书生·浦语第二期实战营笔记

6 班 66 号学员 Jack

第一节 书生·浦语大模型全链路开源体系

主讲人：陈凯老师 上海人工智能实验室 青年科学家

一、大模型成为发展通用人工智能的重要途径

1．2006 年以来专用模型从理论到语音、图象识别、人脸识别、围棋比赛、德州扑克、蛋白质结构预测等领域都取得了重要的进展。

2．ChatGPT 的诞生开启了通用大模型的元年，一个模型对应多种任务、多种模态。

3．延伸概念“模态”：模态的概念有不同的含义和应用领域：在物理学和工程学中，模态指的是结构系统的固有振动特性，每个模态都对应特定的固有频率、阻尼比和模态振型。模态分析是一种计算或试验分析过程，用于确定机械结构在特定频率范围内的振动特性。在计算机科学和软件设计领域，模态通常指的是一种临时的工作状态或模式，通常需要用户执行特定动作来退出。例如，在用户界面设计中，模态窗口是一种对话框或弹出窗口，用于在用户与主要程序交互的同时，专注于完成特定操作。语言学和社会符号学中，模态是指物质媒体经过长时间社会塑造而形成的意义潜能，或用于表征和交流意义的社会文化资源。在人工智能领域中，模态可能指机器对外界信息的感知模式或信息通道（包括文字、图片、视频等等）。因此，模态的具体含义取决于其应用的具体领域和背景。

二、书生·浦语大模型开源历程

1．2023 年 6 月 7 日 InternLM 发布；

2．2023 年 7 月 6 日全面升级并开源了 InternLM-7B，并免费商用，建立了全链条开源工具体系；

3．接下来就是书生万卷 1.0 多模态预训练语料库开源发布，InternLM-Chat-7B v1.1 发布开源智能体框架 Lagent 发布，参数升级到 123B，增强版 InternLM-20B 开源，开源工具链升级；2024 年 1 月 17 日 InternLM-2 开源。

