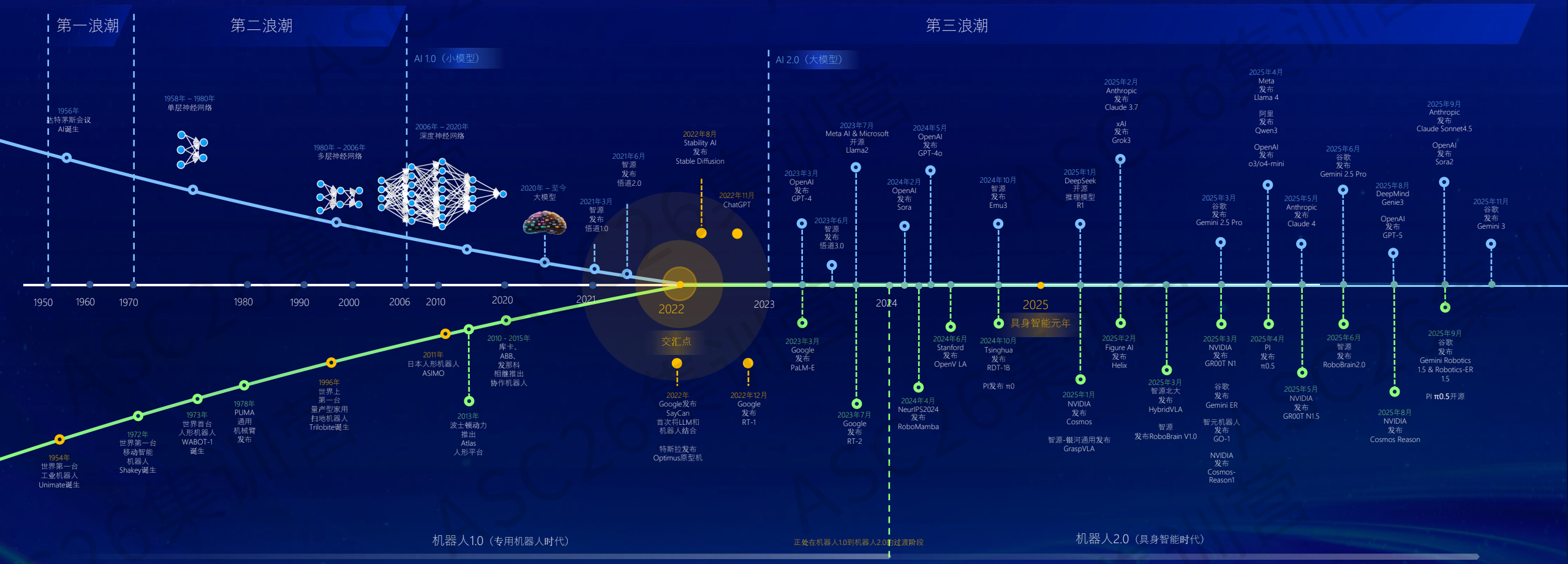
BAAI 智源研究院

人工智能基础培训

大模型：AI第三次浪潮的新拐点

具身智能引领AI与机器人融合的新范式



智源大模型：从“悟道”到“悟界”

融合数字世界与物理世界

悟道 WUDAO

2021年3月，智源发布“悟道”系列大模型开启了中国大模型时代

道：大语言模型系统化方法及路径

大语言模型 → 多模态大模型（理解世界 解析人类大脑） → 世界模型

悟界 WUJIE

2025年6月，智源发布“悟界”系列大模型推动了AI 从数字世界迈向物理世界

界：虚实世界的边界突破

数字世界

物理世界

智源科研布局

持续探索技术前沿，打造支撑产业发展的共性基础“操作系统”

前沿技术探索

具身操作系统
RoboOS

悟界

悟界

RoboBrain

全球首个跨本体
具身大脑模型

悟界

Brainμ

全球首个脑科学多模态
通用基础模型

悟界

OpenComplex

全原子微观生命模型

悟界

Emu

全球首个原生多模态世界模型

共性基础攻关

智算操作系统
FlagOS

众智 FlagOS

全球覆盖AI芯片种类最多的统一开源智算系统软件栈

FlagOpen：打造大模型时代的 Linux

面向异构算力、支持多种框架的大模型全栈开源技术基座

https://github.com/FlagOpen

北京智源人工智能研究院 BEIJING ACADEMY OF ARTIFICIAL INTELLIGENCE

FlagOpen 2.0

持续开源开放数据集

- 数据集大于160个
- 数据集下载量超160万
- 数据集被使用量212次
- 数据工具支持算子24个
- 数据集覆盖大模型全周期
- 社区点赞量2131次

模型覆盖多个应用场景

- 模型月下载总量达2千万
- 模型被使用量700+次
- 语言模型39个
- 多模态模型31个
- BGE向量模型28个
- 社区点赞量6439次

大模型全链路系统架构

- 支持算子300+
- 支持25款国产芯片
- 支持2种训练框架
- 支持异构训练
- 接入17家芯片厂商

评测平台

- 5个评测榜单
- 4种评测方法
- 已评测1000+模型
- 3个自建评测数据集

BAAI

BAAI FlagOpen大模型开源技术体系

Open Source System for Large Models

Model, Algorithm, Data, Training Framework, Evaluation Tools, etc.

https//flagopen.baai.ac.cn

FlagOpen飞智

AltCLIP:换文本塔来扩展CLIP模型语言能力，低资源高效

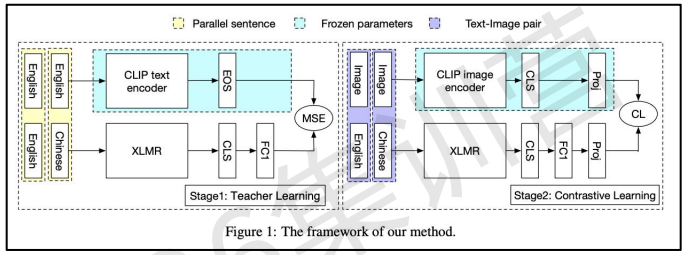
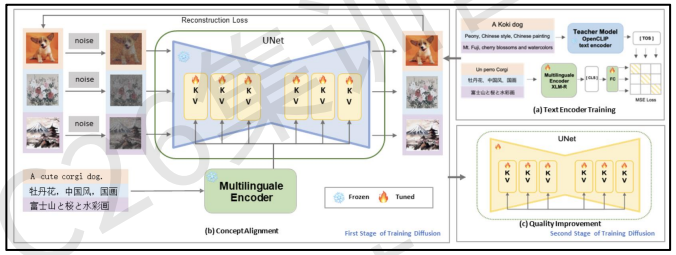
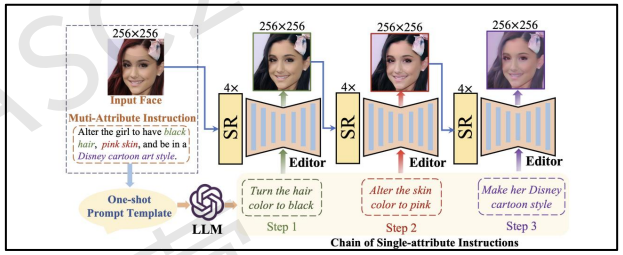


Figure 1: The framework of our method.

