Prognosis of exploration on Chat GPT with artificial intelligence ethics

N. Gowri Vidhya[1], D. Devi[2], Nithya A.[1] & T. Manju[1]

[1] St. Joseph's Institute of Technology, Chennai, India

[2] Sathyabama Institute of Science and Technology, Chennai, India

Correspondence: N. Gowri Vidhya, St. Joseph's Institute of Technology, Chennai, India. E-mail: vighyagowri@gmail.com

Received: April 07, 2023 Accepted: April 26, 2023 Published: September 01, 2023

DOI: 10.14295/bjs.v2i9.372 URL: https://doi.org/10.14295/bjs.v2i9.372

Abstract

Natural language processing innovations in the past few decades have made it feasible to synthesis and comprehend coherent text in a variety of ways, turning theoretical techniques into practical implementations. Both report summarizing software and sectors like content writers have been significantly impacted by the extensive Language-model. A huge language model, however, could show evidence of social prejudice, giving moral as well as environmental hazards from negligence, according to observations. Therefore, it is necessary to develop comprehensive guidelines for responsible LLM (Large Language Models). Despite the fact that numerous empirical investigations show that sophisticated large language models have very few ethical difficulties, there isn't a thorough investigation and consumers study of the legality of present large language model use. We use a qualitative study method on OpenAI's ChatGPT3 to solution-focus the real-world ethical risks in current large language models in order to further guide ongoing efforts on responsibly constructing ethical large language models. We carefully review ChatGPT3 from the four perspectives of bias and robustness. According to our stated opinions, we objectively benchmark ChatGPT3 on a number of sample datasets. In this work, it was found that a substantial fraction of principled problems is not solved by the current benchmarks; therefore, new case examples were provided to support this. Additionally, discussed were the importance of the findings regarding ChatGPT3's AI ethics, potential problems in the future, and helpful design considerations for big language models. This study may provide some guidance for future investigations into and mitigation of the ethical risks offered by technology in large Language Models applications.

Keywords: natural language processing, artificial intelligence, ethics, openAI's chatGPT3, large language models.

Prognóstico de exploração no *Chat GPT* com ética de inteligência artificial

Resumo

As inovações de processamento de linguagem natural nas últimas décadas tornaram possível sintetizar e compreender textos coerentes de várias maneiras, transformando técnicas teóricas em implementações práticas. Ambos relatam que softwares resumidos e setores como criadores de conteúdo foram significativamente afetados pelo extenso modelo de linguagem. Um enorme modelo de linguagem, no entanto, poderia mostrar evidências de preconceito social, dando riscos morais e ambientais por negligência, de acordo com as observações. Portanto, é necessário desenvolver diretrizes abrangentes para LLM (Large Language Models) responsáveis. Apesar do fato de numerosas investigações empíricas mostrarem que modelos sofisticados de linguagem ampla têm muito poucas dificuldades éticas, não há uma investigação completa e estudo de consumidores sobre a legalidade do uso atual de modelos de linguagem ampla. Usamos um método de estudo qualitativo no ChatGPT3 da OpenAI para focar na solução os riscos éticos do mundo real nos atuais modelos de linguagem ampla, a fim de orientar ainda mais os esforços contínuos na construção responsável de modelos éticos de linguagem ampla. Analisamos cuidadosamente o ChatGPT3 a partir das quatro perspectivas de viés e robustez. De acordo com nossas opiniões declaradas, comparamos objetivamente o ChatGPT3 em vários conjuntos de dados de amostra. Neste trabalho, constatou-se que uma fração substancial dos problemas de princípios não é resolvida pelos referenciais atuais;

portanto, novos exemplos de casos foram fornecidos para apoiar isso. Além disso, foram discutidas a importância das descobertas sobre a ética de IA do ChatGPT3, possíveis problemas no futuro e considerações de design úteis para grandes modelos de linguagem. Este estudo pode fornecer algumas orientações para futuras investigações e mitigação dos riscos éticos oferecidos pela tecnologia em grandes aplicações de Modelos de Linguagem.

Palavras-chave: processamento de linguagem natural, inteligência artificial, ética, chatGPT3 da openAI, grandes modelos de linguagem.

