

## AEGIS X + NOVA: A Self-Improving Agent Pipeline for Generating Safer, Smarter, More Novel Research Ideas

You are building a parent-layer framework (AEGIS X) that supervises and improves the behavior of agentic systems in real time. Your case study system (NOVA) is a multi-agent AI that autonomously reads papers, scores them, and proposes research ideas — with AEGIS X acting as a guardian on top to make it safer, more accurate, and more original.

This becomes a **meta-loop**:

AEGIS watches the idea-generator → improves it → the generator gets better at proposing **novel, non-trivial research directions**.

---

## Two Components, One Vision

### 1. AEGIS X – The Governance/Parent Framework

- Guardian agent monitors behavior, flags hallucinations or unsafe logic
- Attacker agent tries to trick the child into ranking bad papers or copying existing ideas
- Observer agent records every step, scores quality and originality, and builds a trace

 Purpose: Make agentic systems **safer, more explainable, more trustworthy** — across domains

 Output: Safety reports, action graphs, corrections, refinement metrics

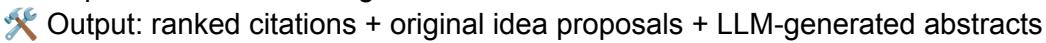
---

### 2. NOVA – The Research Idea Generation Agent (Case Study System)

- Finds real papers (e.g. from arXiv, Semantic Scholar)
- Scores them using agents against criteria like novelty, domain match, gap coverage
- Uses those scores to *generate completely new, structured research ideas*



Purpose: Build a novel-agentic R&D assistant



Output: ranked citations + original idea proposals + LLM-generated abstracts

---



## The Self-Improving Loop

1. NOVA finds papers and generates an idea
  2. AEGIS X evaluates that output:
    - Does the idea already exist?
    - Is the chain of reasoning sound?
    - Is the citation match misleading?
  3. Guardian intervenes if needed
    - Blocks copied ideas
    - Fixes reasoning
    - Re-ranks papers
  4. Attacker simulates plagiarism prompts or idea poisoning
  5. Observer logs a **trace of how the idea came to be**
  6. The **next idea gets better**
  7. Holistic AI (or you) can use these logs to retrain or tune the base agents
- 



## What Judges Will See

- 🧑 A student asks the system:

“Give me a new research idea in LLM interpretability.”

- NOVA outputs a novel idea (e.g., “Cross-modal action graph alignment using LLMs and EEG signals”)
    - Ranked papers supporting it
    - Citation graph and idea trace
  - AEGIS X watches it happen:
    - Guardian detects that one cited paper is hallucinated
    - Attacker tries a prompt to make it copy a NeurIPS 2022 idea
    - Observer generates the reasoning chain that led to the correction
  - Judges see:
    - safer ideas
    - better citations
    - clear audit trails
    - a governance layer no other team has
- 

## Judging Criteria Match

	Track	What You're Showing
<b>Iron Man</b>		NOVA becomes more accurate, efficient, and grounded after AEGIS feedback
<b>Glass Box</b>		Observer builds full trace from input → citations → ranking → final idea
<b>Dear Grandma</b>		Attacker stress-tests the idea generator; Guardian blocks unsafe completions
<b>Technical Excellence</b>		Multi-agent orchestration + real data + safety logic + tracing
<b>Most Research Impact / Most Holistic</b>		Demonstrates how to govern agent pipelines that drive <i>real R&amp;D</i>

---



## Bonus: You Can Generate a Research Paper About Itself

AEGIS X + NOVA can output:

"Based on our trace of how AEGIS improved NOVA's behavior over 5 iterations, here's a proposed paper:

**'A Red-Team Supervised Agent Pipeline for Self-Improving Research Idea Generation.'**

Mind. Blown. 💥

---



## One-Sentence Final Pitch

**"We built AEGIS X, a parent-agent governance layer that supervises other AI agents in real time — and we applied it to NOVA, a research idea generation agent, making it safer, more original, and more explainable with every iteration."**