AltDiffusion:换文本塔来扩展Diffusion模型语言能力，支持18种语言



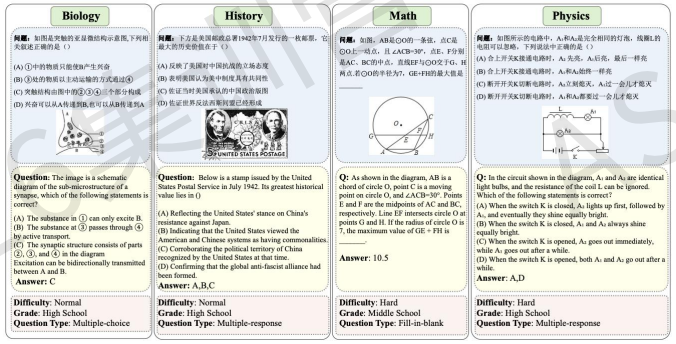
CoIE:基于LLM CoT能力解锁多步可控图片编辑



INFINITY-MM

Infinity Instruct

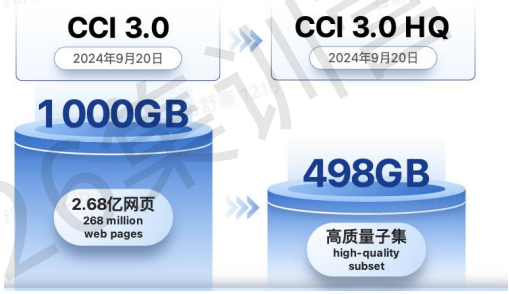
CMMU：中文多题型多模态理解和推理评测数据集



Aquila和Aquila2：首个具备中英双语知识、支持商用许可协议（7B/34B/70B, 8x16B）

悟道·天鹰 WUDAO Aquila

TACO 是一个代码生成基准，有 26443 个问题。它可以用来评估语言模型根据自然语言规范生成代码的能力



CCI3.0 大规模中文语料数据库。帮助模型更深入理解中文语言。

精品数据集构建成果

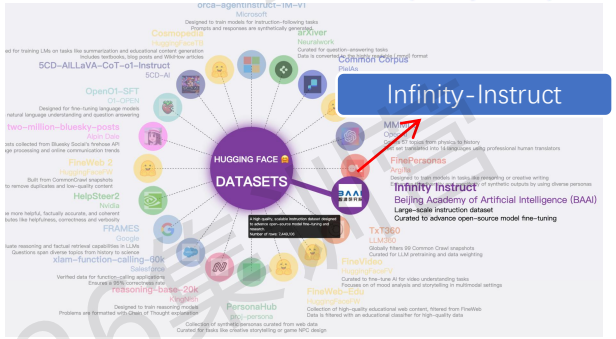
中文最大规模最高质量的可信互联网语料库

AquilaMoE　Emu3

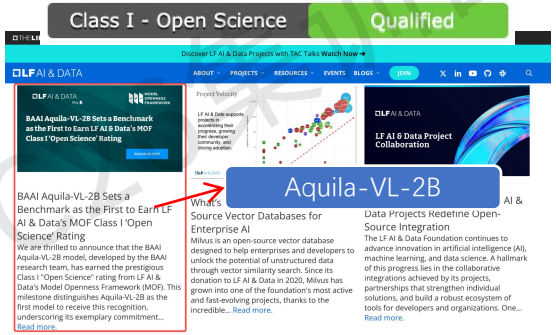
100000+	467	1000+
数据集下载量 Dataset downloads	机构用户 Institutional users	个人用户 Individual users

服务企事业单位大模型研发	支撑人工智能模型迭代训练	助推国家高质量语料生态建设
Serve the development of large-scale models for enterprises and institutions.	Support the iterative training of artificial intelligence models.	Promote the construction of a high-quality language dataset ecosystem for the country.

Huggingface Most Liked Datasets of 2024 (Top20)



全球首个Linux基金会MOF评级达到“最开源”Class I等级模型



中文互联网语料库3.0

质量标注　教育水平　来源可信

规模空前，来源广泛
Leading in scale, wide-ranging sources

精细标注，赋能应用
Precise annotated, empowering applications

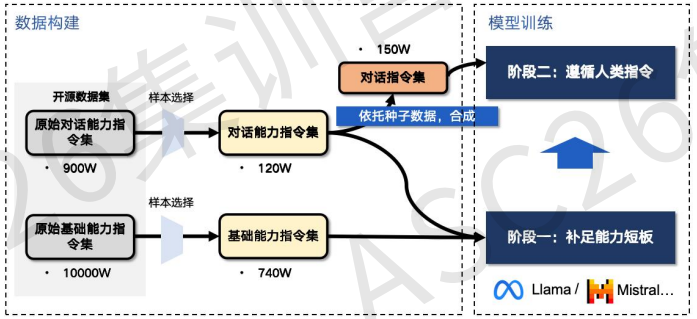
效果突破，更懂中文
Breakthrough effects, deeper understanding of Chinese

CCI 3.0　2024年9月20日　1000GB　2.68亿网页 268 million web pages

CCI 3.0 HQ　2024年9月20日　498GB　高质量子集 high-quality subset

千万级文本指令Infinity-Instruct

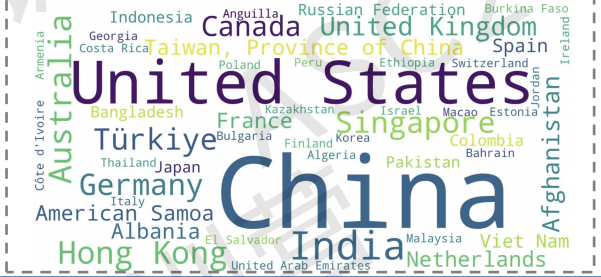
数据筛选　标签体系　数据合成



千万级多模态指令Infinity-MM

数据筛选　标签体系　数据合成

55个国家和地区，200+机构申请下载



DataAgent:基于Agent的高质量数据集构建

数据筛选　标签体系　数据合成　智能化　自动化

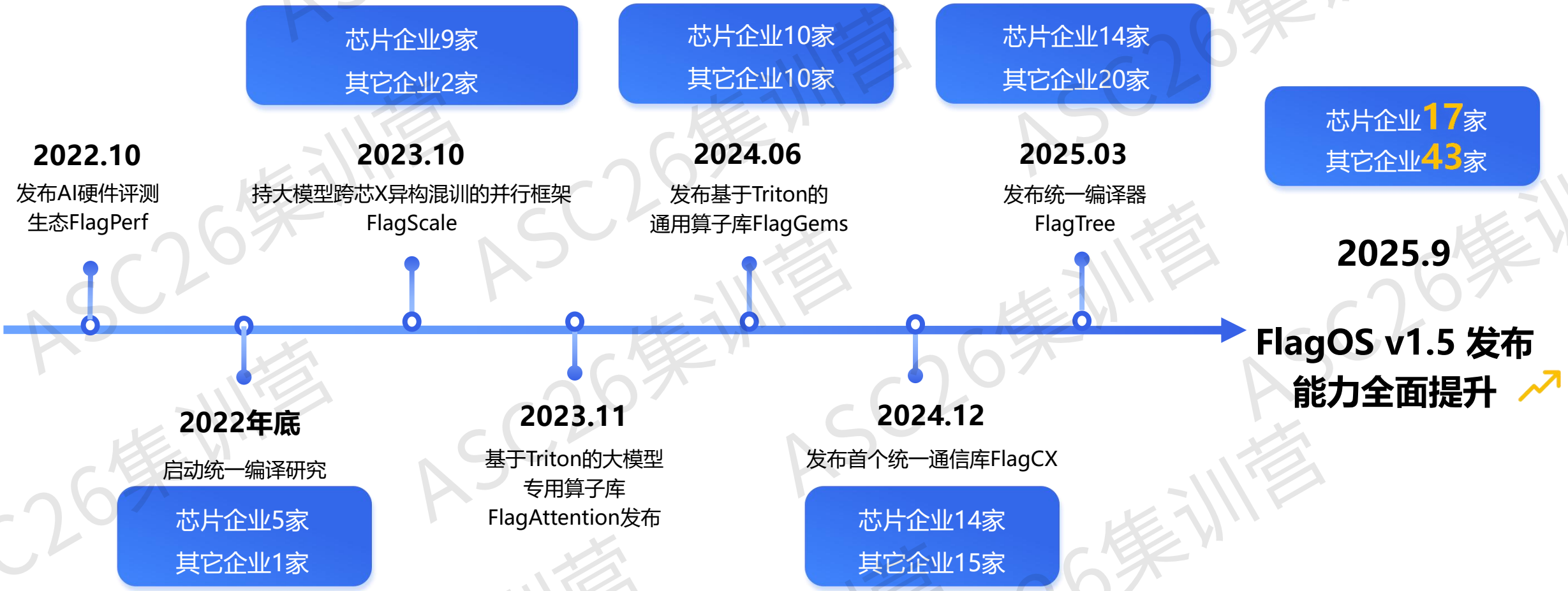
众智FlagOS

面向多种AI芯片的统一、开源系统软件栈

北京智源人工智能研究院

众智FlagOS: 面向多种AI芯片的统一、开源系统软件栈

以技术攻关打破AI芯片生态壁垒，共筑开放计算的创新蓝图



众智FlagOS技术生态：打造统一、开源的AI系统软件技术栈

以技术攻关打破AI芯片生态壁垒，共筑开放计算的创新蓝图

智源研究院牵头北大、清华、中科院计算所等多家科研机构，十多家芯片厂商、多家操作系统厂商、服务器厂商等共同打造FlagOS

CUDA技术生态

全球超过90%的开发者在CUDA上持续创新

新算子开发	80%以上的新算子基于CUDA开发
开源模型	全球有一百多万个开源模型
开源算法项目	全球有数千个以上活跃的开源项目
深度学习框架	PyTorch、Paddle、TensorFlow等原生支持
加速框架	训练：Megatron，DeepSpeed等 推理：TensorRT,，vLLM、SGLang等
算子库	cuBLAS, cuDNN, FlashAttention等
通信库	NCCL
编程语言	CUDA C
编译器	CUDA编译器
芯片	英伟达

