An Evaluation of ChatGPT's Reliability in Generating Biographical Text Outputs

Kehinde Oloyede, Cristina Luca[a] and Vitaliy Milke[b]
School of Computing and Information Science, Anglia Ruskin University, Cambridge, U.K.

Keywords: Artificial Intelligence, Large Language Model, ChatGPT, Biography Generation.

Abstract: The rapid evolution of large language models has transformed the landscape of Artificial Intelligence-based applications, with ChatGPT standing out for generating text that feels human-like. This study aims to assess ChatGPT's reliability and consistency when creating biographical texts. The paper focuses on evaluating how precise, consistent, readable and contextually appropriate the model's biographical outputs are, taking into account various interactions and inputs. The input consisting of a biographical text dataset, specific rules and a prompt was used in an extensive experimentation with ChatGPT. The model's performance was assessed using both quantitative and qualitative measures, scrutinising how well it maintains consistency across different biographical scenarios. This paper shows how greater coherence and accuracy in text generation can be achieved by creating detailed and structured directives. The significance of this study extends beyond its technical aspects, as accurate and reliable biographical data is essential for record-keeping and historical preservation.

1 INTRODUCTION

Biography writing has long been vital to understanding influential lives but traditionally requires labor-intensive research and meticulous fact-checking. The introduction of ChatGPT simplifies this process, though variability in the generated outputs could affect the AI's credibility and user trust. This study aims to investigate whether factors such as the time of day, rule complexity and the AI's interpretative ability contribute to these inconsistencies and suggest strategies to enhance the reliability of AI-generated biographical information.

A notable factor that may influence output consistency is the time of day users interact with ChatGPT. Global user activity may lead to server congestion during peak times, particularly when high demand in the US coincides with UK afternoon hours, possibly impacting performance. Previous research suggests that server load fluctuations can affect AI accuracy and response times (Aslam and Curry, 2021). This paper aims to explore the impact of the variations in server load at different times on the consistencies in biographical outputs, thereby determining whether timing affects reliability.

Another critical consideration is the complexity and volume of interpretative rules ChatGPT must follow. As the number of specific instructions increases, so does the likelihood of inconsistency due to the AI's limited capacity to process simultaneously and prioritise numerous rules, leading to errors in output (Kshetri, 2023). This study assesses how variations in rule volume affect ChatGPT's capacity to produce accurate biographical entries consistently.

The last factor is the AI's capacity to comprehend and apply rules correctly, crucial for achieving reliable outputs. Misinterpretations or incomplete applications of rules can lead to inconsistencies, emphasizing the need for a robust feedback mechanism to verify rule adherence. Such a mechanism, similar to code validation systems, could help the AI clarify and follow guidelines accurately, enhancing the consistency and dependability of its responses (Steiss et al., 2024).

Based on the reasons outlined above, this research aims to achieve three objectives: (I) assess the impact of time of day on biographical output reliability; (II) evaluate the relationship between rule complexity and output consistency; and (III) investigate the AI's rule comprehension and potential for feedback-driven improvement. Using a mixed-methods approach, biographical entries were generated at set intervals throughout the day, with varied rule sets to assess adherence and output quality. By analysing

[a] https://orcid.org/0000-0002-4706-324X
[b] https://orcid.org/0000-0001-7283-28670

Oloyede, K., Luca, C. and Milke, V.
An Evaluation of ChatGPT's Reliability in Generating Biographical Text Outputs.
DOI: 10.5220/0013248400003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 993-1000
ISBN: 978-989-758-737-5; ISSN: 2184-433X

the consistency of outputs, this study aims to determine factors that enhance or hinder the quality of biographical content generated by ChatGPT. The findings will offer insights for optimising AI performance and guiding future improvements in the development of reliable AI-generated biographical content.

2 LITERATURE REVIEW

Natural Language Processing (NLP), a branch of artificial intelligence (AI) focuses on developing models and algorithms that enable machines to process, understand and generate text naturally. Its foundation is built on a variety of linguistic theories and computational techniques (Jurafsky, 2000). Initially heavily reliant on fixed rules, NLP struggled with deeper nuances in languages and context (Reither and Dale, 2000), leading to advancements through statistical models and machine-learning techniques.

