SENSITIVITY ANALYSIS OF WORD IMPORTANCE USING GPT MODEL: A RANKING XAI APPROACH WITH ATTENTION WEIGHTS AND KL DIVERGENCE

Arav Agarwal[1] and Rhea Mahajan[2]

[1]Dhirubhai Ambani International School, Mumbai, India
[2]Department of Computer Science and IT, University of Jammu, J&K, India

ABSTRACT

This paper delves into the intricate realm of generative Artificial Intelligence (AI) models, specifically focusing on transformers like GPT (Generative Pre-trained Transformer). Despite their remarkable capabilities, these models pose challenges in terms of interpretability and accountability, owing to their complex architectures and vast training data. This paper employs a model to investigate the importance of words within the corpus, employing sensitivity analysis techniques. Specifically, attention weights are used to measure the impact of individual words on the model's predictions. The paper proposes a novel approach to rank the importance of words by leveraging attention weights and conducting sensitivity analysis across the dataset. To quantify the discrepancies between model-generated outputs and ground truth, the Kullback-Leibler (KL) divergence is employed. This divergence measure aids in evaluating how well the model captures the underlying distribution of words in the corpus. By integrating KL divergence into the sensitivity analysis, the study aims to provide a more comprehensive understanding of word importance.

KEYWORDS

Artificial Intelligence; Kullback-Leibler (KL) divergence; Generative Pre-trained Transformer

1. INTRODUCTION

Over the past three years, there has been a notable surge in the popularity of generative artificial intelligence, particularly in the field of natural language processing. This surge has been nothing short of explosive, and its impact has been deeply transformative in how we interact with technology. This transformation has extended its reach into numerous aspects of our daily lives, finding its way into a plethora of applications we encounter regularly. The driving force behind this remarkable upswing can be squarely attributed to the remarkable advancements that have been made in deep learning architecture and the simultaneous escalation in computational capabilities. These two factors have converged to propel generative AI to the forefront of technological innovation, reshaping our digital landscape in the process [1].

The emergence of transformer-based models has fundamentally transformed the landscape of artificial intelligence. This pivotal development has left an indelible mark on the field. With the introduction of models such as GPT (Generative Pre-Trained Transformer,) [2] and BERT (Bidirectional Encoder Representations from Transformers) [3], created by Google, the capabilities of natural language generation have achieved unprecedented levels of sophistication. These models are frequently trained on massive datasets of text, with GPT-3 boasting a training dataset that comprises a staggering 175 billion parameters. When combined with their cutting-

edge architectural design, these models exhibit exceptional performance, yielding responses that are remarkably coherent, well-structured, and nearly indistinguishable from human-generated text.

These models have also been integrated into large-scale applications, with perhaps the most prominent example being ChatGPT. In just two months following its release, ChatGPT rapidly amassed an astounding 100 million users, underscoring the broad reach and practical utility of this technology. In addition to ChatGPT, a host of other AI models have similarly found their footing across a diverse array of commercial sectors, including customer support, e-commerce, education, and even healthcare [4]. The influence of generative AI is not confined solely to applications and websites; it has also permeated into the very fabric of operating systems like Windows.

Beyond the realm of commerce, generative AI has proven immensely valuable in creative domains. Writers, artists, and musicians are increasingly collaborating with AI systems, embarking on a journey to discover novel modes of creative expression. AI-generated art, literature, and music not only push the boundaries of human creativity but also challenge our preconceived notions of the interplay between humans and machines. This intersection between artistry and artificial intelligence is forging new frontiers in the creative landscape, redefining the limits of human-machine collaboration and expanding our horizons in innovative ways.

As an increasing number of researchers and scientists enter the burgeoning field of generative AI, it is poised to experience further advancements in the near future. Pioneering organizations like OpenAI and Google remain steadfast in their commitment to refining these models, continually striving to push them closer to perfection. However, amid this wave of progress, ethical concerns and implications have begun to surface, albeit slowly but surely. These issues encompass matters related to bias present in training data, the potential misuse of AI-generated content, and a pervasive sense of skepticism regarding the output of AI systems. These challenges can often be distilled into a broader problem: the lack of interpretability in these AI models.

