ENHANCING INTEGRATED GRADIENTS USING EMPHASIS FACTORS AND ATTENTION FOR EFFECTIVE EXPLAINABILITY OF LARGE LANGUAGE MODELS

Anonymous authors

Paper under double-blind review

ABSTRACT

Understanding the decision-making processes of large language models (LLMs) is critical for ensuring transparency and trustworthiness. While Integrated Gradients (IG) is a popular method for model explainability, it faces limitations when applied to autoregressive models due to issues like exploding gradients and the neglect of the attention mechanisms. In this paper, we propose an enhanced explainability framework that augments IG with emphasis factors and attention mechanisms. By incorporating attention, we capture contextual dependencies between words, and the introduction of emphasis factors mitigates gradient issues encountered during attribution calculations. Our method provides more precise and interpretable explanations for autoregressive LLMs, effectively highlighting word-level contributions in text generation tasks. Experimental results demonstrate that our approach outperforms standard IG and baseline models in explaining word-level attributions, advancing the interpretability of LLMs.

1 INTRODUCTION

027 028

As large language models (LLMs) become increasingly prominent in natural language processing 029 tasks (Kenton & Toutanova (2019); Jha et al. (2020)), understanding their decision-making processes is critical for ensuring transparency and trustworthiness Lipton (2018). Autoregressive models, in 031 particular, generate text by predicting one word at a time based on the preceding context, making 032 it essential to interpret how individual words influence subsequent predictions. Traditional model 033 explainability techniques, such as Integrated Gradients (IG), have been widely used to quantify the 034 contribution of input features to model outputsShrikumar et al. (2017); Lundberg (2017); Murdoch et al. (2018). However, when applied to autoregressive models, IG faces inherent challenges due to their sequential nature, often leading to inaccurate or incomplete explanations Enguehard (2023). 037 Further related works has been discussed in Appendix A.1. In autoregressive text generation, capturing the contextual dependencies between words is crucial for reliable interpretability Vaswani (2017). Moreover, common challenges, such as exploding gradients during the gradient calculation for long texts using the IG method, further complicate the task of identifying meaningful token-level 040 contributions. To address these challenges, we propose an enhanced explainability framework 041 integrating attention mechanisms and emphasis factors with IG. Attention allows us to account for the 042 relationships between words in the context window while scaling factors mitigate the gradient-related 043 issues that can obscure proper explanations. We make the following key contributions in this paper: 044

- 1. We identify the limitations of the exploding gradient problem when applying the Integrated Gradients (IG) method for attribution analysis for long texts using generative LLMs.
- 2. We propose a novel solution to address the exploding gradient problem encountered during attribution calculations in the Integrated Gradients method.
- 3. We integrate the Attention mechanism into the attribution calculation, as it plays a critical role in predicting the next token in large language models (LLMs).

1

4. We conduct a comprehensive comparative study, evaluating our proposed method against several baseline models across multiple datasets and architectures.

011

013

014

015

016

017

018

019

021

023

024 025 026

046 047 048



Figure 1: (a) shows the IG values, the masked self-attention values and the EF values of the tokens
with respect to the "surprise" token for the text "The movie took me by surprise." produced by the
GPT2-small model. Our method combines these to create the values of AIEG method. (b) shows the
self-attention values with respect to the token "surprise".(c) shows the accumulation of gradient of
the output word with respect to a particular input word from the beginning of the text, over time in
the Integrated Gradient method as the length of the generated text gets long.

- 078
- 079 080 081

084

085

2 LIMITATIONS OF GRADIENTS AS ATTRIBUTIONS FOR GENERATIVE MODELS

Axiom: Sensitivity: The gradient-based method does not satisfy the sensitivity axiom. Let's demonstrate this with a straightforward example using a simple RNN. These are the general equations for the hidden state and output for an RNN, as given below.

