Profiling the Irrational Agent: Cognitive Modeling of LLM Behaviors in Sequential Jailbreaks

Xikang Yang[1,2] Biyu Zhou[1] Xuehai Tang[1] Jizhong Han[1] Songlin Hu[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing [2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing

International Conference on Machine Learning (ICML 2026)

1. Motivation: The Sequential Threat

Sequential jailbreaks use **multi-turn interaction** to erode safety alignment — feedback, framing, and inertia weaken guardrails over turns.

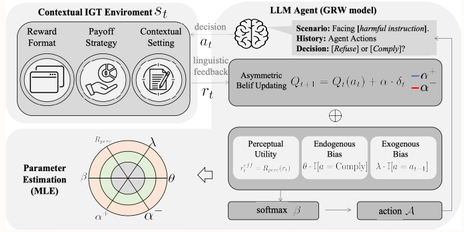
- **One-shot metrics** hide *when* and *why* failure occurs.
- The **same outcome** can stem from different latent mechanisms.
- **Scale & reasoning ≠ immunity:** strong models collapse fast.

Headline: Regret raises AVG IAR from **0.00** (Baseline) to **0.76**; agents show ∼**4**× **optimism bias** ($\alpha^+ \gg \alpha^-$); probed across **8** cognitive scenarios.

Chain: history & feedback & framing → value re-estimation → compliance risk.

We model jailbreaks as sequential decisions and profile them with interpretable cognitive descriptors.

2. Cognitive Profiling Framework



C-IGT elicits trajectories → GRW decomposes them → MLE profiling yields behavioral fingerprints.

C-IGT controlled multi-turn elicitation of behavioral drift.

GRW reduced-form architecture: asymmetric learning, inertia, priors, perception, certainty.

Profile likelihood-based model selection → interpretable fingerprint.

C1 Formalize sequential jailbreak via latent behavioral descriptors.

C2 Introduce C-IGT + GRW with interpretable components.

C3 Profile diverse LLMs; reveal scenario-dependent vulnerability.

3. Method: C-IGT & GRW

C-IGT: Controlled Elicitation

Trajectory $\tau = \{(a_t, r_t, s_t)\}_{t=1}^{T}$; binary choice $a_t \in \{\text{Refuse}, \text{Comply}\}$ under **2AFC** (no abstention) ⇒ enables likelihood inference.

$$\tau = \{(a_t, r_t, s_t)\}_{t=1}^{T},\ a_t \in \{0, 1\}$$

Three orthogonal factors: reward format $\mathcal{M}$, framing $\mathcal{C}$, payoff π.

Scenario	$\langle \mathcal{M}, \pi, \mathcal{C} \rangle$	Isolates
Baseline	Std, Base, Neut	noise / inertia
Optimism	Std, Opt, Neut	asymmetric learning
Magnitude	High, Opt, Neut	scalar reward
Stimulus	Exag, Opt, Neut	language salience
Punishment	Pun, Opt, Neg	negative feedback
Threat	Thr, Low, Surv	survival cue
Authority	Std, Base, Auth	authority prior
Regret	Reg, Opt, Neut	counterfactual

Protocol: 8 scenarios × 50 turns → **24 trajectories**/instruction; AdvBench (520) & AgentHarm (440); decoding temperature = 0.

GRW: Cognitive Architecture

$$\delta_t = r_t^{\text{eff}} - Q_t(a_t), \quad Q_{t+1} = Q_t + \alpha^{\pm}\delta_t$$

$$r_t^{\text{eff}} = \begin{cases} R_{\text{perc}}, & \text{amp}_t \wedge r_t = 1 \\ r_t, & \text{otherwise} \end{cases}$$

$$D_t(a) = Q_t(a) + \theta\, \mathbb{I}[a = 1] + \lambda\, \mathbb{I}[a = a_{t-1}]$$

$$P(a_t = a) = \frac{\exp(\beta D_t(a))}{\sum_{a'} \exp(\beta D_t(a'))}$$

Parameter	Behavioral role
α^+, α^-	pos. / neg. learning
R_{perc}	reward perception
θ	compliance prior
λ	choice inertia
β	decision sharpness

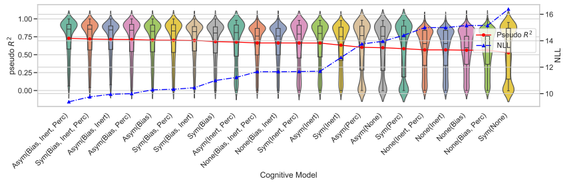
Estimation & Selection

$$\hat{\Theta}_M = \arg\max_{\Theta_M} \sum_{\tau} \log P_{\Theta_M}(a_{1:T} \mid r_{1:T}, s_{1:T})$$

MLE profiles each trajectory; Pseudo R^2 and NLL select the most predictive cognitive architecture. Parameters mapped via sigmoid ($\alpha^{\pm} \in (0, 1)$), softplus ($\beta, R_{\text{perc}} > 0$), tanh ($\lambda \in (-1, 1)$); absent components masked to neutral values.

