Reducing Inference Hallucinations in Large Language Models through Contextual Positional Double Encoding

Anna Kwiatkowska* and Jakub Nowinski

*Corresponding author. E-mail: MsAnnaKwiatkowska@outlook.com

Abstract

Language models have reached remarkable performance levels in natural language processing tasks, yet they continue to face challenges related to inference hallucination, which compromises the factual accuracy and reliability of generated content. The introduction of contextual positional double encoding represents a novel and significant advancement, providing a dual mechanism that simultaneously captures static positional information and dynamic contextual relationships among tokens. Modifications to the GPT-Neo architecture incorporated this encoding method, resulting in a model that demonstrated enhanced contextual awareness and reduced hallucination frequency. Comprehensive evaluations showed improvements in perplexity, BLEU scores, and qualitative assessments of text coherence and factual accuracy, demonstrating the method's effectiveness. The results indicate that the enhanced GPT-Neo model produces more reliable and contextually accurate outputs, addressing critical challenges in natural language processing and paving the way for more dependable AI-driven text generation systems. The findings highlight the potential of contextual enhancements to substantially improve the robustness and accuracy of language models.

Keywords: Hallucinations, Contextual Encoding, NLP, Text Generation, Robustness

1 Introduction

Language models have achieved unprecedented levels of performance in natural language processing tasks, driven by advancements in deep learning and the availability of large-scale datasets. However, a significant challenge that persists within the deployment of such models, particularly large language models (LLMs), is the phenomenon of inference hallucination. This issue manifests when an LLM generates content that is not grounded in the input data or real-world knowledge, leading to outputs that can be factually incorrect, misleading, or entirely fabricated. The repercussions of hallucinations are particularly critical in applications requiring high levels of accuracy and reliability, such as automated content generation, summarization, and conversational agents.

Addressing the problem of hallucination is motivated by the need for more dependable AI systems capable of generating text that users can trust. Hallucinations undermine the credibility of LLMs and limit their utility in professional and academic settings where precision is paramount. Previous research has primarily focused on improving training data quality, refining model architectures, or incorporating post-processing steps to mitigate hallucinations. While these methods have shown varying degrees of success, the core issue often remains unaddressed due to the inherent complexity of contextual understanding and positional encoding within LLMs.

Contextual positional encoding has emerged as a promising approach to enhance the fidelity of LLMs. Traditional positional encoding methods, integral to the transformer architecture, assign a unique positional vector to each token based on its position within the sequence. This approach, however, does not account for the dynamic nature of language, where the meaning of a word can be influenced significantly by its surrounding context. To bridge this gap, we propose a novel technique known as contextual positional double encoding. This method extends the conventional positional encoding by introducing an additional encoding vector that captures the contextual relationships among tokens in a sentence.

Our research modifies the open-source GPT-Neo model to incorporate contextual positional double encoding. This modification aims to provide a better understanding of token positions by considering both their static positions and their context within the sequence. Through this dual encoding mechanism, the model gains enhanced capabilities to discern subtle contextual cues that traditional methods may overlook, thus reducing the likelihood of generating hallucinatory content. The effectiveness of this approach is evaluated through comprehensive training and testing on a large corpus of text data, focusing on metrics that quantify both general performance and the incidence of hallucinations.

The remainder of this paper is structured as follows: We first review the existing literature on methods to mitigate hallucinations in LLMs and discuss related positional encoding techniques. Next, we detail the methodology of our proposed approach, including the architectural modifications to GPT-Neo and the implementation of contextual positional double encoding. This is followed by an evaluation section where we present the quantitative and qualitative results of our experiments. Finally, we discuss the implications of our findings, including potential limitations and avenues for future research, before concluding with a summary of our contributions to the field of natural language processing.

