IEEE AICAS 2025 Grand Challenge - LLM Software and Hardware System Co-optimization

1、Introduction

This track requires deploying and running the LLM under the hardware conditions of the Arm cloud platform, using the CPU processor of Armv9 architecture (T-Head Yitian 710) as the arithmetic platform to carry out the edge-side optimization of the Qwen large language model.

2、Background

In recent years, large language models (LLMs) based on pre-training and Transformer technology have demonstrated outstanding performance in various downstream natural language processing tasks, such as text understanding, text generation, sentiment analysis, machine translation, and interactive question answering. However, due to concerns about data privacy and computational efficiency, achieving efficient LLM inference on the edge has emerged as a key development trend. The deployment of LLMs on edge devices faces significant challenges, as they often have tens of billions or even hundreds of billions of parameters. Furthermore, the growth of model parameter size far outpaces improvements in hardware performance. To address these challenges, academia and industry have begun exploring software-hardware co-optimization methods, such as model compression, dataflow optimization, and operator invocation, to enable efficient deployment and operation of large models on resource-constrained hardware.

Among these approaches, deploying LLMs on CPU-architectured processors has become one of the key development directions. Recently, the industry has been releasing LLMs that can be deployed on edge-side devices, which further promotes this technology trend and attracts more researchers from academia and industry. Currently, how to systematically optimize large models to cope with the lack of hardware performance has become a key issue to achieve efficient inference of LLMs at the edge-side. Since most of the edge-side devices run on CPUs based on Arm architecture, this year's AICAS conference organizes a general LLM performance optimization competition for CPUs based on Arm architecture, with the goal of promoting and advancing the development of related technology.

3、Competition Description

A. Challenge Plan

Participants will base their work on the Qwen2.5 large language model (LLM). Relevant methods (e.g., model compression, parameter sparsity, precision quantization and structure pruning, etc.) can be proposed from multiple perspectives, combined with Arm architecture hardware features and open source software resources (e.g., hardware BF16, vector matrix multiplication, Arm Compute Library, etc.) to systematically optimize and improve the inference performance of

the LLM on hardware. The final score of the optimization method will be obtained through the test plan specified by the organizing committee of the competition.

Qwen2.5 is the LLM officially released by Alibaba Cloud in September 2024, featuring parameter scales ranging from 0.5B to 72B. The model demonstrates comprehensive performance across mainstream benchmark evaluation sets. All models are pretrained on a large-scale dataset containing up to 18T tokens, providing Qwen2.5 with significantly more knowledge compared to its predecessor, Qwen2. Among these, the 0.5B-parameter general-purpose model, Qwen2.5-0.5B, is an open-source and free version. Alibaba Cloud offers global multi-channel access to the model, along with services for training, deployment, and inference.

The preliminary round does not specify a specific hardware platform, the participating teams can independently choose the hardware platform for the verification of the LLM optimization method, the overall performance evaluation, and the comparison of "Ability" and "Efficiency" before and after the optimization of hardware deployment. After the preliminary stage, we will select 16 teams that have submitted valid results and undergone code review to enter the final round. In the final round, the participating teams will use the CPU processor based on Armv9 architecture (T-Head Yitian 710) provided by the organizing committee to carry out hardware and software co-optimization of the LLM, propose targeted algorithmic reasoning and optimization of the deployment strategy, and then carry out the overall performance evaluation and comparison before and after deployment, finally evaluate the first, second, third prizes and the seven winning prizes. The outstanding design solutions will be invited to be published in IEEE conferences or journals.

B. Competition schedule

- Preliminary Round Start: 2024-12-01
- Preliminary Round End: 2025-02-15
- Competition Workshop *：2025-02-22
- Final Round Start: 2025-02-24
- Final Round End: 2025-03-24
- Code and Technical Report Submission：2025-03-30
- Competition Results Announcement: 2025-04-10
- AICAS Conference**：2025-04-28

Competition Workshop *： After the preliminary round, the organizing committee will host a technical workshop in Hangzhou, inviting the top 10 teams from each competition track that advanced to the final round. The workshop will feature keynote speeches by industry experts and include networking activities. The organizing committee will reimburse travel and accommodation expenses incurred by participating teams attending the workshop (if they meet the competition's reward criteria).

