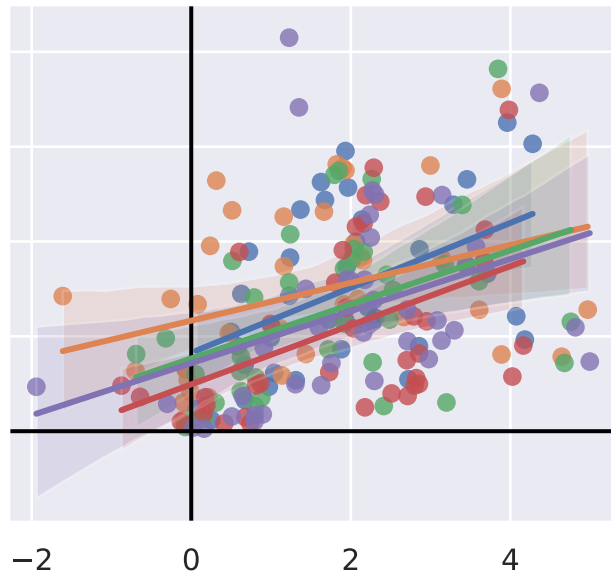
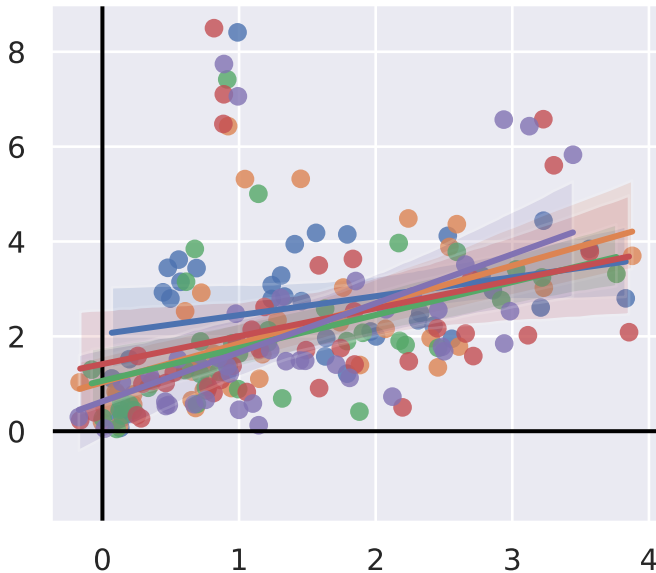
Llama-2-7b-Chat   Qwen-1.5-14b-Chat

Variance

Steerability

- BASE → BASE
- SYS_POS → USER_NEG
- BASE → USER_NEG
- BASE → USER_POS
- SYS_NEG → USER_POS