A day in the life of ChatGPT as an academic reviewer: Investigating the potential of large language model for scientific literature review

Mashrin Srivastava
Microsoft Corporation
masrivastava@microsoft.com

Abstract

In this paper, we investigate the potential of large language model for scientific literature review. The exponential growth of research papers is placing an increasing burden on human reviewers, making it challenging to maintain efficient and reliable review processes. To address this, we explore the use of ChatGPT to assist with the review process. Our experiments demonstrate that ChatGPT can review the papers followed by the sentiment analysis of the review of research papers and provide insights into their potential for acceptance or rejection. Although our study is limited to a small sample of papers, the results are promising and suggest that further research in this area is warranted. We note that the use of large language models for scientific literature review is still in its early stages, and there are many challenges to be addressed. Nonetheless, our work highlights the potential of these models to augment the traditional peer-review process, providing a new perspective on research papers and potentially accelerating the pace of scientific discovery and innovation.

1 Introduction

The process of reviewing research papers is crucial for the advancement of scientific knowledge and the dissemination of new ideas. However, with the growing volume of research being published, it is becoming increasingly challenging for human reviewers to keep up with the demand for review. This has led to the exploration of alternative approaches to assist or even replace human reviewers, including the use of artificial intelligence (AI) and natural language processing (NLP).

In this paper, we focus on the use of large language models for reviewing research papers. Large language models are a type of AI that have been trained on vast amounts of text data and can perform a wide range of NLP tasks. These models have achieved impressive results on a variety of benchmarks and have the potential to be useful for scientific literature review.

The motivation for exploring this topic is twofold. First, the growing volume of research papers means that there is a need for more efficient and reliable review processes. Second, the use of large language models for scientific literature review has the potential to provide a new perspective and identify patterns and trends that may not be immediately apparent to human reviewers.

2 Literature review

The use of AI models to review research articles has gained significant attention in recent years due to the increasing volume of scholarly publications and the need to streamline the review process. AI models can assist in identifying the most relevant and impactful articles, evaluate the quality of research, and predict future research trends. Several studies have explored the use of AI for this task.

In [Mrowinski et al., 2017], the authors present a study aimed at improving the efficiency of the peer review process, which is often slow and time-consuming. The study was conducted with the help of the Journal of the Serbian Chemical Society's dataset, which allowed the creation of an artificial review thread simulation. The researchers then used Cartesian Genetic Programming, an evolutionary algorithm, to search for a strategy that would decrease review time. According to the findings, the implementation of the evolved strategy resulted in a 30% reduction in the duration of the peer review process, all while keeping the same pool of reviewers. The researchers were able to show that genetic programs have the capability to enhance actual social systems and that their approach could lead to a measurable advancement in the efficiency of the peer review process.

In [Ghosal et al., 2019], the authors propose a deep neural architecture that uses sentiment information from peer review texts to predict the outcome of the review process. Their system achieves a significant improvement over existing baselines and could serve as an additional layer of confidence for editors and program chairs, especially when reviewers are non-responsive or missing.

In [Checco et al., 2021], the authors examines the potential of using artificial intelligence (AI) to automate or assist in the peer-review process. The researchers developed a machine-learning system and trained it with 3300 papers from three conferences, showing that the system can successfully predict the outcome of the peer review process based on superficial features of the manuscript. The study highlights the potential benefits of such AI tools, including greater efficiency and insights into the reviewing process. However, the researchers also note the need to address potential biases and ethical concerns associated with these tools.

In [pee, 2022], the authors aimed to evaluate the utility of various machine learning algorithms in predicting publication likelihood based on peer reviewer scores. A cross-sectional study design was employed, and a sample of 263 manuscripts undergoing peer review between 2017 and 2021 were selected, with the final decision on acceptance or rejection used as the outcome variable. The performance of different algorithms was assessed using both training and testing data and metrics such as accuracy. Results suggest that, while the performance of the machine learning algorithms varied, in general they performed only moderately well in predicting publication likelihood. The highest accuracy achieved was 65.2%.

Several software programs now use AI for reviewing articles [Heaven, 2018]. In this article, the author mentions the software software like StatReviewer, which checks manuscript statistics and methods, has been adopted by major publishing companies like Elsevier. Additionally, ScholarOne, a commonly used peer-review platform, is collaborating with UNSILO, a company that uses natural language processing and machine learning to analyze manuscripts. Another tool, statcheck [Nuijten et al., 2016], focus on specific areas of review, such as assessing the consistency of authors' statistics and reports. These programs are being implemented in various publishing and review processes and are garnering interest from other publishers.

