

2025 年全国大学生计算机系统能力大赛

智能计算创新设计赛（先导杯）

北京师范大学校内赛技术方案

一、评价方式说明

- 第 1 条 大赛要求各参赛队综合运用所学知识，完善并优化给定的大模型推理引擎，以展示面向特定目标平台的编程能力和代码优化能力。
- 第 2 条 大赛鼓励各参赛队充分了解目标硬件平台特点，尽可能利用目标硬件平台提高推理引擎运行效率。
- 第 3 条 除本技术方案特别要求、规定和禁止事项外，各参赛队可自行决定代码编写方式、代码优化方案等细节。

二、初赛评分标准

- 第 4 条 比赛内容。轻量级大模型推理引擎完善和优化。
1. 基于 C/C++ 语言实现给定推理引擎代码框架关键算子，能够在给定平台成功编译和测试运行。
 2. 能够根据目标平台架构，对完成的大模型推理引擎进行优化，提升大模型推理引擎在目标平台上的运行效率。
- 第 5 条 功能测试。参赛队完善和优化后的推理引擎应能运行大赛指定的大模型，并对给定的输入测例获得相同的输出结果。所有测例的总运行时间不超过 30 分钟。功能测试要求基于大赛指定的测试模型和输入用例，用自行完善和优化的大模型推理引擎生成输出结果，比对原有输出结果计算得分。若未能生成输出结果，或所有测例输出结果都不同计 0 分；所有测例都通过计 100 分；部分测例通过的，按所通过测例的比例计算功能测试得分。
- 第 6 条 性能测试。在通过 80% 功能测例的前提下，记录每个测例在目标硬件平台上的推理速度作为评价依据。对所有参赛队按照所使用测例的平均生成

速度从高到低排序，速度最高者为 100 分，其余各参赛队依据推理速度与最快速度的比值除 100 计算参赛队在该基准测试程序上的性能得分（性能得分=100 /（推理速度/最快推理速度））。性能测试分值越大越好。

第 7 条 竞赛总成绩 100 分，各分项成绩权重如下：

- 1.大模型推理引擎的功能测试成绩：50%
- 2.大模型推理引擎的性能测试成绩：50%

三、参赛作品提交

第 8 条 各参赛队需按照要求提交完整的设计内容：

- 1.完善和优化后的大模型推理引擎的完整工程文件（必须包含全部 C/C++ 源代码）。
- 2.大模型推理引擎的完善和优化文档。

第 9 条 如果在开发过程中使用了代码大模型等人工智能辅助工具生成部分源码，必须在相关文档、源码注释或项目说明中予以明确说明，包括所使用的工具名称、生成内容的范围及对生成结果所作的人工修改情况等信息，以保障代码的可追溯性与合规性。

第 10 条 参赛队不得针对特定测试用例进行违规优化，不得恶意干扰评测环境或绕过评测系统规则等。一经发现，组委会将取消参赛队参赛资格。

第 11 条 参赛队必须严守学术诚信。一经发现代码抄袭或技术抄袭等学术不端行为，代码重复率在 50%以上，取消参赛队的参赛资格。

四、竞赛平台与测试程序

第 12 条 大赛提供的竞赛平台和测试模型包括：

- 1.曙光算力平台，报名成功后即为每个参赛队分配算力和存储空间。
- 2.竞赛作品评测，技术委员会根据参赛队提交代码生成推理引擎可执行程序，并加载模型和测例进行功能及性能测试。

五、软硬件系统规范

第 13 条 本次竞赛使用 110M 的 Llama 模型作为测试模型，采用 Children-Stories-Collection 数据集中的部分测例作为输入测例。

第 14 条 大赛指定的编译环境用于编译参赛队提交的编译系统源码，参数如下：

- 1.CPU: x86_64 架构; DCU: 9.2.8
- 2.操作系统: NFS Server 3.2 (RTM1)
- 3.C/C++编译器: Clang/Clang++ version 13.0.0
- 4.编译选项:

C 语言: `clang --std=c11 -O2 -lm`

C++语言: `clang++ --std=c++17 -O2 -lm`

六、大赛参考资料

第 15 条 曙光平台的学习资料见: <https://developer.sourcefind.cn/>, 可报名参加课程学习。

第 16 条 课程资料提供但不限于下列内容:

- 1.DCU 使用文档
- 2.DCU 编程实战文档
- 3.曙光平台使用文档
- 4.DTK_CUDA 兼容视频课程
- 5.DTK 开发人员教程
- 6.DCU 异构编程与优化——基础
- 7.DCU 异构编程与优化——进阶
- 8.DCU 异构编程与优化——调试与优化