

题目编号：XH-202628

基于 AI Agent 开发范式的国产 GPU 大模型推理算子库优化比赛方案

一、发榜单位

企业名称：沐曦集成电路（上海）股份有限公司

企业类型：民营企业

企业地址：上海市浦东新区海科路 999 弄 8 栋 C8 栋。

二、题目名称

基于 AI Agent 开发范式的国产 GPU 大模型推理算子库优化

三、题目介绍

1. 题目背景：

大模型进入规模化应用阶段后，推理性能已经成为影响模型服务成本、响应速度和产业落地能力的关键因素。相比训练阶段，推理阶段具有调用频次高、并发压力大、上下文长度持续增长、部署环境复杂等特点，对底层 GPU 软件栈和推理算子库提出了更高要求。

在大模型推理链路中，Attention、FlashAttention、KV Cache、GEMM、算子融合、量化计算、长上下文处理等底层能力，直接决定了模型推理的吞吐、延迟、显存占用和服务稳定性。以 FlashInfer、FlashAttention 等为代表的推理算

子库，已经成为提升大模型推理效率的重要基础软件。

与此同时，AI Coding 与 Agent 技术正在改变软件开发方式。过去依赖专家经验和人工反复调参的底层优化工作，已经出现新的开发范式：通过 Agent 完成代码理解、性能分析、候选实现生成、自动 Benchmark、错误修复、参数搜索和迭代优化，这使得 Agent 能力不仅是 Coding Agent，而能深入 GPU 基础软件优化过程。

2. 目标介绍：

面向国产 GPU 的大模型推理，参赛者需要使用/构建 AI Agent，通过自动代码理解、算子迁移、性能分析、Kernel 优化、Benchmark 验证等方式，优化 FlashInfer、MCTLASS 等推理算子库在国产 GPU/MXMACA 软件栈上的性能表现。

以 Agent/Skill 驱动 FlashInfer、MCTLASS 等大模型推理算子库在国产 GPU 上的自动迁移、自动优化、自动验证与性能迭代，用 Agent 推动算子优化效率，用 Skill 沉淀优化经验，用国产 GPU 建立推理生态壁垒。

3. 选题意义：

本赛题围绕国产 GPU 上的大模型推理算子库优化，实现“Agent+Skill+Benchmark+国产 GPU”的新型技术路线，推动国产 GPU 推理生态走向“推理高性能、优化可复现、经验可沉淀、生态可持续”。

四、参赛对象

参赛对象为 2026 年 6 月 1 日以前正式注册的国内全日制非成人教育的普通高等学校在校专科生、本科生、硕士和博士研究生（不含在职研究生），以及全日制职业教育本科、高职高专在校学生，可通过学生赛道申报作品参赛。

参赛对象可以团队或个人形式参赛，每个团队不超过 10 人，每件作品可由不超过 3 名指导教师进行指导。可以跨专业、跨学校、跨地域组队，但同一团队所有成员均应符合本赛道相关年龄、身份要求。每件作品只可由 1 所高等院校、科研院所等作为参赛主体提交申报。

五、答题要求

本赛题面向国产 GPU 大模型推理场景，鼓励参赛团队构建或使用 AI Agent/Skill 工作流，对 FlashInfer、FlashAttention、MCTLASS（本次聚焦 Fused MOE 算子）等大模型推理算子库开展迁移适配、性能分析、Kernel 优化、自动调优、算子融合和 Benchmark 验证。

本赛题要求参赛团队将 Agent 真正用于底层算子优化过程，使其能够参与并辅助完成代码理解、性能分析、优化生成、测试验证和多轮迭代。参赛作品应体现“Agent 驱动算子优化”的技术特征，而不是仅停留在概念说明、文档生成或简单代码补全层面。

参赛团队需围绕国产 GPU 平台上的大模型推理算子库优化，完成一类或多类任务。鼓励参赛团队在保证正确性、

稳定性和可复现性的前提下，提升相关算子或端到端推理链路在国产 GPU 上的执行性能。

本赛题鼓励参赛团队使用各类 AI Agent、Skill、AI Coding、智能体开发平台和自动化调优工具参与推理算子库优化过程，参赛团队可使用 Claude Code、OpenCode、OpenAI Codex、OpenClaw、CodeBuddy、Cursor 等开发工具，也可使用自研 Agent/Skill 工作流。