AICAS Conference **：After the final round, the organizing committee will invite the **top three** winning teams to present their projects at the 2025 AICAS main conference in France and participate in conference activities. The committee will reimburse the travel and accommodation expenses incurred by the participating teams during the seminar (if they meet the competition's award criteria).

C. Registration

- Log in to the Tianchi：https://tianchi.aliyun.com/competition/entrance/532289
- Complete registration by filling in personal information. Participants and mentors are required to provide real-name verification.
- Each participating team may include up to five participants and two mentors, and each participant can only join one team.
- Participant registration, team formation, and modification operations must be completed by 23:59:59 on February 15, 2025.**（Verification Process：Tianchi official website -个人中心-认证-支付宝实名认证）。**
- Scan the QR code to join the DingTalk group chat.
- This challenge is open to current students and researchers from universities and research institutions.
- Employees of T-Head Semiconductor, Arm, and Alibaba DAMO Academy are not eligible to participate in the award evaluation.

D. Competition's award criteria

- Each track offers a cash prize pool of $5,550 and a travel subsidy of $7,500.

The top 10 teams from the preliminary round will be invited to attend the technical seminar in Hangzhou, and the top 3 teams from the final round will be invited to the AICAS 2025 conference. The organizing committee will provide travel subsidies (pre-tax) as follows:

- Hangzhou Technical Seminar：USD300 or CNY 2100（10 teams）
- AICAS2025：USD1500 or CNY 10500 (3 teams)

The selection criteria are based on their final scores, submitted technical reports, and the quality of their papers.

The distribution of prizes for this competition (pre-tax) is as follows:

- First Place：USD 2000 or CNY 14000（1 team）
- Second Place：USD 1500 or 10500（1 team）
- Third Place：USD 1000 or 7000（1 team）
- Excellence Award：USD 150 or 1050（7 teams）

4、Scoring Criteria

Evaluation Metrics	Evaluation Contents	Evaluation Method
Model Inference Accuracy（Ability）	● Accuracy Capability Testing	Run the local analysis tool lm-evaluation-harness to generate a JSON file.
Model Inference Performance （Efficiency）	● Parameter Compression Ratio ● Throughput Improvement Rate	Run the local analysis tool optimum-benchmark to generate a JSON file.

A. Preliminary Round Evaluation Metrics

1. Model Inference Accuracy（Ability）Evaluation：

Accuracy Testing： Evaluate the model's performance using the open-source evaluation toolkit `lm-evaluation-harness`, and conduct tests locally on the following datasets: ARC_challenge\HellaSwag\ Piqa. The scoring criteria is based on the average test scores of Qwen2.5-0.5B across three datasets, denoted as P_{max}, and the average test scores of Qwen2.5-0.5B-Instruct-GPTQ-INT4 on the same three datasets, denoted as P_{min}. If the model deployed by the contestant achieves an accuracy lower than P_{min}, the score for the inference accuracy will be negative. If the model accuracy is greater than or equal to P_{max}, the highest score for the inference accuracy bonus will be awarded. The specific calculation formula is as follows:

$$Ratio_{accuracy} = \begin{cases} 1, & for\ P >= P_{max} \\ \frac{P - P_{min}}{P_{max} - P_{min}}, & for\ P < P_{max} \end{cases}$$

2. Model Inference Performance（Efficiency）Evaluation ：

Parameter Compression Ratio：The comparison of the memory usage (in bytes) on the hardware system before and after model compression when loading the model onto the computation platform. The initial size of the Qwen2.5-0.5B model is denoted as M_{ori}, and the optimized size of the Qwen2.5-0.5B model is denoted as M_{opt}. The specific calculation formula is as follows:

$$Ratio_{memory} = \frac{M_{ori} - M_{opt}}{M_{ori}} = 1 - \frac{M_{opt}}{M_{ori}}$$

