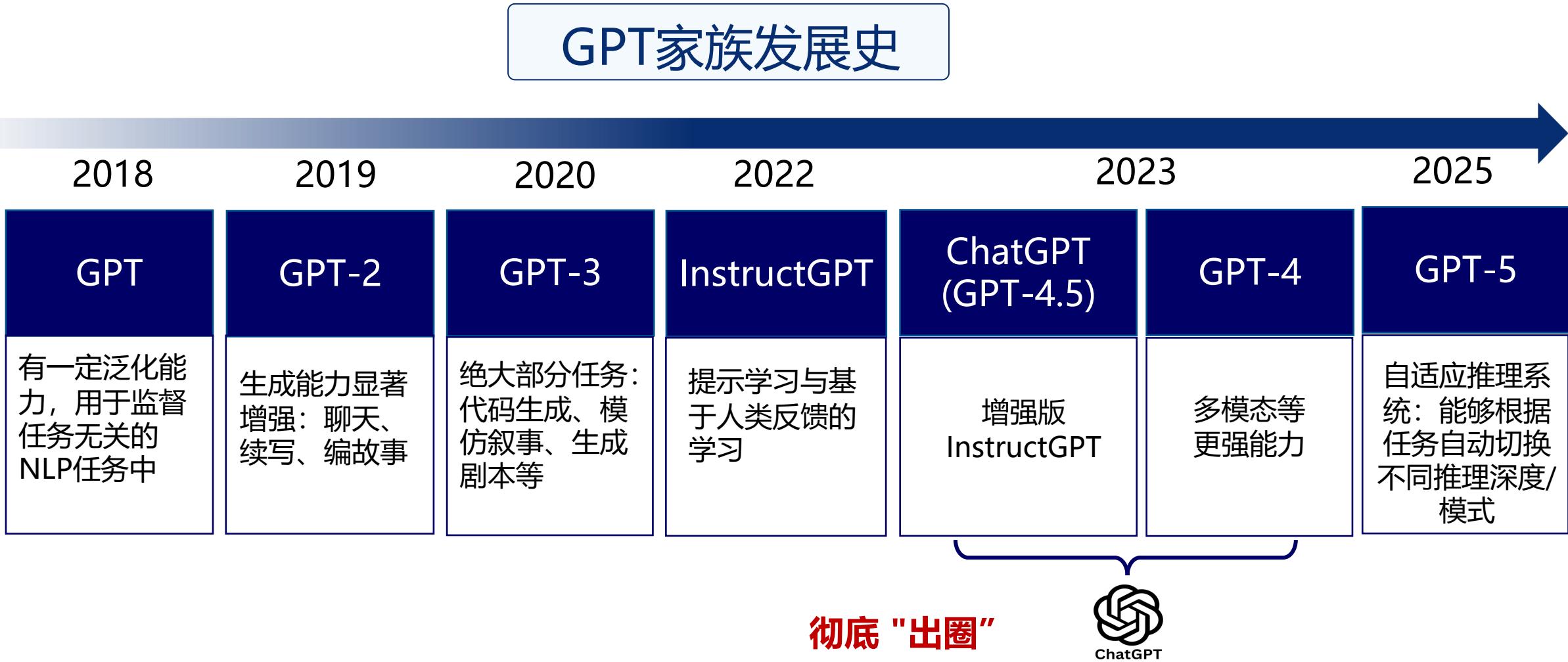
➤ **零样本迁移能力：**模型在没有见过特定任务的训练数据的情况下，仅根据已有的知识和对其他相关任务的学习经验，就能对该新任务进行处理和做出合理预测的能力。而且大模型通常可以在多个任务之间迁移和共享知识，这意味着在特定任务上训练的大模型可以迅速适应其他相关任务。

➤ **涌现能力：**随着模型规模的增大，模型表现出一些原本在较小规模时没有明显出现的新功能或特性。这些能力往往不是通过明确的编程或者训练目标获得的，而是在模型的规模达到一定程度后自发涌现的，称为“涌现能力”。





■ GPT更新推动模型能力跃迁，实现更强理解、推理与自然交互，拓展智能应用边界



■ 按照输入数据类型的不同，大模型主要可以分为以下三大类：



- **语言大模型：**是指在自然语言处理领域中的一类大模型，通常用于处理文本数据和理解自然语言。
- **视觉大模型：**是指在计算机视觉领域中使用的大模型，通常用于图像处理和分析。
- **多模态大模型：**是指能够处理多种不同类型数据的大模型，例如文本、图像、音频等多模态数据。

■ 按照应用领域的不同，大模型主要可以分为 L0、L1、L2 三个层级。

- **L0通用大模型：在多个领域和任务上通用的大模型。**通用大模型就像完成了大学前素质教育阶段的学生，有基础的认知能力。
- **L1行业大模型：针对特定行业或领域的大模型。**它们通常使用行业相关的数据进行预训练或微调，以提高在该领域的性能和准确度。
- **L2垂直大模型：针对特定任务或场景的大模型。**它们通常使用任务相关的数据进行预训练或微调，以提高在该任务上的性能和效果。



- 目前大模型的性能依然在持续不断地刷新，并不断有新的大模型推出，2025年初爆出圈的“国货之光”Deepseek大模型，堪称大模型的性价比典范。



遥遥领先的国产大模型之光



DeepSeek

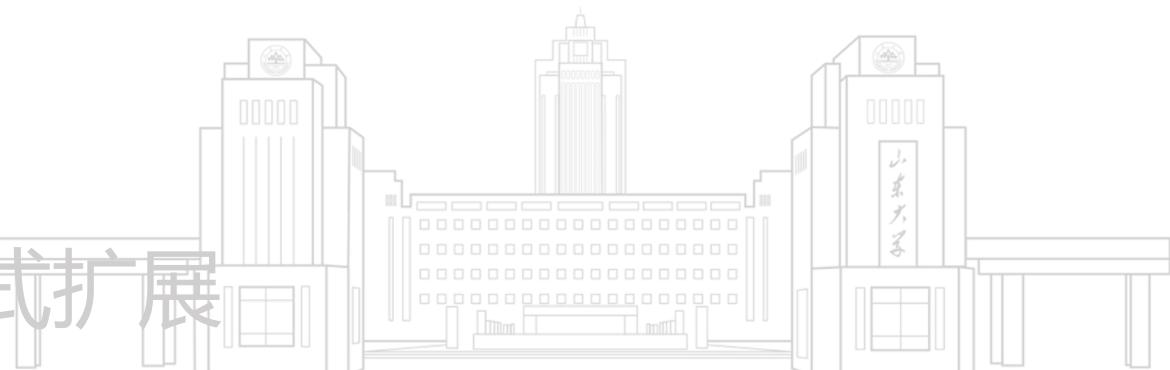
Qwen

以超高的性价比一夜之间火爆全球！

章节目录

CONTENTS

- 01 | 大模型发展与演练
- 02 | 大模型预训练范式**
- 03 | 多模态大模型
- 04 | 基础大模型
- 05 | 大模型应用与范式扩展



■ 大模型的研发主要分为两个阶段，即预训练阶段和微调与对齐阶段。

➤ 预训练是大模型研发的关键阶段，其成效直接关乎模型后续能力的广度与深度。预训练阶段首先要准备的便是大规模、高质量的训练数据。

例句：张明正在\$#&学习大模型的相关内容。张明正在学习大模型的相##关内容。



■ 大模型的研发主要分为两个阶段，即预训练阶段和微调与对齐阶段。

➤ 在模型架构方面，目前国内外主流的大模型主要基于Transformer框架包括以下3种：

- 编码器-解码器 (Encoder-Decoder) —— "翻译官模式"

输入句子：[天气][真][好]

↓ ↓ ↓ (编码器全部可见，蓝色)

编码记忆： 

↑ ↑ ↑ (解码器可随时查看，红色)

输出生成：[The] [weather] [is] [nice]

→ 生成时能参考完整输入

- 因果解码器 (Causal Decoder) —— "写日记模式"

生成顺序：[今天][天气][真][好]

|
↓ |
↓ |
↓ |

可见范围：[今]—[今天]—[今天天气]—[今天天气真]
(红色可见，绿色被遮挡)

■ 大模型的研发主要分为两个阶段，即预训练阶段和微调与对齐阶段。

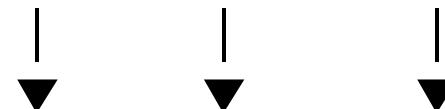
➤ 在模型架构方面，目前国内外主流的大模型主要基于Transformer框架包括以下3种：

- 编码器-解码器 (Encoder-Decoder) —— "翻译官模式"
- 因果解码器 (Causal Decoder) —— "写日记模式"
- 前缀解码器 (Prefix Decoder) —— "命题作文模式"

前缀提示：[写科幻故事：] [火星上...]

 (蓝色，始终可见)

生成内容：[人类][发现了][金字塔]



可见范围：[火]—[火星上]—[火星上人类]—[火星上人类发现了]

(绿色部分仍受因果限制)

- 在进行大规模预训练时，往往需要设计合适的**自监督预训练**任务，使得模型能够从海量无标注数据中学习到广泛的语义知识与世界知识。

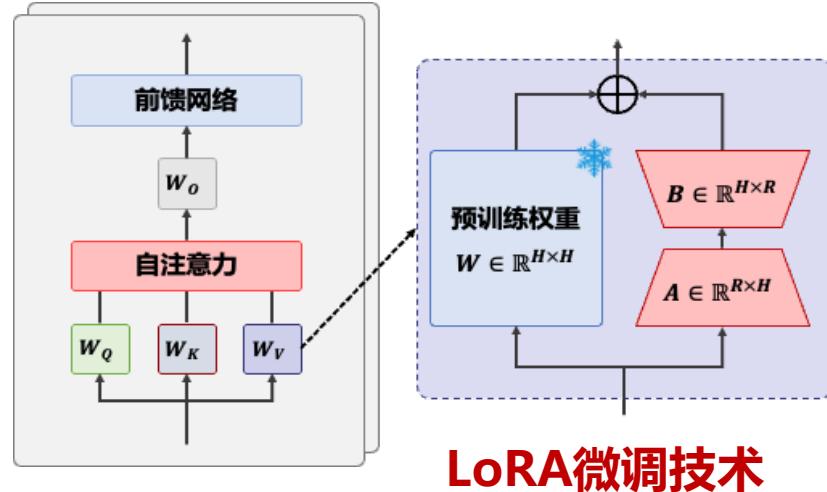
	语言建模	去噪自编码器	混合去噪器
目标	学习词元序列的概率分布	学习从破坏的文本中恢复原始文本	学习从多种破坏模式中恢复原始文本
输入	完整的文本序列	被破坏的文本	被多种方式破坏的文本
输出	无	删除、替换、重排词元等	多种噪声类型 (如删除、替换、重排等)
噪声强度	无	固定或可调	可调
模型结构	编码器	编码器+解码器	编码器+解码器
训练目标	最大化下一个词元的预测概率	最小化重构文本与原始文本的差异	最小化重构文本与原始文本的差异
损失函数	负对数似然	均方误差或交叉熵损失	均方误差或交叉熵损失
应用场景	序列生成、翻译、问答等	文本补全、错误纠正、文本恢复等	多种文本处理任务，增强鲁棒性和泛化能力
鲁棒性	较低	较高	更高
泛化性	中等	较强	更强
优点	学习词元间依赖关系	提高模型对抗噪声的能力	提高模型对抗多种噪声的能力
缺点	可能忽视上下文信息	只能处理特定类型的噪声	训练过程可能很复杂

- 在特定的应用场景下，需要对大模型进一步优化，使其能够有效地应用预训练阶段获取的知识，并理解、适应人类意愿，在不同任务下表现出优秀的准确性和适用性，从而更好地服务于实际应用需求。