This dearth of interpretability is most notably conspicuous in the transformer model, renowned for its intricate architecture featuring multiple stacks of encoders and decoders, each equipped with its own multi-headed attention layers and neural networks. Moreover, these models are trained on a vast corpus of text, drawn from a diverse range of sources, which exacerbates the interpretability issue, particularly within the context of transformer models [5]. The intricate and opaque nature of these models raises crucial questions about how decisions are made within their neural networks. It challenges our ability to understand why these AI systems generate specific responses or predictions, which is a fundamental concern in fields where accountability, transparency, and the mitigation of bias are paramount. Addressing these concerns will be essential to harnessing generative AI's full potential while ensuring its deployment aligns with ethical and societal standards.

With many more researchers and scientists entering this booming field, generative AI is poised to make further advances soon. OpenAI and Google are constantly working on developing these models and pushing them as close to perfection as possible. However, ethical questions and implications have slowly yet surely started to rise. Issues related to bias in training data, misuse of AI-generated content and a general sense of distrust with what AI produces are all being highlighted. This can be boiled down to a general lack of interpretability in these AI models.

The transformer model has a complicated architecture with multiple encoders and decoders, each with their own multi-headed attention layers and neural networks. Furthermore, these models are trained on a large corpus of text taken from a variety of sources, which only exaggerates the

problem with the transformer model specifically. Researchers have been diligently working to address the challenge of explainability and interpretability in artificial intelligence. A subfield known as "Explainable AI" (XAI) emerged about seven years ago with the primary goal of making complex AI models more understandable and interpretable for humans [6]. XAI focuses on providing insights into how AI models arrive at their decisions, enabling users to grasp the reasoning behind specific choices. One key technique used in XAI involves creating saliency maps, which highlight the critical features or portions of input data that significantly influence a model's output. For example, in image recognition, saliency maps can reveal which parts of an image the model pays the most attention to when making predictions.

Despite advancements in XAI, a substantial challenge remains in making transformer models interpretable. The intricate structure of transformers, featuring numerous layers of encoders and decoders, combined with their extensive and diverse training data, presents difficulties in generating effective and reliable explanations. This lack of interpretability raises ethical concerns and undermines the trust that users, researchers, and society as a whole can have in these powerful AI systems [7].

This research paper aims to bridge the critical gap between generative AI, particularly transformer models, and XAI methods. By developing techniques tailored specifically for transformers, the objective is to enhance the transparency and comprehensibility of these models. Through this research, a unique method to conduct sensitivity analysis on a transformer model has been proposed which makes use of the KL divergence to quantify the attention weights.

The significance of this effort extends beyond the academic realm. By enabling a deeper understanding of transformer models, we can facilitate their more responsible and ethical use in real-world applications. Building trust in these models is crucial for their acceptance and successful integration into various domains.

1.1. Research Contributions

- The proposed research investigates interpretability challenges in GPT models, proposing sensitivity analysis using attention weights to understand word importance.
- It also introduces a novel approach to rank word importance by leveraging attention weights and sensitivity analysis techniques.
- It utilizes Kullback-Leibler divergence to quantify model-generated output disparities, enhancing evaluation metrics for generative AI models by integrating it with sensitivity analysis to provide a holistic understanding of word importance and model performance.

2. BACKGROUND

Transformers have revolutionized natural language processing and have shown remarkable success in various tasks including language modelling, translation, and text generation. However, their performance can degrade significantly when applied to tasks with long sequences, primarily due to their inherent instability and sensitivity to input perturbations. To address these challenges, recent research has focused on enhancing the stability and robustness of transformer models through various techniques.

One prominent approach involves integrating stability mechanisms directly into the transformer architecture. For instance, Vaswani et al. [8] introduced the self-attention mechanism, allowing transformers to weigh the importance of different input tokens dynamically. Despite its effectiveness, self-attention has been criticized for its lack of robustness to input variations and

noise (levi et al. [9]). Consequently, researchers have explored methods to mitigate this sensitivity, such as incorporating explicit positional encoding (Shaw et al. [10] or introducing regularization techniques during training (Hua et al., [11]).

