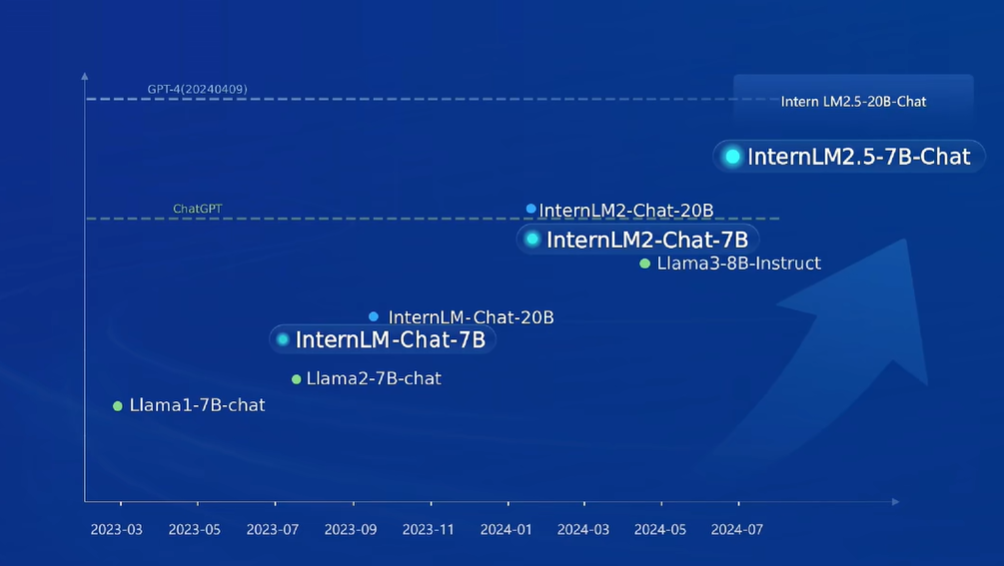
# 书生浦语大模型全链路开源体系

1. 开源之路



* 1. 2023.7.6 InternLM-7B 开源率先免费商用，发布全聊条开源工具体系
  2. 2023.9.20 20B开源，工具链全线升级
  3. 2024.1.17 InternLM2开源
  4. 2024.7.4 2.5开源
  5. 书生浦语大模型2.5



* + 1. 性能提升
       1. 2.5相对于2性能提升20%
       2. 上下文tokens限制提升至100万
    2. 功能增强
       1. 大规模上下文处理
       2. 自主规划和搜索复杂任务能力



1. 核心技术
   1. 模型能力飞轮（当前模型-->更好模型）自身迭代
      1. 当前模型进行数据过滤，智能评估得到预训练数据
      2. 当前模型通过指令生成，辅助标注对齐数据
   2. 高质量合成数据
      1. 基于规则的数据构造
      2. 基于模型的数据扩充
      3. 基于用户反馈的数据生成
2. 全链条开源



* 1. 数据：
     1. 书生·万卷首个精细处理的开源多模态语料库数据



* + 1. 数据处理
       1. 数据提取：MinerU，一站式开源高质量数据提取工具，支持多格式(PDF/网页/电子书)，智能萃取，生成高质量预训练/微调语料。
          1. 复杂版面/公式精准识别
          2. 性能超过商业软件
          3. [Http://github.com/opendatalab/MinerU](http://Http://github.com/opendatalab/MinerU)
       2. 数据标注
          1. Label LLM:专业致力于 LLM 对话标注，通过灵活多变的工具配置与多种数据模态的广泛兼容，为大模型量身打造高质量的标注数据。

支持指令采集、偏好收集、对话评估………

多人协作、任务管理、源码开放可魔改

[Http://github.com/opendatalab/LabelLLM](http://http//github.com/opendatalab/MinerU)

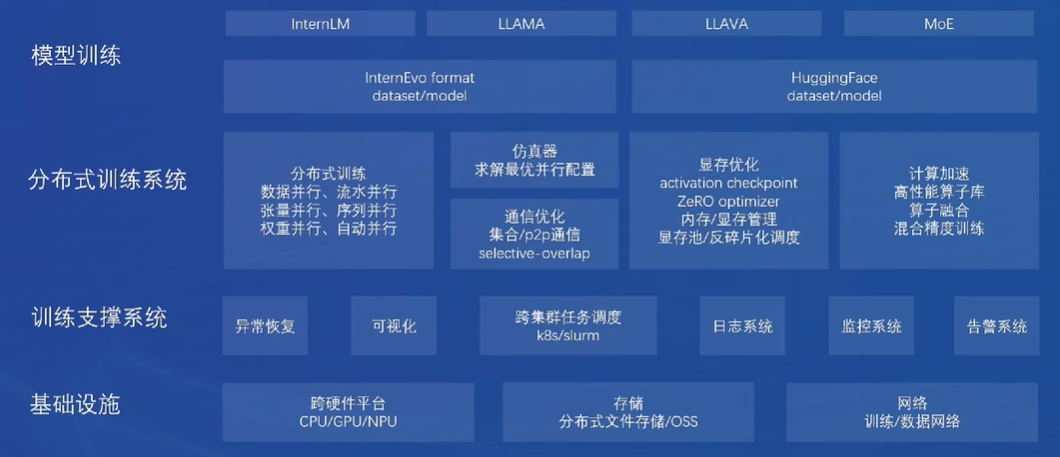
* + - * 1. Label U：一款轻量级开源标注工具，自由组合多样工具，无缝兼容多格式数据，同时支持载入预标注，加速数据标注效率。