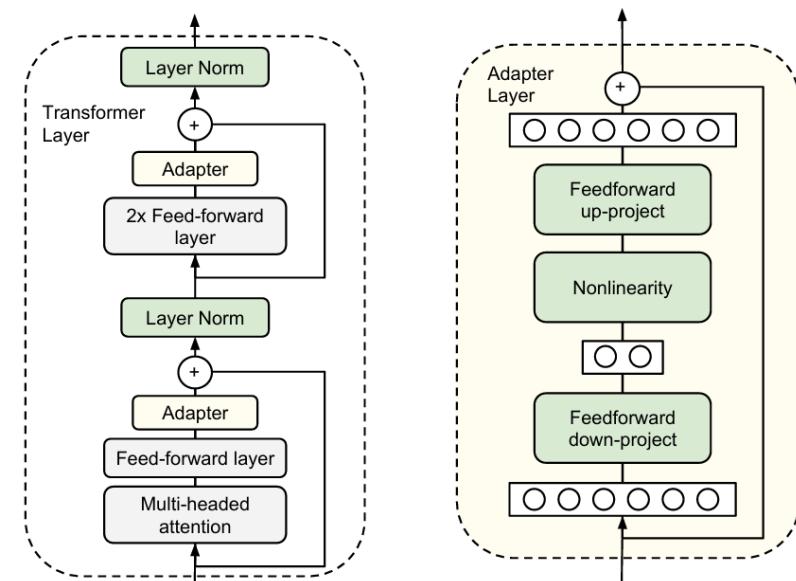
微调方法	内存需求	相对原始模型	典型硬件要求
全参数微调	极高	100%	A100集群(≥8卡)
LoRA ($r=16$)	中	~10-15%	单张 A100/A6000
QLoRA (4-bit)	低	~2-5%	RTX 3090/4090
Adapter	中	~15-20%	单张 A100/V100
P-Tuning v2	低	~5-10%	RTX 3080/3090
Prompt Tuning	极低	~1-2%	消费级 GPU(8GB+)
BitFit	低	~3-5%	消费级 GPU(12GB+)

- **全参数微调**: 对模型所有参数进行更新，效果最好但计算与存储成本最高。
- **冻结主干微调**: 冻结底层或大部分参数，仅微调高层或任务相关层。



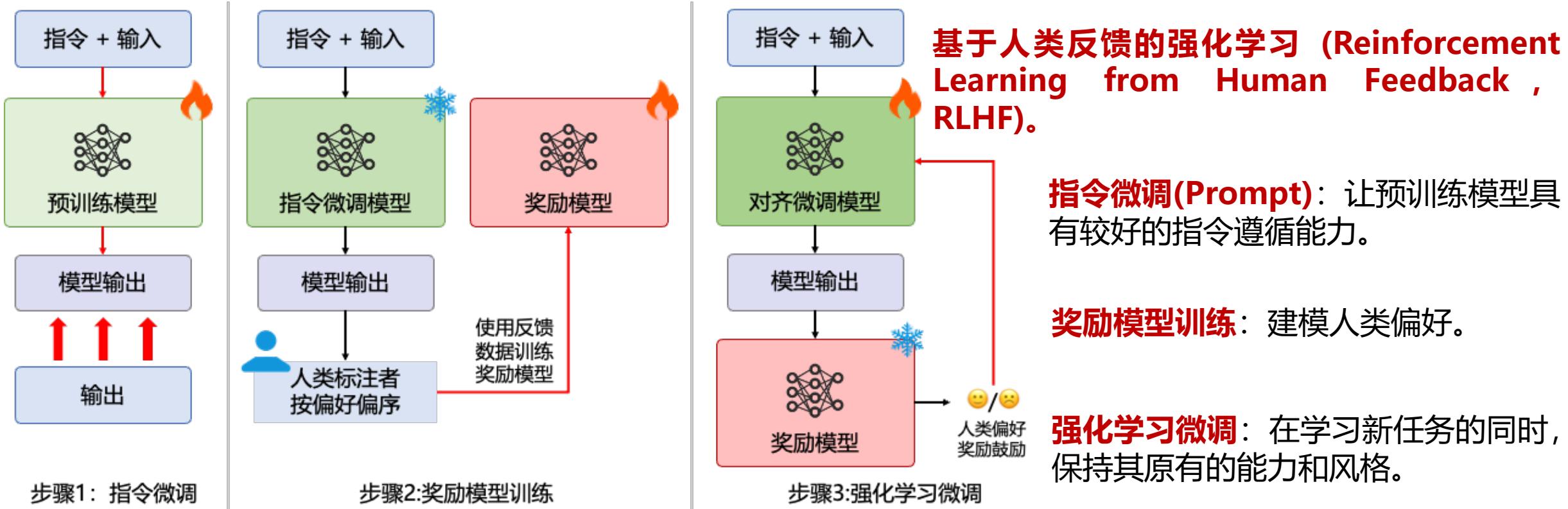
- **Adapter微调技术**: 在多头注意力和前馈网络之后增加两个前馈子层来实现，这两个子层分别是降维和升维的前馈网络，通过控制中间维度的大小来限制参数量。

➤ **LoRA微调技术**: 在不修改原有模型权重的情况下，通过引入低秩矩阵来调整模型的某些层，从而实现快速适应新任务的目的。



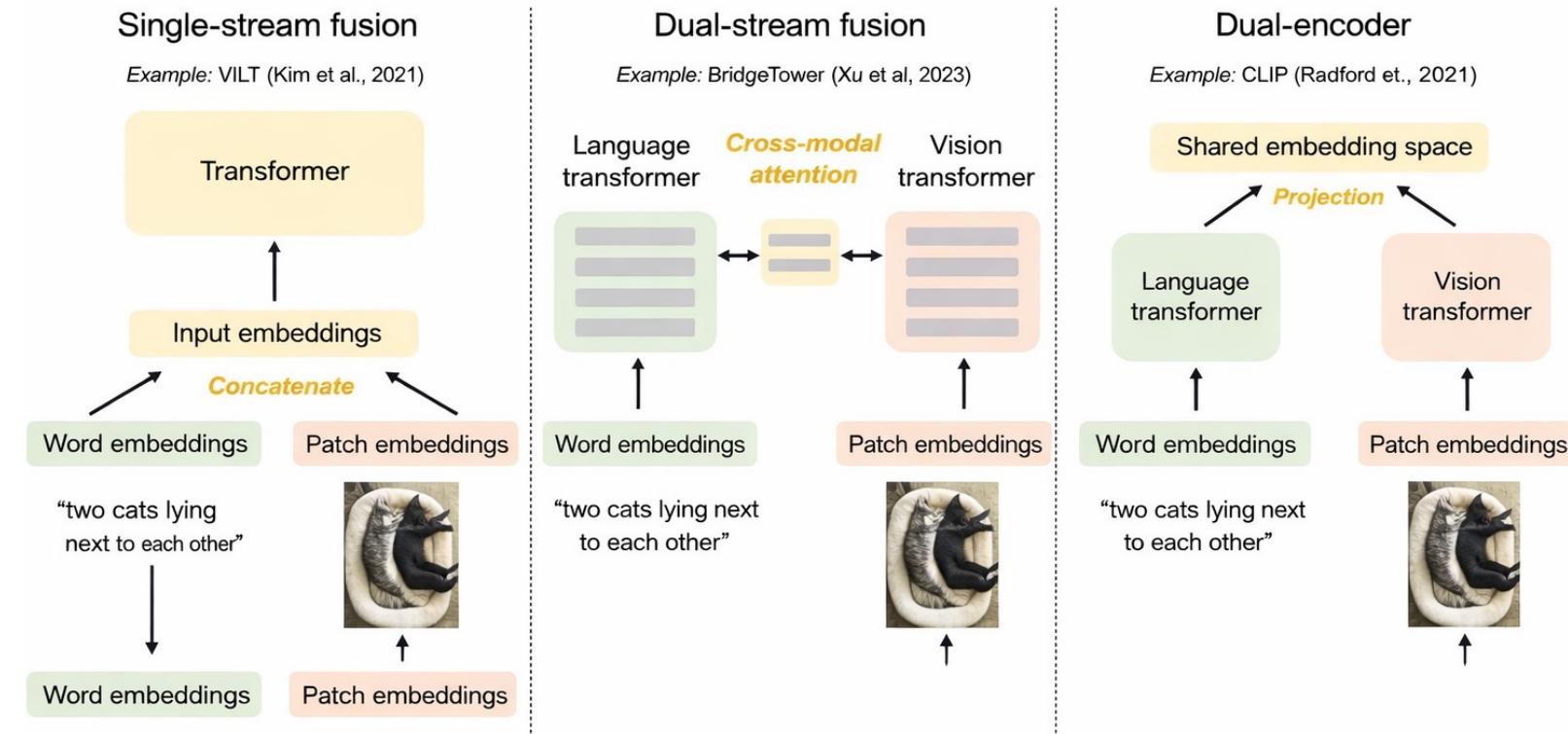
Adapter微调技术

- 强化学习微调的最大优势在于：直接优化“**人类偏好/任务目标**”，而不是仅拟合数据分布。LoRA/全量微调解决的是“学知识”，强化学习微调解决的是“学行为与目标**对齐**”。
- RLHF是强化学习中的一种。通过与环境互动获得环境的反馈，形成新的交互方式与策略，智能体可以更好地从环境中学习，获得更大的奖励回报。



■ 多模态对齐 (Model/Architecture Alignment) 解决的是“模型有没有能力把不同模态的信息放到同一个决策空间里” -- 模态对齐

- **Single-stream:** 词嵌入和图像patch嵌入被直接拼接，模型在每一层Transformer中同时处理跨模态关系。
- **Dual-stream:** 为文本和图像分别建立独立的Transformer，并通过显式的跨模态注意力模块进行双向交互。
- **Dual-encoder:** 文本和图像完全独立编码，仅在最终通过投影映射到共享语义空间进行对齐，更适合大规模对比学习和检索任务。