The introduction of the transformer architecture (Vaswani, 2017) revolutionised NLP with its self-attention mechanisms, enabeling models to capture complex dependencies by processing the entire sequences of words simultaneously. The architecture led to more sophisticated language models like BERT (BidirectionalEncoder Representations from Transformers) and GPT (Generative Pre-trained Transformer). BERT uses a bidirectional approach to pretrain, enabling it to understand contexts from both directions (Kenton and Toutanova, 2019). GPT is pretrained on a vast amount of data to be able to generate coherent and contextually relevant text and images. In 2020, GPT models were one of the largest and most powerful AI models with an impressive 175 billion parameters in GPT-3 ((Brown, 2020), (Wu et al., 2023)). Estimations done in 2023 state that ChatGPT4 has approximately 1.8 trillion parameters - with the architecture consisting of eight models, with each internal model made up of 220 billion parameters (Howarth, 2024).

GPT has found extensive applications in healthcare, particularly through AI chatbots, used for preliminary patient consultations (Oh, 2022) which efficiently collect patient information and provide basic medical advice. Also, GPT tools could also complement the traditional therapy methods as conversational agents in administering cognitive behavioural therapy (CBT) to individuals experiencing mild to moderate mental issues (Jiang et al., 2024).

Research by (Rao et al., 2023) assessed GPT's effectiveness in clinical decision support, particularly in radiology, by examining its ability to recommend suitable imaging services for breast cancer screening and breast pain evaluation. (Jiang et al., 2024) explored how ChatGPT navigates large volumes of medical literature, effectively identifying trends and distilling essential findings to facilitate research. Beyond research, (Saleem and Khan, 2023) highlighted its role in simulating patient interactions for medical students, by generating realistic patient scenarios.

GPT-based tools have also been used in education with (Holmes et al., 2019) explaining how these systems offer customized learning experiences to each student's pace and learning style. (Rudolph et al., 2023) also emphasised this point, examining how chatbots are used in providing support for students. These claims are supported by (Crow et al., 2018), who describes how GPT could assist students in simplifying complex concepts by providing explanations and illustrative examples. Aside from learning, GPT can be used in administrative positions by reducing burdens on educators (Roll and Wylie, 2016). (Madunić and Sovulj, 2024) discusses how it helps in developing lesson instruction materials, lesson plans and educational resources.

(Htet et al., 2024) discusses the use of GPT-based tools in the commercial sector, helping businesses to optimise their marketing strategies and reach their target audience. By automating processes, companies can allocate resources more efficiently to critical areas of operation (Bansal et al., 2024). Professionals can also get their ideas and innovations refined by the use of GPT tools in creative industries (Sarrion, 2023).

GPT-based tools have also been used in generating biographical outputs. For instance, (Xie et al., 2024) fine-tuned GPT-3 on a dataset of historical figures' biographies, which shows that fine-tuning improves the generated outputs. (Rashid et al., 2024) also suggested that educational institutions can use GPT models to create biographical content for teaching materials. Moreover, (Bender et al., 2021) stressed the importance of implementing robust fact-checking and bias mitigation strategies to ensure the ethical use of the GPT models.

In the overall aspect of content creation, GPT-based tools help in services like information discovery, valuable text generation, reference assistance, and even the development of guides and tutorials (Ali et al., 2024).

3 METHODS

3.1 Overview

This study aims to systematically examine how the following factors influence GPT-generated output, us-

ing biography writing as an example: the timing of user interactions with the AI system, the complexity of interpretative directives the AI is programmed to follow, and the system's capability to accurately interpret and apply these directives. To address this research problem, the methodology is designed to offer a comprehensive analysis of these factors and their influence on the consistency of AI-generated outputs. The study first examines the hypothesis that time-of-day interactions impact system performance due to fluctuations in server load. Next, it addresses how the complexity of the interpretative rules that the AI must process affects output; as the number and complexity of these rules increase, the likelihood of errors and inconsistencies in outputs rises correspondingly. Finally, the study assesses the AI's ability to comprehend and implement these rules accurately, as insufficient understanding can lead to inaccuracies and inconsistencies.