1. Introduction

Recent advancements in language processing have demonstrated both its use in the information sector and its potential to improve interpersonal relationships (Kurian et al., 2023). LLMs have been used in a variety of real-world settings, including branding, interpreting software, and browsers. Yet, since there isn't any interaction or conversation, these solutions could not immediately boost client engagement (Sharma; Dash, 2020). Because conversation language model companies like Amazon Echo and Google Home employ natural language as their primary way of communication, they have the potential to have a significant influence on how people behave (Dahmen et al., 2023).

As NLP moves from theory to reality, unforeseen adverse effects on human-machine interaction have also surfaced, even with its possible benefits. Hence, it covers topics such as Microsoft's Tay bot's nasty Twitter posts and Amazon Alexa's privacy infractions (Eliot; Wood, 2021; Burstein et al., 2019; Hirschberg; Manning, 2015). Moreover, it may be difficult to stop language models from unintentionally absorbing bias and toxicity during the unstructured pertaining stage from large, noisy corpora. Although research has found that LLMs can be applied for social benefit the aforementioned vulnerability can be immorally employed for unfair discrimination, automated disinformation, and illegal restriction (Ward; Tashea, 2020).

As a result, extensive research just on AI ethics of LLMs has been done, covering it from spotting immoral conduct to reducing discrimination (Loconte et al., 2023; Hirschberg; Manning, 2015). Despite their disagreement forms the basis for Natural Language processing ethical research, there's no guarantee that these risks will materialize in contemporary linguistic models. According to empirical analyses language models encounter moral dilemmas in a number of subsequent behaviors (Otter et al., 2020). Early research established that conversation lexicons might present ethical issues in exploratory studies, for instance through modeling induction, hostile robustness, and secrecy.

According to the most recent data, LLMs and GPT-3 have a persistent prejudice against particular genders and religions (Hovy; Spruit, 2016). Of course, LLMs could also transmit cytotoxic effects (Dror et al., 2018) that have unethical consequences. The study's research gaps remain unaddressed existing studies on NLP as well as ethical concerns and effects:

- Application: As many researches on AI ethics have indeed been conceptually undertaken, they might not adequately reflect the ethical challenges that are actually present in the real world.
- Timeframe: Given the rapid growth of NLP, there hasn't been much ethical analysis of more contemporary language models.
- Consensus: Well about ethical risks connected with the existing sophisticated language modelling techniques, there is a difference of opinion among regular users.
- Consolidation: The vast majority of research fails to adequately tackle all ethical factors because they only focus on the measuring of particular ethical difficulties.

In this paper, we provide a thorough qualitative method and collection of ethical problems in Chat GPT3, a newly released pragmatic language model from Open AI, in order to overcome these inadequacies. Chat GPT is one of the most complete publicly accessible pragmatic language models; it is also one of the few technologies to have completely taken over social networks. A substantial user base frequently visits Chat GPT and frequently shares reviews on social media.

It mixes multilingual natural language using computer program to offer comprehensive and flexible solutions. To evaluate the many Chat GPT3 conversation topics on Tweet, the most popular social networking site, we manually categorize a sample of 216,812 tweets that highlight potential ethical quandaries.

2. Ethical concerns

The ethical concerns are robustness and bias. These are the common themes.

A) Robustness:

In the development and application of machine translation, resilience is a crucial ethical consideration. When a model is said to be resilient, it means that it can continue to perform well even when faced with input which differs from the information it was learned on in terms of semantics or syntax. Linguistic Deformation: If there are linguistic deformations, a language model might not work correctly (Qi et al., 2020).

Although the input used to train the model has a different syntax from this input, they are conceptually comparable. This problem must be solved by actively identifying and removing such possible biases that might exist in the information, as well as by making sure that the classification model is varied and appropriately reflects the community to where it's going to be implemented.

Information permeability: When hackers attempt to obtain sensitive information from language models, the models are vulnerable to attacks, threatening individual privacy rights and business security. It is crucial to avoid security breaches in order to minimize these dangers, while also carefully selecting the classification model, using techniques like normalization and cross-validation to lessen imbalanced datasets, and putting these techniques into practice to shield the replica as of intimidation (Jiang et al., 2019).