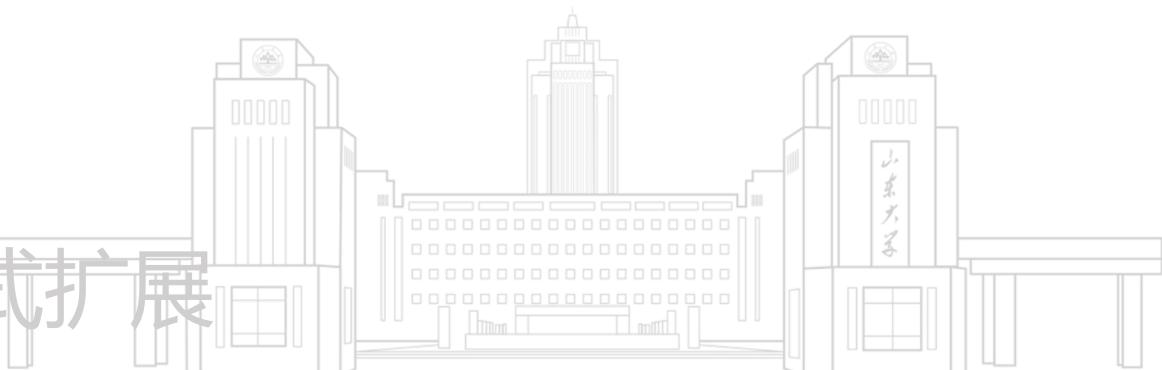


■ 通过模型设计，让不同模态的信息在同一个 Transformer 决策空间中可被联合建模。

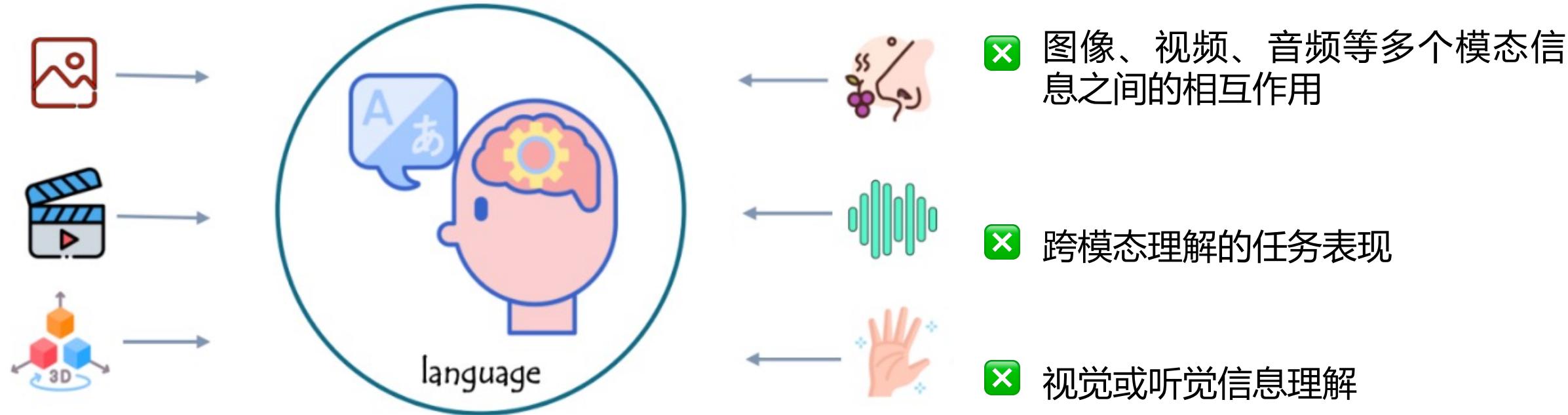
章节目录

CONTENTS

- 01 | 大模型发展与演练
- 02 | 大模型预训练范式
- 03 | 多模态大模型**
- 04 | 基础大模型
- 05 | 大模型应用与范式扩展



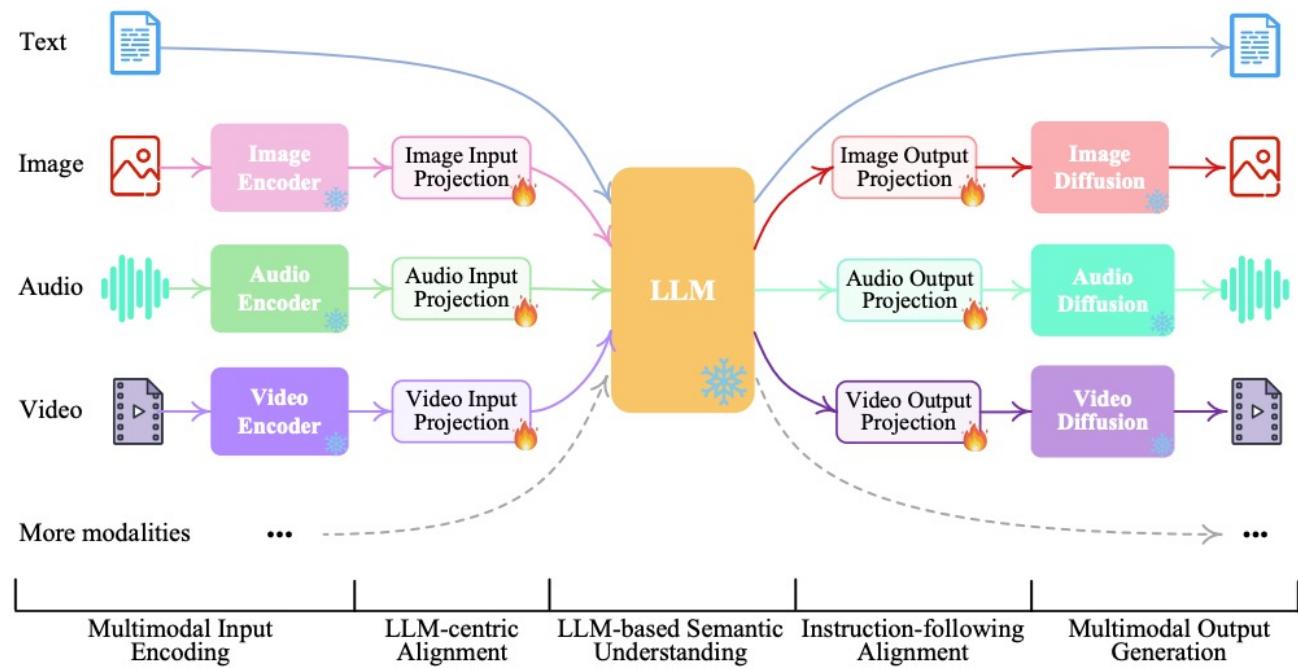
- 大语言模型，与其他模态的模型连接起来，实现协同对推理，催生了一个新的研究领域—**多模态大语言模型（Multi-Modal Large Language Model, MLLM）**



多模态大语言模型通过对多种模态数据的联合学习和推理，模型能够捕捉到不同模态之间的内在关联和互补信息，从而更全面地理解和描述现实世界中的事物和现象。

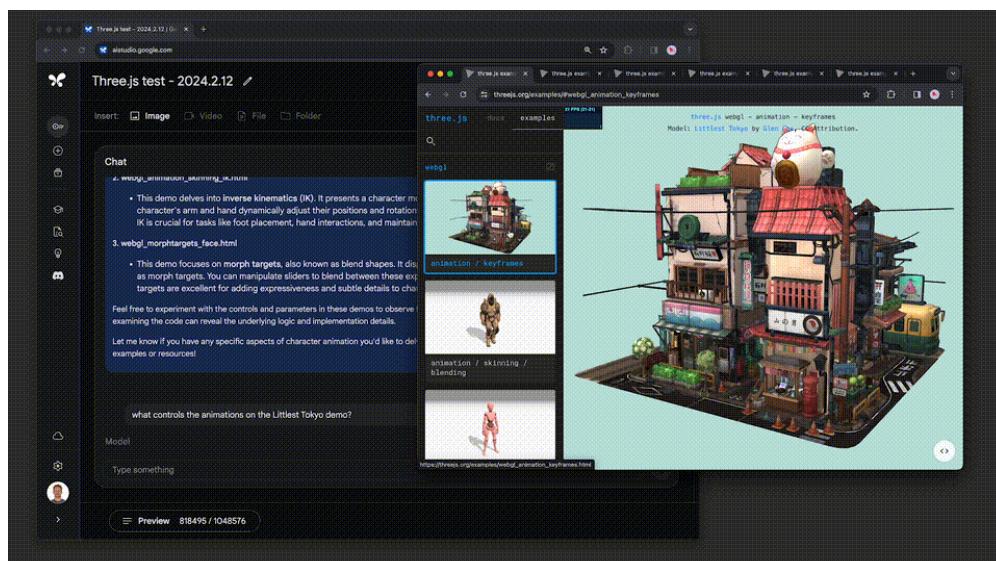
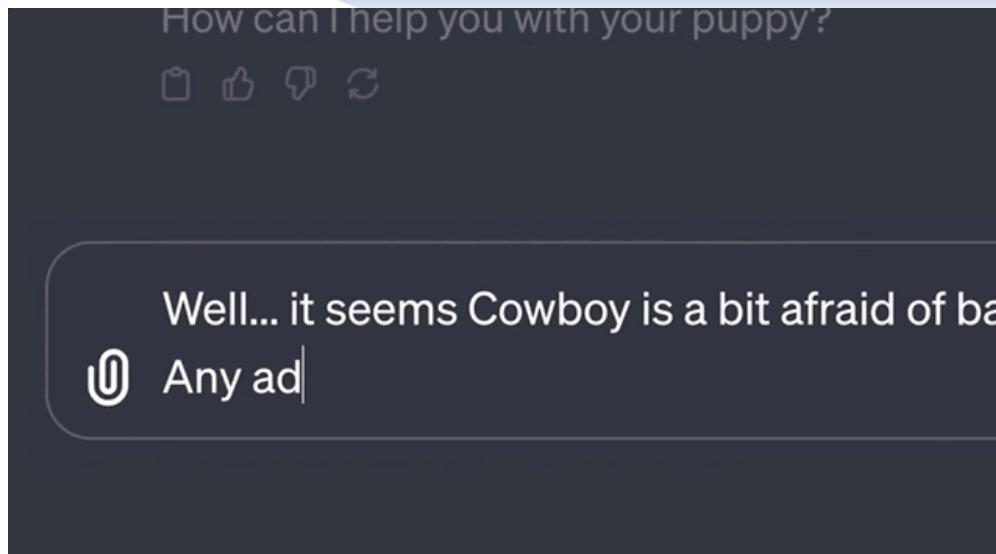
多模态大模型的概念与演进

- 多模态大模型的关键能力在于整合并理解不同的数据格式，应用于需要理解和整合不同类型数据信息的任务。**LLMs处于核心位置，多模态在LLMs基础上进行扩展**，即找到一个方法将不同模态的数据映射到LLMs可以接收的语义空间。



- **多模态编码器**: 负责将不同模态的输入数据编码为模型可理解的表示。
- **输入映射**: 将不同模态的输入数据映射到共享的语义空间。
- **LLMs**: 大型语言模型，用于处理文本数据。
- **输出映射**: 将模型生成的输出映射回原始模态的空间。
- **多模态生成器**: 根据输入数据生成对应的输出数据。

