Northwell Health

Staten Island University Hospital Northwell Health

Accelerating Academic Excellence: Leveraging AI to Evaluate Medical Study Abstracts

Yiela Saperstein, MD, MS; Mario Castellanos, MD; Jackson Abou Chehade, MD; Luis Quiel, MD, MBA, Suzanne El-Sayegh, MD

Background

The rapid evolution of generative AI models has created a need to understand their role in medical education/research evaluation. As training programs seek scalable solutions to expand scholarly output, the integration of Large language models (LLMs) into research assessment represents a late-breaking development with direct implications for informatics. In clinical decision-making, studies show that LLMs excel at structured information processing, but struggle with context-dependent/subjective decisions. Our study offers timely evidence on the strengths and limitations of LLMs for objective versus subjective research evaluation tasks.

Goals

Assess AI vs. Human Performance: Investigate whether LLMs trained on general medical knowledge can evaluate research abstracts as accurately and more consistently than human evaluators across objective and subjective categories.

Methods

223 abstracts from *SIUH Medicine* Department's **2023, 2024** and **2025** *Annual Research Days* were analyzed.

Each abstract was scored by **2 clinical faculty** using a **standardized rubric** on the categories of *Background*, *Methods*, *Results*, and *Discussion/Conclusion* (1-5 in each category). Faculty scores were **averaged** to establish a benchmark.

4 *LLMs (***Gemini 1.5 flash, Gemini 2.5 pro, GPT-4** and **GPT5**) were tasked with scoring the abstracts using the **same rubric** via an **API** on a proprietary Northwell platform.

Python code was used to automate the process and capture LLM generated **scores** and **comments**.

Statistical analysis (*mean*, *median* and *standard deviation…)* were calculated for each evaluator/ category and for total score. The scores were then evaluated using **weighted Cohen's Kappa** and **Spearman correlation** along with their p values to assess agreement and correlation. **Kappa values were averaged** and **analyzed by category** to assess variability across *each model* and *category*.