三、书生·浦语 2.0（InternLM2）的体系

1．轻量级 7B，综合复杂场景 20B 两种规格构成；

2．每种规格分 3 个版本：Base 版（可进一步微调）、InternLM2 及 InternLM2-Chat。

四、回归语言建模的本质，通过新一代数据清洗及过滤技术进行：

1．多维度（文本质量、信息质量、信息密度等）数据价值评估；

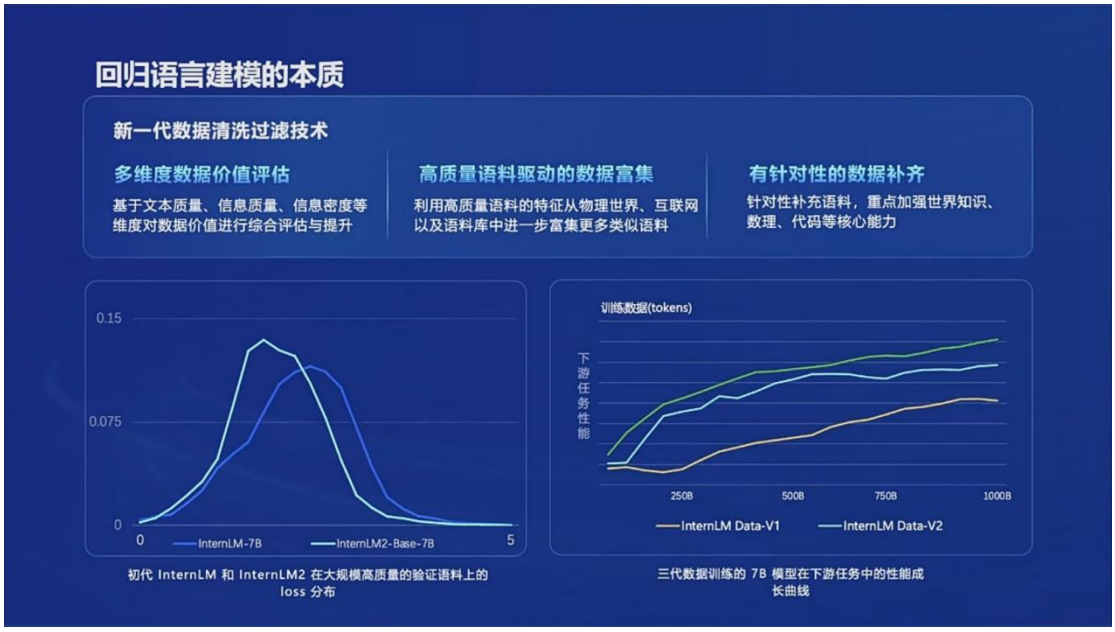
2．高质量语料（源自物理世界、互联网、即有语料库）驱动的数据富集；

3．有针对的数据补齐，加强世界知识、数理、代码等核心能力。

4．InternLM 初代和 2 代 7B 产品 Loss 分布对比。

（1）延伸概念：损失值（Loss Value）是衡量模型预测结果与真实标签之间差异的指标，用于评估模型在训练过程中的性能。低损失值表示模型预测与真实值接近，拟合程度好，通常意味着模型性能较好；而高损失值则表示模型预测与真实值差异较大，拟合程度差，可能存在欠拟合或过拟合问题。损失值的高低反映了模型在当前参数下对训练数据的拟合程度。对于不同的任务和数据集，合适的损失值阈值会有所不同，因此评价损失值是否“好”需要结合模型的性能表现、验证集或测试集上的指标以及实际应用场景进行综合考量。

（2）疑问：对比图中横坐标 0-5 代表什么意思？



5．三代数据训练结果明显优于 1 代和 2 代数据，印证了有针对性的数据补齐是可以提高模型性能的。

五、InternLM2 的主要亮点

1．支持 20 万 Token 上下文；

2．InternLM2-Chat-20B 版本推理能力在重点评测上比肩 ChatGPT3.5；

3．精准指令跟随、结构化创作在 AlpacaEval2 超越 GPT-3.5 和 Gemini Pro；

（1）延伸概念：AlpacaEval 是一种用于评估大语言模型性能的指标，它被设计用于对抗生成式预训练（GPT）模型的弱点，并更全面地反映模型的真实能力。这一指标由 OpenAI 提出，旨在解决以往评估指标的一些局限性，特别是在理解和生成多样性方面存在的问题。AlpacaEval 的名称中，“Alpaca” 代表一种动物，是一种灵活而多才多艺的动物，象征着模型在不同任务和语境中的灵活性和多功能性。而“Eval” 则是 evaluation 的缩写，表示评估。因此，AlpacaEval 旨在通过更全面的评估，更准确地捕捉模型的综合表现。AlpacaEval 主要关注以下几个方面：多样性（Diversity）： 衡量模型生成文本的多样性，避免单一或刻板的输出。多样性是指模型在生成不同样本时的差异程度。在实际应用中，我们期望模型不仅能够生成准确的内容，还能够呈现出多样的表达方式，以适应不同场景和需求。一致性（Consistency）： 评估模型在处理相似输入时生成的输出是否一致。一致性是指当模型面对相似的问题或请求时，其回应应该是稳定和一致的。这有助于确保模型在类似场景下能够提供可靠的结果。相关性（Relevance）： 衡量生成文本与输入之间的语义相关性。相关性

是指模型生成的文本是否与给定的输入有明确的关联，以及是否符合预期的语境。这有助于确保模型的输出在语境上是合理的，而不是简单地生成无关或荒谬的内容。

4．工具调用能力整体升级（通过 Latent Web Demo 展示）；可以更可靠地支持复杂智能体的搭建，支持对工具进行有效的多轮调用，完成较复杂的任务。

5．计算能力往往是大模型的短板，造成最终推理结果的错误。InternLM2 内生计算能力增强，配合代码解释器，在 GSM8K 和 MATH 数据集达到和 GPT4 相仿的水平。配合代码解释器，20B 版本已经能够完成积分求解等大学数学题目。