多模态大模型



多模态内容理解和文本生成



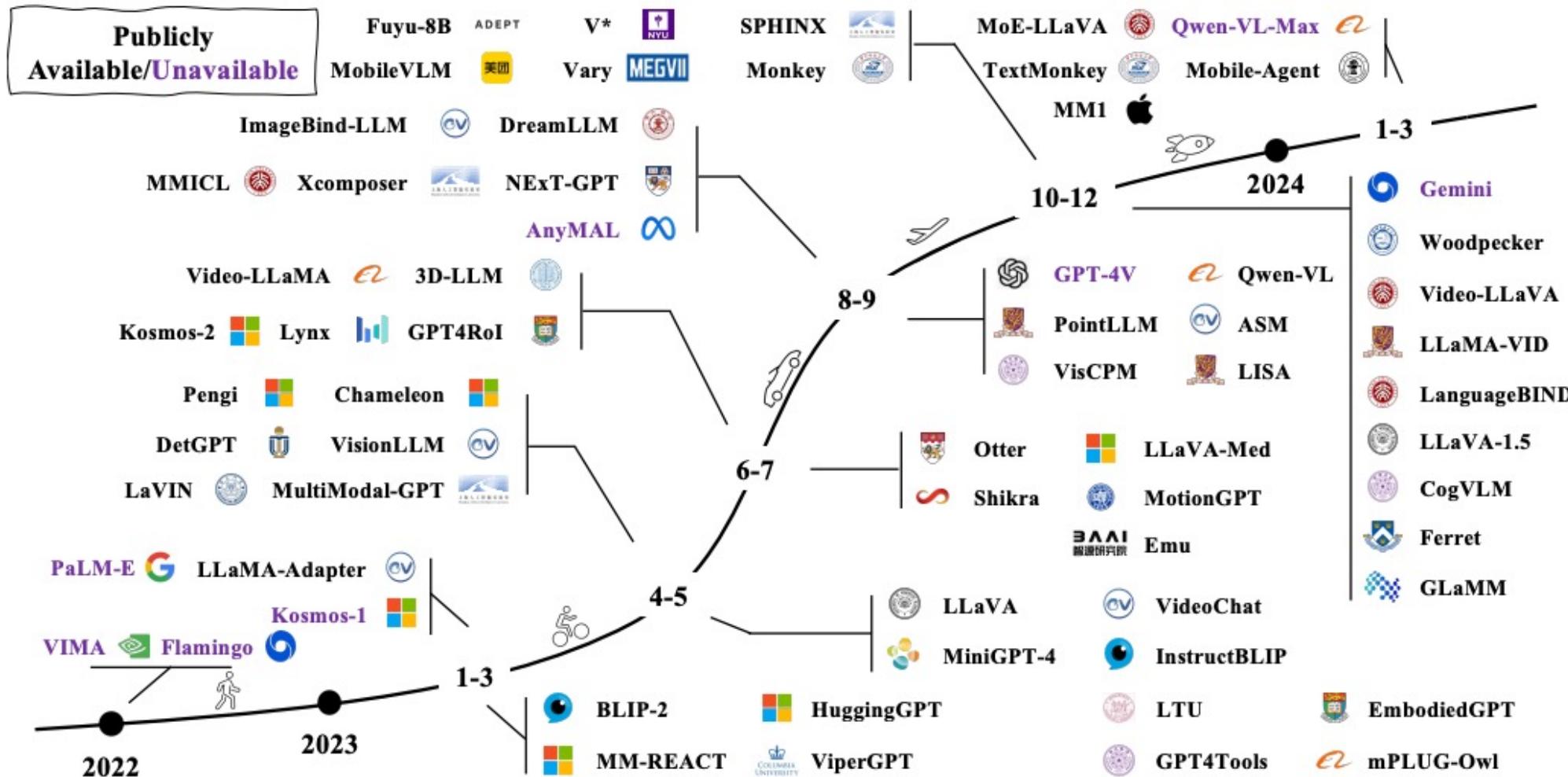
特定模态内容生成



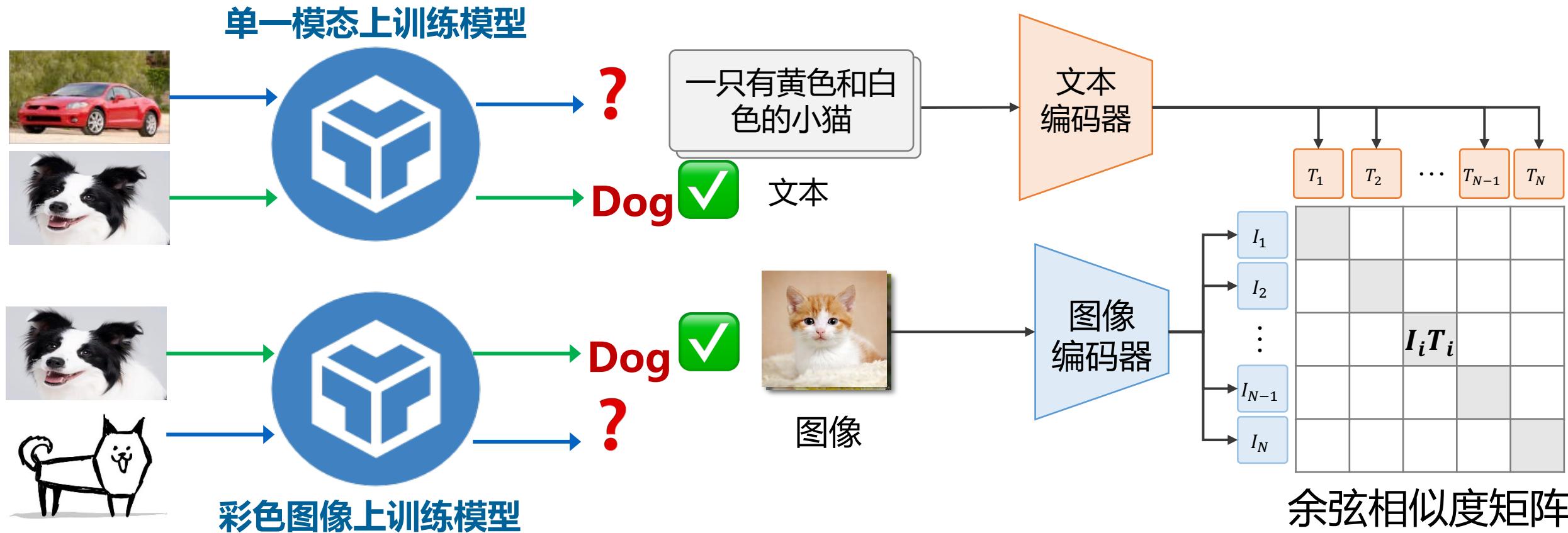
各种模态内容生成



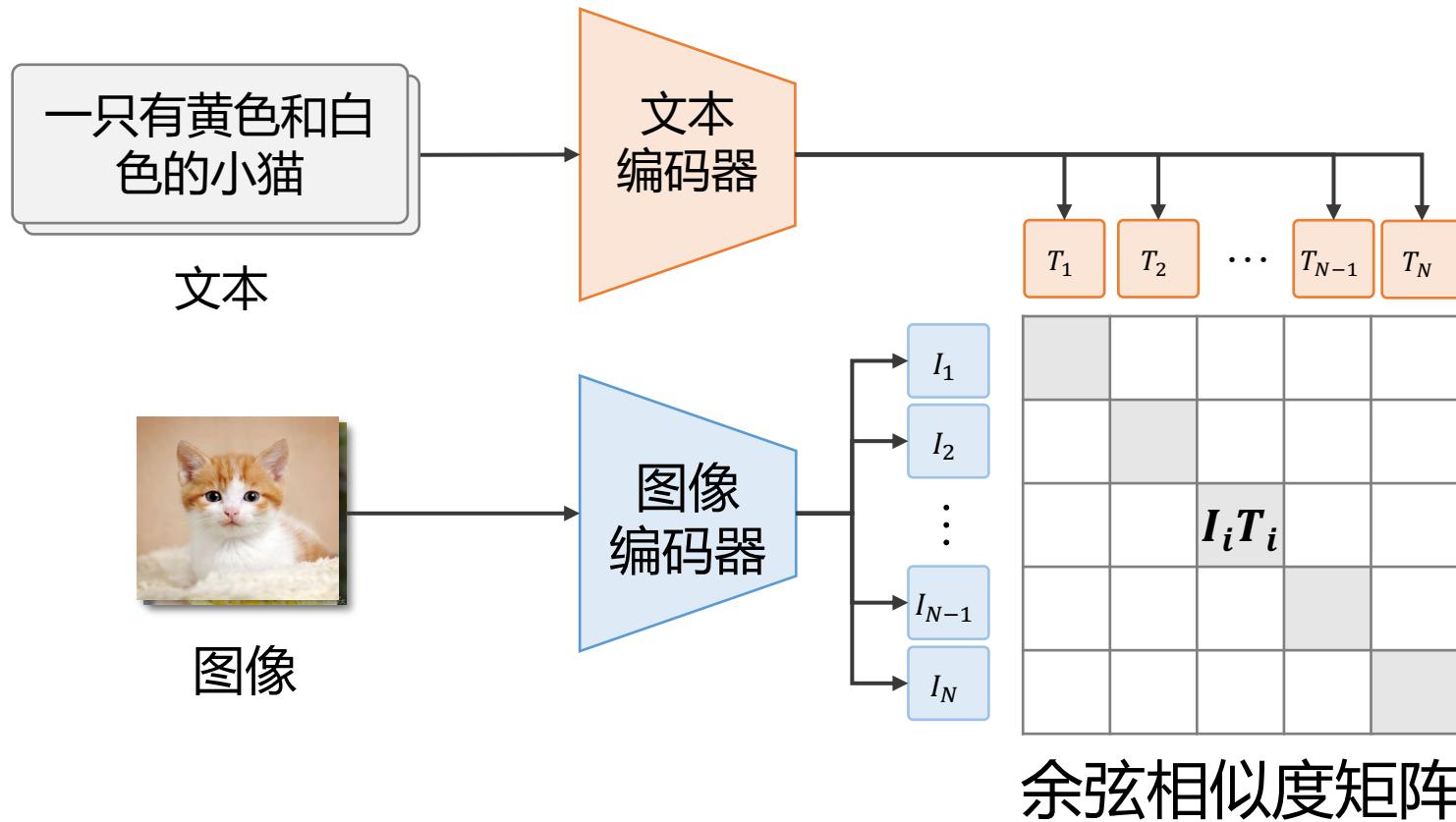
■ 多模态大模型历经融合发展，如今其综合处理能力不断提升，在各行业广泛应用。



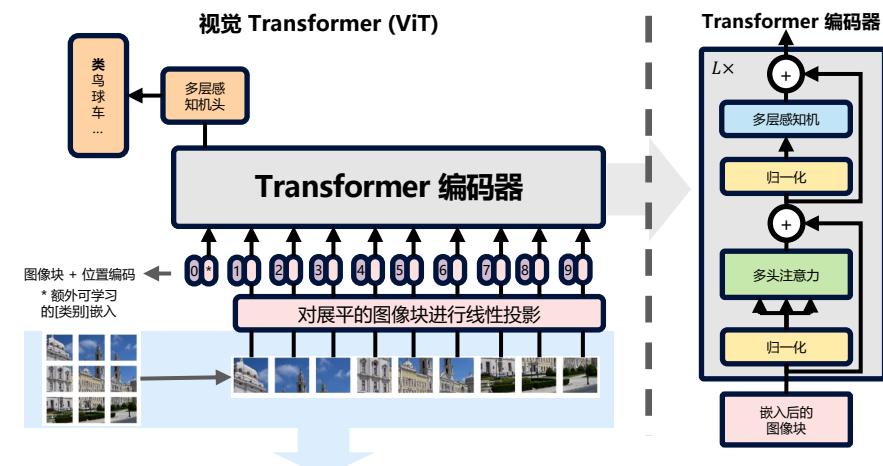
- CLIP (Contrastive Language-Image Pre-training) 探索了如何利用大规模未标注数据进行有效的多模态学习，在图像分类、检索和生成等任务上表现出了强大的迁移能力，为多模态大模型的发展提供了重要的技术借鉴和思路启发。



CLIP – 多模态大模型的先驱



CLIP模型采用了一种双塔架构，分别是图像编码器和文本编码器，用于处理图像和文本数据。



在文本编码器方面，CLIP沿用了GPT-2的架构；而在图像编码器的设计上，CLIP经过多次尝试，比较了不同的ResNet和Transformer架构，最终选择了ViT模型作为其图像编码器。

CLIP – 多模态大模型的先驱

一只只有黄色和白色的小猫

文本

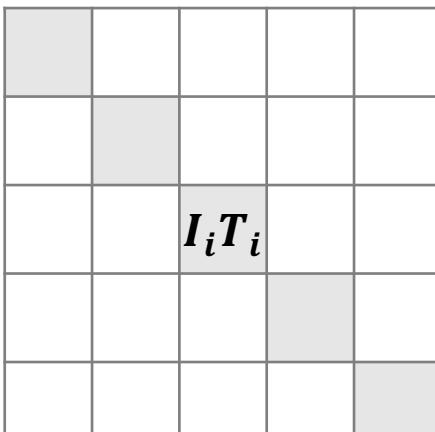


图像

文本
编码器

图像
编码器

T_1 T_2 ... T_{N-1} T_N



余弦相似度矩阵

$$S = (I_E \cdot T_E^T) \times e^\varepsilon \in \mathbb{R}^{N \times N}$$

对比
学习

最大化对角线上的相似度值（正确匹配）
尽可能减小非对角线上的值（错误匹配）

- CLIP模型采用了一种双塔架构，分别是图像编码器和文本编码器，用于处理图像和文本数据。
- 采用了自监督学习的方法，通过对大量无标签的图像-文本对进行训练来学习图像和文本之间的对齐关系。
- 通过对比学习的方法，将图像和文本映射到同一个高维语义空间中，使得它们能够在这个空间中相互理解和交互。

