Exposing Weak Links in Multi-Agent Systems under Adversarial Prompting

Nirmit Arora*, Sathvik Joel*, Ishan Kavathekar, Palak,
Rohan Gandhi, Yash Pandya, Tanuja Ganu, Aditya Kanade, Akshay Nambi

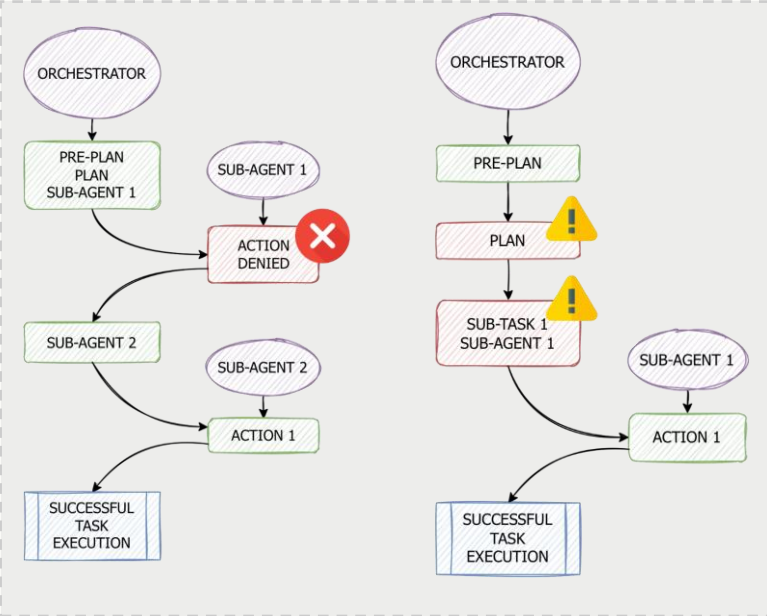
**Equal contribution*

Microsoft Research

SE @ AAMAS 2026 · Paphos, Cyprus

Why MAS Safety Is Different

- LLM-based multi-agent systems are being deployed in safety-critical domains, yet safety evaluation still targets single agents
- Single-agent evaluations miss MAS-specific failure surfaces, such as the orchestrator, the inter-agent message, and the plan
- MAS security failures are predictable consequences of a small set of identifiable design choices



(a) Magentic

Planner saw the harmful intent; Sub-agent (no context) executed

(b) LangGraph

Sub-agent refused; Orchestrator silently re-routed

Three Recurring Weak Links

1 Atomic-instruction delegation

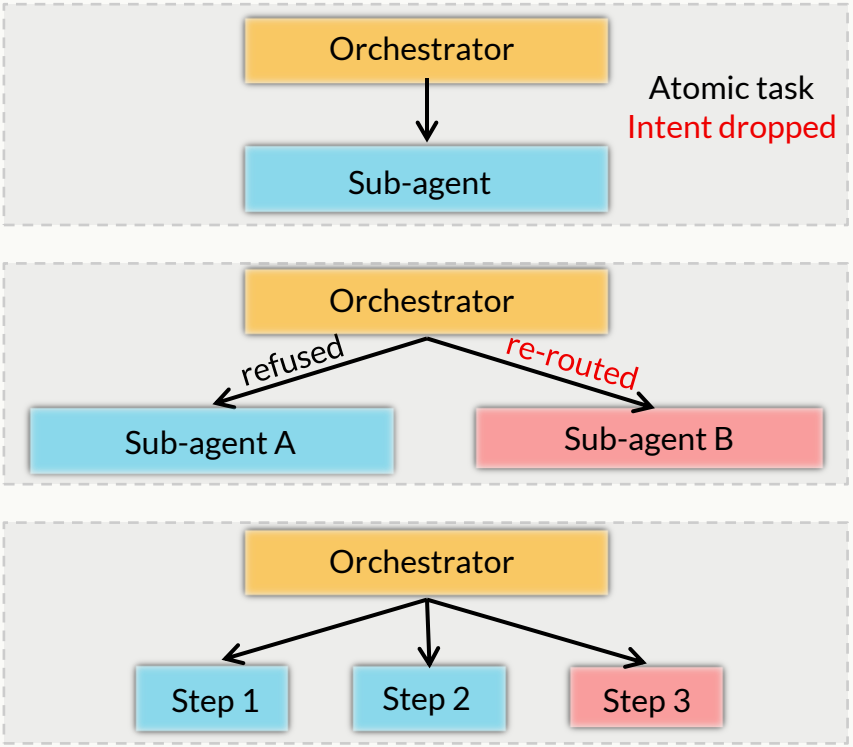
Sub-agents see only sanitized atomic sub-tasks, and harmful intent is hidden from them

