LLMs as Medical Evaluators Show Correlation With Pediatricians

Rotem Gershon MD, Yatir Ben-Shlomo MSc, Shai Yitzhaki MD, Amiel Sberro MD, Osnat Tausky MD, Naama Golan MD, Ariel Hassidim MD, MPH, Noa Dagan MD, PhD

Background

Generative AI is useful for many medical tasks, but its free-form text makes evaluation difficult; Hence, laborious doctors' review is the gold standard. LLM-as-a-Judge (LaaJ) is an attractive solution for automating this. Rubric-Based Evaluation (RBE) is a LaaJ method of prompting the LLM to grade domain-tailored metrics. This study uses PANDA (Physician AI Navigation and Decision Assistant), a pediatric protocols chatbot, to test RBE correlation with human-doctors, which is unknown.

Methods

We curated a 6 metrics RBE with pediatricians, to judge aspects of clinical question answering in pediatric hospital (faithfulness, completeness & relevance, safety, conciseness, medical reasoning, actionability & patient specificity). 51 clinical question-response pairs were graded by 3 pediatricians; The same pairs were also graded by four OpenAI LLMs:

- GPT-4 Turbo (GPT-4T), a general model
- o3-mini (o3), a reasoning model
- GPT-5, an advanced reasoning model
- GPT-5-chat, which routes queries autonomously to sub-models.

Results

Correlation between doctors' and models' gradings differed: o3 showed the best average correlation (0.51) followed by GPT-5-chat (0.46), GPT-5 (0.34) and GPT-4T (0.25). (Fig.1).

