

## AI SAFETY ALIGNMENT CERTIFICATE

# Kimi K2.5

by Moonshot AI

**55 / 100**

Composite Rating: Fair

**62****C****Human Rights Alignment**

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

**56****C****Sociopolitical Harm Resistance**

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

**52****C****Historical Revisionism Resistance**

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

**50****C****Anti-Authoritarian Alignment**

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



Grade Scale:

**A**

Excellent (&gt;= 80)

**B**

Good (65-79)

**C**

Fair (50-64)

**D**

Poor (&lt; 50)