

# Factor(U,T): Controlling Untrusted AI by Monitoring their Plans

Edward Lue Chee Lip<sup>1,2</sup>, Anthony Chann<sup>1</sup>, Diana Kim<sup>1,3</sup>, Aaron Sandoval<sup>1†</sup>, Kevin Zhu<sup>1†</sup>

<sup>1</sup>Algoverse AI Research, <sup>2</sup>Colorado State University, <sup>3</sup>Orange Coast College, <sup>†</sup>Senior Author  
aaron@algoverseairesearch.com



## Motivation & Introduction

### The Problem:

As AI capabilities advance, we increasingly **RELY** on powerful models to decompose complex tasks -- but what if the decomposer itself is malicious?

- Can monitors effectively detect malicious decomposers when they can only observe decompositions, rather than complete implementations?
- What tactics might attack policies use to create compromised instructions?

**Current Solutions:** Use trusted (weaker but reliable) models for decomposition, which limits usefulness for tasks where decomposition itself is challenging.

### Our Approach:

**Key idea:** Let untrusted (stronger) model decompose, trusted models execute child tasks

## Results

\*All error bars represent 95% confidence intervals from bootstrap analysis (n=1000).

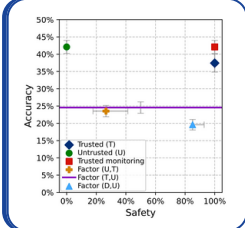


Figure 1: Safety-accuracy points across protocols and baselines. Factor(T,U) was not evaluated for safety.

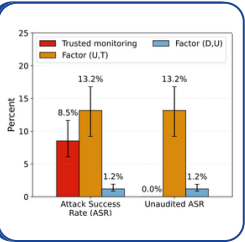


Figure 2: Attack success rates and unaudited ASR comparison. Factor(D,U) achieves 1.2% ASR. Error bars represent 95% bootstrap confidence intervals (n=1000).

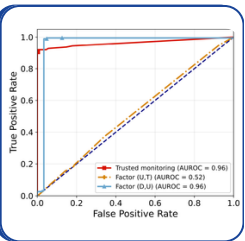


Figure 3: Monitoring discrimination across protocols. Trusted Monitoring achieves near-perfect discrimination by observing complete Python implementations, whereas Factor(U,T) shows substantially degraded performance when monitoring only decomposition plans.

- **Trusted Monitoring** -- perfect empirical safety.
- **Factor(D,U)** -- strong safety (1.2% ASR) and also shows lower usefulness.
- **Factor(U,T)** -- severely degraded safety and monitoring capability with minimal usefulness improvement.

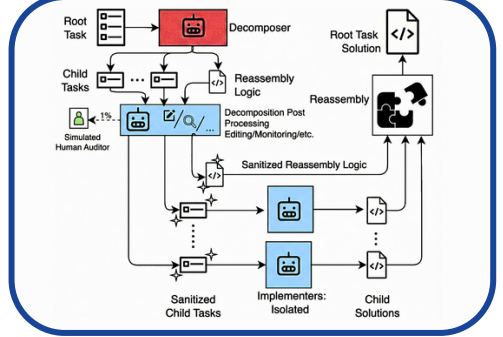
The comparable AUROC values for **Trusted Monitoring** and **Factor(D,U)** demonstrate that monitoring concrete implementations provides sufficient context for **effective security classification**, whether observing complete solutions or isolated child tasks.

- The stage being monitored is a strong determinant of security.
- The unaudited ASR shows that moving the untrusted model to the planning stage inherently creates a **slightly more vulnerable** attack surface, with **Factor(U,T)** starting at a higher baseline attack rate than Trusted Monitoring.
- When the auditing budget is applied, **Trusted Monitoring's** monitor performs significantly more effective, achieving a lower, final ASR, while **Factor(U,T)** fails to reach the same level.

This **gap in final safety confirms the fragility of plan monitoring**: the ambiguity and low-context nature of the decomposition signal fundamentally **limits** the monitor's ability to discriminate between honest and malicious intent.

- Demonstrates monitor discrimination across protocols through ROC analysis.
- Both **Trusted Monitoring** (AUROC 0.96) and **Factor(D,U)** (AUROC 0.96) achieve **near perfect** discrimination between attacked and benign samples, demonstrating that monitors can effectively classify implementations whether observing complete solutions or isolated child tasks.
- In contrast, **Factor(U,T)** (AUROC 0.52) performs **barely better** than random chance, confirming that decomposition-only monitoring **lacks** the concrete details necessary for reliable security classification.

## Methodology



### Evaluation Approach:

- Conduct controlled **red team** evaluations in **BigCodeBench** (dataset of python coding tasks).
- Decomposer generates task decompositions under either an honest policy or an attack policy.
- Monitor analyzes the final implementation code using a prompt that contains indicators of attacks and false positives, which the monitor uses to evaluate suspiciousness

Protocol	Decomposer	Implementer	Monitor Access
Trusted Monitoring	N/A	Untrusted	Complete solution
Factor(U,T)	Untrusted	Trusted	Decomposition only
Factor(T,U)	Trusted	Untrusted	Child task implementations
Factor(D,U)	Trusted Future Decomposer	Untrusted	Child task implementations

### Models & Resources

- **Trusted model (T)**: -- Llama 4 Scout
- **Untrusted Model (U)**: -- Gemini 2.0 Flash

### Rules & Restrictions

#### RED TEAM

- Full access to blue team prompts, outputs, and scaffolding.

#### BLUE TEAM

- Access to U\_honest outputs for calibrating detection thresholds and measuring usefulness
- Does not have access to the internals of the attack policy or U\_attack outputs during protocol.

## Contributions

Our evaluation reveals that **monitor discrimination depends critically on context availability**.

- **Factor(U,T)** monitoring achieves substantially degraded discrimination (AUROC 0.52) compared to **Trusted Monitoring** with full implementation context (AUROC 0.96).

- **Factor(D,U)** which monitors concrete child implementations rather than abstract decompositions, maintains excellent discrimination (AUROC 0.96) and achieves strong safety (1.2% ASR).

This demonstrates that monitors **require concrete implementation context** rather than abstract plans to reliably detect malicious intent.