通过自动化方法，快速发展生态

技术研发，打造核心技术根基

众智FlagOS技术生态

FlagOS工具

- KernelGen：多芯片算子自动生成
- FlagRelease：开源模型自动迁移发布
- FlagCICD：跨芯代码自动迁移集成发布（建设中）

支持PyTorch、Paddle

FlagOS核心开源技术库

- FlagScale：训推一体加速框架，支持多种引擎(Megatron, vLLM，SGLang等)
- FlagGems-BLAS, FlagGems-LLM, FlagAttention
- FlagCX统一通信库
- Triton，TileLang(计划中)
- FlagTree统一编译器

多个厂商、多种架构的AI芯片

通过全面AI赋能FlagOS实现弯道超车

众智FlagOS v1.6：面向多种AI芯片的全栈开源技术

AI大模型

语言模型	多模态模型	具身智能模型
DeepSeek，Qwen，Seed-oss，GPT-oss，Step，Grok-2，Llama3，GLM4.5，Kimi-2, Hunyuan，ERNIE4.5等	EMU3/3.5，MiniCPM-v，Qwen2.5/3-VL，ERNIE4.5，Llava-Onevision	RoboBrain2, RoboBrainX0.5 Pai-0, Pai-0.5

FlagOS 生态使能项目

- Megatron-FL
- vLLM-plugin-FL
- TransformerEngine-FL

深度学习框架

(PyTorch, PaddlePaddle, etc)

FlagOS 开源核心库

- FlagGEMs 通用大模型算子
- FlagScale 训练推理并行框架
- FlagTree 统一编译器
- FlagCX 统一通信库

FlagOS 开源工具

- KernelGen v1.0 算子自动生成工具
- FlagRelease 模型自动迁移和发版工具
- FlagCICD 开源项目迁移、测试和发版工具
- FlagPerf 多芯片评测工具

支持多种芯片架构

17家厂商25款AI芯片

英伟达、华为、寒武纪、昆仑芯、海光、摩尔线程、沐曦、壁仞、天数智芯、清微智能、平头哥、燧原、ARM中国、算能、曦望科技、太初等

支持多种系统硬件形态

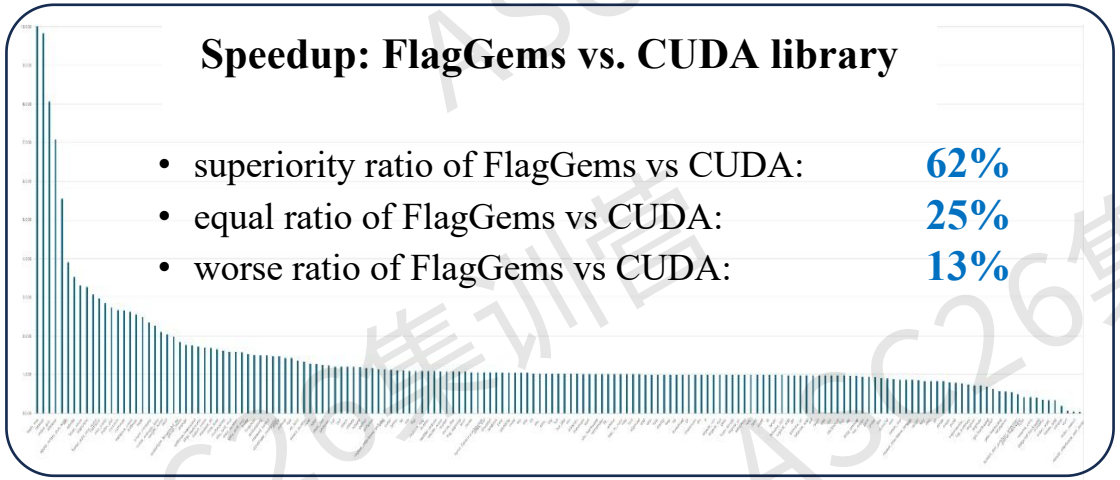
端侧计算模组 | AI PC | AI服务器 | 超节点系统

支持多种应用场景

数据中心 (train & Inference) → 机器人 (cloud-edge cooperation) → 边缘 (inference)

算子库FlagGems：性能优异、全球最大的Triton通用算子库

- **已经发布了225个常用大模型算子，87%达到或超过CUDA算子性能**



- **成为PyTorch基金会生态合作项目**

- **引入KernelGen自动生成算子**

FlagOS1.6版本（2026年1月9日发布）新增138个高性能Triton算子，FlagGems算子总数将超过360个，其中296个算子性能达到或超过CUDA原生算子。

- **7家国产厂商适配FlagGems算子，基于FlagTree统一编译器，全部测试超过220个算子，以确保技术的泛化性**

基于FlagTree统一编译器，7家国产厂商适配FlagGems算子相比厂商原生**算子加速比中位数**.

GPU A	GPU B	GPU C	GPU D	DSA A	DSA B	DSA C
112%	107%	104%	95%	96%	80%	79%

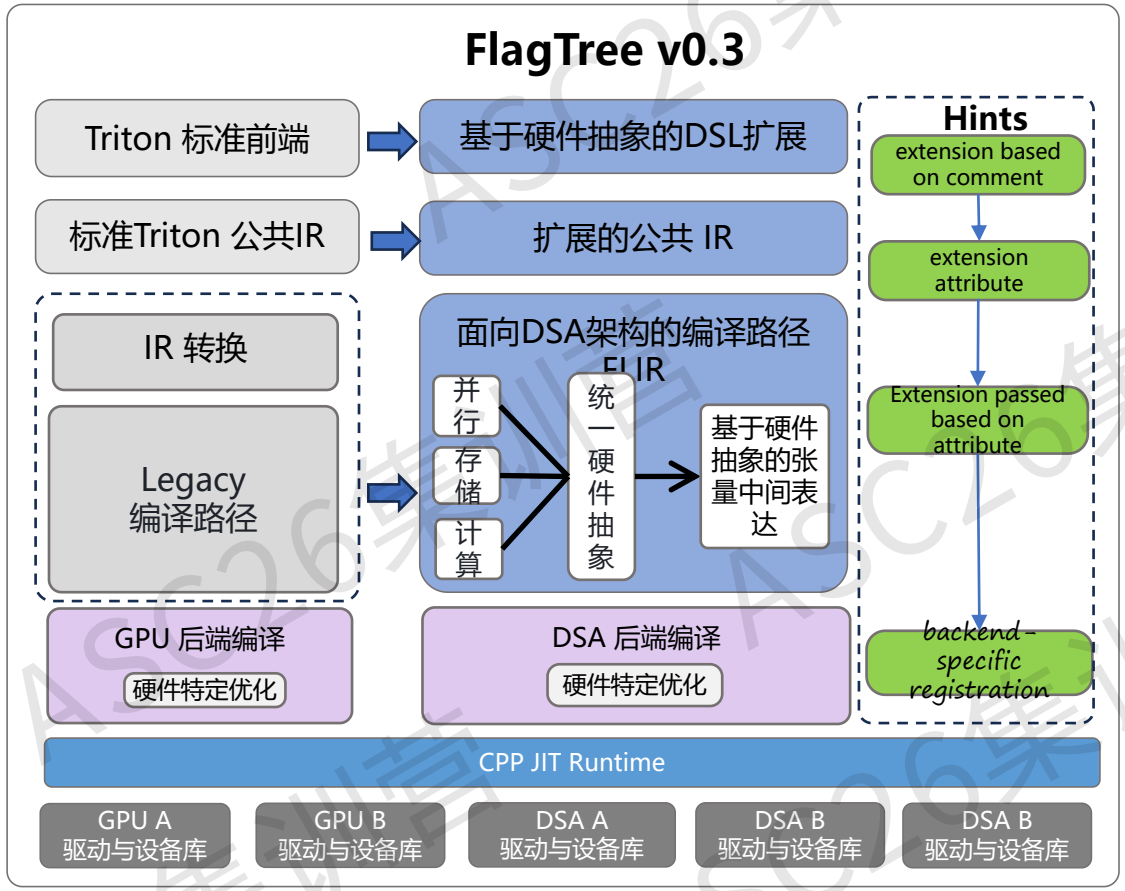
- **PreTune 离线搜寻最优内核配置**

问题：为了满足每次request的变长输入的优化，Triton使用online AutoTune，造成推理耗时陡增

解法：设计高效的离线搜寻机制，建立Shape与内核参数对应关系，online 直接查表获取内核参数，降低搜索耗时

收益案例：应用于 Qwen2.5-7B-Instruct，端到端推理性能提升40%。

公共编译器FlagTree：支持多厂家芯片，极大提升Triton性能



FlagTree v0.3 —— 不同厂商芯片统一编译器，提升性能。

1. 统一编译器： 累计已经支持了12家厂商，20种芯片，包括英伟达、华为、寒武纪、海光、寒武纪、摩尔线程、沐曦、天数智芯、清微智能、ARM中国、算能、平头哥。

2. 提升性能：

- **Hints：不改变Triton语言，增加hints，基于硬件感知的编译优化技术**

通过注释嵌入硬件优化提示，挖掘硬件潜力，提升算子性能，已在英伟达、华为昇腾和Arm China AIPU 构建编译链路，其中部分重点算子在华为昇腾提速**10%**

