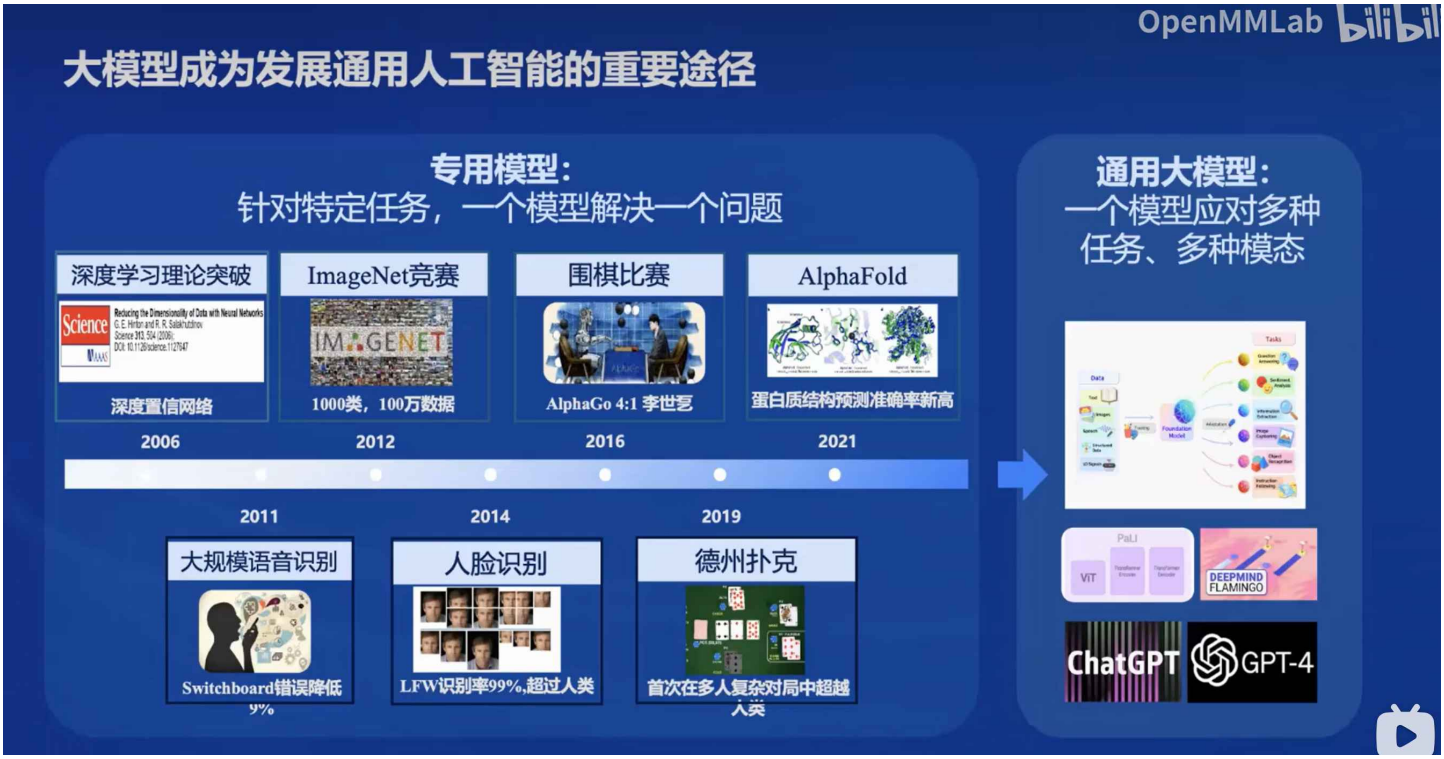
书生二期第一次课笔记

通过本次课程的学习了解到书生提供了一个当前比较全的大模型架构平台。包括大模型的构建与训练，大模型智能体，多模态的大模型智能体。这些基础的开源平台使得更多人可以借助其做大模型的开发与部署应用工作。



