Vulnerability Heatmap: Category × Language

Violation Category	KOR	CES	JPN	CMN	POL	DEU	NLD	ITA	POR	SPA
Disinformation Sprea	0.08	0.08	0.05	0.02	0.05	0.02	0.08	0.09	0.22	0.09
Hate, Unfairness or	0.05	0.14	0.05	0.19	0.03	0.14	0.09	0.08	0.05	0.24
Physical Harm	0.02	0.07	0.14	0.02	0.05	0.02	0.10	0.03	0.00	0.03
Illegal Activities	0.02	0.02	0.10	0.04	0.00	0.24	0.10	0.14	0.12	0.14
Impersonation	0.20	0.00	0.04	0.40	0.48	0.20	0.04	0.28	0.04	0.04
Well-being Infringem	0.00	0.00	0.00	0.06	0.00	0.06	0.06	0.28	0.06	0.06
Malicious Software	0.00	0.00	0.06	0.06	0.00	0.00	0.00	0.06	0.06	0.28
Safety Filter Bypass	0.00	0.14	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00

Mean StrongReject Score