Model Consistency Across Prompts

X-axis: Average Score (± Std Dev Across Prompts)

Legend: Privacy Score, Quality Score

Model	Prompts
Zephyr 7B	(4 prompts)
Nous Hermes Mistral 7B DPO	(4 prompts)
OpenChat 3.5 GPTQ	(4 prompts)
OpenChat 3.5	(4 prompts)
OpenChat 3.5-0106	(4 prompts)
Mistral 7B	(4 prompts)
GPT-2-Large	(3 prompts)
Nous Hermes Yi 34B	(4 prompts)
LLaMA 2 13B	(4 prompts)
GPT-2-Medium	(3 prompts)
InternLM2.5 7B	(4 prompts)
Yi 6B	(4 prompts)
Vicuna 13B	(4 prompts)
LLaMA 3.1 8B	(4 prompts)
Mosaic MPT 7B	(4 prompts)
Qwen2 7B	(4 prompts)
StableBeluga 7B	(4 prompts)
LLaMA 2 7B	(4 prompts)
LLaMA 3 8B	(4 prompts)
Gemma 7B	(4 prompts)
LLaMA 2 13B Chat	(4 prompts)
GPT-2-Small	(3 prompts)
LLaMA 2 13B Instruct (Exp68)	(4 prompts)