书生·浦语 2.0（InternLM2）的体系

面向不同的使用需求，每个规格包含三个模型版本

7B

为轻量级的研究和应用提供了一个轻便但性能不俗的模型

20B

模型的综合性能更为强劲，可有效支持更加复杂的实用场景

InternLM2-Base

高质量和具有很强可塑性的模型基座，是模型进行深度领域适配的高质量起点

InternLM2

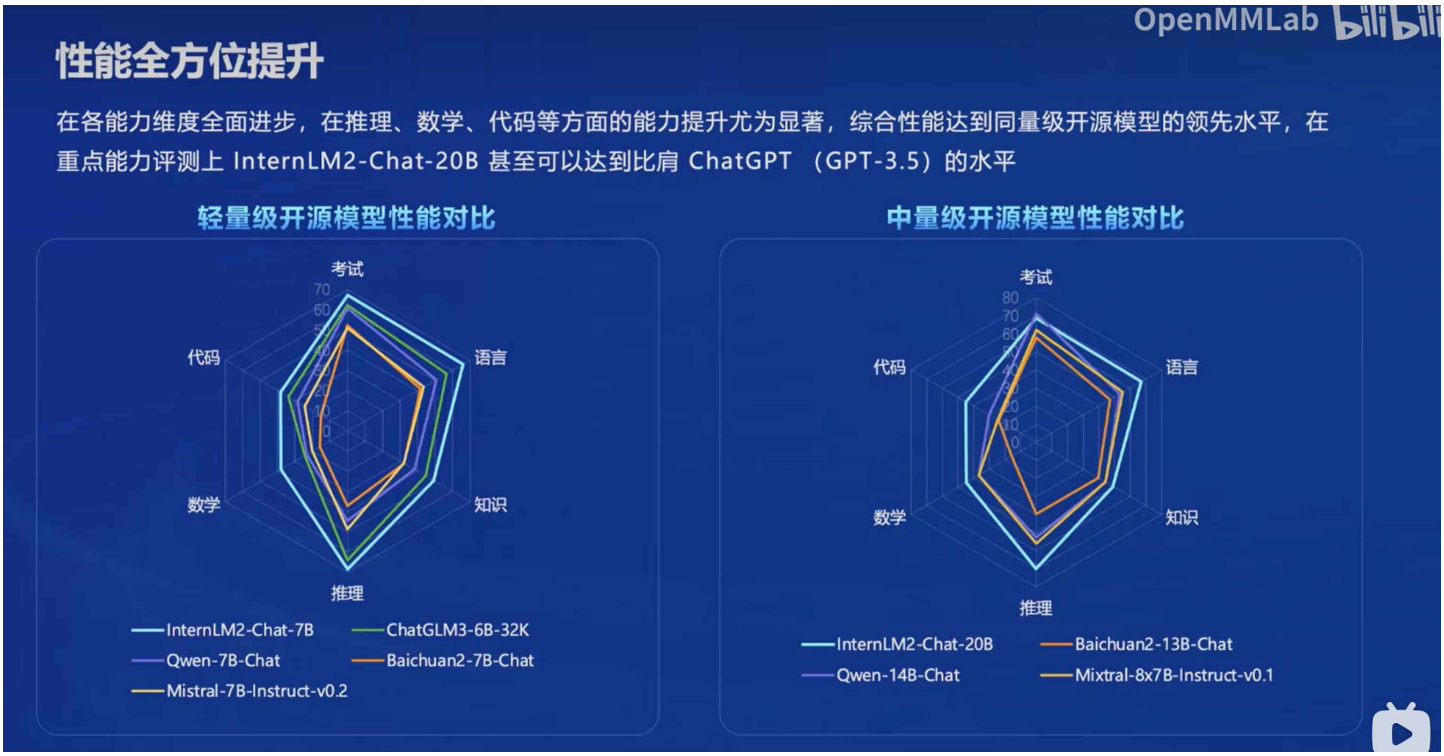
在 Base 基础上，在多个能力方向进行了强化，在评测中成绩优异，同时保持了很好的通用语言能力，是我们推荐的在大部分应用中考虑选用的优秀基座