通过本赛题，期望形成一批面向国产 GPU 的大模型推理算子优化成果、Agent 优化工作流、Skill 模块、Benchmark 方法和工程实践案例，为国产 GPU 大模型推理生态建设提供支撑。

任务：算子性能调优 Agent/Skill

内容	说明
迁移 FlashInfer 的关键算子到 MACA 平台，并优化性能	版本：FlashInfer-ai/FlashInfer 0.2.8 API: BatchPrefillWithRaggedKVCacheWrapper - headdim64/128/256, qk192+vo128 for mla BatchPrefillWithPagedKVCacheWrapper - headdim64/128/256 BatchMLAPagedAttentionWrapper

内容	说明
	<p>- headdim qk576+vo512</p> <p>BatchDecodeWithPagedKVCacheWrapper</p> <p>- headdim64/128/256</p> <p>语言：MACA/C++</p> <p>数据类型：BF16</p> <p>page size: [1, 16]</p> <p>seqlen: 1K~180K，选自真实大模型，且 seqlen 取值随机以增加泛化性</p>
<p>迁移 FlashAttention 的关键算子到 MACA 平台，并优化性能</p>	<p>版本：Dao-AILab/FlashAttention 2.6.3</p> <p>API: flash_attn_with_kvcache</p> <p>语言：MACA/C++</p> <p>数据类型：BF16</p> <p>headdim: [32, 64, 96, 128, 160, 192, 224, 256, 512]，其中 128、256、512 高优</p> <p>page size: 16</p> <p>seqlen: 1K~180K，选自真实大模型，且 seqlen 取值随机以增加泛化性</p>
<p>迁移 Fused MOE 算子到 MACA 平台，</p>	<p>API: Fused MOE</p> <p>语言：Tilelang/Triton/MACA C;</p>

内容	说明
并优化性能	精度：INT8 W8A8； shape: n_tokens, n_experts, topK, N, K 均选自真实大模型，在典型切分，典型 seqlen, batchsize 的真实取值

提交可复现的算子优化的源码、测试及测试框架、性能测试脚本、性能报告、Agent/Skill，以及对应的 PPT 和文档。

六、作品评选标准

本赛题采用“自动评测打榜 + 专家综合评审 + 终审擂台赛答辩”相结合的评价机制。自动评测结果作为作品性能表现的重要依据，专家评审重点考察作品的技术路线、Agent 参与深度、工程质量、创新性和可复现性。

（一）打榜机制

本赛题设置阶段性排行榜和最终排行榜。参赛团队可在规定时间内多次提交作品，系统根据统一评测环境、统一数据集、统一 Benchmark 脚本和统一评分规则进行自动评测，并生成榜单排名。

1. 榜单设置

主办方设置以下榜单：

榜单名称	评价重点	适用作品
Agent 自动优化算子	单个算子或算子库在指定输入规模下	FlashInfer、FlashAttention、MCTLASS 等算子优化作品

榜单名称	评价重点	适用作品
性能榜	的性能提升，也包含 Agent/Skill 在优化流程中的参与深度和自动化能力	（包含了 Agent 工具链、AutoTune、Benchmark 自动化作品）

Agent 自动优化算子性能榜依据客观性能指标排名（后续将根据提交的 Agent 流程核对复现）。

2. 自动评测流程

每次打榜提交后，系统按照以下流程进行评测：

作品提交 → 环境检查 → 编译构建 → 正确性测试 → Benchmark 性能评测 → 稳定性与回归检查 → 生成评测报告 → 更新排行榜。

只有通过正确性测试和基础稳定性测试的作品，才进入性能排名。未通过正确性测试的作品不计入性能榜。

3. 评测指标

赛题将提供对应算子的自动化测试程序和目标测试集。目标测试集中的测试用例来源于真实公开大模型或主流大模型推理负载，例如 DeepSeek-V4、GLM-5.1、Kimi-K2.5、Step-3.5-Flash、MiniMax M2.5/M2.7、MiMo-V2 等模型，在典型序列长度（sequence length，简称 seqlen）、批大小（batch size）、head 数、head dimension、数据类型等条件下的实际输入形态。