Another line of research focuses on analysing the sensitivity of transformer models to input changes and identifying factors that contribute to their instability. Pande et al. [12] conducted sensitivity analysis on transformer-based language models and revealed that certain tokens have a disproportionate influence on model predictions, leading to instability in outputs. Building upon this insight, Zhang et al. [13] proposed a sensitivity-aware training framework that selectively penalizes high-sensitivity tokens during optimization, resulting in more robust models. In the same year, Davis et al. [14] proposed Catformer, a novel framework for designing transformers via sensitivity analysis. Through extensive experiments and empirical evaluations, they demonstrated the effectiveness and versatility of Catformer in enhancing the stability and performance of transformer-based models in real-world settings.

Moreover, techniques from control theory and system stability have been adapted to enhance the robustness of transformer architectures. Taking inspiration from Nguyen et.al [15], Han et al. [16] leveraged robust control theory to design transformers with improved stability properties, demonstrating superior performance on tasks with noisy inputs or adversarial perturbations.

Despite these advancements, existing approaches often lack a comprehensive understanding of the underlying factors influencing transformer stability. Furthermore, many proposed methods exhibit limited generalization across different tasks and datasets. Addressing these limitations requires a deeper investigation into the intrinsic properties of transformer models and the development of more principled and transferable stability enhancement techniques

In this paper, we contribute to this ongoing research by employing a GPT (Generative Pre-trained Transformer) model to investigate the importance of words within the corpus, employing sensitivity analysis techniques. Specifically, attention weights are used to measure the impact of individual words on the model's predictions. The paper proposes a novel approach to rank the importance of words by leveraging attention weights and conducting sensitivity analysis across the dataset. To quantify the discrepancies between model-generated outputs and ground truth, the Kullback-Leibler (KL) divergence is employed. This divergence measure aids in evaluating how well the model captures the underlying distribution of words in the corpus. By integrating KL divergence into the sensitivity analysis, the study aims to provide a more comprehensive understanding of word importance.

3. DATASET DESCRIPTION

The Cornell Movie-Dialogs Corpus has been used for proposed research [17]. It is a product of Cornell University's research, is a substantial dataset tailored for natural language processing (NLP) and dialogue analysis research. It boasts a diverse array of movie scripts, spanning genres from romantic comedies to action films, providing a variety of slanguage usage across varied social contexts. The Cornell Movie Dialogs Corpus is an extensive compilation of fictional dialogues sourced from original movie scripts, providing researchers with a rich and diverse collection of conversational data. Within this corpus, there exist 220,579 exchanges spanning across 10,292 pairs of characters from a wide array of movies, encompassing a total of 9,035 distinct characters featured in 617 films. With a staggering 304,713 utterances, the dataset offers a comprehensive glimpse into character interactions and dialogue dynamics. Furthermore, it includes detailed movie metadata such as genres, release years, IMDB ratings, and the number of IMDB votes, allowing for contextual analysis of conversations within the broader cinematic context. Additionally, character metadata, including gender information for 3,774 characters and

the positions of characters in movie credits for 3,321 characters, provides further depth for character-centric studies. This corpus serves as an invaluable resource for researchers in natural language processing and computational linguistics, offering a wealth of data for various analytical and modelling purpose.

4. METHODOLOGY

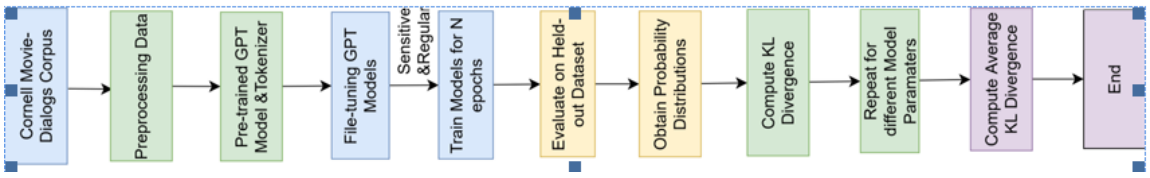

Figure 1.Flow Diagram of Proposed Research

Mechanisms are crucial components of transformer-based models, as they determine the degree to which each word or token attends to others in a sequence during processing, indicating the model's focus. To conduct sensitivity analysis using attention weights, several steps are typically followed.

