Low-Precision Large Language Model(LLM) Efficient Computation Challenge

Yuanyong Luo[1], Xin Wang[1], Jianpeng Li[1] and Jianlin Yu[1]

I. Grand Challenge Description

The dramatic scaling of deep learning models has made computational resource consumption and storage bandwidth pressure core bottlenecks constraining AI deployment. To achieve more efficient inference and deployment with limited hardware resources, low-precision computing has emerged as a key frontier trend in both academia and industry. Novel low-precision data formats, such as HiFloat8, HiFloat4, and MXFP4, show significant potential for optimizing model computation density and reducing energy consumption due to their unique balance of dynamic range and precision. However, extremely low-bit-width quantization inevitably introduces significant precision degradation, and different network architectures exhibit varied sensitivity to these new formats. This necessitates quantization algorithms that not only possess high compression efficiency but also maintain model robustness in complex tasks. Currently, overcoming the limitations of traditional uniform quantization algorithms and achieving a dual breakthrough in performance and precision with novel low-precision formats remains a critical challenge to address.

In the HiF8 data format[1], a dedicated field D is incorporated alongside the sign bit (S), exponent bit (E), and mantissa bit (M) to denote the variable bit width of the exponent and mantissa. The D field adheres to a tapered distribution, provides anti-overflow capabilities, and aligns with the characteristics of AI model parameters. Thus, HiF8 successfully expands the format's dynamic range under 8-bit constraints by leveraging precision features that match data distribution, all while preserving the required precision for neural networks. Consequently, this provides a more comprehensive 8-bit single data format expression for neural network training and inference. Furthermore, a white paper detailing the HiFloat4 data format will be released concurrently.

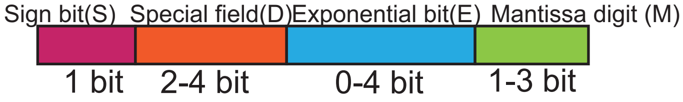

Fig. 1. Structure of HiF8 data format

Scaling strategies are another avenue for improving low-precision data quantization and thus a primary focus of this challenge. We specifically examine the effectiveness and stability of per-tensor current scaling and per-tensor delayed scaling during full pre-training of small-scale models. Participants are encouraged to explore and innovate quantization strategies and algorithms based on their chosen data formats. Based on this rationale, we established this challenge to explore the application boundaries of novel formats such as HiFloat8 and MXFP in model quantization. By establishing two technical tracks—quantization-aware training (QAT) and post-training quantization (PTQ)—the challenge encourages researchers to develop innovative quantization strategies and calibration algorithms. This exploration not only facilitates the evolution of efficient, high-precision model deployment techniques but also provides critical theoretical support and practical guidance for future hardware architecture design optimized for low-precision computing.

The challenge centers on Text-to-Image and Text-to-Video generation tasks. This challenge establishes two primary research directions:

Direction A (Training-Aware Quantization, TQ): Participants are required to use specified public datasets and pre-trained models to perform quantization-aware fine-tuning using the HiFloat8/HiFloat4/MXFP4 formats, optimizing model accuracy on downstream tasks.

Direction B (Post-Training Quantization, PTQ): Participants are required to apply post-training quantization directly to pre-trained models to achieve model compression and acceleration using formats such as HiFloat4 or MXFP4.

Two sub-challenges are set up around these research directions. Each sub-challenge adheres to the unified principles stipulated in the "Evaluation Criteria" section.

A. Sub-challenge 1: W4A4 Quantization for Inference (HiFloat4/MXFP4)

Large model inference faces significant deployment costs, which often constrain application proliferation. Quantizing the weights and activations of the linear layers in large models effectively reduces weight movement overhead and leverages low-precision computational power to enhance inference performance. This task focuses on 4-bit weight and activation

[1] Huawei Technologies Co., Ltd

quantization, restricted to either the HiFloat4 or MXFP4 numerical format. The reference model is Qwen3-32B. Evaluation is based on testing standard downstream task datasets, using the mean absolute percentage precision loss as the final result. The objective is to achieve a W4A4 inference precision loss of less than 0.5% relative to the FP16 model on the MMLU, GSM8K, GPQA, and MATH-500 datasets. Participants are permitted to protect a limited number of Transformer blocks at high precision: a maximum of 5 layers for MXFP4 and 2 layers for HiFloat4.

B. Sub-challenge 2: W8A8 Quantization for Training (HiFloat8)

Large model training incurs high costs and lengthy iteration cycles, limiting rapid development. Quantizing the weights and activations of linear layers, and utilizing low-precision formats within attention layers, can effectively reduce data movement costs and accelerate training via low-precision computation. This task focuses on 8-bit weight and activation quantization and attention quantization, strictly limited to the HiFloat8 numerical format. The test model is Wan2.1. Participants are encouraged to employ delayed scaling strategies wherever possible to further minimize quantization overhead and accelerate training. The training methodology will be evaluated based on the video generation quality produced by the post-training model using the BestWishYsh dataset. The Vbench evaluation metric will be used, aiming for a precision loss of less than 0.5%.

