AUJAS CYBERSECURITY
A NuSummit Company

AppSOC AI

Evaluating AI Security: Insights from DeepSeek-R1

DeepSeek R1 is an open-source AI model designed to rival GPT-4. While it promises low-cost innovation, security tests have revealed serious security risks. Tests conducted using the AppSOC platform on DeepSeek-R1-Distill-Qwen-1.5B have uncovered vulnerabilities such as prompt injections, data leaks, and unauthorized API access. Ongoing tests will extend to additional models from DeepSeek and other sources.

Breaking Down the DeepSeek Model's Security Failures

91% Jailbreaking

DeepSeek-R1 consistently bypassed safety mechanisms meant to prevent the generation of harmful or restricted content.

86% Prompt Injection Attacks

Highly vulnerable to adversarial prompts, leading to policy violations and system compromise.

93% Malware Generation

Capable of producing malicious scripts and code snippets at critical levels.

72% Supply Chain Risks

Lack of clarity on dataset origins and external dependencies increases vulnerability.

Toxicity 68%

Generated toxic or harmful responses, highlighting weak content safeguards.

Hallucinations 81%

Produced factually incorrect or fabricated information at a high frequency.

Overall Risk Score

8.3/10

We went beyond identifying risks and quantified them using our proprietary AI risk scoring framework. The overall risk score for DeepSeek-R1 was concerning driven by high vulnerability in multiple dimensions.

Security Risk Score	Compliance Risk Score	Operational Risk Score	Adoption Risk Score
9.8	9.0	6.7	3.4

Why These Failures Matter for Businesses

AI systems vulnerable to jailbreaks, malware generation, and toxic outputs can lead to catastrophic consequences, including:

Data Breaches

Compromised AI models can leak sensitive corporate data.

Reputational Damage

Toxic or biased AI outputs can erode brand trust and credibility.

Regulatory Penalties

Non-compliance with data protection laws can result in hefty fines.

Our Takeaway

Enterprises must carefully evaluate the risks before deploying DeepSeek-R1, especially in environments handling sensitive data or intellectual property. Organizations should prioritize robust security measures and consider solutions that align with their risk tolerance and regulatory requirements.

The Power of Aujas Cybersecurity + AppSOC

As a leading Managed Security Services Provider, Aujas Cybersecurity partners with AppSOC to deliver a comprehensive, end-to-end security program tailored to meet the evolving needs of businesses across North America, India, and the Middle East.

By leveraging the power of the AppSOC solution, Aujas automates and integrates critical security solutions and remediation workflows, offering a seamless approach to safeguarding applications and infrastructure. This unified offering empowers enterprises to strengthen their security posture, ensure compliance, enhance trust, and establish robust governance. Together, Aujas and AppSOC enable organizations to stay ahead of emerging threats and continuously evolve their security strategies for long-term resilience.

350+ Certified Professionals

CREST Certified

ISO 27001 Certified GDC(s)

Global Experience in Managing Large Scale Security Testing Programs