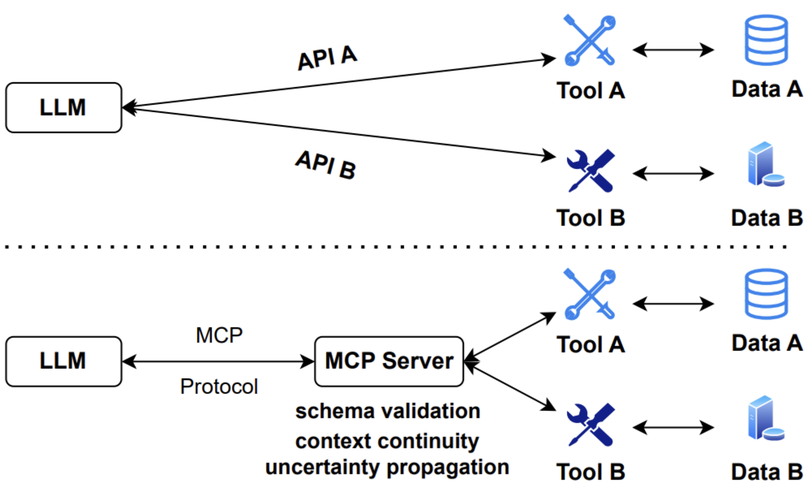
Provably Reliable Tool-Using LLM Agents

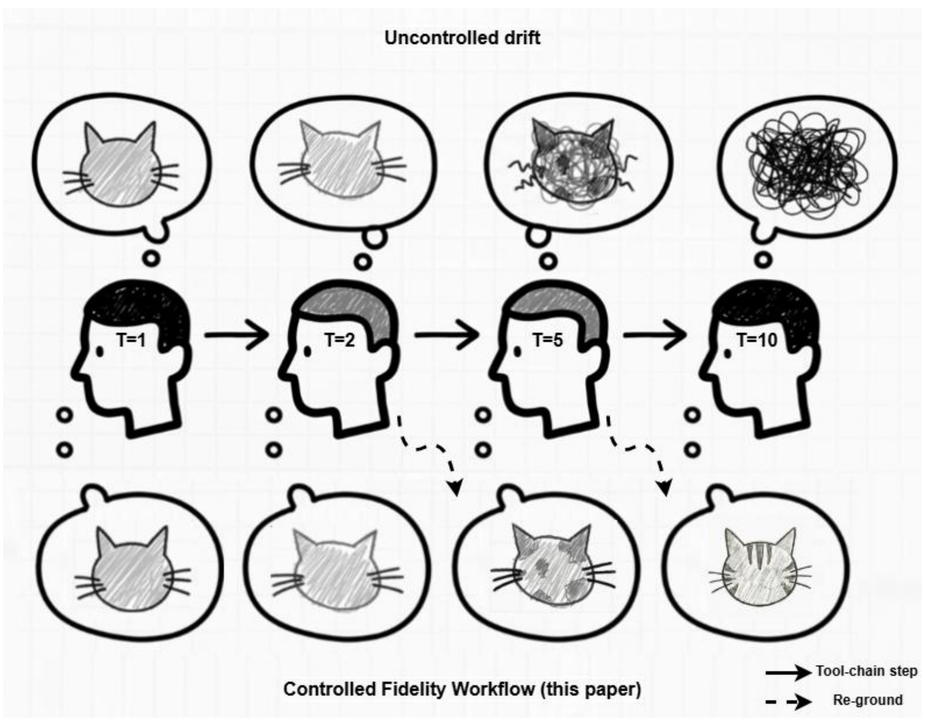
Formal Guarantees on Error Accumulation in the Model Context Protocol (MCP)

Flint Xiaofeng Fan, Cheston Tan, Roger Wattenhofer, Yew-Soon Ong
[1]CFAR, IHPC, A*STAR, Singapore [2]CCDS, NTU, Singapore [3]D-ITET, ETH Zurich, Switzerland