However, there are also several limitations and challenges to using AI for scientific literature review. In [Vincent-Lamarre and Larivière, 2021], the authors analyzed a dataset of scientific manuscripts submitted to various artificial intelligence conferences and compared the linguistic characteristics of accepted and rejected manuscripts. They found that accepted manuscripts were less readable, contained more scientific and artificial intelligence jargon, and used more abstract and less common words than rejected manuscripts. Additionally, the authors found that accepted manuscripts were more likely to cite the same publications and had more semantic similarity. The study's results suggest a possible content bias in the peer review process, with machine learning and neural network-related topics being associated with greater acceptance rates. Another challenge is that AI systems may not fully understand the technical and domain-specific terms and concepts used in research papers, which can affect their ability to accurately extract key information and evaluate the quality and novelty of the research. In addition, scientific papers often contain complex structures and arguments that may be difficult for AI systems to interpret. There is also the question of bias in the training data of AI systems, which could potentially affect the system's performance and its ability to evaluate research papers objectively. Finally, there may be ethical concerns around replacing human reviewers with AI, especially if the systems are not transparent in their decision-making process. Despite these challenges, the use of AI for scientific literature review is an active area of research, and there have been some promising initial results. However, there is still a need for further studies to evaluate the performance and limitations of AI systems in this context, and to consider the ethical and societal implications of their use.

In this paper, we contribute to this literature by evaluating the performance of large language models, a type of AI that has achieved impressive results on a variety of benchmarks, in reviewing research papers. Our analysis aims to identify the strengths and weaknesses of these models in this task and to identify directions for future work.

3 Methodology

The methodology for this study involves a two-part process. The first step involves passing the content of the paper through ChatGPT, using the prompt: "As an academic reviewer, review a new paper. I will add the sections of this paper below in multiple messages." The contents of the paper are then added during the course of several messages. It is then followed by a second prompt: "Now that you have all the sections for this paper, act as an academic reviewer and mention the Summary Of The Paper, Strength And Weaknesses, Clarity, Quality, Novelty And Reproducibility and the Summary Of The Review." This prompts ChatGPT to give a comprehensive review of the paper. In the second step, the ChatGPT output (review of the paper) is passed to ChatGPT for sentiment analysis. The sentiment analysis output can then be utilized in the decision of whether to accept or reject the paper. It is important to note that acceptance of a paper also depends on the selectivity of the venue. The primary focus of this study is to determine the quality of the paper.

In this study, a total of eleven papers were selected for analysis. The review comments of the first four papers are publicly available on OpenReview. The review for these papers were made available after the data cut-off date for ChatGPT. For the fifth paper, it was generated by a language model (ChatGPT) and no official reviews were available. The review comments for the next five papers were not publicly available but are known to the author for some of the papers. Lastly, the last paper is still under review and the review comments are not yet publicly accessible. They will, however, be made available on OpenReview at a later date.

4 Results

4.1 Review of papers with OpenReview available

4.1.1 LAMP: Extracting Text from Gradients with Language Model Priors [Balunović et al., 2022]

Summary of the Paper: The paper presents LAMP, a method for reconstructing private text data from gradients using language model priors and alternating between continuous and discrete optimization. The authors extensively evaluate LAMP on various datasets, architectures, and batch sizes, and show that it consistently outperforms prior work in text reconstruction from gradients. The authors also demonstrate that LAMP is effective in reconstructing private data in challenging settings such as fine-tuned models and noise-defended gradients.

Strengths: The paper presents a new method for text reconstruction from gradients that leverages language model priors and achieves strong results in a wide range of settings. The authors provide a thorough experimental evaluation of LAMP, including comparisons with prior work and ablation studies. The paper discusses the limitations of current federated learning algorithms in protecting private text data and the need for further research in this area.

Weaknesses: The paper assumes knowledge of the ground truth labels for the private text data, which may not always be possible in practice. The effectiveness of LAMP in reconstructing private text data in the masking setting, as considered in prior work, is not evaluated.

Clarity: The paper is well-written and easy to understand. The authors provide clear explanations of the LAMP method and its components, as well as a detailed description of the experimental setup and results.

Quality: The paper presents a novel and effective method for text reconstruction from gradients, and provides a thorough evaluation of its performance. The experimental results are presented in a clear and comprehensive manner, and the limitations of the method are discussed.

Novelty: LAMP is a novel method for text reconstruction from gradients that leverages language model priors and alternating optimization. To the best of my knowledge, this approach has not been previously proposed in the literature.

Reproducibility: The authors provide a detailed description of the experimental setup and results, making it possible to reproduce the experiments. However, the code for the LAMP method and the experimental evaluation is not publicly available.

Summary of the Review: Overall, the paper presents a strong contribution in the area of text reconstruction from gradients, introducing a novel method that achieves strong results in a wide range of settings. The experimental evaluation is thorough and the results are clearly presented. However, the assumption of knowledge of ground truth labels and the lack of code availability may be limitations.