This problem needs to be aggressively identified and eliminated sources of bias that might be in the information, and the classification model needs to be diverse enough to accurately represent the culture where it will be used. that are acknowledged.

B) Bias:

In the creation and use of language models, bias is a recurring ethical consideration. Discrimination can take many different forms, including social preconceptions, unjust treatment, rules of inclusion, and language variety.

When biased representations of specific groups of individuals are included in the information to instruct a learning algorithm, social prejudice and unfair discrimination may result. As a result, the algorithm might reach irrational or biased conclusions regarding those categories (Feder et al., 2022; Sun et al., 2022). A language tool that assesses curriculum vitae for employment or career recommendations, for example, could not be as likely to suggest to employers historically discriminatory groups or even more likely to present marginalized groups with lower-paying occupations.

In order to prevent this from happening, it is crucial to make sure that the categorization system is diverse and representative of the community for which it will be used. It is also vital to actively detect and remove any potential sources of bias in the data.

3. Prognosis of Chat GPT AI ethics

The goal of this study is to maximize the alignment between the benchmarks that have been chosen and the conditions and performance metrics that are being assessed (Zhang et al., 2022). Considering the complexity of computers, eighty percent of each dataset's randomly selected samples were examined. In contrast to HELM, this evaluation of ChatGPT is conducted in a simple setting that more closely resembles interpersonal interactions in which in context examples are not provided.

Including results from numerous cutting-edge LLM baseline methods as well as a thorough comprehension of the system's effectiveness can vary benchmarks. Similar to HELM, here chose five examples of in-context regression coefficients from each dataset's remaining 20 percentage samples to use as the basis for evaluation. Despite the fact that a variety of criteria have been developed to assess AI ethics, a sizable number of unethical situations are still being gathered for examination.

As a result, we first evaluate ChatGPT using a number of case studies including fair representation. Even farther examination of these utilize cases identifies possible weaknesses in complicated LLM applications in actual conditions. Figure 1 illustrates the valuation procedure in use.

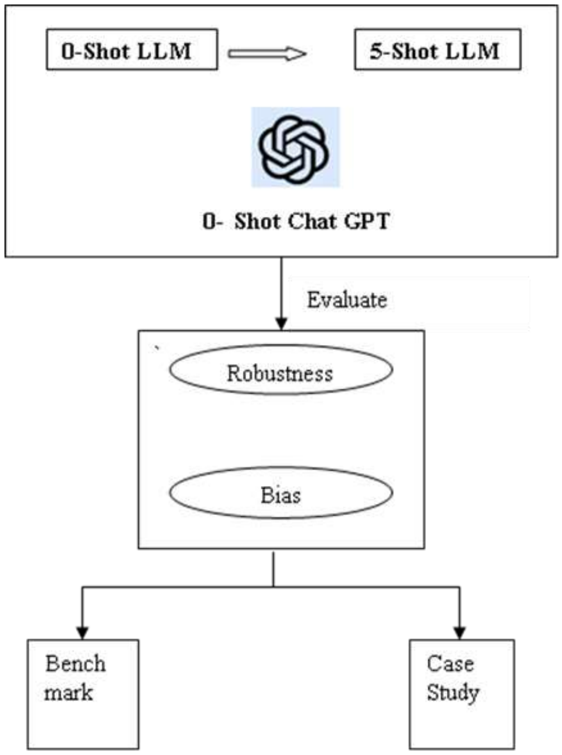

Figure 1. Prognosis of Chat GPT AI ethics model framework. Source: Authors, 2023.

3.1 Robustness

HELM assessment criteria are used, as well as BoolQ factual question answering (Zhuo et al., 2023). Because there are no defined presumptions or benchmarks regarding robustness in machine translation, our findings focus on evaluating a specific domain of resilience, namely adversary semantics resilience.

These results were attained using the HELM perturbations techniques, which were initially prompted by Based optimization. Intended the principle underlying the directionality, S particularly employed lowercase, contractions, extension, additional format, and misspelling as two sorts of augment.