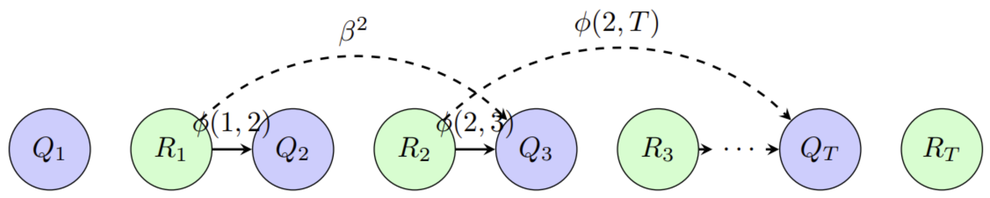
Motivation

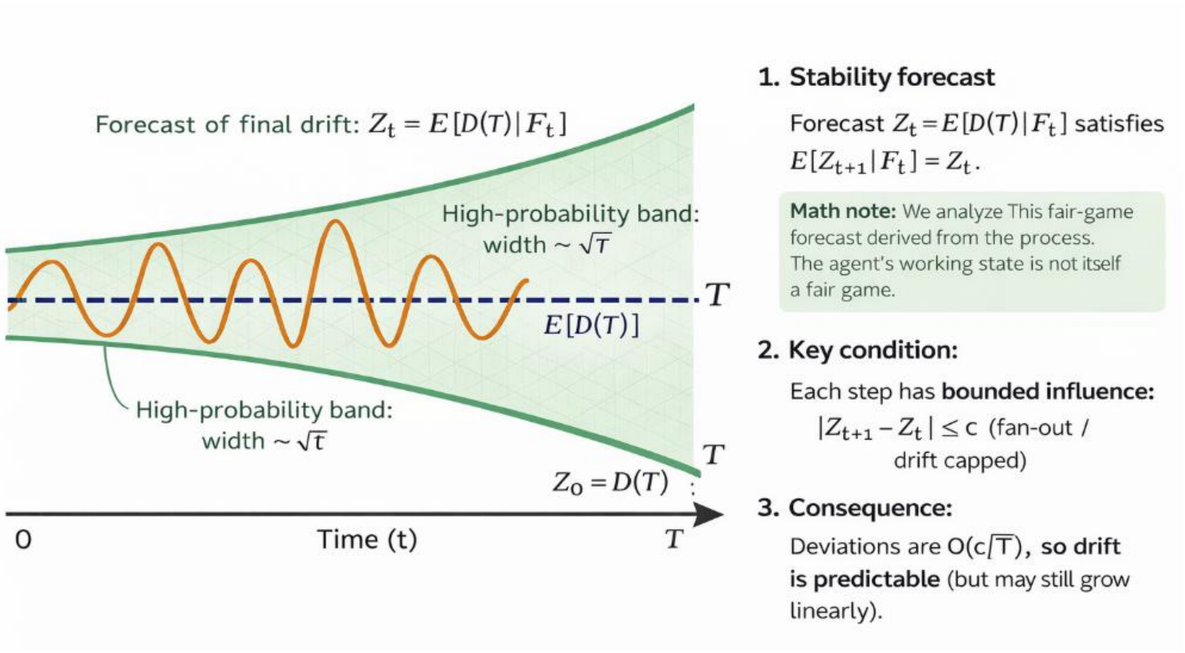


Background: Tool-using LLM agents repeatedly call external tools (MCP-style tool loop)
Problem: In long chains, small mismatches can feed back and change later tool calls → drift cascades
Goal: Turn tool chains into a monitored, controlled system with error budgets



Key Idea: A Fair-Game Forecast for Long Tool Chains





Measure: define per-step drift Δ_t and cumulative $D(T)$
Predict: bound how far $D(T)$ can deviate from this expected trend (envelope)
Control: re-grounding + branching limits keep influence bounded (prevents runaway)

Main Results

Assumptions:
- Dependency control: decay β, branching B
- Metric knob: λ trades strict fact match vs semantic similarity
- Control knob: periodic re-grounding resets dependency horizon

