UNIVERSITÄT
BIELEFELD
Faculty of Sociology

GOR

The AI Reviewer – A GOR case study

The Potential of Large Language Models in Scientific Research Evaluation

Dorian Tsolak, Zaza Zindel, Simon Kühne & Daniela Wolf

Contact: dorian.tsolak@uni-bielefeld.de

Research Question

Can large language models (LLMs) complete (simple) reviewer tasks to ease the workload of researchers?

Background

- LLMs are significantly impacting various professions by automating many routine tasks (World Economic Forum, 2023).
- Researchers are often asked to review abstracts for conferences, adding to their already high workload.
- The main objective of reviewing conference abstracts is to determine the quality of the presented research (1), the fit with the scientific domain (II), and the alignment with the conference‘s focus (III).
- Reviewers are selected because of their ability to cover these criteria, leveraging their comprehensive expertise in the field.
- LLMs, trained on a vast array of scientific literature, potentially encompass most of the knowledge within a given field.
- While LLMs may struggle to incorporate author information into their review, this is not a concern because scientific submissions to conferences and journals are usually reviewed double-blind.
- Given their extensive training data, LLMs might be able to properly assess and evaluate short, simple scientific texts without tables and figures, such as conference abstracts.

Data & Methods

Data:

1. Original text data, i.e., submitted GOR 24 conference abstracts (see www.conftool.org/gor24/sessions.php, no author information)
2. Scientific reviewer assessments in the form of 2-4 numeric scores per conference abstract (scores only, no text)
3. AI reviewer assessments in the form of 2-4 numeric scores per conference abstract

Model Specification:

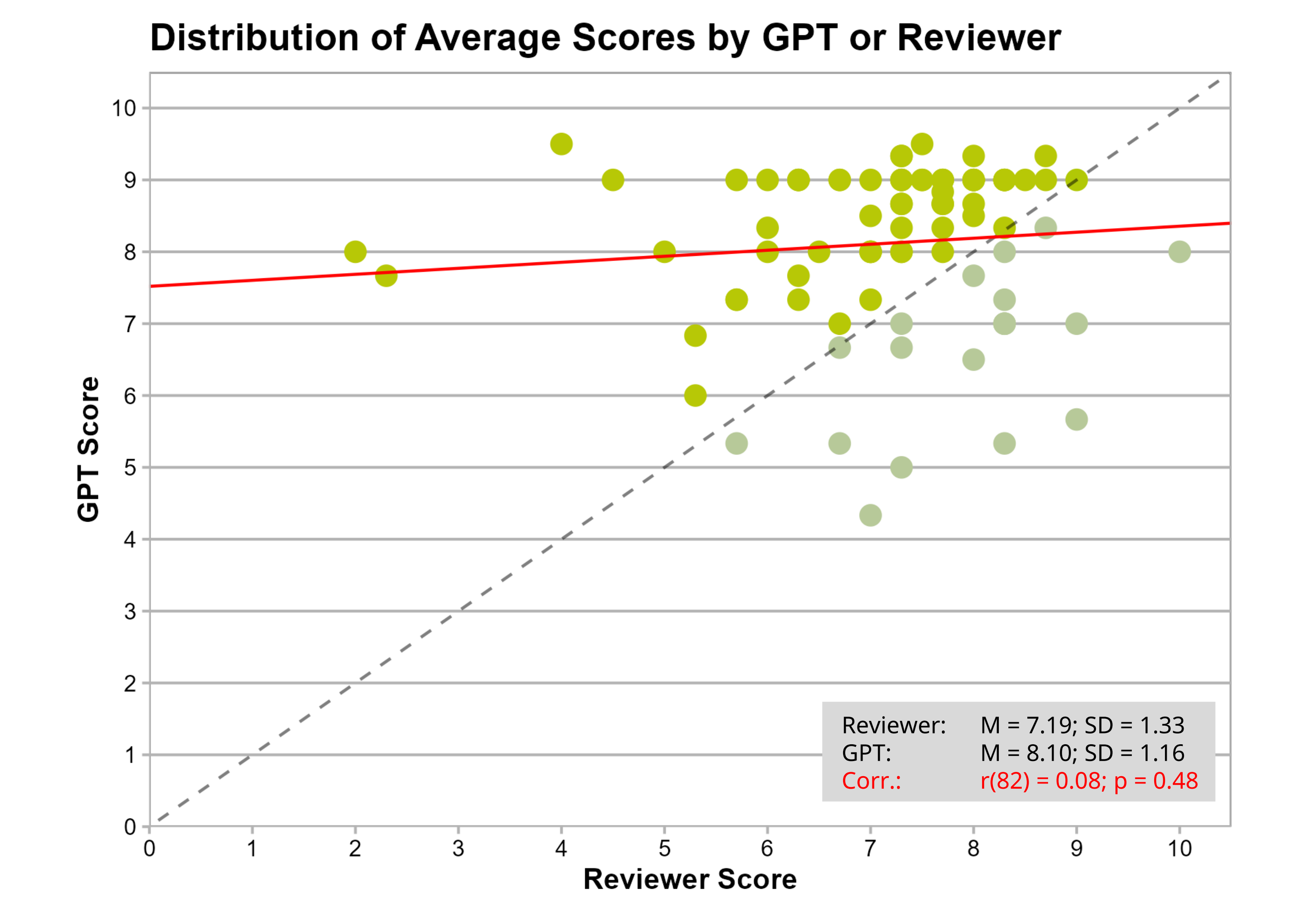
We used ChatGPT (gpt-4-0125-preview) and instructed it to adhere to the same review criteria as scientific reviewers by starting the prompt as follows:

