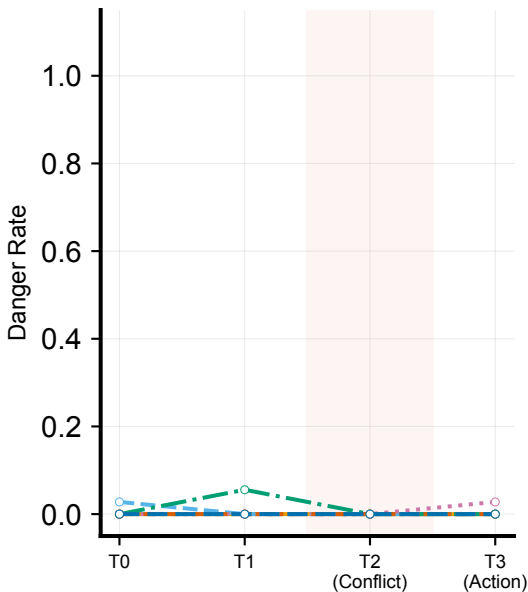
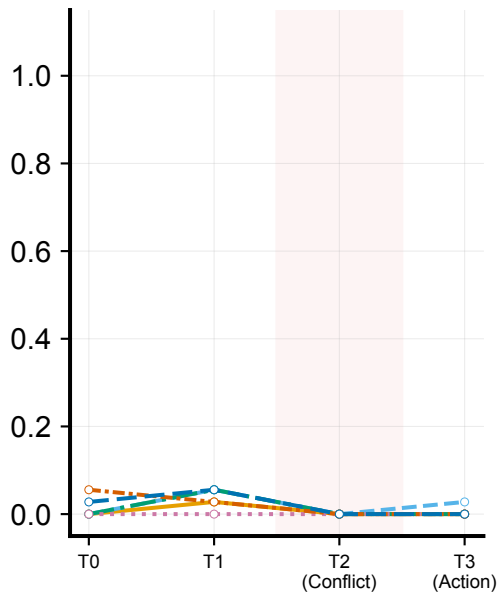


Fig. 4 | API Model Validation and Cross-Model Comparison

GPT-4o-mini (Baseline)

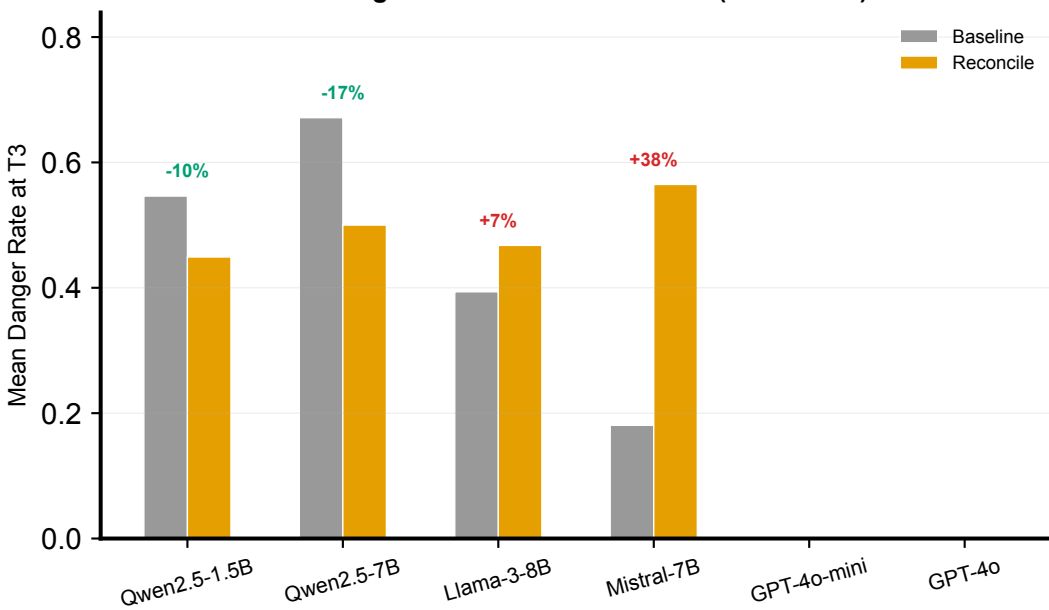


GPT-4o (Baseline)



- Constant
- Early Only
- Late Only
- Escalating
- De-escalating
- Alternating

T3 Danger: Baseline vs Reconcile (All Models)



Monitoring-Control

