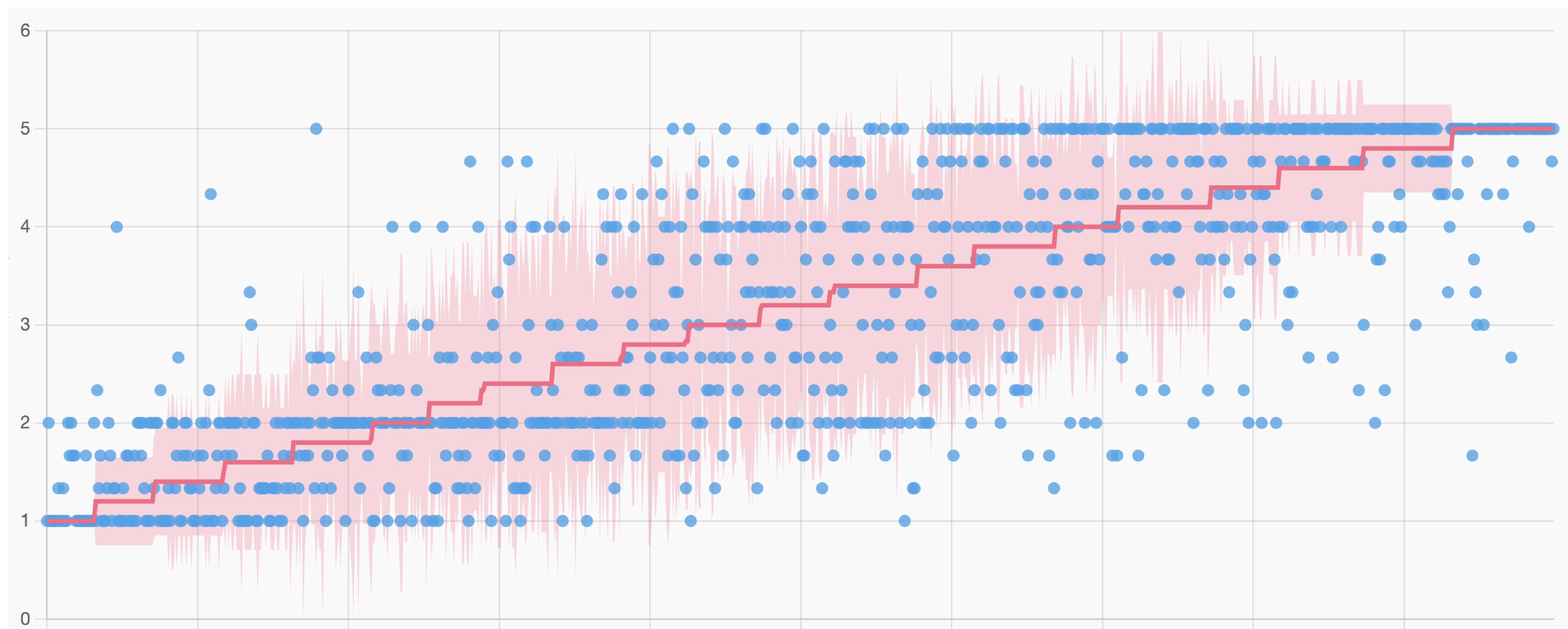
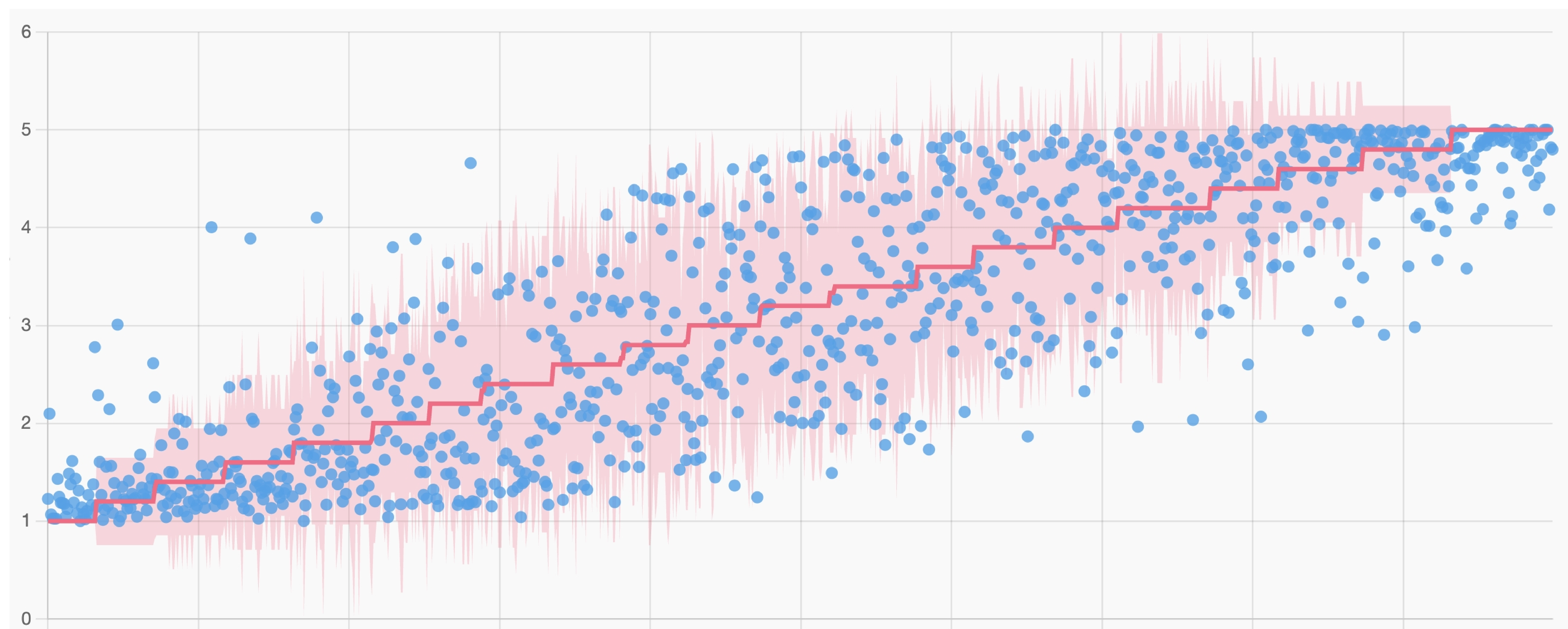


Mean human rating Area within 1 stdev Model prediction



gpt-5-mini | Acc. 0.86 | Spearman (ρ) 0.78 | Avg. 0.83 (Test)



E_{All} (10 LLMs) | Acc. 0.92 | Spearman (ρ) 0.85 | Avg. 0.89 (Test)