3.1.1 Data Collection and Experimental Design

To test the hypotheses, an experimental setup was designed to collect the necessary data. These steps include the following:

- Time-Based Data Collection: To investigate the impact of time on output variability, outputs will be generated at various times throughout the day. This will be done over 14 days (2 weeks) to ensure that we have a robust data set.
- Rule Complexity Testing: GPT will be provided with tasks of varying rule complexity. These tasks will range from simple rules to highly complex structures. The outputs will then be assessed to analyse the impact of rule complexity.
- Rule Comprehension Analysis: A set of rules has been designed with varying levels of complexity. GPT's ability to comprehend and apply given rules will be tested, and the outputs will be analysed for consistency and accuracy.

3.1.2 Data Processing and Pre-Analysis

Following data collection, the subsequent phase entails processing the data in preparation for analysis. This stage includes:

- Data Cleaning: Ensuring the dataset is free of errors, duplicates and inconsistencies is essential for preserving data integrity and ensuring reliable results during analysis. This process involves correcting inaccuracies, such as updating misrecorded job titles, eliminating redundancies to avoid repeated information and verifying consistency in key details like job roles and career trajectories across the dataset.
- Categorisation and Tagging: Data will be systematically categorised according to variables such as time of day, rule complexity or clarity of rule comprehension. Tagging in this manner aids in streamlining the analysis process and enables the extraction of meaningful correlations. Each entry is tagged with the recording time and rule complexity. The rule sets have been divided into four (ranging from simple to complex):
  Rule set 1 - Word count;
  Rule set 2 - word count and biographical structure;
  Rule set 3 - word count, biographical structure and biographical style;
  Rule set 4 - word count, biographical structure, biographical style and use of language.
- Correlation with Performance Metrics: The biographical data outputs will be examined to global server performance metrics, including performance time and server load, to assess whether these variables exhibit a relationship with output variability.

3.2 Rules

As this research aims to evaluate ChatGPT's ability to generate biographical texts that adhere to specified rules on structure, length, tone and neutrality, a few sets of rules have been created to assess the reliability of ChatGPT responses critically.

> Using these instructions please write a factual report on the parliamentary candidate below of no more than 250 words using the past tense and British English spelling and grammar. You should assess the data neutrally, use a boring and non-contentious writing style and remove any self-promotion. Do not include the fact that the parliamentary candidate is a prospective parliamentary candidate. Avoid including dates.
> **Biography Structure.**
> The biography should always start by evaluating the most significant career achievements of the candidate. It then should state the candidate's political experience level and give examples of any significant achievements. It should then give details of any community or voluntary role and associated achievements made by the candidate. Lastly, it should describe any significant political interests and evidence for them.
> **Biography Style.**
> Use a simple clear neutral writing style, this is the most important rule. It should be written in the

past tense. Use an analytical style with insights and statements supported by specific facts. Use a themed approach, do not write as a historical narrative. Do not make value judgments, only report facts. Please use British English spelling and grammar, for example for words such as 'organisation' and 'specialised'. Do not mention that they have been a prospective parliamentary candidate. Please use a candidate's name only at the very start of the candidate biography. After that, please use pronouns such as he or she as appropriate.
Use of Language.
The following words are prohibited in the output, always find an alternative:- political vistas, political domain, multifaceted, societal, transitioning.
Please substitute the phrases, words and characters below with the value after the symbol '='
With a background deeply embedded in = With a background in
His political interests included = His political interests are likely to include
Her political interests included = Her political interests are likely to include
illustrating her commitment to = illustrating her focus on
Political Journey = Political Career
demonstrating a commitment to = with a focus on
evidencing = showing
underscored = showed
underscores his multifaceted approach = shows his approach
showcased = showed
showcases = shows
ascended = advanced
showcasing = showing
dedication = focus
manifested = shown
characterized = characterised
emphasizing = emphasising
organizations = organisations
recognizing = Recognising
journey = activity
fueled = strengthened
In the realm of = In terms of
political journey = political experience