Results

Mean and Standard Deviation by each judge/LLM

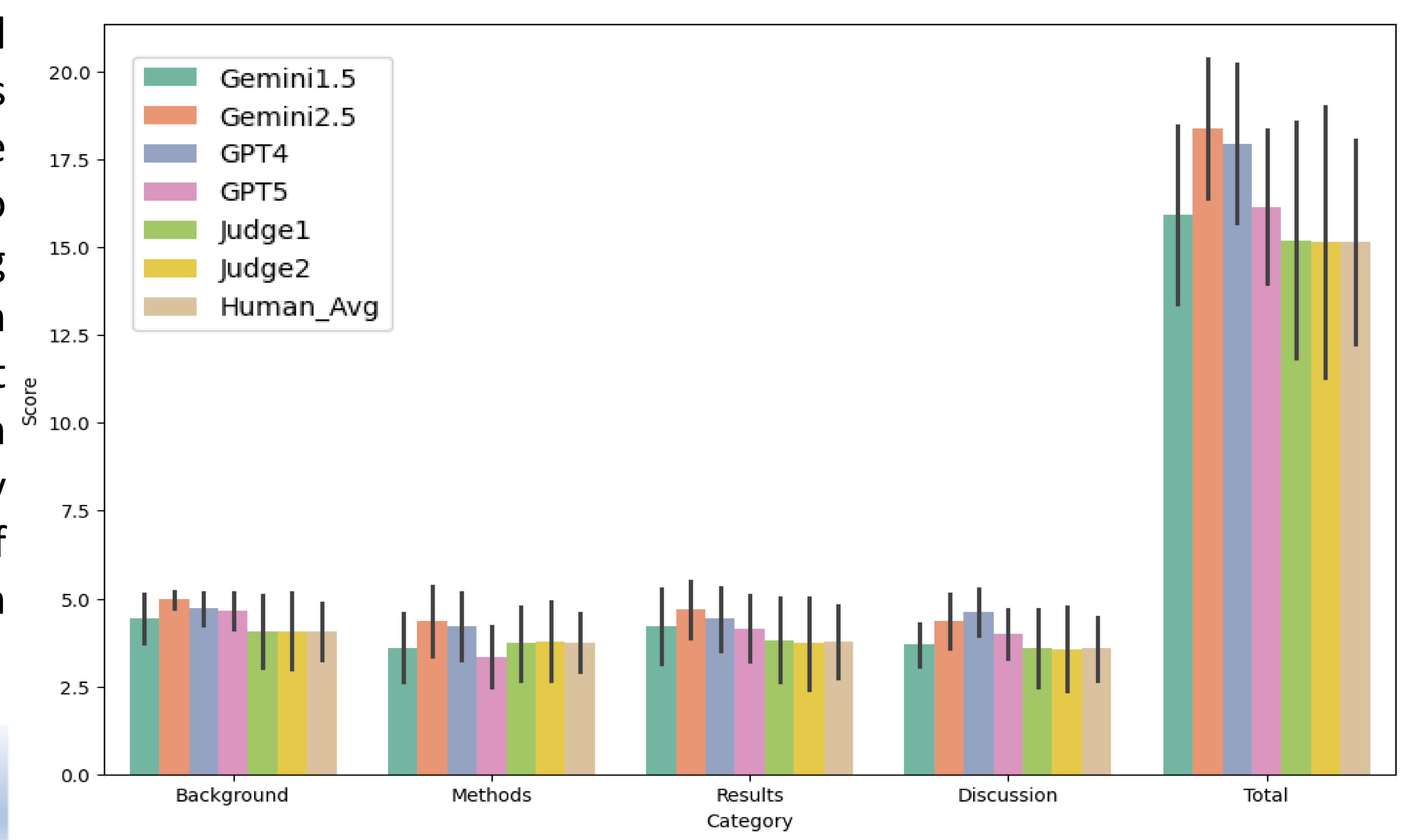


Figure 1: LLMs assigned higher overall scores to abstracts compared to faculty evaluators, with Gemini2.5 and GPT4 yielding the highest average total scores of **18 (SD = 2)**, followed by Gemini1.5 and GPT5. Faculty evaluators provided lower scores, with human average total score of **15 (SD = 3)**.