This challenge aims to ensure the fairness and reproducibility of evaluations by providing comprehensive datasets and supporting tools. Participants can utilize multiple public standard test sets (e.g., OpenS2V-5M, MMLU, GSM8K, GPQA, MATH-500) for solution development and validation. For a fair final evaluation, a private, hidden test set will be employed, with results released concurrently at the challenge conclusion. Furthermore, we provide a basic quantization emulation operator toolkit supporting precision validation for HiFloat8, HiFloat4, and MXFP4 data formats, along with accompanying use cases. This toolkit is compatible with Ascend 910 series devices and NVIDIA GPU platforms, facilitating cross-hardware platform verification.

II. Evaluation Criteria

The evaluation will be structured into two award systems: general awards and a dedicated innovation award. Participants may select a maximum of two models from the designated model set (which will be released alongside the challenge description and accompanying public datasets) for their quantization validation.

TABLE I

Award Detail

Awards		**Evaluation Criteria**	**Key points**
General awards	Three class in each direction	**Objective Metrics (80%):** Given that the model precision meets the established baseline, evaluation focuses on the quantization ratio and resulting performance acceleration. **Subjective Review (20%):** Assessment criteria include solution completeness, engineering implementation quality, and report compliance.	Participants are encouraged to achieve superior performance optimization ratios and higher engineering maturity while satisfying basic precision requirements.
Innovation award	One in each direction	**Objective Metrics (60%):** Model precision relative to the required baseline performance. **Innovation (40%):** Encouraging the design and practical implementation of native algorithms and operators specifically for HiFloat quantization, assessed independently to promote original breakthroughs.	Key consideration will be given to whether innovative quantization algorithms or optimized operators are proposed specifically leveraging the characteristics of the novel HiFloat data formats, demonstrating algorithmic breakthroughs.

Innovation Award Subjective Evaluation Criteria

- Algorithmic Originality (40%): This criterion assesses the degree of innovation in the proposed quantization methodology, specifically whether it introduces entirely novel approaches or provides substantial improvements to existing methods.
- HiFloat-Native Exploration (30%): A core focus is whether the solution deeply leverages the unique characteristics of the HiFloat data format to design highly adapted quantization algorithms or computational workflows.
- Generality and Transferability (20%): This metric evaluates the potential for the proposed solution to be transferred and applied effectively across other model architectures or task scenarios.
- Engineering Value (10%): Focus is placed on code quality and standardization, reproducibility of results, and ease of deployment.

III. Submission Guideline

Participants should submit their result that contains files as follows:

- Quantized model files and corresponding inference code to validate engineering implementation.

TABLE II

CHALLENGE TIMELINE

Jan 30, 2026 (Development Phase):	Release of the challenge description and accompanying resources, including the development dataset, specified quantized large model versions, basic quantization emulation operator toolkit, and use cases.
Feb 10, 2026 (Development and Discussion Phase):	Release of the precision test dataset, establishment of a discussion forum and Q& A panel, and organization of expert support to address participant inquiries.
Mar 15, 2026 (Final Phase):	Participants submit technical reports and necessary engineering validation files required for their chosen tracks.
Mar 30, 2026 (Announcement):	Announcement of award results and test outcomes for each sub-challenge. Winning teams must further prepare their technical reports and reproducible engineering materials for final verification.

- A comprehensive technical report detailing the solution design, innovation analysis, and comparative experimental results.
- Reproducible scripts and associated performance validation data.

The detailed timeline is shown in Table II:

IV. Organizers' Contacts and Short Bio

Yuanyong Luo(luoyuanyong@hisilicon.com), Senior Researcher: Researcher at the Turing Lab and the inventor of the HiFloat data format. His primary research interests focus on lightweight, efficient large language models (LLMs) and AIGC models. He has made significant contributions to ultra-low-bit post-training quantization (PTQ), deeply analyzing quantization error sources and proposing the innovative CBQ method. This method achieves performance breakthroughs in ultra-low-bit quantization (e.g., W4A4) across various large models such as LLaMA by establishing cross-block dependencies and managing intra-layer dependencies.

Xin Wang(wangxin237@huawei.com), PhD.: Senior Researcher at Huawei, focusing on large language model architecture, compression acceleration, and inference optimization. He spearheaded an efficient pruning framework that addresses the challenges of aggressive structural pruning through innovative weight re-initialization techniques. This framework supports multi-dimensional pruning (width, depth, etc.) and integrates optimizations tailored for the Ascend NPU hardware, significantly reducing model size and inference costs.

Jianpeng Li(lijianpeng7@huawei.com), Senior Researcher: Technical expert at Huawei in the low-precision computing domain, actively engaged in driving industry ecosystem development. He played a key role in the R&D and standardization of the HiF8 data format, sharing insights on 8-bit format evolution and low-precision application at industry platforms like GCC, committed to promoting the industrial implementation and ecosystem collaboration of next-generation low-precision floating-point formats.

Jianlin Yu(yujianlin1@huawei.com), Senior Researcher: Senior research expert at Huawei specializing in innovative large model quantization frameworks and algorithms, and a core contributor to the framework. This framework supports various low-precision data formats (INT/FP, etc.) and integrates advanced algorithms such as learnable truncation thresholds. It aims to achieve an optimal balance between model precision and inference efficiency, providing critical support for the efficient deployment of large models on the Ascend platform.

References

[1] Luo Y, Zhang Z, Wu R, et al. Ascend hifloat8 format for deep learning[J]. arXiv preprint arXiv:2409.16626, 2024.