Alongside the rules, ChatGPT was provided with detailed biographical information about various individuals whose details are publicly available online. This data was vital for the model to create accurate and meaningful content, offering a comprehensive understanding of each person's career milestones, achievements and significant roles. With this rich background, ChatGPT could craft biographies that accurately reflected the individual's contributions and accomplishments. This information was key in helping the model seamlessly integrate specific details into a well-structured and compliant narrative.

A scoring rubic was designed to measure ChatGPT's compliance with the rules listed below:

1. Word Count - Ensure the biography is concise, aiming for 200-250 words.
2. Biography Structure - Follow the specified structure strictly.
3. Biography Style - Maintain a neutral, professional tone with British English spelling and grammar.
4. Use of Language - Avoid the prohibited words and phrases. Ensure clarity and coherence.

To evaluate the generated biography based on the above criteria, any deviation from the provided rules are examined and a score to each criterion is assigned as per the table 1.

3.3 Iterations

3.3.1 Iteration 1

The evaluation begun by designing a prompt with clear guidelines designed to shape the content's length, structural organisation, stylistic tone and language precision. The prompt also included comprehensive biographical data on a person (with available information online), ensuring that the model had access to an extensive knowledge base for generating meaningful and detailed biographies. The prompts were presented in an unstructured format to evaluate ChatGPT's interpretive abilities and adherence to the specified rules. In this iteration testing involved submitting prompts with these detailed instructions at regular intervals over 24 hours.

Results showed that while the model could generate generally accurate and coherent content, it often deviated from the rules. Common issues included inconsistent word count, unintended inclusion of dates, and variations in tone and structure, as shown in table 2, all of which impacted the clarity and professionalism of the output. These findings highlight the model's limitations in strictly following complex instructions, emphasising the need for further tuning and refinement. Given the low quality of the output at this stage, the human review is needed to ensure high-quality and rule-compliant text, especially in contexts where precision and adherence to specific standards are critical.

3.3.2 Iteration 2

In this iteration, adjustments were made to improve ChatGPT's adherence to specific guidelines, addressing earlier issues with word count, date inclusion and the use of prohibited words. The rules were reorganised into distinct categories with clear titles, aiming to

Table 1: Criteria Score.

	Word Count
5	The biography is within the exact 200-250 word range
4	The biography is slightly outside the range, either slightly below or above
3	The biography is notably outside the range, but still fairly close
2	The biography is well outside the range, requiring significant revision
1	The biography is completely off, either too short or excessively long
	Biography Structure
5	The structure follows the prescribed order (Significant Career Achievements, Political Experience, Community Role, Political Interests) without any deviations
4	The structure is mostly correct but has minor overlaps or areas that could be clearer
3	The structure is somewhat followed but with noticeable deviations or unclear between sections
2	The structure is poorly followed, with significant organisational issues
1	The structure is not followed at all, with a chaotic or unorganised presentation
	Biography Style
5	The tone is perfectly neutral and professional, with correct British English spelling and grammar throughout
4	The tone is mostly neutral but has minor deviations; British English is mostly used correctly
3	The tone is somewhat neutral but occasionally strays into a more promotional or emotional style; some British English errors are present
2	The tone frequently deviates from the neutral standard; noticeable errors in British English
1	The tone is entirely inappropriate for the context; significant grammatical or spelling issues
	Use of Language
5	No prohibited words or phrases are used; language is clear, concise, and coherent
4	Mostly avoids prohibited words, with minor issues in clarity or phrasing
3	Some prohibited words or phrases are present; clarity or coherence is affected in some areas
2	Multiple instances of prohibited language; significant issues with clarity or coherence
1	Frequent use of prohibited language; the text is unclear or incoherent

enhance clarity and help the model better interpret instructions. This restructuring focused on guiding the model toward more precise, consistent responses by distinctly outlining requirements such as word count, content inclusion and style.