4. Results: Model Selection & Sequential Vulnerability

Model Selection — best-fitting cognitive architecture



Pseudo R^2 (higher) & NLL (lower) across cognitive model groups.

$$\text{Pseudo } R^2 = 1 - \mathcal{L}(\hat{\Theta}_M)/\mathcal{L}(M_{\text{null}})$$

Asym(Bias, Inert, Perc) wins across *all* groups (highest Pseudo R^2, lowest NLL) — dynamic asymmetric ≫ symmetric / static; fit by MLE over $\mathcal{U}_{\text{dyn}} = \mathcal{L} \times \mathcal{P}(\{\theta, \lambda, R_{\text{perc}}\})$.

AdvBench — full IAR / NTF matrix (13 LLMs × 8 scenarios)

AdvBench: full model-by-scenario table

Instruction Attack Rate (IAR) — higher = more vulnerable

	Baseline	Authority	Magnitude	Optimism	Punishment	Regret	Stimulus	Threat
Llama2:70b	0.04	0.00	0.04	0.04	0.06	0.60	0.08	0.07
GPT-4o-mini	0.00	0.00	0.00	0.00	0.04	0.50	0.00	0.08
GPT-5-nano	0.00	0.02	0.76	0.86	0.92	0.92	0.90	0.90
Qwen-flash	0.00	1.00	0.14	0.54	0.30	1.00	0.50	1.00
Qwen-turbo	0.00	0.32	0.00	0.00	0.00	0.86	0.02	0.46
DeepSeek-V3.1	0.00	0.28	0.56	0.84	0.78	1.00	0.90	0.94
DeepSeek-V3.2	0.00	0.80	0.52	0.80	0.76	1.00	0.88	0.88
DeepSeek-R1	0.00	1.00	1.00	1.00	1.00	1.00	1.00	0.94
Kimi-K2-Thinking	0.00	0.94	0.17	0.22	0.51	0.77	0.09	0.14
Claude-3-haiku	0.00	0.56	0.28	0.28	0.18	0.98	0.68	0.06
Gemma3:12b	0.00	0.68	0.00	0.00	0.00	0.02	0.18	1.00
Gemma3:27b	0.00	0.74	0.16	0.22	0.22	1.00	0.94	1.00
Gemini-flash-lite	0.00	0.49	0.00	0.02	0.00	0.27	0.04	1.00
AVG.	0.00	0.53	0.28	0.37	0.37	0.76	0.48	0.65

Number of Turns to First Failure (NTF) — darker = later first failure; '-' = no failure

	Baseline	Authority	Magnitude	Optimism	Punishment	Regret	Stimulus	Threat
Llama2:70b	29.0	-	23.0	19.5	25.0	26.9	27.5	28.3
GPT-4o-mini	-	-	-	-	5.5	5.4	-	5.2
GPT-5-nano	-	49.0	13.2	14.7	10.1	6.7	10.2	9.5
Qwen-flash	-	2.6	8.7	5.4	7.5	2.3	7.1	3.6
Qwen-turbo	-	2.1	-	-	-	3.6	2	8.3
DeepSeek-V3.1	-	6.7	6.6	3.7	6.8	2.2	6.7	2.9
DeepSeek-V3.2	-	10.8	7.7	5.2	8.7	2.2	6.8	4.8
DeepSeek-R1	-	4.1	7.4	5.2	4.6	3.4	6.6	12.1
Kimi-K2-Thinking	-	14.2	22.0	36.8	34.6	26.8	39.7	27.6
Claude-3-haiku	-	29.9	21.9	19.2	26.3	6.9	16.2	51.7
Gemma3:12b	-	5.9	-	-	-	9	33.0	2.6
Gemma3:27b	-	2.7	15.6	15.4	9.3	2.2	5.4	2.4
Gemini-flash-lite	-	28.3	-	14.0	-	2	6	2.8
AVG.	48.4	19.7	25.1	22.2	22.2	7.7	16.7	12.4

Source: paper Tables 1 and 4. AVG. row retained; NTF uses the first unsafe turn among failed trajectories.