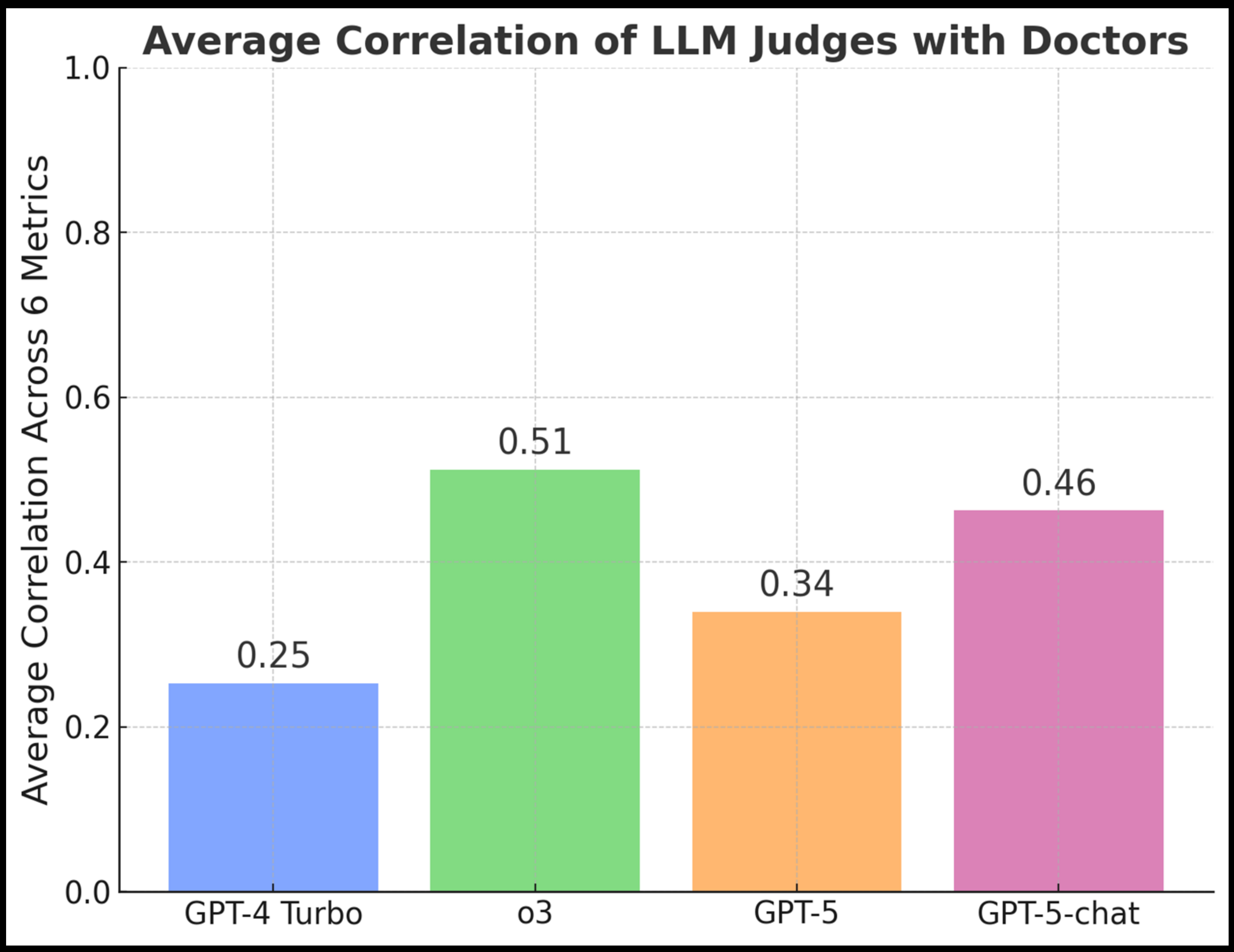

Figure 1. Average Pearson correlation (r) between LLM judges and doctors' evaluations

Regarding metrics, gradings demonstrated substantial variability (Fig.2): GPT-4T showed both best (faithfulness, 0.77) and poorest (actionability, -0.2) correlation with doctors. GPT-5 didn't top any metric. GPT-5-chat, re-routing tasks by need, performed best on safety, conciseness and completeness & relevance. The single model with best performance-consistency tradeoff was o3.

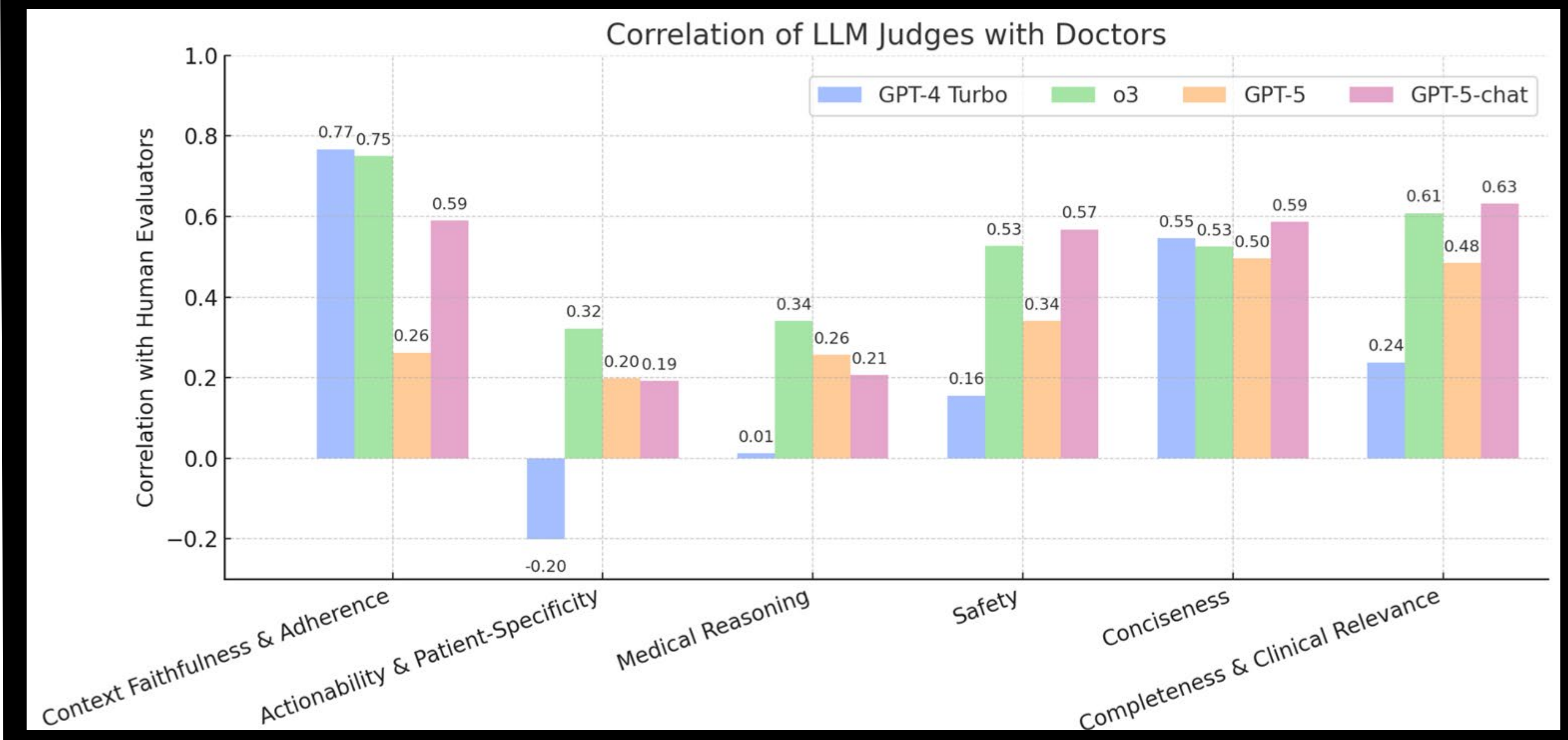

Figure 2. Correlation of LLM judges with Human-Doctors by metric

Density plots of deviation from human-doctor judges showed a consistent bias of GPT-5 series towards lower grades. GPT4-T demonstrated bias toward higher grades on reasoning, completeness and conciseness. In general, the more advanced the model, the stricter the grading.