The throughput improvement rate for prefill and decoding：The throughput (tokens per second) of the model before and after optimization is analyzed using the optimum-benchmark open-source tool. The throughput of the initial Qwen2.5-0.5B model on the hardware platform is denoted as T_{ori}, and the throughput of the optimized Qwen2.5-0.5B model is denoted as T_{opt}. The results for the prefill and decoding stages are calculated separately. The specific calculation formula is as follows:

$$Ratio_{throughput\_P/D} = \frac{T_{opt\_P/D} - T_{ori\_P/D}}{T_{opt\_P/D}} = 1 - \frac{T_{ori\_P/D}}{T_{opt\_P/D}}$$

The total score of the preliminary round will be calculated with weighting, and each metric will be normalized in the weighted formula based on the highest score for that metric across all participating teams. Contestants will upload their results to the Tianchi platform, which will then provide the total score and ranking online. The specific calculation formula is as follows:

$$Tot_{1st} = 0.4 \times \frac{Ratio_{accuracy}}{MAX(Ratio_{accuracy})} + 0.2 \times \frac{Ratio_{memory}}{MAX(Ratio_{memory})} + 0.2 \times \frac{Ratio_{throughput\_P}}{MAX(Ratio_{throughput\_P})} + 0.2 \times \frac{Ratio_{throughput\_D}}{MAX(Ratio_{throughput\_D})}$$

The current requirements for the competition entries include:

- Test Report：Introduce the test results after optimizing QWEN-0.5B by your team, and provide documentation of the results from the specified third-party testing software.
- Technical Documentation: Describe the optimization methods implemented by your team for the LLM.
- Source Code of the Optimization Method: Provide the source code for the implemented optimization method.

B. Final Round Evaluation Metrics

In the final round, participants need to deployed and optimized LLM on the Yitian platform on AliCloud, using Qwen2.5-0.5B for performance evaluation and comparison, the scoring criteria for the final round will be provided subsequently. The final score of the competition will be weighted by the score of the preliminary round and the score of the rematch.

After the final round of the competition, participating teams are required to submit:

- Paper: A paper that presents the team's LLM application scenarios, the hardware-software co-design approach, as well as a comparison of the evaluation metrics and performance summary before and after optimization.
- Source Code: The actual source code used for model optimization must be submitted. *(The competition organizers have the right to use the code and development cases from the GC participating teams as promotional material in open-source communities such as GitHub, Huggingface, etc., or for inclusion in open-source tools related to sponsor products.)*

Note: The maximum number of submissions per day for the preliminary and final rounds is 5.

5、Committee

- Xi'an University of Electronic Science and Technology: Wei Mao, Bo Li
- Shanghai Jiao Tong University: Yongfu Li, Zhezhi He
- Nanjing University: Li Du, Yuan Du

- University of Electronic Science and Technology of China: Liang Chang
- T-head (Shanghai) Semiconductor Technology Co., Ltd.: Xiaohan Ma, Guosheng Yu
- Arm Technology (China) Co., Ltd.: Fengzhi Pan, Xile Yang

6、Relevant Links

- 2025 AICAS: http://www.aicas2025.org/
- T-Head Semiconductor Co., Ltd.: https://www.t-head.cn/
- Tongyi Qianwen: https://qianwen.aliyun.com
- Arm Technology (China): https://www.armchina.com
- Huggingface (Global download link for Qwen model): https://huggingface.co/Qwen/Qwen2.5-0.5B
- Modelscope (Chinese download link for Qwen model): https://modelscope.cn/models/qwen/Qwen2.5-0.5B/summary
- Qwen Github (Download links for all open source versions): https://github.com/QwenLM/Qwen2.5
- Optimum-benchmark: https://github.com/huggingface/optimum-benchmark