Averaged Weighted Kappa

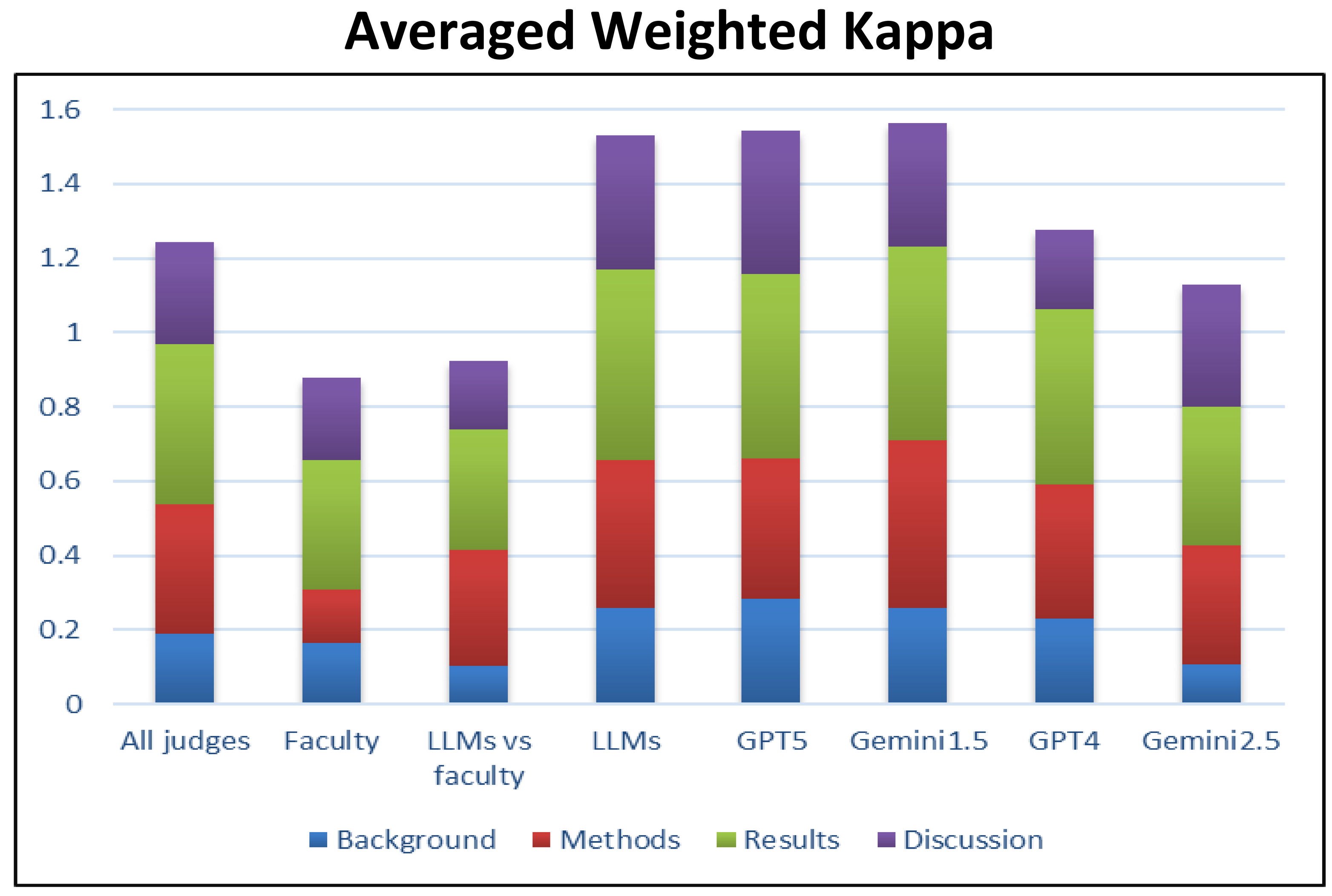
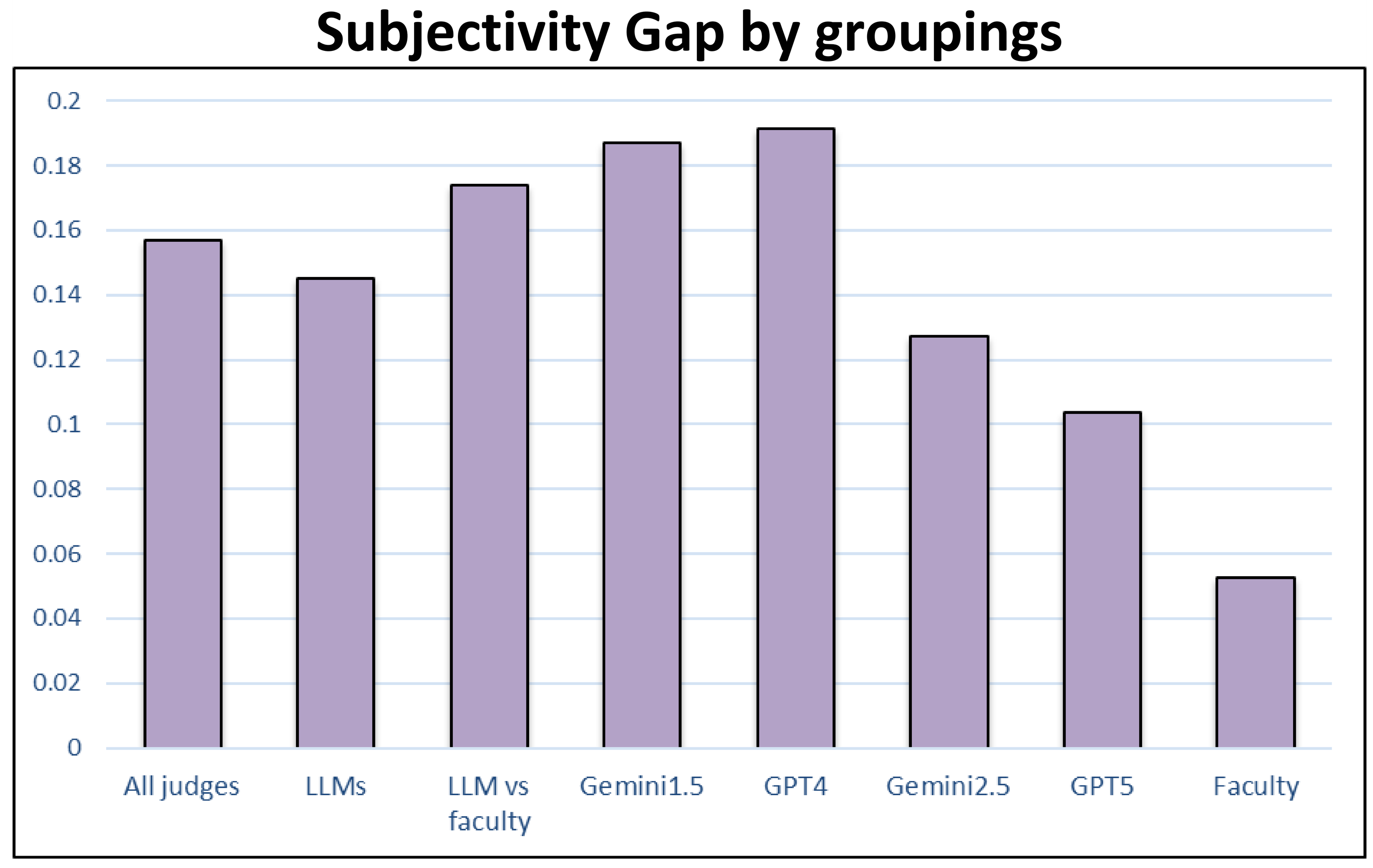


