



Agentic Confidence Calibration

Jiaxin Zhang, Caiming Xiong, Chien-Sheng Wu | Salesforce AI Research

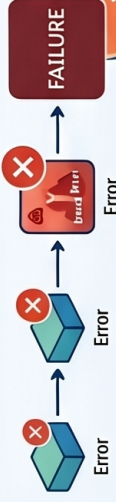
THE CHALLENGE OF AGENTIC SYSTEMS

The Barrier of Overconfidence

AI agents often remain highly confident even when failing, posing a fundamental barrier to their deployment in high-stakes, real-world settings.

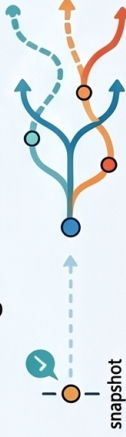


Compounding Trajectory Errors



Unlike static models, agents face unique challenges including errors that build up over multiple steps and opaque failure modes from external tool interactions.

Static vs. Agentic Calibration



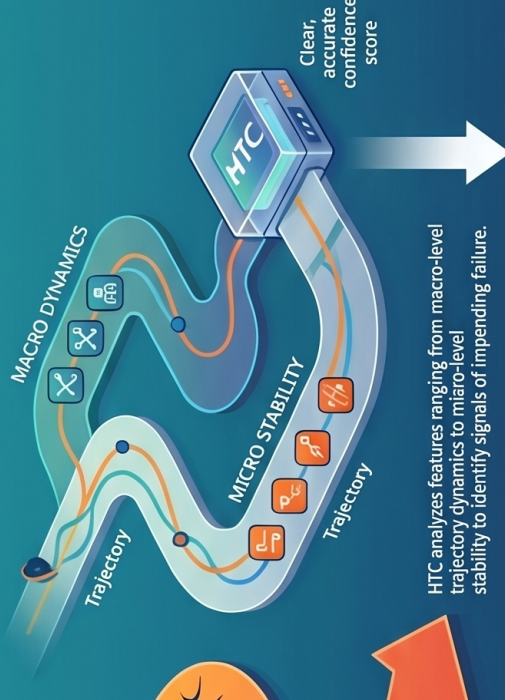
Existing calibration methods are built for single-turn outputs and cannot address the dynamic, process-heavy nature of autonomous trajectories.

THE SOLUTION: HOLISTIC TRAJECTORY CALIBRATION (HTC)

Introducing HTC

A novel diagnostic framework that extracts rich process level features across an agent's entire trajectory to assess confidence accurately.

Macro to Micro Analysis



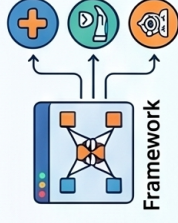
HTC analyzes features ranging from macro-level trajectory dynamics to micro-level stability to identify signals of impending failure.

Simple and Interpretable

Powered by an interpretable model, HTC reveals the specific signals behind agent failure rather than acting as a "black box."

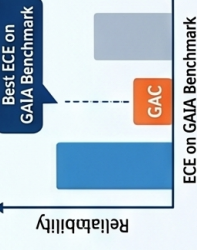


CORE ADVANTAGES & PERFORMANCE



High Transferability

The framework can be applied across different domains without the need for extensive retraining.



Superior Calibration (Lowest ECE)

The General Agent Calibrator (GAC) achieved the best calibration results on the out-of-domain GAIA benchmark.



Broad Validation

HTC consistently outperformed strong baselines across 8 benchmarks, multiple Large Language Models (LLMs), and diverse agent frameworks.

Scope of Validation:

- 8 Diverse Benchmarks
- Best ECE on GAIA Benchmark (Out-of-Domain Success)
- Multiple LLMs & Diverse Agent Frameworks (Framework Compatibility)