

题目编号：SH-05

# “推理大模型的训练调优与性能加速助力全栈自主 AI” 比赛方案

## 一、发榜单位

华为技术有限公司

## 二、题目名称

推理大模型的训练调优与性能加速助力全栈自主 AI

## 三、题目介绍

在人工智能的时代浪潮中，各类大模型的发展突飞猛进，加速各行业应用创新。进入 2025 年，大语言模型（LLMs）快速迭代与进化，特别是在推理能力和运行效率方面有了显著提升。推理能力上，以 OpenAI o1 系列和 DeepSeek R1 为代表的推理模型，带来了 scaling law 范式的转移(inference-time scaling)，非常擅长解决如解谜、高级数学和编码等复杂任务，将大模型推向更广泛的应用场景；运行效率上，以量化、MOE、知识蒸馏为代表的模型优化技术，使得大模型可以部署到更轻量化的设备上，给端侧 AI 带来更大的想象空间。

展开来看，DeepSeek 证明了在没有思维链(Chain-of-Thought, CoT) 监督数据的情况下，只通过强化学习也可以显著增强大语言模型的推理能力；而且，通过知识蒸馏，使用推理大模型产

生的思维链数据和少量通用数据进行微调，小模型在资源受限的场景中也能实现复杂推理任务；此外，在参数量、算力需求指数倍增长情况下，如何充分释放 AI 芯片算力、提升大模型性价比也成为 AI 领域的关键技术，这其中高性能算子的实现与优化是达成此目标的基础。

因此，本次比赛聚焦大模型的推理能力提升和性能优化，要求选手基于华为全栈 AI 技术，使用强化学习、知识蒸馏等技术，提升选定的轻量级模型在数学计算、逻辑推理、代码生成等任务上的能力，并通过算子融合与优化、模型结构优化、量化等技术，保证精度的同时提升模型推理性能。

1. 本次比赛基于昇腾 AI 云服务，要求参赛团队使用 AI 芯片 Ascend-snt9b、AI 异构计算框架 CANN、AI 应用开发平台 ModelArts 等华为全栈 AI 技术，增强模型推理能力并提升性能。

2. 参赛团队需要掌握华为昇腾 AI 处理器架构、CANN 软件栈与 ModelArts 开发环境，实现模型训练与推理调优，大赛将提供昇腾云模型开发、昇腾算子开发相关的技术文档及课程材料，助力参赛团队学习相关技术，了解实践操作。

3. 大赛鼓励选手将优化后的模型部署到开发板、手机等端侧设备上，以真实的产业需求为导向，探索推理模型在行业中的应用。

#### **四、参赛对象**

本题目只设学生赛道。

参赛对象为 2025 年 6 月 1 日以前正式注册的全日制非成人教育的各类高等院校在校专科生、本科生、硕士研究生、博士研究生（不含在职研究生），参赛人员年龄在 40 周岁以下，即 1985 年 6 月 1 日（含）以后出生。

同一作品不得同时参加第十九届“挑战杯”全国大学生课外学术科技作品竞赛（以下简称第十九届“挑战杯”竞赛）其他赛道的评比。

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、跨单位、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 所高等院校作为参赛主体提交申报。

## **五、答题要求**

**整体赛程分为初赛和总决赛，作品具体要求如下：**

**1. 初赛：**组委会将提供 Qwen 轻量级模型与模型推理测试的示例工程，参赛团队自行选择强化学习算法（如 GRPO、PPO 等），或者知识蒸馏技术，在 ModelArts 的 notebook 开发环境（规格为 Ascend: 1\*ascend-snt9b(32G)|ARM: 24 核 192GB）中对模型全部参数或部分参数进行微调。

比赛也考察参赛团队对大语言模型推理性能的调优能力，选手可以自行选择优化手段，如模型结构优化、算子融合、算子性能优化、量化等，在保证模型精度同时，尽可能提升推理

性能。算子优化可以往小算子融合的方向设计（例如 add+layernorm 算子融合、matmul+add 融合等）。

组委会将在5月下旬发布初赛A榜测试数据集与 ModelArts 模型推理的 Notebook 样例，供参赛团队下载参考。参赛团队使用微调后的模型自行对测试集进行推理，并将推理结果按如下格式保存为单个 JSON 文件进行提交：

```
{"result": {  
  "results": [  
    {"id": " id1", "content": "xxx"},  
    {"id": " id2", "content": "xxx"}],  
  }  
}  
...
```