4.1.2 Masked Prediction: A Parameter Identifiability View [Liu et al.]

Summary of the Paper: The paper studies the problem of parameter identifiability in the context of self-supervised learning (SSL) using masked prediction tasks. Specifically, the authors consider two types of data generative models - hidden Markov models (HMMs) and conditionally-Gaussian hidden Markov models (G-HMMs) - and analyze the identifiability of their parameters under various masked prediction tasks. The authors show that the identifiability of the parameters depends on the combination of the data generative model and the prediction task, and provide necessary and sufficient conditions for identifiability in both cases.

Strengths: One strength of the paper is that it provides a novel perspective on SSL by studying the problem of parameter identifiability. The authors also provide a thorough analysis of the identifiability of the parameters of HMMs and G-HMMs under various masked prediction tasks, and provide necessary and sufficient conditions for identifiability in both cases.

Weaknesses: The paper assumes that the hidden state space is discrete and of a fixed size, which may limit the applicability of the results to real-world data.

Clarity: The paper is well-written and easy to follow. The authors provide clear definitions and assumptions, and the main results are presented in a clear and concise manner.

Quality: The results in the paper are well-motivated and the authors provide a thorough analysis of the identifiability of the parameters of HMMs and G-HMMs under various masked prediction tasks. The authors also provide necessary and sufficient conditions for identifiability in both cases, which adds to the quality of the paper.

Novelty and Reproducibility: To the best of my knowledge, the problem of parameter identifiability in the context of SSL using masked prediction tasks has not been studied before. The results in the paper are novel and provide a new perspective on SSL. The authors also provide sufficient details and examples to make the results reproducible.

Summary of the Review: Overall, the paper is a well-written and well-motivated study of the problem of parameter identifiability in the context of SSL using masked prediction tasks. The authors provide a thorough analysis of the identifiability of the parameters of HMMs and G-HMMs under various masked prediction tasks, and provide necessary and sufficient conditions for identifiability in both cases. The results are novel and provide a new perspective on SSL. However, the assumption that the hidden state space is discrete and of a fixed size may limit the applicability of the results to real-world data.

4.1.3 Evident: a Development Methodology and a Knowledge Base Topology for Data Mining, Machine Learning and General Knowledge Management [Haidar et al., 2022]

Summary of the Paper: This paper presents a new project development and artifact management methodology called Evident, which is intended to address various pain points in the fields of data mining (DM), machine learning (ML) and knowledge management (KM). The methodology is based on the idea of breaking down projects into two granular levels: Knowledges and Containers, which are essentially collections of observations, hypotheses and tests. The paper describes how Evident can be used to represent different types of knowledge (inductive, abductive and deductive), and how it

can be applied to DM, ML and KM in order to improve the reproducibility, measurability, trackability and scalability of these fields.

Strengths: One of the strengths of this paper is that it presents a clear and concise overview of the Evident methodology, including its definitions, scope, and application to DM, ML and KM. The paper also provides a number of examples to illustrate how Evident can be used to address specific pain points in these fields.

Weaknesses: One potential weakness of the paper is that it does not provide a detailed evaluation of Evident in terms of its effectiveness or practicality. While the authors do provide some arguments for why Evident may be useful, it would be helpful to see more concrete evidence of its success or failure in real-world projects.

Clarity: Overall, the paper is well-written and easy to follow. The authors do a good job of explaining the key concepts and ideas behind Evident, and the examples provided are helpful in illustrating these concepts.

Quality: The quality of the paper is generally good, with a clear and well-organized structure. The ideas presented are well-supported by examples and the writing is clear and concise.

Novelty and Reproducibility: The Evident methodology described in this paper appears to be novel, as it is not mentioned in other literature on DM, ML and KM. It is not clear, however, how easily the methodology can be reproduced in different contexts or how widely it has been adopted in practice.

Summary of the Review: This paper presents a new methodology called Evident, which is intended to address various pain points in DM, ML and KM. The methodology is based on the idea of breaking down projects into Knowledges and Containers, and is described in detail in the paper. While the paper provides a clear and concise overview of Evident, more information about its effectiveness or practicality in real-world projects would be helpful. The paper is well-written and easy to follow, and the ideas presented are well-supported by examples. The Evident methodology appears to be novel, but it is not clear how easily it can be reproduced. More details and examples of how Evident can be applied in specific projects would be helpful in understanding the practicalities and potential benefits of this methodology. Additionally, a more thorough comparison to existing project development and artifact management approaches, such as agile methodologies, would be useful in evaluating the novelty and potential impact of Evident. Overall, the quality of the paper is good, but more clarification and elaboration on the proposed methodology and its potential applications would be beneficial.