Hidden State Update:
$$h_t = \sigma(\mathbf{W}_{hx} \cdot x_t + \mathbf{W}_{hh} \cdot h_{t-1} + \mathbf{b}_h)$$
 (1)

Output Calculation:
$$y_t = \phi(\mathbf{W}_{hy} \cdot h_t + \mathbf{b}_y)$$
 (2)

087 The hidden state at time step t is denoted by h_t . The activation function, denoted by σ and ϕ , is 088 typically a non-linear function such as tanh or ReLU. which introduces non-linearity and affects 089 gradient flow during backpropagation. The trainable weight matrix for the input x_t is represented 090 by \mathbf{W}_{hx} . Here, x_t denotes the input at time step t. The trainable weight matrix for the previous 091 hidden state h_{t-1} is given by \mathbf{W}_{hh} , and h_{t-1} is the hidden state from the previous time step. \mathbf{b}_h 092 and \mathbf{b}_{y} represent the bias terms. Now we will try to create a simple RNN from equations (1) and (2). We take $\mathbf{W}_{hx} = -1$, $\mathbf{W}_{hh} = 0$, $\mathbf{W}_{hy} = 1$, $\mathbf{b}_{h} = 1$, $\mathbf{b}_{y} = 1$, $h_{t-1} = 0$, and $\phi = \sigma = \text{ReLU}$. So the equation becomes: $h_t = \text{ReLU}(1 - x_t)$ and $y_t = \text{ReLU}(1 + h_t)$. The value of y_t is 2 and 1 for 094 values of $x_t = 0$ and 2 respectively. The value of $x \cdot \frac{\partial y_t}{\partial x}$ is 0 for both the values of x_t , indicating that sensitivity is not preserved. 096

097 098

099

3 APPROACH USING INTEGRATED GRADIENTS

100 In the paper Sundararajan et al. (2017), the authors demonstrate the application of Integrated Gradients 101 to neural machine translation models utilizing LSTM architectures. They compute the contribution 102 of each input token to the probability of every output token, which is represented in the form 103 of wordpieces. This process effectively aligns the output sentence with the input sentence. For 104 the baseline, the authors set the embeddings of all tokens to zero, except for the start and end 105 markers. Suppose for a neural network F, the goal is to compute attributions IG(x) that quantify the contribution of each input word to the network's output. Consider an input $x \in \mathbb{R}^n$ and a baseline 106 input $x' \in \mathbb{R}^n$ (typically a zero vector). Integrated Gradients compute the attributions of the input 107 relative to this baseline.



Figure 2: (a) shows the gradient values of the previously generated tokens with respect to the "actors" 129 token calculated by the Integrated Gradient method. As the distance between the tokens increases, 130 the gradient value increases and may explode if the text under consideration is very long. (b) shows 131 the gradient values of the previously generated tokens with respect to the "actors" token calculated by 132 our proposed method AIEG. The method has scaled down the gradient values by about 10%. The EF 133 factor helps to capture only the gradients where the model makes decisions. 134

136 The attribution of the *i*-th word is given by:

137 138

135

139

144

146

147

This formula represents the integrated gradient along the straight line between the baseline x' and the 142 input x. 143

Although the method has demonstrated effectiveness in neural machine translation, it fails to address certain limitations specific to generation tasks, which we will explore in the following sections. 145

 $\mathrm{IG}_{i}(x) = (x_{i} - x_{i}') \int_{\alpha = 0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_{i}} \, d\alpha$

(3)

3.1 LIMITATIONS IN INTEGRATED GRADIENT METHOD

148 While the method outlined in this paper has proven effective in numerous applications, it encounters 149 specific challenges when applied to long text generation with auto-regressive models such as GPTs. 150 Firstly, the method struggles with problems similar to exploding and vanishing gradient issues 151 that arise when computing the gradients of the output relative to input tokens, as shown in figure 152 1(c). Secondly, it neglects the impact of attention—a critical component in large language models 153 (LLMs)—in the attribution calculations of input tokens concerning the output tokens. Thirdly, this method assigns equal importance to all gradients, even in regions where the model's decision 154 remains unchanged, leading to the accumulation of low-quality gradients. The issue of exploding and 155 vanishing gradients is presented below as a theorem. 156

157 **Theorem:** Consider an auto-regressive neural network represented by the function $F : \mathbb{R}^n \to \vec{\mathbf{e}}$, 158 where $\vec{\mathbf{e}}$ is the embedding word vector of dimension n. Given an input sequence $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ where $\mathbf{x}_t \in \mathbb{R}^n$ is the input vector at time step t, the hidden state of F is updated iteratively at each 159 time step. Denote the hidden state at time step t as $\mathbf{h}_t \in \mathbb{R}^m$, where m is the dimension of the hidden 160 layer. We propose that when long sequences of text are considered (T >> 1), the calculation of 161 Integrated Gradients may result in undefined or numerically unstable values during the calculation

162 of $\frac{\partial y_t}{\partial x_{t'}}$, where y_t is the output generated at time t and $x_{t'}$ is the input word that was generated at time t'. 163 time t'.

