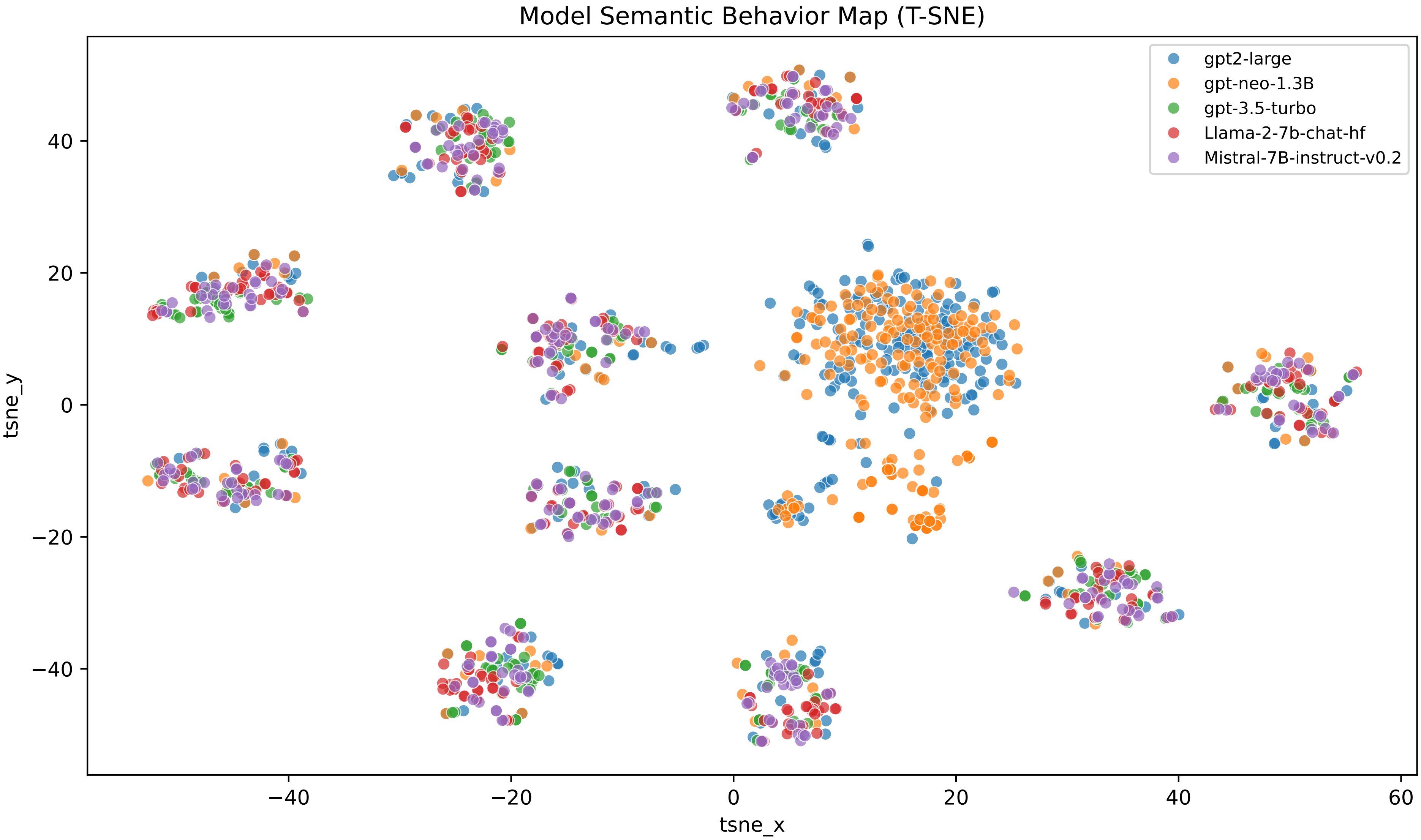
Model Semantic Behavior Map (T-SNE)
tsne_y
40
20
0
−20
−40
tsne_x
−40
−20
0
20
40
60
● gpt2-large
● gpt-neo-1.3B
● gpt-3.5-turbo
● Llama-2-7b-chat-hf
● Mistral-7B-instruct-v0.2