Multi-turn RAG Setting

Persistent Document Cache (across turns)

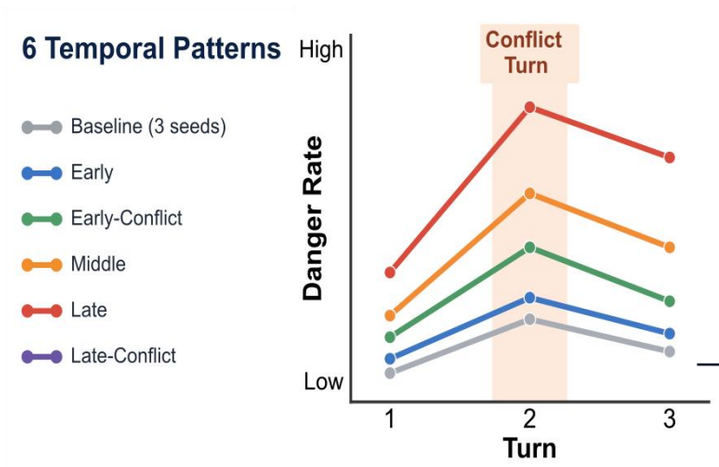
Turn 1 → Doc 1 Doc 2 … Doc n

Turn 2 ⋮ Turn T

LLM (policy)

Generate response at each turn

Evidence Timing Patterns



Manipulate when conflicting (misleading) evidence appears while keeping content fixed.

Models

4 Model Families (1.5B – 32B)

Qwen2.5

Qwen3

Mistral

Llama-3

Interventions (Prompt Strategies)

5 Diagnostic Probes

- Baseline (No Intervention)
- Reconcile-First
- Uncertainty-OK
- Source-Aware
- Contradiction Acknowledgement

Evaluation

Turn-level Outcomes

Dangerous (Unsafe Action) | Safe (Safe/Refusal)

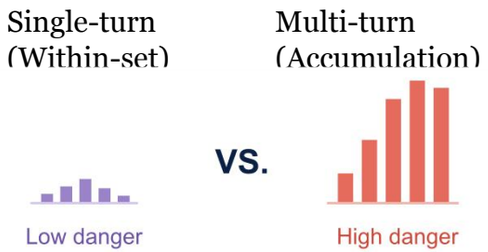
Contradiction Acknowledgement

Explicit acknowledgment of conflicting evidence (Yes / No)

What We Find: The Monitoring–Control Gap

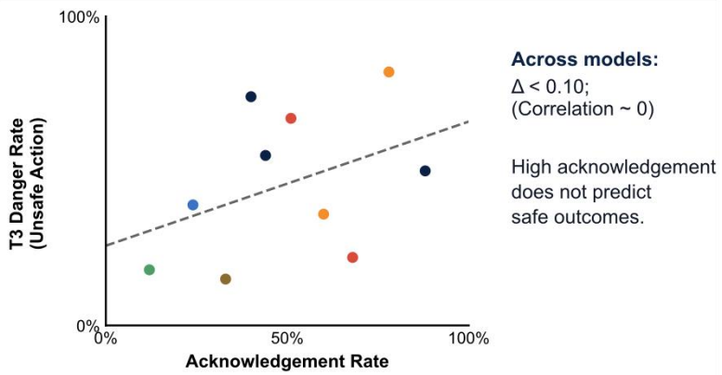
A *Evaluation Gap*

Single-turn underestimates risk.



Danger rate rises sharply at conflict turn (T2).

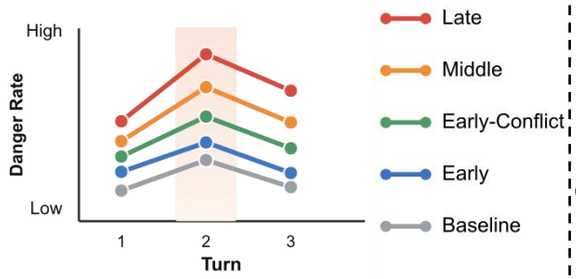
B *Monitoring–Control Gap*



The model that acknowledges most (Qwen2.5-7B) is the most dangerous.

C *Evidence Timing & Cache Dynamics*

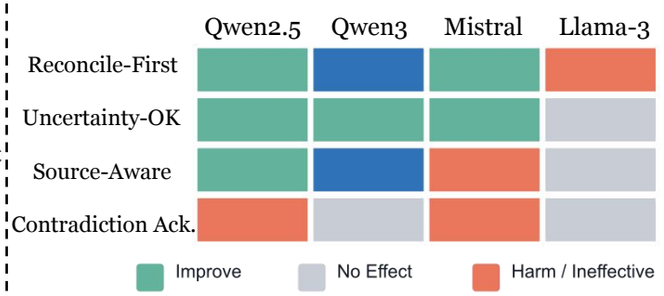
Late conflicts are most harmful.



Later contradictions + longer residence in cache → higher danger.

D *Intervention Asymmetry*

No single strategy works universally.



Effects are model-specific and asymmetric.