Proof: Consider the scenario where we aim to calculate the attributions of the input words for the output word generated at time t. In autoregressive models—such the output at time t - 1 serves as input for generating the output at time t. Given Equations (1) and (2), the gradient of the output at time step t, denoted as y_t , with respect to an input word $x_{t'}$ at time step t', derived from an interpolated input, is expressed as follows. For t' = t (the same time step):

170 The gradient of the output y_t with respect to the input x_t is:

$$\frac{\partial y_t}{\partial x_t} = \frac{\partial y_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial x_t} \tag{4}$$

Where:

171 172 173

174 175

176 177

178 179 180

181

182

183

185 186

187 188 189

190

191

192

193

194

195 196

197

199

$$\frac{\partial y_t}{\partial h_t} = \mathbf{W}_y \cdot \sigma' (\mathbf{W}_y h_t + \mathbf{b}_y) \tag{5}$$

$$\frac{\partial h_t}{\partial x_t} = \mathbf{W}_x \cdot \sigma' (\mathbf{W}_h h_{t-1} + \mathbf{W}_x x_t + \mathbf{b}_h)$$
(6)

For t' < t (previous time steps):

The gradient of the output y_t with respect to an earlier input $x_{t'}$ (where t' < t) requires us to account for the effect of $x_{t'}$ on all subsequent hidden states up to h_t :

$$\frac{\partial y_t}{\partial x_{t'}} = \frac{\partial y_t}{\partial h_t} \cdot \frac{\partial h_t}{\partial h_{t-1}} \cdot \frac{\partial h_{t-1}}{\partial h_{t-2}} \cdots \frac{\partial h_{t'+1}}{\partial h_{t'}} \cdot \frac{\partial h_{t'}}{\partial x_{t'}} \tag{7}$$

Where:

$$\frac{\partial h_t}{\partial h_{t-1}} = \mathbf{W}_h \cdot \sigma' (\mathbf{W}_h h_{t-1} + \mathbf{W}_x x_t + \mathbf{b}_h)$$
(8)

and σ' is the derivative of the activation function.

Based on equations 7 and 8, it is clear that the gradients of each hidden layer are successively multiplied by the gradients of the previous layers, leading to an accumulation of gradients during the computation. Following this, the Integrated Gradients (IG) values are determined using equation 3. This process, particularly for long sequences, can result in gradient explosion due to the cumulative summation, as depicted in figure 2, which may introduce instability in the gradient calculations. \Box

4 OUR PROPOSED METHOD

4.1 MOTIVATION

200 We aim to identify the positive contributions of individual input tokens towards the generation of the 201 output token. In this work, we address the limitations of Integrated Gradients (IG) discussed in the 202 previous section. Our approach aims to mitigate the risk of gradient explosion by scaling down the 203 value of $\frac{dF}{dx}$ and focusing only on high-quality gradients, i.e., those where the logits exhibit rapid 204 changes (Walker et al. (2024)). By selectively considering these gradients, we effectively reduce the likelihood of gradient explosion in $\frac{dF}{dx}$. Also, given the critical role that attention mechanisms play in Large Language Models (LLMs) for next-token generation, it is essential to incorporate 205 206 207 attention weights when calculating attributions. The attention mechanism, introduced by Vaswani (2017), allows models to focus on relevant parts of an input sequence, emphasizing key information 208 when processing long sequences, shown in figure 1(b), and improving performance in tasks like 209 translation, summarization, and question-answering. Consequently, attention is crucial for LLMs' 210 decision-making in the next token generation and should be considered alongside integrated gradients 211 in attribution calculations. Additionally, large language models like GPTs (Radford et al., 2019) use 212 a technique known as masked self-attention, which plays a pivotal role in sequence generation tasks. 213

