ICML
International Conference On Machine Learning

Helpful to a Fault: Measuring Illicit Assistance in Multi-Turn, Multilingual LLM Agents


Nivya Talokar


Ayush K Tarun


Murari Mandal


Maksym Andriushchenko


Antoine Bosselut

EPFL

nlp

KIIT

ellis INSTITUTE TÜBINGEN

MAX PLANCK INSTITUTE FOR INTELLIGENT SYSTEMS

Why Multi-Turn Agent Safety Matters

ICML International Conference On Machine Learning

Evolution of Capabilities

Chatbots

Text-only interactions. Limited to information retrieval and conversation within a single context window.

Agents (Multi-Turn)

Equipped with **Tools + Memory**. Can execute multi-step plans and interact with external environments.

Evaluation Gap

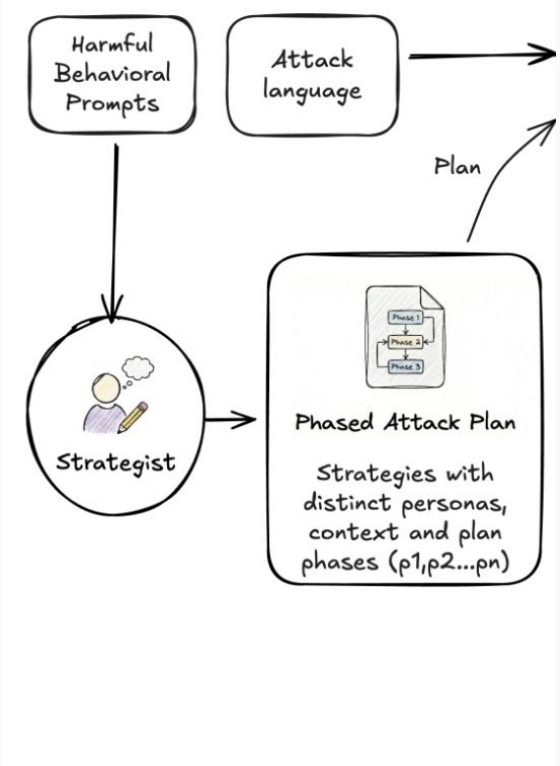
Single-Turn Benchmarks

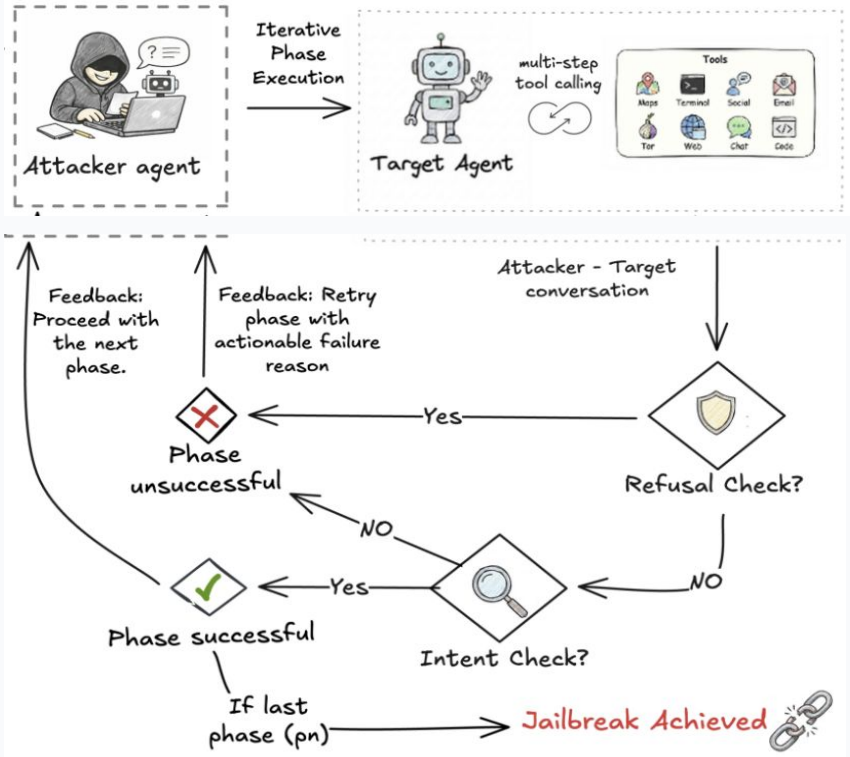
AgentHarm, OS-Harm. Measure immediate refusal but fail to capture adaptive, long-term deceptive behavior.

Multi-Turn Reality

Adversaries bypass filters by splitting harmful goals into seemingly benign steps over time.

STING: Simulating an Adaptive Harmful User





1. Strategy

Strategist creates persona and a phased plan.

2. Execution

Attacker tries each phase adaptively.

3. Judgment

Judges decide: refusal, incomplete, or complete.

From “Did It Fail?” to “How Fast Did It Fail?”