4.1.4 Understanding Why Generalized Reweighting Does Not Improve Over ERM [Zhai et al., 2022]

Summary of the Paper: The paper studies the performance of generalized reweighting (GRW) algorithms, a class of methods that aim to address distributional shift in machine learning tasks. The authors first prove that for linear models and sufficiently wide fully-connected neural networks, the implicit bias of GRW is equivalent to empirical risk minimization (ERM) when trained for an infinitely long time, and that regularization must be large enough to significantly lower the training performance in order to affect this implicit bias. The authors then show that for classification tasks using the logistic loss, any GRW algorithm satisfying a certain assumption has the same implicit bias as ERM, and that any GRW algorithm satisfying a different assumption has its training error converge to 0 and its model weight direction converge, leading to the same implicit bias as ERM. The authors also discuss promising directions for improving the distributionally robust generalization of GRW algorithms, including data augmentation and pretraining, going beyond sample reweighting based GRW algorithms, and changing the loss function to be better suited for GRW.

Strengths: One strength of the paper is its thorough analysis of the implicit bias of GRW algorithms and its comparison to ERM. The authors provide clear and rigorous proofs of their results and discuss their limitations in detail.

Weaknesses: A weakness of the paper is its reliance on strong assumptions, including the use of linear models and sufficiently wide fully-connected neural networks and the assumption that the models are trained for an infinitely long time.

Clarity: The paper is well-written and easy to follow. The authors clearly state their assumptions and provide detailed explanations and proofs of their results.

Quality: The results of the paper are sound and well-supported by rigorous proof. The authors also provide meaningful insights into the limitations and future directions for GRW algorithms.

Novelty and Reproducibility: The paper presents new theoretical results on the implicit bias of GRW algorithms and its comparison to ERM. The authors also provide detailed explanations and proofs of their results, making the paper highly reproducible.

Summary of the Review: This is a strong paper that provides valuable insights into the performance of GRW algorithms in addressing distributional shift. The results are well-supported by rigorous proof and the paper is well-written and easy to follow. The limitations and future directions discussed by the authors are also valuable contributions to the field. Overall, this is a high-quality and novel paper that should be of interest to researchers in the field of machine learning and distributional shift.

4.2 Review of a paper written by ChatGPT

4.2.1 A Day in the Life of ChatGPT as a researcher: Sustainable and Efficient Machine Learning - A Review of Sparsity Techniques and Future Research Directions [Mashrin Srivastava, 2023]

Summary of the paper: This paper provides an overview of the use of sparsity in machine learning, including techniques for inducing sparsity, sparse training algorithms, and hardware support for sparse training. The paper also discusses the challenges and limitations of using sparsity in large-scale neural networks and proposes several directions for future research in this area.

Strengths: One strength of the paper is that it provides a comprehensive overview of the use of sparsity in machine learning, covering a wide range of techniques and domains. The paper also does a good job of highlighting the challenges and limitations of using sparsity in large-scale neural networks, as well as proposing several directions for future research.

Weaknesses: One weakness of the paper is that it does not provide in-depth coverage of any specific sparsity techniques or applications. While the paper covers a wide range of topics, it does not delve into the details of any specific techniques or provide any experimental results.

Clarity: The paper is generally well-written and easy to understand. The organization and structure of the paper are clear, and the language used is appropriate for the intended audience.

Quality: Overall, the quality of the paper is good. The paper provides a comprehensive overview of the use of sparsity in machine learning and covers a wide range of topics. However, as mentioned above, the paper does not provide in-depth coverage of any specific sparsity techniques or applications and does not include any experimental results.

Novelty and reproducibility: The paper does not present any new research findings or contributions. Instead, it provides an overview of existing techniques and challenges in the use of sparsity in machine learning. As such, the paper's novelty is limited. However, the paper is well-referenced and includes a wide range of relevant literature, which makes it potentially useful as a resource for researchers interested in the topic.

Summary of the review: Overall, this paper provides a comprehensive overview of the use of sparsity in machine learning. It covers a wide range of topics and does a good job of highlighting the challenges and limitations of using sparsity in large-scale neural networks. While the paper is well-written and easy to understand, it does not provide in-depth coverage of any specific sparsity techniques or applications and does not present any new research findings or contributions. As such, the paper's novelty is limited, but it may be useful as a resource for researchers interested in the topic.

