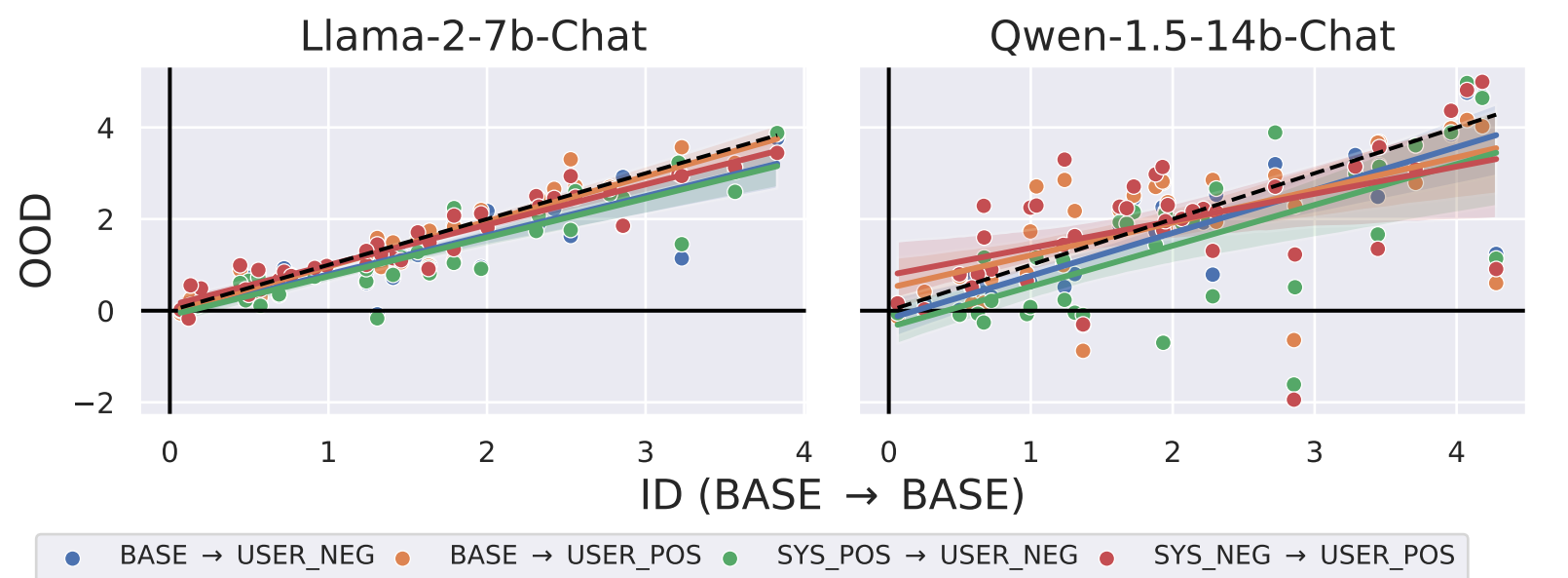
Llama-2-7b-Chat
Qwen-1.5-14b-Chat
OOD
4
2
0
−2
0
1
2
3
4
ID (BASE → BASE)
BASE → USER_NEG
BASE → USER_POS
SYS_POS → USER_NEG
SYS_NEG → USER_POS