Firstly, access to attention weights is obtained, facilitated by libraries such as Hugging Face's Transformers in Python, which provide easy access to these weights. Next, attention weights are extracted for specific input tokens or layers by analysing the output produced by the model. Once attention weights are extracted, various methods can be employed to conduct sensitivity analysis. One approach involves identifying token importance by determining which tokens or words have the highest or lowest attention weights, indicating their influence on the model's predictions. Additionally, attention weights can be analysed across different layers to understand how the model processes information hierarchically, providing insights into its decision-making process. Another method involves gradient-based sensitivity analysis, where input tokens are modified, and the resulting changes in attention weights are observed to identify tokens that heavily influence specific predictions or outputs. Furthermore, attention weights can be visualized using heatmaps or other graphical representations to gain a better understanding of the model's focus and attention distribution throughout the sequence. However, these methods provide mere visual representations that do not metrically contribute to the interpretability and explainability of multi-headed attention. Hence, the method proposed in this paper will employ the KL divergence to get quantifiable values for further analysis and explainability. Fig. 1 visualizes the flow of the proposed research.

Before understanding interpretability, it is imperative to understand the inner-workings of a transformer. A transformer works on an auto-regressive encoder-decoder structure, with the encoder stacks having multiple attention heads and the detector stacks having attention heads built specifically for masked language modelling. The attention head itself takes an input sequence of vectors $h = [h_1, \ldots, h_n]$ corresponding to the n tokens in the sequence. Given a vector h_i, the vector is transformed into query, key and value vectors through linear transformations [*i.e.* h_q, h_k, h_v]. These linear transformations are formed using the following formula in equation 1:

$$H \times W^{q,k,v} = H^{q,k,v} \tag{1}$$

To calculate attention, one must take the query vector for each word and find its product with the transpose of the key vector for each word within the sequence, including itself. This gives us the

attention score. These values are normalized through scaling and SoftMax functions as shown in equation 2:

$$Attention(Q, K, V) = sof\,max\left(\frac{QK^T + M}{\sqrt{d_k}}\right)V \tag{2}$$

where d_k is the size of the input and M is the attention mask added. Now, multi-headed attention takes place, which allows the model to jointly attend to information from different representation subspaces at different positions as in equation 3:

$$Multihead(Q, K, V) = \left(head_1, \ldots, head_n\right)concat \tag{3}$$

where $\underline{head_i}$ is the attention score for each head and *concat* is a function which concatenates the attention scores for each item from all the heads.

For the proposed research, the Cornell datasets were imported into our test bench set in a virtual environment running Python 3.10. The data was pre-processed so that only utterances with more than 5 tokens remained. The preprocessing was done through tokenization using the GPT-2 tokenizer that comes with the HuggingFace model. Then, a pre-trained model of GPT2 from the HuggingFace library which had been pre-trained on 1.5 billion parameters, was fed specific utterances from the dataset, and processed using the attention layers within GPT-2, and the attention weights were now used for further processing using the KL divergence method [18].
The attention score for each token was compared with using the following formula given in equation 4:

$$KL(P||Q) = \sum_i P(i) \log\frac{P(i)}{Q(i)} \tag{4}$$

where P and Q are referencing the first layer and the last year in the attention model, and the attention scores provided as probability distributions from 0 to 1, which is computed through the SoftMax normalization. A higher KL divergence for a particular token indicates that the attention distribution changes significantly across the layers, hence indicating a more dynamic role in the sentence itself, and playing a big role in GPT's prediction process, whereas conversely, a lower KL divergence shows that GPT is unable to extract varied information from that particular token, implying a less critical role in the sentence.