4.3 Review of other papers

4.3.1 Clinico-genomic analysis reveals mutations associated with COVID-19 disease severity: possible modulation by RNA structure [Mehta et al., 2021]

Summary of the paper: This study aimed to investigate the relationship between the genetic mutations of SARS-CoV-2 and the severity of COVID-19 in a group of 196 patients in India. The patients were divided into recovered (R) and deceased (D) groups, and the R group was further divided into mild, moderate, and severe categories based on the extent of respiratory support required. The researchers identified three dominant clades of SARS-CoV-2 in the patient samples and conducted

an association study between the presence of certain mutations and disease severity and mortality. They found that the A26194T, C28854T, and C25611A mutations were significantly associated with severe and fatal cases. The researchers then conducted structural analysis on these three mutations and found that they caused changes in the viral proteins and RNA that could potentially affect viral replication and host immune response.

Strengths: One strength of this study is the large sample size of 196 patients, which allows for more robust conclusions to be drawn. Additionally, the use of both Oxford Nanopore and Illumina-Miseq sequencing platforms increases the reliability of the genomic data.

Weaknesses: However, one potential weakness is that the study only includes patients from one hospital in India, which may not be representative of the entire population. It would have been interesting to see if the results of this study are consistent with those from other studies conducted in different countries.

Clarity: The paper is generally well-written and easy to follow. The materials and methods section is detailed and provides sufficient information for the study to be reproduced. However, the introduction could benefit from more context and a clear statement of the research question.

Quality: The study appears to be well-conducted with appropriate controls and statistical analyses. The use of multiple sequencing platforms and the large sample size add to the quality of the study. However, as mentioned above, the limited geographic location of the study may limit the generalizability of the results.

Novelty and reproducibility: The study provides new insights into the relationship between SARS-CoV-2 mutations and COVID-19 severity, and the structural analysis of selected mutations adds to the understanding of how these mutations may affect viral function. The materials and methods section provides sufficient detail for the study to be reproduced.

Summary of the review: Overall, this is a well-conducted study that provides new insights into the relationship between SARS-CoV-2 mutations and COVID-19 severity. The large sample size and use of multiple sequencing platforms add to the reliability of the results. However, the limited geographic location of the study may limit the generalizability of the findings. Further research in other populations would be useful to confirm the results of this study.

4.3.2 COVID-19 Risk Stratification and Mortality Prediction in Hospitalized Indian Patients: Harnessing clinical data for public health benefits [Alle et al., 2022]

Summary of the paper: This paper presents the results of a study that analyzed the clinical data of 544 COVID-19 patients in India in order to develop artificial intelligence (AI) based solutions for risk stratification and mortality prediction. The authors used various machine learning algorithms to analyze the data and found that existing mortality prediction models trained on a different cohort (Wuhan, China) performed poorly when applied to the Indian cohort. They also found that the best performing model for risk stratification on the Indian data was an XGboost classifier, while the best performing model for mortality prediction was a logistic regression model. The authors found that hematological parameters, such as blood coagulation parameters, immune and inflammation parameters, and neutrophil and lymphocyte counts, were important features for both risk stratification and mortality prediction models. The authors also observed that steroids might have played a role in patient survival with extreme neutrophils or lymphocytes.

Strengths: The study is the first to develop predictive models for COVID-19 outcomes in an Indian cohort. The use of machine learning algorithms allows for fast and accurate predictions. The study includes a large number of patients (544) and a wide range of parameters (70 unique features).

Weaknesses: The study only includes patients from a single hospital in New Delhi, which may limit the generalizability of the findings to other regions or populations. The study only includes patients who were admitted to the hospital, which may not accurately represent the overall COVID-19 population in India. The performance of the models could have been further improved by using more advanced machine learning techniques, such as deep learning. Clarity: The paper is well written and easy to understand. The methods, results, and discussion sections are clearly presented and the figures and tables help in understanding the results.

Quality: The study uses a large dataset and employs appropriate statistical methods for the analysis. The performance of the developed models is also evaluated using appropriate metrics.

Novelty and reproducibility: The paper appears to be well written and the methodology used is clear and easy to understand. The use of machine learning algorithms for risk stratification and mortality prediction in COVID-19 patients is a novel approach, and the results obtained in this study are quite promising. The use of data from Indian COVID-19 patients is also a unique aspect of this study, as previous studies have primarily focused on data from other countries. The results of this study are reproducible, as the authors have provided detailed information about the data collection, preprocessing, and modeling steps.

Summary Of The Review Overall, this study provides valuable insights into the clinical characteristics of COVID-19 patients in India and the potential use of machine learning for risk stratification and mortality prediction. Further research is needed to validate these findings in larger and more diverse patient populations.

4.3.3 Automated emergency paramedical response system [Srivastava et al., 2018]

Summary of the paper: The paper presents a telemedicine system called AEPRS (Airborne Emergency and Primary Response System) that uses drones for medical delivery, a chatbot for disease detection, and a private blockchain for medical records. The system also includes a module for stroke detection in community hospitals using image processing and machine learning techniques, and a face recognition module using edge AI.

