书生·浦语大模型全链路开源体系

书生浦语2.0的体系包含三个模型版本

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，在测评中成绩优异，同时保持了很好的通用语言能力，是推荐的大部分应用中考虑选用的基座

InterLM2-Chat：在 `Base` 的基础上，经过 `SFT` 和 `RLHF` ，面向对话交互进行优化，具有很好的指令遵循、公顷聊天和调用工具等能力。

回归语言建模的本质

多维度数据价值评估，基于文本质量、信息质量、信息密度等维度对数据价值进行综合评估

书生浦语2.0的主要亮点

超长上下文：模型在 `20` 万的 `token` 上下文中，几乎完美实现”大海捞针“

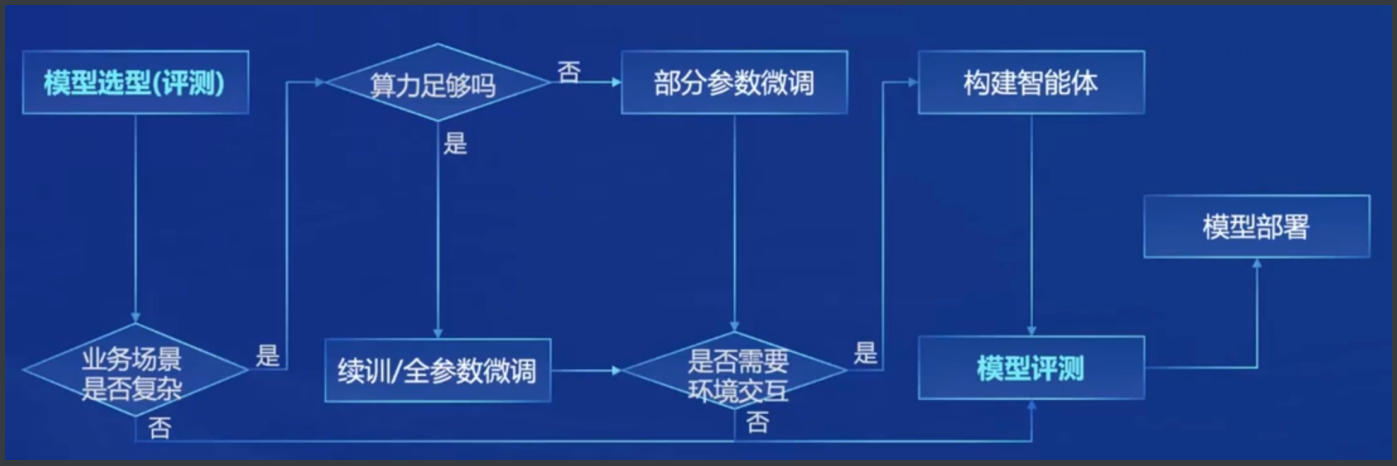
综合性能全面提升：推理、数学、代码提升显著

优秀的对话和创作体验：精准指令跟随、丰富的结构化创作，在 `AlpacaEval2` 超越 `GPT-3.5` 和 `Gemini Pro`

工具调用能力整体升级：可靠支持工具多轮调用，复杂智能体搭建.

突出的数理能力和使用的数据分析

从模型到应用的流程



书生浦语全链条开源开放体系

数据：2TB数据，涵盖多种模态与任务

预训练：InterLM-Train，并行训练，极致优化

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，每秒生成2000+ tokens

评测：OpenCompass，性能可浮现100套评测级，50w的题目

应用：LagentAgentLego，支持多种智能体，支持代码解释器多种工具

开放高质量语料数据

书生万卷1.0总数据量有2TB，符合中国价值观的中文语料；数据构成主要包括文本数据5个亿文档，数据量超1T；图像-文本数据集，超2200w个文件，数据量超140G；视频数据：超2200w个文件，数据量超140GB；

书生万卷CC总数据量有400GB，时间跨度长，横跨2013-2023年互联网公开内容；来源丰富，从90个dumps的1300亿原始数据中萃取1.38%内容；安全密度高：唯一在毒性、色情和个人隐私都进行安全加固处理。

