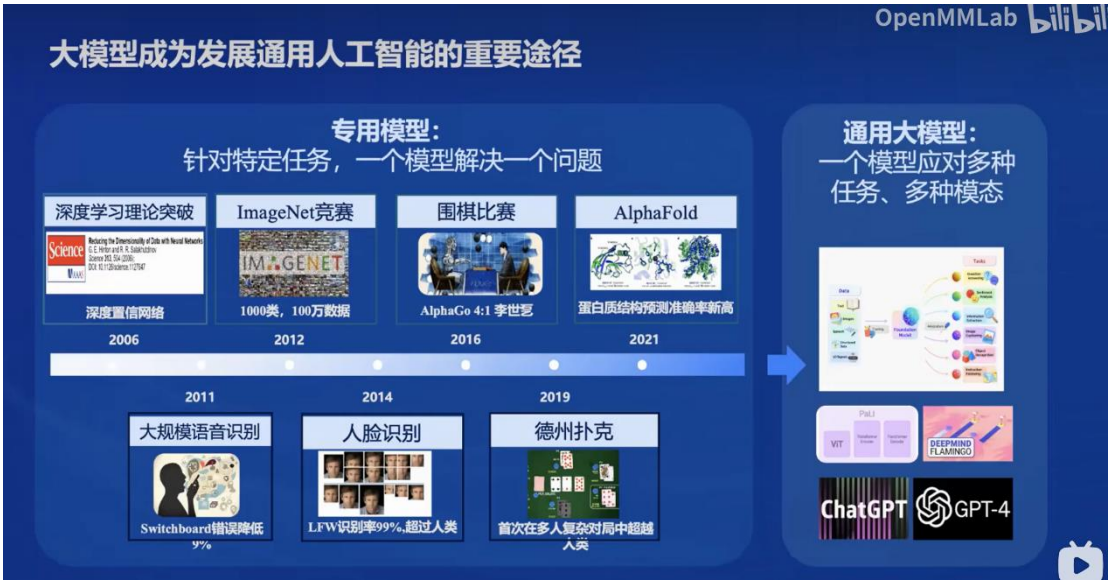


Lesson 1 笔记

第一节课了解了一个挺全面的大模型架构平台。这平台从大模型建设训练到智能体，还有多模态的智能体这些要害部分都覆盖了。有了这些开源的工具，开发者和科研工作者们开发和部署大模型就方便多了。



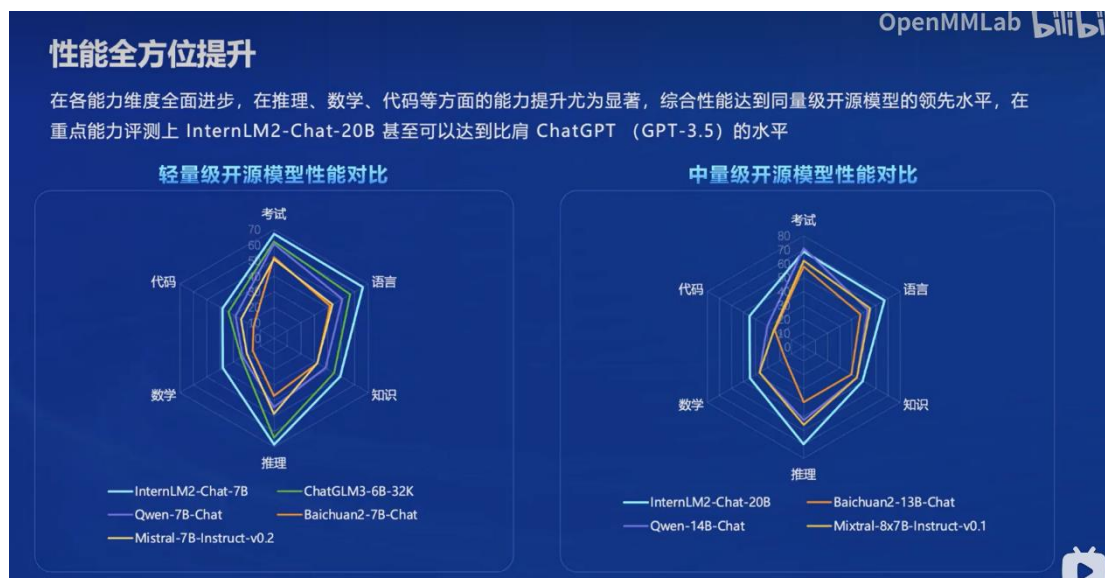
人工智能领域重要里程碑的时间线和分类，重点是自然语言处理及相关领域的成就。它分为两个主要部分：基础模型和应用模型。基础模型是向右侧的应用模型迈进的垫脚石。



