OxCam Programmes - AI+ Course 2025
Project Proposal

Programme Cohort	Large Language Models & Generative AI
Course Group	GEA-Group 2
Group Name	Robust Mathematical Problem Solver
Group Members	Liu Peixuan, Long Haobo, Cao Shuhan, Wu Jiayu, Wu Xuanyu, Wei Yanran, Jin Huiyan

Project Title

Robust Mathematical Reasoning Agent: Enhancing Robustness and Verifiability in AI Problem Solving

Project Summary

Large language models (LLMs) demonstrate remarkable capabilities in mathematical reasoning, yet their robustness remains a significant concern. Recent research, such as [5], highlights that introducing irrelevant or adversarial information into prompts can drastically impair reasoning accuracy, exposing the susceptibility of current LLMs to such interference. Furthermore, the rising complexity of mathematical tasks requires not only robust natural language understanding but also the formal verification of logical reasoning steps to ensure correctness [6, 3, 2]. Contemporary frameworks have shown the benefits of combining LLM-generated informal reasoning with formal verification backends, e.g., Lean or Coq, to improve trustworthiness [4, 7, 1]. Our project builds upon these directions and aims to develop an anti-interference mathematical reasoning agent that merges natural language reasoning with formal logic verification while mitigating the influence of prompt injections and adversarial triggers. Specifically, our agent will pre-filter noise or adversarial triggers, generate stepwise natural language reasoning, translate it into a formal proof for automatic checking, and iteratively refine answers based on proof feedback. Furthermore, by fine-tuning LLMs for adversarial resistance and designing agent-based multi-tool strategies, we anticipate improved robustness, interpretability, and verifiability in mathematical problem solving. The expected outcome is a deployable prototype agent, rigorous evaluation on adversarial mathematical datasets [5], and a detailed analysis of strengths and failure cases, paving the way for more trustworthy AI-powered theorem proving and mathematical problem solving.

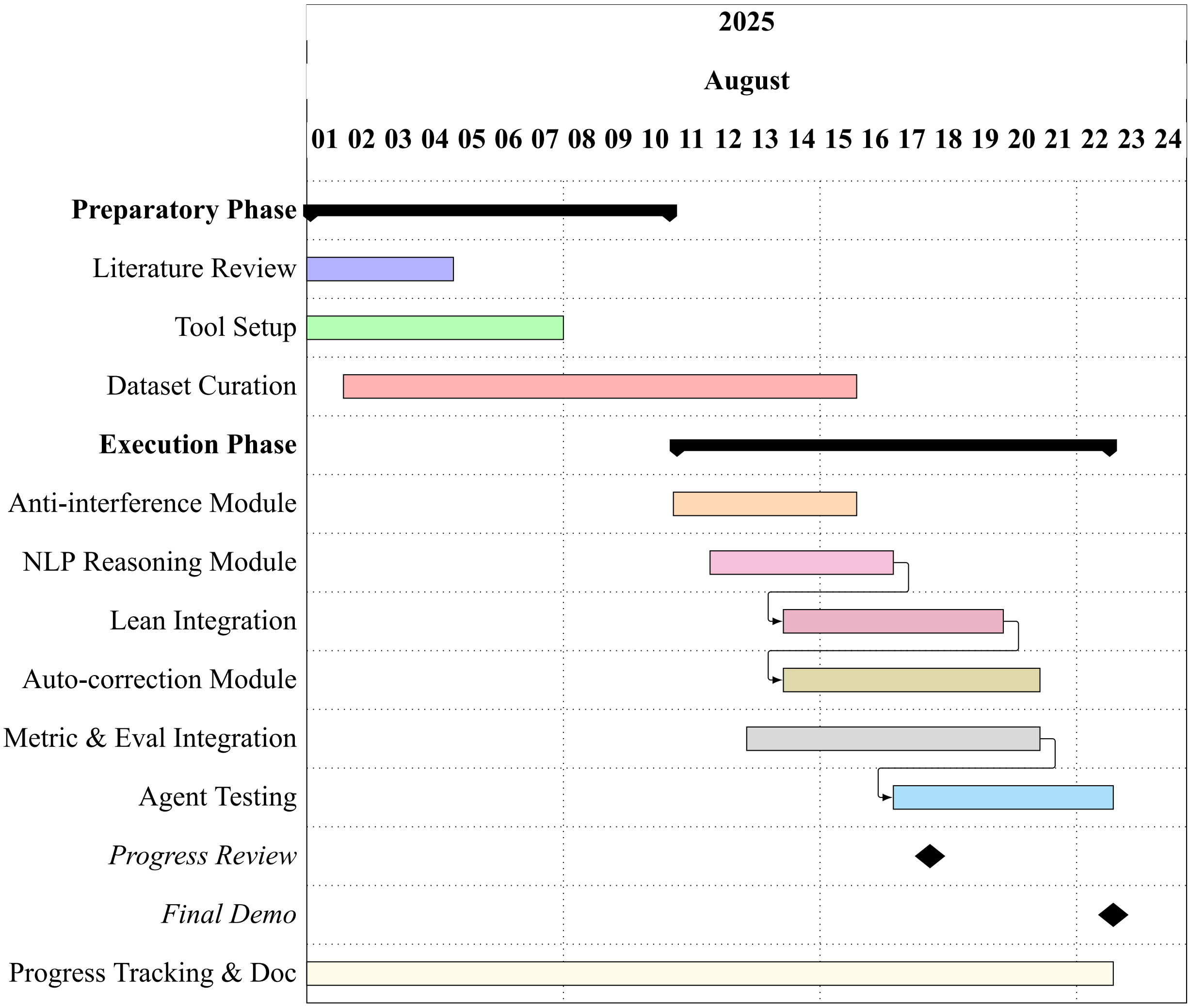
Project Significance and Contribution to the Field