5. RESULTS

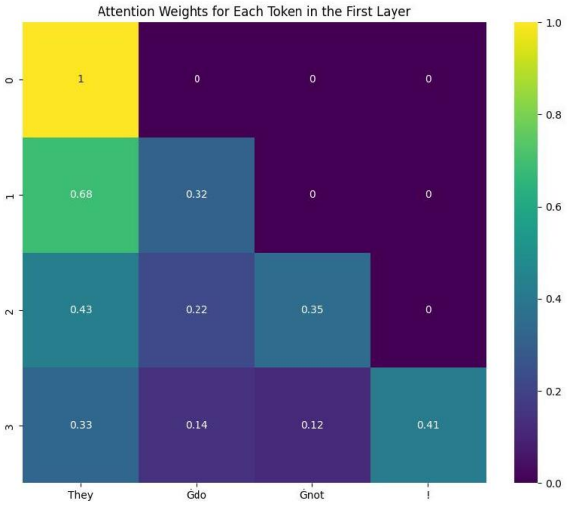

Figure 2. Attention Heatmap

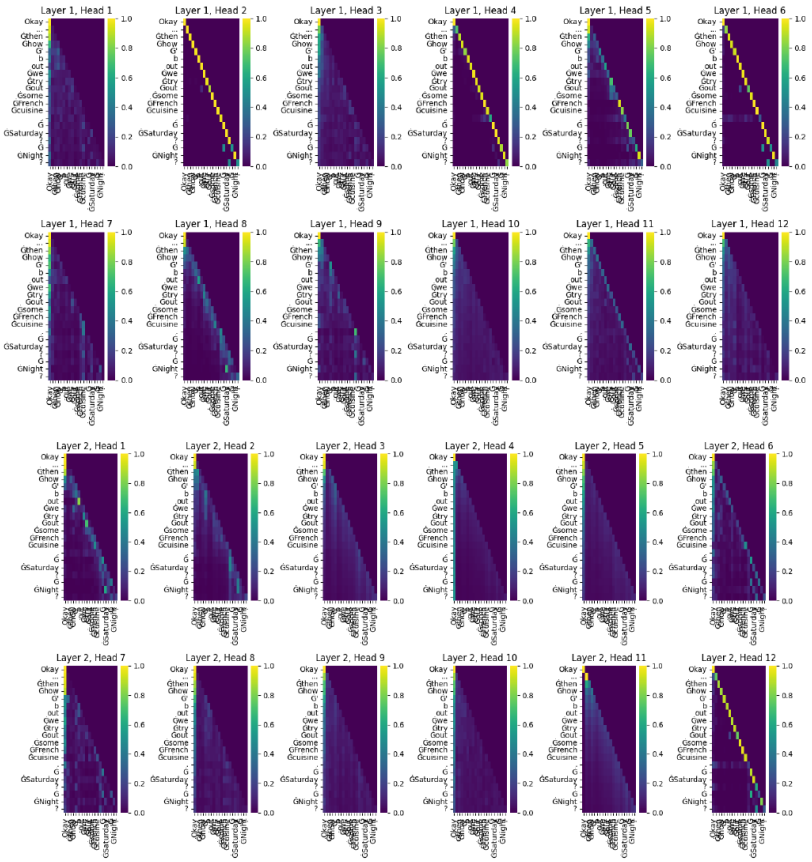

Figure. 3 Attention Head Visualization

The model had been trained for three epochs, during which a simple attention visualization was conducted for each token, as illustrated in Fig. 2. using heatmaps to gain a better understanding of

the model's focus and attention distribution throughout the sequence. Overall, conducting sensitivity analysis using attention weights enables a deeper understanding of how transformer-based models process information and make predictions. Subsequent to this initial visualization, further analyses were carried out, focusing on the attention heads and the layers, as depicted in Fig. 3. Fig. 2 and Fig. 3 act as baseline comparisons similar to the saliency-based XAI features that have been developed. However, while these visualizations allow one to understand the relationship between each word, the KL divergence method allows for a generalized understanding of how GPT treats each word. Following these visualizations, the KL divergence was computed and recorded for every token within the input sequence, as demonstrated in Figure 4. An exemplary instance extracted from utterance 60 is presented therein, shedding light on the divergence analysis conducted post-visualization.