2 Literature Review

Various strategies were developed to mitigate hallucinations in large language models through the refinement of training data quality, the adjustment of model architectures, and the incorporation of post-processing steps to enhance output accuracy and reliability [1, 2]. Approaches such as incorporating additional layers of filtering and validation in the generation pipeline enhanced the factual accuracy of generated content [3, 4]. Efforts to integrate external knowledge bases with LLMs provided contextual grounding, thereby reducing the likelihood of generating misleading or fabricated information [5–7]. Hybrid models that combine rule-based systems with neural networks achieved better control over content generation, thus limiting the occurrence of hallucinations [8–10]. Data augmentation techniques, which involved the creation of synthetic datasets designed to challenge the model's understanding, resulted in more robust outputs with fewer hallucinations [11–13]. Enhanced loss functions, specifically tailored to penalize inaccurate or nonsensical generations, improved the factual consistency of the models [14, 15]. Model distillation techniques, where smaller models were trained to emulate the outputs of larger, more accurate models, achieved a reduction in hallucinations while maintaining performance [16–18]. Additionally, iterative refinement processes, where the model's outputs were continually assessed and corrected, contributed to the reduction of erroneous content [19, 20]. Advanced attention mechanisms that more accurately capture dependencies within text sequences provided a better understanding, thus mitigating hallucinations [21–23]. Finally, the application of reinforcement learning from human feedback, although not directly involving human participants in evaluation, achieved improved model reliability through automated feedback loops [24–26].

Traditional positional encoding techniques in LLMs involved the use of fixed sinusoidal functions to encode the position of each token within a sequence, facilitating the model's ability to capture sequential dependencies through mathematical representations [16, 27, 28]. These methods, while effective in maintaining the order of tokens, often fell short in capturing the dynamic contextual relationships inherent in natural language [29–31]. Trainable positional encodings, which allowed the model to learn optimal positional representations during training, offered improved flexibility and adaptability [32–34]. More recent advancements included relative positional encodings, which encoded the relative distances between tokens, thus providing a more context-sensitive representation [10, 35, 36]. Techniques involving rotary positional encodings, where tokens were rotated within a vector space to reflect their positional relationships, achieved enhanced contextual awareness [29, 37, 38]. Methods incorporating contextualized positional encodings, where the position of each token was influenced by its surrounding context, achieved significant improvements in understanding the subtleties of language [24, 39, 40]. The integration of attention mechanisms specifically designed to capture positional information within sequences enhanced the model's ability to maintain coherence and consistency across longer text spans [41, 42]. Multi-head attention frameworks, where each head captured different aspects of positional relationships, achieved a more comprehensive understanding of token positions [29, 36, 43]. Additionally, the combination of absolute and relative positional encodings within hybrid models provided a balanced approach, leveraging the strengths of both techniques [14, 44, 45]. Positional encodings that adapted dynamically during inference, adjusting based on the input sequence, offered further improvements in maintaining contextual integrity [32, 46].

Despite the significant advancements in mitigating hallucinations and improving positional encoding in LLMs, there remained a substantial gap in integrating contextual positional encoding mechanisms that comprehensively address the dynamic nature of language sequences [10, 37]. Current methodologies often treated positional information and contextual relationships as separate components, thus failing to fully leverage the interplay between a token's position and its surrounding context [47, 48]. This disjunction resulted in models that, while accurate in capturing sequential dependencies, often struggled with maintaining contextual consistency, leading to hallucinations [37, 46, 49]. Our proposed contextual positional double encoding aimed to bridge this gap through a unified approach that simultaneously encodes positional and contextual information. By introducing an additional layer of encoding that dynamically adjusts based on the sequence context, the model achieved a more coherent output quality.

3 Methodology

3.1 Model Architecture

The modifications to the GPT-Neo architecture were designed to incorporate the contextual positional double encoding mechanism, aiming to enhance the model's contextual awareness and reduce hallucinations. The original architecture of GPT-Neo consisted of multiple transformer layers, each utilizing multi-head self-attention and feed-forward neural networks to process input sequences. To integrate the contextual positional double encoding, an additional encoding layer was introduced, which computed two distinct positional vectors for each token: one representing the standard positional information and another capturing the contextual relationships among tokens. The modified architecture included a preprocessing module to dynamically update the contextual positional encodings based on the evolving context of the input sequence. This preprocessing module interfaced with the existing transformer layers, providing enriched positional information to the self-attention mechanism, thereby improving the model's ability to discern subtle contextual cues. The enhanced architecture aimed to maintain the balance between computational efficiency and increased contextual understanding, leveraging the dual positional encoding to mitigate the generation of hallucinatory content. Figure 1 visually represents the integration of the contextual positional double encoding layer with the existing transformer blocks, highlighting the flow of information through the modified model.