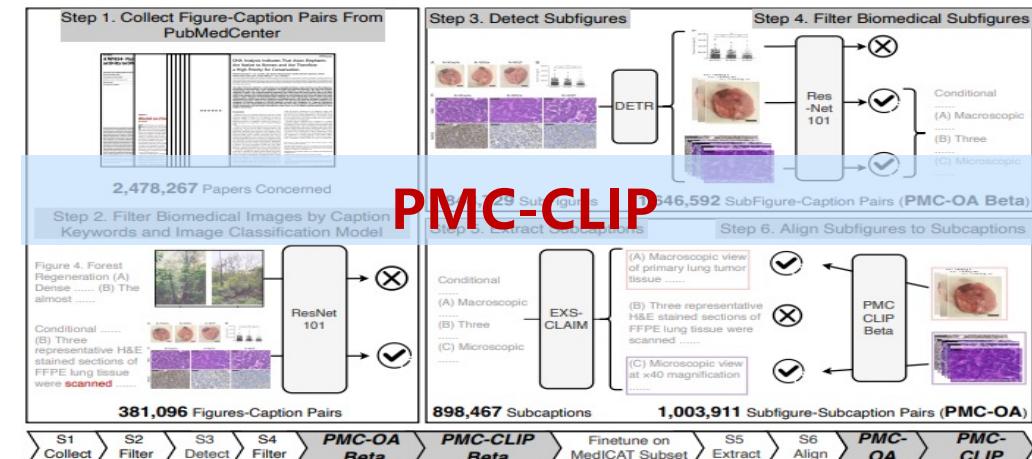
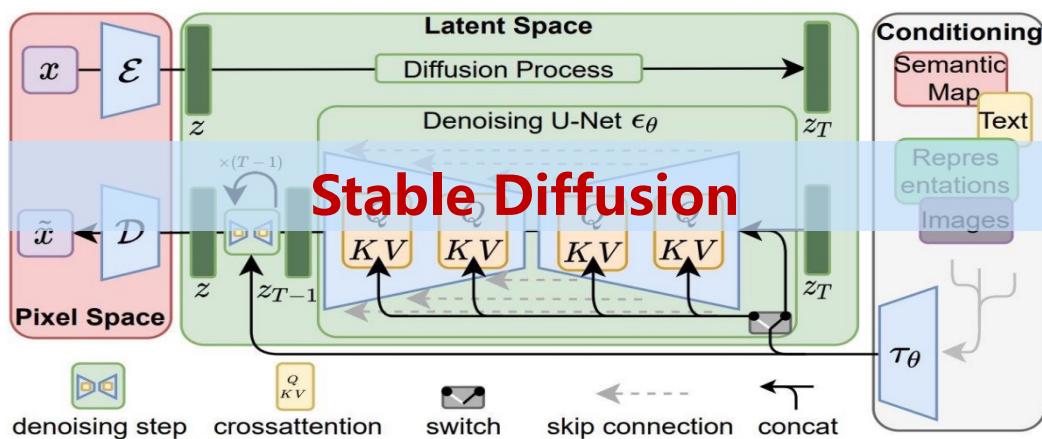
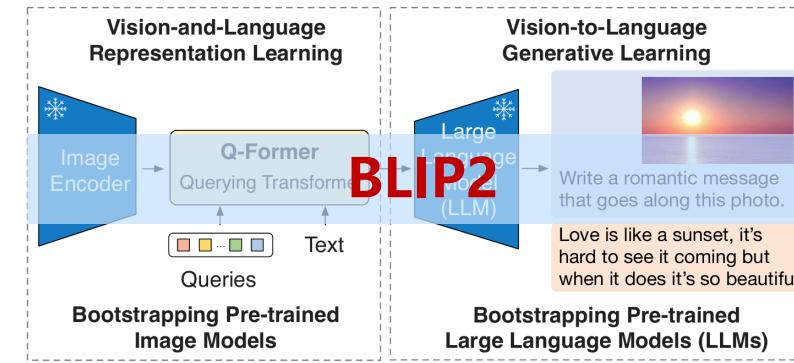
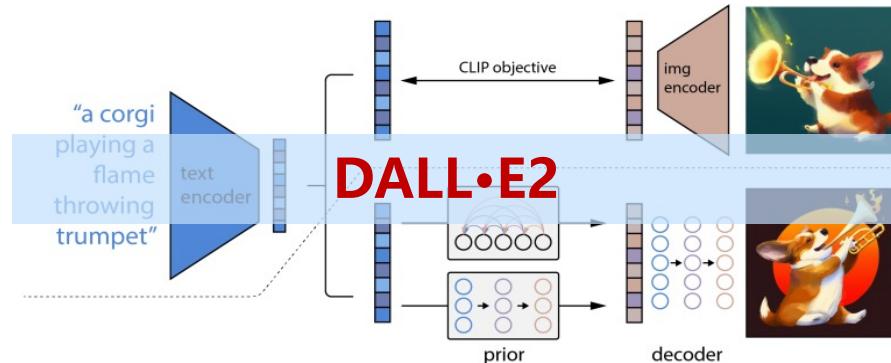
对比学习是一种自监督学习方法，通过将相似样本的特征拉近、不相似样本的特征推远，来学习数据的有效表示。具体来说，模型会构建正样本对（相似样本）和负样本对（不相似样本），通过优化损失函数，使正样本对在特征空间中更接近，负样本对更远离。



对比
学习

CLIP – 多模态大模型的先驱

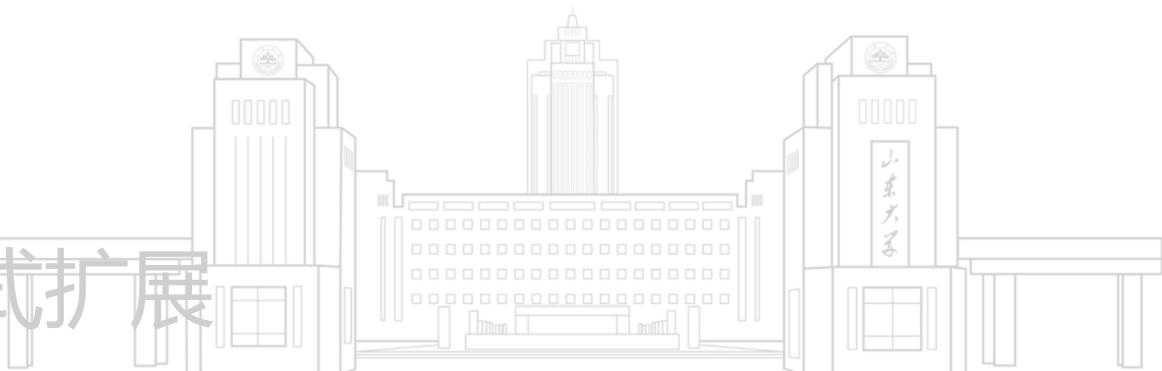
- CLIP通过对比学习的方式学习到了图像与文本之间的潜在关联，实现了扩模态数据的有效融合，并开创自监督预训练典范，提升迁移学习能力。



章节目录

CONTENTS

- 01 | 大模型发展与演练
- 02 | 大模型预训练范式
- 03 | 多模态大模型
- 04 | 基础大模型**
- 05 | 大模型应用与范式扩展



■ 基础大模型是基石，是底座，是基座。



■ 垂域大模型是在特定领域精耕细作的“尖兵”。



- 基础大模型是基于**大规模通用数据**预训练的模型，具备通用语言理解与生成能力。
 - 从规模上看，它拥有海量参数，常达数十亿甚至数万亿，像 GPT-3 就有 1750 亿参数，如此规模能捕捉数据中复杂模式。
 - 在通用性上，它不是针对特定任务，而是可处理多种类型任务和数据，比如文本、图像、语音等，像 BERT 模型能用于文本分类、问答等多种自然语言处理任务。
 - 从功能上，它为下游任务和领域模型提供基础，通过微调适配具体应用场景，减少重新训练成本。



- 基础大模型采用Transformer架构，依赖海量数据处理与预训练技术，利用分布式训练策略在大规模集群上进行训练。

01

模型架构

基础大模型采用Transformer架构，基于注意力机制，适合并行化训练，提升训练效率。



02

数据处理与预训练

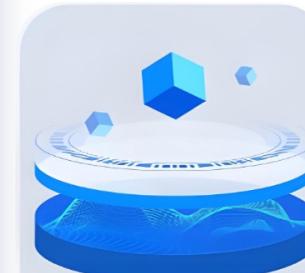
基础大模型对海量数据进行清洗与去噪，去除无用或错误信息。



03

模型训练与优化

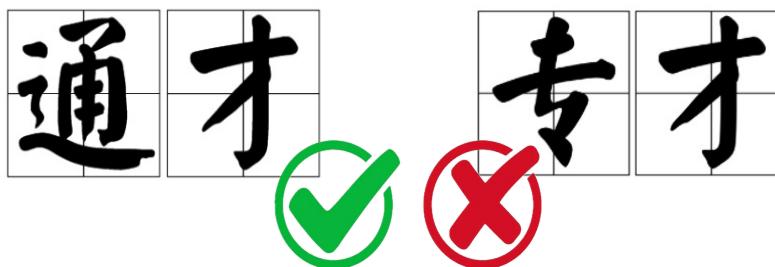
基础大模型采用数据并行和模型并行，加速模型训练过程。



- 基础大模型虽然通用性强，但是专业性不足，并且存在数据安全与隐私问题，以及脏/假数据对于模型的损害风险，同时面临计算资源与推理效率挑战。

通用性与专业性

基础大模型通用性强，但在专业领域表现不够精准，难以满足精细化需求。在处理专业术语和复杂任务时存在局限，难以满足特定领域的专业需求。



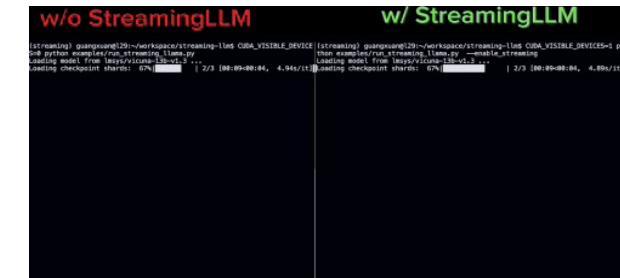
数据安全与隐私

基础大模型使用大量通用数据，涉及数据安全和隐私问题。需加强数据加密和隐私保护，确保数据安全，数据安全是基础大模型的重要问题。



计算资源与效率

基础大模型训练和推理需要大量计算资源，计算成本较高。计算资源是基础大模型的瓶颈，优化资源使用可以提升模型训练和推理效率。



- 2025年DeepSeek-V3以**低成本、高性能**登顶基础大模型排行榜，其在多个领域的基准测试中表现出色，为开源模型的发展提供了新的强大示例，推动了整个基础大模型应用的技术进步和爆发。

指标	DeepSeek-V3	GPT-4o
训练成本	557.6万美元	约1亿美元
参数量	671亿	1750亿
训练时长	278.8万GPU小时 (H800)	约2500万GPU小时 (A100) 
训练硬件	2048块英伟达H800 GPU 	约2.5万个英伟达A100 GPU 
训练效率	高效负载均衡、FP8混合精度训练 	未公开具体优化技术
性价比	显著高于GPT-4o	较低

Benchmark (Metric)	DeepSeek-V3	Qwen2.5-72B-Inst.	Llama3.1-405B-Inst.	Claude-3.5-Sonnet-1022	GPT-4o-0513
Architecture	MoE	Dense	Dense	-	-
# Activated Params	37B	72B	405B	-	-
# Total Params	671B	72B	405B	-	-
MMLU (EM)	88.5	85.3	88.6	88.3	87.2
MMLU-Redux (EM)	89.1	85.6	86.2	88.9	88
MMLU-Pro (EM)	75.9	71.6	73.3	78	72.6
DROP (3-shot F1)	91.6	76.7	88.7	88.3	83.7
English	IF-Eval (Prompt Strict)	86.1	84.1	86	86.5
	GPQA-Diamond (Pass@1)	59.1	49	51.1	65
	SimpleQA (Correct)	24.9	9.1	17.1	28.4
	FRAMES (Acc.)	73.3	69.8	70	72.5
	LongBench v2 (Acc.)	48.7	39.4	36.1	41
	HumanEval-Mul (Pass@1)	82.6	77.3	77.2	81.7
	LiveCodeBench(Pass@1-COT)	40.5	31.1	28.4	36.3
Code	LiveCodeBench (Pass@1)	37.6	28.7	30.1	32.8
	Codeforces (Percentile)	51.6	24.8	25.3	20.3
	SWE Verified (Resolved)	42	23.8	24.5	50.8
	Aider-Edit (Acc.)	79.7	65.4	63.9	84.2
	Aider-Polyglot (Acc.)	49.6	7.6	5.8	45.3
	AIME 2024 (Pass@1)	39.2	23.3	23.3	16
	MATH-500 (EM)	90.2	80	73.8	78.3
Math	CNMO 2024 (Pass@1)	43.2	15.9	6.8	13.1
	CLUEWSC (EM)	90.9	91.4	84.7	85.4
	C-Eval (EM)	86.5	86.1	61.5	76.7
	C-SimpleQA (Correct)	64.1	48.4	50.4	51.3
					59.3

