Article:

Thelwall, M. orcid.org/0000-0001-6065-205X (2025) ChatGPT for complex text evaluation tasks. Journal of the Association for Information Science and Technology, 76 (4). pp. 645-648. ISSN 2330-1635

https://doi.org/10.1002/asi.24966

Received: 6 October 2024 | Revised: 4 November 2024 | Accepted: 4 November 2024

DOI: 10.1002/asi.24966

BRIEF COMMUNICATION

JASIST WILEY

ChatGPT for complex text evaluation tasks

Mike Thelwall

Information School, University of Sheffield, Sheffield, UK

Correspondence
Mike Thelwall, Information School, University of Sheffield, Sheffield, UK.
Email: m.a.thelwall@sheffield.ac.uk

Abstract

ChatGPT and other large language models (LLMs) have been successful at natural and computer language processing tasks with varying degrees of complexity. This brief communication summarizes the lessons learned from a series of investigations into its use for the complex text analysis task of research quality evaluation. In summary, ChatGPT is very good at understanding and carrying out complex text processing tasks in the sense of producing plausible responses with minimum input from the researcher. Nevertheless, its outputs require systematic testing to assess their value because they can be misleading. In contrast to simple tasks, the outputs from complex tasks are highly varied and better results can be obtained by repeating the prompts multiple times in different sessions and averaging the ChatGPT outputs. Varying ChatGPT's configuration parameters from their defaults does not seem to be useful, except for the length of the output requested.

1 | INTRODUCTION

This brief communication reports some lessons learned from applying ChatGPT to complex text processing research projects, some simpler text processing projects, and as a research assistant. Large language models (LLMs) seem to be uniquely flexible in their capabilities, and now challenge or outperform bespoke artificial intelligence (AI) solutions for tasks like grammar correction, text summarization, and translation. As a result of this flexibility and the opaque nature of their operation (their algorithms are broadly known but are too large and complex for outputs to be meaningfully traced back to inputs), insights about how to use them effectively in a particular context may be obscured by the variety of other contexts in which they are used.

This article focuses on one type of task: complex text evaluation in the sense of tasks requiring at least a paragraph of system instructions and that are applied exclusively or primarily to text, with the objective of performing an evaluation. Expert or peer-review summative evaluations of academic documents fit this definition because they require system instructions that explain the review criteria. In addition, the documents that they are applied to would usually be primarily text, perhaps with a few figures. More generally, evaluations of defined aspects of quality for any type of report would be a similar task. In contrast, sentiment analysis is less complex because a LLM should “understand” sentiment well enough for a short prompt like the following to work “Does the following sentence contain positive, negative, or neutral sentiment?” From a different perspective, text summarization is not in scope because it is not evaluative.

2 | HOW ChatGPT WORKS

ChatGPT uses LLM technology, which means that it has ingested huge amounts of text in a way that enables it to predict likely future text from any given prompt. It has ingested enough text to be able to respond to questions rather than just guessing future words in a sentence or paragraph. It has also been through additional training by being fed with tasks or questions and then getting feedback on its responses (Ouyang et al., 2022). It is probably paired with a wide range of other technologies to present its results and perhaps also to deal with specialist types of task, such as programming and image generation (DALL-E).

3 | TASK DESCRIPTION AND OUTPUT PLAUSIBILITY

I have used ChatGPT for various complex document evaluation tasks, including peer-review evaluations of conference, journal, and platform submissions (Thelwall & Yaghi, 2024a), post-publication expert review evaluations of published journal articles (Thelwall & Yaghi, 2024b) and evaluations of Impact Case Study (ICS) documents describing the tangible societal benefits accrued from academic research (Kousha & Thelwall, 2024). In all cases, ChatGPT was fed several paragraphs of instructions, either as part of the customization setup of a Custom GPT in the web interface or as part of the system instructions for an Applications Programming Interface (API) request. These seem to achieve the same eventual goal, but the API is essential for submitting large numbers of requests.