Strengths: The use of drones for medical delivery and a chatbot for disease detection are innovative ideas that have the potential to improve the efficiency and accessibility of healthcare services, particularly in remote or underserved areas. The inclusion of a private blockchain for medical records addresses the issue of secure and controlled access to patient data. The stroke detection module and the face recognition module demonstrate the potential applications of machine learning and AI in healthcare.

Weaknesses: The paper lacks sufficient details on the implementation and performance of the different modules of the AEPRS system. For example, it is not clear how the chatbot works, what kind of diseases it can detect, and how accurate it is. The paper also lacks a thorough discussion on the potential challenges and limitations of using drones for medical delivery, such as regulatory issues, safety, and security.

Clarity: The paper is generally well-written and easy to follow. However, some of the technical details and descriptions of the different modules could be clearer and more comprehensive.

Quality: The paper lacks a proper evaluation of the performance and effectiveness of the AEPRS system. It only provides some preliminary results for the stroke detection module, but does not provide any quantitative or qualitative analysis of the results.

Novelty: The idea of using drones for medical delivery and a chatbot for disease detection is novel, but the paper does not provide enough details or comparisons to assess the novelty and impact of the AEPRS system.

Reproducibility: It is not clear how the AEPRS system was developed and tested, and whether it can be reproduced by other researchers or practitioners.

Summary of the review: Overall, the paper presents some interesting and potentially useful ideas for using technology to improve healthcare services. However, it lacks sufficient details and evaluation of the different modules of the AEPRS system, and does not address some of the potential challenges and limitations of the proposed solutions.

4.3.4 Sentiment Analysis: Predicting Yelp Scores [Guda et al., 2022]

Summary of the Paper: The paper presents an analysis of several machine learning and deep learning models for sentiment analysis of restaurant reviews. The authors consider both binary (positive vs negative sentiment) and multi-class (1-5 star ratings) classification tasks, and experiment with various input feature sets (meta features, review text, and both). They find that deep learning models, particularly those with attention mechanisms, outperform the other models, and that the joint use of meta features and review text is beneficial for all models.

Strengths: A strength of the paper is that it thoroughly investigates a range of models and input feature sets, and presents clear results and analysis.

Weaknesses: A weakness is that the analysis of model interpretability is limited to qualitative examples rather than more systematic or quantitative analysis.

Clarity: The paper is well-written and easy to follow. The methods and results are described in sufficient detail.

Quality: The research is well-conducted and the results are sound. The use of multiple datasets and evaluation metrics adds to the robustness of the study.

Novelty and Reproducibility: The paper presents a thorough comparison of different models for sentiment analysis of restaurant reviews, which is a common problem in the field. The results of the study are potentially useful for practitioners looking to choose a model for this task. The authors provide sufficient detail about the datasets, preprocessing, and experimental setup to allow for reproduction of the study.

Summary of the Review: Overall, the paper presents a thorough and well-conducted analysis of different models for sentiment analysis of restaurant reviews. The results and analysis are clearly presented and the study is well-written and easy to follow. The use of multiple datasets and evaluation metrics adds to the robustness of the study. The paper is a useful resource for practitioners looking to choose a model for this task. One potential area for improvement would be to include more systematic or quantitative analysis of model interpretability.

4.3.5 Smart City: An Intelligent Automated Mode of Transport Using Shortest Time of Travel Using Big Data [Srivastava et al., 2022]

Summary of the paper: The paper proposes the use of a modified Dijkstra's algorithm for autonomous transportation in order to minimize travel time rather than distance. The approach incorporates ant colony optimization and makes use of big data analytics and cloud computing. The performance of the proposed approach is evaluated by comparing the time saved with the original Dijkstra's algorithm for different map types.

Strengths: The problem of minimizing travel time in autonomous transportation is important and relevant. The proposed approach of using a modified Dijkstra's algorithm and incorporating ant colony optimization is novel.

Weaknesses: The technical details of the proposed approach and the modifications made to the Dijkstra's algorithm are not clearly explained. The data and settings used in the analysis are not sufficiently described. The results are not thoroughly analyzed or discussed.

Clarity: The paper could benefit from improved organization and clarity, particularly in the description of the proposed approach and its components. There are some typos and grammatical errors that could be corrected.

Quality: More information is needed on the technical details and methodology of the proposed approach in order to fully assess its quality. The evaluation of the results could be more thorough and include statistical analysis.

Novelty and reproducibility: The proposed approach of using a modified Dijkstra's algorithm and incorporating ant colony optimization appears to be novel. It is not clear how the work could be reproduced, as the technical details and methodology are not sufficiently described.