预训练

高可扩展：支持从8卡到千卡训练，千卡加速效率达92%

极致性能优化：hybrid zero 独特技术 + 极致优化，加速50%

兼容主流：支持HuggingFace等技术生态，支持各类轻量化技术。

开箱即用：支持多种规格语言模型，修改配置即可训练

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用到的两种方式。

增量续训

使用场景：让基座模型学习到一些新的知识，比如某个垂直领域知识；

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：高质量对话、问答数据

Xtuner 支持多种微调算法、使用多种开源生态、自动优化加速；训练方案覆盖NVIDIA 20 系以上所有显卡，最低下只需8G显存就可以微调7B模型。

评测

CompassRank：性能榜单

CompassKit：大模型评测全站工具链

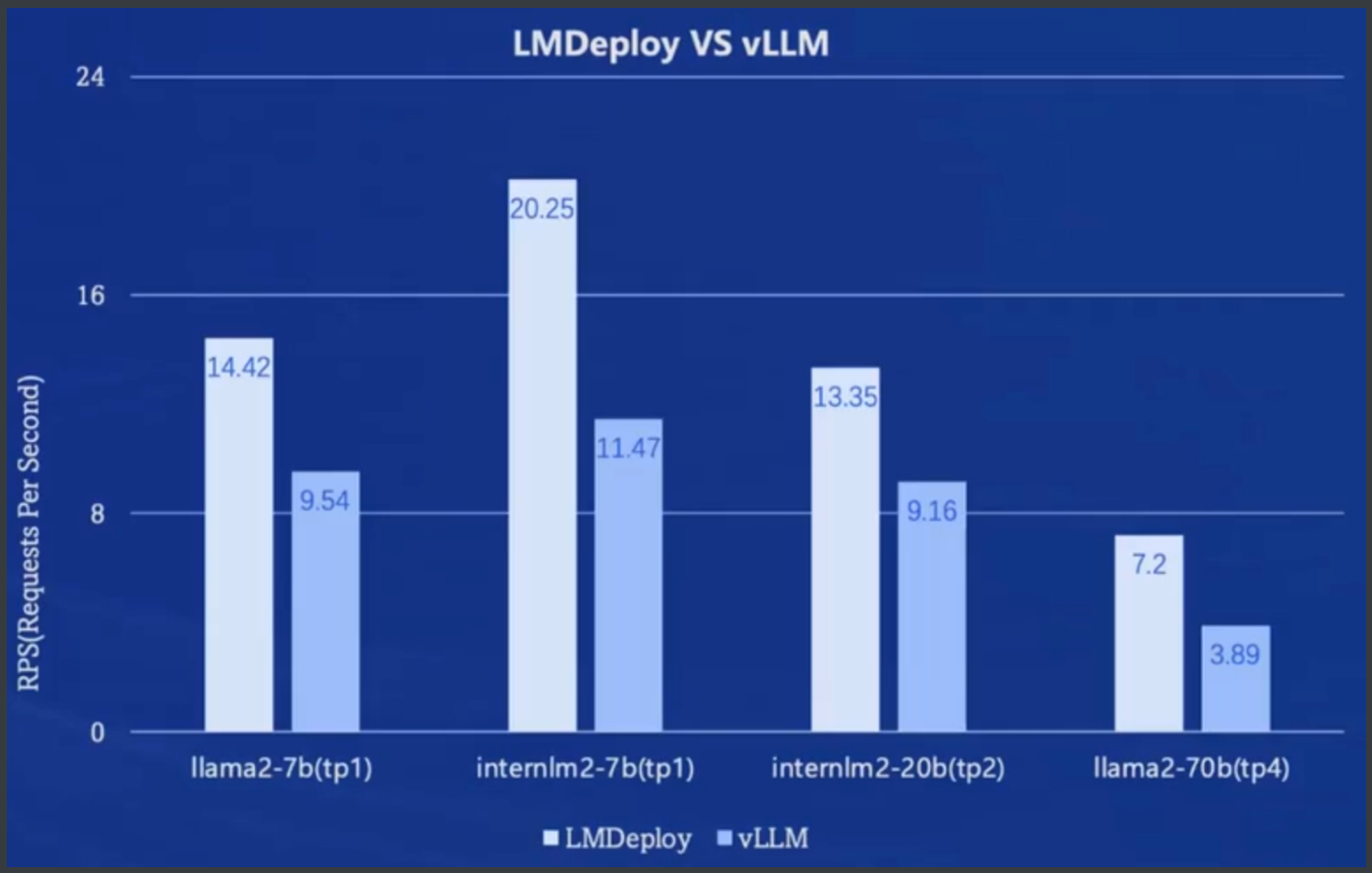
包含数据污染检查，模型推理借入、长文本能力评测、中英文双语主观评测

CompassHub：评测基准社区

OpenCompass：获得Meta官方推荐唯一国产大模型评测体系

部署

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



智能体

轻量级智能体框架Lagent

多模态智能体工具箱 `AgentLego`

总结

整体听下来，让大模型能够简单化，快速入门；另外从数据预处理到微调在到评测全链路有了一个新的体系认知，希望后面这些工具能够更加的简单化，去编程化，人人均能够快速训练自己垂直领域的模型。