In all cases, the system instructions given to ChatGPT were lightly adapted from the instructions for the human reviewers, with the main change being the use of the pronoun "you" for the ChatGPT instructions (imitating examples on the ChatGPT website), replacing the indirect style of the original instructions (e.g., starting the ChatGPT system instructions with, "You are an expert at evaluating academic research"). ChatGPT always produced plausible outputs from the first attempt at this and so it seems to be very good at following detailed system instructions. This contrasts with traditional machine learning AI which needs substantial technical expertise to setup to produce appropriate outputs. The following illustrates a short session with a system configuration designed to evaluate academic rigor alone. Most outputs are five times longer than this and include detailed score rationales.

> >>Score this: https://arxiv.org/abs/2408.06752
>
> The article titled "*Evaluating Research Quality with Large Language Models: An Analysis of ChatGPT's Effectiveness with Different Settings and Inputs*" investigates the feasibility of using ChatGPT for evaluating academic research quality. The study examines how various inputs (e.g., full text, abstract, title) affect the LLM's scoring accuracy, finding that abstracts and titles yield the highest correlations with human evaluations. The research highlights the potential but also the limitations of LLMs in academic assessment contexts, with ChatGPT models slightly improving results over time.
>
> **Rating: 3***
>
> This study is internationally excellent but does not achieve the highest level of rigour due to its emerging application of LLMs in evaluation processes.

While the immediate production of plausible outputs from ChatGPT suggests that it understands the instructions in some sense, it does not mean that the output is accurate or even meaningful. For example, it can write plausible outputs about made-up papers from their titles alone, so the prompt, "Score the paper: Mice learn Spanish faster than Japanese" produced a serious evaluation that did not mention the implausibility of the premise, awarding 3 out of 4 for originality. Asking ChatGPT directly, "Can mice learn Spanish?," produced the correct response, "Mice, like most animals, do not have the cognitive ability to learn human languages such as Spanish." This is an extreme example of a meaningless context that will not occur in normal evaluations, but it confirms that ChatGPT does not always apply its relevant knowledge to given tasks.

4 | OUTPUT EVALUATION

As mentioned above, producing a plausible output is not the same as producing a correct output and therefore it is essential to find a mechanism to evaluate the accuracy of the ChatGPT outputs. In the situation where the output includes a summary score or a category then the meaningfulness of the results can be checked against a gold standard of human scores/categories with standard metrics, such as correlation, mean absolute deviation, precision, recall, or F-measure.

5 | PROMPT REPETITION

LLMs are essentially probabilistic models of language. They work by calculating the most likely tokens (words or parts of words) to follow from those already present. Current versions include a randomness or creativity parameter to allow some variation in the token selected at each stage. Thus, asking ChatGPT to suggest a word to end the sentence, “the sky is” might produce “blue” sometimes and “cloudy,” “vast,” or “limitless” at other times but not “be” unless the creativity parameter setting was extremely high.

For simple tasks with short inputs and instructions, such as sentiment analysis, the same result might be given most of the time, especially if there is clearly a correct answer. For complex evaluation tasks, however, the results can vary substantially as the random parameter is repeatedly evoked to create a long answer, leveraging probabilities related to the long instructions. In this case, the output from the same prompt is likely to vary each time, including in length, overall structure, and summary evaluation (if any).

Two important facts follow from the above observation. First, non-systematic experiments with variations in the inputs or instructions are pointless if the aim is to improve the results. This is because the natural system variations make it impossible to know the effect of an individual change from a single test. Instead, systematic larger-scale testing is needed. Second, results that more accurately reflect the underlying LLM probability model can be gained by repeating a prompt (in a separate session if using the web interface, otherwise it learns from its previous result) many times and averaging the results. Two studies have shown that averaging up to 30 repetitions gives much more accurate results than individual tests (Thelwall, 2024a, 2024b).

6 | SYSTEM PROMPT VARIATION

There seems to be only one systematic comparison of the effectiveness of different system prompts for a complex text evaluation task, and this found that shorter versions of the instructions produced worse results (Thelwall, 2024b). This suggests that complex instructions may be manageable by ChatGPT and there is little to be gained by attempting to substantially alter human instructions for ChatGPT.