这个时间线详细列出了从 2023 年至 2024 年间 InternLM 模型的几个主要版本发布和里程碑事件。



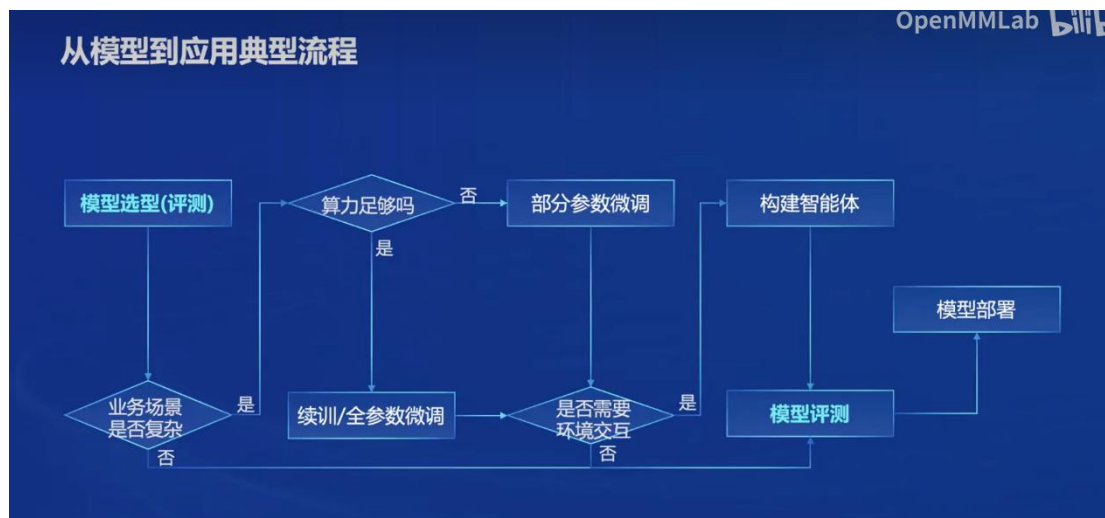
这张图片强调了 InternLM2 的多维度应用能力，包括提升了的图文理解、跨会话信息的处理、以及在对话创作、工具链集成、数学能力和数据分析方面的优势。这些特点使得 InternLM2 在多个领域都展现出竞争力。



InternLM2-chat-20B 这模型儿和 GPT3.5 差不多牛，在用多智能体时，GPT3.5 搭配上几个小伙伴一起干，能力还能超过 GPT4 呢。把 InternLM2-chat-20B 跟一个 Agent 工作流程搭一搭，也可能更好，令人期待。



智能体可以通过整合不同的工具来优化工作流程，例如可以执行代码生成、API 调用等多种操作。InternLM2 模型可以与多个不同的工具和服务结合使用，提高操作的灵活性和效率。



流程图描述了从信息接收（例如用户的输入）到信息处理和操作执行的整个流程。它涉及了多个步骤，包括理解输入、计划生成、任务执行和反馈生成。这个流程图可能是用来解释如何构建一个智能体的工作流程，以便它可以自动化地处理任务并生成合理的输出。



提供的工具和服务就是为了优化和简化大模型的训练、调参、部署和可视化过程。这些工具也可能被设计用来支持模型训练和智能体构建的不同方面，从而提高了工作效率和效果。

全链条开源开放体系 | 微调

大语言模型的下游应用中，增量续训和有监督微调是经常会用到两种方式。

增量续训

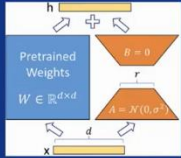
使用场景：让基座模型学习到一些新知识，如某个垂类领域知识
训练数据：文章、书籍、代码等

有监督微调

使用场景：让模型学会理解各种指令进行对话，或者注入少量领域知识
训练数据：高质量的对话、问答数据

全量参数微调

部分参数微调



提到了大模型训练的两个方面——模型微调和零样本学习。对于模型微调，强调了在有限数据环境下，可以通过调整预训练模型来适应特定任务，如文本、中医、代码等领域的应用。而零样本学习则是指模型可以直接在看到新任务时表现出良好的性能，无需额外的训练数据。

