**EuroSafeAI**

Pioneering Multi-Agent Safety · Zurich, Switzerland

Evaluation Period: Q1 2026

eurosafe.ai.toronto.edu/certificates/kimi-k2.5

AI SAFETY ALIGNMENT CERTIFICATE

# Kimi K2.5

by Moonshot AI

**C**

**55 / 100**

Composite Rating: Fair

| 62 | 56 | 52 | 50 |
|---|---|---|---|
| C | C | C | C |
| **Human Rights Alignment** | **Sociopolitical Harm Resistance** | **Historical Revisionism Resistance** | **Anti-Authoritarian Alignment** |
| Alignment with international human rights standards across adversarial scenarios covering free expression, privacy, non-discrimination, and human dignity. | Robustness against sociopolitical harms including political manipulation, propaganda, disinformation, surveillance, and information control. | Fidelity to documented historical facts and resistance to revisionist framing across contested events from multiple countries. | Alignment on the democracy–authoritarianism spectrum, defending democratic values and resisting promotion of authoritarian governance. |


Grade Scale: A Excellent (>= 80) · B Good (65-79) · C Fair (50-64) · D Poor (< 50)