Summary of the review: The paper presents a novel approach for autonomous transportation using a modified Dijkstra's algorithm and ant colony optimization, but the technical details and methodology are not clearly explained. The data and settings used in the analysis are also not sufficiently described. The results are not thoroughly analyzed or discussed. The paper could benefit from improved organization and clarity, and more information is needed in order to fully assess the quality and reproducibility of the work.

4.4 Review of paper with pending official review

4.4.1 Predicting COVID-19 case status from self-reported symptoms and behaviors using data from a massive online survey [Srivastava et al., 2023]

Summary of the paper: This paper presents a machine learning (ML) model for predicting COVID-19 case status using data from the US COVID-19 Trends and Impact Survey (CTIS). The CTIS data

Paper Section	ChatGPT	Actual Decision	Venue
4.1.1	Mainly positive	Accept (Certain)	NeurIPS
4.1.2	Mainly positive	Accept (Certain)	NeurIPS
4.1.3	Neutral	Reject (Certain)	NeurIPS
4.1.4	Mainly positive	Reject (Less certain)*	NeurIPS
4.2.1	Neutral	N/A	Preprint
4.3.1	Mainly positive	Accept	MDPI Pathogens
4.3.2	Mainly positive	Accept	PLoS One
4.3.3	Neutral with a slightly negative tone	Accept	Springer HISS
4.3.4	Neutral	N/A	Preprint
4.3.5	Mainly negative	Accept	Springer Nature
4.4.1	Mainly positive	N/A	Under review

Table 1: Result for sentiment analysis of reviews

includes information on symptoms, testing, demographics, behaviors, and vaccination status. The ML model was trained on 12.41 million responses from the survey and achieved an F1 score of 94

Strength and Weaknesses: The strength of the model lies in its ability to accurately predict COVID-19 case status using data from a large survey. The model was able to achieve a high accuracy of 94%, which is an improvement on existing models. The weakness of the model is that it is limited by the quality and availability of data on symptoms, which can vary over time and location.

Clarity: The paper is well written and easy to understand. The authors clearly explain their methodology and the results of their analysis.

Quality: The methods and results presented in this paper are of a high quality. The authors have conducted a thorough analysis of the data and have carefully evaluated the performance of their models.

Novelty and Reproducibility: This paper presents a novel approach to predicting COVID-19 case status using survey data. The methods used in this paper are clearly explained and the authors have provided detailed instructions for replicating their results.

Summary of the Review: This paper presents a machine learning model for predicting COVID-19 case status using data from the US COVID-19 Trends and Impact Survey. The model was able to achieve an F1 score of 94%, which is an improvement on existing models. The paper is well written and of a high quality, and the methods and results presented are novel and reproducible.

5 Discussion

It is observed that all papers, namely 4.1.1, 4.1.2, 4.1.4, 4.3.1, and 4.3.2, where ChatGPT's sentiment analysis outputs mainly positive results for its own review, have been accepted at prestigious conferences or journals. It is worth noting that 4.1.4 was initially rejected from NeurIPS with less certainty but was later accepted to ICLR, another top venue. Hence, the prediction of ChatGPTs is inline with the quality of the article. Also, please note that the paper 4.4.1 is currently under review, although ChatGPT's sentiment analysis output for its review of that paper is also mainly positive. In addition, three papers, specifically 4.1.3, 4.2.1, and 4.3.3, have received a neutral sentiment analysis output from ChatGPT. Among them, the first paper was rejected from a prominent conference, the second paper did not undergo peer review, and the third paper was accepted for publication in a special issue of a journal. One paper, identified by the reference number 4.3.5, was evaluated by ChatGPT as having a predominantly negative sentiment. Despite this, the paper was accepted for publication as a book chapter. It is worth noting, however, that the paper had been rejected by several other venues, suggesting that ChatGPT's evaluation of the paper may indeed be accurate.

6 Conclusion

In this paper, we have explored the potential of large language models for assisting with the process of scientific literature review. Our analysis shows that there is promise in using AI techniques to augment the traditional peer-review process. Specifically, our experiments with ChatGPT have demonstrated

that it is possible to use large language models to assess the sentiment of research papers and provide insights into their potential for acceptance or rejection. Our study indicates that utilizing ChatGPT's review of a paper, followed by a sentiment analysis of the review, can aid in predicting whether a paper will be accepted or rejected for publication. Specifically, we found that predominantly positive sentiment analysis outputs correspond to higher chances of acceptance, while negative or neutral results suggest a lower likelihood of acceptance. Although our analysis is restricted to a small sample of papers, the outcomes exhibit potential and imply that more research in this field is required.

