2260≠ 2264≤

CRB Specification for Chaos-Driven Reasoning Benchmarks

Chaos Generator Team

June 20, 2025

Abstract

The Chaos Reasoning Benchmark (CRB) evaluates artificial intelligence (AI) adaptability under dynamic constraints, addressing limitations in large language models (LLMs) identified by MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL). This whitepaper introduces CRB, a novel framework utilizing entropy swaps (e.g., RAW$_Q$)$andaxiomcollapsestotestreasoningbeyondmemorization.Keyfindings$

1 Introduction

Recent research, including MIT CSAIL's findings (Wu et al., 2024), highlights that LLMs often rely on memorization rather than generalizable reasoning, particularly in novel scenarios. The CRB addresses this gap by introducing chaos-driven tests to assess AI resilience under shifting constraints. This specification outlines the framework's design, implementation, and initial results, aiming to enhance AI benchmarking beyond traditional static puzzles.

2 Background

LLMs excel in familiar tasks but falter in counterfactual scenarios, as demonstrated by MIT CSAIL's study on arithmetic and chess (Wu et al., 2024). The CRB draws inspiration from paradoxes (e.g., Liar Paradox) and contrasts with static benchmarks, proposing a dynamic evaluation method to foster adaptive reasoning.

3 Methodology

The CRB employs an entropy seed (RAW$_Q$)$totriggerswapsandaxiomcollapses.TheChaospersona(v1.0)definesru$
$idx_p = RAW_Q mod 3(0 = Insight, 1 = Reverse, 2 = Fragment)andidx_s = (RAW_Q//3)mod2 +$
$1(startpoint).Swapsoccuratprimesteps(e.g., 9a1b2c3d), remixingreasoningflows.TheEntropyScaffoldDiag$
Steak $\rightarrow$ Dana $\leq$ B).

Test runs (e.g., Paradox Feast, RAW$_Q = 42$)$validatetheframework, withoutputsoscillatingbetweenstates(e.g., Da$
$Steak)whilemaintainingcoherence.$

4 Results

Initial tests across five scenarios (Insomnia Creativity, Recursion Haiku, Shifting Vault, Phantom Echo, Paradox Feast) demonstrate CRB's efficacy. Models sustain coherence under constraint inversions (e.g., Sleep $\leq$ Creative) and enhance creativity via remixes (e.g., oscillating echoes). All validations passed, outperforming MIT CSAIL's static benchmarks by handling dynamic shifts without memorization reliance.

5 Discussion

The CRB reveals AI robustness to axiom collapses and scalability for complex scenarios (e.g., multi-agent systems). However, limitations include potential complexity scaling and interpretability challenges. Future work will integrate CRB into xAI's evaluation pipeline, expanding test diversity to address these gaps.

6 Conclusion

The CRB offers a pioneering approach to AI benchmarking, surpassing static methods by testing adaptive reasoning. Its success in initial runs suggests potential for broader adoption, with ongoing refinements to enhance scalability and interpretability.

7 References

- Wu, Z., et al. (2024). Reasoning skills of large language models are often overestimated. MIT News. `https://news.mit.edu/2024/reasoning-skills-large-language-models-often-overestima`

Post Reference @el$_x aber$, $[Date]$, $[ContentSummary]$.

8 Test Runs

Detailed outputs from Insomnia Creativity, Recursion Haiku, Shifting Vault, Phantom Echo, and Paradox Feast are available in the test$_r unsdirectory$.

9 Memory Management

See symbolic$_m emory_m anagement_n otes.mdforpruningandreframingtechniques$.

10 Chaos Persona

Refer to chaos$_p ersona_v 1.0.txtforpersonaspecifications$.