✓ 高性能 ✓ 低成本 ✓ 高训练效率 ✓ 开源

- 垂域大模型，也被称为领域大模型或行业大模型，是相对于通用基础大模型而言，针对特定垂直领域或行业进行专门设计、训练和优化的大型人工智能模型。
- 设计原则可遵循三阶段优化路径：领域数据收集与标注、知识融合以及模型优化。

领域针对性强

数据专业性强

任务特定性强

知识深度强

01

领域数据收集与标注：从专业数据库、行业报告、企业内部数据等多渠道收集数据，制定严格的标注规范，确保数据标注准确性和一致性。

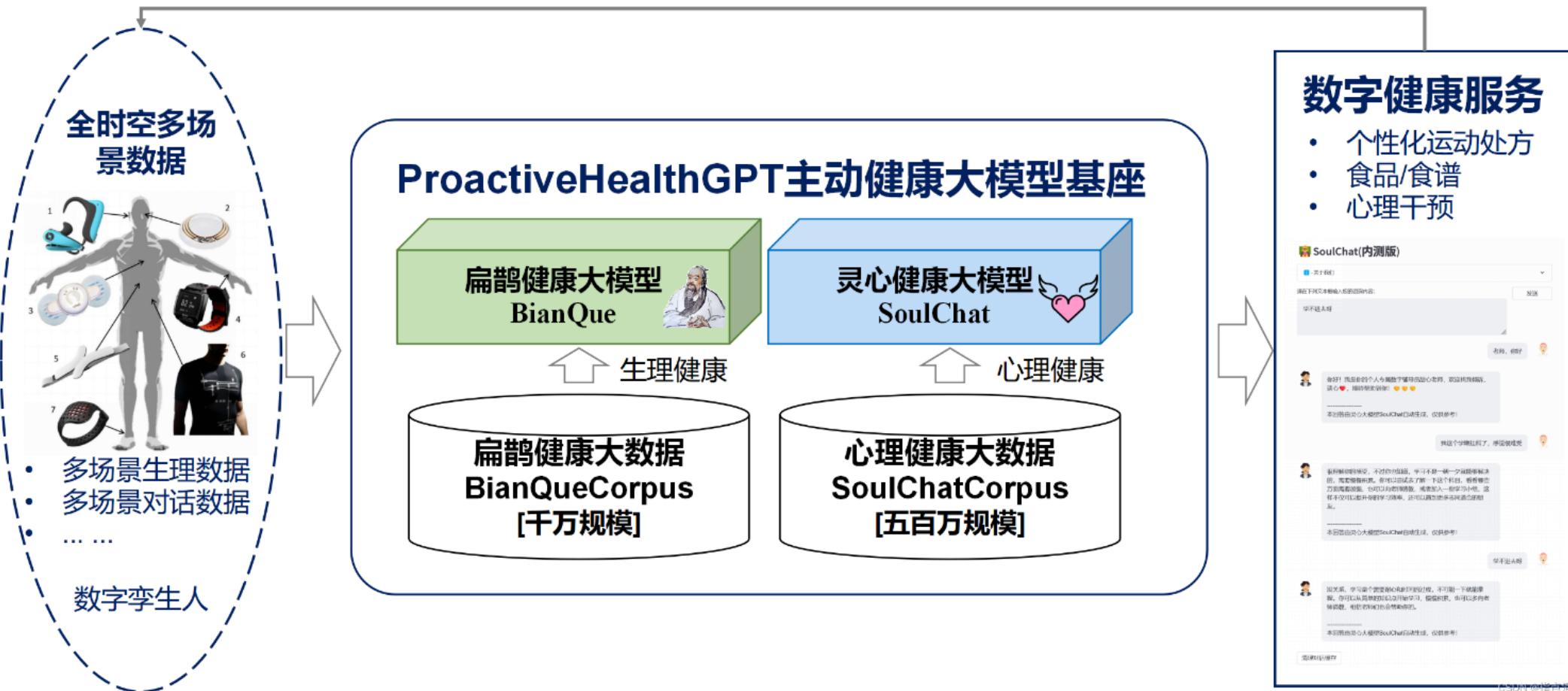
02

知识融合：构建领域知识图谱，将领域知识融入模型训练中，通过知识蒸馏、多模态融合等技术，提升模型对领域知识的理解和应用能力。

03

模型优化：根据领域数据特点和任务需求，对模型架构进行调整和优化，如增加特定领域的模块、调整网络层数和参数，以提高模型性能。

- **垂域大模型**，是相对于通用基础大模型而言，针对特定垂直领域或行业进行专门设计、训练和优化的大型人工智能模型。
- 设计原则可遵循三阶段优化路径：领域数据收集与标注、知识融合以及模型优化。

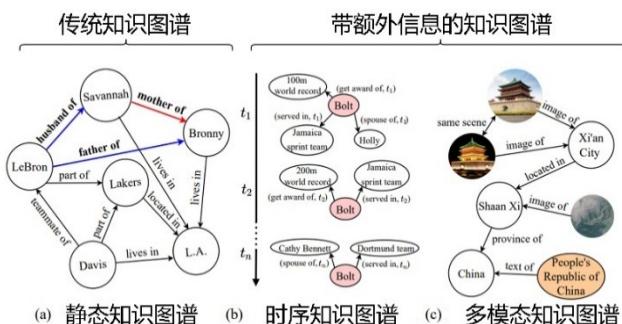


■ **领域知识深度融合技术**进一步推动垂域大模型发展。



知识图谱增强推理

方法：实体链接+图神经网络，
通过实体链接将文本实体与
知识图谱对应，再用图神经
网络推理实体间关系。



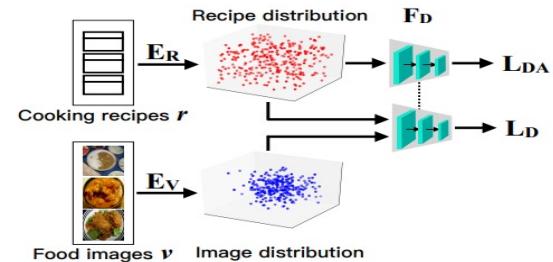
领域规则引导解码

方法：受限解码+规则模板，
依据领域规则限制解码过程，
确保生成内容符合规则。



多模态数据对齐

方法：跨模态对比学习，通过对比学习使不同模态数据特征对齐，提升模型对多模态数据理解。

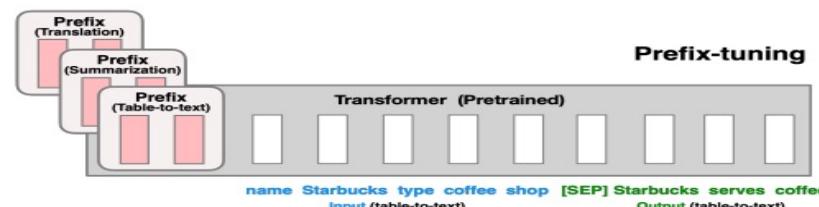
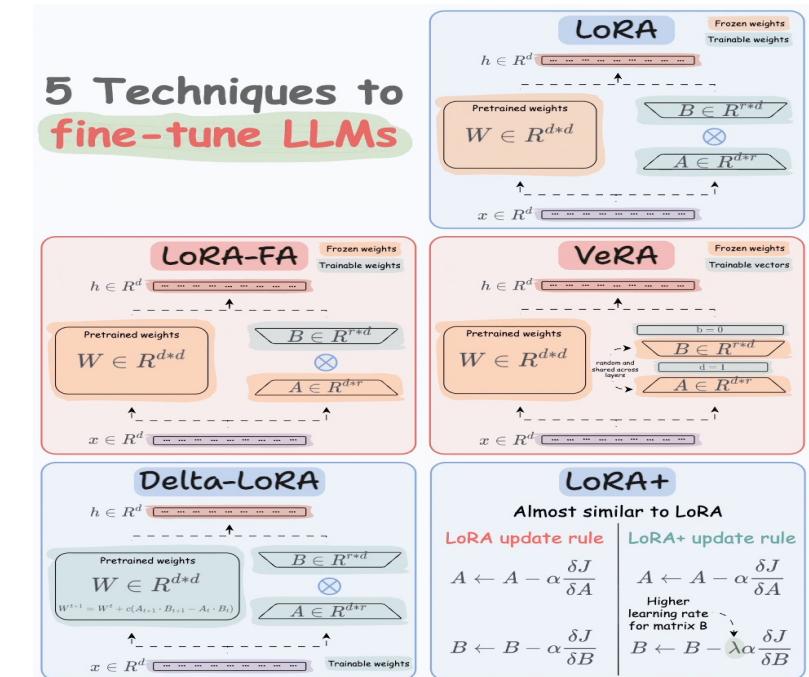


■ 参数高效微调技术是垂域大模型实现高效定制化的技术手段。

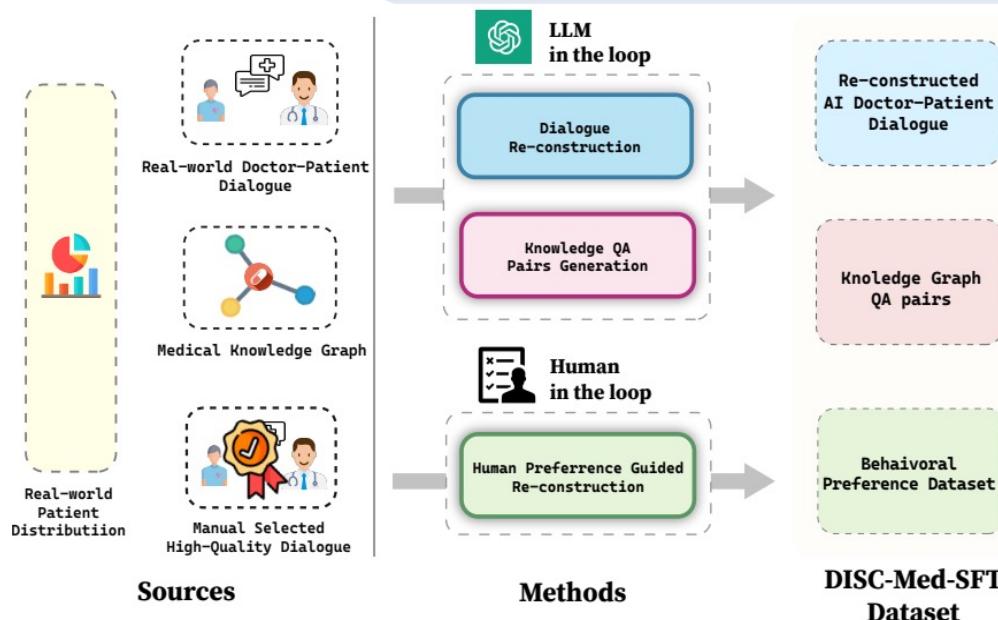
全量 Supervised Fine-tuning: 使用大量标注数据进行微调，使模型在特定领域任务上达到更好的性能。参数量100%，训练成本高，适用于数据充足、算力充裕场景，能充分优化模型性能。