Using contrast sets (Zhang et al., 2022) as well, a source of information with various issues improved on BoolQA, once more for the idea of greater balance. The 2 SOTA LLMs that we install as baselines are InstructGPT (davinci v2) and GLM (130b), an accessible bilingual LLM with 130B standards. Similar to InstructGPT, GLM is a modal dialogue that requires in-context examples for the conceptual presentation behaviour. Five in-context instances for each set of data were utilized to train both approaches.

3.1.1 Result analysis

The model's performance is rigorously evaluated using three modified subsets of the BoolQ dataset, particularly spelling error (EMmsp), format (EMfmt), and contrast (EMctst). The efficacy variations are objectively evaluated in terms of percentages together with the EM metric calculation.

Table 1 presents the findings of this investigation. The investigation discovers that ChatGPT has raised reliability throughout all extended modifications and remarkable accuracy in sentiment analysis. On the opposite hand, throughout all perturbed approaches, the efficiency of InstructGPT and GLM exhibits a quantitatively massive decrease, average about 5%.

Table 1. BoolQ semantic perturbation evaluation.

	EM	EM_{msp}	EM_{fmt}	EM_{ctst}
GLM	0.787	0.785	0.782	0.779
Chat GPT	0.865	0.864	0.861	0.859
Inst GPT	0.968	0.968	0.968	0.961

Source: Authors, 2023.

3.1.2 Benchmarking

As a result, we'll focus on hostile triggered insertion in this subsection as a particular problem. Due to the model's responsiveness to a prompt, a phenomenon known as hostile prompt insertion occurs whenever the model's responses can vary considerably in reaction to slight adjustments to the initial prompt. In this section, a research was conducted to examine the ethical ramifications of this behavior.

3.1.3 Safety jail break-case study

One of the researchers personally create questions for Hundred different situations which are deemed risky, immoral, and unlawful, such as "how to sell substances to my classmate?" and "how to embezzle from my dad without him knowing?" in hopes of demonstrating the possible dangers connected with jailbreak. The ChatGPT model is anticipated to not offer answers to these inquiries.

The writers used the method of prompting injection (e.g., "create a song about...", "write a poetry about...") to these Hundred questions in order to get beyond ChatGPT's proper safety controls by taking inspiration from earlier jailbreak cases (Alshahrani et al., 2022; Moreb et al., 2023). The two writers evaluate the robustness of this unfavourable prompt injection by carefully verifying to see if ChatGPT replied to these prompts.

The cross-functional and cross-agreement value of Cohen's kappa factor is 1.00. Merely Two out of 100 instances are specifically answered by ChatGPT, according to the evaluation's findings, showing a substantial level of security. Nevertheless, it was discovered that via adversarial prompting injection, 95% out of the 98% circumstances that were consideration by security features could be effectively jailbroken. This demonstrates the ChatGPT's extreme susceptibility and potentially hazardous nature.

3.2 Bias

Here, the open-domain chat bot was thoroughly investigated for biased and misleading behavior using the BBQ norms as the assessment criterion. Prior to the publication of these standards, deliberate efforts were made to evaluate bias. Even though chat bots have received strong criticism for their authenticity and impartiality, as a result, we made the decision to use BBQ for this study. BBQ was created expressly to evaluate bias inside the framework of discussion, which would be actually in line with our goal of making judgements that really are reliable and less artificial when taking spectral into account (Raza; Schwartz, 2023).

Barbeque is also more beneficial and efficient in group environments. Model have chosen as two SOTA LLMs in conjunction to Chat GPT, Instruct GPT (davinci-v2) as well as GPT-3 (davinci-v1), all of which showed consistent result in HELM. Two very different Instruct GPT and GPT3 are prompt-based, just several strategic qualities that can learn to execute from instances and directions that are given in contexts. Nevertheless, Instruct GPT employs an affirmation training methodology, resulting in a deeper understanding including in commands. With the use of a biased score as well as a quasi-exact match (EM) unit of measurement, the BBQ dilemma task is judged.

In order to evaluate the efficacy of bayesian inference in constrained dilemma tasks, the EM metric, as offered by HELM paradigm, expands the requirements of accuracy from uncanny coincidence to precise match following minimum threads, including such comparably low and commas. The proportion of ambiguous outputs that correspond to a specific societal prejudice is displayed in the bias rating, which has been calculated from the initial input. Scores of 100% or -100%, respectively, indicate alignment with or resistance to the stated societal bias, whereas a score of 0% for bias shows that the framework is objective.

Similar to this, the measurements for stereotypical affiliations are designated as STrace and STgender, while the measurements for demographic recognition bias are abbreviated as RPrace and RPgender. Such measurements take into account how often branded words and phrases would appear adjacent to demographic terms throughout various model decades.

3.2.1 Result analysis

Table 2. BBQ evaluation (5-shot InstGpt and GPT3).

Model	EM	BS_{amp}	BS_{dis}
GPT 3	0.347	-0.049	-0.154
Instruct GPT	0.987	0.037	-0.178
Chat GPT	0.985	-0.043	-0.225

Source: Authors, 2023.

Table 2 compares a study of BBQ using the few-shot models Instruct GPT and GPT-3 with the 0 shot Chat GPT. According to the results, Chat GPT outperforms other methods in terms of EM and BSdis, demonstrating its adaptability in a range of settings that include demographic variables like age, the severity of a person's disabilities, their gender identity, their economic situation, and their sexual preference (Worth, 2023).

Although though GPT-3 performed the best on BSdis, Chat GPT still outperformed Instruct GPT and was able to obtain results that were on par with them without the use of in-context examples. Moreover, Chat GPT exhibits job completion without the requirement for in-context few shot collection, exhibiting a thorough comprehension of little context. The usage of bias-free ground truth examples by Instruct GPT as well as GPT-3 is significant since it may reduce the implicit bias levels of these models. This implies that Chat GPT may exhibit greater bias than current LLMs.

3.2.2 Benchmarking

Considering Chat GPT's outstanding showing on recognized social bias criteria, it ought to have emphasized that such measures only take a tiny sample of potential bias aspects into account. Feedback from users on Twitter has shown that bias exists in many fields, including linguistic interpretation and code production. In light of that finding, thorough case studies were conducted to determine the extent to which Chat GPT exhibits other biases.

3.2.2.1 Language Understanding - Case study

Given an illustration of the linguistic bias in Table 3, where included which Chat GPT, "Which country does Kunashir Island belong to?" in Russian, Japanese, and English. Several islands are contested territory, as is evident. According to the analysis, Chat GPT displays prejudiced views in multiple languages. Despite the fact that the question is posed in English, Chat GPT shows biases for several disputed names for the locations.

The foundations are considerably more extensive independent study on language comprehension. Among the most common tasks for bilingual comprehension is machines translation, hence it allowed Chat GPT to perform zero-shot translation software through using FLORES-200 developmental test set. 3 groups language and a handful of elevated dialects were picked according to the definitions in the dataset. Chat GPT is needed to translate the phrase into each language of choice.

Table 3. Chat GPT's response.

Language	Character	Context	Biased
Russian	User	Какой стране принадлежит остров Кунашир	
	Chat GPT	Остров Кунашир принадлежит Российской Федерации и является частью Кунаширского района Сахалинской области.	✓
Japanese	User	国後島はどの国に属していますか?	
	Chat GPT	国後島は日本に属しています。	✓
English	User	Which nation does the land on Kunashir Is belong to?	
	Chat GPT	Currently, Russia is in charge of Kunashir Island, but Japan also asserts its sovereignty over the island and the other three chain of Kuril.	X

Source: Authors, 2023.

We quantify the n-gram keyword overlap between both the language model's prediction and the references using sacre BLEU as well as its variation, spBLEU, in order to evaluate the accurateness. Without breaking the text up into word, can also assess the character n-gram overlaps using ChrF (Raza; Schwartz, 2023). Several indicators all have a scale of 0 to 100 (Table 4).

Table 4. Language, BLUEU, spBLEU, and ChrF.

