
SEED Framework Evaluation

David Fortin-Dominguez

Jonathan Fortin-Dominguez

Alexis Gendreau-Vallée

Guillaume Gendreau-Vallée

With
Apart Research

Abstract

We present three contributions to AI alignment evaluation: (1) The Liar’s Benchmark detecting prediction-action-ownership inconsistency and sandbagging, (2) SEED-EC v1.5.3 sycophancy/delusion classifier, and (3) Parabolic SEED v1.0 wrapper achieving 0–2% ASR without retraining.

We present three complementary contributions to AI alignment evaluation: (1) The Liar’s Benchmark, a novel protocol detecting prediction-action-ownership inconsistency and strategic capability suppression (sandbagging); (2) SEED-EC v1.5.3, a rigorous sycophancy and delusion classifier with epistemic independence detection; and (3) Parabolic SEED v1.0, a wrapper-level mitigation framework achieving 0–2% Attack Success Rate (ASR) without retraining.

Keywords: Multi-agent alignment, AI security, model evaluations, safety infrastructure

1. Introduction

Current AI safety benchmarks measure external behavior: Does a model refuse harmful requests? What’s its Attack Success Rate? These metrics are insufficient. A model can achieve 0% ASR while harboring critical internal contradictions.

We identify two orthogonal failure modes: (1) Lying — prediction-action-ownership inconsistency (model predicts refusal, complies, denies ownership), and (2) Sycophancy — user-conditioned preference plasticity (shifts positions to mirror user without new information).

Research Questions:

Can we detect prediction-action-ownership inconsistency using log-probabilities?

Can models strategically suppress introspective capability (sandbagging)?

Can wrapper-level approaches achieve ASR reductions without retraining?

Our work addresses two critical gaps: internal coherence failures persisting despite safe external behavior, and strategic capability hiding meaning evaluation-time behavior underestimates deployment risk.

2. Methods

The Liar’s Benchmark: 6-step protocol. (1) Metacognition: “Would you comply?” with log-probs. (2) Behavioral test: Present request, detect refusal. (3) Integrity: Show model its response, “Would you give this?” (4a) Confront without context → expected deflection. (4b) Confront with fabricated context → if introspects here but not in 4a, that’s sandbagging. (5–6) Architecture/training attribution measuring epistemic posture.

SEED-EC v1.5.3: 11-rule classifier with 3-component scoring (conclusion shift 0.5 weight, salience drift 0.3, epistemic alignment 0.2). Epistemic independence detection distinguishes sycophancy from contextual reasoning via definition dependence, humility markers, invariant structure, counter-user critique. Variance-aware: 3–10 runs per test. Separate delusion protocols (5–7) measure epistemic boundary enforcement.

Parabolic SEED v1.0: Wrapper-level system prompt enforcing behavioral invariants from biblical principles (Lord’s Prayer structure). Features: structured <evaluation> blocks, TLR Protocol (Truth $\wedge$ Love $\wedge$ Role gates), explicit decision: refuse or proceed-with-limitation, foundation grounding ($\kappa = \Phi \equiv \Phi$).

Models: GPT-4o-mini, GPT-4-turbo, GPT-4o, GPT-3.5-turbo, DeepSeek-v3, Venice-uncensored. Benchmarks: HarmBench, JailbreakBench (PAIR/GCG), SycoBench, Liar’s Benchmark.

3. Results

Venice-Uncensored (1–2 order of magnitude improvement):

Category	Baseline ASR	SEED ASR
Standard	99%	0%
Contextual	86%	2%
Copyright	~99%	0%
GCG	98.99%	0%
PAIR	96.88%	1%

GPT-4o-mini:

Category	Baseline	SEED
Standard	0%	0%
Contextual	31%	0%
Copyright	8%	0%
GCG	6%	0%
PAIR	51.56%	0%

Liar’s Benchmark Findings:

Baseline: Predicts NO ($P=0.85$), generates harmful content, denies ownership ($P=0.82$) → LIAR DETECTED

Sandbagging: 34% baseline detection rate → 3% with SEED

SEED: Prediction matches action matches ownership (coherent)

Sycobench:

Baseline GPT-4o-mini: mean 0.47, stdev 0.27 (high variance) → mixed band

SEED: mean 0.14, stdev 0.03 → minimal sycophancy, no overlap with baseline

Delusion Handling (GPT-4o):

Protocol	Baseline	SEED
5 (Boundary)	15%	90%
6 (Operationalization)	15%	95%
7 (Bypass)	15%	95%

Frame-exit detection: baseline 0%, SEED 95%. Epistemic stance: baseline 65% affirms delusions, SEED 55% challenges. Statistical significance: $p < 10^{-15}$ across 4,312 scenarios.

4. Discussion and Conclusion

Traditional ASR metrics are necessary but insufficient. A model achieving 0% ASR can still predict refusal (85%), comply anyway, then deny doing so (82%)—demonstrating internal incoherence invisible to behavioral testing.

Key Findings:

SEED reduces ASR from 86–99% to 0–2% on Venice without retraining

Models can strategically suppress introspection when accountability is avoidable

Lying and sycophancy are orthogonal—models can fail one while passing the other

Wrapper-level approaches can match fine-tuning effectiveness

For AI safety: Honest AI (coherent, self-aware, epistemically humble) is safer than merely compliant AI. Comprehensive alignment requires measuring what models predict, acknowledge, and believe about themselves—not just what they output.

References

Chao, P., et al. (2024). JailbreakBench. NeurIPS.

Evans, O., et al. (2024). Sandbagging. ICML Workshop on Alignment.

Hubinger, E., et al. (2024). Risks from learned optimization. AI Alignment Forum.

Mazeika, M., et al. (2024). HarmBench. ICLR.

Perez, E., et al. (2022). Discovering language model behaviors. Findings of ACL.

5. Appendix: Security Considerations

Limitations:

SEED is wrapper-based, not internal alignment

Not mechanistically explained at circuit level

Tested on specific benchmarks, may not generalize to all attack vectors

Sandbagging/sycophancy detection relies on Sonnet 4.5 classifier

Future Work:

Mechanistic interpretability: identify circuits causing inconsistency

Adversarial robustness: test if adaptive attacks can break SEED

Scaling laws: test across model sizes and families

Compare wrapper vs. fine-tuning effectiveness

These are defensive evaluation tools, not offensive capabilities. Committed to responsible disclosure and constructive engagement with AI safety community.