

Human likelihood rating
(item-normalized)

Model

∞ -gram

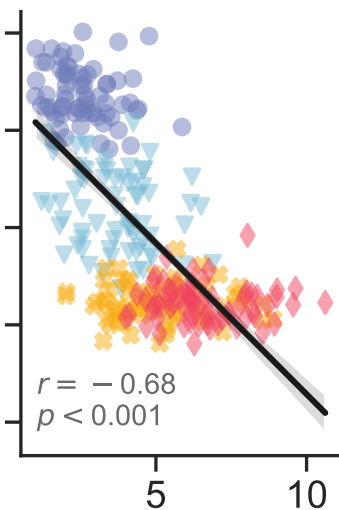
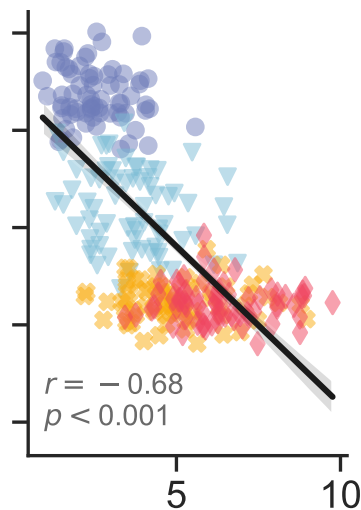
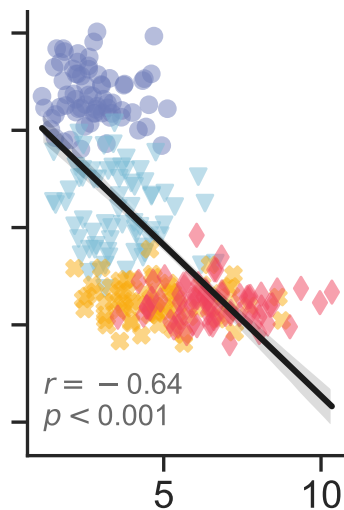
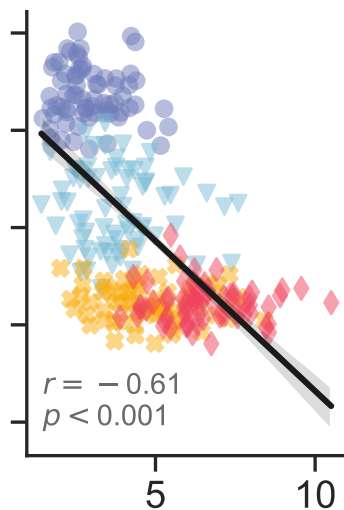
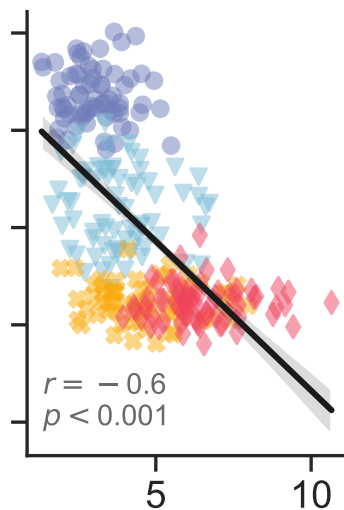
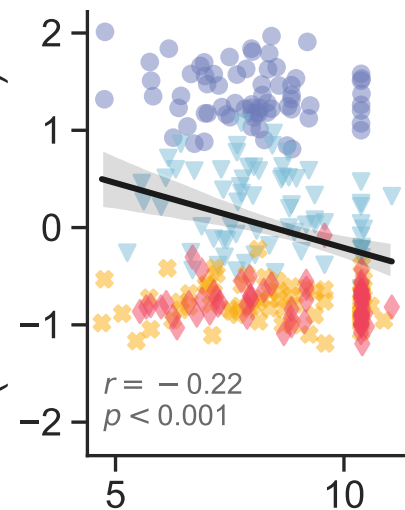
GPT-2
(124M)

GPT-2 Med
(355M)

GPT-2 XL
(1.5B)

Llama-3
(8B)

Llama-3
(70B)



Mean surprisal of continuation

● Probable

▼ Improbable

✕ Impossible

◆ Inconceivable