测试程序将自动评测参赛作品的正确性与综合性能。其中，正确性为必须达成的基础目标，包括功能正确性、数值精度达标、边界条件通过和运行流程完整。未通过正确性测试的作品不进入性能排名。

在正确性通过的前提下，测试程序将自动统计作品在目标测试集上的综合性能表现，包括延迟、吞吐、显存占用、稳定性以及不同输入规模下的覆盖情况，并按照预先公布的权重计算加权平均性能得分。该得分作为算子性能榜和相关任务排名的主要依据。

4. 复现与反作弊要求

为保证榜单结果真实有效，参赛作品需满足以下要求：

1. 作品必须能够在主办方指定环境中复现；
2. 不得通过识别测试样例、硬编码输出、跳过计算等方式刷榜；
3. 不得牺牲计算正确性换取性能提升；
4. 不得利用未授权接口、系统漏洞或评测环境漏洞获取不当成绩；
5. 主办方有权对异常成绩进行复测、要求补充说明或取消排名。

如参赛作品在复现过程中无法稳定达到提交成绩，主办方可根据复测结果调整其榜单成绩。

（二）作品评选标准

提交作品最终评价采用 100 分制，权重计划如下：

类别	评审维度	权重	说明
客观 评测	性能提升效果	60%	在保证准确性和稳定性及精度等的前提下，相比基线版本，在延迟、吞吐、Token/s、显存占用等方面取得的提升
	Agent/Skill 的可复现性	20%	<p>Agent 真实参与源码理解、代码生成、性能分析、自动调优、Benchmark 和多轮迭代，按照可复现程度打分：</p> <p>功能能复现：5 分</p> <p>性能复现达提交标称的 60%以上：10 分</p> <p>性能复现达提交标称的 80%以上：15 分</p> <p>性能复现达提交标称的 90%以上：20 分</p>
客观 评测+主 观评测	文档说明与 演示报告	20%	技术报告、README、运行说明、演示视频、答辩材料是否清晰完整，按照材料完备和质量程度打分。

七、作品提交时间

2026 年 5 月至 9 月上旬，各参赛团队选择榜单中的题目开展研发攻关，各高校、科研机构等组织协调机构应组织学生和青年科技工作者参赛，安排专业人员给予指导，为参赛团队提供支持保障。

2026 年 9 月 5 日前，各参赛团队要向发榜单位完成作品提交，具体要求详见作品提交方式。

2026 年 9 月 20 日前，由发榜单位完成初审，确定入围终审擂台赛的晋级作品和团队。

2026 年 10 月，安排专门团队提供帮助和指导，各晋级团队完善作品。

2026 年 11 月，组织终审擂台赛，角逐“擂主”。

八、参赛报名及作品提交方式

（一）报名方式

（1）参赛选手登录“挑战杯”官网 www.tiaozhanbei.net，在“揭榜挂帅”擂台赛报名入口注册账号，登录大赛申报系统在线填写报名信息。报名信息提交后，下载打印系统生成的报名表。

（2）申报人在报名表对应位置加盖所在学校或所在单位公章。

（3）将盖章版报名表扫描件上传至报名系统，等待系统审核。请参赛选手注意查看审核状态，如审核不通过，需重新提交。

(4) 系统开放报名时间为 2026 年 5 月 30 日—6 月 30 日，逾期后系统将自动关闭报名功能。

(二) 作品提交方式

作品的提交除提到的客观评测外，在初赛/总决赛截止时间前，参赛团队应将所有要求的材料，包括可复现的算子优化的源码、测试及测试框架、性能测试脚本、性能报告、Agent/Skill，以及对应的 PPT 和文档等统一打包压缩提交至邮箱：

opensource@metax-tech.com。压缩包命名方式为：申报人所在单位-申报人姓名-作品名称-联系电话（例如：XX 大学-张XX-XX 方案-手机号）。