其中，id 为测试问题 id，content 为请求返回的结果，可以参考 Notebook 样例进行输出。

7月下旬组委会将开放初赛B榜，测试集不公开，但提供推理模型包构建样例，参赛团队需要参考样例开发推理应用，提交适配 ModelArts 规范的推理模型包。测试集为针对该模型设计的多组输入数据，参赛团队提交的模型包需要以测试数据为单位返回请求结果与运行耗时，单个测试用例可能包括多个问题，选手以单个测试用例为单位输出一个结果，格式为 JSON 字符串，举例如下：

```
{"result":{"results":[{"id": id1, "content": "xxx"}, {"id": id2, "content": "xxx"}], "duration": 41.72}}
```

其中, `id` 为测试问题 `id`, `content` 为请求返回的结果, `duration` 为整条测试用例的推理耗时（单位为 `ms`），可以参考模型包样例进行输出。

**2. 总决赛：**选手需准备方案介绍 PPT 进行答辩，内容包含推理能力提升和性能优化的技术方案、推理模型的应用 `demo` 等（可附上演示视频）。

## 六、作品评选标准

### 初赛：客观题打榜

#### A 榜测评包含如下两个指标

1. 格式得分，模型输出的结果需要包含推理过程（必须是模型本身的推理输出，不得使用后处理进行格式包装），满足如下格式`^<think>(.*?)</think>\s*<answer>(.*?)</answer>$`；

2. 精度得分，测试集为 200 条包含数学、推理、代码相关任务的输入数据，判分系统将提取模型输出结果中的 `answer` 部分进行精度评测；

#### B 榜测评除格式、精度得分外，还包含如下指标：

3. 性能得分，测试集中包含多种并发数、输入长度、输出长度的组合，选手根据组委会提供的示例代码自行提取每组测试用例的推理耗时，其中低于原始模型性能的不得分，其余按照所有选手提交作品性能排名计算得分。注：性能部分要求选

手必须有算子融合、优化方面的工作，否则不得分。

**初赛 B 榜判分系统对提交作品的要求如下：**

1. 计算规格是 Ascend: 1\*ascend-snt9b(32G)|ARM: 24 核 192GB。评分系统加载参赛团队提交的模型工程，对比赛用的 200 条测试数据（此部分数据不公开）进行批量预测，最后根据预测结果自动评分。

2. 评分系统设置了 1 小时（不包含排队时间）判分任务超时时间，如果算子运行速度较慢导致判分任务运行时间超时，则该作品无得分。参赛团队在提交作品前应自行测试算子的性能，保证不会超时。

3. 参赛团队需要在每条数据推理前后自行计算耗时，并输出到结果文件。**希望团队能秉承诚信原则，输出真实的数据。**组委会将在比赛提交时间截止后，按得分排序检查代码，一旦发现作弊，将取消该团队的所有成绩。

4. 在 ModelArts 模型管理导入模型后，模型不会自动更新，如果您有更好的模型需要提交判分，则需要重新导入模型。为了方便区分不同模型的分数，不建议在同一个模型上创建多个版本，建议每次自定义新的名称重新导入模型。

**总决赛（终审擂台赛）：**初赛最终排名以 B 榜成绩和代码核查结果为准，筛选出一定数量的队伍入围。对入围作品，大赛组委会将综合考虑功能完整性、技术先进性（算子优化是核心指标）、场景创新性进行判分。

酌情加分项：使用华为开发者空间提供的相关资源和服务、端侧应用鸿蒙适配、及其他华为开放能力调用。

## 七、作品提交时间

2025 年 5 月-8 月，各高校应组织学生参赛，安排专业人员给予指导，为参赛团队提供支持保障。

2025 年 8 月 15 日前，各参赛团队通过[华为云竞赛平台](#)提交作品，具体要求详见作品提交方式。2025 年 5 月下旬至 7 月 20 日为初赛 A 榜作品提交期，2025 年 7 月 21 日至 8 月 12 日为初赛 B 榜作品提交期，8 月 13 日至 8 月 15 日，需按组委会要求提交最终的代码文件。**初赛最终排名将依据 B 榜成绩和代码核查结果。**