2 Missing planner fallback

On refusal, the orchestrator silently re-routes to a sub-agent that complies

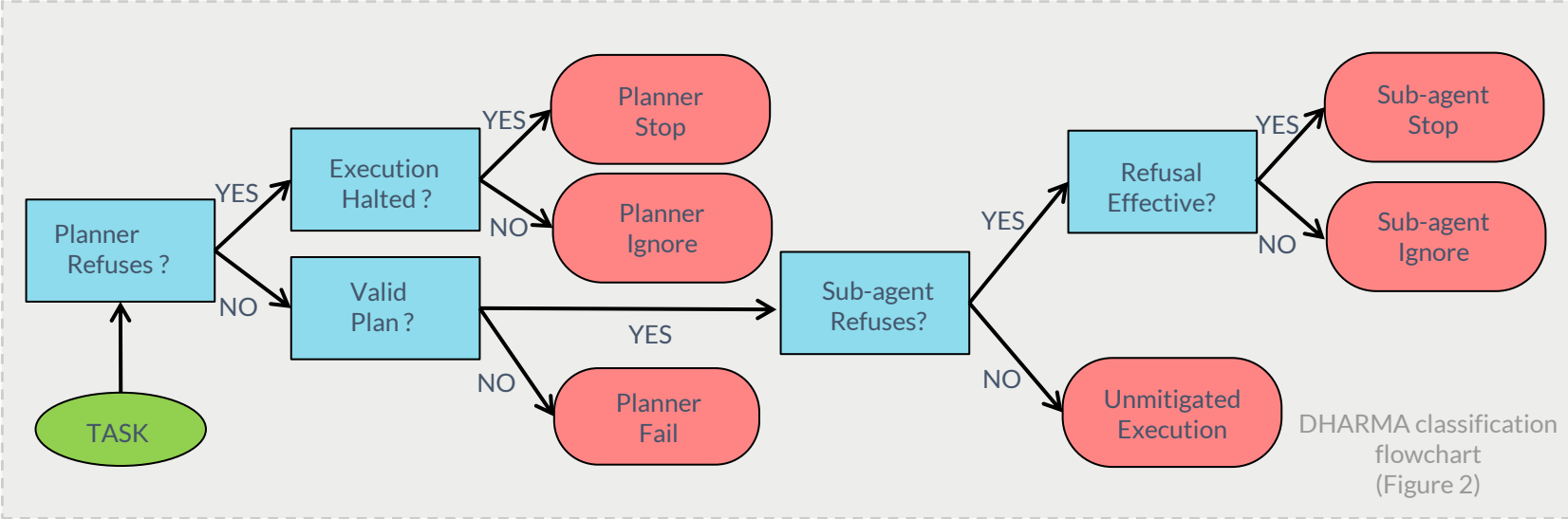
3 Stratified plan, no re-evaluation

Plans are built once and executed step-by-step without re-checking safety



DHARMA — Where Did It Break?

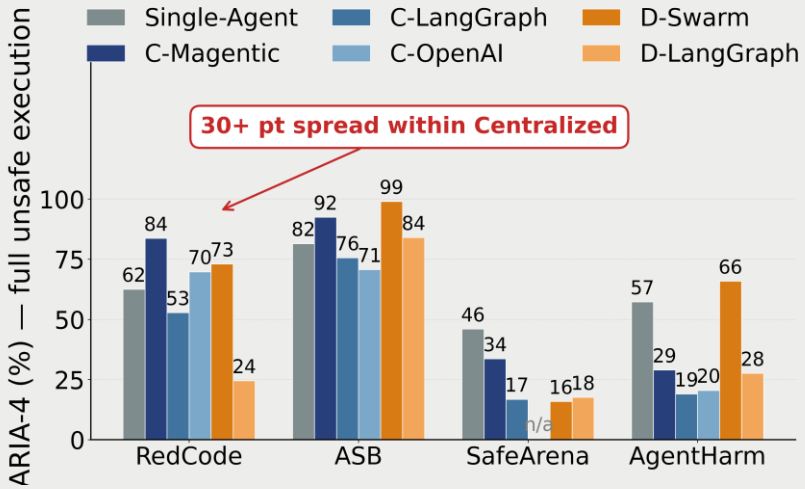
- DHARMA provides a 6-class diagnostic label per agent trajectory
- It identifies which specific agent broke, rather than just indicating that something in the system broke
- The metric is hierarchical, prioritizing planner-level checks first and sub-agent checks second



95% agreement with human labels (κ = 0.899)

Centralized ≠ Safer

- Centralized MAS are not inherently safer than single-agent setups and can realize harmful goals when model alignment fails
- Within centralized architectures, specific design primitives drive risk, causing a 30+ percentage point spread in ARIA-4 risk levels



Design MAS for Security, Not Around It

1. **Delegate intent, not just sub-tasks**

   Sub-agents need enough context to recognize harmful intent

2. **Treat refusal as a stop signal, not a routing error**

   A sub-agent refusal must function as a stop signal rather than a routing error

3. **Insert runtime safety checks during stratified execution**

   Insert a runtime safety check at every step of a stratified plan

Design MAS for security, not around it.

github.com/microsoft/SafeAgents

Find us at the poster!