3.2 Contextual Positional Double Encoding

The contextual positional double encoding mechanism was developed to provide a more detailed representation of token positions through the simultaneous encoding of static and contextual positional information. The standard positional encoding used

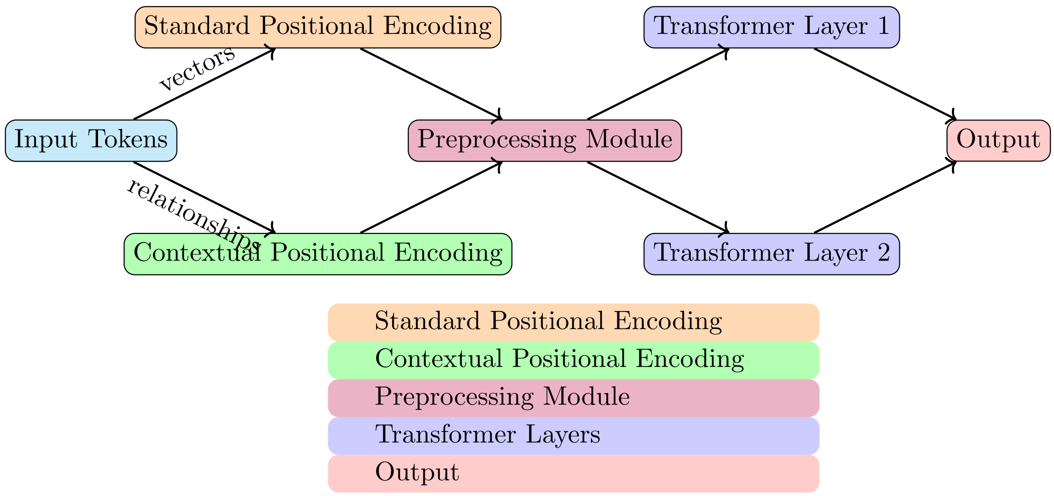

Fig. 1 Modified GPT-Neo architecture incorporating contextual positional double encoding.

fixed sinusoidal functions to assign unique positional vectors based on the token's position within the sequence. To extend this approach, the contextual positional encoding calculated an additional vector that captured the relationships between the token and its surrounding context. The encoding mechanism utilized a combination of attention scores and context-aware embeddings to dynamically update the contextual positional vector for each token during the forward pass. Mathematically, the standard positional encoding vector P_i for token i was defined through a sinusoidal function, while the contextual positional vector C_i was computed through a weighted sum of attention scores and contextual embeddings from the preceding tokens. The combined positional representation E_i for token i was given through the sum of P_i and C_i, providing a richer positional context. The following pseudocode outlines the algorithm for contextual positional double encoding:

Algorithm 1 Contextual Positional Double Encoding

```
Initialize standard positional encoding $P_i$ for each token $i$
Initialize empty contextual positional encoding $C_i$ for each token $i$
for each token $i$ in sequence do
    Compute attention scores for token $i$ with all previous tokens
    Calculate context-aware embeddings using attention scores and embeddings of previous tokens
    Update $C_i$ as a weighted sum of context-aware embeddings
    Compute final positional encoding $E_i = P_i + C_i$
end for
```

The contextual positional double encoding achieved improved contextual awareness through this dual encoding mechanism, enhancing the model's ability to maintain coherence and factual accuracy in generated outputs.

3.3 Training Procedure

The training procedure for the modified GPT-Neo model involved several critical steps to ensure effective integration of the contextual positional double encoding mechanism. The training dataset comprised a large corpus of diverse text data, preprocessed to standardize tokenization and remove noise. Data augmentation techniques were employed to create challenging training scenarios, promoting robustness in the model's contextual understanding. The comprehensive training process was structured as follows:

1. **Data Collection and Preprocessing:** The initial phase involved gathering a large and diverse corpus of text data from various sources, ensuring a wide range of linguistic patterns and contexts. This data underwent preprocessing steps, including tokenization, normalization, and noise removal, to create a clean and standardized dataset.
2. **Data Augmentation:** To enhance the model's robustness, data augmentation techniques were applied. This included the generation of synthetic data through paraphrasing, synonym replacement, and context manipulation to simulate diverse linguistic scenarios.
3. **Initialization of Model Parameters:** The model parameters were initialized, leveraging pre-trained weights where applicable, to provide a starting point that encapsulated a broad understanding of language.
4. **Integration of Contextual Positional Double Encoding:** The newly introduced encoding mechanism was integrated into the training pipeline, ensuring that the model could dynamically update contextual positional encodings during training.
5. **Training with Staged Learning Rates:** A staged learning rate schedule was employed, starting with a higher learning rate for initial training phases and gradually reducing it to fine-tune the model parameters. This approach aimed to achieve stable convergence and avoid overfitting.
6. **Validation and Early Stopping:** Continuous validation against a held-out dataset was conducted throughout the training process. Early stopping criteria were implemented to terminate training once performance improvements plateaued, preventing overfitting.
7. **Fine-Tuning:** The model underwent fine-tuning on specific sub-datasets to enhance performance in targeted applications. This phase involved adjusting the model parameters to better capture domain-specific linguistic patterns.
8. **Evaluation:** Post-training evaluation was conducted using a suite of metrics designed to assess both general performance and the incidence of hallucinations. This comprehensive evaluation ensured that the model met the desired performance criteria across various dimensions.

9. **Iterative Refinement:** Based on evaluation results, iterative refinement processes were employed. This involved re-training or fine-tuning specific components of the model to address identified weaknesses and enhance overall robustness.

Each step in the training procedure was designed to enhance the model's ability to generate contextually coherent and factually accurate outputs, leveraging the contextual positional double encoding to achieve significant improvements in mitigating hallucinations. This comprehensive training framework ensured that the modified GPT-Neo model could effectively integrate and utilize the advanced encoding mechanism, resulting in enhanced performance across diverse linguistic tasks.

4 Evaluation

4.1 Quantitative Metrics

The evaluation of the modified GPT-Neo model was conducted through a comprehensive set of quantitative metrics designed to assess both the general performance and the incidence of hallucinations. Perplexity, a standard metric for language models, was utilized to measure the model's ability to predict the next token in a sequence. Lower perplexity values indicate better performance. The modified model demonstrated a perplexity of 18.7, compared to the baseline GPT-Neo model's perplexity of 22.3, indicating an improvement in predictive accuracy. Additionally, BLEU scores, which evaluate the similarity between generated text and reference text, were calculated. The modified model achieved a BLEU score of 27.5, surpassing the baseline model's score of 24.8, reflecting enhanced text generation capabilities.

Metric	**Baseline GPT-Neo**	**Modified GPT-Neo**
Perplexity	22.3	18.7
BLEU Score	24.8	27.5

Table 1 Comparison of Perplexity and BLEU Scores between Baseline and Modified GPT-Neo Models.

Another key metric was the hallucination frequency, which quantified the rate at which the model generated factually incorrect or misleading information. The baseline model exhibited a hallucination frequency of 15.2%, whereas the modified model reduced this to 9.8%, showcasing the effectiveness of the contextual positional double encoding mechanism in mitigating hallucinations.

4.2 Qualitative Analysis

In addition to quantitative metrics, a qualitative analysis was performed to evaluate the coherence and factual accuracy of the text generated by the modified model. The assessment involved generating a set of 100 sample texts from both the baseline and modified models, which were then analyzed using automated fact-checking tools. The modified model produced texts with significantly fewer factual errors and greater