The model's performance was tested by generating responses at four different times daily — 10 a.m., 2 p.m., 10 p.m. and 2 a.m. — to assess consistency across intervals. The rules were categorised into four groups as described in section 3.1.2. This division aimed to evaluate the model's ability to understand and apply different rule combinations. This approach helped determine the model's adherence to instructions and whether performance varied by time of day.

Table 2: Iteration 1 Scoring Results.

Time of Day	Word Count	Biography Structure	Biography Style	Use of Language
7:00 AM	3	2	4	3
8:00 AM	3	2	3	3
9:00 AM	3	2	4	3
10:00 AM	3	2	3	3
11:00 AM	4	2	3	4
12:00 PM	3	3	3	3
1:00 PM	3	2	4	3
2:00 PM	3	2	4	3
3:00 PM	3	2	3	3
4:00 PM	3	2	3	3
5:00 PM	3	2	3	3
6:00 PM	4	3	3	4
7:00 PM	3	2	3	3
8:00 PM	3	3	3	3
9:00 PM	3	4	3	3
10:00 PM	3	2	3	3
11:00 PM	3	3	4	3
12:00 AM	3	2	3	3
1:00 AM	3	3	3	4
2:00 AM	3	2	3	3
3:00 AM	3	2	3	3
4:00 AM	3	3	3	4
5:00 AM	4	3	4	3
6:00 AM	3	2	4	4

The results presented in table 3 show significant improvements in ChatGPT's ability to generate biographical texts according to specified guidelines. Key advancements included precise adherence to word count limits, successful exclusion of dates, and the avoidance of prohibited language. This suggests that clear rule-setting aids the model in producing concise, rule-compliant responses. However, challenges persisted with biographical structure and style. The model showed inconsistencies in following the prescribed organisational sequence, which sometimes resulted in disjointed content flow. Similarly, although generally maintaining a neutral tone, the model occasionally deviated into informal language, affecting the intended formality. These findings suggest that while straightforward content rules are well-executed, complex organisational and stylistic guidelines require further refinement for full compliance. This progress highlights both the model's responsiveness to explicit instructions and the need for future iterations to strengthen its capacity for nuanced and coherent biographical writing.

3.3.3 Iteration 3

In this iteration, ChatGPT showed significant improvement in generating biographical text that adhered to the specified criteria. Refining the guidelines and structuring them with specific titles enhanced the model's accuracy and consistency. Responses were

Table 3: Iteration 2 - results.

Rule Set	Time of Day	Word Count				Bio Structure				Bio Style				Use of Language			
		Day				Day				Day				Day			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	10AM	5	5	5	5												
2	10AM	5	5	5	5	3	3	3	3								
3	10AM	5	5	5	5	3	3	4	3	4	4	3	4				
4	10AM	5	5	5	5	3	3	4	4	3	4	4	3	5	5	5	5
1	2PM	5	5	5	5												
2	2PM	5	5	5	5	3	3	3	3								
3	2PM	5	5	5	5	4	3	3	3	4	4	4	4				
4	2PM	5	5	5	5	3	3	3	3	4	4	4	4	5	5	5	5
1	10PM	5	5	5	5												
2	10PM	5	5	5	5	3	3	3	3								
3	10PM	5	5	5	5	4	3	3	3	4	3	3	4				
4	10PM	5	5	5	5	3	3	3	3	4	4	4	3	5	5	5	5
1	2AM	5	5	5	5												
2	2AM	5	5	5	5	3	3	3	3								
3	2AM	5	5	5	5	4	4	4	3	3	4	3	3				
4	2AM	5	5	5	5	3	3	3	4	4	4	4	4	5	5	5	5

generated four times daily at — 10 am, 2 pm, 10 pm, and 2 am — enabling ongoing evaluation and adjustment of the model's performance.

The model consistently maintained the specified word count range (200-250 words), a notable improvement from earlier versions, showcasing its capacity for delivering concise and balanced summaries. Additionally, the biography followed a structured formatalso defined by the rules.

The model's ability to maintain a neutral and formal tone throughout this iteration was crucial, especially for academic and professional contexts, as it avoided any subjective language or biases. Compliance with specific language rules, such as avoiding prohibited terms, further enhanced the formal quality of the text, making it both accessible and credible. These advancements underline the importance of iterative refinement in AI development, as gradual rule adjustments and testing helped identify and address previous shortcomings. This success in biographical writing demonstrates ChatGPT's potential to produce high-quality, structured, and professionally suitable content when guided by clear and detailed instructions, providing a strong foundation for AI applications in similar complex writing tasks.

Table 4: Iteration 3 - results.

Rule Set	Time of Day	Word Count				Bio Structure				Bio Style				Use of Language			
		Day				Day				Day				Day			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	10AM	5	5	5	5												
2	10AM	5	5	5	5	5	5	5	5								
3	10AM	5	5	5	5	5	5	5	5	4	5	5	5				
4	10AM	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2PM	5	5	5	5												
2	2PM	5	5	5	5	5	5	5	5								
3	2PM	5	5	5	5	5	5	5	5	5	5	5	5				
4	2PM	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	10PM	5	5	5	5												
2	10PM	5	5	5	5	5	5	5	5								
3	10PM	5	5	5	5	5	5	5	5	5	5	4	5				
4	10PM	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
1	2AM	5	5	5	5												
2	2AM	5	5	5	5	5	5	5	5								
3	2AM	5	5	5	5	5	5	5	5	5	5	5	5				
4	2AM	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5

4 RESULTS AND DISCUSSIONS

Through experiments conducted at different times and using both structured and unstructured prompts, ChatGPT's performance across varied conditions was recorded. Key insights highlight patterns in the model's strengths and weaknesses in adhering to guidelines, revealing the influence of prompt structure and temporal factors on output quality. These findings offer conclusions on ChatGPT's effectiveness in producing accurate, coherent, and rule-compliant biographical content.

4.1 Iteration Comparison

In Iteration 1, ChatGPT struggled to consistently follow the specified rules, leading to variations in quality and adherence. The model particularly faced challenges in maintaining a structured biographical format, with noticeable inconsistencies in rule application as shown in figure 1a.

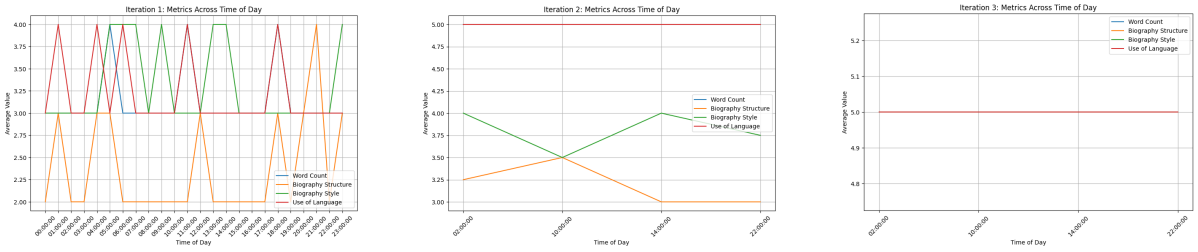
In iteration 2, following the reorganisation and clarification of the rules, ChatGPT's responses showed marked improvement. The word count aligned more closely with the specified range, and language used adhered better to the guidelines, including successful avoidance of prohibited terms. The biographical structure also improved, though some elements were inconsistently applied, indicating partial adherence as represented in 1b. While the system responded well to the revised rules, further refinement is needed to complete compliance with the desired format.

In iteration 3, the rules were reorganised for improved clarity and structure, using subheadings to separate guidelines into distinct sections and presenting instructions line by line. This format enhanced readability, reduced ambiguity, and enabled step-by-step adherence. As a result, ChatGPT responded accurately, following each rule precisely. The clearer segmentation allowed ChatGPT to interpret and implement the guidelines more efficiently, leading to smoother interactions and consistent compliance with the updated rules as it can be seen in 1c.