7 | INPUT VARIATION

In contrast to the situation for system instructions, it seems to be possible to give ChatGPT too much information to evaluate. In a comparison of evaluations where the input was (a) article titles, (b) article titles and abstracts, and (c) article titles, abstracts, and full text (without references and tables), the second option produced the best results (Thelwall, 2024b). While evaluating an article based on its title alone is nonsensical, it is surprising that ChatGPT 4o-mini performed better on titles and abstracts combined than on full text inputs. This suggests that the condensed summary of an abstract provides the key information needed for an evaluation whereas lengthy full texts might overload ChatGPT with too much information that is less relevant to its task. While a human reviewer would presumably benefit from checking the full text for rigor in particular, ChatGPT does not seem to.

The most logical explanation for the above phenomenon seems to be that while ChatGPT can ignore irrelevant information in text, it may perform better with more condensed inputs. The fact that it performs better without the information needed for a proper evaluation also underlines the fact that its plausible outputs are not evaluations but only mimic evaluation with the available information. Thus, even for cases where the input is full text, it should not be assumed that ChatGPT is performing a meaningful evaluation, but only an approximation.

8 | PARAMETER AND MODEL VARIATION

LLMs have some parameters that can be varied in the API, such as for the creativity/probability component. There are also multiple models, including for different issues of each ChatGPT model and mini variants, which are less accurate (and cheaper) versions. These are tricky to compare for complex text evaluation tasks because the size of the datasets evaluated are likely to be small enough that only major improvements in the performance of a ChatGPT model would result in statistically significantly better results. It can also be financially expensive to compare many variations because the medium length system prompts, medium or long documents to analyze, and the need for up to 30 iterations increases the cost of API calls.

Experiments with different models and parameter variations on complex text evaluation tasks so far suggest that the default parameters do not need to be changed, but there are substantial differences between models. In general, the results are consistent with the expectation that newer and more complete models (e.g., 4o rather than 4o-mini) perform better (cf., Saad et al., 2024; Thelwall, 2024b). Nevertheless, the cut down versions of models seem to have accuracy that is close to that of the full models and are much cheaper so are a reasonable practical choice.

9 | FINE TUNING

Fine tuning is the process of producing a customized variant of ChatGPT that has evolved to learn to perform better on a particular task from being fed examples of it. Fine tuning works well for tasks with simple outputs (e.g., a single sentiment score). It does not seem promising for complex text evaluations because the outputs are varied and complex. For example, peer-review reports on the same journal article are never the same, so it is not clear that ChatGPT could meaningfully learn patterns from being fed sets of articles and their peer-review reports, unless there were common errors that reviewers often identified. This is an open question, however.

10 | CONCLUSIONS

This brief communication has attempted to provide some insights into the use of LLMs for complex text evaluation tasks. It is based on limited evidence from a single system, ChatGPT, mainly 4o-mini, and a narrow range of academic tasks. These may serve as a starting point for future research designs.

ORCID

Mike Thelwall https://orcid.org/0000-0001-6065-205X

REFERENCES

Kousha, K., & Thelwall, M. (2024). Assessing the societal influence of academic research with ChatGPT: Impact case study evaluations. arXiv preprint arXiv:2410.19948.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Saad, A., Jenko, N., Ariyaratne, S., Birch, N., Iyengar, K. P., Davies, A. M., Vaishya, R., & Botchu, R. (2024). Exploring the potential of ChatGPT in the peer review process: An observational study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *18*(2), 102946. https://doi.org/10.1016/j.dsx.2024.102946

Thelwall, M. (2024a). Can ChatGPT evaluate research quality? *Journal of Data and Information Science*, *9*(2), 1–21.

Thelwall, M. (2024b). Evaluating research quality with large language models: An analysis of ChatGPT's effectiveness with different settings and inputs. https://arxiv.org/abs/2408.06752

Thelwall, M., & Yaghi, A. (2024a). Evaluating the predictive capacity of ChatGPT for academic peerreview outcomes across multiple platforms. Submitted.

Thelwall, M., & Yaghi, A. (2024b). In which fields can ChatGPT detect journal article quality? An evaluation of REF2021 results. https://arxiv.org/abs/2409.16695

How to cite this article: Thelwall, M. (2024). ChatGPT for complex text evaluation tasks. *Journal of the Association for Information Science and Technology*, 1–4. https://doi.org/10.1002/asi.24966