六、性能全方位提升

1．在能力维度上与 40 多个大模型进行了对比，能力维度包括考试、语言、知识、推理、数学、代码。

2．实用数据分析功能，如读取 Excel 表，画柱状图、进行线性预测等等。

七、AI 助手举例

1．规划旅游行程攻略；

2．充满人文关怀的对话；

3．流浪地球 3 剧本创作；

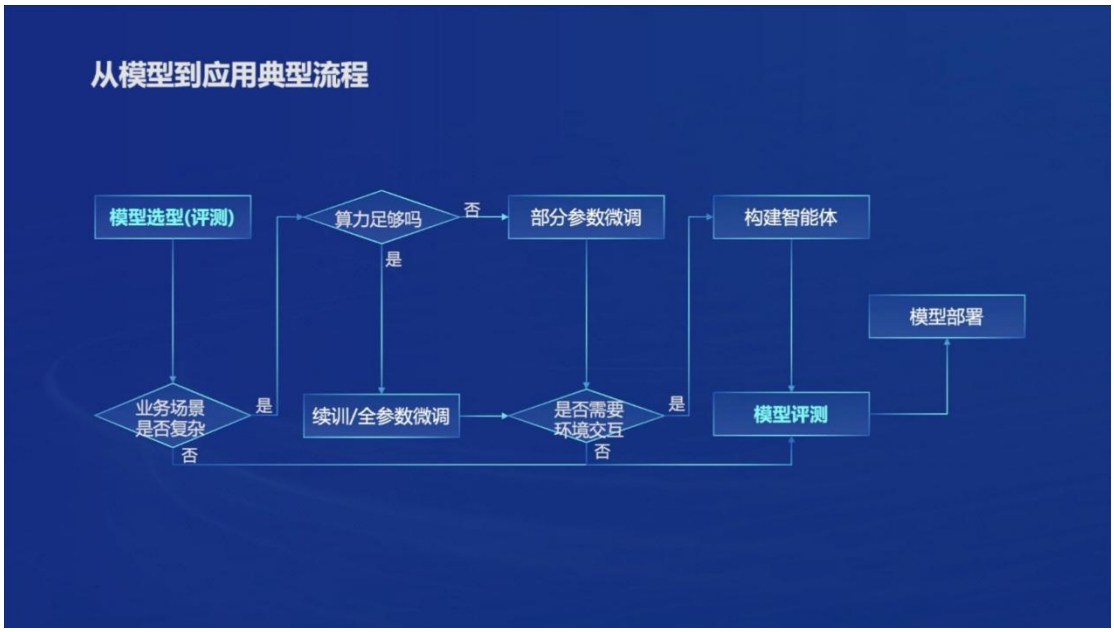
八、从模型到应用

1．智能客服

2．个人助理

3．行业应用

4．典型流程图



（1）疑问：如何判断算力是否足够？

九、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，速度达到 3600 token/sec/gpu；
- 微调 XTuner，支持全参数微调，支持 LoRA 等低成本微调；
- 部署 LMDeploy，每秒生成 2000+tokens；
- 评测 OpenCompass，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。

1. 开放高质量语料数据（https://opendatalab.org.cn）；

（1）书生万卷 1.0 2TB

（2）书生万卷 CC400 GB InternLM2 预训练语料

2. InternEval 预训练开源框架

（1）8 卡到千卡；

（2）Hybrid Zero 加速；

（3）兼容 HuggingFace 等技术生态，支持各类轻量化技术；

（4）开箱即用，支持多种语言模型，修改配置即可训练；

3. 微调

(1) 支持增量续训和有监督微调(全量参数、部分参数)

(2) 高效微调框架 XTuner 横向对比、适配多种生态（HuggingFace、ModelScope 等）和硬件（Nvida20 以上所有显卡，最低 8GB 显存可微调 7B 模型）

4. 评测 2024.1.30OpenCopass 发布思南大模型评测体系

（1）CompassRank 中立全面性能榜单（大语言模型及多模态模型榜单）

（2）CompassKit 大模型评测全栈工具链

①支持包括 GSM-8K、MMLU 等主流数据集上的污染检测；

②更丰富的模型推理接入，支持近 20 个商业模型 API，支持 LMDeploy、vLLM LightLLM 等推理后端；

③支持 200K 大海捞针测试，支持多个主流长文本评测基准；

④支持基于大模型评价的中英文双语主管评测，提供模型打分、模型对战多种能力，灵活切换上百种评价模型。

（3）通过（MMBench、MMLU 等）共建开源开放 CompassHub 高质量评测基准社区，荟聚优秀评测集；

（4）OpenCompass 广泛应用于头部大模型企业（阿里、华为、百度、腾讯、微软等）和科研机构（如复旦大学），并获得 META 官方推荐唯一国产大模型评测体系；是有 100+评测集，50 万+题目；在 OpenCompass2.0 的评测下发布了 OpenCompass 年度榜单（综合客