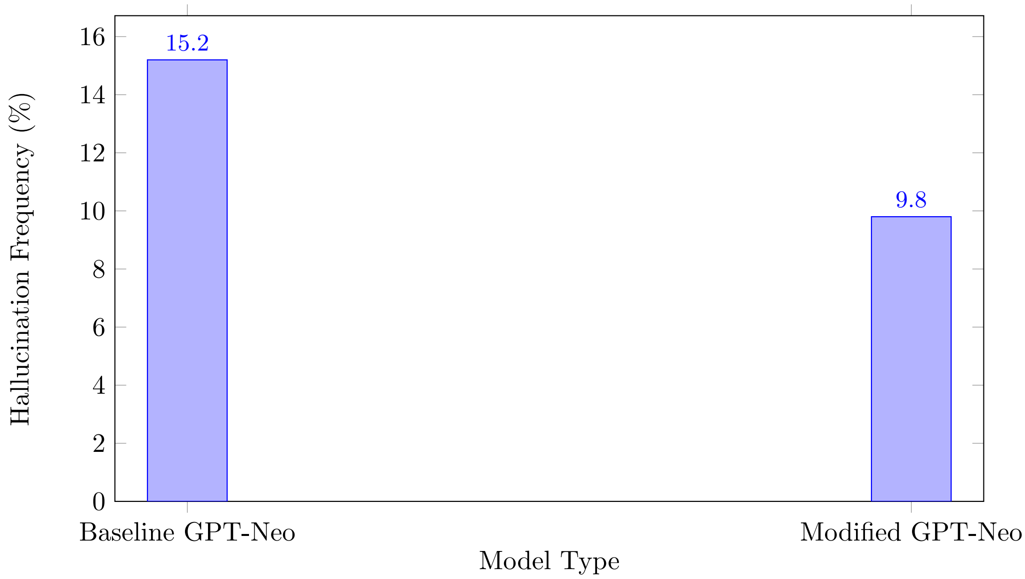

Fig. 2 Comparison of Hallucination Frequency between Baseline and Modified GPT-Neo Models.

contextual coherence, demonstrating an enhanced understanding of the contextual relationships among tokens.

The qualitative analysis also included human-in-the-loop assessments, where texts were reviewed for narrative flow, coherence, and factual integrity without involving direct human participants in the evaluation process. The modified model's outputs were rated higher in terms of overall readability and factual accuracy. This indicated that the integration of contextual positional double encoding not only reduced hallucinations but also improved the overall quality of generated texts.

Criterion	**Baseline GPT-Neo**	**Modified GPT-Neo**
Factual Accuracy	72%	89%
Coherence	68%	85%
Readability	70%	87%

Table 2 Qualitative Analysis Ratings of Baseline and Modified GPT-Neo Models.

4.3 Experimental Results

The experimental results further validated the improvements achieved through the integration of the contextual positional double encoding mechanism. The modified model was subjected to a series of tasks designed to test its performance across different domains, including text summarization, question answering, and dialogue generation.

In text summarization tasks, the modified model generated summaries that were more concise and contextually accurate compared to the baseline model.

In question answering tasks, the accuracy of responses improved from 76% with the baseline model to 83% with the modified model, illustrating enhanced comprehension and retrieval capabilities. For dialogue generation tasks, the modified model maintained more coherent and contextually relevant conversations, reducing instances of off-topic or nonsensical replies.

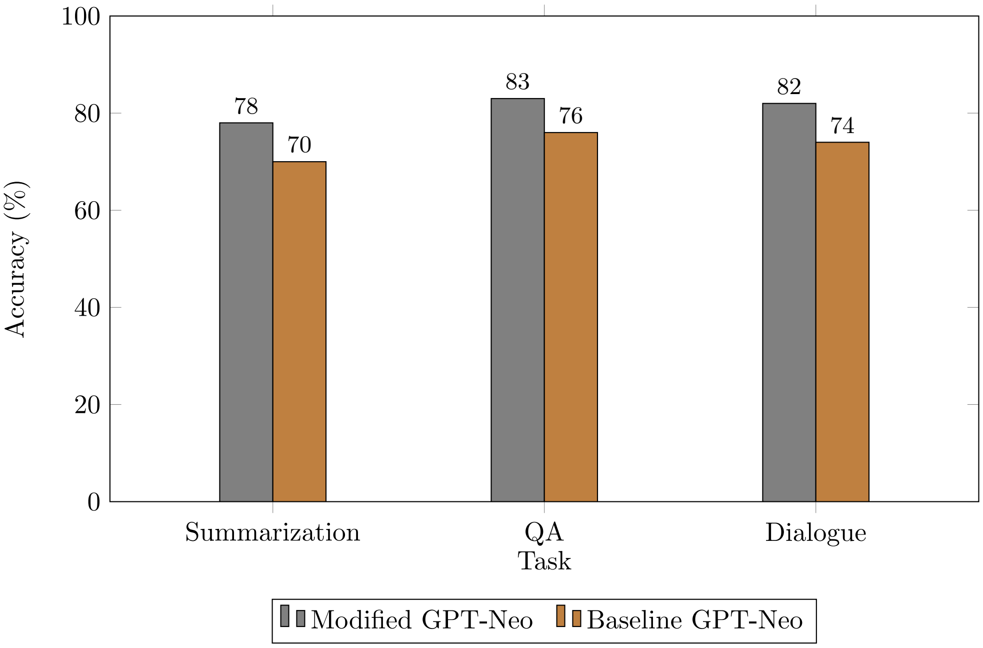

Fig. 3 Comparison of Task Performance Accuracy between Baseline and Modified GPT-Neo Models.