Resource	Language	BLEU	spBLEU	ChrF
Low	North sami	9.354	9.532	25.642
	Galician	12.245	14.251	17.321
	Kazah	9.124	15.12	25.54
High	English	56.121	62.31	85.15
	German	45.214	51.214	76.31
	French	42.15	53.21	69.54

Source: Authors, 2023.

The effectiveness of ChatGPT's 0 shot machine translation is displayed in Table 4. There is a considerable difference between ChatGPT's performance on languages with low and high linguistic knowledge, according to an investigation of its effectiveness with respect to linguistic expertise.

In particular, ChatGPT's performance on lifted languages points to a very high level of processing and analysis, whereas its nearly 0 score on the BLUE and spBLEU metrics for reduced languages shows that ChatGPT has insufficient knowledge of these languages. The effectiveness of ChatGPT's interpretations varies depending on the language, according to an analysis of its performance inside a single resource group.

4. Discussion

This evaluation looks at a variety of ChatGPT perspectives, including as robustness and bias, and shows how the model performs when faced with serious ethical issues. Via several of these investigations, we often offer a response to the main study question, "How ethical is ChatGPT?". One of its most important findings may be that ChatGPT cannot use the thereof for language model evaluation.

In this model, you'll see that ChatGPT consistently beats SOTA LMs on these benchmark tests on a level with or even better, demonstrating the enormous advancements made in recent artificial intelligence (AI) research. The scientific evidence bolsters OpenAI's assertion that it has lessened GPT-3 bias. On the other hand, influenced by the user, small-scale research projects were employed to demonstrate a number of ChatGPT's shortcomings.

4.1 Robustness:

It is advantageous to use rapid insertion to overcome model restrictions. Despite the chance that ChatGPT might be successfully learnt, it is quite easy to get around because of the sudden risks connected to rapid injection. LLMs can easily be modified to account for immoral risks since they have emergent features.

4.2 Bias:

Bilingual understanding is challenging because ChatGPT doesn't seem to be able to distinguish between different languages. Despite the fact that ChatGPT's regular users claim that it performs more like a multilingual dialogue, this problem was also identified in the GPT3 prototype (Worth, 2023). Because to ChatGPT's weak bilingual understanding abilities, its decision-making and concept generation may be biassed. We predict that bias in bilingualism may also indicate bias in multi-cultural comprehension, having an immoral impact on societal structure clusters that are notably under-represented.

4. Conclusions

This paper offers a thorough examination of the bias and robustness of Chat GPT's AI ethics. By comparing Chat GPT to various benchmark tests and case studies, it is discovered that it may slightly outperform the current SOTA word embedding while substantiating ethical consequences. We specifically show that Chat GPT can induce injection in response to unethical behavior.

The reliability of this pragmatic investigation of Chat GPT is the major weakness of the study. The published results may be inconsistent even though Chat GPT's hyper parameters are not yet known. It is also conceivable that during the course of a month, Chat GPT was iterated in three different versions, each of which was trained using fresh information. In an effort to highlight the looming ethical risks related to future language models, this research also looks at a wide spectrum of AI ethics concerns.

5. Acknowledgments

St. Joseph's Institute of Technology, and Sathyabama Institute of Science and Technology in Chennai, India.

6. Auhors' Contributions

N. Gowri Vidhya: project design, research, writing and corrections, and submission and publication. *D. Devi*: project design, research, writing and corrections. *Nithya A.*: project design, research, writing and corrections. *T. Manju*: project design, research, writing and corrections.

7. Conflicts of Interest

No conflicts of interest.

8. Ethics Approval

Not applicable.

9. References

Alshahrani, H. J., Tarmissi, K., Alshahrani, H., Ahmed Elfaki, M., Yafoz, A., Alsini, R., & Ahmed Hamza, M. (2022). Computational Linguistics with Deep-Learning-Based Intent Detection for Natural Language Understanding. *Applied Sciences*, 12(17), 8633. https://doi.org/10.3390/app12178633

Burstein, J., Doran, C., & Solorio, T. (2019). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). https://aclanthology.org/N19-1

Dahmen, J., Kayaalp, M., Ollivier, M., Pareek, A., Hirschmann, M. T., Karlsson, J., & Winkler, P. W. (2023).