观评测）及 OpenCompass 年度榜单（中文主管评测-对战胜率），通过更加准确的循环评测策略分析，发现：

①综合客观评测下

1）得分最高的 GPT-4-Turbo 也仅仅达到了 61.8 分（满分 100 分）；

2）国内多个模型（如阿里巴巴的通义千问 Qwen-Max、百度 ERNIE Bot-Pro、通义千问 Qwen-72B-Chat、书生 •浦语的 InternLM2-Chat-20B）综合能力和 GPT-4-Turbo 接近，但复杂推理仍是短板，与模型尺寸存在强相关性；

3）各个模型打分结果显示，在语言和知识“文科”维度，轻量级模型和重量级/闭源商业模型与差距较小，但在数学、推理、代码等“理科”维度上，性能和尺寸强相关；

4）大量开源模型和 API 模型的客观性能和主观性能存在较大的偏差，社区需要在夯实客观能力基础、偏好对齐和对话体验上下功夫。

②中文主管评测-对战胜率评测下

1）国内近期发布的闭源大模型（如智谱轻言 GLM-4、零一万物 Yi-34B-Chat、百川智能 Baichuan2-Turbo、书生 • 浦语 InternLM2-Chat-7B 等）表现优异，多个维度上缩小了与 GPT-4-Turbo 的差距。

2）国内模型在中文场景具有性能优势，单个维度超越 GPT-4-Turbo；

3）零一万物 Yi-34B-Chat、书生 • 浦语 InternLM2-Chat-20B 以中轻量级的尺寸，展示出优秀的综合性对话体验，并接近商业闭源模型的性能，未来可期。

5. 部署，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，支持 Python、gRPC、RESTful 接口调用。

(1) 轻量化

① 4bit 权重；

② 8bit k/v；

(2) 高效推理引擎（turbomind 引擎及基于 pytorch 的引擎）

① 持续批处理技巧（后续课程是否会介绍？）；

② 深度优化的低比特计算 Kernels；

③ 模型并行；

④ 高效的 k/v 缓存管理机制（8bit k/v）

(3) 完备易用的工具链

① 量化、推理、服务全流程，接口可兼容 openai-server、gradio、triton inference server 使用；

1) 延伸概念：Gradio 一种 AI 展示工具。在人工智能飞速发展的今天，向世界展示你的 AI 模型变得越来越重要。这就是 Gradio 发挥作用的地方：一个简单、直观、且强大的工具，让初学者到专业开发者的各个层次的人都能轻松展示和分享他们的 AI 模型。

2) 深度学习部署神器——triton inference server。

a. triton 可以充当服务框架去部署你的深度学习模型，其他用户可以通过 http 或者 grpc 去请求，相当于你用 flask 搭了个服务供别人请求，当然相比 flask 的性能高很多了

b. triton 也可以摘出 C-API 充当多线程推理服务框架，去除 http 和 grpc 部分，适合本地部署多模型，比如你有很多模型要部署，然后分时段调用，或者有 pipeline，有了 triton 就省去你处理显存、内存和线程的麻烦。

② 无缝对接 OpenCompass 评测推理精度；

③ 多维度推理速度评测工具；

(4) 支持交互式推理，不为历史对话买单

(5) 部署框架性能对比 LMDeploy VS vLLM，结果 LMDeploy 遥遥领先。

6. 轻量级智能体框架 Lagent

(1) 支持多种智能体能力（ReAct、ReWoo、AutoGPT）

(2) 灵活支持多种大语言模型（GPT-3.5/4、InternLM、Hugging Face Transformers、Llama）

(3) 简单易拓展、工具丰富

① AI 工具：文生图、文生语言、图片描述等；

② 能力拓展：搜索、计算器、代码解释器等；

③ Rapid API：出行 API、财经 API、体育资讯 API 等。

(4) 举例：代码解数学题、零样本泛化：多模态 AI 工具使用。

(5) 多模态智能体工具箱 AgentLego

① 提供了大量视觉、多模态相关领域的前沿算法功能；

② 支持多个主流智能体系统，如 LangChain，Transformers Agent，lagent 等；

③ 多模态工具调用接口灵活，支持各类输入输出格式的工具函数；

④ 工具检索及一键式远程工具部署。

总结，六大模块构成了书生·浦语的全链条开源开放体系。