Conceptual similarity judgements

Openchat vs. Starling

Mistral-Instruct vs. Zephyr-Mistral

Gemma-Instruct vs. Zephyr-Gemma

Llama/Tulu vs. Llama-Chat/Tulu-DPO

P(multiple concepts)

Prompt manipulations

none, default

persona

random

nonsense

Temperature manipulations

1.5

2.0

1.0

0.5

0.0

NA

animals

politicians

Concept category

Non-aligned model

Aligned model

Human baseline (animals)

Human baseline (politicians)