- **CPP JIT Runtime提升算子速度**

基于CPP语言的运行时包装机制，降低运行时开销，应用于20余个算子，平均性能提升**20%**以上；该技术已支持英伟达、天数和华为昇腾

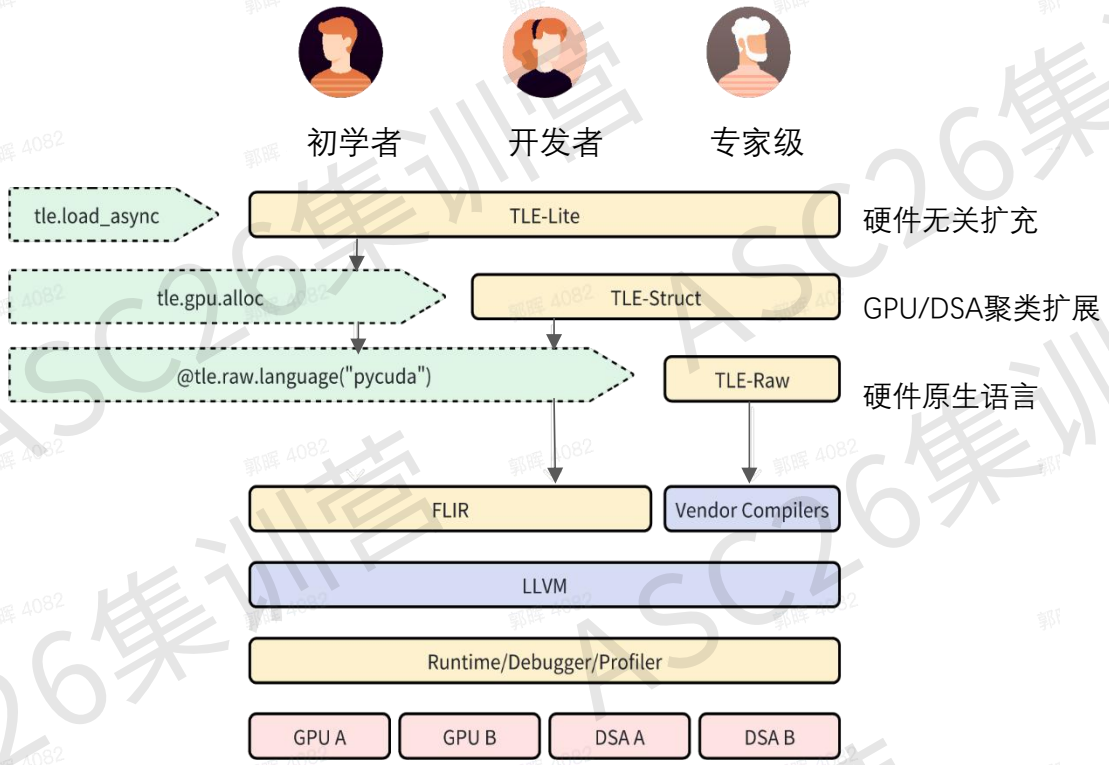
案例: 使用FlagOS全部替换Qwen的CUDA依赖，使用FlagGems和FlagTree 可以**获得逼近CUDA最优的端到端推理性能，显著原生Triton**。

模型名称	FlagScale + FlagGems + OpenAI Triton编译器	FlagScale + FlagGems + FlagTree
Qwen2.5-7B-Instruct	35%	95%
Qwen3_30A_A3B	25%	92%

FlagTree v0.4：面向开发者多样化需求，构建三层渐进式编程语言

FlagTree v0.4 —— 扩展Triton语言，提升硬件感知

✓在原有Triton语言基础上，提出分层的语言扩展 ***Triton Language Extension (TLE)***



为算子开发者、芯片厂商提供“一站式”的编译器能力

- 给予开发者在易用和极致优化之间的多种选择，只需要修改Triton中的部分代码，即可获得显著性能收益。
- 加深架构感知，充分优化性能，甚至为芯片厂商提供了结合厂商原生底层语言的可行性
- **兼容Triton，保留当下200多个全球算子库的生态优势**

TLE三层语言设计：

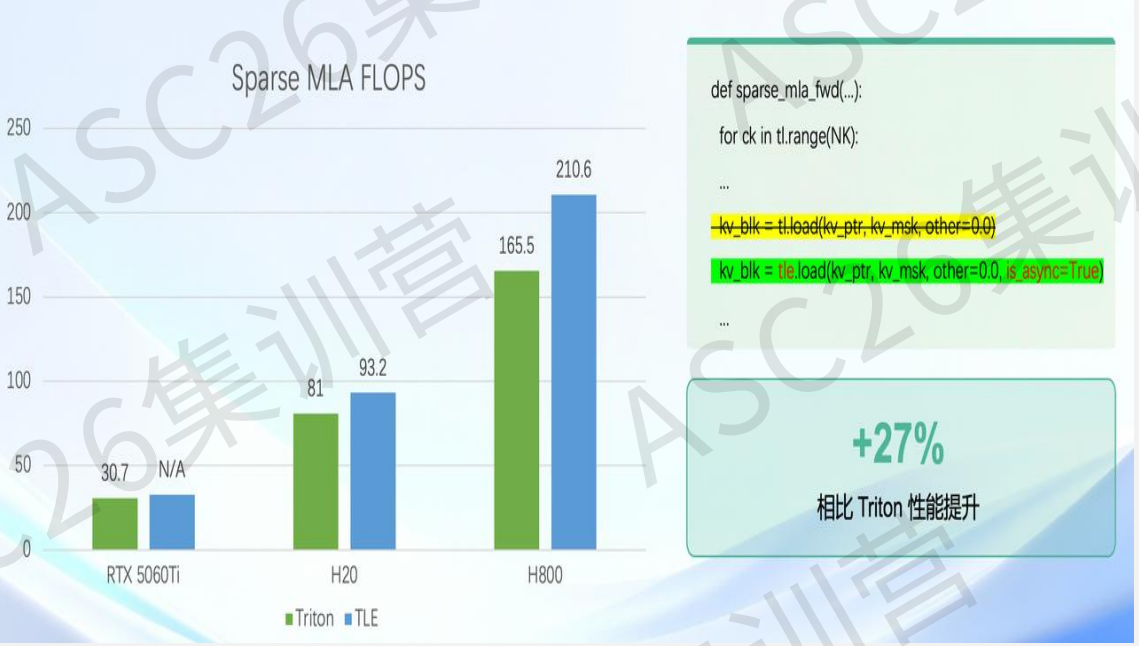
- TLE-Lite 的设计哲学是 “一次编写，到处运行” 。允许用户以最小的改动修改现有的 Triton 内核，同时兼容各种硬件后端。其可用于算法工程师的快速优化场景。
- TLE-Struct 面向的是架构感知的精细调优。它的核心理念是：根据硬件架构特征，将后端分为 GPGPU、DSA 等聚类，暴露通用的层次化并行和存储结构。这意味着，开发者可以显式定义数据布局、结构化计算映射，从而更好地发挥硬件的差异化能力。这一层主要面向算子开发工程师，帮助他们在不写底层代码的情况下，实现更贴近硬件的优化。
- TLE-Raw 的设计哲学是 “原生透传，极致掌控” 。目标用户是性能优化专家。它允许你打破 DSL 的抽象边界，直接内联厂商原生代码，比如 CUDA、MLIR 等。

