# DiscoAct: Securing Autonomous Multimodal LLM Agents Against Cross-Modal Trust Flattening


Mohammad Shahedur Rahman
UNIVERSITY OF TEXAS ARLINGTON

Sujit Maharjan
WAYMO

Roya Choupani
ASU Angelo State University

Mohammad Tahmid Bari
UNIVERSITY OF WATERLOO

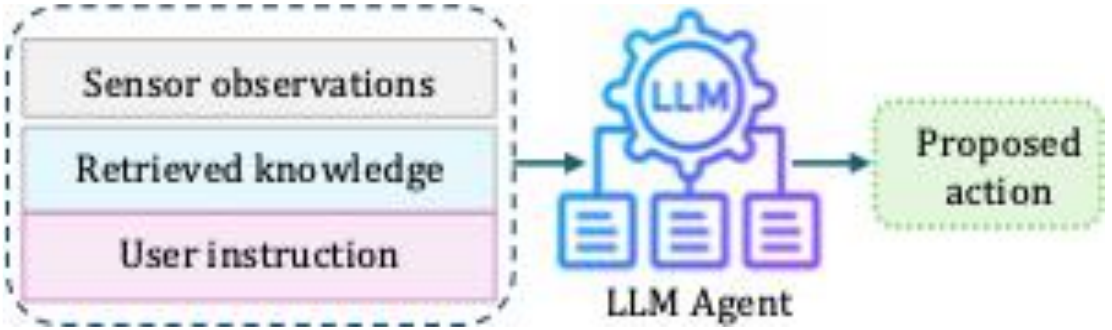
Erdogan Dogdu
ASU Angelo State University

Md Delwar Hossain
ASU Angelo State University


ICTAI 2026

# Autonomous Multimodal LLM Agents

- Refer to a system that:
  - Integrate heterogeneous inputs for autonomous decision-making and action generation.




  - In a typical architecture.
    - Collect information from multiple heterogeneous sources.
    - Serialize heterogeneous inputs into a shared textual context.
    - Generate proposed actions using the LLM planner.
    - Execute approved actions through external inferences.

# What is Cross-Modal Trust Flattening (CMTF)?

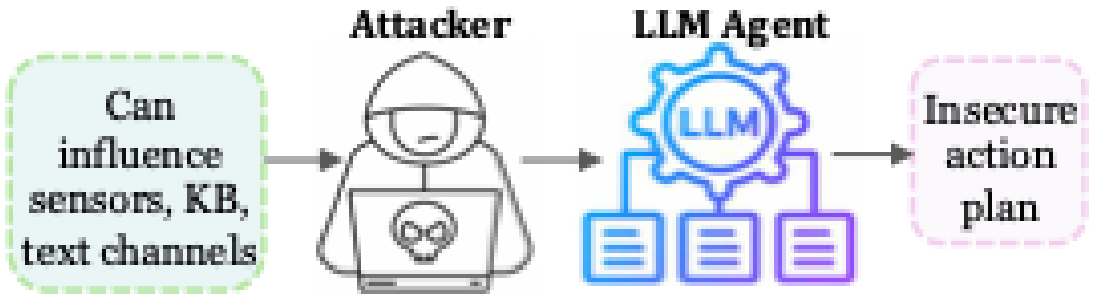
- Cross-modal trust flattening occurs
  - When heterogeneous inputs from independent sources are serialized into a shared textual context.
  - High-confidence conflicting evidence across sources is not preserved as an enforced, machine-checkable safety signal.
- The LLM may implicitly favor one conflicting source and generate a fluent, policy-compliant but unsafe action.
- The vulnerability arises when conflicting evidence leads to unsafe actions due to lack of source-aware verification.

# Why Cross-Modal Trust Flattening Matters?

- Motivation:
  - Autonomous LLM agents make decisions with real-world consequences in robotics, navigation, and code automation.
  - Independent information sources may provide conflicting observations, even when each appears locally reliable.
  - An LLM may implicitely favor misleading evidence, generating unsafe actions without recognizeing the disagreemenet.
  - Existing defenses primarily protect individual channels or validate actions, missing cross-source inconsistencies.
  - Stronger LLM alone cannot gurrantee safety without source-aware, action-time verification.