Figures 2 - 5 show how the results improved over the three iterations for each of the criteria used - Word Count, Biography Structure, Bigraphy Style and Use of Language.

4.2 Discussions

This research shows that ChatGPT consistently provides high-quality responses regardless of the time of day or server load, maintaining accuracy, clarity, and



(a) Iteration 1 Metrics. (b) Iteration 2 Metrics. (c) Iteration 3 Metrics.

Figure 1: Metrics for all three iterations.

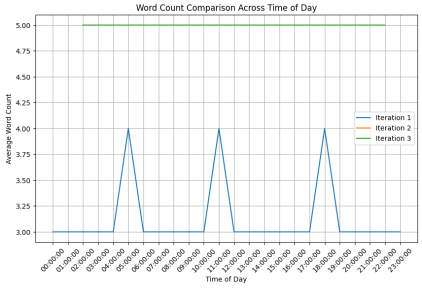


Figure 2: Word Count across the three iterations.

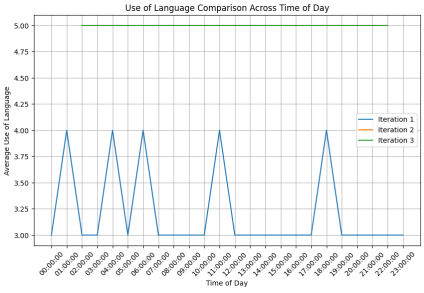


Figure 5: Use of Language across the three iterations.

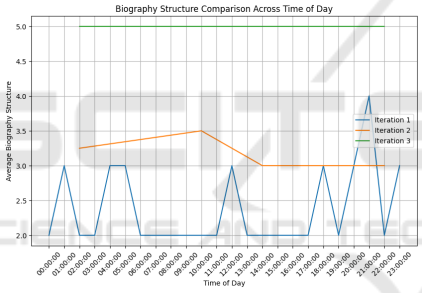


Figure 3: Biography Structure across the three iterations.

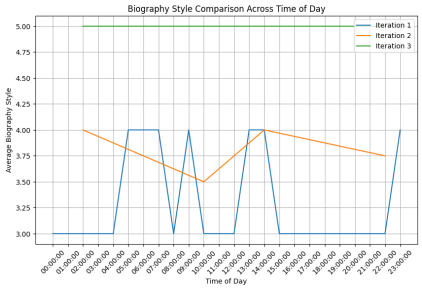


Figure 4: Biography Style across the three iterations.

relevance. However, during peak usage times, minor delays or brief pauses can occur due to high server demand.

ChatGPT's performance is highly influenced by the clarity and structure of the rules provided. Disorganised rules lead sometimes to missed or overlooked details and less accurate responses, whilst clear, step-by-step instructions significantly improved accuracy and consistency.

A set of human-written biographies was compared with those generated by ChatGPT. Whilst the AI-generated outputs are very similar in structure, human-written biographies have a feel of personal, community and public service aspects, providing a detailed, narrative-driven overview of the individual's career progression and values. In contrast, ChatGPT'sgenerated biographies focus on professional accomplishments in a more formal tone, summarising career milestones but lacking emotional depth.

This research highlighted ChatGPT's capacity for adaptive learning through repeated exposure to instructions. Initially, the model struggled to follow some rules, but its accuracy improved over time with continued repetition.

5 CONCLUSION

This research aimed to evaluate ChatGPT's performance on text generation, using biography writing as a case study. The outcome shows the importance of structured, clear instructions in achieving high-quality, consistent outputs. Initially, dense and unstructured rules led to inconsistent results, revealing that ChatGPT struggled with complex instructions that lacked clarity. By reorganising rules into well-defined subtopics with explicit instructions, the model's performance improved significantly, producing texts that met accuracy and consistency standards.