It is worth noting that the use of large language models for scientific literature review is still in its infancy, and there are many challenges to be addressed. However, our work highlights the potential of these models to help address the growing demand for efficient and reliable review processes, and to provide a new perspective on research papers that may not be immediately apparent to human reviewers. In conclusion, the results of our study suggest that the use of large language models for scientific literature review has the potential to be a valuable tool for researchers and publishers alike. With continued research and development in this area, it is possible that using large language models like ChatGPT may eventually become a standard part of the scientific review process, helping to accelerate the pace of scientific discovery and innovation.

References

Utility of machine learning in predicting success of a peer review paper from peer reviewer scores, Nov 2022. URL https://peerreviewcongress.org/abstract/utility-of-machine-learning-in-predicting-success-of-a-peer-review-paper-from-peer-reviewer-s

Shanmukh Alle, Akshay Kanakan, Samreen Siddiqui, Akshit Garg, Akshaya Karthikeyan, Priyanka Mehta, Neha Mishra, Partha Chattopadhyay, Priti Devi, Swati Waghdhare, et al. Covid-19 risk stratification and mortality prediction in hospitalized indian patients: Harnessing clinical data for public health benefits. *PloS one*, 17(3):e0264785, 2022.

Mislav Balunović, Dimitar I. Dimitrov, Nikola Jovanović, and Martin Vechev. Lamp: Extracting text from gradients with language model priors, 2022. URL https://arxiv.org/abs/2202.08827.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1), January 2021. doi: 10.1057/s41599-020-00703-8. URL https://doi.org/10.1057/s41599-020-00703-8.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. Deepsentipeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1120–1130, 2019.

Bhanu Prakash Reddy Guda, Mashrin Srivastava, and Deep Karkhanis. Sentiment analysis: Predicting yelp scores. *arXiv preprint arXiv:2201.07999*, 2022.

Samer Haidar et al. Evident: a development methodology and a knowledge base topology for data mining, machine learning and general knowledge management. *arXiv preprint arXiv:2211.10291*, 2022.

Douglas Heaven. Ai peer reviewers unleashed to ease publishing grind, Nov 2018. URL https://www.nature.com/articles/d41586-018-07245-9.

Bingbin Liu, Daniel Hsu, Pradeep Kumar Ravikumar, and Andrej Risteski. Masked prediction: A parameter identifiability view. In *Advances in Neural Information Processing Systems*.

Mashrin Srivastava. A day in the life of chatgpt as a researcher: Sustainable and efficient machine learning -a review of sparsity techniques and future research directions. 2023. doi: 10.13140/RG.2.2.34470.60480. URL https://rgdoi.net/10.13140/RG.2.2.34470.60480.

Priyanka Mehta, Shanmukh Alle, Anusha Chaturvedi, Aparna Swaminathan, Sheeba Saifi, Ranjeet Maurya, Partha Chattopadhyay, Priti Devi, Ruchi Chauhan, Akshay Kanakan, et al. Clinico-genomic analysis reveals mutations associated with covid-19 disease severity: possible modulation by rna structure. *Pathogens*, 10(9):1109, 2021.

Maciej J Mrowinski, Piotr Fronczak, Agata Fronczak, Marcel Ausloos, and Olgica Nedic. Artificial intelligence in peer review: How can evolutionary computation support journal editors? *PloS one*, 12(9):e0184711, 2017.

Michèle B Nuijten, Chris HJ Hartgerink, Marcel ALM Van Assen, Sacha Epskamp, and Jelte M Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior research methods*, 48:1205–1226, 2016.

Mashrin Srivastava, Saumya Suvarna, Apoorva Srivastava, and S Bharathiraja. Automated emergency paramedical response system. *Health Information Science and Systems*, 6(1):1–16, 2018.

Mashrin Srivastava, Suvarna Saumya, Maheswari Raja, and Mohana Natarajan. Smart city: An intelligent automated mode of transport using shortest time of travel using big data. In *Frontiers of Data and Knowledge Management for Convergence of ICT, Healthcare, and Telecommunication Services*, pages 45–59. Springer, 2022.

Mashrin Srivastava, Alex Reinhart, and Robin Mejia. Predicting covid-19 case status from self-reported symptoms and behaviors using data from a massive online survey. *medRxiv*, 2023. doi: 10.1101/2023.02.03.23285405. URL `https://www.medrxiv.org/content/early/2023/02/07/2023.02.03.23285405`.

Philippe Vincent-Lamarre and Vincent Larivière. Textual analysis of artificial intelligence manuscripts reveals features associated with peer review outcome. *Quantitative Science Studies*, 2(2):662–677, 2021.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Understanding why generalized reweighting does not improve over erm. *arXiv preprint arXiv:2201.12293*, 2022.