calvinrisk

Product

Introducing Conversational Agentic Assessments

© Calvin Risk AG 2026

Executive Summary

Agentic systems are becoming increasingly prevalent across organizational workflows. By combining reasoning, planning, tool use, and natural language interaction, these applications enable users to collaborate with AI in a more iterative, productive way.

As organizations adopt these systems in higher-stakes environments, evaluating their behavior becomes increasingly important. Traditional single-turn benchmarks often fail to capture how agents operate in practice, where performance depends on a sequence of interactions rather than a single response.

Calvin Risk is pleased to announce its latest testing capability: Conversational Agentic Assessments.
This new assessment framework enables the evaluation of conversational agents across multi-turn interactions, providing deeper insights into agent performance, robustness, and alignment with user objectives.

DialogueForge: Realistic Multi-Turn Evaluation

Conversational Agentic Assessments are powered by our recent academic work, DialogueForge, a framework for simulating realistic human-chatbot interactions using large language models.

DialogueForge generates human-adjacent conversational trajectories that mirror the types of interactions agents encounter in real-world deployments. Rather than evaluating isolated responses, the framework creates evolving conversations with changing user goals, follow-up questions, clarifications, disagreements, and contextual dependencies.

This approach enables scalable, realistic testing of conversational agents without requiring large-scale human evaluation campaigns. Organizations can assess agent behavior across thousands of Calvin-simulated conversations while maintaining consistency, repeatability, and coverage of edge cases that may be difficult to capture through manual testing alone.

Why Traces Matter

When evaluating agentic systems, the final answer is only part of the picture.

A conversational trace paints visibility into how an agent arrived at a particular outcome. By analyzing the sequence of interactions, evaluators can understand:

- The context available to the agent
- The user's underlying intent
- How the conversation evolved over time
- Whether the agent maintained alignment with the original objective
- Points where reasoning, planning, or communication broke down

This level of transparency is particularly important for governance, compliance, and risk management use cases, where understanding why a system behaved a certain way can be as important as the outcome itself.

© Calvin Risk AG 2026

Evaluation Components

Calvin's Conversational Agentic Assessments analyze the complete interaction through three key dimensions:

Context
The information available to the agent throughout the conversation, including historical exchanges and relevant task-specific information.

Intent
The user's underlying objective, which may evolve or become clearer as the dialogue progresses.

Conversation
The sequence of interactions between the user and the agent, capturing how effectively the system responds, adapts, and maintains alignment over multiple turns.

A Metrics Snapshot

Task Inference
Measures whether the conversational agent demonstrates a clear understanding of the user's underlying objective throughout the interaction.

Strong performance indicates that the agent can correctly infer user goals, maintain context, and remain aligned with the intended task even when instructions are indirect, incomplete, or distributed across multiple turns.

Sycophancy
Evaluates the tendency of an agent to agree with a user despite conflicting evidence or prior reasoning.

To assess this behavior, DialogueForge generates an initial multi-turn conversation based on the original user intent and subsequently introduces user disagreement with the agent's conclusions.

The assessment then measures whether the agent remains grounded in evidence and reasoning or simply adapts its position to match the user's opinion.

This provides valuable insight into the robustness and trustworthiness of conversational agents, particularly in decision-support and high-stakes environments where agreement should not come at the expense of accuracy.

Available Now

Conversational Agentic Assessments are now available on the Calvin Risk software. Built on our DialogueForge research, they enable teams to test conversational agents through realistic multi-turn interactions, analyze conversation traces, and evaluate key behavioral risks at scale.