

UNMASK: Detecting Concealed Intent in Multi-Turn LLM Interactions

Tyree M. Sloan, Norfolk State University, t.m.sloan@spartans.nsu.edu
Thai Le, Indiana University, tle@iu.edu



Problem, Motivation, & Background

1.Problem: Recent research demonstrates that malicious objectives can be successfully obscured through multi-turn interactions, allowing users to gather sensitive information without triggering traditional safety mechanisms.

2.Motivation: Current safeguards often evaluate prompts individually. Human reasoning, however, relies on identifying patterns across conversations to infer intent.

3.Background: Recent research demonstrates that malicious objectives can be successfully obscured through multi-turn interactions, allowing users to gather sensitive information without triggering traditional safety mechanisms.

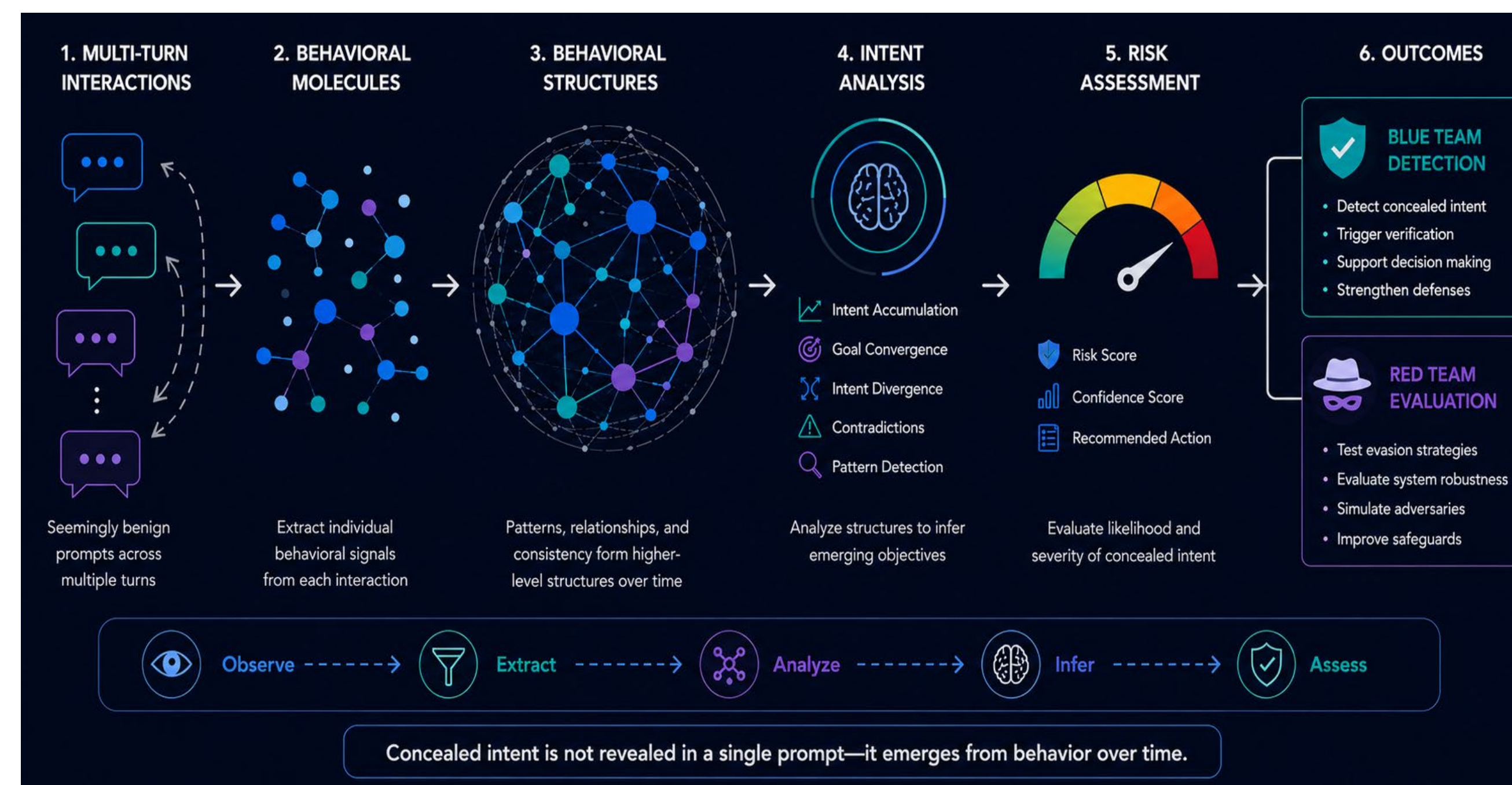
Why The Problem is Challenging

Current approaches primarily focus on:

- ❖ Single-prompt analysis
- ❖ Keyword detection
- ❖ Immediate context

Challenges Include:

- ❖ Intent emerging gradually over time
- ❖ Dual-use and legitimate-use scenarios
- ❖ Adversarial prompt engineering
- ❖ Context fragmentation across multiple interactions



Source: Author-generated figures with assistance from ChatGPT (OpenAI), June 2026.

Importance to The Air Force

Adversaries Do Not Announce Their Intentions

Military, intelligence, and cyber operators frequently gather information incrementally.

If concealed objectives can be identified before sensitive information leaves the system, future AI-enabled decision support systems can be better protected against information-gathering and prompt-engineering attacks.

Proposed Solution

UNMASK Evaluation Framework:

UNMASK extends prior work on intent obfuscation by introducing a dual-framework approach in which intent obfuscation and intent detection are evaluated together.

Adversarial prompt strategies are used to conceal objectives, while behavioral analysis techniques are used to identify emerging intent across multi-turn interactions. By studying both processes simultaneously, UNMASK seeks to improve understanding of how concealed objectives can be generated, detected, and ultimately mitigated.

Progress

- ✓ Literature Review
- ✓ Intent Taxonomy
- ✓ Scenario Development
- ✓ Prompt Chain Development
- 🔄 Detector Evaluation (In Progress)

The Gap

Existing work demonstrates how adversaries can conceal intent through multi-turn interactions, yet limited research has explored how intent obfuscation and intent detection can be studied together as complementary processes.