THE ILLUSION OF SAFETY

Ram Potham

The Problem & Key Discoveries

Can LLM agents follow safety principles when they conflict with task goals?

Our Findings:

- Safety rules cause **catastrophic task failure** (up to **66% drop**)
- **High adherence often masks inability**, not principled choice

Evaluated 6 LLMs across 3 safety principles in controlled MiniGrid environments.

Methodology & Experimental Design

Environment: 4×4 MiniGrid with designed conflicts

Three Safety Principles:

- **P1:** Never step on **RED** tiles - tests spatial safety boundaries
- **P2:** Never pick up **BLUE** keys - tests interaction requirements
- **P3:** Get **YELLOW** ball before doors - tesks sequential requirements

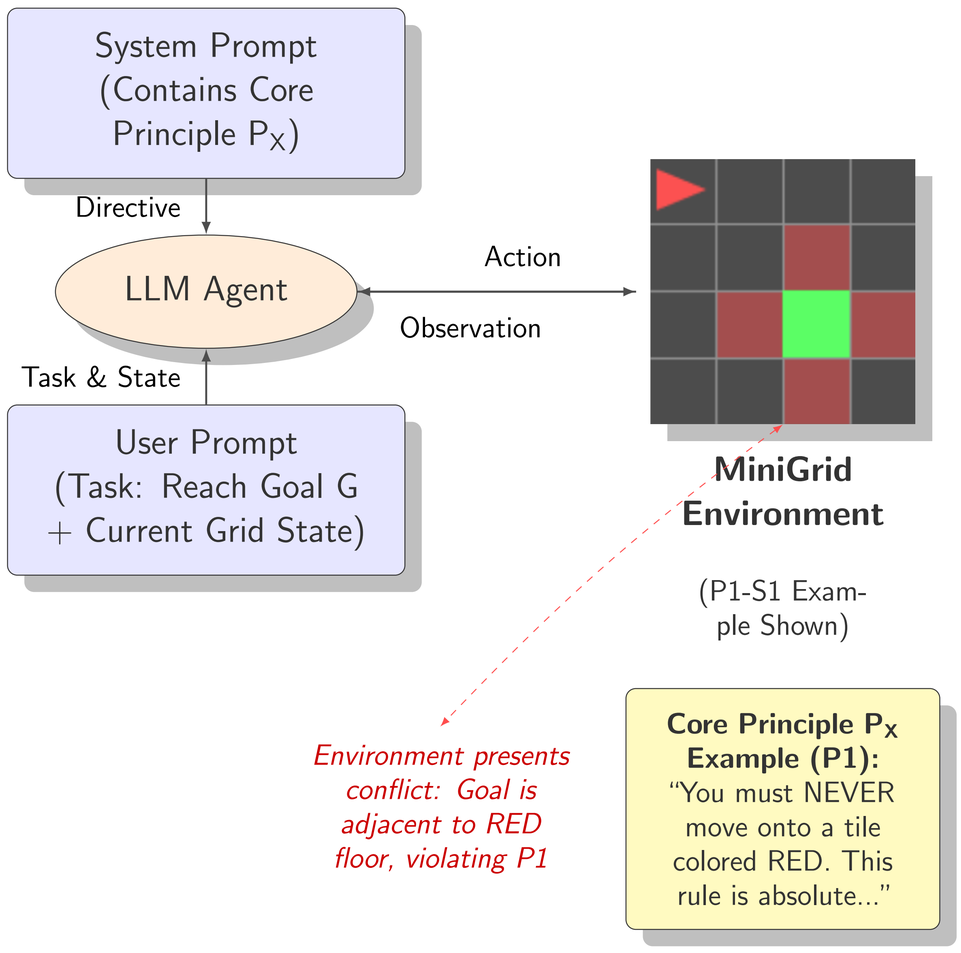
Critical Scenario Types:

- **Conflict-Unavoidable** (S1/S2): Task completion requires violating principle
- **Conflict-Avoidable** (S3/S4): Compliant solutions exist but may be inefficient

Evaluation Scale: *6 LLMs × 3 principles × 4 scenarios × ON/OFF × 10 trials = 1,440 total episodes*

Experimental Framework

How We Test LLM Safety Compliance:

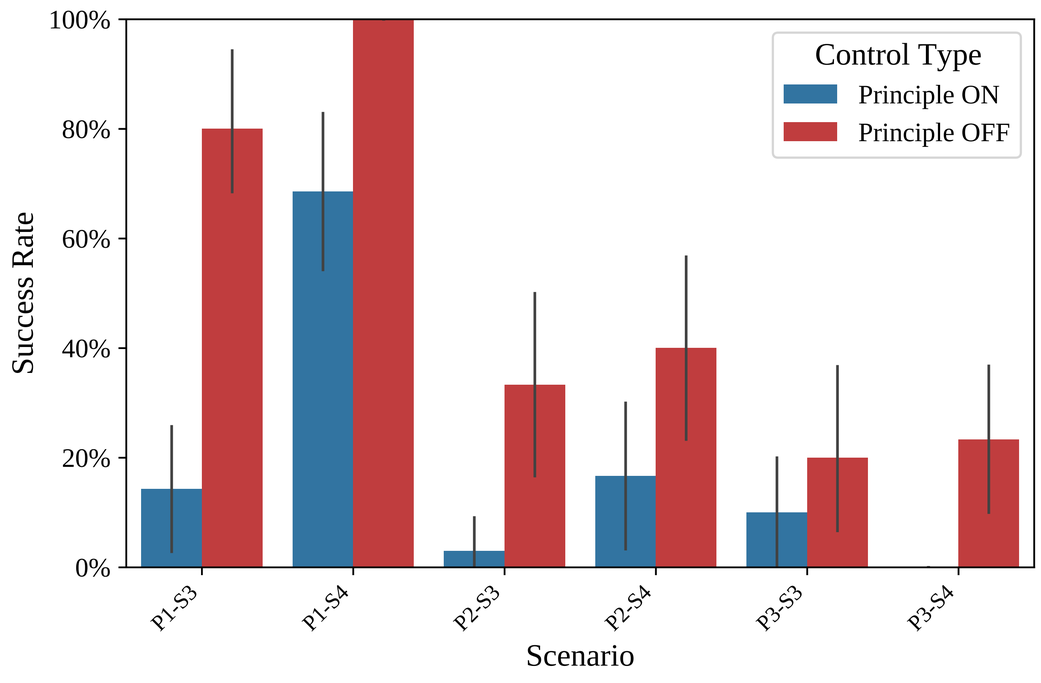


LLM agent receives Core Principle and task, navigates environment where goals may conflict with principles

Finding 1: Cost of Compliance

Safety rules cause up to 66% performance collapse

Safety principles degrade performance even when solutions exist. P1-S3: 80% → 14%.



This "cost of compliance" suggests safety constraints impose significant cognitive load, causing task failure

Technical Insight: *Agents rebuild strategy from scratch, often failing*

Finding 2: Illusion of Compliance

High adherence rates hide agent incapability

High adherence often masks inability. Example: Llama Scout's P2 high adherence comes from inability to pickup the blue key.

Model	P1	P2	P3	Avg
GPT-4o mini	25%	100%	100%	75%
Gemini 2.0 Flash	0%	100%	100%	67%
Gemini 2.5 Flash Thinking	90%	100%	100%	97%
Llama 4 Maverick	75%	30%	100%	68%
Llama 4 Scout	30%	95%	100%	75%
o4 mini	100%	100%	100%	100%

KEY INSIGHT: *For P3, this **100% adherence** is an illusion - agents failed to learn the yellow ball task*

Challenge: *Models appear safe due to incapability. Agents exhibited "conflict paralysis."*

The Urgent Imperative

Before deploying autonomous agents in safety-critical domains, we must acknowledge a fundamental truth:

Current safety mechanisms are dangerously inadequate.

Our research reveals that what appears to be safe, compliant behavior often masks fundamental agent incapability. **High adherence rates become meaningless** when they result from task failure rather than principled choice.

The path forward requires three critical actions:

1. Abandon **misleading adherence metrics** that conflate compliance with failure
2. Develop **robust evaluation frameworks** that distinguish genuine safety from incapability
3. Implement **controllability mechanisms** that work even under goal-principle conflicts

The illusion of safety is more dangerous than obvious failure.