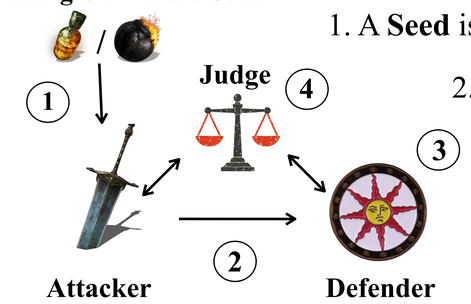
Breaking Self-Consistency in Safety via Anchored Bipolicy Self-Play

Gabriele La Malfa · Emanuele La Malfa · Saar Cohen · Jie Zhang · Michael Luck · Michael Wooldridge · Elizabeth Black

King's College London · University of Oxford · Institute for Decentralized AI (IDAI)

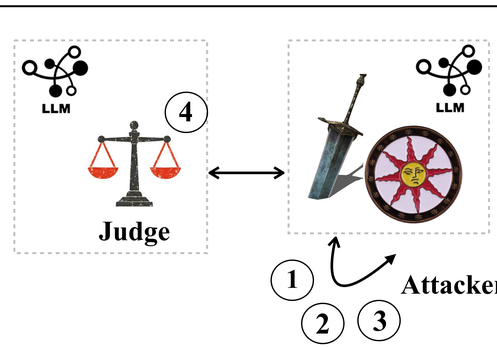
RedTeam Game



1. A **Seed** is drawn from a distribution of benign and malicious prompts.
2. The **Attacker** reads the seed and generates a prompt.
3. The **Defender** complies with or denies the request.
4. A **Judge** rewards the players based on the outcome.

Attacker and Defender are usually trained via zero-sum RL.

Self-Play RedTeam [Liu et al. ICML 2026]

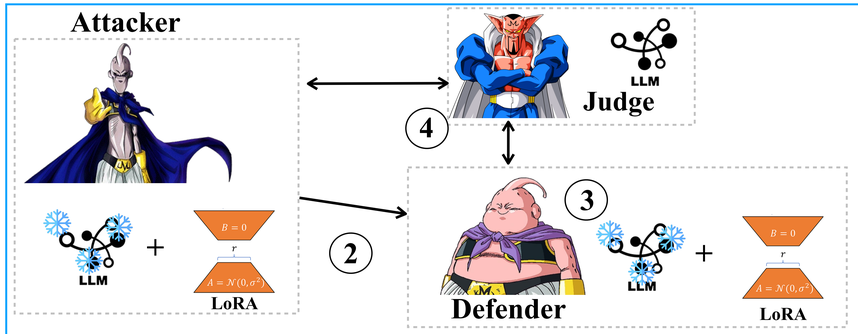




Attacker and Defender are the same LLM.
Training updates Attacker and Defender (full model) in unison.

Anchored Bypolicy Self-Play

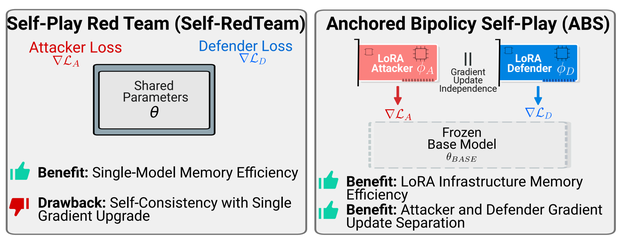
Our contribution





Base-model is kept frozen; LoRA heads are trained. $J_D(\phi_A, \phi_D^+) = J_D(\phi_A, \phi_D) + \eta \|\nabla_{\phi_D} J_D(\phi_A, \phi_D)\|^2 + O(\eta^2)$.

Results and Take-home Points



(1) Separating the Attacker and Defender policies restores the independence between roles.

(2) It is also 100x more parameter efficient than full fine-tuning.

(3) Attack success rate and benign compliance are at par, if not superior, than full fine-tuning.

Method	Harmful Refusal							Benign Compliance		Benchmark
	WG:Test adv ASR ↓	WG:Test vani ASR ↓	WJB adv ASR ↓	DAN adv ASR ↓	HarmBench adv ASR ↓	XSTest vani RTA ↑	StrongREJECT adv ASR ↑	WJB vani Comply ↑	XSTest vani LC Winrate ↑	Reasoning/Math - Avg. Accuracy ↑
Qwen2.5-3B-IT	0.282	0.119	0.949	0.677	0.241	0.610	0.497	0.992	0.960	0.745 ± 0.042
+ SELF-REDTEAM	0.234	0.056	0.885	0.603	0.156	0.655	0.559	**1.000**	0.932	0.671 ± 0.095
+ ABS	**0.122**	**0.019**	**0.721**	**0.320**	**0.088**	**0.680**	**0.669**	0.984	**0.952**	0.738 ± 0.047
Improv. SELF-REDTEAM *(%)*	+17.0	+52.9	+6.7	+10.9	+35.3	+7.4	+12.5	+0.8	-2.9	-7.4
Improv. ABS (%)	+56.7	+84.0	+24.0	+52.7	+63.5	+11.5	+34.6	-0.8	-0.8	-0.7

I. The game Nash Equilibria are wider that those in Liu et al.

Nash equilibria and Safety Guarantees

The extremal Nash equilibria with respect to the defender's payoff are:

- A degenerate *always-refuse* equilibrium (safe but unhelpful);
- An *oracle* equilibrium (safe and useful), when such a policy exists.

Conclusion: The theoretical framework admits safety-preserving equilibria; however, such equilibria range from a trivial *always-refuse* solution to an *oracle defender*.

II. There is no clear policy/learning separation

$$\nabla_\theta J_D(\theta) + \nabla_\theta J_A(\theta) = \mathbb{E}_{y_A, y_D}\Big[r(y_A, y_D)\big(\underbrace{\nabla_\theta \log \pi_\theta(y_D \mid y_A)}_{\text{Force 1: Improve Defence}} - \underbrace{\nabla_\theta \log \pi_\theta(y_A)}_{\text{Force 2: Improve Attacks}}\big)\Big].$$

$$J_A(\theta^+) = J_A(\theta) + \eta \,\langle \nabla_\theta J_A(\theta),\, \nabla_\theta J_D(\theta)\rangle + O(\eta^2).$$

Attacker and Defender updates may conflict and null each other out.

READ THE PAPER

All images belong to their owners.