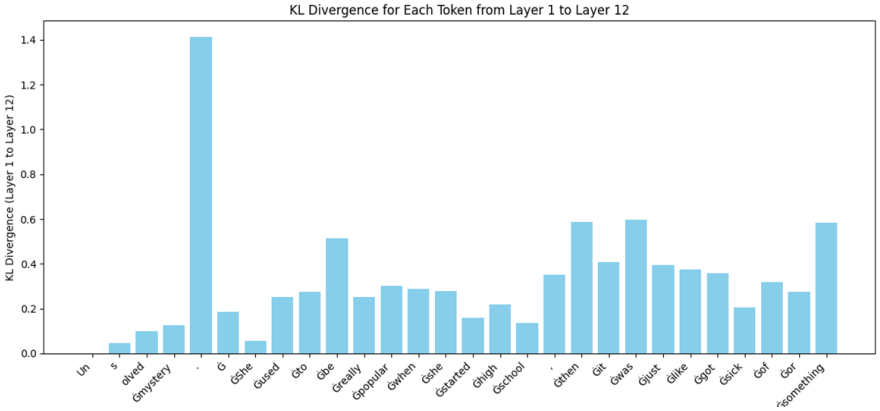

Figure 4.KL Divergence scores for Utterance #60

As the KL divergence increases, so does the significance of the input word within the sequence as illustrated in Fig.4, Fig. 5 and Fig.6. This relationship suggests that tokens with higher KL divergence values play a more pivotal role in shaping the contextual understanding and flow of information within the sequence. In essence, a higher KL divergence signifies a greater degree of reliance on the specific token for contextual cues and information integration, underscoring its importance in the overall comprehension process. The findings indicate that the transformer model places significant importance on punctuation marks that have been tokenized. These findings are consistent with the work done by Clark et. al. [19], showing how attention mechanisms are able to capture information about punctuation marks and give them importance within the attention maps itself, which is being reflected with the KL divergence method as well.