支持图片、视频、音频多种数据标注0

小巧灵活，AI标注导入二次人工精修

[Http://github.com/opendatalab/labelU](http://http//github.com/opendatalab/MinerU)

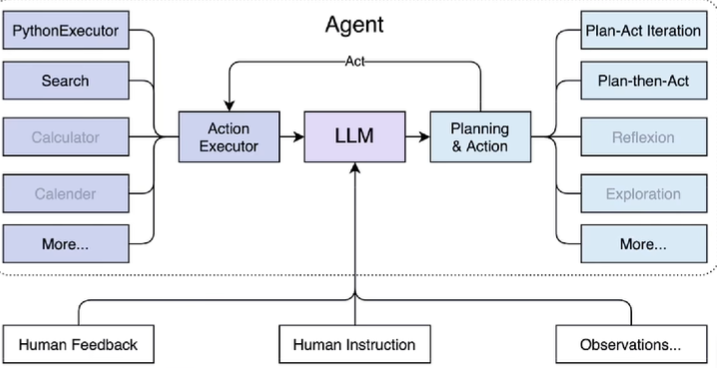
* 1. 预训练：InternEvo，性能超过国际主流训练框架DeepSpeed



* 1. 微调：XTuner，多种微调和偏好对齐算法，支持干亿参数+百万上下文



* + 1. 适配多种生态：多种微调算法多种微调 & 偏好对齐算法，覆盖各类应用场景。
    2. 适配多种开源生态：支持加载 HuggingFace、ModelScope 模型或数据集。
    3. 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支持干亿参数 +百万上下文训练。
  1. 部署：LMDeploy，性能超过国际主流推理框架 vlm
     1. 高效的推理引擎：
        1. TurboMind 推理引擎：LMDeploy 使用 TurboMind 作为高性能的推理引擎，能够显著提升模型的推理速度。
        2. PyTorch 推理后端：支持使用 PyTorch 作为推理后端，确保与现有生态系统的兼容性。
     2. 可靠的量化技术：
        1. Weight-only 量化：通过只对权重进行量化，减少模型的存储和计算需求，同时保持较高的推理精度。
        2. K/V Cache 量化：对 K/V Cache 进行量化，进一步优化内存使用和推理速度。
     3. 多样的接口支持：
        1. Python 推理接口：提供 Python 接口，方便开发者在 Python 应用中调用模型。
        2. RESTful 接口：支持 RESTful API，使 Web 应用和其他系统可以通过 HTTP 请求调用模型。
        3. gRPC 接口：支持 gRPC 协议，提供高性能的 RPC 服务，适用于微服务架构。
     4. 卓越的兼容性：
        1. 类 OpenAI 服务：提供类似于 OpenAI 的服务接口，方便开发者迁移和集成。
        2. Gradio：支持 Gradio，一个用于快速构建和共享机器学习应用的工具。
        3. Triton 推理服务：支持 NVIDIA Triton 推理服务器，提供高性能的多模型推理能力。
  2. 评测：OpenCompass，社区最全面的开源评测体系
  3. 应用：大模型具有局限性



* + 1. MindSearch 思索式开源搜索应用
    2. Lagent 首个支持代码解释器的智能体框架，轻量级
       1. 支持多种类型智能体能力
          1. ReAct
          2. ReWoo
          3. AutoGPT
       2. 支持多种大语言模型：GPT-3.5/4，IntrenLM，HuggingFace Transformers，Llama
       3. 支持丰富的工具
          1. AI工具：文生图、文生语音、图片描述
          2. 能力拓展：搜素、计算机、代码解释器
          3. Rapid API：出行API、财经API、体育资讯API
    3. MinerU 高效文档解析工具
    4. HuixiangDou 基于专业知识库的群聊助手，企业级知识库构建工具，群聊场景 LLM 知识助手，为即时通讯(如微信、飞书)群聊场景设计。
       1. 利用检索增强生成(RAG)，非参数记忆，利用外部知识库提供实时更新的信息。
       2. 利用知识图谱KG，结构化知识库匹配意图行为可解释。
       3. 场景特点
          1. 无关问题不吭声
          2. 明确回答直接回复
          3. 不违背核心价值观