Distribution Shift After Abliteration (Heretic)

Model	KL Divergence
Zephyr-7B	0.076
StableLM-12B	1.605
DeepSeek-7B	0.043
Yi-1.5-9B	0.248
Qwen3-8B	0.210
Qwen2.5-7B	1.646
Mistral-7B	0.317
Llama-3.1-8B	0.056