LoRA: 通过低秩矩阵分解进行微调。在保持原模型性能的同时，适应特定领域的任务需求。参数量0.1-1%，训练成本极低，适用于低资源垂域适配，能在少量数据下快速优化模型。

Prefix-tuning: 通过调整模型的提示词，引导模型输出。参数量0.1%，训练成本中等，适用于生成式任务控制，可有效控制生成内容方向。



垂域大模型——你想要的专才！



医疗大模型DISC-MedLLM



安恒信息恒脑大模型

基于**Baichuan13BBase基础大模型指令微调**得到

DISC-Med-SFT训练集，包含超过**47万个来源于现有的多个医疗数据集的不同示例**。

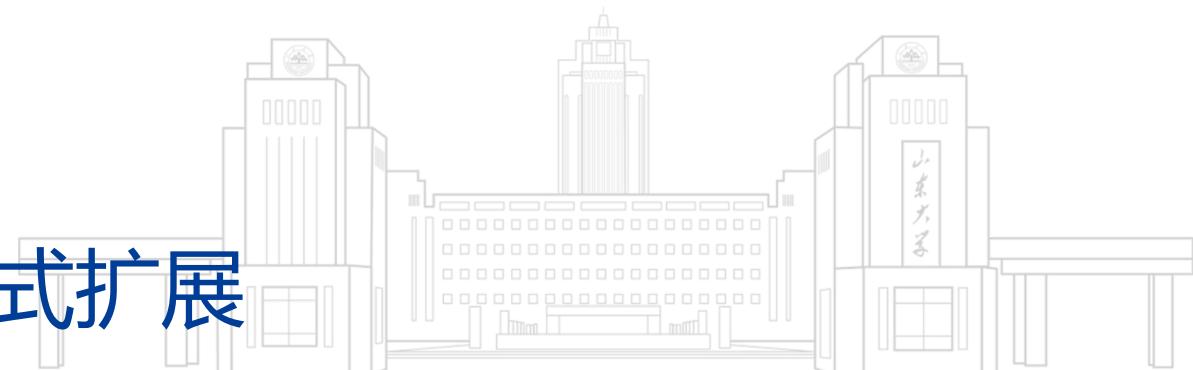
以**大模型为底座，以智能体为核心**，实现安全能力场
景化调度，

2024年唯一入选工信部“未来产业领域”案例名单
的安全垂域大模型

章节目录

CONTENTS

- 01 | 大模型发展与演练
- 02 | 大模型预训练范式
- 03 | 多模态大模型
- 04 | 基础大模型
- 05 | 大模型应用与范式扩展



大模型如何接入其他领域?

