Spearman correlation with El Human Study 1.000 Human Study -0.917 LMC-EI (gpt-3.5-turbo-0125) -0.917 LMC-EI (council.mean pooling) -0.917 LMC-EI (dbrx-instruct) -0.900 LMC-EI (qwen1.5-72B-Chat) -0.883 LMC-EI (council.no agg) -0.883 LMC-EI (flagships council.no agg) -0.883 LMC-EI (smallest council.no agg) -0.883 LMC-EI (best council.no agg) -0.883 LMC-EI (gpt-4-0613) -0.883 LMC-EI (council.majority) -0.867 LMC-EI (gemini-1.5-pro) -System 0.867 LMC-EI (gemini-1.0-pro) -LMC-EI (command-r-plus) -0.867 0.867 LMC-EI (claude-3-opus) -0.850 LMC-EI (gwen1.5-110B-Chat) -0.833 LMC-EI (claude-3-haiku) -0.817 LMC-EI (gpt-4o-2024-05-13) -0.800 LMC-EI (llama-3-70b-chat) -0.750 LMC-EI (gwen1.5-32B-Chat) -0.728 LMC-EI (mixtral-8x7b) -0.650 MMLU -0.517 Chatbot Arena mined EQ subset (n=100) 0.483 Chatbot Arena -0.450 LMC-EI (Ilama-3-8b-chat) -0.383 EQ-Bench -

0.2

0.0

0.4

0.6

0.8

1.0

