Internship Proposal

QCRI Research Internship Program

Project Title

GYM: A Benchmarking Framework for Multimodal Reasoning, Mathematics, Code Generation, and Spatial Understanding

Project Description

This project aims to develop GYM — a comprehensive evaluation and benchmarking framework ("gymnasium") designed to rigorously test and advance the capabilities of multimodal AI models. As large language models increasingly process multiple input modalities (text, images, diagrams, code), there is a critical need for robust benchmarks that systematically assess their performance across diverse reasoning tasks.

GYM will cover one of the four core challenge tracks:

- Multilingual Reasoning — evaluating cross-lingual inference, arithmetic question answering, and logical deduction from mixed text-image inputs.
- Multimodal Mathematics — assessing the ability of models to solve mathematical problems presented in visual, diagrammatic, and textual forms.
- Code Generation — benchmarking the generation, understanding, and debugging of code from natural language and visual specifications.
- Spatial Reasoning — measuring a model's capacity to interpret and reason about spatial relationships, images, maps, and geometric configurations.

Interns will contribute to dataset curation, task design, evaluation pipeline development, and baseline model assessments, helping produce an open benchmark resource for the research community.

Project Type

Research

Internship Batch

- Batch 1: May 10 to July 9 — suitable for Education City students (CMUQ, TAMUQ, and HBKU students)

Duties / Activities

- Review existing multimodal benchmarks and identify gaps in the literature.
- Curate, annotate, and quality-check datasets across the four GYM tracks (reasoning, math, code, spatial).
- Design and implement evaluation tasks, rubrics, and scoring pipelines.
- Run baseline evaluations on state-of-the-art multimodal models (e.g., GPT-4o, Gemini, Claude).
- Analyze model performance, produce error analyses, and summarize findings.

- Contribute to a technical report or paper documenting the GYM framework.

Required Skills

Required:

- Proficiency in Python programming.
- Familiarity with machine learning concepts and deep learning basics.
- Strong analytical and problem-solving skills.

Preferred:

- Experience with NLP or computer vision libraries (e.g., HuggingFace, OpenCV, PyTorch).
- Familiarity with large language models or multimodal models (API usage or fine-tuning).
- Background in mathematics or formal reasoning tasks.
- Experience with data annotation or benchmark construction.

Preferred Intern Academic Level

Undergraduate (preferred, in Computer Science, AI, Data Science, Mathematics, or related fields)

Learning Opportunities

- Hands-on exposure to cutting-edge multimodal AI research and state-of-the-art large language models.
- Experience in benchmark design, data curation, and evaluation methodology — highly valued skills in AI research.
- Opportunity to co-author or contribute to a research paper or technical report.
- Mentorship from active AI researchers at QCRI.
- Development of skills in Python, ML frameworks, and scientific writing.
- Networking with peers and researchers in Qatar's AI community.

Expected Team Size

2–4 interns (team projects are strongly preferred to encourage collaboration and division of tasks across the four GYM tracks).

Mentor

Name: Md Rizwan Parvez

Email: mpravez@hbku.edu.qa

Affiliation: QCRI