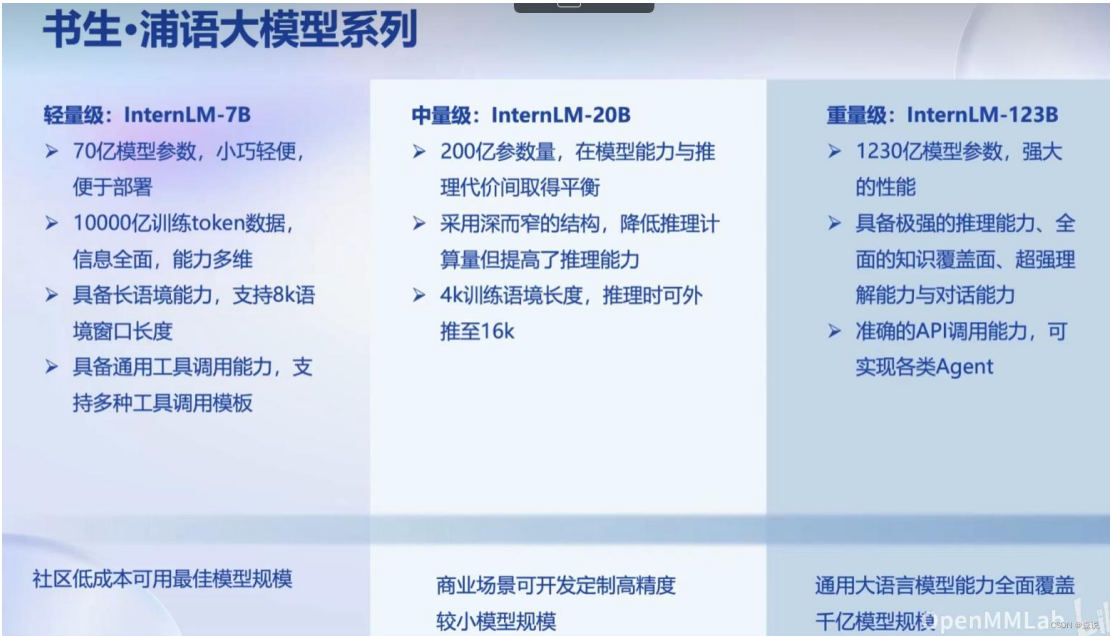
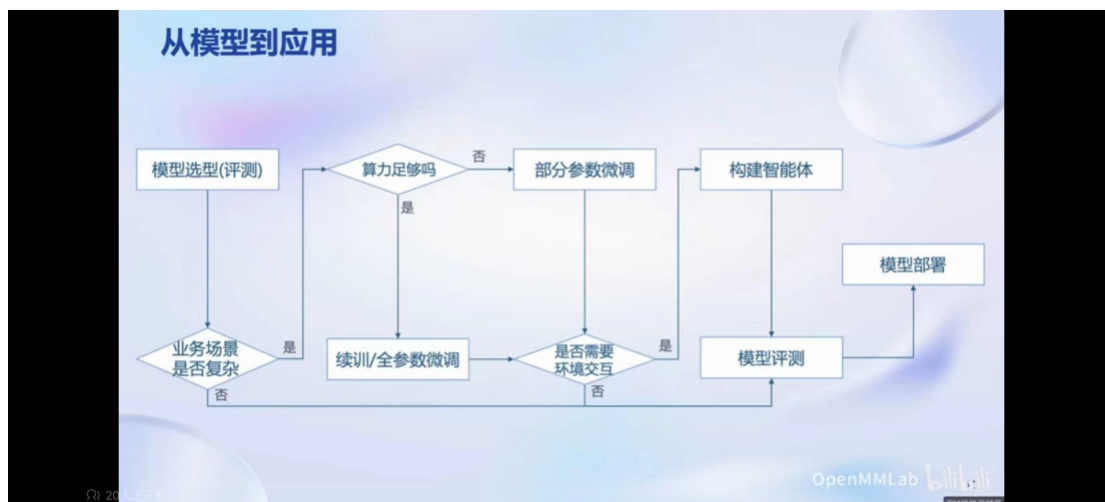