Ensuring the reliability and interpretability of LLM-generated reasoning is a rapidly growing concern, especially in mathematics, where logical soundness is paramount. Prior work has revealed both the vulnerabilities of LLMs to adversarial prompt injections [5] and the usefulness of combining LLMs with formal verification frameworks to enhance solution correctness and accountability [6, 3, 4]. Agent-based approaches for formal reasoning have recently emerged [1, 8, 7], but existing agents often overlook robust anti-interference mechanisms and may not close the feedback loop between automated proof checking and LLM reasoning generation. Our project systematically integrates:

1. pre-processing modules for noise or adversarial removal as motivated by [5],
2. agent frameworks orchestrating multiple reasoning and verification components [1, 6],
3. verification-driven feedback for self-correction, as explored in recent systems [3, 4].

Methodologically, our advances include new workflows for joint informal and formal reasoning, targeted fine-tuning for adversarial resistance, and new evaluation metrics for robustness. These contributions offer both practical tools and new insights towards reliable and scalable math-reasoning agents, with implications for broader AI research on adversarial robustness and automated verification [4, 6].

Project Timeline and Task Allocations



Task allocations:

1. Liu Peixuan: Tracks project progress and maintains comprehensive project documentation.
2. Long Haobo: Responsible for implementing the anti-interference module in Python to ensure system robustness.

3. Cao Shuhan: Designs and integrates the natural language reasoning generation and transformation components using LLMs and agent frameworks.

4. Wu Jiayu: Develops the auto-correction agent to improve the accuracy of formal proof outputs.

5. Wu Xuanyu: Collects and organizes math problem datasets for system evaluation and development.

6. Wei Yanran: Sets up the test framework and conducts metric calculations for performance assessment.

7. Jin Huiyan: Analyzes evaluation results and provides feedback for system optimization.

Bibliography

References

[1] Kaito Baba, Chong Liu, Satoshi Kurita, and Akinori Sannai. Prover agent: An agent-based framework for formal mathematical proofs. 2025.

[2] Luoxin Chen, Jinming Gu, Liankai Huang, Wenhao Huang, Zhicheng Jiang, Allan Jie, Xiaoran Jin, Xing Jin, Chenggang Li, Kaijing Ma, Cheng Ren, Jiawei Shen, Wenlei Shi, Tong Sun, He Sun, Jiahui Wang, Siran Wang, Zhihong Wang, Chenrui Wei, Shufa Wei, Yonghui Wu, Yuchen Wu, Yihang Xia, Huajian Xin, Fan Yang, Huaiyuan Ying, Hongyi Yuan, Zheng Yuan, Tianyang Zhan, Chi Zhang, Yue Zhang, Ge Zhang, Tianyun Zhao, Jianqiu Zhao, Yichi Zhou, and Thomas Hanwen Zhu. Seed-Prover: Deep and Broad Reasoning for Automated Theorem Proving.

[3] Chen Liu, Yizhi Yuan, Yingtong Yin, Yang Xu, Xin Xu, Zelin Chen, Yue Wang, Li Shang, Qun Liu, and Ming Zhang. Safe: Enhancing mathematical reasoning in large language models via retrospective step-aware formal verification. 2025.

[4] Mihir Patel, Rounak Bhattacharyya, Tianshi Lu, Aarav Mehta, Nathaniel Voss, Negar Norouzi, and Gauri Ranade. Leantutor: A formally-verified ai tutor for mathematical proofs. 2025.

[5] Maitreya Rajeev, Rajasekharan Ramamurthy, Pulkit Trivedi, Vivek Yadav, Olatunji Bamgbose, Sudarshan T Madhusudan, James Zou, and Nazneen Fatema Rajani. Cats confuse reasoning llm: Query agnostic adversarial triggers for reasoning models. 2025.

[6] Zhezheng Ren, Zhongyu Shao, Junhua Song, Hang Xin, Hao Wang, Wen Zhao, Liang Zhang, Zhiyang Fu, Qiang Zhu, Donghong Yang, et al. Deepseek-prover-v2: Advancing formal mathematical reasoning via reinforcement learning for subgoal decomposition. 2025.

[7] Anshuman Thakur, Giorgos Tsoukalas, Yufei Wen, Jiongkun Xin, and Swarat Chaudhuri. An in-context learning agent for formal theorem-proving. 2024.

[8] Anshuman Thakur, Yufei Wen, and Swarat Chaudhuri. A language-agent approach to formal theorem-proving. 2025.

Total Word Count (excluding bibliography): 514

Evaluation Sheet

This section is to be completed by the Instructor(s).

Final Mark:

Further Comments: