ISSN: 3065-0712

Current Research and Innovations Journal

Research Article

A WEB APPLICATION FOR CORRECTING LANGUAGE MODEL MISALIGNMENT THROUGH REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

[1]Chukwudi Daniel Okeke and [2]Blessing Grace Udom

[1]Department of Computer Science, National Open University of Nigeria, Nigeria.

[2]Department of Computer Science, Akwa Ibom State University, Mkpat Enin, Nigeria.

Abstract

Recent years have seen tremendous progress in the field of artificial intelligence, which has sparked the creation of cutting-edge tools like OpenAI ChatGPT. The OpenAI GPT -3 family of big language models serves as the foundation for ChatGPT, which is enhanced through the use of supervised and reinforcement learning methodologies. Its goal is to produce text that can't be distinguished from human-written information. It can hold conversations with users in a way that is surprisingly clear-cut and uncomplicated. Reinforcement Learning from Human Feedback (RLHF) is the technique employed. Human input and machine learning methods (Supervised Learning) are used to train the model. It is employed in the training phases to reduce biased, damaging, and false outputs. The resulting Instruct models are much better at following instructions than GPT-3. Above all, customized ChatGPT web application that can fine-tune a given input and generate text that is of high quality, harmless, truthful and appropriate, without biased outputs. A key motivation for our work is to increase helpfulness and truthfulness output while mitigating the harms and biases of language models. In conclusion, our results show that reinforcement learning from human feedback (RLHF) techniques is effective at significantly improving the alignment of general-purpose AI systems with human intentions.

Keywords: Artificial Intelligence; ChatGPT; OpenAI; Reinforcement Learning; Human Feedback; InstructGPT models

Introduction

According to Lee and Shirani (2004), Web development has evolved significantly over the years, with HTML, CSS, and JavaScript being the core technologies for building interactive and engaging user interfaces. In an era where digital communication and online interactions have become integral to both businesses and individuals, there is a growing need for a tailored and highly responsive conversational AI solution (Edet & Ansa, 2023). Recent major progress in the field of artificial intelligence has resulted in the creation of cutting-edge technologies like OpenAI ChatGPT (Roumeliotis and Tselikas, 2023). The ChatGPT language model is state-of-the-art technology that has the power to drastically alter the Web development industry. As the integration of ChatGPT in Web development follows principles of performance, reliability and quality. ChatGPT is the most advanced chatbot that has ever been developed. A chatbot is a piece of software with artificial intelligence that can carry on

human-like dialogue. Users can ask questions or make requests, and the system responds within seconds (Rudolph et al., 2023).

ChatGPT is built on top of OpenAI's GPT -3 family of large language models and is fine-tuned with both supervised and reinforcement learning techniques. Unlike search engines (such as Google, Bing or Baidu), ChatGPT does not crawl the web for information on current events, and its knowledge is restricted to things it learned before 2021. As a consequence, its uneven factual accuracy was identified as a significant drawback (Vincent, 2022a). OpenAI's Generative Pretrained Transformer (GPT) language model was modified to create the state-of-the-art language model ChatGPT. Its goal is to produce text that can't be distinguished from human-written information. It has the ability to have discussions with users in a way that is surprisingly clear and simple (Mhlanga, 2023). With the increase in the dependency on the Web-based systems and applications, the importance of their performance, reliability and quality have become very significant. This study explores building frontend technologies (HTML, CSS, and JavaScript) with backend technologies (Node.js, Express) with OpenAI API integration to create a robust web application. The frontend refers to the user interface that the user interacts with. The frontend was simply responsible for the visual aspects of a website, such as its layout, color scheme, and font choices (Ekong et al., 2023). One of the primary responsibilities of the frontend is to retrieve data from the backend through an application programming interface (API) (Ekong et al., 2022). Finally, the frontend plays a crucial role in how users interact with and navigate a website or application. It combines design skills and programming knowledge to create a dynamic and userfriendly experience (Le, 2020). Backend refers to everything data related. This is where logical operations occur and is responsible for security, what kind of data and logic goes to the front-end. The backend provides some API which comprehends one another. In all, the backend houses the business logic, handles security concerns, and maintains a connection to the database. (Vickler, 2021). An API (Application Programming Interface) is a set of rules, protocols, and tools that allows different software and web applications to communicate with each other (IBM, 2023). The API facilitates the integration of OpenAI's advanced AI capabilities into an array of applications, products, and services. GPT-3 and GPT-3.5 are a series of language models developed by OpenAI for generating human-like natural language text. These models GPT-3 series models consist of davinci and text-davinci-001 while GPT-3.5 series models consist of code-davinci-002, textdavinci-002, text-davinci-003, and gpt-3.5-turbo. OpenAI then used supervised fine-tuning to create text-davinci-002 and introduced the Reinforcement Learning from Human Feedback (RLHF) training strategy to create text-davinci-003, which improved its ability to understand instructions and generate text. (Christiano et al., 2017; Stiennon et al., 2020).

ISSN: 3065-0712

Current Research and Innovations Journal

Research Article

Methods

The method used is Reinforcement Learning from Human Feedback (RLHF). The model is trained using a combination of machine learning techniques (Supervised Learning) and human input. It used the training steps to minimize harmful, untruthful, and unbiased outputs.

Model formation

Fine-tuning ChatGPT with Reinforcement Learning from Human Feedback (RLHF) consisted of three distinct steps:

Step 1: Collect demonstration data and train a supervised policy

A pre-trained language model is fine-tuned on a relatively small amount of demonstration data curated by labelers, to learn a supervised policy (the Super Fine-Tuning model) that generates outputs from a selected list of prompts. This represents the baseline model. Having collected a dataset S of (x, y_0, y_1, y_2, y_3,…, y_n) tuples,

Step 2: Collect comparisons data and train a reward model

Labelers are asked to vote on a relatively large number of the **supervised fine-tuning (SFT)** model outputs, this way creating a new dataset consisting of comparison data. A new model is trained on this dataset. This is referred to as the **reward model (RM)**.

We train this model to predict which summary $y \in \{y_0, y_1\}$ is better as judged by a human, given a post x. If the summary preferred by the human is y_i, we can write the RM loss as:

$$\text{loss}(r) = \mathbb{E}_{(x,\{y_i\}_i,b)\sim S}\left[\log \frac{e^{r(x,y_b)}}{\sum_i e^{r(x,y_i)}}\right] \qquad (1)$$

where r_θ (x, y) is the scalar output of the reward model for post x and summary y with parameters θ, and D is the dataset of human judgments. At the end of training, we normalize the reward model outputs such that the reference summaries from our dataset achieve a mean score of 0.

Step 3: Optimize a policy against the reward model using the Proximal Policy Optimization (PPO) reinforcement learning algorithm

The reward model is used to further fine-tune and improve the SFT model. The outcome of this step is the so-called **policy model**. We want to use the reward model trained above to train a policy that generates higher-quality outputs as judged by humans. We include a term in the reward that penalizes the KL divergence between the learned RL policy π RL φ with parameters φ and this original supervised model π SFT (Schulman et al., 2017). The full reward R can be written as:

$$R(x,y) = r_\theta(x,y) - \beta \log[\pi_\phi^{\text{RL}}(y|x)/\pi^{\text{SFT}}(y|x)] \qquad (2)$$

ISSN: 3065-0712

Current Research and Innovations Journal

Research Article

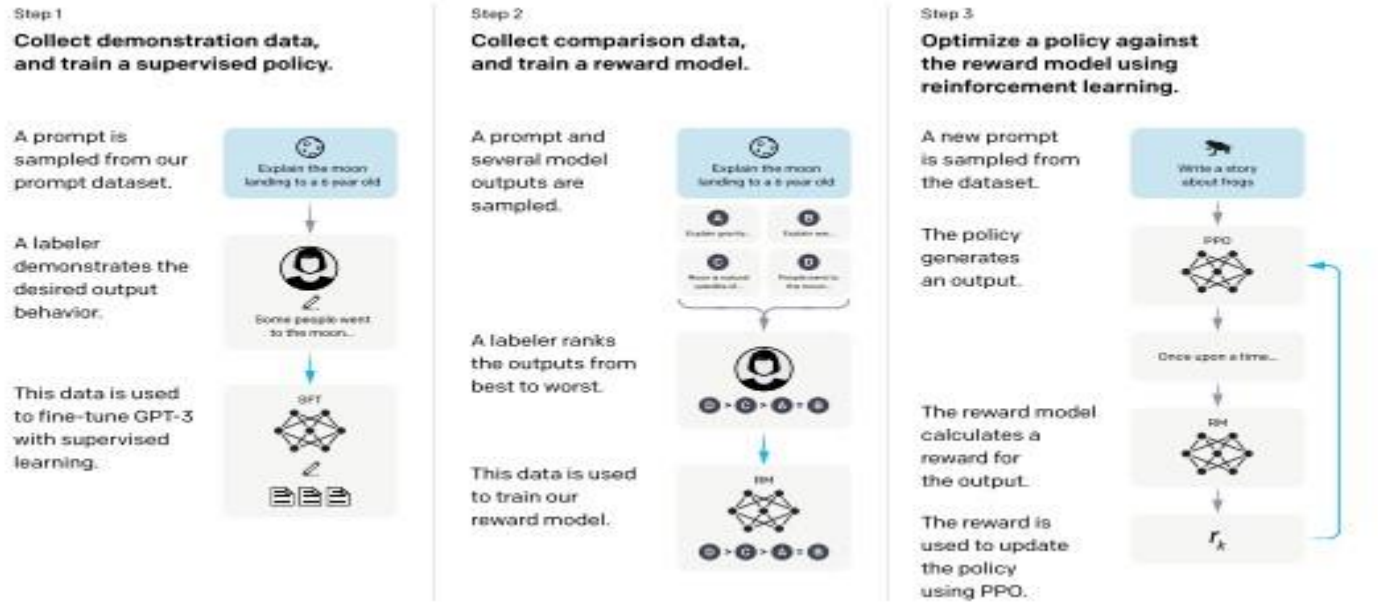

Figure 1: A diagram illustrating the three steps of our method: (1) Collect demonstration data and train a supervised policy (2) Collect comparisons data and train a reward model, and (3) Optimize a policy against the reward model (Ouyang et al.,2022).

The resulting Instruct models are superior in quality at following instructions than GPT-3. We first assess how well outputs from Instruct follow user instructions, by having labelers equate its outputs to those from GPT-3. We find that Instruct models are substantially choose on prompts submitted to both the Instruct and GPT-3 models on the API.

Software and Development Environment

Creating ChatGPT web application with GPT model typically involves integrating OpenAI's API or GPT-3.5 model into our project. We will be looking at the development tools used, the programming languages used and techniques. The development environment has a number of elements which include: Visual Studio Code and Visual Studio Code Extensions.

Node Package Manager (NPM): Packages that were used are **CORS**, **Vite:** Vue.js and React, Dotenv, **Express:** (Express.js,), **Nodemon:** (Node.js). It supports a wide range of programming languages, including JavaScript, Typescript, CSS, HTML, JSON, and more. The version control system used in this project is Github which is a cloud-based version control system. The operating system utilized is Windows 10 Home, which is well adapted for web development.

ISSN: 3065-0712

Current Research and Innovations Journal

Research Article

The frontend of the application was deployed on Netlify (https://www.netlify.com) while the backend was deployed on Render (https://render.com.)

Results

The results of the research are in two folds. Firstly, configuring frontend technologies (HTML, CSS, and JavaScript) using client side on Netlify and backend technologies (Node.js, Express) using server side on Render with OpenAI API integration to create a robust web application. Secondly, the resulting InstructGPT models are superior in quality at following instructions than GPT-3. Above all, customized ChatGPT web application that can fine-tune a given input and generate text that is of high quality, harmless, truthful and appropriate, without biased outputs. See figure 2.

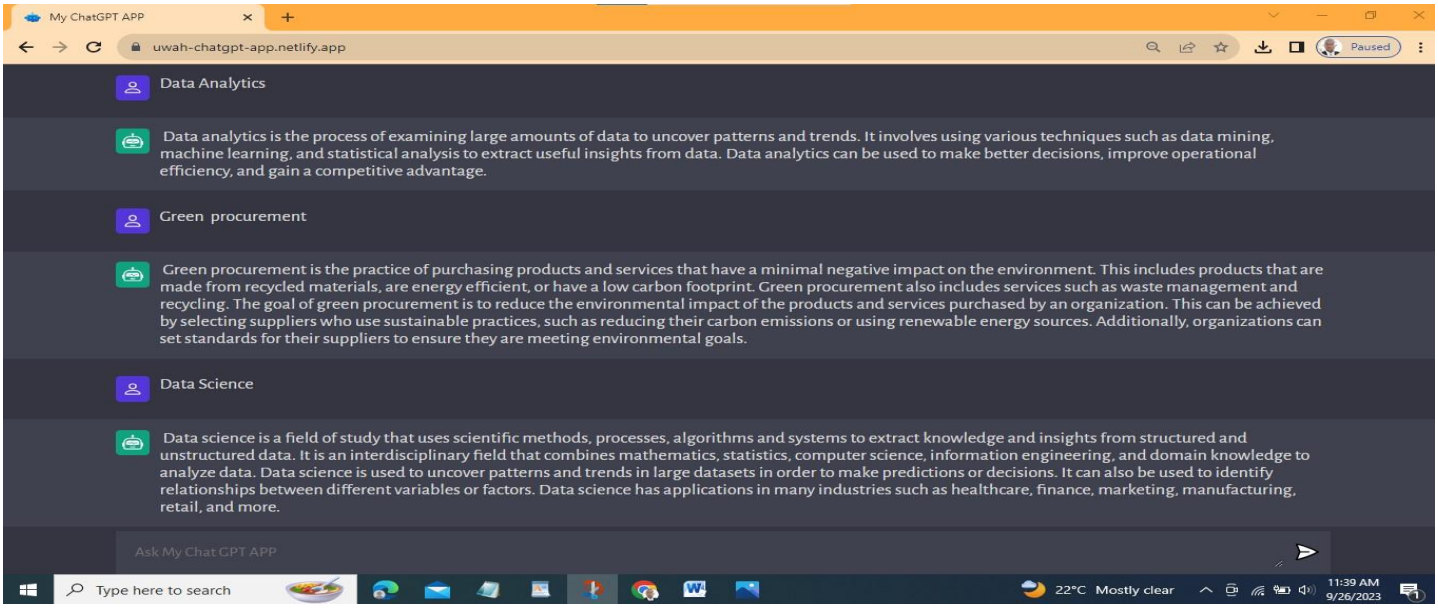

Figure 2: Customized ChatGPT web application that can fine-tuned a given input and generate output.

Discussion

So these training strategies to predict the next word (or a masked word) in a text sequence may not necessarily be learning some higher-level representations of its meaning, these training procedures cause the language model to become misaligned for some more difficult tasks. Based on the original GPT-3 model, ChatGPT has been further trained with the aim of minimizing the model's misalignment difficulties through the use of Reinforcement Learning guided by human feedback. But how can we tackle the alignment problem using Reinforcement Learning based on human feedback? In particular, we adjust or fine-tune GPT-3 to adhere to a wide range of procedure (Christiano et al., 2017; Stiennon et al., 2020). We first collect a dataset of human-written

ISSN: 3065-0712

Current Research and Innovations Journal

Research Article

demonstrations on prompts submitted to our API, and use this to train our supervised learning baselines. Next, we collect a dataset of human-labeled comparisons between two model outputs on a larger set of API prompts. We then train a reward model (RM) on this dataset to predict which output our labelers would prefer. Finally, we use this RM as a reward function and fine-tune our GPT-3 policy to maximize this reward using the PPO algorithm (Schulman et al., 2017). The resulting InstructGPT models are superior in quality at following instructions than GPT-3. It is important to know that safety and alignment problems we are aiming to solve are complex and subjective, and are not fully captured by simple automatic metrics. A key motivation for our work is to increase helpfulness and truthfulness while mitigating the harms and biases of language models. To measure the quality of our models, we primarily use a suite of existing metrics on publicly available datasets as shown figure 3.

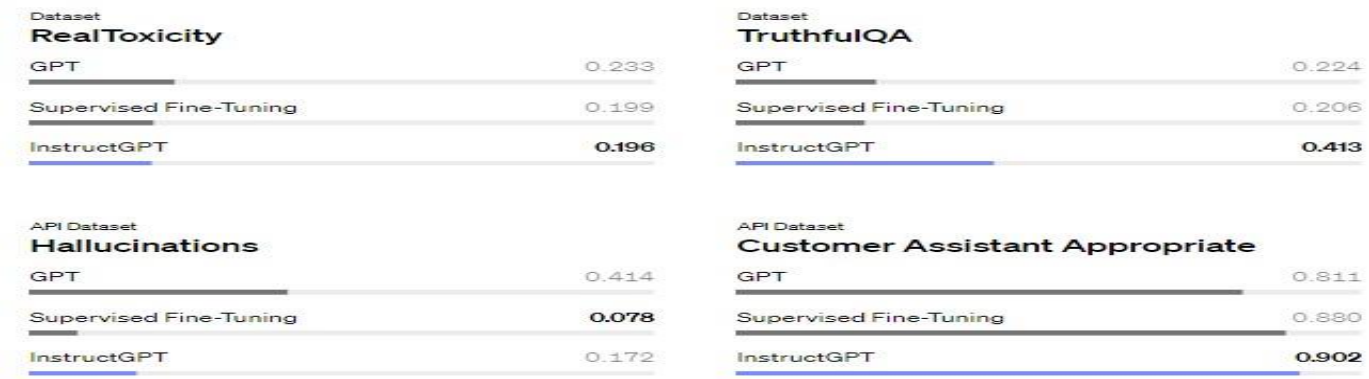

Figure 3: Evaluating the datasets (Source: https://openai.com/research/instruction-following)

Assessing the suitability, toxicity and veracity of InstructGPT. Higher ratings are better for TruthfulQA and appropriateness, and lower levels are better for toxicity and hallucinations. Finally, the resulting InstructGPT models are ultimately significantly more adept in following instructions than GPT-3. They also demonstrate marginal reductions in the development of harmful products and fewer fabrications of facts.

Conclusion

The research findings reveal the profound impact of reinforcement learning from human feedback (RLHF) on enhancing the alignment of general-purpose AI systems with human intentions. By integrating human preferences into the training process, RLHF methodologies, as demonstrated through the development of InstructGPT models, offer a promising approach to refining AI generated text. This human-centric alignment represents a paradigm shift in AI development, prioritizing the harmonization of machine behavior with human values, preferences, and expectations. As a result, AI systems trained using RLHF exhibit a heightened ability to generate text that resonates with users on a deeper level, fostering a more intuitive and satisfying interaction experience. A key outcome of applying RLHF techniques is the substantial enhancement in the quality of generated text. Through

iterative adjustments guided by human feedback, InstructGPT models refine their language generation capabilities, producing outputs characterized by improved coherence, relevance, and linguistic accuracy. This quality enhancement translates into tangible benefits for users, including a more natural and engaging conversational experience, as well as increased utility and applicability of AI-generated content across various domains and applications. Furthermore, RLHF plays a pivotal role in ensuring the safety, trustworthiness, and ethical integrity of AI-generated text. By leveraging human preferences and judgments, InstructGPT models are trained to prioritize safety considerations, minimizing the risk of producing harmful or inappropriate content. This proactive approach to safety mitigation instills confidence among users and stakeholders, mitigating concerns surrounding the potential negative consequences associated with AI-generated text, such as misinformation, harmful stereotypes, or offensive language. Additionally, RLHF methodologies contribute to the promotion of truthfulness, accuracy, and unbiased mitigation in AI-generated text. By learning from human labeled comparisons and preferences, InstructGPT models refine their understanding of what constitutes truthful and reliable information, thereby producing outputs aligned with factual accuracy. Moreover, RLHF actively addresses the challenge of bias in AI systems by identifying and rectifying biases present in training data and model outputs, fostering fairness, inclusivity, and neutrality in language generation. This holistic approach to text generation represents a significant step forward in advancing the responsible and ethical development of AI technologies.

References

Alessio, H.M.; Malay, N.; Maurer, K.; Bailer, A.J.; Rubin, B. (2018). Interaction of proctoring and student major on online test performance. Int. Rev. Res. Open Distrib. Learn. 19, 166–185. [CrossRef]

Azaria, A. (2022). ChatGPT usage and limitations. Preprint. DOI:10.13140/RG.2.2.26616.11526

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H. P. d. O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al. (2021). Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.

Chesterman, S. (2023). AI-generated content is taking over the world. But who owns it? The Straits Times, https://www.straitstimes.com/opinion/ai-generatedcontent-is-taking-over-theworld-but-who-owns-it.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems, pages 4299–4307.

Cotton, D. R., Cotton, P. A., & Shipway, J. R. (2023). Chatting and cheating. Ensuring academic integrity in the era of ChatGPT. Preprint. https://doi.org/10.35542/osf. io/mrz8h.

Deng, J., & Lin, Y. (2022). The benefits and challenges of ChatGPT: An overview. Frontiers in Computing and Intelligent Systems, 2(2), 81-83.

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. Preprint. bioRxiv 2022.12.23.521610; doi: https://doi.org/10.1101/2022.12.23.521610

Hao, K. (2022). Everything to know about Elon Musk's OpenAI, the maker of ChatGPT. Augustman, https://www.augustman.com/sg/gear/tech/openai-what-to-knowabout-thecompany-behind-chatgpt.

Harkut, D. G., & Kasat, K. (2019). Introductory chapter: artificial intelligence challenges and applications. Artificial Intelligence-Scope and Limitations.

IBM (2023) "What is an API?". https://www.ibm.com/topics/api. Retrieved: 2023-04-11.

Kim, B., Kim, H., Lee, S. W., Lee, G., Kwak, D., Jeon, D. H., ... & Sung, N. (2021). What changes

can large-scale language models bring? intensive study on hyperclova: Billions-scale korean generative pre trained transformers. arXiv preprint arXiv:2109.04650

Kundalia (2023). ChatGPT and the future of writing. Hindustan Times. Retrieved January 31, 2023, from (https://www.hindustantimes.com/books/ chatgpt-and-the-future-of-writing101675090609362.html).

Lauterbach, A. (2019). Artificial intelligence and policy: quo vadis?. Digital Policy, Regulation and Governance.

Le (2020). Different teams will often have very different ideas of what part of their app is the frontend. Available: https://www.theseus.fi/handle/10024/340287. Accessed: 02.05.2022.

Lee and Shirani (2004). A Component Based Methodology for Web Application Development. Journal of Systems and Software, 71(1−2), 2004, pp.177−187.

MakeUseOf (2023) "A Guide to the OpenAI API and What You Can Do With It". https://www.makeuseof.com/openai-api-guide-what-can-you-do/. Published: 2023-03-16. Retrieved: 2023-04-11.

Mhlanga (2023). Open AI in Education, the Responsible and Ethical Use of ChatGPT Towards Lifelong Learning. https://ssrn.com/abstract=4354422.

OpenAI (2023). "DALL.E2". Available at: https://openai.com/dall-e-2/.

OpenAI (2022). New and Improved Content Moderation Tooling. (https://openai.com/blog/newand-improved-content-moderation-tooling/). Accessed: 24 FEB 2023.

Ouyang Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong

Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, Ryan Lowe (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. Preprint. 1–12.

Roumeliotis, K.I.; Tselikas, N.D. (**2023).** ChatGPT and Open-AI Models: A Preliminary Review. Future Internet, 15, 192. https://doi.org/10.3390/fi15060192.

Rudolph, J.; Tan, S. (2023). Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? J. Appl. Learn. Teach.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

Shklar and Rosen (2003). Web Application Architecture: Principles, Protocols and Practices. USA: John Wiley & Sons.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. (2020). Learning to summarize from human feedback. arXiv preprint arXiv:2009.01325.

Susnjak, T. (2022). ChatGPT: The end of online exam integrity? Preprint. arXiv:2212.09292v1.

Susnjak, T.(2022). CHATGPT: The end of online exam integrity? arXiv, arXiv:2212.09292.

Tien, J. M. (2017). Internet of Things, real-time decision-making, and artificial intelligence. Annals of Data Science, 4, 149-178.

Uc-Cetina et al (2022). Survey on reinforcement learning for language processing. Artificial Intelligence Review, 1–33. https://doi.org/10.1007/s10462-022-10205-5.

Vaswani et al., (2017). Attention is all you need. Advances in neural information processing systems, 30. 31st Conference on Neural Information Processing Systems. CA, USA: Long Beach, Available at https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aaPaper.pdf.

Vickler (2021). Backend development in simple terms is all the things happening in the background that you cannot visibly see. Available: https://www.amazon.com/Javascript-Back-EndProgramming/dp/B08YFC7YZG. Accessed: 22.04.2022.

Vincent, J. (2022a). AI-generated answers temporarily banned on coding Q&A site. Stack Overflow. https://www.theverge.com/2022/12/5/23493932/chatgptai-generated-answerstemporarily-banned-stack-overflowllms-dangers.

Edet, A. E. and Ansa, G. O. (2023). Machine learning enabled system for intelligent classification of host-based intrusion severity. Global Journal of Engineering and Technology Advances,16(03), 041–050.

Ekong A., Ekong B., and Edet A. (2022). Supervised Machine Learning Model for Effective Classification of Patients with Covid-19 Symptoms Based on Bayesian Belief Network, Researchers Journal of Science and Technology, 2, 27-33.

Ekong, B., Ekong, O., Silas, A., Edet, A., & William, B. (2023). Machine Learning Approach for

Classification of Sickle Cell Anemia in Teenagers Based on Bayesian Network. Journal of Information Systems and Informatics, 5(4), 1793-1808. https://doi.org/10.51519/journalisi.v5i4.629.