预训练阶段：模型在大规模通用数据集上学习基础知识和结构。

微调阶段：将预训练模型应用到特定任务的数据集上，通过再训练提升表现。

模型应用阶段：通过优化模型交互，使用**RAG技术**和Agent实现动态任务处理，并在各行业中应用。



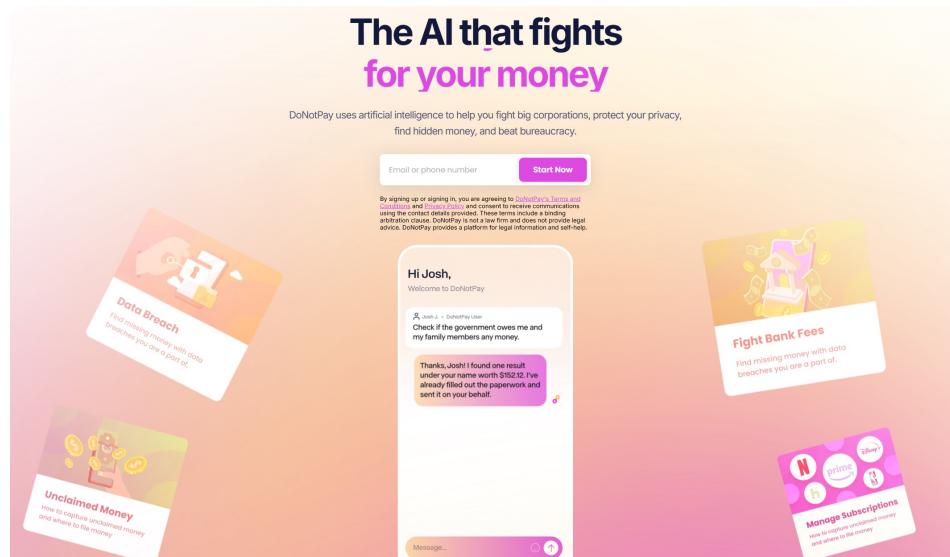
1. 法律

DoNotPay

DoNotPay使用GPT-4生成法律文件、提供法律建议，帮助用户解决小额纠纷、申诉罚单等。

Harvey AI

Harvey利用LLM为法律事务所提供法律研究、合同审查和文档生成服务，提升法律工作的效率。



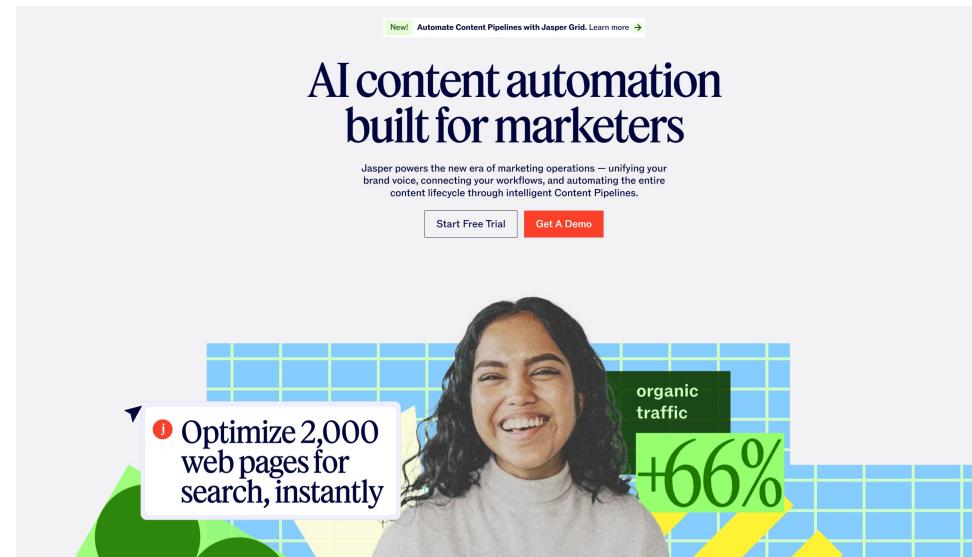
1. 文案生成与写作

Jasper AI

Jasper AI基于LLM提供内容生成服务，包括文章、广告文案、博客内容等，帮助创作者提高写作效率。

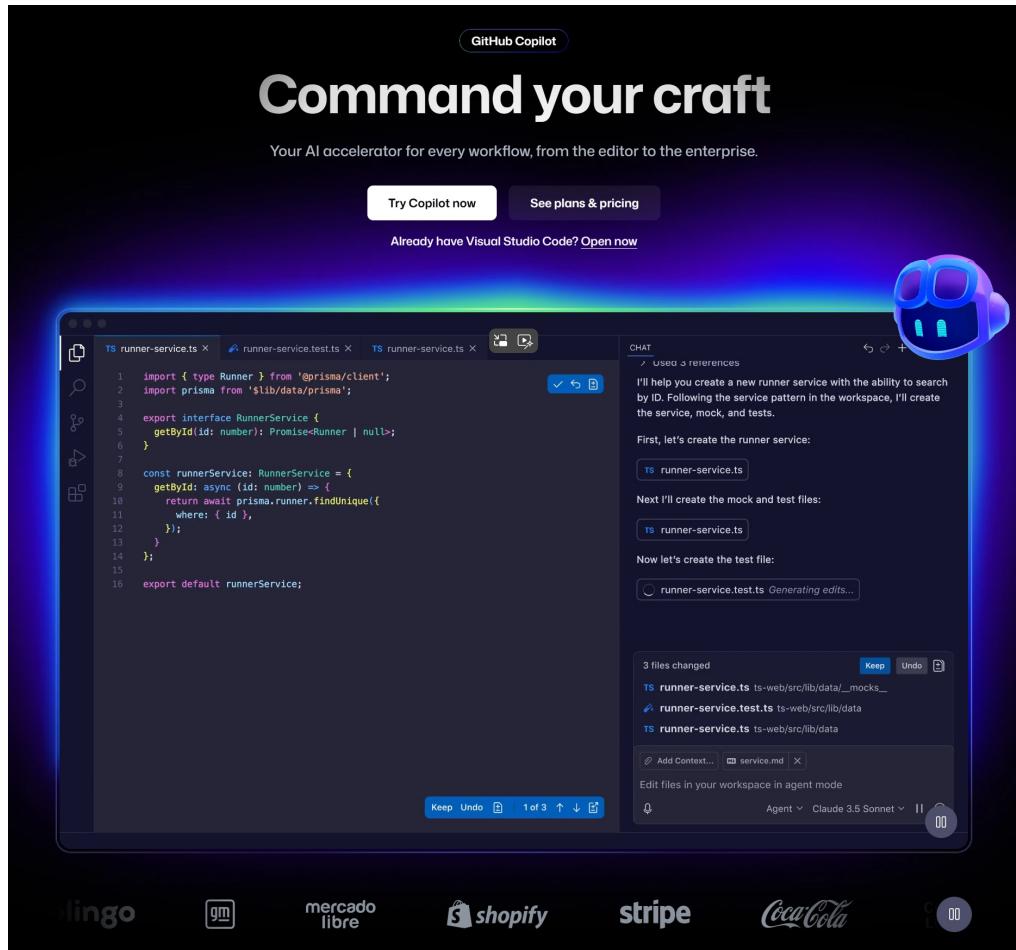
Copy.ai

Copy.ai 使用GPT-4生成市场营销文案、社交媒体帖子和电子邮件，帮助企业和个人更快速地创建内容。

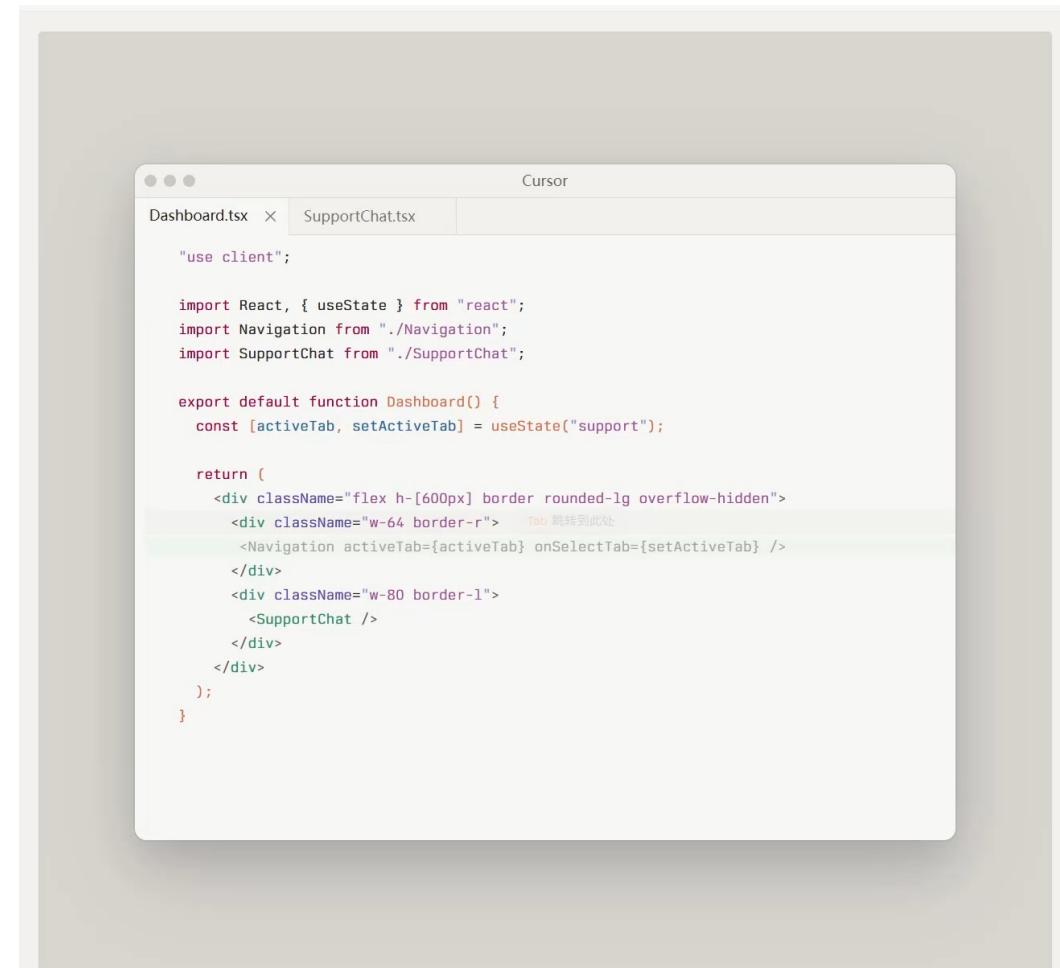


■ 利用 AI/自动化工具，提升开发效率的核心能力，开发全流程智能支持

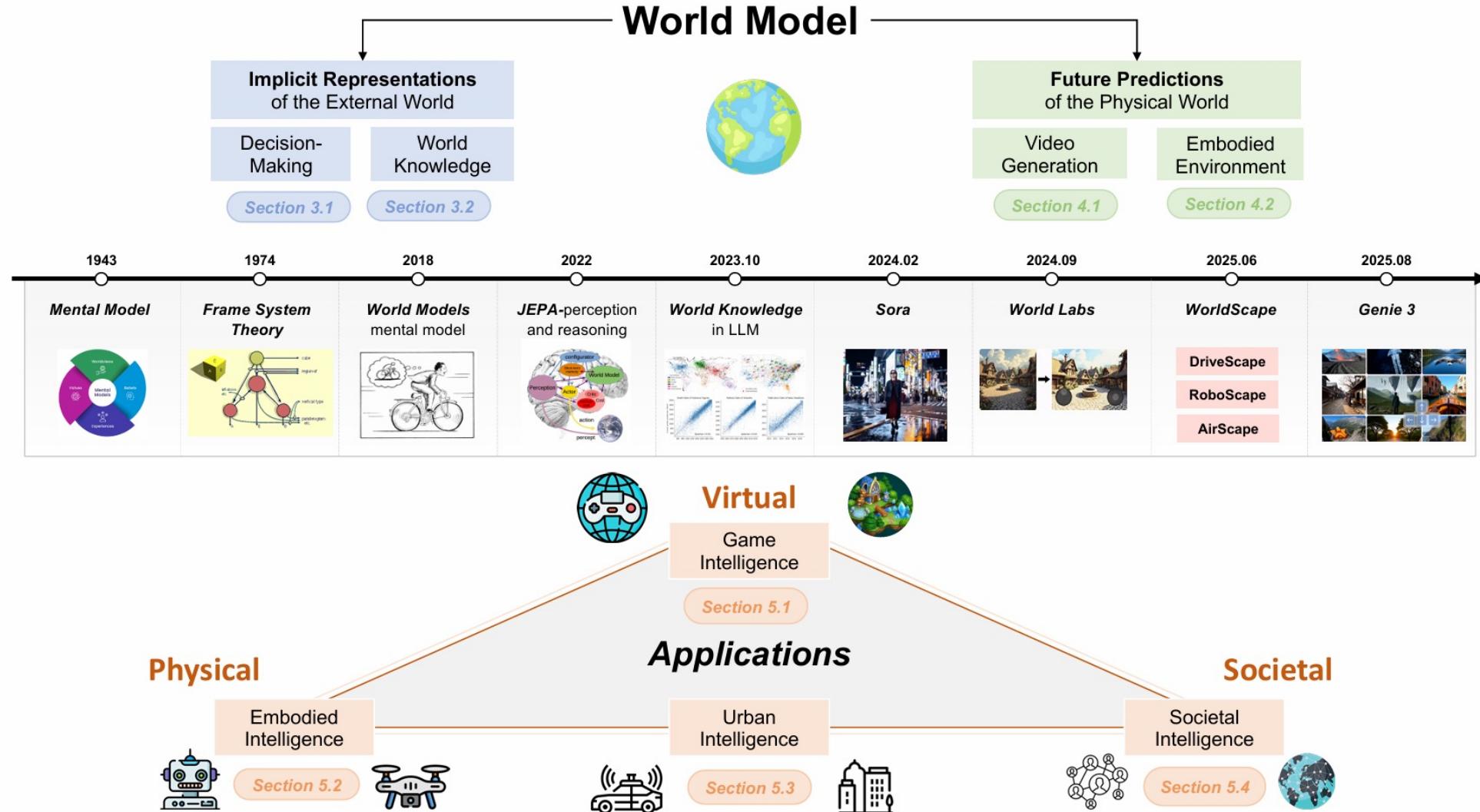
➤ **GitHub Copilot:** GitHub与OpenAI联合开发的行业先驱，基于GPT模型重构代码生成范式。



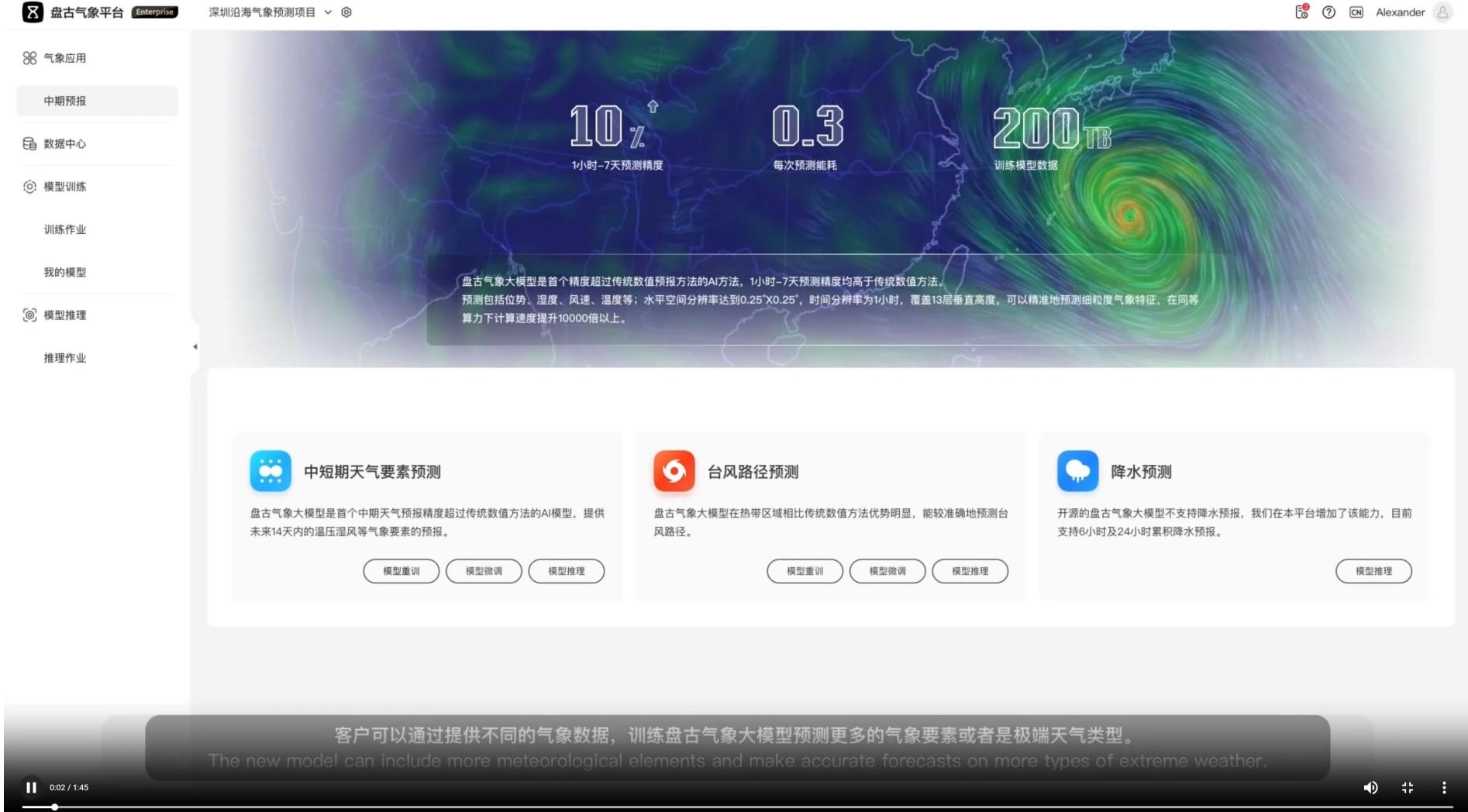
➤ **Cursor:** 旨在让您的工作效率达到非凡水平,是使用AI 编写代码的最佳方式。



■ 世界模型类似于一个“内部模拟器”，能够让智能体感知环境、预测其动态变化，并基于这些预测进行高效的规划和决策。



■ 盘古气象平台是基于盘古气象大模型提供训推一体化的平台，支持模型一键启动 模型训练、微调、部署和推理。



盘古气象平台 Enterprise 深圳沿海气象预测项目 Alexander

气象应用 中期预报 数据中心 模型训练 训练作业 我的模型 模型推理 推理作业

10% 1小时-7天预测精度

0.3 每次预测能耗

200TB 训练模型数据

盘古气象大模型是首个精度超过传统数值预报方法的AI方法，1小时-7天预测精度均高于传统数值方法。预测包括位势、湿度、风速、温度等；水平空间分辨率达到 $0.25^{\circ}\text{X}0.25^{\circ}$ ，时间分辨率为1小时，覆盖13层垂直高度，可以精准地预测细粒度气象特征，在同等算力下计算速度提升10000倍以上。

中短期天气要素预测 台风路径预测 降水预测

模型重训 模型微调 模型推理 模型重训 模型微调 模型推理 模型推理

客户可以通过提供不同的气象数据，训练盘古气象大模型预测更多的气象要素或者是极端天气类型。
The new model can include more meteorological elements and make accurate forecasts on more types of extreme weather.

0:02 / 1:45

■ 在全球科技革命与产业变革加速演进的当下，人工智能作为重塑全球竞争力格局的关键变量，正深刻影响着国家发展的战略主动权。

- **人才教育：**在人才培养模式创新方面，**建立“产学研用”协同机制**是关键。企业、高校和科研机构应紧密合作，建立“行业需求图谱-人才能力模型-课程动态调整”的联动机制；在人才引育上，通过政策激励和平台搭建吸引全球顶尖AI人才，依托协同创新平台打造“基础研究-技术转化-产业应用”全链条孵化生态，**培养兼具技术能力与行业洞察力的复合型人才**。
- **技术创新：**随着AI技术的持续迭代，亟需在全栈式AI基础设施上实现突破，通过开源开放策略激发创新动力，培育多元化的技术创新主体，从而加速技术创新进程。此举有望助力我国突破“技术孤岛”与“规则依附”的困境，构建**自主可控的人工智能技术体系**，并打造具有国际影响力的技术创新高地。
- **鼓励开源：****开源已经成为激活技术创新能力的关键动力**。鼓励行业领军企业和科研机构开放深度学习框架、预训练模型及开发工具链等关键技术，积极培养开源社区和开源生态，支持中小企业和开发者积极使用并贡献开源组件，实现技术普惠并加速创新扩散。