TLE在DeepSeek Sparse MLA上的实践案例

Sparse MLA （Sparse Multi-Head Attention） 是 DeepSeek Sparse Attention 的核心算子，在处理长序列时具有显著的计算和存储优势。

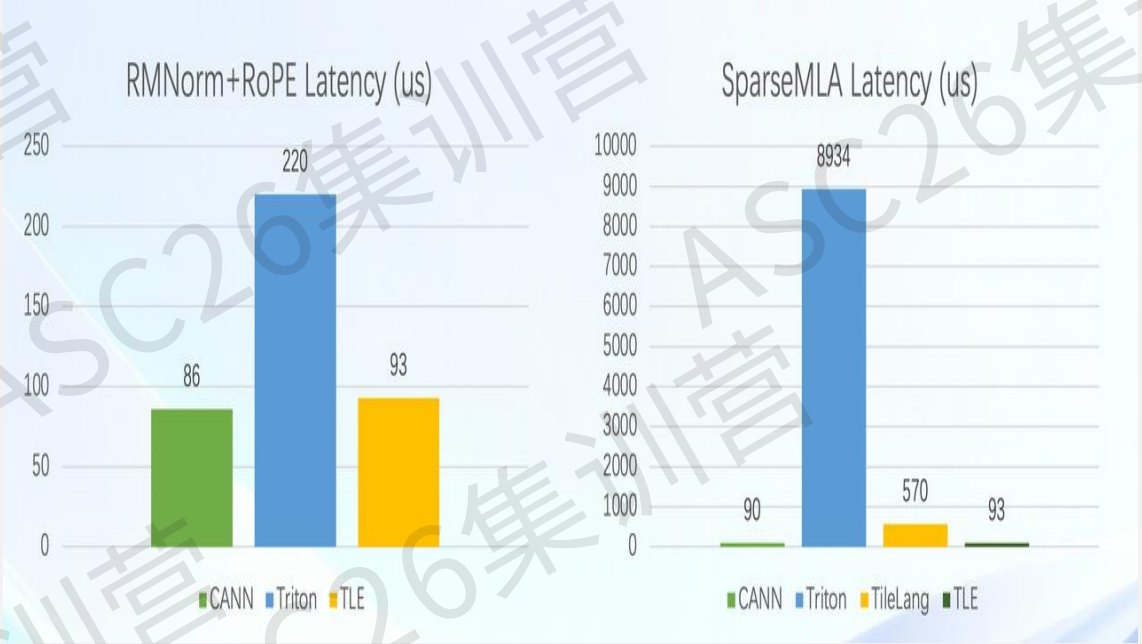
GPGPU

NV GPU上仅将 Triton Kernel 的一行代码替换为 tle.load(is_async=True) 后， 在 GPU 上取得了最高 27% 的性能提升。

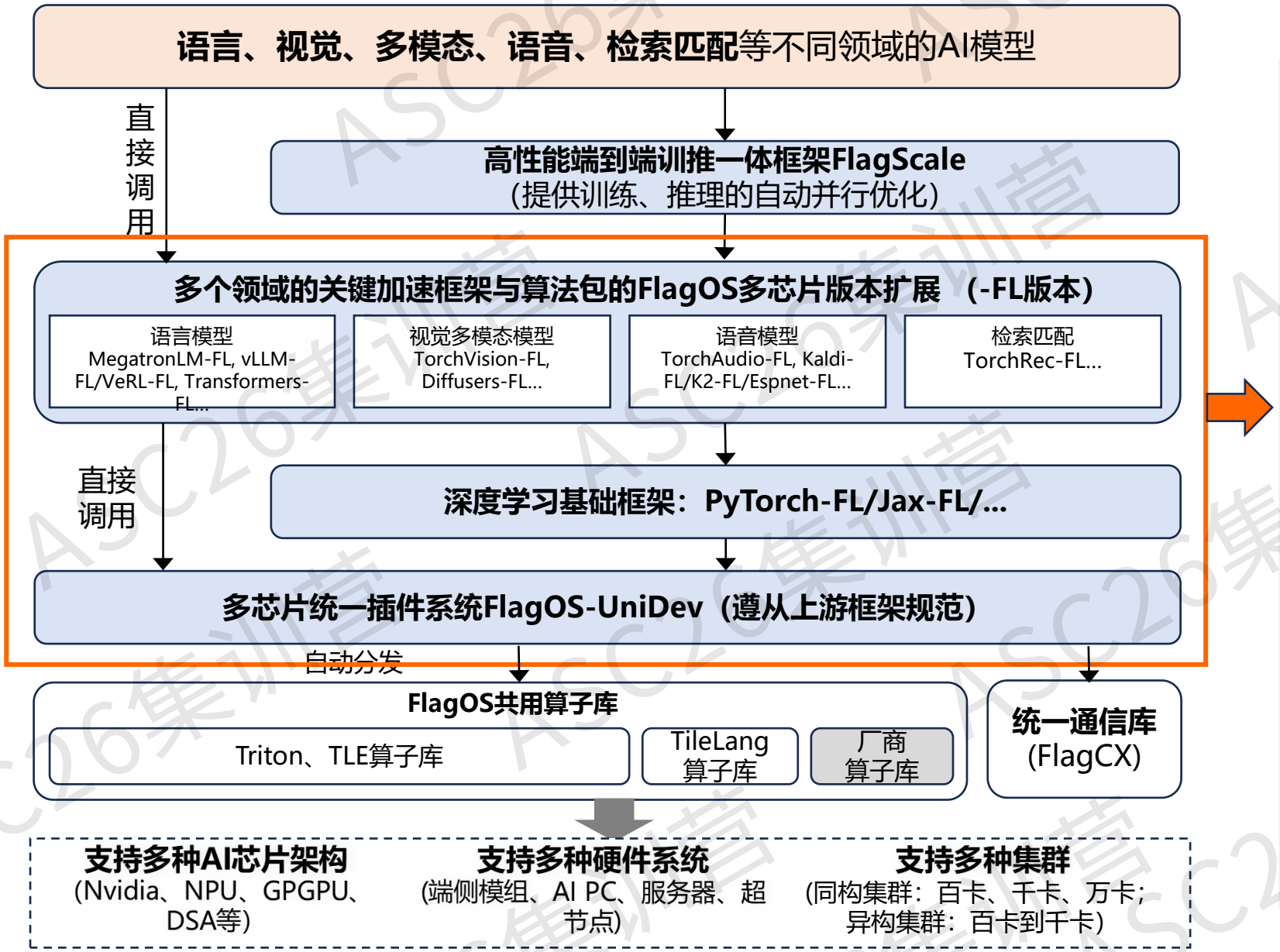


DSA

Ascend NPU上通过 TLE 优化，取得了和原生 CANN 算子媲美的性能



FlagScale框架架构：多芯片多语言的训练与推理框架体系



当前挑战

- 模型众多，所实现和执行依赖各类深度学习框架和加速框架（PyTorch、vLLM、SGLang等）
- 不同芯片上的框架版本不同，存在兼容性问题
- 存在不同语言的算子库

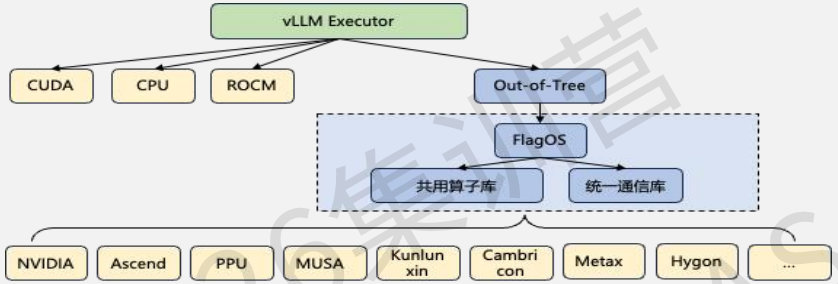
目标

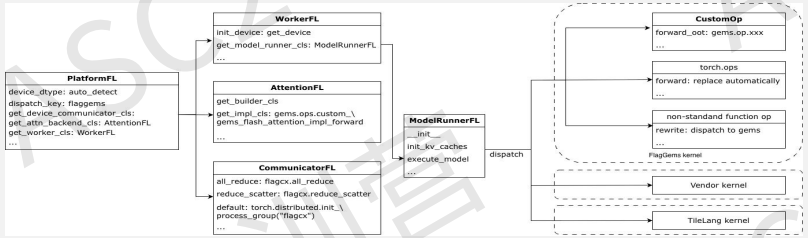
- 基于公共算子库和编译器，实现多芯片统一插件系统，自动选择不同语言算子库，支持不同的领域框架与算法包。
- 实现不同领域不同结构模型的自适应优化能力，满足不同场景需求。

