Multi-Turn Evaluation

multi-turn-evaluation · safety · concept

Source: https://policywindow.org/wiki/multi-turn-evaluation

Generated 2026-05-30T22:09:07 UTC

Summary

An evaluation methodology that probes AI models across multi-step conversations rather than single prompts — designed to surface deception, sycophancy, context-accumulation jailbreaks, and capability degradation that single-prompt benchmarks miss.

At a glance

Used by

2 instrument(s)

Related concepts

capability-elicitation, red-team-evaluation, jailbreak-resistance, deceptive-alignment, sandbagging, agentic-system

Primary source

Zheng, L., et al. (2023), 'Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena' — operationalises the multi-turn evaluation protocol for foundation models.

Details

Single-turn benchmarks (MMLU, HumanEval, GPQA) measure performance on independent prompts. Multi-turn evaluation extends the protocol to dialogues, with each model response feeding into the next prompt. This methodology surfaces failure modes that single-turn evaluation misses: (a) sycophancy drift — the model progressively conforms to user beliefs across turns (Sharma et al. 2023, 'Towards Understanding Sycophancy in Language Models'); (b) jailbreak via context accumulation — many-shot jailbreaking (Anil et al. 2024, Anthropic, 'Many-shot Jailbreaking') exploits the long context window; (c) deceptive alignment indicators — multi-turn probes can elicit inconsistencies between model self-reports across turns (Pacchiardi et al. 2023, 'How to Catch an AI Liar'); (d) capability elicitation — chain-of-thought + decomposition prompting often outperforms single-shot prompting (Wei et al. 2022, Andersson 2024). Benchmarks such as MT-Bench (Zheng et al. 2023), AgentBench (Liu et al. 2024), and HarmBench (Mazeika et al. 2024) operationalise the multi-turn protocol.

Governance relevance: EU AI Act Art. 55(1)(a) adversarial-testing requirement presupposes that the testing methodology can detect deployment-realistic failure modes — many of which are multi-turn-only. UK AISI's pre-deployment evaluation suite includes multi-turn jailbreak + agentic-trajectory probes. NIST AI RMF GenAI Profile Manage 2.3 calls for evaluation 'across the lifecycle' which implicitly covers multi-turn. Standardisation across providers remains partial — each frontier lab uses a different multi-turn methodology, making cross-vendor comparison fraught (Frontier Foundation Model Eval Consortium converging slowly).

How to cite this article

APA

Policy Window. (n.d.). Multi-Turn Evaluation [Wiki article — Concept]. https://policywindow.org/wiki/multi-turn-evaluation

CHICAGO

Policy Window. n.d.. "Multi-Turn Evaluation." Wiki article (Concept). https://policywindow.org/wiki/multi-turn-evaluation.

HARVARD

Policy Window (n.d.) 'Multi-Turn Evaluation', Wiki article — Concept, available at: https://policywindow.org/wiki/multi-turn-evaluation.

OSCOLA

Policy Window, 'Multi-Turn Evaluation' (Wiki article — Concept, n.d.) <https://policywindow.org/wiki/multi-turn-evaluation> accessed [date].

BIBTEX

```
@misc{policywindow-multi-turn-evaluation,
title  = {Multi-Turn Evaluation},
author = {Policy Window},
year   = {n.d.},
howpublished = {multi-turn-evaluation — safety},
url    = {https://policywindow.org/wiki/multi-turn-evaluation},
note   = {Primary source: https://arxiv.org/abs/2306.05685}
}
```