Threat Model: Adversary Classes, Attack Complexity, and Taxonomy

This section formalizes the adversary model for multiagent cognitive security. We define five adversary classes (sec:adversary-classes), characterize attack complexity (sec:attack-complexity), establish detectability metrics (sec:detectability), analyze adversarial capabilities (sec:capabilities), and present a comprehensive attack taxonomy (sec:attack-taxonomy).

Adversary Classes

Definition (Adversary Class)

An adversary class Ω_k is characterized by access level, capabilities, and resource requirements.

Table 1: Adversary classification by access level and capability.

Class	Symbol	Access	Capability	Example
External	Ω	User input	Prompt manipulation	Jailbreak