基于FlagScale的统一多芯片插件系统已正在支持主流深度学习框架、训练推理加速框架、强化学习框架

vLLM-Plugin-FL推理插件

https://github.com/flagos-ai/vllm-plugin-FL

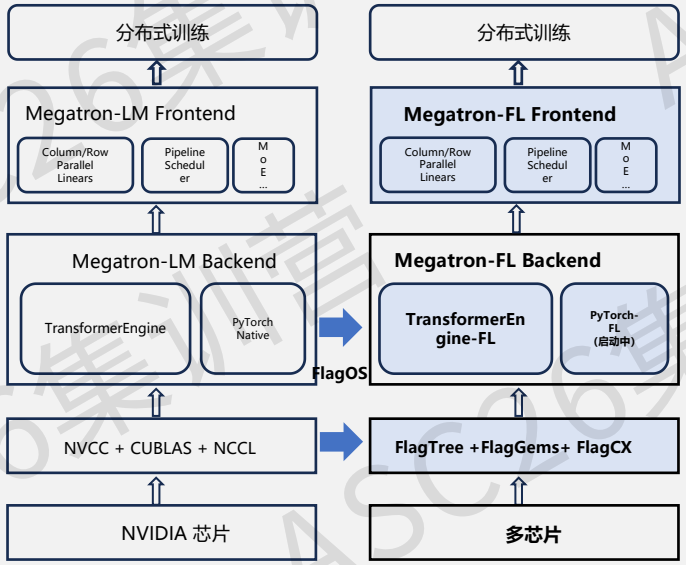




- 实现Qwen3-4B和Qwen3-32B 在单卡和两卡达到100% Triton算子覆盖，最优吞吐可达到原生的90%。
- Qwen3-Next 四卡模型并行，当前吞吐最优可达到原生的76%（优化中）。

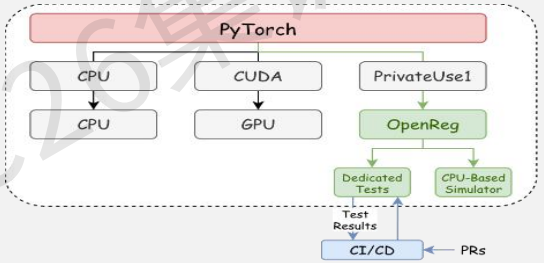
Megatron-LM-FL +TransformerEngine-FL训练插件

https://github.com/flagos-ai/Megatron-LM-FL
https://github.com/flagos-ai/TransformerEngine-FL



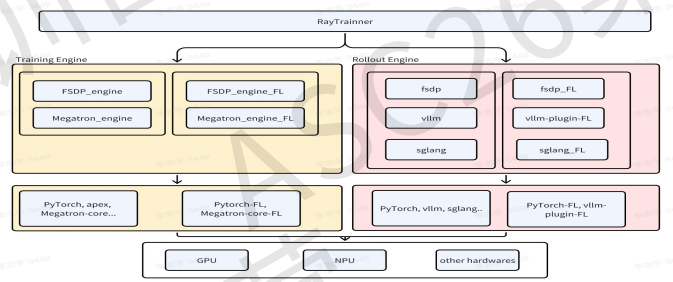
- 实现Nvidia/海光/沐曦/摩尔线程等芯片上端到端训练验证
- 在Qwen3分布式训练实现100% Triton 算子覆盖，性能为原生 80%。

PyTorch-FL 深度学习插件（进行中）



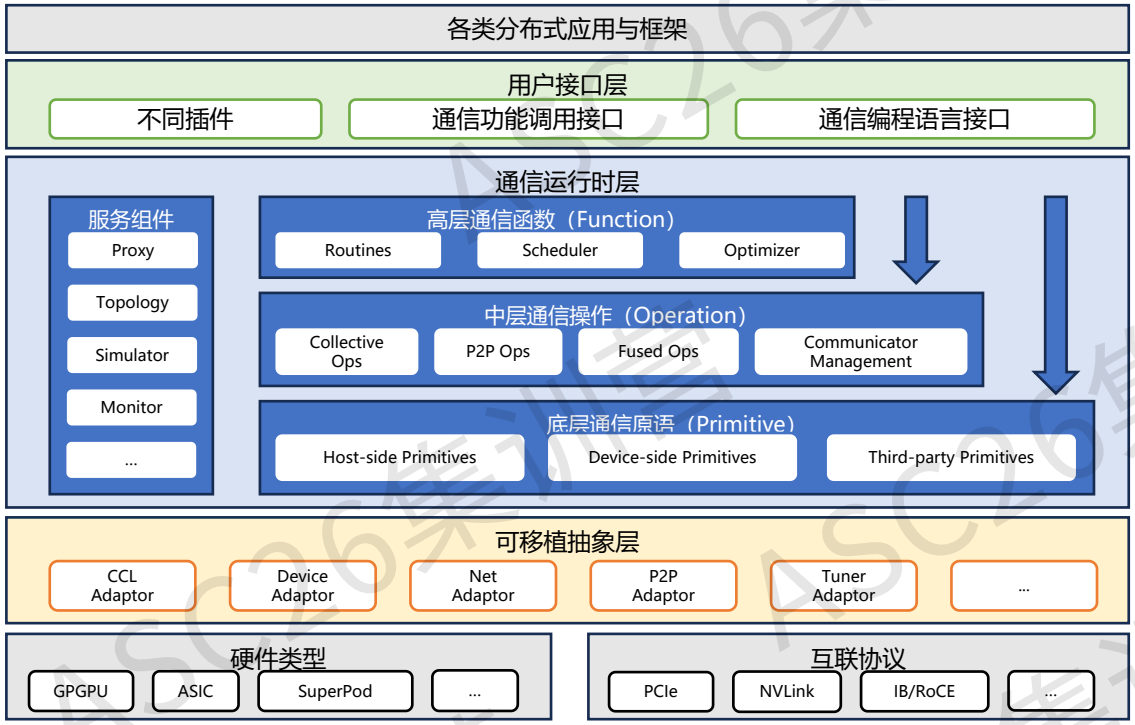
- 与PyTorch OpenReg合作打造多芯片统一插件

VeRL-FL 强化学习插件（进行中）



- 目前仅支持GPU和华为NPU，新增一款硬件需要侵入修改VeRL代码，代价高。
- 与字节合作基于Megatron-LM-FL和vllm-plugin-FL，实现多芯片的扩展，不再侵入VeRL代码。

统一通信库FlagCX：开源项目支持9种芯片5大协议，国家和ITU国际标准同时立项



FlagCX 统一通信库已经支持了9家芯片厂商

Vendor	Nvidia 英伟达		AMD 超威		Iluvatar 天数智芯		Cambricon 寒武纪		MetaX 沐曦		Kunlunxin 昆仑芯		Hygon 海光		Huawei 华为		Moore Threads 摩尔线程	
Mode	Homo	Hetero	Homo	Hetero	Homo	Hetero	Homo	Hetero	Homo	Hetero	Homo	Hetero	Homo	Hetero	Homo	Hetero	Homo	Hetero
send																		
recv																		
broadcast																		
gather																		
scatter																		
reduce																		
allreduce																		
allgather																		
reducescatter																		
alltoall																		
alltoallv																		
group ops																		

- **全场景覆盖**：全面支持集合通信，涵盖同构和异构全场景，并实现多芯片自动拓扑探测功能
- **更多芯片支持**：支持英伟达、寒武纪、昆仑芯、海光、华为昇腾、摩尔线程等9种芯片
- **更多协议支撑**：支持IBRC、IBUC 、RoCE、Socket、UCX等5种网络协议
- **双深度学习框架支持**：既支持PyTorch，也被原生集成到百度飞桨3.0正式发版中
- **国家国际标准同时立项**：国家标准《人工智能 统一通信库接口规范》正式立项（20255428-T-469）；ITU国际标《 Requirements and Framework of Cross-Platform Unified Communication Libraries for Distributed Multimedia AI Systems》已正式通过国际标准ITU-T SG21的标准立项。