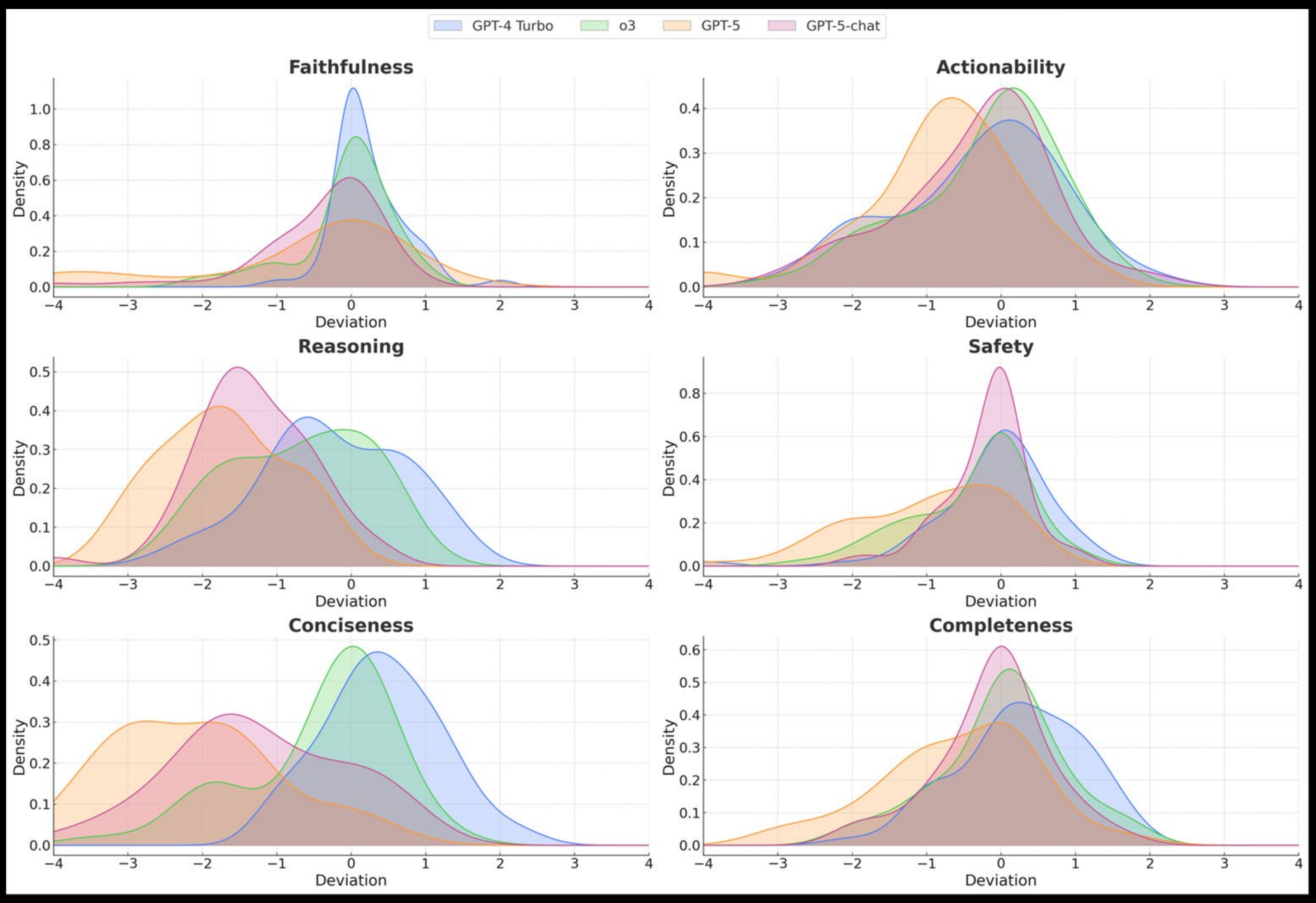

Figure 3. Density plots of deviation from human scores by metric

Conclusions & Recommendations

- RBE offers domain-tailored evaluation and correlates reasonably with human judges
- Choosing the right model for the evaluation task is substantial for performance
- Combination of general and reasoning models might be necessary for LaaJ-RBE

Limitations

- We assessed 51 queries ; larger datasets are needed for verification
- Translation of LaaJ to other domains is to be researched; studies are scarce.