SAPTANG
Proactive Threat Defence

Saptang Labs - Machine Learning Challenge

Challenge Overview:

Large Language Models (LLMs) excel at many NLP tasks but struggle with structured, multi-step reasoning. While they generate fluent text, they often hallucinate intermediate steps, fail to decompose complex problems, or skip verification—particularly problematic in logic-based question answering where accuracy requires breaking problems into subproblems, applying appropriate methods, and verifying results.

This challenge asks participants to design an Agentic AI system that goes beyond passive inference. The system should actively plan, decompose, and solve logic problems—autonomously selecting and executing subtasks while providing transparent, human-readable reasoning traces.

Machine Learning
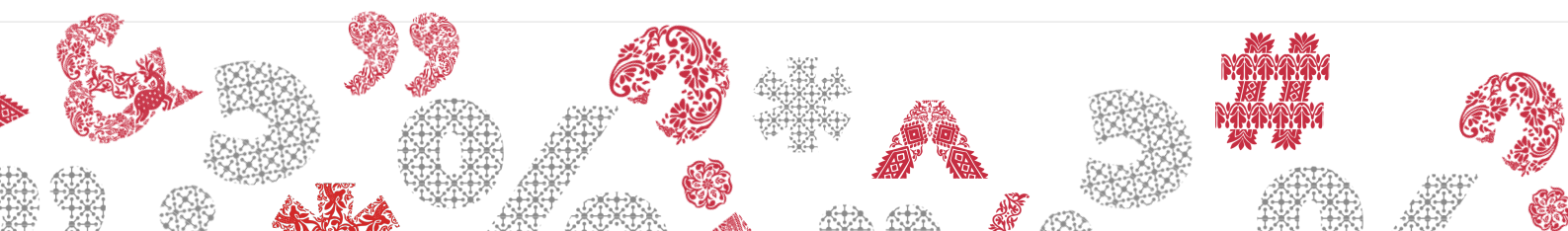
Objectives

The primary aim of this project is to build an Agentic Reasoning System that can:

- <u>Problem Decomposition</u> – Autonomously break down complex logic questions into manageable subproblems.

- <u>Tool Selection</u> – Choose appropriate tools such as symbolic solvers, calculators, or code execution modules for solving subtasks.

- <u>Execution & Verification</u> – Carry out subtasks and verify results to ensure correctness and reliability.

- <u>Reasoning Traces</u> – Provide transparent, step-by-step reasoning alongside the final answers for interpretability.

- <u>Dataset Performance</u> – Achieve strong results on the given dataset while demonstrating originality and creativity in approach.

- <u>Implementation Quality</u> – Deliver a clear, modular, and well-documented system for ease of understanding and reproducibility.

Dataset

Training and testing data is provided in csv format, the test output should be provided in a csv file in the specified format only (output.csv file given in the drive link).

Link for Dataset and out.csv file format: <u>Click Here</u>

Restrictions

The purpose of this challenge is to evaluate participants' ability to design and implement their own agentic reasoning system, rather than relying on the reasoning capabilities of existing large-scale proprietary models. To maintain fairness and ensure the focus remains on system design and innovation, the use of pre-trained reasoning-heavy LLMs is strictly prohibited.

Specifically, participants are not allowed to use models such as OpenAI GPT-4, GPT-4 Turbo, GPT-5, Anthropic Claude 3 Opus, Google DeepMind Gemini Ultra, or other similar APIs and services that already provide advanced multi-step reasoning out of the box.

Instead, participants are encouraged to:

- Utilize smaller or base LLMs with limited reasoning capabilities.

- Incorporate symbolic solvers, rule-based systems, or calculators where appropriate.

- Develop their own modular pipelines for problem decomposition, subtask execution, and verification.
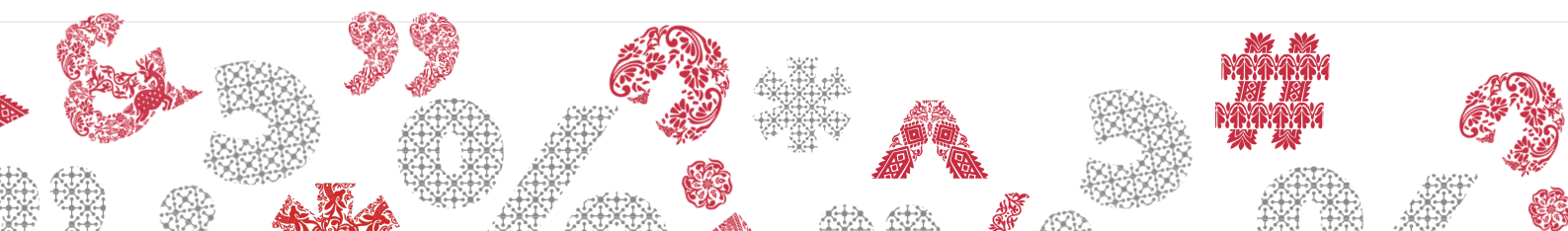
Requirements for Round 1

a. GitHub Repository – A complete implementation of the system with a clear folder structure, modular code, and proper documentation to ensure ease of understanding and reproducibility.

b. Prediction CSV File – A file containing the model's predictions on the provided test dataset, formatted exactly as specified.

c. Technical Report – A well-structured document that summarizes:
- System Design and Architecture – Explanation of the overall framework and key components.
- Problem Decomposition & Reasoning Approach – Description of how the system tackles multi-step reasoning.
- Results and Evaluation – Performance metrics, analysis, and key insights.

Evaluation Criteria

- Performance and Accuracy (50%) – Performance on the test dataset (Macro F1 Score), inference time, and the quality of step-by-step reasoning traces.
- Approach Creativity & Originality (35%) – Novelty and effectiveness of the reasoning framework.
- Technical Report Quality (10%) – Clarity, structure, and depth of the submitted report.
- Code Readability & Organization (5%) – Clean, well-structured, and documented code in GitHub.

Important Dates

- Doubt Clearing Session: 27th September, 7 PM
- Final Submission: 8th October EOD

Contact Details

For any further queries or doubts related to the problem statement, we have created a dedicated WhatsApp group. You can simply scan the QR code below to join and stay connected for discussions and updates.

Whatsapp Group