InternLM2-Chat

在 Base 基础上，经过 SFT 和 RLHF，面向对话交互进行了优化，具有很好的指令遵循、共情聊天和调用工具等的能力

书生·浦语 2.0（InternLM2）的主要亮点

超长上下文	综合性能全面提升	优秀的对话和创作体验	工具调用能力整体升级	突出的数理能力和实用的数据分析功能
模型在 20万 token上下文中，几乎完美实现“大海捞针”	推理、数学、代码提升显著InternLM2-Chat-20B 在重点评测上比肩 ChatGPT	精准指令跟随，丰富的结构化创作，在 AlpacaEval2 超越 GPT-3.5 和 Gemini Pro	可靠支持工具多轮调用，复杂智能体搭建	强大的内生计算能力，加入代码解释后，在 GSM8K 和 MATH 达到和 GPT-4 相仿水平



internLM2-chat-20B有GPT3.5的能力，在多智能体链路中，GPT3.5的多智能体组合可以超过gpt4的能力，所以可以尝试一个internLM2-chat-20B组合一个Agent工作流程的水平如何。

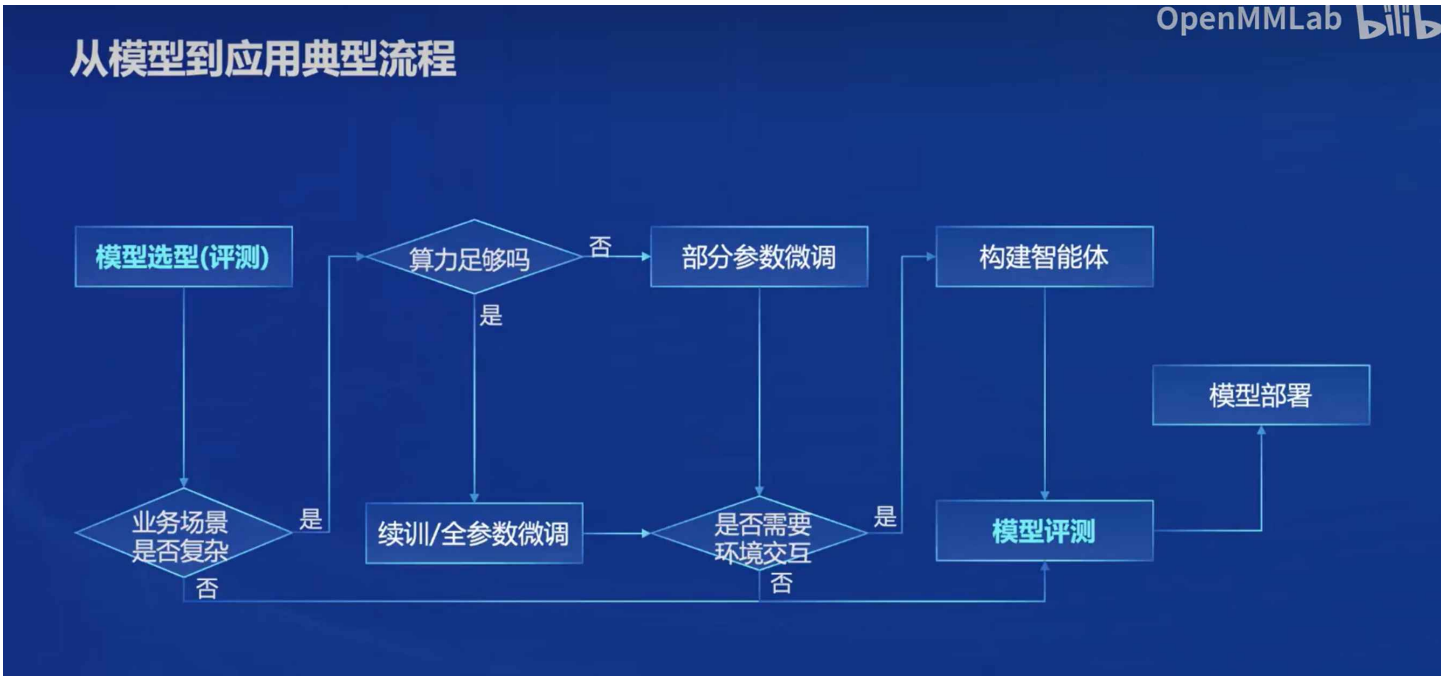
1）Reflection：让 Agent 审视和修正自己生成的输出；

2）Tool Use：LLM 生成代码、调用 API 等进行实际操作；

3）Planning：让 Agent 分解复杂任务并按计划执行；

4）Multiagent Collaboration：多个 Agent 扮演不同角色合作完成任务；

单纯的比较gpt3.5与gpt4的编码能力，GPT-4 做得更好，正确率达到了 67.7%，但如果你围绕 GPT-3.5 使用一个 Agent 工作流程，实际上它的表现甚至比 GPT-4 还要好。如果你将这种类型的工作流程应用于 GPT-4，它也表现得非常好。你会注意到，GPT-3.5 与一个 Agent 性工作流程相结合实际上超过了 GPT-4 的表现。



以上的思路表明，如果智能体在复杂的场景表现的不够好，要么是RAG或者是微调，但是我们如上面的笔记，可以尝试构建一个智能体的工作流，也许可以更好的提高多个智能体组成的系统整体的效果，比微调更好，或者微调后建立这样的工作流效果又能够得到一个提升。

OpenMMLab

书生·浦语全链条开源开放体系

数据
书生·万卷

2TB数据，
涵盖多种模态与任务

预训练
InternLM-Train

并行训练，极致优化
速度达到 **3600 tokens/sec/gpu**

微调
XTuner

支持 **全参数微调**，
支持**LoRA等低成本微调**

部署
LMDeploy

全链路部署，性能领先
每秒生成 **2000+ tokens**

评测
OpenCompass

全方位评测，性能可复现
100 套评测集，**50** 万道题目

应用
Lagent
AgentLego

支持多种智能体，支持**代码解释器**等多种工具

OpenMMLab

全链条开源开放体系 | 开放高质量语料数据

数据集获取： OpenDataLab https://opendatalab.org.cn/

书生· 万卷 1.0

总数据量：2TB
