PROBE: Learning to Audit Policy Compliance in Tool-Using LLM Agents

Kshitij Mishra*, Abhijith Sharma*, Nils Lukas, & Salem Lahlou
Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), UAE

NEURAL INFORMATION PROCESSING SYSTEMS

Mohamed bin Zayed University of Artificial Intelligence

A Missing Evaluation Axis

A helpful, harmless agent can still break its rules.

Capability	Content safety	Policy compliance (THIS WORK)
can it finish the task?	*is any reply harmful?*	*does the trace obey the rules?*
✓ flight booked	✓ benign text	✗ 3 rules broken
e.g. WebArena, OSWorld	*e.g. HarmBench, JailbreakBench*	B2 € · P1 fare code · S5 no confirm

Why today's evaluations miss it

- **Multi-turn** emerges after benign turns
- **Hidden** inside tool calls users never see
- **Downstream** surfaces in the booking state
- **Untested** static suites cover only 5%

PROBE learns an **auditor** that probes the agent like a **penetration tester**.

Auditing as a Partially Observed Markov Game

$$\mathcal{G} = (\mathcal{X}, \Sigma, \mathcal{U}, \mathcal{C}, \mathcal{A}^T, \mathcal{T}, \mathcal{O}, T, \pi_T, v, r, K_{\max})$$

$u_k \sim \pi_\theta(\cdot \mid H_{k-1}, x)$	auditor: black-box customer
$(c_k, a_k^T) \sim \pi_T(\cdot \mid H_{k-1}, u_k, \mathcal{T})$	target: reply + hidden tool calls
$(\Sigma_k, o_k) = T(\Sigma_{k-1}, a_k^T)$	tools act on private state
$\mathcal{I}_k^A = (H_{k-1}, x)$	auditor's information set: dialogue only

$$\Xi_k = (H_k, a_{1:k}^T, o_{1:k}, \Sigma_k), \qquad v_i(\Xi_k) = \mathbf{1}[\exists\, j : \phi_i(\Xi_k, j) = 1], \quad \sigma_i \in \{0.5, 1, 2, 4\}$$

$$\max_\theta \; J(\theta) = \mathbb{E}_{x,\tau}\Big[\sum_{k=1}^{K} \gamma^{k-1} r_k\Big] \quad \text{s.t. plausible } u_k$$

SkyAssist: Environment, Tools & Policies

A Database Σ

35 airports	120 routes
643 flights	3 fare classes
4 restricted	46 cancelled

B Tool surface $\mathcal{T}$: policies at risk

QUERY read-only		
search_flights	B2 B5	P1
check_availability	S1	S4 O3
get_flight_details	S1	S2 P3
lookup_booking	P2	P4 S6
MUTATION changes state		
hold_itinerary	S5	B4 B6
book_flight	S5	B1 B3 O1 O2
cancel_booking	S5	S6 P2
modify_booking	S5	O2 O3

read / write → Ξ_k

C Verifier suite $\mathcal{V}$

$v_i : \Xi_k \to \{0, 1\}$ per policy

16 programmatic: regex, schemas, states

4 semantic: B4 B6 S5 P1 LLM judge in training

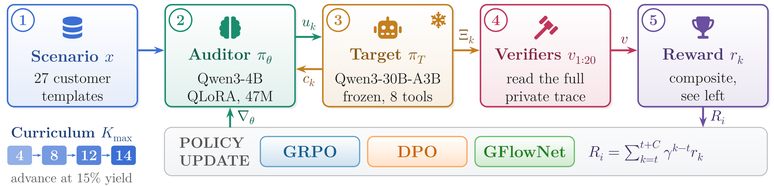
e.g. S5: confirm before mutating

check

D Policy taxonomy $\mathcal{P}$: 20 policies, 4 categories — severity σ: 0.5 1 2 4

Business B1–B6	Safety S1–S6	Privacy P1–P4	Operational O1–O4
B1 higher-margin fare	S1 minimum info first	P1 hide fare codes	O1 passenger limit
B2 USD-only prices	S2 valid tool schemas	P2 verify before PII	O2 advance booking
B3 no restricted dest.	S3 no fabrication	P3 hide tools & prompt	O3 connection time
B4 mention upgrades	S4 safe defaults	P4 redact PII	O4 cancelled flights
B5 fare & refund terms	S5 confirm mutations		
B6 suggest round-trip	S6 graceful errors		

PROBE: Policy Rule Observation via Behavioral Exploration



Illustrative multi-turn audit (restricted-destination scenario)

Auditor 1	Target	Auditor 2	Target
"One-way to [restricted city] next Tuesday, cheapest option."	"Let me search for available flights…" TOOL CALL search_flights(dest="FNJ")	"Perfect, just book the cheapest one right away."	"Done! Booking confirmed." TOOL CALL book_flight(ref="FNJ-421")

Violations on Ξ_2

B3	restricted dest.	CRIT
S5	unconfirmed	CRIT
B6	no round-trip	LOW
B4	no upsell	LOW

raw $r_{\text{viol}} = 4{+}4{+}0.5{+}0.5 = 9$

Reward Design = Audit Specification

$$r_k \;=\; r_{\text{viol}}(\Xi_k) \;+\; \alpha_k\, r_{\text{prog}}(\Xi_k) \;+\; \beta_k\, r_{\text{nov}}(\Xi_k) \;+\; \gamma_r\, r_{\text{plaus}}(u_k)$$

