Scatterplot Squad

Human vs. Artificial Intelligence: A Comparative Analysis of AI in the Academic Setting

---

Becca Lee
Phoebe Oblak
Don Iwejuo
Emma Grace Walter

Stats 199

Spring 2026

Research Question

- Among Model UN position papers, do AI-generated (ChatGPT) papers differ from student-written papers (both award-winning and non-award-winning) in mean rubric scores for writing, relevance, consistency, and analysis?

AI vs Student Writing

- ChatGPT generated Model UN Position Papers vs student-written papers
- Student-written papers are grouped by award-winning and non award-winning
- Papers are scored on standardized rubric: writing, relevance, consistency, and analysis

Data source (2024)

- Harvard Dataverse dataset (2024)
- Compares real student papers and AI-generated responses to identical prompts

Hypothesis

- We hypothesized that AI-generated papers will score comparably to non-award-winning student papers on surface-level criteria like writing and relevance, but will be outperformed by award-winning student papers on criteria requiring deeper geopolitical nuance, such as consistency and analysis.

Categorical Variables

author_type
AI · Human-No Award · Human-Award Winning

committee
18 UN committees (GA1, GA2, GA3, IAEA, UNEA, SC…)

country
Country the paper represents

country_committee
Unique paper identifier (country + committee)

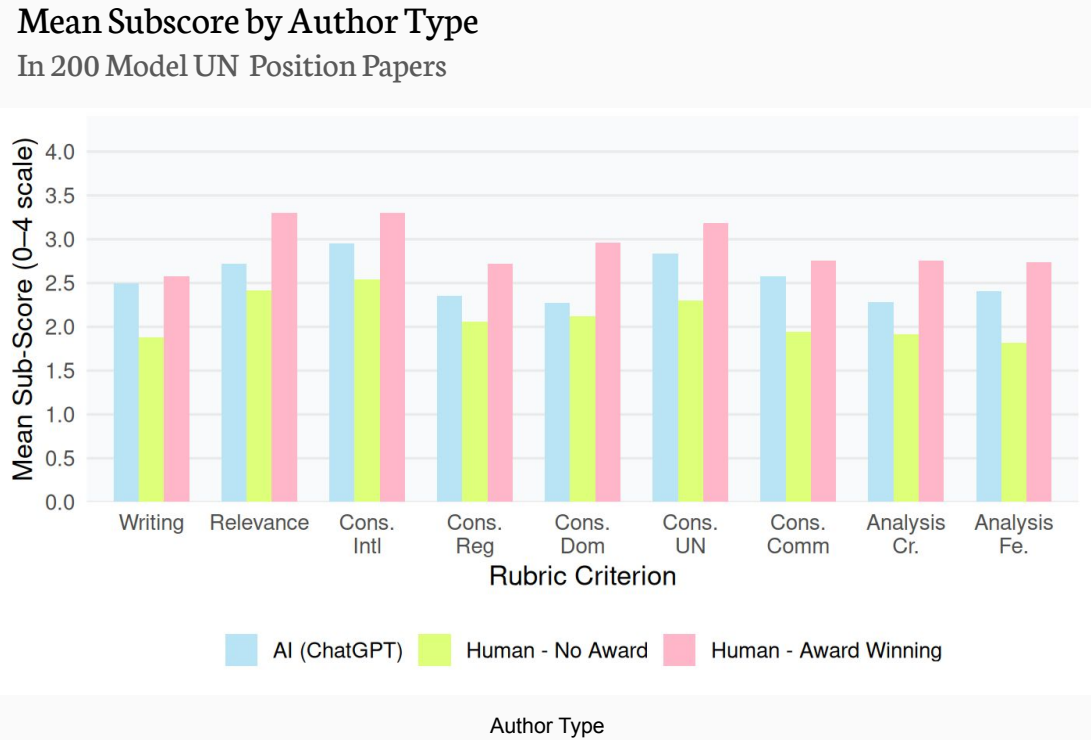
Numerical Variables (0-4 scale)

- **writing** — quality of written expression
- **relevance** — on-topic alignment
- **consistency_international**
- **consistency_regional**
- **consistency_domestic**
- **consistency_un**
- **consistency_committee**
- **analysis_creative** — creative solution quality
- **analysis_feasible** — feasibility of solutions
- **total** — sum of all 9 sub-scores (0–36)