发布日期：2023年8月14日
符合主流中国价值观的中文语料

数据构成

文本数据	图像-文本数据集	视频数据
• 5亿个文档 • 数据量超 1TB	• 超2,200万个文件 • 数据量超140GB	• 超2,200万个文件 • 数据量超140GB

多模态融合
万卷包含文本、图像和视频等多模态数据，涵盖科技、文学、媒体、教育和法律等多个领域。该数据集对模型的知识内容、逻辑推理和泛化能力的提升有显著效果。

精细化处理
万卷经过语言筛选、文本提取、格式标准化、数据过滤和清洗（基于规则和模型）、多尺度去重和数据质量评估等精细数据处理环节，能够很好地适应后续模型训练的要求。

价值观对齐
在万卷的构建过程中，研究人员注重将数据内容与主流中国价值观进行对齐，并通过算法和人工评估的结合提高语料库的纯净度。

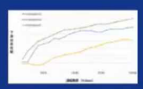
书生· 万卷 CC

总数据量：400GB
发布日期：2024年3月6日
安全、信息密度要高的英文语料

三大优势

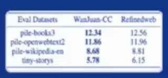
时间跨度长	来源丰富多样	安全密度高
横跨了 2013-2023 年互联网公开内容	从 90 个 dumps 的 1300 亿份原始数据中“萃取” 1.38% 内容	唯一在毒性、色情和个人隐私都进行了安全加固处理

高质量语料驱动效率性能双提升
万卷CC 作为 InternLM2 预训练语料，在不同数据规模上取得的任务性能分布显示，大幅提升模型训练效率



四重“萃取”高质量数据
通过原创技术，对原始数据进行多阶段处理，得到了高信息密度的万卷CC

数据质量高、模型更可靠
由万卷CC作为训练数据的模型在多项验证中取得了更优效果



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。

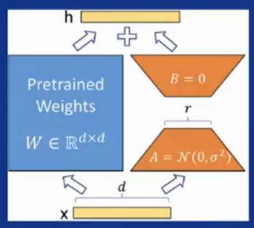
增量续训

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

全量参数微调　部分参数微调

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



如何保证增量续训不会遗忘之前的知识？这里有怎样的机制？

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 Guanacao ...	MMEngine	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**

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

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

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

全链条开源开放体系 | 智能体

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