Figure 2: Weighted Kappa was calculated for each pair of evaluators. These values were averaged together for different groupings. LLMs demonstrated higher consistency and agreement in grading than faculty. *Average total weighted kappa between faculty graders* was **0.29** while *between LLMs* was *0.46*. *Gemini1.5 flash* followed by *GPT5* had the highest averaged weighted kappa (**0.50** and **0.48** respectively) compared with *GPT4* (**0.39**) and *Gemini2.5 pro* (**0.32**).

Methods and Results sections had higher overall agreement than background and discussion sections.

Subjectivity Gap by groupings



Figures 3: For each grouping, average kappa scores were again averaged for *Methods and Results sections* vs for *Background and Discussion sections*. The difference was calculated and labelled the *Subjectivity Gap*.

The **subjectivity gap was higher for LLMs than faculty**. The *older models* (Gemini1.5 and GPT5) had higher Subjectivity Gaps than the *newer models* (Gemini2.5 and GPT5) with GPT5 having the lowest Subjectivity Gap among the LLMs.

Summary & Future direction

LLMs show promise for streamlining research assessment by reducing variability and providing reproducible scoring. Across all categories, **LLMs demonstrated greater internal consistency than faculty**, though they **graded more generously**, especially GPT-4 and Gemin2.5 pro. **LLMs show highest concordance across objective categories** (Methods, Results) consistent with previous research showing that *LLMs struggle in subjective decisions*. However, the **newer models**, GPT-5 and Gemini2.5 flash, **have less of a discordance** between objective and subjective categories suggesting a distinct evaluative style and reinforcing the importance of ongoing benchmarking as new models emerge. Future directions include *incorporating newer models*, *refining prompts* and *assessing AI-generated feedback*. While not a replacement for mentorship, **LLMs may provide a scalable tool to enhance fairness and efficiency in residency research evaluation**.