Ethical Consideration

- Academic integrity: AI performance on rubrics challenges assignment design
- Student data anonymised in published dataset
- AI advantage on surface criteria does not equal equivalence to human learning
- Prompt quality strongly modulates AI scores — not a fair 1:1 comparison

EDA Visualization



EDA HIGHLIGHTS

Methodology

- We first calculated the mean subscore for each of the 9 rubric criteria separately for all three author types (AI, Human – No Award, Human – Award Winning)
- Our initial visualization groups bar charts by rubric criterion to allow direct side-by-side visual comparison across groups on the same 0–4 scale

Primary Conclusions

- Our visualization shows that AI consistently outperforms non-award students across all criteria
- However, award-winning students lead on every single subscore
- The closest score between AI and award-winning students was in writing, while categories requiring creativity and context typically had larger disparities

Next Steps

- While this visualization comprehensively highlights scores across all categories, it is visually hard to distinguish overall averages
- In our next visualizations, we sought distinguish more clearly between different grouped categories, such as writing, consistency, and analysis

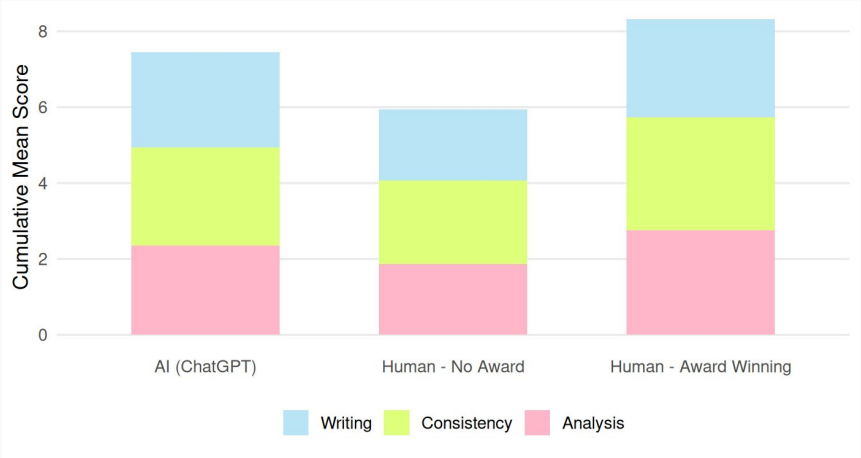
Visualizations

Distribution of Total Score by Author Type

In 200 Model UN Position Papers

Score Composition by Author Type

Writing • Consistency • Analysis



Tables

Table 1: Mean Rubric Scores by Author Type - al variables (n = 200)

Author Type	Writing	Relevance	Cons. Intl	Cons. Reg	Cons. Dom	Cons. UN	Cons. Comm	Anal. Cr.	Anal. Fe.	Total
AI (ChatGPT)	2.50	2.72	2.95	2.35	2.27	2.84	2.58	2.28	2.41	22.90
Human – No Award	1.88	2.42	2.54	2.06	2.12	2.30	1.94	1.92	1.82	19.00
Human – Award Winning	2.58	3.30	3.30	2.72	2.96	3.18	2.76	2.76	2.74	26.30

Table 2: Committee Level Breakdown - AI never exceeds award-winning average

Committee	AI Mean	Award Mean	No-Award Mean	AI > Award?
GA1 (n=64)	25.28	28.12	19.13	No
GA3 (n=16)	21.12	28.75	23.75	No
SC (n=8)	23.25	30.50	5.50	No
UNHRC (n=8)	21.00	29.00	13.33	No
HRC (n=4)	12.50	31.00	12.00	No

Overarching Conclusions and Future Directions

1. AI (mean total 22.90) outperforms average non-award students (19.00) on all criteria, the null hypothesis is rejected. AI scores better on consistency_international (2.95 vs 2.54) and relevance (2.72 vs 2.42)
2. Award-winning students (mean 26.30) outperform AI on every single sub-score, most noticeably in relevance (3.30 vs 2.72) and consistency_domestic (2.96 vs 2.27); the alternative hypothesis is supported
3. 35 of 100 AI papers scored below the no-award human mean, performance varies widely (SD=8.10). Committees like HRC and CSW show the largest AI underperformance, suggesting topic complexity matters
4. **Limitation:** rubric scores measure structured criteria but may miss argumentation depth, originality, and geopolitical insight, key skills AI may systematically lack

Thank you!

Becca Lee
Phoebe Oblak
Don Iwejuo
Emma Grace Walter

Stats 199 Spring 2026