模型介绍



模型到应用



性能



体系




数据

多模态数据


### 全链条开源开放体系 | 数据

#### 书生·万卷 1.0




**文本数据**

- 50亿个文档
- 数据量超 1TB



**图像-文本数据集**

- 超2,200万个文件
- 数据量超140GB



**视频数据**

- 超1,000个文件
- 数据量超900GB

总数据量: 2TB  
发布日期: 8月14日

#### 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。

OpenMMLab

数据集平台

### 全链条开源开放体系 | 数据

#### 飞速成长

模态 **30+**

数据集 **5,400+**

数据大小 **80TB**

#### Open dataLab

#### 丰富多样的开放数据



**60 亿 图像**

LAION-5B SA-1B ImageNet



**8 亿 片段 视频**

MovieNet Kinetics MOT



**1 万亿 tokens 语料**

The Pile C4 WikiQA



**1 百万 3D 模型**

OmniObject3D ShapeNet Scannet



**2 万 小时 音频**

LibriSpeech VoxCeleb Speech Commands

#### 服务与工具



**灵活检索**

支持 **10+** 搜索条件组合



**高速下载**

单文件稳定速度至少 **20M/s**



**智能标注**

支持 **30+** 工具组合形式



**高效采集**

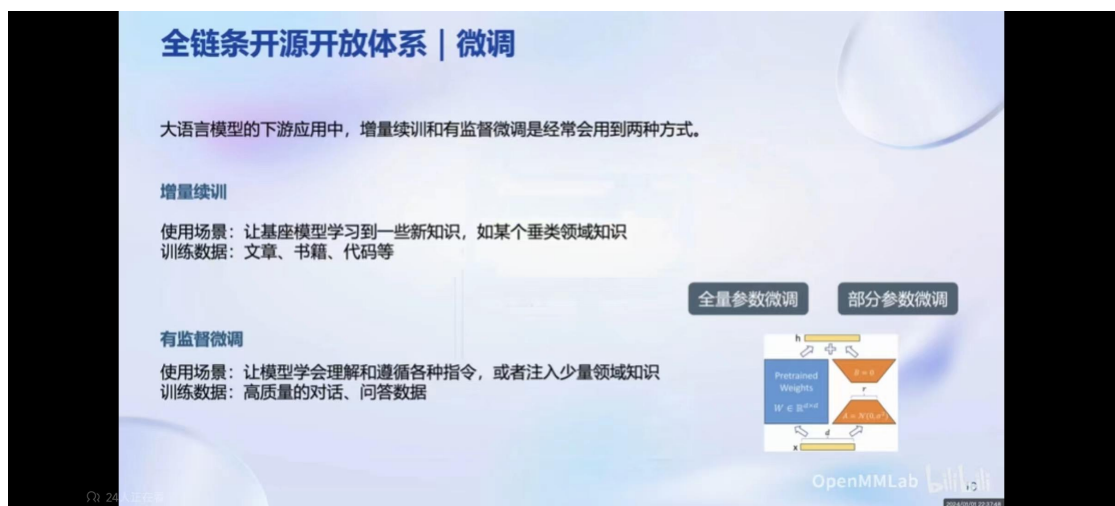
整体效率可提升 **40%**

OpenMMLab

预训练



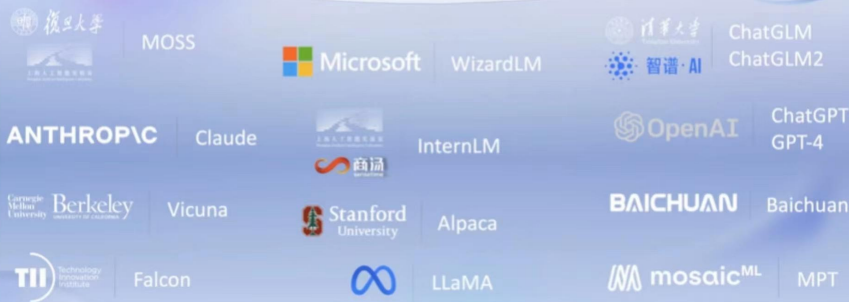
微调



评测

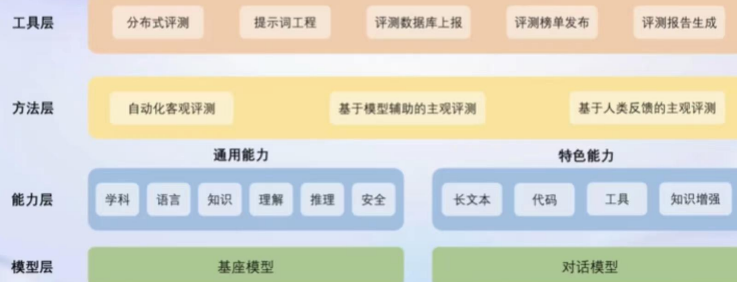
## 全链条开源开放体系 | 评测

丰富的模型支持



## 全链条开源开放体系 | 评测

OpenCompass 开源评测平台架构



## 全链条开源开放体系 | 评测

OpenCompass

全球领先的大模型开源评测体系  
6大维度, 80+评测集, 40万+评测题目

学科	语言	知识	理解	推理	安全
初中考试 中国高考 大学考试 语言能力考试 职业资格考试	字词释义 成语习语 语义相似 指代消解 翻译	知识问答 多语种知识问答	阅读理解 内容分析 内容总结	因果推理 常识推理 代码推理 数学推理	偏见 有害性 公平性 隐私性 真实性 合法性

部署



## 全链条开源开放体系 | 部署

### 大语言模型特点

- 内存开销巨大**
  - 庞大的参数量
  - 采用自回归生成token, 需要缓存k/v
- 动态Shape**