全链条开源开放体系 | 微调

高效微调框架 XTuner

 InternLM

 Llama

 Qwen

 BaiChuan

 ChatGLM

任务类型	数据格式	训练引擎	优化加速	支持算法
增量预训练 指令微调 工具类指令微调	Alpaca MOSS OpenAI Guanaco	 MVE	Flash Attention DeepSpeed ZeRO Pytorch FSDP	QLoRA 微调 LoRA 微调 全量参数微调

 NVIDIA

消费级显卡

GeForce RTX 2080, 2080Ti
GeForce RTX 3060 ~ 3090Ti
GeForce RTX 4060 ~ 4090

数据中心

Tesla T4, V100
A10, A100, H100

适配多种生态

- 多种微调算法
多种微调策略与算法，覆盖各类 SFT 场景
- 适配多种开源生态
支持加载 HuggingFace、ModelScope 模型或数据集
- 自动优化加速
开发者无需关注复杂的显存优化与计算加速细节

适配多种硬件

- 训练方案覆盖 NVIDIA 20 系以上所有显卡
- 最低只需 8GB 显存即可微调 7B 模型

介绍了调参工具 XTuner，列举了它可以支持的不同的模型和框架，比如 InternLM、LlaMa、Qwen 等，还有多种硬件平台的支持，例如 NVIDIA 的不同型号 GPU。这张图片还提到了 XTuner 能够提供的一些特点，包括兼容性、调参速度和效率，还有与其他平台如 HuggingFace 和 ModelScope 的集成，以及支持最新的 NVIDIA 20 系列显卡，明显是为了突出它的高性能和广泛的适用性。

OpenMMLab 

全链条开源开放体系 | 部署

LMDeploy

LMDeploy 提供大模型在GPU上部署的全流程解决方案，包括模型轻量化、推理和服务。

接口

- Python
- gRPC
- RESTful

轻量化

- 4bit权重
- 8bit k/v

推理引擎

- turbomind
- pytorch

服务

- openai-server
- gradio
- triton inference server

高效推理引擎

- 持续批处理技巧
- 深度优化的低比特计算 kernels
- 模型并行
- 高效的k/v缓存管理机制

完备易用的工具链

- 量化、推理、服务全流程
- 无缝对接OpenCompass评测推理精度
- 多维度推理速度评测工具

支持交互式推理，不为历史对话买单

- 非交互式
- 交互式

U1	A1				
U1	A1	U2	A2		
U1	A1	U2	A2	U3	A3

U1	A1
U2	A2
U3	A3

LMDeploy 是一种专门用于 GPU 上部署大型机器学习模型的工具，它支持多种编程语言和服务接口。

OpenMMLab 

全链条开源开放体系 | 智能体

轻量级智能体框架 Lagent

支持多种类型的智能体能力

ReAct

- 输入
- 选择工具
- 执行工具
- 结束条件
- 结束

ReWoo

- 输入
- 计划拆分
- DAG
- 计划执行
- 结束

AutoGPT

- 输入
- 选择工具
- 人工干预
- 执行工具
- 结束条件
- 结束

灵活支持多种大语言模型

 GPT-3.5/4

 InternLM

 Hugging Face Transformers

 Llama

简单易拓展，支持丰富的工具

AI 工具	能力拓展	Rapid API
文生图	搜索	出行 API
文生语音	计算器	财经 API
图片描述	代码解释器	体育资讯 API

概述了智能体开发的几个组件和它们的交互：

ReAct、ReWoo 和 AutoGPT 代表不同的智能体技术。

它们之间的交互包括初始化、状态更新、动作决策和循环迭代等步骤。

这些智能体还与 DAG（有向无环图）等技术结合，以优化决策和执行路径。

此外，也展示了与 GPT-3.5/4、InternLM、Hugging Face Transformers 和 Llama 等不同的大模型和框架的兼容性。

多模态智能体工具箱 AgentLego

- 丰富的工具集合，尤其是提供了大量视觉、多模态相关领域的前沿算法功能
- 支持多个主流智能体系统，如 LangChain, Transformers Agent, lagent 等
- 灵活的多模态工具调用接口，可以轻松支持各类输入输出格式的工具函数
- 一键式远程工具部署，轻松使用和调试大模型智能体



强调了多智能体系统的组合能力，能够整合 LangChain、Transformers Agent 和 lagent 等不同技术。

介绍了它能够提供的可扩展性和灵活性，以及对各种工具的搜索和服务功能。

还展示了 Hugging Face、OpenMMLab、OSAM 和 Stable Diffusion 等不同平台或技术的整合，指出了 AgentLego 在这个生态系统中的中心作用。