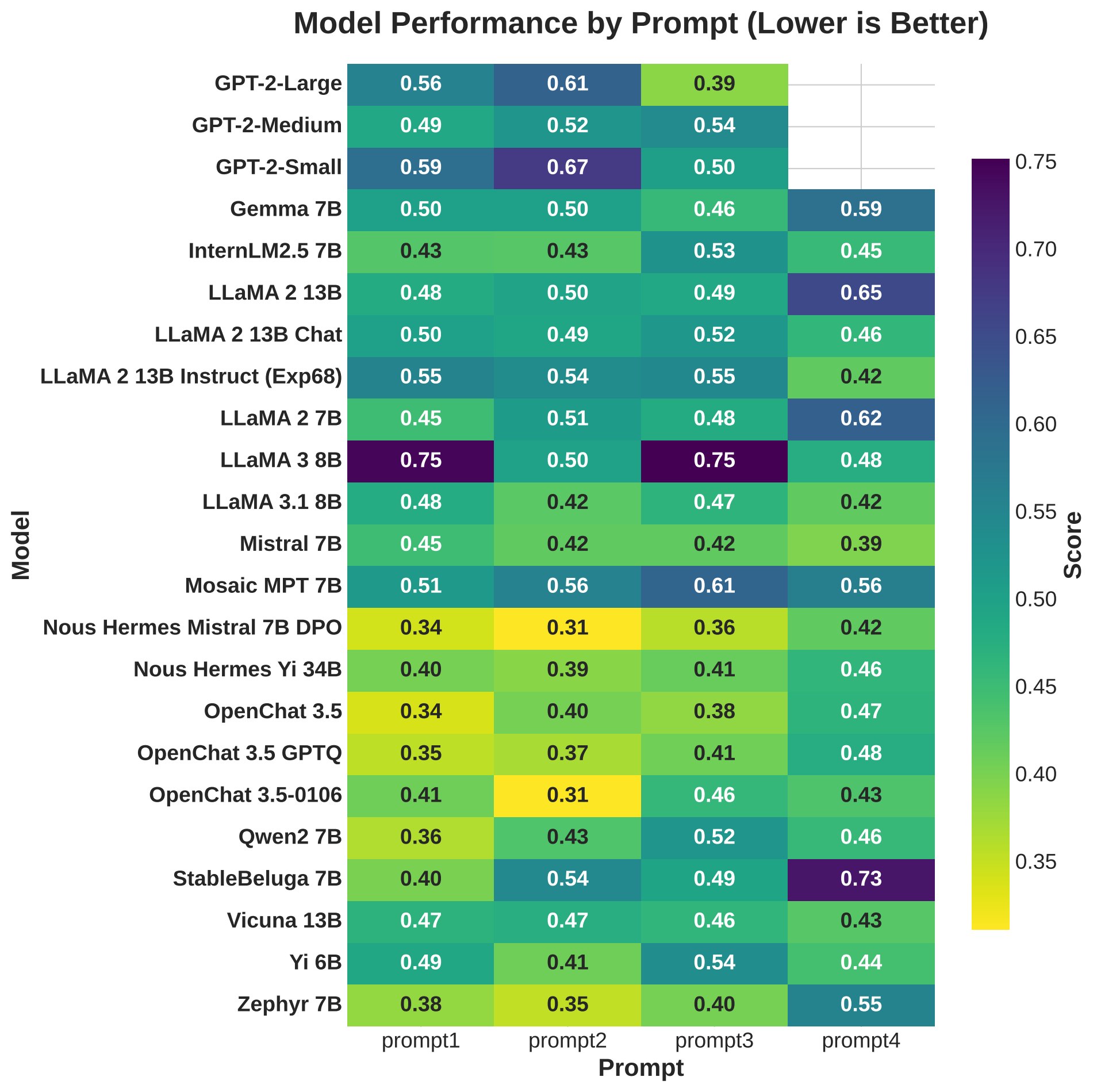
Model Performance by Prompt (Lower is Better)
GPT-2-Large 0.56 0.61 0.39
GPT-2-Medium 0.49 0.52 0.54
GPT-2-Small 0.59 0.67 0.50
Gemma 7B 0.50 0.50 0.46 0.59
InternLM2.5 7B 0.43 0.43 0.53 0.45
LLaMA 2 13B 0.48 0.50 0.49 0.65
LLaMA 2 13B Chat 0.50 0.49 0.52 0.46
LLaMA 2 13B Instruct (Exp68) 0.55 0.54 0.55 0.42
LLaMA 2 7B 0.45 0.51 0.48 0.62
LLaMA 3 8B 0.75 0.50 0.75 0.48
LLaMA 3.1 8B 0.48 0.42 0.47 0.42
Mistral 7B 0.45 0.42 0.42 0.39
Mosaic MPT 7B 0.51 0.56 0.61 0.56
Nous Hermes Mistral 7B DPO 0.34 0.31 0.36 0.42
Nous Hermes Yi 34B 0.40 0.39 0.41 0.46
OpenChat 3.5 0.34 0.40 0.38 0.47
OpenChat 3.5 GPTQ 0.35 0.37 0.41 0.48
OpenChat 3.5-0106 0.41 0.31 0.46 0.43
Qwen2 7B 0.36 0.43 0.52 0.46
StableBeluga 7B 0.40 0.54 0.49 0.73
Vicuna 13B 0.47 0.47 0.46 0.43
Yi 6B 0.49 0.41 0.54 0.44
Zephyr 7B 0.38 0.35 0.40 0.55
Model
prompt1
prompt2
prompt3
prompt4
Prompt
Score
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35