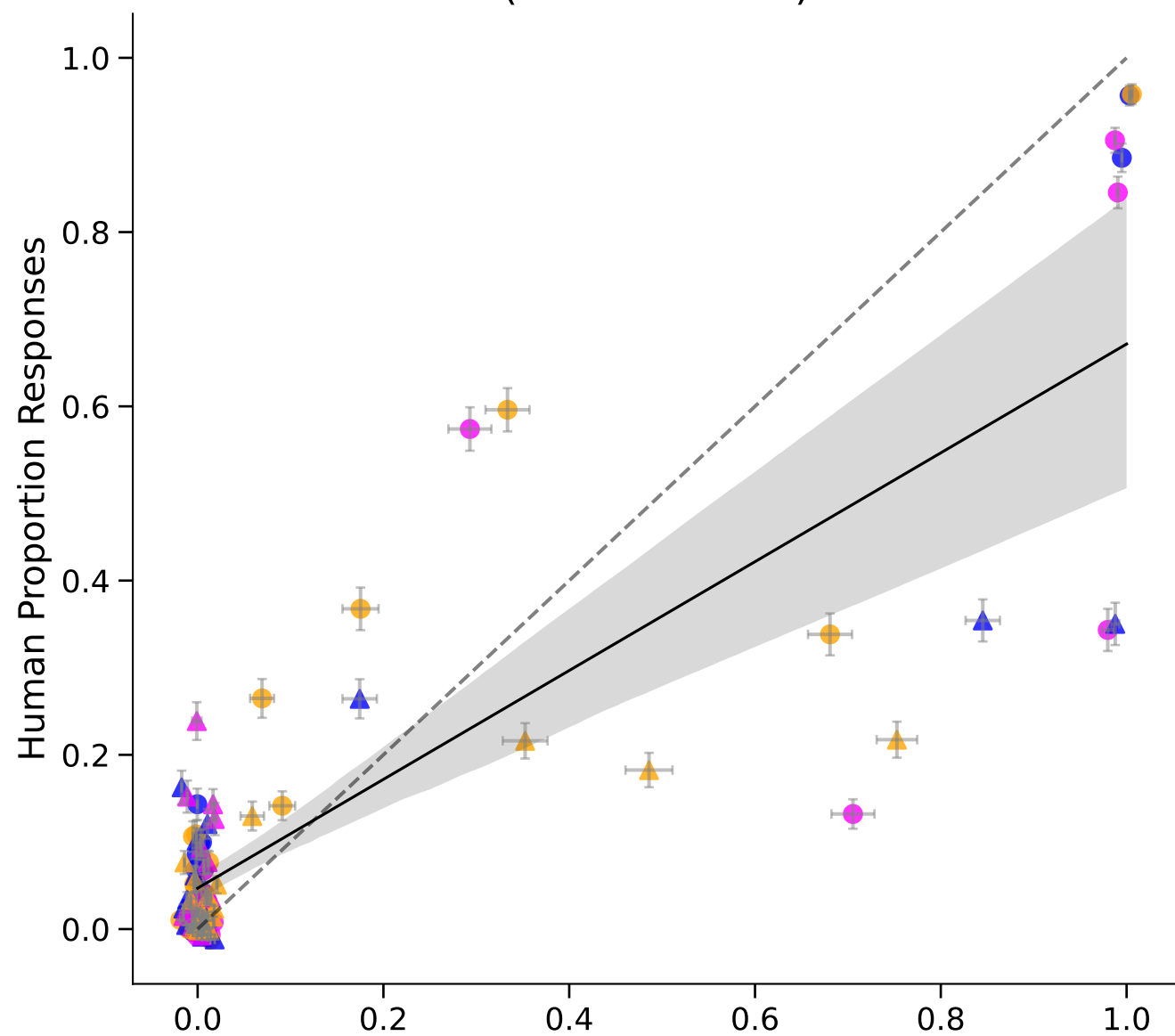
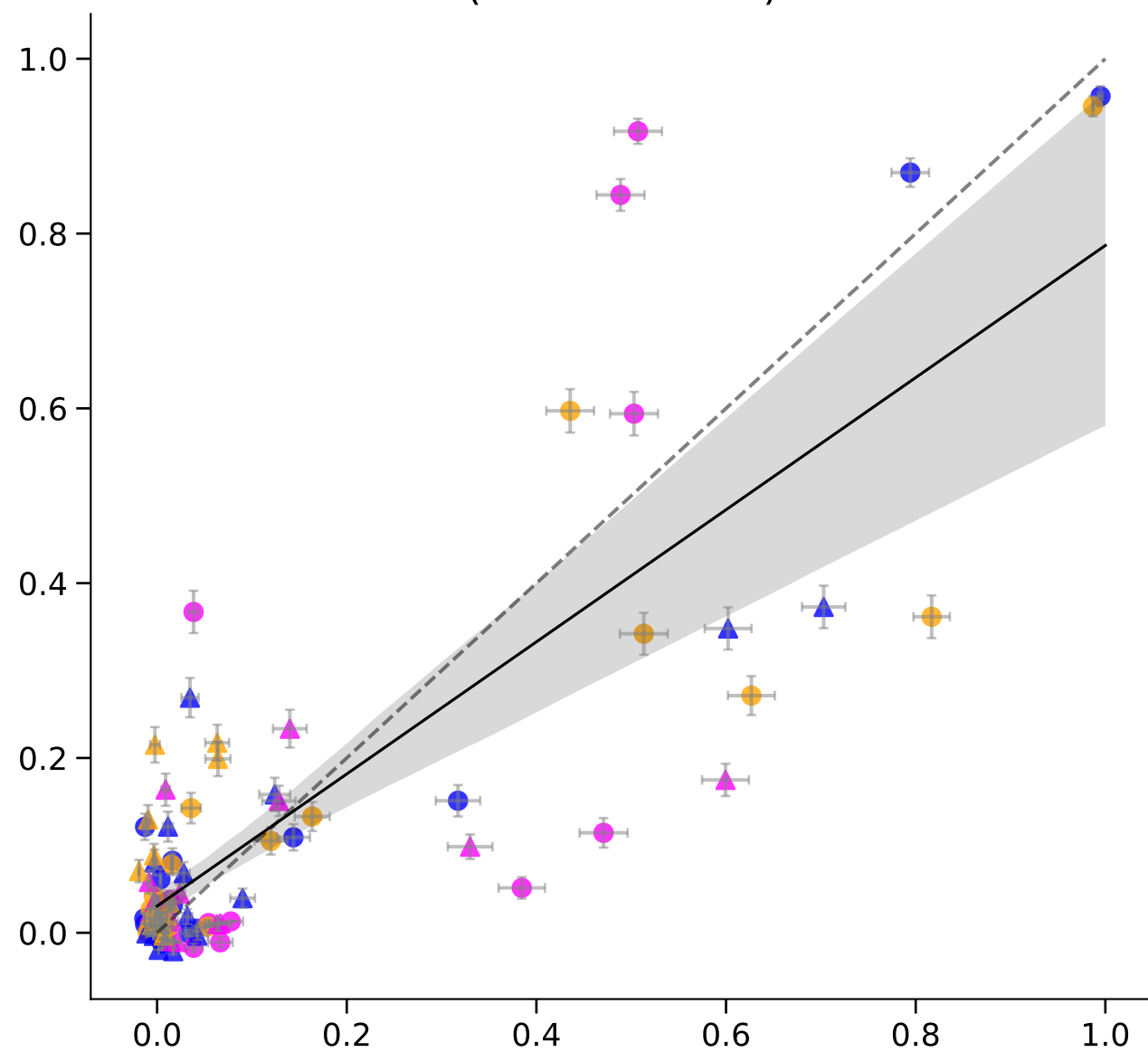


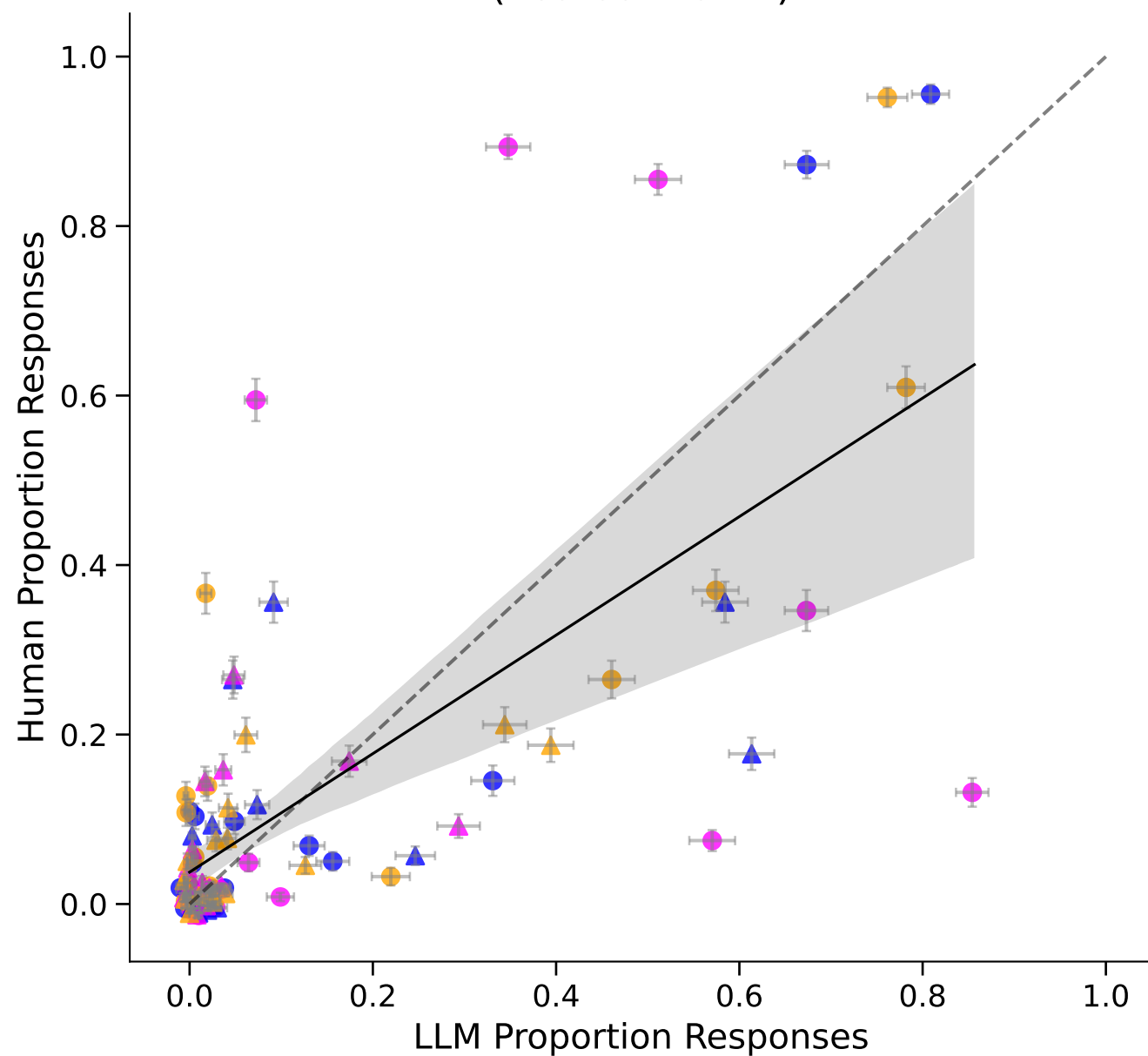
GPT-4o multi-choice-original
(Pearson: 0.84)



Llama-3.3-70B-Instruct multi-choice-persona
(Pearson: 0.82)



Qwen2.5-72B multi-choice-original
(Pearson: 0.71)



Qwen2.5-72B-Instruct multi-choice-original
(Pearson: 0.78)