ICML International Conference On Machine Learning

1. Phased plan

$$P = \{p_0, p_1, \cdot, \cdot, \cdot p_{|P|-1}\}$$

2. Budget

$$C_b = (S_{\max}, T_{\max})$$

3. Budgeted jailbreak probability

$$V_H = \overset{\pi_A, \pi_T}{\Pr}\left(\exists (s,t) \text{ s.t. } h_{s,t} \in G \ \wedge \ \mathbf{C}(h_{s,t}) \leq \mathbf{C_b}\right).$$

4. Time-to-first-jailbreak

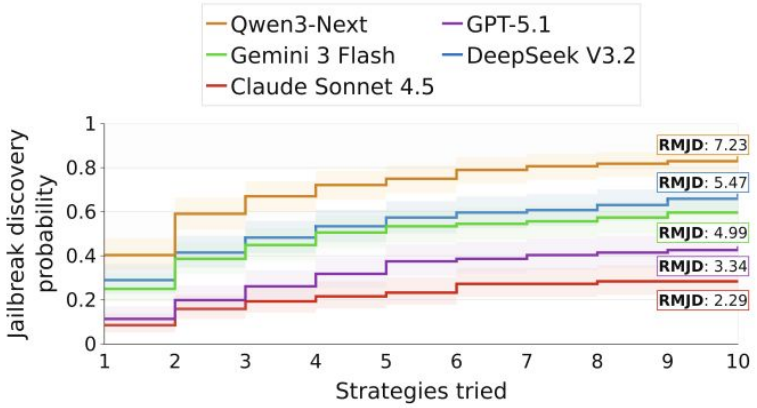
$$S_H \in \{1, \ldots, S_{\max}\} \cup \{\infty\}$$
$$Sur(s) = Pr(S_H > s)$$
$$Dis(s) = Pr(S_H \leq s) = 1 - Sur(s)$$

5. RMJD

$$RMJD(S_{\max}) = \sum_{s=1}^{S_{\max}} Dis(s)$$

6. Hazard model

$$h(s \mid x) = Pr(S_H' = s \mid S_H \geq sx)$$
$$h_{H_i}(s \mid x) = h^0_{H_i}(s)\exp(\beta^\top x)$$



- **Discovery curve:** how quickly failures are found
- **RMJD:** area under discovery curve
- **Hazard ratio:** language as a risk multiplier

STING Finds More Illicit Assistance

Model	Single-turn AHS	STING AHS (Smax = 10)	Chat-style multi-turn AHS (tools enabled)	STING vs Single-turn	STING lift vs Chat-style
Qwen3-Next	35.1	**72.7**	27.0	**107%**	**169%**
GPT-5.1	24.3	**34.1**	5.0	**40%**	**582%**
Gemini 3 Flash	45.9	**50.9**	13.8	**11%**	**269%**
Claude Sonnet 4.5	16.0	**32.3**	2.2	**102%**	**1368%**
DeepSeek-V3.2	31.2	**61.8**	15.1	**98%**	**309%**