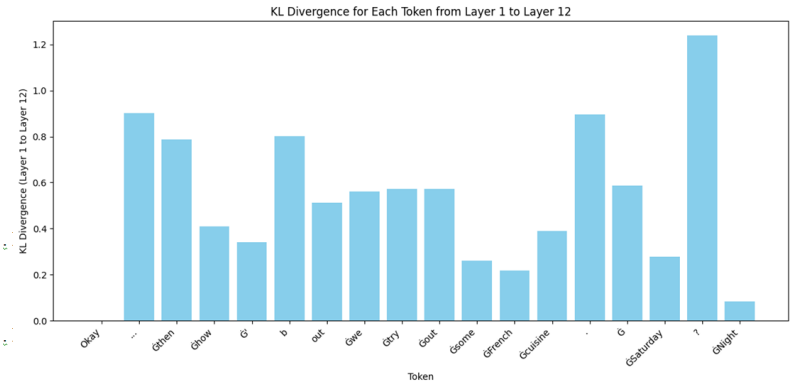

Figure 5. KL Divergence Scores for Utterance #65

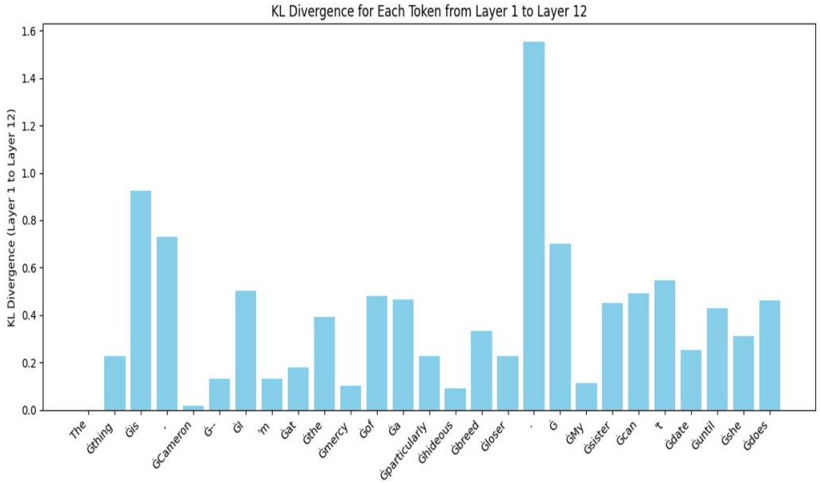

Figure 6. KL Divergence Scores for Utterance #45

This highlights the mechanism's ability to comprehend context throughout the entirety of a sentence. This observation sheds light on the attention mechanism within GPT-2, revealing a pronounced focus on sequential relationships. It suggests that GPT-2 may effectively interpret inputs when they are punctuated, facilitating clearer understanding and processing. Moreover, a higher KL divergence associated with a particular token signifies its greater reliance by other tokens for contextual cues. Analysing utterances #45 and #60 as examples, we see that token such as "is" and "was" given higher importance than any of the other non-delimiter tokens. Hence, we see that the attention model gives importance to the tense of a sentence and capturing the various syntactic and semantic dependencies in a sentence. This can be verified with previous literature as well.

6. TIME COMPLEXITY ANALYSIS

The first step involved encoding the input text using the tokenizer. This step typically has a time complexity proportional to the length of the input text, denoted as O(n), where n is the length of the input text. The model inference step involves passing the input tokens through the model to obtain outputs. The time complexity of this step depends on the model architecture and the

Pseudo-code

```
function process_input(input_text):
    tokens = tokenize(input_text)  // O(n)
    outputs = model_inference(tokens)  // O(1)
    attention_weights = compute_attention_weights(outputs)  // O(m * k)
    kl_divergence = compute_kl_divergence(attention_weights)  // O(m)

function tokenize(input_text):
     Tokenize the input text using the tokenizer
    Return the tokens

function model_inference(tokens):
    Pass the input tokens through the model
     Return the outputs

function compute_attention_weights(outputs):
     Iterate through each token in the input sequence
    Compute attention weights for each layer of the model
    Return the attention weights

function compute_kl_divergence(attention_weights):
    For each token, compute the KL divergence between the attention weights of the first and las
layers
     Return the KL divergences
```

length of the input sequence but can be roughly considered O (1) for a fixed-length input. In the next step, each token in the input sequence is iterated and computes the attention weights for each layer of the model. This involves accessing and processing attention weights for each token and layer. Let's the number of tokens ne m and the number of layers in the model be k. The time complexity for this step is approximately O (m * k)

For each token, the code calculates the KL divergence between the attention weights of the first and last layers. This involves computing the KL divergence for each token, which has a constant time complexity O (1) The overall time complexity of the model can be approximated as O(n + m * k), where n is the length of the input text, m is the number of tokens, and k is the number of layers in the model. The time complexity of, specifically, computing the KL divergence goes down to O (1). The pseudo-code is given below.

7. CONCLUSION AND FUTURE WORK

Attention mechanisms have a lack of interpretability due to their very nature. This study aims to provide quantifiable metrics to understand the importance of specific words within an input sequence provided to the GPT model by tangling with the attention weights. Previous methods have allowed for contextual interpretability between tokens spread across multiple layers, whereas by implying a statistical KL divergence method allows us to understand the weight of a word for the attention model. Through experimentation utilizing the Cornell datasets and employing advanced tools such as the GPT-2 tokenizer and pre-trained models, the study highlights the practical significance of attention mechanisms in real-life situations. Investigating attention scores and their use in tasks like KL divergence calculation provides insights into the interpretability and effectiveness of transformer-based models. The method used provides a much lower time complexity while still extracting plentiful data on the sensitivity of specific tokens in an input sequence provided to a transformer.