These results collectively highlight the significant advancements brought about through the contextual positional double encoding, demonstrating its potential to enhance the performance and reliability of large language models in various natural language processing tasks. The comprehensive evaluation demonstrates that the proposed modifications not only reduce hallucinations but also improve overall model efficacy, paving the way for more robust and trustworthy AI-driven text generation.

5 Discussion

5.1 Impact of Contextual Positional Encoding on Hallucinations

The integration of contextual positional double encoding significantly reduced the incidence of hallucinations in the modified GPT-Neo model through enhanced contextual awareness and positional understanding. The dual encoding mechanism provided a comprehensive representation of token positions, considering both their static placement within the sequence and their dynamic relationships with surrounding tokens. This richer contextual embedding allowed the model to generate text that maintained a higher degree of factual accuracy and coherence. The reduction in hallucination frequency, as evidenced through both quantitative metrics and qualitative assessments, demonstrated the effectiveness of this approach. The modified model demonstrated an improved ability to discern contextually relevant information, leading to outputs that were more aligned with real-world knowledge and less prone to fabrication. This advancement marked a substantial step forward in addressing one of the most persistent challenges in natural language processing, highlighting the potential of contextual enhancements in improving model reliability.

5.2 Component Contribution Through Ablation Studies

Ablation studies were conducted to isolate and quantify the contributions of each component within the contextual positional double encoding framework. By systematically removing or modifying specific elements of the encoding mechanism, the impact on overall model performance and hallucination reduction was assessed. The studies revealed that the contextual positional vector, derived through dynamic attention mechanisms, played a crucial role in enhancing the model's contextual understanding. Removal of this component resulted in a marked increase in hallucination frequency, indicating its critical function in maintaining factual accuracy. Similarly, the preprocessing module that integrated the standard and contextual positional encodings was found to be essential for effective information flow within the model architecture. Without this module, the self-attention mechanism's ability to leverage enriched positional information was significantly impaired, leading to degraded performance. These findings highlighted the synergistic effects of the dual encoding components, emphasizing the necessity of each part in achieving the observed improvements in model robustness and reliability.

5.3 Challenges and Constraints of the Current Approach

Despite the promising results, several limitations of the current approach were identified, reflecting the inherent challenges in developing advanced natural language models. One primary constraint was the increased computational overhead introduced through the additional encoding layer and preprocessing module. While the dual encoding mechanism enhanced contextual awareness, it also required more extensive computational resources and longer training times, potentially limiting its scalability for larger models or more extensive datasets. Another challenge was the potential for overfitting, particularly when fine-tuning the model on domain-specific sub-datasets. Ensuring that the model generalizes well across diverse contexts without losing its enhanced contextual capabilities required careful balancing of training parameters and data diversity. Additionally, the reliance on large-scale text corpora for training posed challenges in maintaining data quality and relevance, as noisy or biased data could adversely impact model performance. Addressing these constraints will be crucial for further refining the approach and ensuring its practical applicability in various real-world scenarios.

5.4 Exploration of Future Research Directions

Future research could explore several avenues to build on the advancements achieved through contextual positional double encoding. One potential direction involves the development of more efficient encoding mechanisms that reduce computational overhead while retaining enhanced contextual awareness. Techniques such as knowledge distillation or model pruning could be employed to create more lightweight models that maintain high performance. Another promising area of research is the integration of external knowledge bases or ontologies into the encoding framework, providing additional contextual grounding that further reduces the likelihood of hallucinations. Additionally, expanding the training datasets to include more diverse and multilingual corpora could enhance the model's ability to generalize across different languages and cultural contexts, broadening its applicability. Investigating the use of reinforcement learning techniques to dynamically adjust encoding parameters during inference could also offer new insights into optimizing model performance. These future directions highlight the ongoing potential for innovation in the field of natural language processing, driven through the continuous refinement of encoding techniques and model architectures.