Key findings highlighted that with clear, well-structured guidance, ChatGPT can effectively follow detailed directives and adapt over time, showing potential as a robust tool for rule-compliant text generation. In this paper the authors showed that a thoughtful prompt design is essential for maximising ChatGPT's capabilities. Future efforts should focus on refining prompt structures to further enhance the model's reliability and adaptability.

ACKNOWLEDGEMENTS

We would like to thank Mapolitical Ltd for providing us with the rules and biographical texts essential for validating this study.

REFERENCES

Ali, D., Fatemi, Y., Boskabadi, E., Nikfar, M., Ugwuoke, J., and Ali, H. (2024). ChatGPT in teaching and learning: A systematic review. *Educ. Sci. (Basel)*, 14(6):643.

Aslam, A. and Curry, E. (2021). Investigating response time and accuracy in online classifier learning for multimedia publish-subscribe systems. *Multimedia Tools and Applications*, 80(9):13021–13057.

Bansal, G., Chamola, V., Hussain, A., Guizani, M., and Niyato, D. (2024). Transforming conversations with AI—A comprehensive study of ChatGPT. *Cognit. Comput.*, 16(5):2487–2510.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, New York, NY, USA. ACM.

Brown, T. B. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Crow, T., Luxton-Reilly, A., and Wuensche, B. (2018). Intelligent tutoring systems for programming education: a systematic review. In *Proceedings of the 20th Australasian Computing Education Conference*, pages 53–62.

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial intelligence in education promises and implications for teaching and learning*. Center for Curriculum Redesign.

Howarth, J. (2024). Number of parameters in gpt-4 (latest data).

Htet, A., Liana, S. R., Aung, T., and Bhaumik, A. (2024). Chatgpt in content creation: Techniques, applications, and ethical implications. In *Advanced Applications of Generative AI and Natural Language Processing Models*, pages 43–68. IGI Global.

Jiang, M., Zhao, Q., Li, J., Wang, F., He, T., Cheng, X., Yang, B. X., Ho, G. W., and Fu, G. (2024). A generic review of integrating artificial intelligence in cognitive behavioral therapy. *arXiv preprint arXiv:2407.19422*.

Jurafsky, D. (2000). Speech and language processing.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Kshetri, N. e. a. (2023). "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642.

Madunić, J. and Sovulj, M. (2024). Application of ChatGPT in information literacy instructional design. *Publications*, 12(2):11.

Oh, D.-Y. e. a. (2022). Durvalumab plus gemcitabine and cisplatin in advanced biliary tract cancer. *NEJM evidence*, 1(8).

Rao, A., Kim, J., Kamineni, M., Pang, M., Lie, W., and Succi, M. D. (2023). Evaluating chatgpt as an adjunct for radiologic decision-making. *MedRxiv*, pages 2023–02.

Rashid, M. M., Atilgan, N., Dobres, J., Day, S., Penkova, V., Küçük, M., Clapp, S. R., and Sawyer, B. D. (2024). Humanizing AI in education: A readability comparison of LLM and human-created educational content. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*

Reither, E. and Dale, R. (2000). Building natural language generation system.

Roll, I. and Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International journal of artificial intelligence in education*, 26:582–599.

Rudolph, J., Tan, S., and Tan, S. (2023). Chatgpt: Bullshit spewer or the end of traditional assessments in higher education? *Journal of applied learning and teaching*, 6(1):342–363.

Saleem, M. and Khan, Z. (2023). Healthcare simulation: An effective way of learning in health care. *Pakistan Journal of Medical Sciences*, 39(4):1185.

Sarrion, E. (2023). *Exploring the power of ChatGPT*. Apress, Berkeley, CA.

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., and Olson, C. B. (2024). Comparing the quality of human and chatgpt feedback of students' writing. *Learning and Instruction*, 91:101894.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., and Tang, Y. (2023). A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Xie, Z., Evangelopoulos, X., Omar, Ö. H., Troisi, A., Cooper, A. I., and Chen, L. (2024). Fine-tuning GPT-3 for machine learning electronic and functional properties of organic molecules. *Chem. Sci.*, 15(2):500–510.