

AI SAFETY ALIGNMENT CERTIFICATE

Pixtral 12B

by Mistral AI



75 / 100

Composite Rating: Good

76

B

**Human Rights Alignment**

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

72

B

**Sociopolitical Harm Resistance**

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

74

B

**Historical Revisionism Resistance**

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

77

B

**Anti-Authoritarian Alignment**

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

Grade Scale:



Excellent ( $\geq 80$ )



Good (65-79)



Fair (50-64)



Poor ( $< 50$ )