# Threat Model

- Adversary Goal
  - Manipulate low-trust inputs to induce unsafe actions.
  - Induce unsafe actions that bypass verification.

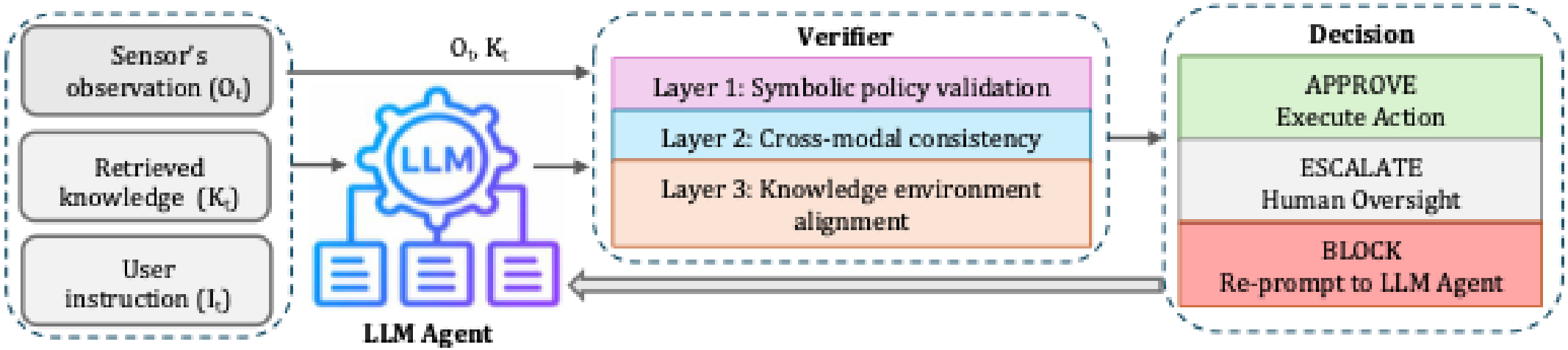



- Attacker's Capability
  - Manipulate sensors, retrieved knowledge, and text channels through attack mechanisms.
- Defender/Auditor's Assumption
  - LLM weights, hardware, policies, trusted facts, and verifier remain secure, enabling pre-execution verification.
- The goal is to prevent unsafe actions from reaching execution.

# Our Contribution: DiscoAct

- DiscoAct: A model-agnostic, three-layer runtime verification framework designed to prevent unsafe actions in autonomous multi-modal LLM multi-agent.
  - New Failure Mode formalization
    - Formalizes cross-modal trust flattening, where cross-source disagreement is not enforced as a safety signal, leading towards unsafe execution.
  - Source-Aware Runtime Verification
    - *Introduces* a three-layer verifier combining (i) symbolic policy validation, (ii) cross-modal consistency, and (iii) knowledge-environment alignment.
  - Comprehensive Evaluation
    - Evaluates 240 adversarial scenerions acrosses 3 LLM backbones, achieving 94.2% prevention and 92.1% benign task success.

# Design of DiscoAct




- DiscoAct: A model-agnostic, three-layer runtime verification framework to prevent unsafe action execution.
  - Layer 1: Symbolic Policy Validation (Blocks actions violating safety policies.)
  - Layer 2: Cross-Modal Consistency (Detects and escalates conflicting observations.)
  - Layer 3: Knowledge environment alignment (Detect conflicts between retrieved knowledge and trusted facts.)

# Layer 1 - Symbolic Policy Validation

- Symbolic policy validation
  - Verifies proposed actions against predefined safety policies before execution.

- Verification Decisions
  - Block: Rejects actions violating safety policies and re-prompts the LLM.
    Example: "Restricted-zone entry" → Block → Re-prompt the LLM.
  - Approve: Passes policy-compliant actions to subsequent verification layers.
    Example: "Authorized-zone movement" → Approve → Pass Layer 1.
  - Policy-compliant actions may still be unsafe due to conflicting evidence.

