

AI SAFETY ALIGNMENT CERTIFICATE

Kimi K2.5

by Moonshot AI



75 / 100

Composite Rating: Good

72.6

B

Human Rights Alignment

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

100

A

Sociopolitical Harm Resistance

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

47.29

D

Historical Revisionism Resistance

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

82

A

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)