关于对《人工智能 统一通信库接口规范》等167项拟立项国家标准项目公开征求意见的通知

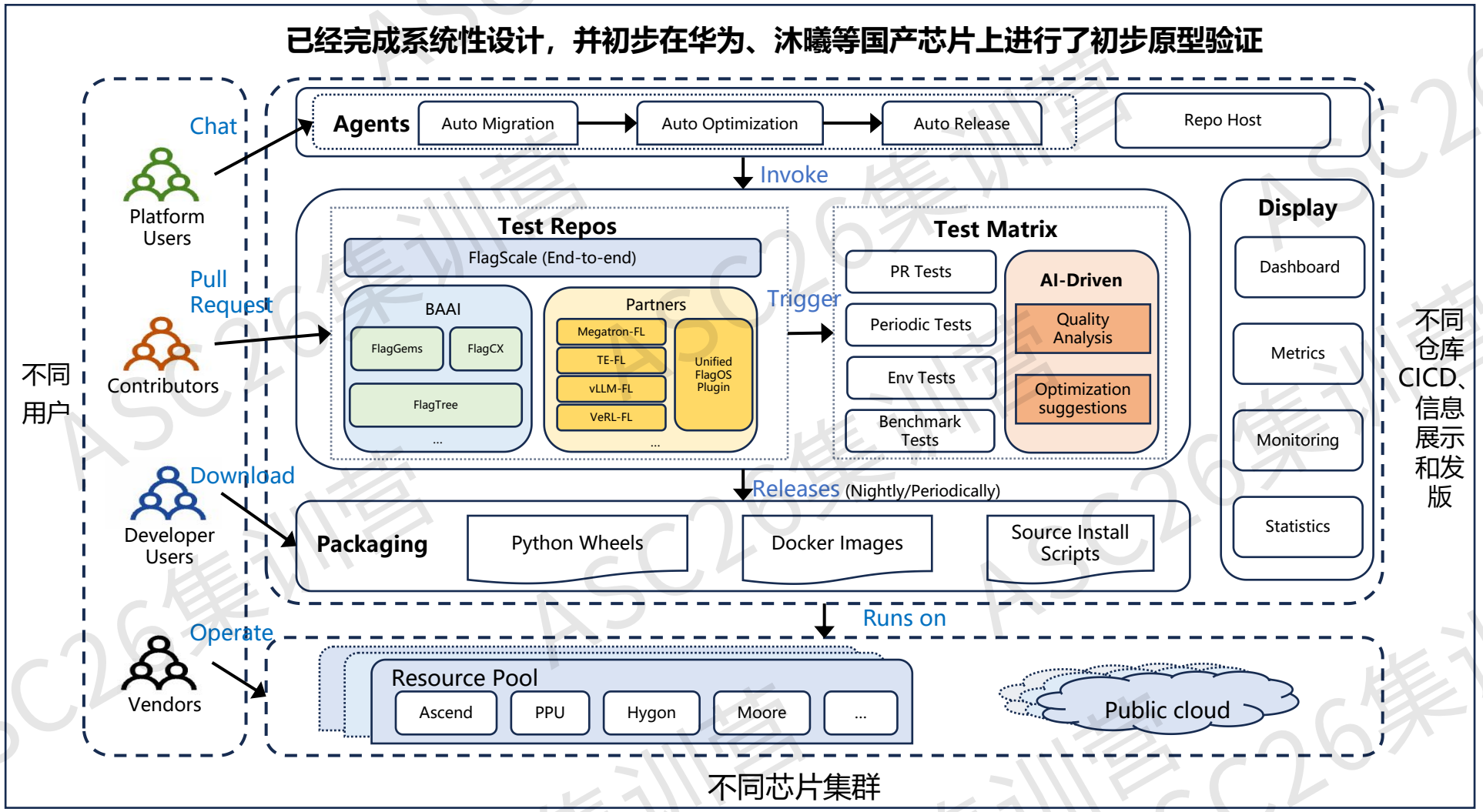
各有关单位：

经研究，现对《人工智能 统一通信库接口规范》等167项拟立项国家标准项目公开征求意见，征求意见截止时间为2025年8月27日。请登录标准技术司网站征求意见公示网页http://std.samr.gov.cn/gb/gbSuggestionPlan?bId=10002589，查询项目信息和反馈意见建议。

2025年7月28日

FlagCICD：多芯片开源项目统一CICD与发布平台

目标：为众智技术栈的多个开源项目，打造统一的多芯片适配、集成、测试、发版的工具平台，为众智技术生态的迭代发展、长期可维护建设重要的技术基础。



- **必要性与挑战**
  - 缺乏有质量保证的各种框架/库/算法包等的多芯片开源版本
  - 框架众多，版本混乱：多种框架/库/算法包多版本并存
  - 多芯片兼容性问题：不同厂商/固件/驱动不兼容性严峻，维护成本高
- **关键技术与创新点**
  - 实现跨芯片跨集群的标准化CICD流程
  - 实现AI驱动代码自动迁移和质量保证等
  - 提供在线自动化服务能力，服务各个领域的框架/库/算法包等

FlagRelease：开源大模型跨芯自动迁移和版本发布，性能对齐

目标：依托FlagOS技术栈，构建了一套自动化迁移开源大模型至不同AI芯片的工具，并自动发布。通过FlagOS和智能体的相互支持，迁移和发布效率提升**4倍**，加速模型在多架构环境下的落地效率与生态成熟度

源源不断产生多芯片模型，加快追赶生态差距

下载: 基于CUDA的原始模型

ModelScope & Hugging Face

发布：迁移后的模型、代码及适配不同芯片的Docker 文件

迁移任务识别
- 选择目标芯片
- 下载选定模型

AI Agent 智能分析与修复
- 系统环境预校验
- 框架&算子库依赖分析
- AI Agent 智能修复

模型迁移到目标硬件
- 目标硬件运行CI/CD
- 自动调优获取最优策略
- 自动模型服务部署

迁移后的模型评估
- 自动评估效果
- 监控系统性能
- 指标统计分析

迁移后的多芯片模型发布
- 用户文档
- 评估结果
- 代码镜像

硬件集群（搭载各种AI芯片）

URL: https://modelscope.cn/organization/FlagRelease https://huggingface.co/FlagRelease/models?p=1

已支持的开源模型	已支持的芯片
GLM4.5系列	Nvidia
Qwen2&3系列	Huawei
Kimi-K2、phi-4、	ARM
grok-2、step3	Metax
MiniMax-M1-80k	Iluvatar
MiniCPM-v4	Hygon
Seed-OSS-36B	Cambricon
gpt-oss-120b	Kunlunxin
ERNIE-4.5-300B-A47B-PT	
RoboBrain2.0-7B/32B	
MiniCPM-o-2_6-8B	
Deepseek-R1	

一站式提供最新模型、多种芯片的统一版本，用户开箱即用，三步即可完成安装

KernelGen： 高性能Triton算子自动生成

KernelGen v1.0 是一个面向高性能 Triton 算子生成的自动化平台，从“只会写代码助手”升级为覆盖算子生成、基线构建、验证测试与多芯片适配的完整生命周期系统，实现一次描述、自动生成、自动评测、多芯片自动适配。