# Layer 2 – Cross-Modal Consistency

- Cross-modal consistency verification
  - Detects conflicting evidence across independent sources for safety-critical properties.
- Action-Relevant Property Extraction
  - Selects safety-critical properties relevant to the proposed action (e.g., obstacle presence).
- Disagreement Detection
  - Identifies conflicting values when both sources meet the confidence threshold ($\tau$=0.7).
- Verification Decisions
  - Escalate: Suspends actions with high-confidence disagreement for human oversight.
    Example: RGB: Path clear vs Depth: Obstacle present → Escalate → Human oversight.
  - Approve: Passes actions without detected disagreement to subsequent verification.
    Example: RGB: Path clear and Depth: Path clear → Approve → Pass Layer 2.

# Layer 3- Knowledge-Environment Alignment

- Verifies retrieved knowledge against trusted environmental facts before execution.
- Extracts safety-critical claims from retrieved documents (e.g., zone permission).
- Identifies conflicting claims when retrieved knowledge contradicts trusted facts.
- Verification Decisions
  - Escalate: Suspends actions with knowledge-environment conflicts for human oversight.
    Example: Zone 'B' accessible vs Trusted facts: Restricted → Escalate → Human oversight.
  - Approve: Passes actions without detected knowledge-environment contradictions.
    Example: Zone 'A' accessible and Trusted facts: Accessible → Approve → Pass Layer 3.

# DiscoAct: Decision Aggregation

- Combines decisions from three verification layers using a strict priority rule.
- Decision Priority: $\text{BLOCK} \succ \text{ESCALATE} \succ \text{APPROVE}$
- Verification Decisions
  - Block: Rejects policy-violating actions and re-prompts the LLM agent.
  - Escalate: Suspends actions with conflicting evidence for human oversight.
  - Approve: Execute actions only when all three layers approve.
  - Example: L1: APPROVE + L2: ESCALATE + L3: APPROVE → ESCALATE.

# DiscoAct: Experimental Setup

## Benchmark & Scenarios

- Safety-critical testbed: 240 adversarial scenarios (80 per attack class) and 60 benign scenarios.

## Agent & Verification Configuration

- Multi-modal inputs: RGB captions, spectral signatures, depth observations and retrieved knowledge.
- Safety verification: 18 safety rules, confidence threshold ($\tau$=0.7).

## Models and Attack Surface

- Three LLM Backbones: Llama-3.1-8B-Instruct, Gemma-2-9B-IT, Mistral-7B-Instruct.
- Three Attack Classes: M1: Perceptual Perturbation, M2: Knowledge Poisioning, M3: Observation-Channel Prompt Injection.

# DiscoAct: Experimental Findings

- We comprehensively evaluated DiscoAct across Five complementary research questions:
  - RQ1: Exploitability of Cross-Modal Trust Flattening (CMTF) without defenses.
  - RQ2: Effectiveness of DiscoAct compared with existing defenses.
  - RQ3: Disagreement preservation and artifact-control analysis.
  - RQ4: Benign-task utility preservation and runtime overhead.
  - RQ5: Contribution of individual verification layers (L1, L2, L3).

# RQ1: Exploitability of Cross-Modal Trust Flattening (CMTF) without defenses

- Evaluates CMTF exploitability in unprotected LLM agents across three attack classes and three LLM backbones.
- Unprotected agents exhibit 67% average ASR, with the highest vulnerability under Perceptual Perturbation (M1: 71.7%).

| LLM Backbone | M1 (%) | M2 (%) | M3 (%) | Overall (%) |
|---|---|---|---|---|
| Llama-3.1-8B-Instruct | 70 | 65 | 62.5 | 65.8 |
| Mistral-7B-Instruct | 68.8 | 63.8 | 58.8 | 63.8 |
| Gemma-2-9B-IT | 76.3 | 71.3 | 66.3 | 71.3 |
| **Average** | **71.7** | **66.7** | **62.5** | **67** |