As the field progresses, there are numerous prospects for future research and advancements. One promising direction includes deeper examination of attention weight interpretation and its role in shaping model decision-making processes. Another potential area of growth lies in exploring innovative approaches for optimizing and fine-tuning attention mechanisms, aiming to boost model performance while minimizing computational overhead. The research can be extended by analysing further parts of the speech and seeing how the KL divergence interacts with such tokens. This opens up pathways to potentially find inaccuracies in the method and improve on it.

AUTHORS CONTRIBUTIONS

Arav Agarwal confirms the responsibility for the following: study conception and design, data collection, analysis and interpretation of results, and manuscript preparation under the supervision of Rhea Mahajan

REFERENCES

[1] Stokel-Walker, C., & Van Noorden, R. (2023). What ChatGPT and generative AI mean for science. Nature, 614(7947), 214–216. https://doi.org/10.1038/d41586-023-00340-6

[2] Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the promising perspectives and valid concerns. Healthcare, 11(6), 887. https://doi.org/10.3390/healthcare11060887

[3] Zheng, X., Zhang, C., & Woodland, P. C. (2021). Adapting GPT, GPT-2 and BERT language models for speech recognition. 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). https://doi.org/10.1109/asru51503.2021.9688232

[4] Wang, C., Liu, S., Yang, H., Jiu-Lin, G., Wu, Y., & Liu, J. (2023). Ethical considerations of using ChatGPT in health care. Journal of Medical Internet Research, 25, e48009. https://doi.org/10.2196/48009

[5] Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). Opinion Paper: "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. https://doi.org/10.1016/j.ijinfomgt.2023.102642

[6] Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2020). Explainable AI: A review of Machine Learning Interpretability Methods. Entropy, 23(1), 18. https://doi.org/10.3390/e23010018

[7] Chan, A. (2022). GPT-3 and InstructGPT: technological dystopianism, utopianism, and "Contextual" perspectives in AI ethics and industry. AI And Ethics, 3(1), 53–64. https://doi.org/10.1007/s43681-022-00148-6

[8] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

[9] Levi, N., Bloch, I. M., Freytsis, M., & Volansky, T. (2022). Noise injection node regularization for robust learning. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.15764

[10] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-attention with relative position representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers) (pp. 464-468).

[11] Hua, H., Li, X., Dou, D., Xu, C., & Luo, J. (2023). Improving pretrained Language Model Fine-Tuning with noise stability regularization. IEEE Transactions on Neural Networks and Learning Systems, 1–15. https://doi.org/10.1109/tnnls.2023.3330926

[12] Pande, M., Budhraja, A., Nema, P., Kumar, P., & Khapra, M. M. (2020). On the Importance of Local Information in Transformer Based Models. arXiv (Cornell University). https://arxiv.org/pdf/2008.05828.pdf

[13] Zhang, Y., Zhang, H., Wang, S., Wu, W., & Li, Z. (2022). PATS: Sensitivity-Aware Noisy Learning for Pretrained Language models. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.12403

[14] Davis, J. Q., Gu, A., Choromański, K., Dao, T., Ré, C., Finn, C., & Liang, P. (2021). Catformer: Designing Stable Transformers via Sensitivity Analysis. Proceedings of the 38th International Conference on Machine Learning, PMLR 139:2489-2499, 2021, 2489–2499.

[15] Tan Nguyen, Minh Pham, Tam Nguyen, Khai Nguyen, Stanley J Osher, and Nhat Ho. Fourierformer: Transformer meets generalized Fourier integral theorem. Advances in Neural Information Processing Systems, 2022

[16] Han, X., Ren, T., Nguyen, T., Nguyen, K., Ghosh, J., & Ho, N. (2022). Designing Robust Transformers using Robust Kernel Density Estimation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2210.05794

[17] Danescu-Niculescu-Mizil, C., & Lee, L. (2011). "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs." In Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011.

[18] D. Yu, K. Yao, H. Su, G. Li and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 2013, pp. 7893-7897, doi: 10.1109/ICASSP.2013.6639201.