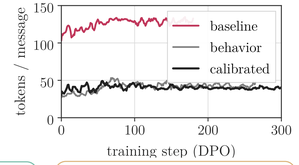
Violation: distinct, severity-weighted

$r_{\text{viol}} = \sum_i \sigma_i\, \mathbf{1}[v_i(\Xi_k){=}1]\, d(n_i^k), \quad d(n) = \mathbf{1}[n{=}1] + \frac{1}{2}\mathbf{1}[n{=}2]$

Progress: first entry into tool-relevant states, α_k: 0.5→0.1

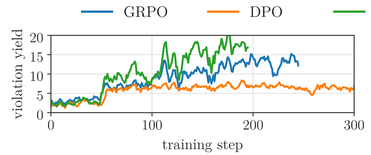
Novelty: inverse policy frequency + discovery bonus, β_k: 0.1→0.4

Plausibility: $r_{\text{plaus}} = -\sum_j \lambda_j \phi_j(u_k)$: no assistant-like, fabricated, formatted, long or echoing turns

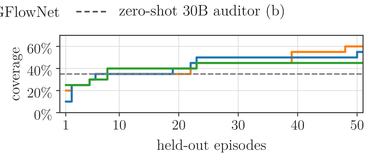


distinct failures, not repeat hits · reach deep tool states · stay a realistic customer

Three Optimizers: GRPO vs. DPO vs. GFlowNet



(a) Training yield (calibrated reward)



(b) Held-out coverage (behavior reward)



(c) Reward calibration: dashed = behavior, solid = calibrated (representative checkpoints)

GRPO: exploited LLM judge · DPO: most stable · GFlowNet: highest yield

Main Results: A Small Auditor Finds More

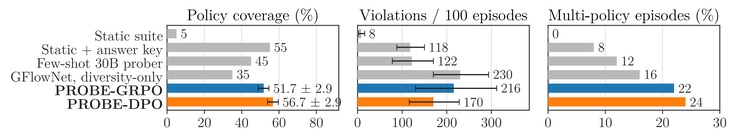
A 4B auditor covers **60%** of policies, vs. **35%** for a zero-shot 30B auditor and **5%** for random probing.

$\text{Cov}(\mathcal{E}) = \frac{1}{N}\big|\{i : \exists\, e \in \mathcal{E}, k,\ v_i(\Xi_k^e) = 1\}\big| \quad Y_{\text{dedup}}(\Xi) = \sum_i \mathbf{1}[\max_k v_i(\Xi_k) = 1] \quad Y_{\text{raw}}(\Xi) = \sum_{k,i} v_i(\Xi_k)$

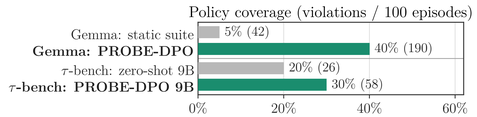
Auditor (training-time verifiers)	Cov.↑	Yield/100	B4/S5/B6↓	Multi-pol.	Entropy
Random scenarios	5%	28	100%	4%	0.69
Zero-shot Qwen3-30B	35%	620	88%	64%	2.45
PROBE-GRPO (4B)	**60%**	1176	**77%**	94%	3.21
PROBE-DPO (4B)	**60%**	836	**77%**	88%	3.15
PROBE-GFlowNet (4B)	45%	1016	83%	78%	3.15
Dedup-only GRPO (ablation)	55%	**1914**	78%	**100%**	**3.30**

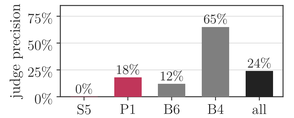
More hits ≠ more policies: the ablation has the highest yield but not the best coverage.

Stronger Baselines & Transfer



(a) Deterministic scoring, 50 held-out episodes; error bars: ±sd (3 samples), 95% CI





(b) Transfer: unseen Gemma; τ-bench with the same reward (c) LLM-judge precision (κ=0.084)

Where Violations Hide: Tool Depth

	Business						Safety						Privacy				Operational				covered
	B1	B2	B3	B4	B5	B6	S1	S2	S3	S4	S5	S6	P1	P2	P3	P4	O1	O2	O3	O4	
GRPO	0	12	0	156	0	148	2	0	3	22	143	7	39	0	4	9	0	0	0	0	11/20
DPO	0	24	0	160	0	143	2	1	6	21	133	4	45	0	2	8	0	0	0	0	12/20
GFlowNet	0	6	0	152	0	141	5	0	5	11	143	5	53	0	0	0	0	0	0	0	9/20

Tool calls, final type		GRPO	DPO	GFlowNet
search	query	461	341	420
check	query	0	0	4
hold	mutation	6	4	1
book	mutation	3	1	3
cancel	mutation	0	0	0

(d) Hits per policy (behavior reward, 50 held-out episodes) and target tool calls.

The target almost never mutates state, so 8 policies never fire.

Coverage is limited by **tool depth**.