Spend Less and Do More

An LLM Efficiency Challenge

Tech Arena 2025 - **Phase 1**

Huawei Ireland Research Center

This document contains the instructions to compete in the Huawei Ireland Research Center Tech Arena 2025 challenge. Please take the time to read this document carefully and do not hesitate to contact us if you have any questions.

Problem

Your objective is to create an efficient Large Language Model (LLM) inference pipeline for single-round question answering. Therefore, the pipeline does not need to retain any conversational history. Common LLM inference optimization techniques include:

- **Context and Orchestration** - RAG, knowledge graphs, tools (e.g. calculators), prompt engineering, etc.
- **Deployment and Serving** - specialized accelerators, asynchronous serving, prefix and response caching, etc.
- **Model Design** - quantization, pruning, optimized attention mechanisms, token fusion, model merging, low-rank factorization, etc.

The proposed solution will be assessed against the following performance metrics:

- **Accuracy through an LLM-as-a-Judge**
- **End-to-End Latency**

1 Challenge Description

1.1 Business Context

Several startups dream of leveraging the power of Large Language Models (LLMs) to build the next successful business. However, the cost associated with both LLM training and serving, primarily due to expensive computing resources, represents a core challenge. With a constrained budget, startups typically have just one opportunity to train, fine-tune, or deploy an LLM. Even a minor error could consume their entire budget without any tangible results.

This lack of flexibility can hinder innovation, placing remarkable pressure on every product design choice. As a result, startups navigating this environment must make their “single shot” count, turning significant constraints into a quest for precision and efficiency.

In this challenge, you are the founding development team of a startup that aims to launch an LLM-powered educational application. Your goal is to build an LLM-powered learning platform that makes standard school curriculum subjects easily accessible to everyone. When developing your AI infrastructure, focus on what truly matters for a new business: create a foundation that is cost-effective, agile, scalable, and sustainable. The best solution will support your company's future growth, not just showcase the “coolest” technology.

1.2 Question Answering Requirements

You will leverage the available computational resources to implement a pipeline that can answer single-round questions on three subjects: **algebra**, **geography**, and **history**.

Please ensure each answer is no longer than 5000 characters. In terms of LLMs, you can use one or more of the models from the HuggingFace platform, which are listed below. The required question and answer formats are shown in Examples 1 and 2.

1. embeddinggemma-300m
2. gemma-3-270m-it
3. gemma-3-1b-it
4. gemma-3-4b-it
5. Llama-3.2-1B-Instruct
6. Llama-3.2-3B-Instruct
7. Qwen3-Embedding-0.6B
8. Qwen3-0.6B
9. Qwen3-1.7B
10. Qwen3-4B

```
{"questionID": "2d4fbd50-c20a-4e2b-aeb4-06f3f8b49d8c",
    "question": "What is the capital of the world?"}
```

Example 1: Example Question

```
{"questionID": "2d4fbd50-c20a-4e2b-aeb4-06f3f8b49d8c",
    "answer": "There is no official capital of the world."}
```

Example 2: Example Answer

2 Computing Environment

Your solution must be deployed to the environment detailed in Table 1 for testing. The evaluation will consist of approximately 500 questions of varied difficulty, spread across the subjects.

You will submit a single “zip” file that contains an “inferencePipeline” folder. Make sure your submission can work within the directory structure shown in Figure 1. As detailed in Figure 2, we will test your submission using a “run.py” script that imports a “loadPipeline” function from your “inferencePipeline” folder. This function should load your pipeline and return a callable. The returned callable must accept a list of questions as input and produce a list of answers in the formats specified in Examples 1 and 2. The evaluation will be conducted in an environment without internet access; therefore, you must load the required LLM models from a local cache (“/app/models”), as shown in Figure 3. Finally, your submission must include a “requirements.txt” file listing all the necessary dependencies, as shown in Figure 4.

If your solution requires a database, please ensure the database is pre-installed and included within your solution's directory.

To help with your first submission, we will provide you with a working example through the challenge platform.

CPU	AMD x86_64, 16 cores	**Memory**	128 GB, DDR4
OS	Linux	**Prog. Lang.**	Python 3.12

Table 1: Main Computing Environment Features

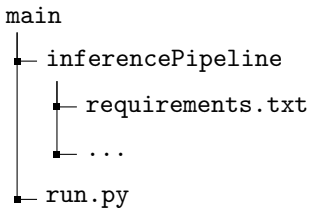

Figure 1: Evaluation Directory Structure

```python
from inferencePipeline import loadPipeline

# Load the inference pipeline
pipeline = loadPipeline()

# Start time
start_time = time.perf_counter()

# Run the inference pipeline
answers = pipeline(questions)

# Elapsed time
elapsed_time = time.perf_counter() - start_time
```

Figure 2: How “run.py” Calls and Times your Pipeline

```python
from transformers import AutoTokenizer, AutoModelForCausalLM

model_name = 'Qwen/Qwen3-1.7B'
cache_dir = '/app/models'

tokenizer = AutoTokenizer.from_pretrained(model_name,
                                          cache_dir=cache_dir)

model = AutoModelForCausalLM.from_pretrained(model_name,
                                             cache_dir=cache_dir)
```

Figure 3: How your Pipeline will Load Models from the Local Cache

```
transformers==4.49.0
pandas==2.3.2
...
```

Figure 4: Example of “requirements.txt” File

3 Solution Submission

You will submit your solution through a dedicated page on the challenge management platform. Further details about the submission process are provided in Table 2.

Format	A “zip” file that contains a single folder named “inferencePipeline”. Inside this folder, you must provide a “requirements.txt” file listing all necessary dependencies.
Size	The maximum submission file size is 1 GB. Please exclude LLM models from your submission folder as they will be provided via a local cache (see Section 2).
Frequency	You are allowed **one submission per day**. If your evaluation fails within the first 15 minutes, you may resubmit.
Evaluation Time	Any evaluation that exceeds 2-hour time limit will be terminated and will result in a null score.
Debugging	The event management platform will provide logs to assist you in debugging your code.
Deadline	The competition ends at 23:59 IST (Irish Standard Time) on Friday, November 7th. Please be aware that any submissions still processing at this time will be disregarded. Your final score will be determined by your last successfully completed submission on the leaderboard.

Table 2: Submission Process Details

4 Evaluation Metrics

The solution will be evaluated using a weighted sum of accuracy (60%), and end-to-end latency (40%) as detailed in Table 3.

The scoring for each metric on the leaderboard will be normalized. The best performance for a given metric will establish the 100% benchmark, and all other entries will be scored as a percentage relative to this result. The final score is a weighted average of the two metrics. The group with the highest final score will be ranked as the winner. Finally, it should be noted that the minimum accuracy required to enter the leaderboard is 10%.

Metric	**Details**
Accuracy through an LLM-as-a-Judge	Percentage of correct answers, as evaluated by an LLM with a context window of up to 1,047,576 tokens and that supports up to 32,768 output tokens per request.
End-to-End Latency	This is the total latency to answer all questions in the test set. It does not include I/O time (for loading and saving data) or the initial setup and deployment time for the inference pipeline.

Table 3: Evaluation Metrics

5 Restrictions

1. External API calls are not permitted.
2. Leaking challenge evaluation data is strictly forbidden.

Disclaimer

- This document is only intended to enable the “Tech Arena 2025” event hosted by Huawei Ireland Research Center. Under no circumstances should the information presented herein be interpreted as representative of any real entity or organization.
- This document is for “Tech Arena 2025” participants only and should not be distributed to external parties.