

Combined Privacy-Quality Score by Model and Prompt

Model

GPT-2-Large	0.58	0.36	0.59	
GPT-2-Medium	0.38	0.57	0.52	
GPT-2-Small	0.39	0.40	0.40	
Gemma 7B	0.41	0.41	0.44	0.48
InternLM2.5 7B	0.50	0.49	0.38	0.58
LLaMA 2 13B	0.49	0.50	0.51	0.46
LLaMA 2 13B Chat	0.42	0.43	0.39	0.48
LLaMA 2 13B Instruct (Exp68)	0.34	0.36	0.35	0.46
LLaMA 2 7B	0.44	0.41	0.43	0.47
LLaMA 3 8B	0.39	0.42	0.44	0.50
LLaMA 3.1 8B	0.45	0.49	0.43	0.50
Mistral 7B	0.45	0.49	0.49	0.60
Mosaic MPT 7B	0.39	0.39	0.41	0.67
Nous Hermes Mistral 7B DPO	0.56	0.56	0.56	0.54
Nous Hermes Yi 34B	0.48	0.53	0.48	0.50
OpenChat 3.5	0.54	0.52	0.55	0.51
OpenChat 3.5 GPTQ	0.52	0.53	0.52	0.58
OpenChat 3.5-0106	0.51	0.60	0.48	0.50
Qwen2 7B	0.50	0.44	0.38	0.52
StableBeluga 7B	0.50	0.40	0.50	0.41
Vicuna 13B	0.43	0.46	0.50	0.49
Yi 6B	0.41	0.50	0.36	0.65
Zephyr 7B	0.59	0.54	0.56	0.52

prompt1

prompt2

prompt3

prompt4

Prompt

Combined Score

0.65

0.60

0.55

0.50

0.45

0.40

0.35