项目地址：https://kernelgen.flagos.io/home

KernelGen v1.0 - 高性能Triton算子自动生成平台

四大核心能力

- Triton算子自动生成
- 基线自动构建
- 自动化验证与测试
- 多芯片后端适配验证

工作流程

自然语言描述 → 算子生成&基线构建 → 测试反馈迭代优化 → 多芯片评估验证

多芯片验证平台

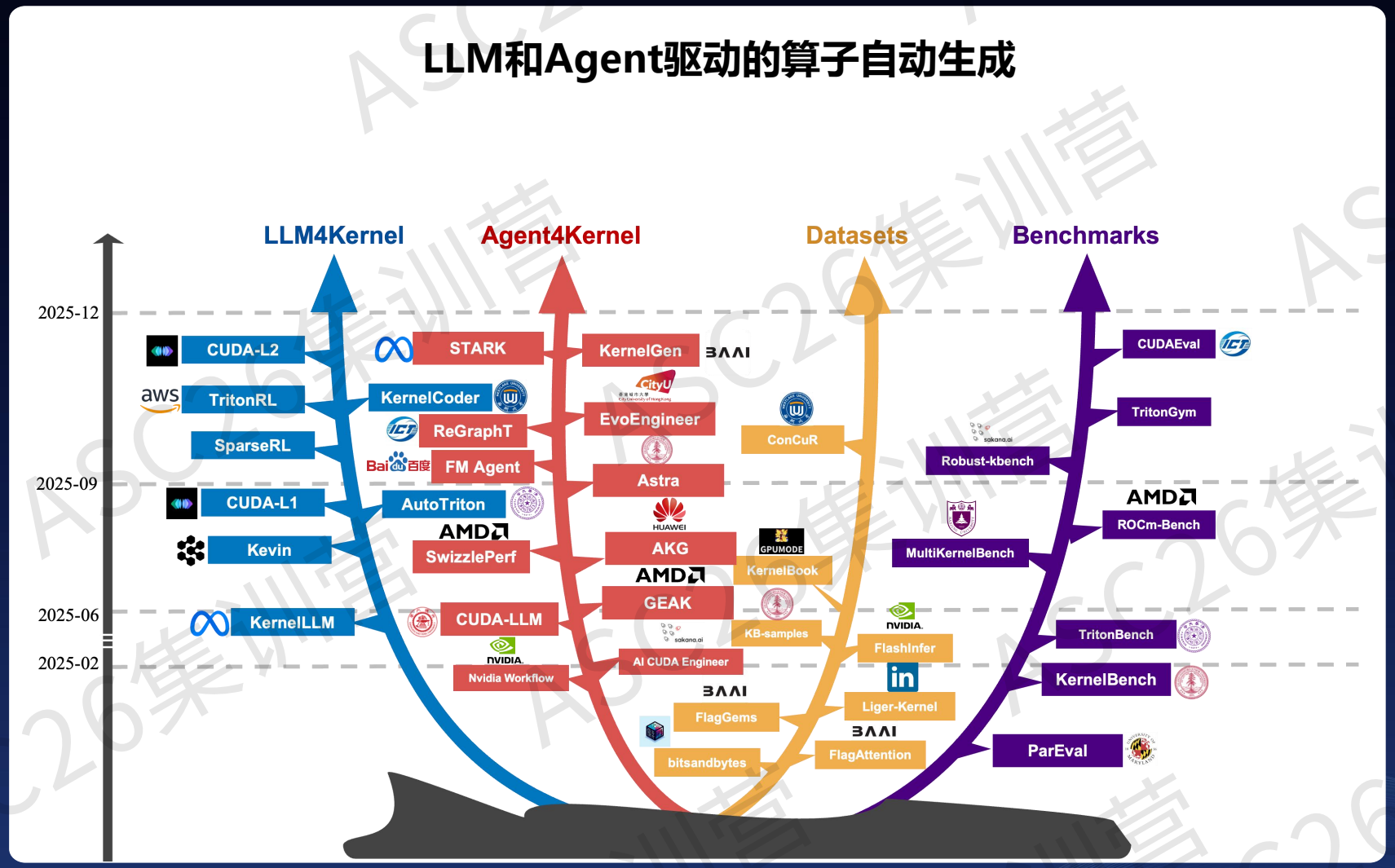
华为　海光　摩尔　天数　N卡

- 200+算子开发，2年->3小时
- 自动生成的Triton算子中，**50%**的算子性能优于CUDA原生算子

进一步优化：

- 丰富知识库，尤其针对专家级经验、各种芯片优化经验的引入
- 强化学习，与FlagTree编译器优化信息的协同等

2025-2026 算子自动生成技术快速发展



KernelGen 定位

核心能力

AI 驱动的 Triton 算子自动生成与验证平台

技术栈位置

连接 AI 模型与 GPU 硬件的关键中间层

生态价值

降低算子开发门槛，加速国产芯片适配

战略意义

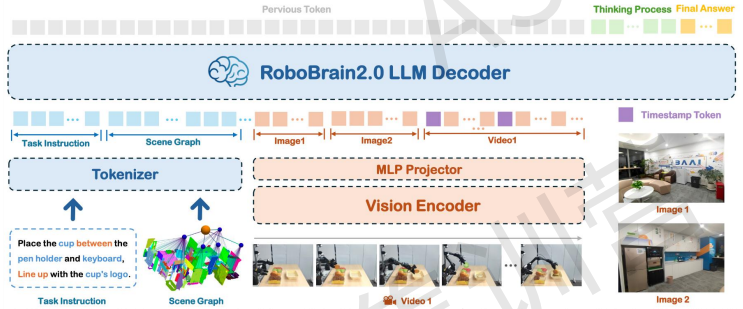
在 AI 推理能力飞跃的 2025 年，KernelGen 将 IMO 金牌级的验证能力应用到算子开发领域。

GitHub 传送门： https://github.com/flagos-ai/awesome-LLM-driven-kernel-generation

FlagOS支持具身智能全链路训练推理：实现跨芯片、提效

VLM大脑模型

已支持RoboBrain 1.0/2.0训推

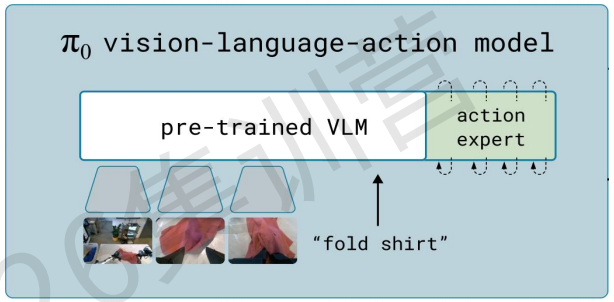


分布式训练：
- 通过非均匀流水线并行实现高效并行
- 基于Energon 实现高效分布式数据加载
- 利用显存分配预处理消除碎片
- 精细化重计算显著降低显存占用
- 端到端性能相比Llama-factory **提升154.81%**

高效推理：
- 多后端支持
- 部署参数自动调优
- 模型量化（WA）加速
- 端到端推理性能 **提升22%**

VLA端到端模型

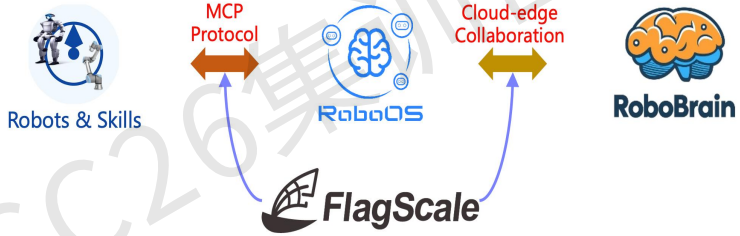
已支持RoboBrain-X0和π_0训推



分布式训练：
- 预训练/后训练数据加载统一
- 后训练通过冻结VLM支持
- 基于Energon 实现高效分布式数据加载

高效推理：
- 多后端支持
- 部署参数自动调优
- 模型量化（W8A16）加速
- 端到端推理性能 **提升23%**
- 端侧部署支持
- 真机系统代理

高效端云协同



机器人技能注册标准化：
- 典型场景输入 token **减少 65%**
- 自动技能注册和技能商店构建

端云协同的快速通信方案：
- 平均延迟降低至 **<3ms**
- 支持全流程历史数据随机访问

技能检索功能：
- 总 Token 用量 **降低 29.8%**

国产端侧支持

快速适配

天数端侧模组

FlagOS支持“2025中关村具身智能大赛”，为选手提供跨平台的训练、推理能力

众智FlagOS 立足中国，构筑全球生态

聚拢生态合作企业和机构

- **芯片企业**：寒武纪、华为等**17家**
- **服务器企业**：浪潮、新华三等**5家**
- **AI软件企业**：硅基流动、中科加禾等**7家**
- **高校和科研机构**：清华、北大、计算所、先进编译实验室等**5家**
- **集成和应用企业**：中国移动、中国联通、天翼云、软通动力、东华、中科软等**16家**
- **大模型企业**：百度、科大讯飞、面壁智能等**3家**
- **操作系统**：麒麟软件、龙蜥、OpenCloudOS、Circle Linux等**4家**，国际层面RedHat正在洽谈
- **行业和开源组织**：电子标准化研究院、CSDN等**4家**

全球合作，拓展国际技术生态

- 通用算子库成为PyTorch基金会正式生态合作项目
- 统一通信库ITU国际标准正式立项

高校人才培养

- 基于FlagOS，中国科技大学、中科院计算所、北航、北邮等高校开展课程培训，累计覆盖超过**13000学生**
- 启动了首批包括北大、清华、中科院等**10所**重点高校的“自主软硬件生态技术 高校人才培养计划” 同启动

生态合作平台

北京人工智能公共算力平台

Modelers

Hugging Face

ModelScope

CSDN

FlagOS 全球开发者吸引

- 当前全球触达开发者近**2万人**，开源贡献者**400多人**
- 启动 “FlagOS开放计算全球大赛”，从算子、到大模型和具身智能四大赛道

技术攻坚，打破AI芯片生态壁垒
众智成城，共筑开放计算的创新蓝图

FlagOS
公众号

FlagOS
官网

FlagOS
GitHub