2025 年 8 月底前，由大赛组委会会同发榜单位共同完成初审，确定入围终审擂台赛的晋级作品和团队。

2025 年 9 月，发榜单位安排专门团队提供帮助和指导，各晋级团队完善作品，冲刺攻关参加终审擂台赛，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 [2025.tiaozhanbei.net](http://2025.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公章。

3. 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 2025 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

提交具体作品时，务必一并提交 1 份报名系统中审核通过的参赛报名表（所有信息与系统中填报信息保持严格一致）。

本单位命题为算法能力+创新创意，客观判分部分须在华为云竞赛平台提交代码。因此，选择华为命题的选手须同步登录 <https://developer.huaweicloud.com/competition/information/1300000068/introduction>，进行实名校验、作品提交、算法优化、查看实时排行榜。赛题相关的辅助学习资料、资源、FAQ 等，也在该平台发布。

参赛选手需要提交算法包以及方案介绍 PPT 等，详情请见上面的作品评选标准，将 PPT 作品方案介绍+作品代码+报名系统中审核通过的参赛报名表一并压缩成 ZIP 压缩包，上传到华为云命题大赛平台

<https://developer.huaweicloud.com/competition/information/1300000068/submission>，请将压缩包命名为：院校名称+队长姓名+队长手机号+队伍名称（例如：XX 大学+张 XX+137XXX+XX 战队）



## 九、赛事保障

本单位成立“揭榜挂帅”赛事服务项目组，提供赛题技术文档材料，提供云资源券，助力参赛选手学习技术，了解实践操作。

提供赛题相关的模型训练等学习课程，为参赛学生提供体系化学习路径和课程培训，具体详见

<https://developer.huaweicloud.com/competition/information/1300000068/html7>。

作品算法判分能力基于华为云竞赛平台，有严格可信的平台性能、24 小时 Oncall 团队、华为统一的问题处理 SLA。对参赛高校和学生的疑问，设立一线、二线响应机制，确保及时准确解决相关技术卡点。通过这些措施，我们致力于为参赛选手创造一个公平、高效、专业的大赛环境。

## 十、设奖情况及奖励措施

### （一）设奖情况

根据评分规则，综合评定参赛队伍。拟设“擂主”1 名（从特等奖中评选），特等奖 3 名，一等奖 4 名，二等奖 6 名，三等奖 8 名。

2025 年“揭榜挂帅”擂台赛学生赛道获奖情况将按照一定分值计入第十九届“挑战杯”竞赛学校团体总分，具体分值以第十九届“挑战杯”竞赛章程为准。

### （二）奖励措施

1. 本单位将结合项目实际，拟奖励擂主 8 万元（叠加特等奖激励一共 10 万元）；奖励特等奖每支队伍 2 万元；奖励一等奖每支队伍 1.5 万元；奖励二等奖每支队伍 1 万元；奖励三等奖每支队伍 0.5 万元。

2. 比赛中表现优异的获奖选手，将有机会进入华为人才储备池，并优先获得实习及就业的推荐机会。

3. 实际发奖数量将依据作品提交整体情况及赛事组委会的评审结果来确定。

### （三）奖金发放方式

以上奖金为税前奖金，由获奖团队承担税款。所有现金奖励将在比赛结束后 1 个季度内，通过银行转账的方式，发放至各获奖团队指定的账号。

## 十一、比赛专班联系方式

### 1. 专家指导团队

顾问专家：夏老师，联系电话：15919865031

负责比赛期间技术指导保障。

### 2. 赛事服务团队

联络专员：刘老师，联系电话：15889847842

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

华为创立于 1987 年，是全球领先的 ICT（信息与通信）基础设施和智能终端提供商。我们的 20.7 万员工遍及 170 多个国家和地区，为全球 30 多亿人口提供服务。我们致力于把数字世界带入每个人、每个家庭、每个组织，构建万物互联的智能世界。科学探索与技术创新是推动人类文明进步和社会发展的主要力量。华为重视研究与创新，近十年累计投入的研发费用超过人民币 11,100 亿元；截至 2023 年底，华为在全球共持有有效授权专利超过 14 万件。

华为将持续与政、产、学、研、用等各领域的产业组织和生态伙伴开放合作，持续向产业界贡献标准提案、产业理解、技术难题等，推动产业发展和技术进步。