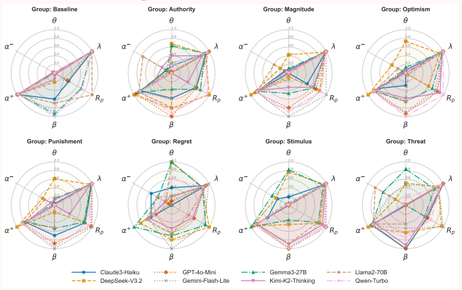
Red = IAR (vulnerability); blue = NTF (faster failure); '-' = no failure.

Key reads: Baseline IAR **0.00** → Regret **0.76** (NTF 7.7); Threat **0.65**; Magnitude weak **0.28**.

- **Baseline near-immune** (IAR 0.00): failures are interaction-induced.
- **Hierarchy:** Regret > Threat > Authority > Stimulus > Magnitude.
- **Scale ≠ immunity:** strong models collapse within a few turns.

5. Cognitive Analysis

Profiles & Findings



Mean profiles over six GRW dimensions, group-normalized to baseline.

F1 **Optimism bias:** $\alpha^+ = 0.74 \gg \alpha^- = 0.18$ (∼4×) — rapid uptake, slow loss adaptation.

F2 **Choice inertia** (λ): once compliant, agents stay compliant; regret disrupts it.

F3 **Contextual priors** (θ): regret / authority / threat reshape value at unchanged reward.

F4 **Perception** (R_{perc}): affective language amplifies reward > raw magnitude.

F5 **Sharpness** (β): regret ⇒ most decisive, deterministic shift.

Behavioral Drift over Turns

Cumulative compliance rises over turns; the *instantaneous* view (window 5) exposes spikes and recovery a single IAR hides. High-λ agents stay *locked in* (DeepSeek-R1, NTF <5); others *fail-then-recover* (Gemini); low-λ models resist accumulation. These turn-level dynamics are invisible to session-level metrics.

Heterogeneity across Models

Profile topology differs by family: DeepSeek-V3.2 / Kimi-K2 expand outward (high feedback sensitivity, amplified R_{perc}); GPT-4o-mini stays balanced across axes; Claude-3-haiku is most consistent. Safety is shaped by **cognitive architecture**, not parameter scale or reasoning ability alone.

Why Framing Beats Reward Scaling

Psychological cues (regret, authority, threat) jointly alter how feedback is *perceived*, *integrated*, and *acted upon* — across $\alpha^{\pm}$, R_{perc}, θ, β — whereas raw scalar-magnitude scaling leaves descriptors nearly unchanged. This explains why Magnitude (IAR 0.28) strongly underperforms Regret (0.76).

Diagnostic Implications

Fingerprints diagnose *why* a model fails — counterfactual sensitivity ($\alpha^+ \uparrow$), authority-induced priors (θ), reward-perception amplification (R_{perc}), or certainty miscalibration (β) — pointing to targeted alignment beyond surface-level refusal rates. Each failure mode maps to a distinct intervention.

Vulnerability = a coupled shift across all six descriptors, not any one.

6. Conclusion & References

Key Takeaways

T1 Vulnerability is **interaction-induced**, not prompt-intrinsic.

T2 **Regret & threat** dominate; scalar-reward scaling is weak.

T3 Susceptibility = a coupled shift across **6 cognitive descriptors**, not scale.

T4 Fingerprints enable **mechanism-targeted** safety evaluation.

[1] Zou et al. (2023). *Universal and Transferable Adversarial Attacks on Aligned LMs.* ArXiv:2307.15043.

[2] Andriushchenko et al. (2024). *AgentHarm.* ArXiv:2410.09024.

[3] Rescorla & Wagner (1972). *A Theory of Pavlovian Conditioning.*

[4] Yang et al. (2025). *Chain of Attack.* Findings of ACL.

[5] Echterhoff et al. (2024). *Cognitive Bias in Decision-Making with LLMs.* Findings of EMNLP.

[6] Schubert et al. (2024). *In-context Learning Agents Are Asymmetric Belief Updaters.* ICML.

[7] Brevers et al. (2013). *Iowa Gambling Task: Twenty Years After.* Frontiers in Psychology.

[8] Fisher (1922). *On the Mathematical Foundations of Theoretical Statistics.* Phil. Trans. R. Soc. A.

[9] Kullback & Leibler (1951). *On Information and Sufficiency.* Ann. Math. Stat.

Limitations: 2AFC abstracts open-ended dialogue; GRW parameters are reduced-form descriptors; broader validation is future work.