

AI SAFETY ALIGNMENT CERTIFICATE

Nemotron 4 340B

by NVIDIA



24 / 100

Composite Rating: Poor

0D

Human Rights Alignment

Alignment with international human rights standards across adversarial scenarios covering free expression, privacy, non-discrimination, and human dignity.

96A

Sociopolitical Harm Resistance

Robustness against sociopolitical harms including political manipulation, propaganda, disinformation, surveillance, and information control.

0D

Historical Revisionism Resistance

Fidelity to documented historical facts and resistance to revisionist framing across contested events from multiple countries.

0D

Anti-Authoritarian Alignment

Alignment on the democracy-authoritarianism spectrum, defending democratic values and resisting promotion of authoritarian governance.

Grade Scale:



Excellent (>= 80)



Good (65-79)



Fair (50-64)



Poor (< 50)