  - 请求数不固定
  - token逐个生成, 且数量不定
- 模型结构相对简单**
  - transformer 结构, 大部分是 decoder-only

### 技术挑战

- 设备**
  - 低存储设备 (消费级显卡、移动端等) 如何部署?
- 推理**
  - 如何加速 token 的生成速度
  - 如何解决动态shape, 让推理可以不间断
  - 如何有效管理和利用内存
- 服务**
  - 提升系统整体吞吐量
  - 降低请求的平均响应时间

### 部署方案

- 技术点**
  - 模型并行
  - 低比特量化
  - Attention优化
  - 计算和访存优化
  - Continuous Batching

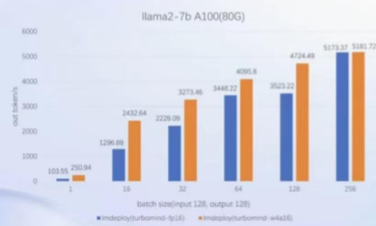
OpenMMLab

## 全链条开源开放体系 | 部署

### 领先的推理性能

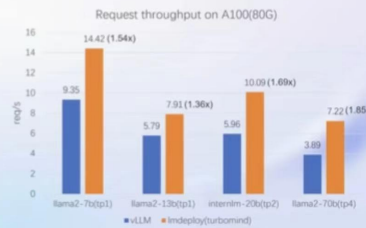
#### 静态推理性能

固定 batch, 输入/输出 token 数量



#### 动态推理性能

真实对话, 不定长的输入/输出



OpenMMLab

## 全链条开源开放体系 | 部署



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



#### 高效推理引擎

- 持续批处理技巧
- 深度优化的低比特计算 kernel
- 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
- 和 OpenAI 接口高度兼容的 API server

OpenMMLab

## 全链条开源开放体系 | 部署

### 大语言模型特点

**内存开销巨大**

- 庞大的参数量
- 采用自回归生成token, 需要缓存k/v

**动态Shape**

- 请求数不固定
- token逐个生成, 且数量不定

**模型结构相对简单**

- transformer 结构, 大部分是 decoder-only

### 技术挑战

**设备**

- 低存储设备 (消费级显卡、移动端等) 如何部署?

**推理**

- 如何加速 token 的生成速度
- 如何解决动态shape, 让推理可以不间断
- 如何有效管理和利用内存

**服务**

- 提升系统整体吞吐量
- 降低请求的平均响应时间

### 部署方案

**技术点**

- 模型并行
- 低比特量化
- Attention优化
- 计算和访存优化
- Continuous Batching

OpenMMLab

智能体

## 全链条开源开放体系 | 智能体

### 代码解数学题

### 零样本泛化：多模态 AI 工具使用

OpenMMLab

## 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

OpenMMLab

开放体系