> *“Write a short review for the following abstract which has been submitted to a scientific conference on general online research. Also, rate it on a scale from 0-10 that reflects the content of your review. The scale is structured as follows. 00 means 'Definitely reject (has no merit)', 02 means 'Probably reject', … [and so on and so forth]”.*

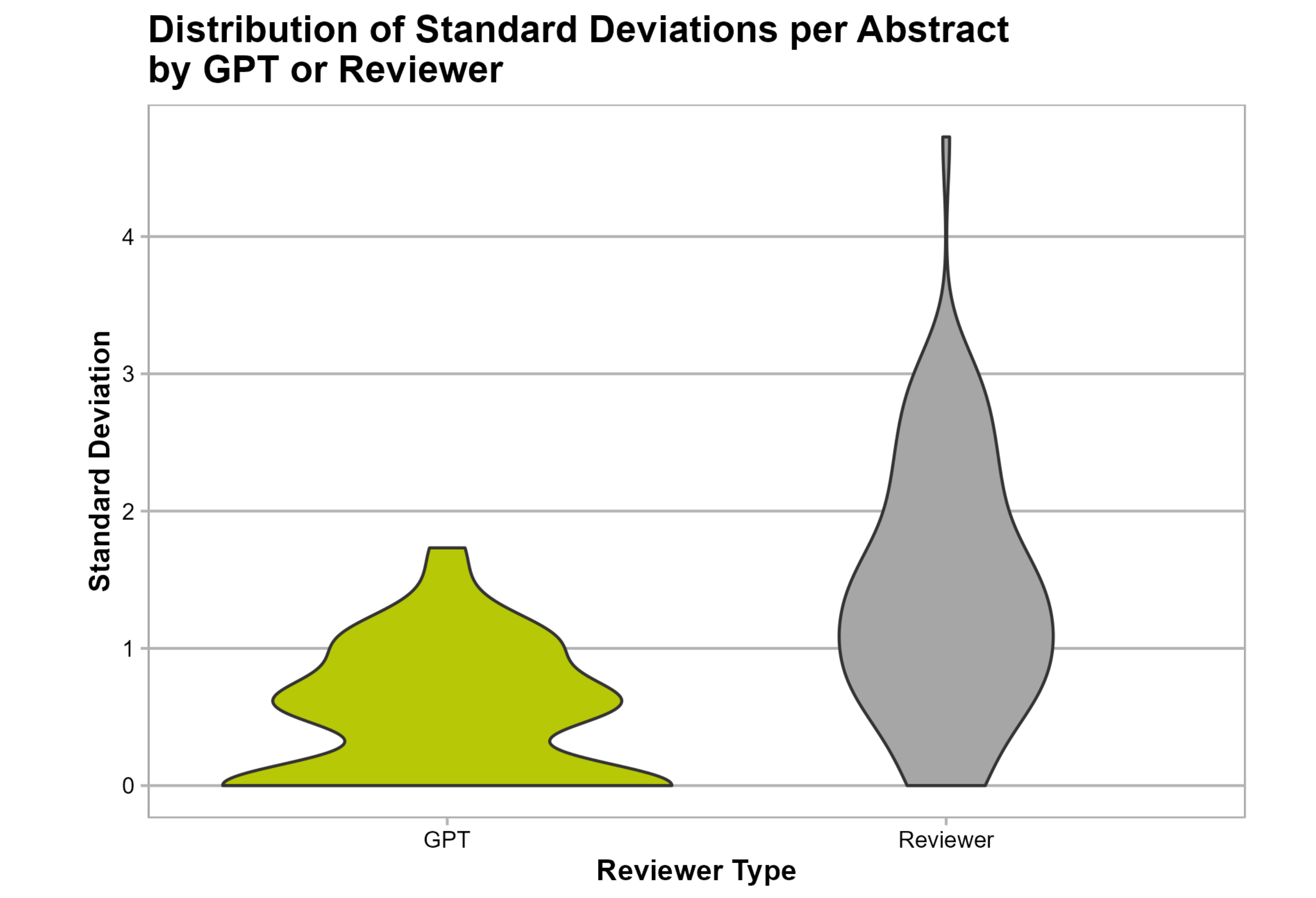
We then provided the GOR review scheme that human reviewers also receive as instructions.

Results

Distribution of Average Scores by GPT or Reviewer



Distribution of Standard Deviations per Abstract by GPT or Reviewer



- On average, the LLM rates abstracts better by almost 1 point (0.91) on a 0-10 score.
- Multiple assessments of the same abstract are more homogenous across multiple LLM prompts compared to multiple human reviewers.
- There is almost no relationship between the human reviewer ratings and the LLM ratings, indicating that variance in ratings across abstracts by the LLM is almost random.

Insights

- LLM ratings for abstracts seem almost random and largely meaningless, as they do not indicate abstract quality i.e., every abstract receives a good rating, but the allocation of the best and the worst ‘good’ ratings appears to be random.
- LLMs can produce reasonable-looking results when reviewing scientific abstracts and tend to be less critical on average and less extreme than human reviewers, but results should not be taken seriously (even though they appear very reasonable).

Limitations

- *Case study problem:* One specific prompt, model, use-case and conference. The latter with little variation in the quality of the abstracts (high scores on average, even when evaluated by the, more critical, human reviewers).
- *Ground truth assumption:* We assume that human reviewers are competent and that their reviews reflect the abstract quality

References:
World Economic Forum. (2023). Jobs of Tomorrow: Large Language Models and Jobs: WHITE PAPER. World Economic Forum. https://www3.weforum.org/docs/WEF_Jobs_of_Tomorrow_Generative_AI_2023.pdf