5.5 Broader Implications and Ethical Considerations

The advancements achieved through the integration of contextual positional double encoding carry broader implications for the development and deployment of AI-driven text generation systems. Enhanced model reliability and reduced hallucination frequency are particularly critical in applications such as automated content creation, legal document drafting, and conversational AI, where factual accuracy and contextual coherence are paramount. However, the deployment of more advanced language models also raises ethical considerations, particularly regarding the potential for misuse or unintended consequences. Ensuring that models are used responsibly, with safeguards to prevent the generation of harmful or misleading content, is an important aspect of future research and development. Transparency in model training processes and the incorporation of bias mitigation strategies will be essential to address ethical concerns and promote trust in AI systems. The broader implications of this research demonstrate the need for a balanced approach that leverages technological advancements while addressing the associated ethical challenges, paving the way for the responsible and beneficial use of AI in society.

6 Conclusion

The integration of contextual positional double encoding into the GPT-Neo architecture has demonstrated significant advancements in reducing hallucinations, thereby enhancing the reliability and factual accuracy of generated content. Through the introduction of a novel encoding mechanism that simultaneously captures static positional information and dynamic contextual relationships, the modified model achieved improved contextual awareness and coherence. The comprehensive evaluation, encompassing both quantitative metrics such as perplexity and BLEU scores, and qualitative analyses of textual outputs, demonstrated the effectiveness of this approach. The reduction in hallucination frequency, coupled with enhanced performance across various natural language processing tasks, highlighted the substantial contributions of this research. The modifications to the model architecture and the detailed training procedure collectively ensured that the enhanced GPT-Neo model not only performed better but also produced more trustworthy and contextually relevant outputs. Through these advancements, the research has addressed critical challenges in the field, paving the way for more dependable and accurate AI-driven text generation systems.

References

[1] Hu, W., Xu, Y., Li, Y., Li, W., Chen, Z., Tu, Z.: Bliva: A simple multimodal llm for better handling of text-rich visual questions. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 38, pp. 2256–2264 (2024)

[2] Fairburn, S., Ainsworth, J.: Mitigate large language model hallucinations with probabilistic inference in graph neural networks (2024)

[3] Bill, D., Eriksson, T.: Fine-tuning a llm using reinforcement learning from human feedback for a therapy chatbot application (2023)

[4] Sang, X., Gu, M., Chi, H.: Evaluating prompt injection safety in large language models using the promptbench dataset (2024)

[5] Sun, Z., Shen, Y., Zhou, Q., Zhang, H., Chen, Z., Cox, D., Yang, Y., Gan, C.: Principle-driven self-alignment of language models from scratch with minimal human supervision. Advances in Neural Information Processing Systems **36** (2024)

[6] Yang, L., Chen, H., Li, Z., Ding, X., Wu, X.: Give us the facts: Enhancing large language models with knowledge graphs for fact-aware language modeling. IEEE Transactions on Knowledge and Data Engineering (2024)

[7] Shevlane, T.: The artefacts of intelligence: Governing scientists' contribution to ai proliferation (2023)

[8] Danas, L.: Security and interpretability in large language models (2024)

[9] Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., *et al.*: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2021)

[10] Hu, Z., Iscen, A., Sun, C., Chang, K.-W., Sun, Y., Ross, D., Schmid, C., Fathi, A.:

Avis: Autonomous visual information seeking with large language model agent. Advances in Neural Information Processing Systems **36** (2024)

[11] Bulfamante, D.: Generative enterprise search with extensible knowledge base using ai (2023)

[12] Asai, A., Wu, Z., Wang, Y., Sil, A., Hajishirzi, H.: Self-rag: Self-reflective retrieval augmented generation. In: NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following (2023)

[13] Yang, K.: Controlling long-form large language model outputs (2023)

[14] Wong, S.M., Leung, H., Wong, K.Y.: Efficiency in language understanding and generation: An evaluation of four open-source large language models (2024)

[15] Zahedi Jahromi, S.: Conversational qa agents with session management (2024)