Main Guarantee (informal): $\mathrm{w.h.p} \geq 1 - \eta$,

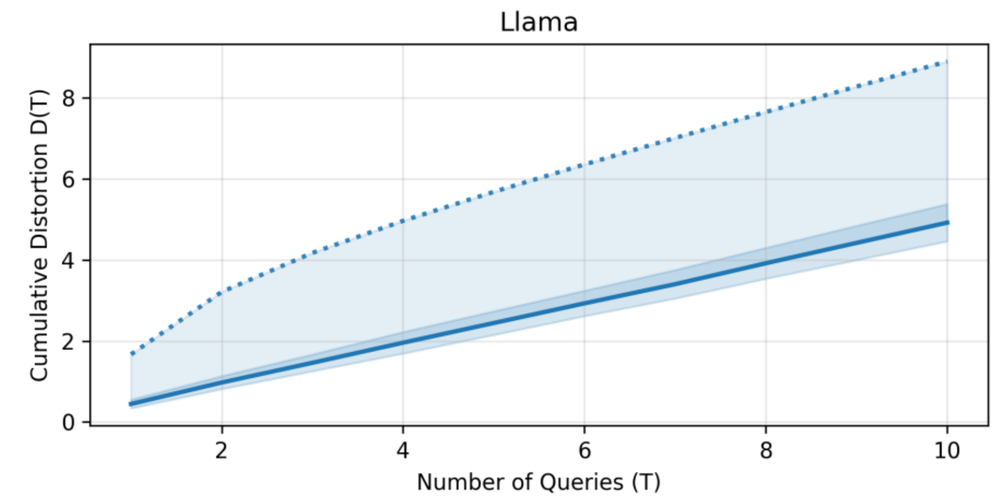
$D(T) \simeq E[D(T)]$ (linear in T), $|D(T) - E[D(T)] \leq O(\sqrt{T \log(1/\eta)})$

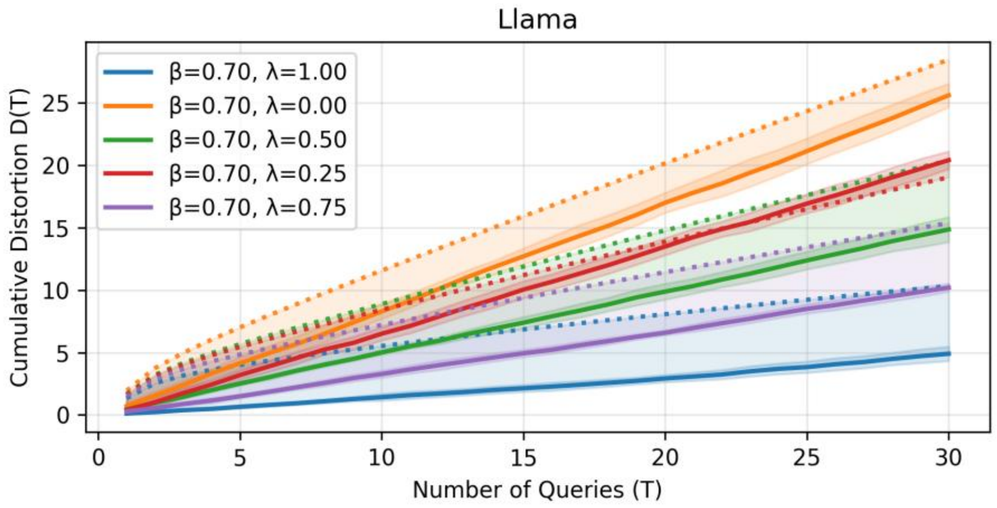
When: bounded influence / decay ($\beta B < 1$), stable responses.

Interpretation: No exponential blow-up; deviations grows like $\sqrt{T}$.

Experiments

Setup: agent retrieves & answers from pre-cached datasets
One chain: the agent retrieves → answers → retrieves → answers for 10-60 steps
Distortion score: at each step, we calculate how far the answer moved away from the tool-anchored reference
Models: Qwen2-7B, Llama-3-8B, Mistral-7B
Goal: Can we forecast cumulative distortion over long tool chains (e.g., $T \leq 30$)?





Takeaways

Predictability: distortion tracks a linear trend; deviations are $O(\sqrt{T})$ under bounded influence.
Tuning: larger λ substantially reduces measured distortion (up to ~80%).
Monitoring: fit $\hat{\beta}$ from autocorrelation to set envelopes and alarms.
Control: re-ground to cap propagation

Future Work

Relax assumptions: extend guarantees beyond bounded influence / decay.
Production validation: test MCP monitoring and re-grounding on live toolchains with retrieval freshness, API failures, and non-stationarity.
Multi-agent systems: extend monitoring to multi-agent settings with shared tools and shared state, where cross-agent propagation creates additional dependency paths.

ETH zürich
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE
Centre for Frontier AI Research
A*STAR CFAR