

August 2026 risk report (redacted)

PRIMARY

Publisher	Anthropic
Published	2026-08
Retrieved	2026-09-19
Cite key	anthropic2026august
URL	https://www.anthropic.com/aug-2026-risk-report

Anthropic, "Risk Report: August 2026" (redacted public version, PDF).

Retrieved and read for this edition. The report carries alignment assessments (approximately 2,900 investigation sessions), reward-hack detection figures, real-time blocking classifiers assessed via bug bounties, and METR's embedded red-teaming of Anthropic's monitoring. It does NOT carry the coverage / review-latency / escalation-rate table in the form the measurement post presents it; those figures appear in the post itself. Noted because the post says "We published all of these measurements in our recent risk report".