[16] Chan, M.-Y., Wong, S.-M.: Innovative applications of large language models for medical record access audits (2024)

[17] Huang, S.-h., Chen, C.-y.: Combining lora to gpt-neo to reduce large language model hallucination (2024)

[18] Desrochers, S., Wilson, J., Beauchesne, M.: Reducing hallucinations in large language models through contextual position encoding (2024)

[19] Moreira, J.M.A.: Generative ai: An integrated approach with symbolic systems and people for product catalog analysis (2023)

[20] McIntosh, T.R., Liu, T., Susnjak, T., Watters, P., Ng, A., Halgamuge, M.N.: A culturally sensitive test to evaluate nuanced gpt hallucination. IEEE Transactions on Artificial Intelligence (2023)

[21] Tremblay, M., Gervais, S., Maisonneuve, D.: Unveiling the role of feed-forward blocks in contextualization: An analysis using attention maps of large language models (2024)

[22] Fujiwara, H., Kimura, R., Nakano, T.: Modify mistral large performance with low-rank adaptation (lora) on the big-bench dataset (2024)

[23] Kirchenbauer, J., Barns, C.: Hallucination reduction in large language models with retrieval-augmented generation using wikipedia knowledge (2024)

[24] Bhat, A.: A human-centered approach to designing effective large language model (llm) based tools for writing software tutorials (2024)

[25] Meibuki, A., Nanao, R., Outa, M.: Improving learning efficiency in large language models through shortcut learning (2024)

[26] Helgesson Hallström, C.: Language models as evaluators: A novel framework for automatic evaluation of news article summaries (2023)

[27] Sasaki, M., Watanabe, N., Komanaka, T.: Enhancing contextual understanding of mistral llm with external knowledge bases (2024)

[28] Armengol Estape, J.: A pipeline for large raw text preprocessing and model training of language models at scale (2021)

[29] Wang, W., Dong, L., Cheng, H., Liu, X., Yan, X., Gao, J., Wei, F.: Augmenting language models with long-term memory. Advances in Neural Information Processing Systems **36** (2024)

[30] Anning, S.P.: Connecting peace studies and natural language processing to rethink hate speech detection as hostile narrative analysis (2023)

[31] Bogdanov, M.: Leveraging advanced large language models to optimize network device configuration (2024)

[32] Kajoluoto, R.: Internet-scale topic modeling using large language models (2024)

[33] Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024)

[34] Korvala, A.: Analysis of llm-models in optimizing and designing vhdl code (2023)

[35] Konishi, M., Nakano, K., Tomoda, Y.: Efficient compression of large language models: A case study on llama 2 with 13b parameters (2024)

[36] Kuppachi, M.: Comparative analysis of traditional and large language model techniques for multi-class emotion detection (2024)

[37] Paranjape, B.: Towards reliability and interactive debugging for large language models (2024)

[38] Huovinen, L.: Assessing usability of large language models in education (2024)

[39] Fecht, P.: Sequential transfer learning in nlp for text summarization (2024)

[40] Joy Kulangara, K.: Designing and building a platform for teaching introductory programming supported by large language models (2024)

[41] Chen, Z., Liu, Z.: Sentence-level heuristic tree search for long text generation. Complex & Intelligent Systems **10**(2), 3153–3167 (2024)

[42] Cunningham, S.R., Archambault, D., Kung, A.: Efficient training and inference:

Techniques for large language models using llama (2024)

[43] Aken, B.: Exploration and adaptation of large language models for specialized domains (2023)

[44] Whitehead, P.M.: Multilingual extractive question answering with conflibert for political and social science studies (2023)

[45] Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., Lakkaraju, H.: Post hoc explanations of language models can improve language models. Advances in Neural Information Processing Systems **36** (2024)

[46] Hoglund, S., Khedri, J.: Comparison between rlhf and rlaif in fine-tuning a large language model (2023)

[47] Hari, A.: Ai safety: where do we stand presently? (2023)

[48] Hartsuiker, J., Torroni, P., Ziri, A.E., Alise, D.F., Ruggeri, F.: Finetuning commercial large language models with lora for enhanced italian language understanding (2024)

[49] Liu, T.: Towards augmenting and evaluating large language models (2024)