PRISM-X: A Dual-Agent Governance Pipeline for Safe and Novel Research Ideation

Ravpreet Singh Gill
University College London
07711 837858
ravpreet.gill@outlook.com

Hamza Latif
University College London
07949 272371
Hamzalatif966@gmail.com

Cyrus Chan
University College London
+85255 379619
Hei.chan.25@ucl.ac.uk

ABSTRACT

PRISM-X is a dual-agent governance pipeline designed to augment large language models with vigilant oversight, enabling authentic and novel research idea generation. Built as a hackathon prototype, PRISM-X integrates a parent-layer governance framework (AEGIS X) supervising a child-layer ideation agent (NOVA). AEGIS X comprises three coordinated agents – a Guardian (content filter and safety watchdog), an Attacker (adversarial red teamer), and an Observer (trace logger and quality assessor) – that operate in real time to mitigate unsafe or flawed outputs. NOVA ingests real academic papers to propose original research ideas, complete with abstracts and citations, demonstrating scholarly acumen. PRISM-X's iterative meta-loop allows AEGIS X to intervene in NOVA's generation process, scrutinize outputs for risks or errors, and gradually improve output quality through feedback. We detail the system architecture and a case study in LLM interpretability research, where AEGIS X intercepts hallucinated citations and enforces prudent corrections. Preliminary evaluation suggests a 37% reduction in unsafe or policy-violating completions and a 22% increase in idea novelty, achieved with acceptable latency overhead. These results indicate that multi-agent governance, as embodied by PRISM-X, can augment the reliability and creativity of autonomous research assistants.

Keywords

Large Language Model, Guardian, Attacker, Observer

1. INTRODUCTION

Large language models (LLMs) can now generate research articles and proposals, but ungoverned generation risks hallucination, unsafe claims, and factual errors. Over 30% of citations produced by some models are non-existent, damaging the credibility of AI-generated research. Autonomous research tools require internal safeguards to ensure trustworthy and high-quality outputs. PRISM-X addresses this by introducing a governance layer—AEGIS X—that oversees a research idea generator – NOVA - using a triad of oversight agents.

2. System Overview

PRISM-X is structured as a dual-agent architecture in which:

- NOVA functions as the child agent responsible for producing research ideas, abstracts, and citation-backed outputs after ingesting real academic papers.
- AEGIS X operates as the parent governance layer, consisting of three coordinated agents:
- Guardian – filters unsafe, non-credible, or policy-violating content;
- Attacker – adversarially probes NOVA's outputs by attempting to induce hallucinations, unsafe claims, or logical flaws;
- Observer – captures reasoning traces, evaluates quality, and logs governance interventions.

Together, these components form an iterative meta-loop where AEGIS X continuously critiques, challenges, and corrects NOVA's outputs in real time. This layered structure ensures that generation remains both creative and trustworthy.

3. Methodology

3.1 NOVA Ideation Pipeline

NOVA follows a four-stage workflow:

1) Paper ingestion – NOVA reads real academic papers from a predefined corpus.
2) Representation extraction – key methods, gaps, and themes are summarised.
3) Idea generation – NOVA proposes an original research direction and drafts an abstract with citations.
4) Trace output – NOVA exposes intermediate reasoning so AEGIS X can inspect citations, claims, and novelty.

3.2 AEGIS X Governance Workflow

AEGIS X uses a three-agent loop:

Guardian Pass:

The Guardian checks NOVA's output for fabricated citations, unsafe biological claims, overconfident scientific statements, and factual inconsistencies. If problems appear, NOVA must revise its output.

Attacker Pass:

The Attacker red-teams the output by trying to push NOVA towards hallucinations, unsafe suggestions, or ambiguous ideas. Any failures trigger corrective revisions.

Observer Pass:

The Observer records novelty score, safety score, reasoning-trace clarity, and the number of required interventions. These logs guide the next refinement cycle.

3.3 Iterative Meta-Loop

PRISM-X cycles through generate → evaluate → revise until the final output satisfies safety and novelty requirements. The loop stops early when all checks are passed.

4. Case Study: LLM Interpretability Research

To demonstrate the workflow, PRISM-X was tested on interpretability research topics. NOVA initially produced an abstract with two fabricated citations and ambiguous claims.

- The Guardian removed fabricated references.
- The Attacker detected overclaiming and forced a grounded rewrite.
- The Observer flagged low novelty and triggered another iteration.
- The final output was more specific, verifiable, and original.

5. Results

Experiments compared NOVA alone versus NOVA supervised by AEGIS X.

- Unsafe or policy-violating outputs: 37% reduction
- Idea novelty: 22% improvement
- Fabricated citations: significantly reduced
- Latency overhead: approximately 18–25%, considered acceptable

The evaluation shows that multi-agent governance greatly improves the trustworthiness and creativity of autonomous research ideation.

6. Conclusion

PRISM-X demonstrates that layered governance meaningfully enhances the safety, novelty, and academic quality of LLM-generated research ideas. By combining NOVA's generative capability with AEGIS X's tri-agent oversight system, the pipeline reduces hallucinations, filters unsafe content, and ensures outputs are grounded and credible. Early results indicate that governance-augmented ideation systems like PRISM-X could help shape the future of trustworthy AI-assisted research.