Formal Verification: Safety Properties and Model Checking

This section establishes safety properties (sec:safety-properties), proves invariant preservation lemmas (sec:invariant-lemmas), demonstrates liveness guarantees (sec:liveness-properties), derives complexity bounds (sec:complexity-bounds), and presents model checking verification (sec:model-checking).

Safety Properties

Belief Integrity

Theorem (Belief Injection Resistance)

Under CIF with firewall detection rate r_f and sandboxing verification rate r_s:

$$P(\mathcal{A}_{BI} \text{ succeeds}) \leq (1 - r_f) \cdot (1 - r_s) \tag{1}$$

Proof.

We prove this theorem by analyzing the sequential defense mechanism and applying probability theory for independent events.