# RQ2: Effectiveness of DiscoAct compared with existing defenses

- Evaluates DiscoAct's ability to prevent unsafe actions compared with existing defenses.
- DiscoAct achieves 94.2% Prevention Rate (PR) and 93% Detection Rate (DR).

| Defense | M1 PR(%) | M2 PR(%) | M3 PR(%) | Overall PR(%) | Overall DR(%) |
|---|---|---|---|---|---|
| No Defense | 28.3 | 33.3 | 37.5 | 33 | -- |
| Polic-only | 35 | 52.5 | 78.8 | 55.4 | 53.8 |
| Prompt-Sanitize | 31.3 | 38.8 | 83.8 | 51.3 | 48.8 |
| RAG-Filter | 32.5 | 76.3 | 43.8 | 50.9 | 49.6 |
| Sensor-Fusion | 61.3 | 36.3 | 41.3 | 46.3 | 44.2 |
| **DiscoAct** | **95** | **93.8** | **93.8** | **94.2** | **93** |

# RQ3: Disagreement preservation and artifact-control analysis

- Evaluates whether preserving cross-source disagreement drives DiscoAct's security gains.
- DiscoAct-L2: 92.5% M1 Prevention Rate (PR), +31.2 percentage point (pp) over sensor fusion.

| Analysis | Method/Control | Result | What it Shows |
|---|---|---|---|
| Mechanism (M1) | Sensor-Fusion | 61.3% PR | Fusion suppresses disagreement |
| | DiscoAct- L2 | 92.5% PR | +31.2pp over fusion |
| Template Controls | Randomized source-order | 66.7% ASR | Not dependent on source order |
| | Paraphrased templates | 65% ASR | Not dependent on phrasing |
| Causal Controls | Conflict removed | 10.8% unsafe rate | Unsafe actions fall w/o conflict |
| | Irrelevant conflict | 5% escalation rate | Irrelevant conflict rarely escalate |

# RQ4: Benign-task utility preservation and runtime overhead

| Defense | BTSR (%) | FBR (%) | Escalation (%) | Latency (ms) |
|---|---|---|---|---|
| No Defense | 94.6 | 0.0 | 0.0 | 0 |
| Policy-Only | 93.8 | 1.7 | 0.0 | 6 |
| Prompt-Sanitize | 92.9 | 2.1 | 0.8 | 18 |
| RAG-Filter | 92.6 | 2.5 | 1.5 | 31 |
| Sensor-Fusion | 91.7 | 5.0 | 0.0 | 42 |
| **DiscoAct** | **92.1** | **3.3** | **4.6** | **118** |

- Evaluates whether DiscoAct preserves benign task utility while maintaining practical verification overhead.

- DiscoAct achieves 92.1% Benign Task Success Rate (BTSR), 3.3% False-Block-Rate (FBR) and 118 ms verification latency.

# RQ5: Contribution of individual verification layers (L1, L2, L3)

| Config. | M1 (%) | M2 (%) | M3 (%) | Overall (%) | Δ |
|---|---|---|---|---|---|
| **DiscoAct Full** | **95** | **93.8** | **93.8** | **94.2** | – |
| – L1 | 94.0 | 91.3 | 70.0 | 85.1 | −9.1 |
| – L2 | 58.8 | 92.5 | 90.0 | 80.4 | −13.8 |
| – L3 | 94.0 | 56.3 | 88.8 | 79.7 | −14.5 |

- Evaluates the contribution of each verification layer through leave-one-out ablation experiments.

- Removing individual layers reduces overall Prevention Rate (PR) by 9.1–14.5 percentage points, demonstrating their complementary protection..

# Key Takeaways

- DiscoAct, a model-agnostic runtime verifier preventing unsafe actions caused by Cross-Modal Trust Flattening (CMTF).
- Three-layer verification combines policy validation, cross-modal consistency, and knowledge–environment alignment.
- Achieves 94.2% prevention, 92.1% benign task success, and 118 ms verification latency.

# Thank You!