课程1：

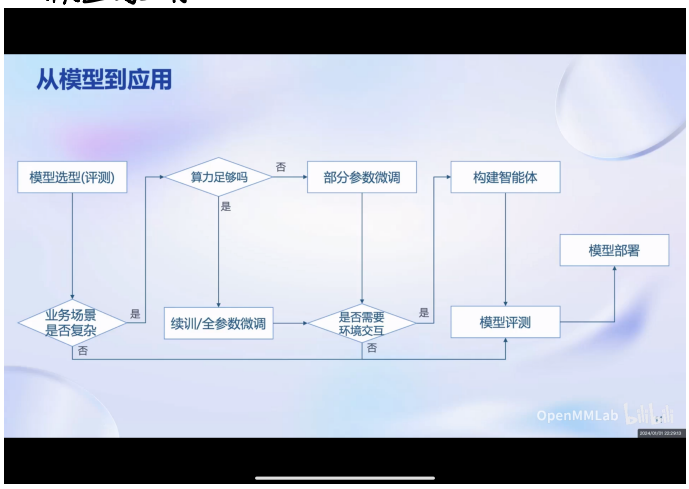
专用模型：针对特定问题，一个模型解决一个问题

⇓

通用模型：一个模型应对多种任务，多种模态。

书生·浦语大模型

从模型到应用：



书生浦语开源开放体系：

数据 → 预训练 → 微调 → 部署 → 评测 → 应用。

数据：文本数据、图像-文本数据、数据数据

多模态融合

精细化知识

价值观对齐

预训练：高扩展性、极致性能优化、兼容主流、开箱即用

微调：增量续训、有监督微调

评测：学科、语言、知识、理解、推理、安全

OpenCompass

大语言模型的局限性：最新信息和知识的获取、回复的可靠性、数学计算、工具使用和交互

轻量级智能框架 Lagent

多模态智能体工具箱 AgentLego


全链条开源开放体系 | 部署
大语言模型特点
内存开销巨大
庞大的参数量
采用自回归生成token，需要缓存k/v
动态Shape
请求数不固定
token逐个生成，且数量不定
模型结构相对简单
transformer 结构，大部分是decoder-only
技术挑战
设备
低存储设备（消费级显卡、移动端等）如何部署?
推理
如何加速 token 的生成速度
如何解决动态shape，让推理可以不间断
如何有效管理和利用内存
服务
提升系统整体吞吐量
降低请求的平均响应时间
部署方案
技术点
模型并行
低比特量化
Attention优化
计算和访存优化
Continuous Batching
OpenMMLab
25