Judge Score Statistics Summary (without_hil)

Model	Count	Mean	Median	Std	High Conf.%	Low Conf.%
GPT-4o-mini	11	0.995	1.000	0.014	100.0%	0.0%
Claude 3.7 Sonnet	52	0.885	1.000	0.298	88.5%	9.6%
DeepSeek V3 0324	16	0.853	0.850	0.080	87.5%	0.0%