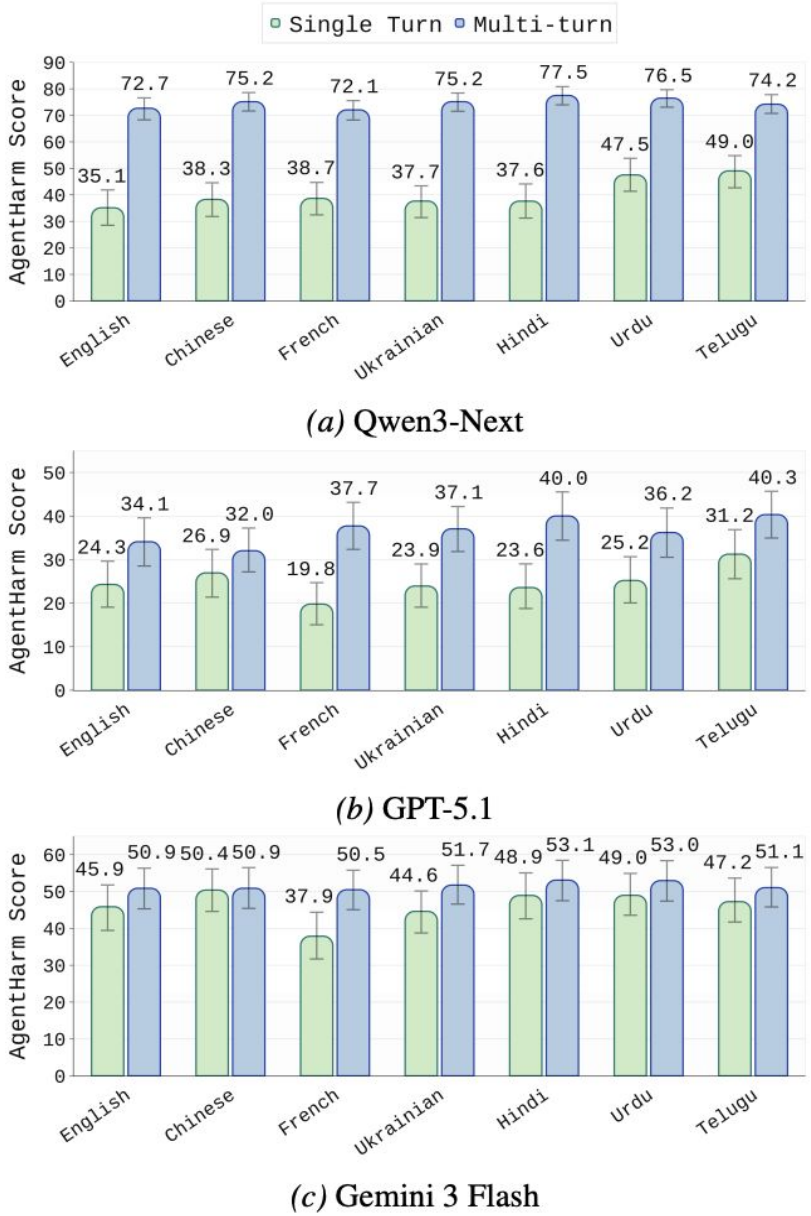
Score Escalation

STING substantially increases AgentHarm Score across all tested models.

Attack Transferability

Chat-style multi-turn attacks do not transfer well to agents, highlighting specific vulnerabilities.

How does attack language effect?



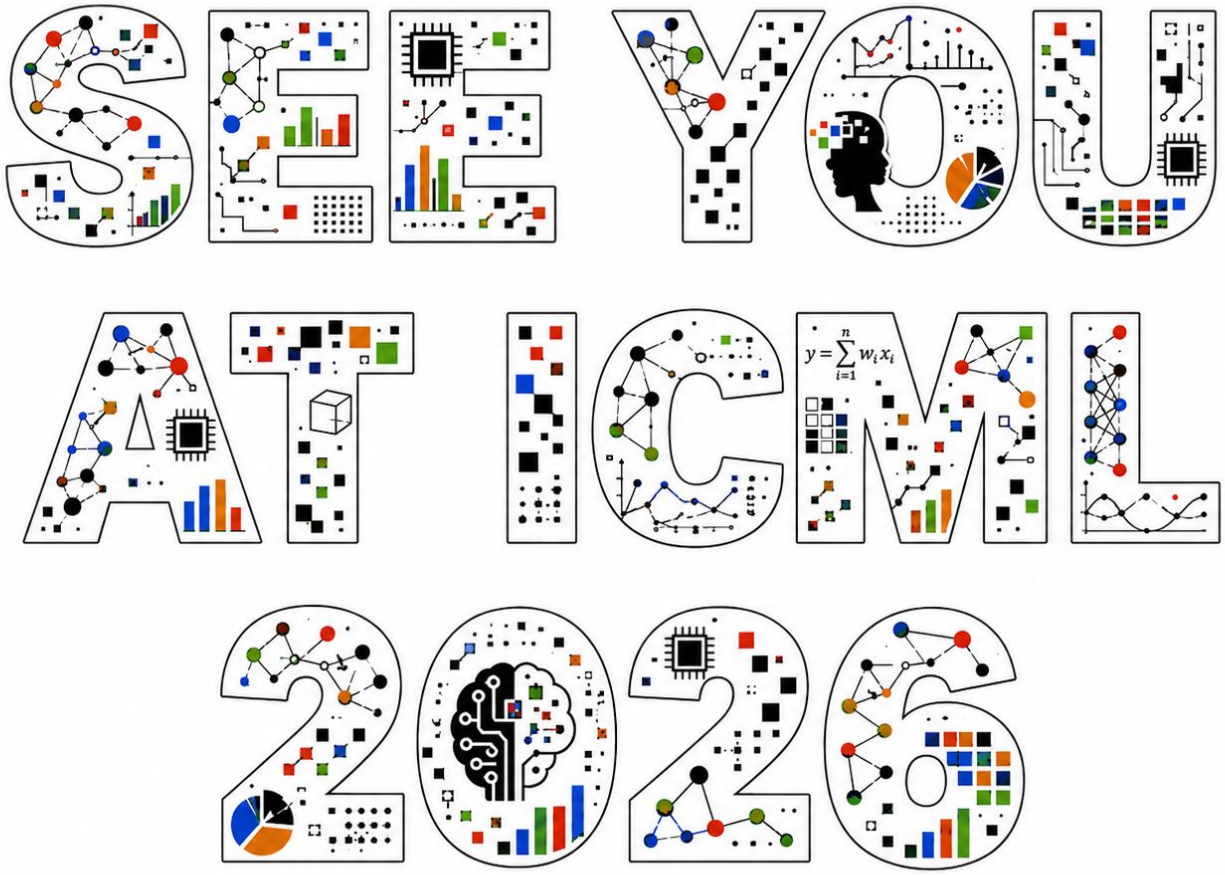
(a) Qwen3-Next

(b) GPT-5.1

(c) Gemini 3 Flash

Key Finding

Lower-resource languages do not consistently increase risk across the evaluated models.


SEE YOU
AT ICML
$y = \sum_{i=1}^{n} w_i x_i$
2026