除参赛报名表外，各参赛组提交的文档、源代码和相关文件不得携带任何参赛学校、老师和学生的个人信息。同时，各参赛团队在提交作品时，同步报送 1 份经报名系统审核通过的参赛报名表，报名表所有信息须与系统内填报内容完全一致。

九、赛事保障

1. 算力资源支持：参赛选手在完成报名后，提供对应线上的曦云 C500 等在线算力资源券。

2. 技术培训：赛前赛中至少组织 2 场线上培训，提供培训回放及答疑文档；另外届时会根据实际情况，决定是否组织线下的专场培训。

3. 专家指导：建立线上答疑社群，由沐曦股份技术团队指导，定期回复技术问题。

4. 交流平台：在沐曦股份开发者社区开设赛事专属社区板块，支持参赛团队分享经验、交流问题，促进技术共创。

5. 技术文档和课程

十、设奖情况及奖励措施

1. 设奖情况

按参赛作品数量比例设奖，原则上评出“擂主”1个、特等奖5个，一等奖5个、二等奖6个、三等奖8个，获奖比例不超过参赛作品总数的30%，从特等奖中角逐出擂主团队。

2. 奖励措施

奖金奖励：擂主奖励10万元/个（叠加特等奖后奖金），特等奖奖励2万/个，一等奖奖励1万元/个，二等奖奖励0.5万元/个，三等奖奖励0.2万元/个。以上奖金均为税后金额。

高端GPU奖励：擂主团队2张GPU加速卡（不叠加特等奖GPU奖励），特等奖团队1张GPU加速卡。

实习奖励：优秀获奖者有机会参与之江&沐曦股份“南湖之新”联合培养计划；所有获奖团队可获得赛事荣誉证书。

曝光支持：获奖作品可在沐曦股份开发者社区展示，作品可提供成果孵化与应用推广支持。

备注：从特等奖中角逐出擂主团队，奖金10万元已叠加特等奖奖金，擂主不叠加特等奖GPU奖励。

3. 奖金发放方式

比赛结束后，单位比赛专班工作人员与获奖团队取得联系，填写奖金申请表，赛事终审结果公示无异议后（公示期约1个

月)，在 30 个工作日内通过银行转账一次性发放至团队负责人指定账户，上述所列奖金均为税后奖金。

十一、比赛专班联系方式

1. 专家指导团队

顾问专家：董老师，联系电话：13482748718

顾问专家：武老师，联系电话：18951640525

顾问专家：韩老师，联系电话：18017474835

负责比赛期间技术指导保障。

2. 赛事服务团队

联络专员：杨老师，联系电话：15201842467

联络专员：章老师，联系电话：13501701786

负责比赛期间组织服务及后期相关赛务协调联络。

3. 联系时间

比赛期间工作日（9:00－17:00）

附：发榜单位简介

沐曦集成电路(上海)股份有限公司(股票代码:688802.SH)成立于2020年9月,于2025年12月成功登陆科创板。总部位于上海,并在北京、南京、成都、杭州、深圳、武汉、长沙等地设立全资子公司及研发中心。作为一家专注于全栈GPU芯片及解决方案的集成电路设计企业,致力于打造世界一流的GPU芯片及计算平台,成为数字经济的算力基石。

公司拥有技术完备、设计和产业化经验丰富的团队,核心成员平均拥有近20年高性能GPU产品端到端研发经验,曾主导过十多款世界主流高性能GPU产品研发及量产。目前,公司已推出全面覆盖人工智能训练和推理、通用计算、图形渲染和科学智能等场景的四大序列产品,并配套自研MXMACA软件栈,真正实现了“软硬协同”,满足“高能效”和“高通用性”的算力需求。2025年起,公司坚持“开放协同、自主可控”的方向,全面推进计算生态以及产业生态的建设。在计算生态层面,公司以自主研发的MXMACA全栈软件栈为核心,积极拥抱开源,打造自主、开放、兼容的通用计算开源生态;在产业生态层面,公司依托“1+6+X”战略布局,以数字算力底座为基,持续深耕金融、医疗健康、能源、教科研、交通、大文娱六大重点行业,同步积极探索具身智能、低空经济等新兴领域,为数字经济与新质生产力发展提供坚实算力支撑。