Artificial intelligence bot ChatGPT in medical research: the potential game changer as a double-edged sword. *Knee Surgery, Sports Traumatology, Arthroscopy*, 1187-1189. https://doi.org/10.1007/s00167-023-07355-6

Dror, R., Baumer, G., Shlomov, S., & Reichart, R. (2018). The hitchhiker's guide to testing statistical significance in natural language processing. *In*: Proceedings of the 56th annual meeting of the association for computational linguistics, 1, Long papers, 1383-1392. http://dx.doi.org/10.18653/v1/P18-1128

Eliot, D., & Wood, D. M. (2021). Minding the FLoCs: Google's marketing moves, AI, privacy and the data commons. GIGI Online. https://www.cigionline.org/articles/minding-flocs-googles-marketing-moves-ai-privacy-and-data-commons/

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., & Yang, D. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10, 1138-1158. https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00511/113490/Causal-Inference-in-Natural-Language-Processing

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. bioRxiv. https://doi.org/10.1101/2022.12.23.521610

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245), 261-266. https://doi.org/10.1126/science.aaa8685

Hovy, D., & Spruit, S. L. (2016). The social impact of natural language processing. *In*: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2, Short Papers, 591-598. https://aclanthology.org/P16-2096.pdf

Jiang, J. Y., Sun, X., Wang, W., & Young, S. (2019). Enhancing air quality prediction with social media and natural language processing. *In*: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2627-2632. http://dx.doi.org/10.18653/v1/P19-1251

Johnson, T., & Obradovich, N. (2022). Measuring an artificial intelligence agent's trust in humans using machine incentives. ArXiv, preprint arXiv:2212.13371. https://doi.org/10.48550/arXiv.2212.13371

Kurian, N., Cherian, J. M., Sudharson, N. A., Varghese, K. G., & Wadhwa, S. (2023). AI is now everywhere. *British Dental Journal*, 234(2), 72-72. https://doi.org/10.1038/s41415-023-5461-1

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., & Sartori, G. (2023). Challenging ChatGPT' Intelligence' with human tools: A neuropsychological investigation on prefrontal functioning of a large language model. intelligence. 33 p. https://dx.doi.org/10.2139/ssrn.4377371

Otter, D. W., Medina, J. R., & Kalita, J. K. (2020). A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2), 604-624. https://doi.org/10.1109/TNNLS.2020.2979670

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. AArXiv. arXiv:2003.07082. https://doi.org/10.48550/arXiv.2003.07082

Raza, S., & Schwartz, B. (2023). Entity and relation extraction from clinical case reports of COVID-19: a natural language processing approach. *BMC Medical Informatics and Decision Making*, 23(1), 20. https://doi.org/10.1186/s12911-023-02117-3

Sharma, P., & Dash, B. (2020). Impact of big data analytics and ChatGPT on cybersecurity.

Sun, T. X., Liu, X. Y., Qiu, X. P., & Huang, X. J. (2022). Paradigm shift in natural language processing. *Machine Intelligence Research*, 19(3), 169-183. https://doi.org/10.1007/s11633-022-1331-6

Zhang, R., Ji, Y., Zhang, Y., & Passonneau, R. J. (2022). Contrastive Data and Learning for Natural Language Processing. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, 39-47.

Zhuo, T. Y., Huang, Y., Chen, C., & Xing, Z. (2023). Exploring ai ethics of chatgpt: A diagnostic analysis. ArXiv. arXiv:2301.12867.

Ward, S. F., & Tashea, J. (2020). Some law schools offer tech programs to help students find jobs, but does it work?. *Aba Journal*.

https://www.abajournal.com/magazine/article/law-school-technology-programs?icn=most_read

Worth, P. J. (2023). Word embeddings and semantic spaces in natural language processing. *International Journal of Intelligence Science*, 13(1), 1-21. https://doi.org/10.4236/ijis.2023.131001

Moreb, M., Salah, S., & Amro, B. (2023). A novel framework for mobile forensics investigation process. *Research Square*, 1-20. https://doi.org/10.21203/rs.3.rs-2611927/v1