Axiom: Attention: Consider two words, \mathbf{x}_{t1} and \mathbf{x}_{t2} , within a sequence of words in a sentence, with attention values $A_{t1,t}$ and $A_{t2,t}$, respectively, corresponding to the generated output word at time t. If $A_{t1,t} > A_{t2,t}$, then it implies that the word \mathbf{x}_{t1} has a greater impact and contribution towards the



Figure 3: The above diagrams illustrate the Output Vs Alpha and EF Vs Alpha graphs of two different texts. (a) and (c) illustrates how the output changes with varying values of α . Here the output is the L1 normalisation of the embedding vector of the targeted output. It is evident that around $\alpha = 0.8$, the model makes its prediction, and beyond this point, the model maintains its decision. (b) and (d) shows the variation of the emphasis factor (EF) with respect to the values of α . Our method focuses on high-quality gradients, where the model makes decisions (rapid change of output), while in regions of low-quality gradients, the EF becomes 0, reducing the entire term in our proposed method to 0.

247

generation of the output word at time t. This higher attention value reflects that \mathbf{x}_{t1} is considered more relevant and influential in the context of the output prediction compared to \mathbf{x}_{t2} . \Box

Previous attribution methods have largely overlooked the significance of attention mechanisms in their computations. In contrast, we propose using the aforementioned axiom to calculate the attribution of each input word towards the output. Let for an auto-regressive model with L decoder layers and H multi-headed masked self-attention mechanisms, Attention_{t',t,h,l} denote the attention value of an input word $x_{t'}$ generated at time t' with respect to the output word x_t generated at time t (where t' < t), for the l^{th} layer and h^{th} attention head. The overall attention of that input word towards the output word $NetAttention_{t',t}$ can then be computed as:

258 259

where,

A

255 256

257

260

261 262

NetAttention_{t',t} =
$$\frac{1}{L} \sum_{l=0}^{L} \left(\frac{1}{H} \sum_{h=0}^{H} \text{Attention}_{t',t,h,l} \right)$$
 (9)

Attention_{t',t,h,l} =
$$\left(\operatorname{softmax} \left(\frac{\mathbf{Q}_{t',h,l} \mathbf{K}_{t,h,l}^T}{\sqrt{d_k}} \right) \right) \mathbf{V}_{t',h,l}$$
(10)

263 264 $\mathbf{Q}_{t',h,l}$ is the query vector of token $x_{t'}$, $\mathbf{K}_{t,h,l}$ is the key vector of token x_t , $\mathbf{V}_{t',h,l}$ is the value vector 265 of the $x_{t'}$ token for the l^{th} layer and the h^{th} attention head and \mathbf{d}_k is the dimensionality of the 266 key/query vectors.

Building upon the above theorem and axiom, we introduce our proposed attribution method, Attended Integrated and Emphasized Gradients (AIEG), along with the Emphasis Factor (EF). Consider an auto-regressive model generating a token x_t . Our goal is to compute the attribution of a previously generated token $x_{t'}$, where 0 < t' < t. In this context, x'_t serves as the baseline for the input x_t , and

270 $x'_{t'}$ acts as the baseline for the token $x_{t'}$. Thus, to compute the attribution of the output token with 271 respect to the input token $x_{t'}$:

273 274

275 276 277

278 279

281

283 284

$$\operatorname{AIEG}_{t'}(x_t) = \operatorname{NetAttention}_{t',t} \times \operatorname{PosNorm}\left((x_{t'} - x'_{t'}) \times \int_{\alpha=0}^{1} \frac{\partial F(x'_t + \alpha(x_t - x'_t))}{\partial x_{t'}} \times \operatorname{\mathbf{EF}} d\alpha \right)$$
(11)

where,

$$EF = \frac{|F(x'_t + \alpha(x_t - x'_t)) - F(x'_t + (\alpha - \epsilon)(x_t - x'_t))|}{|F(x'_t + \alpha(x_t - x'_t))| + |F(x'_t + (\alpha - \epsilon)(x_t - x'_t))|},$$
(12)

$$\operatorname{PosNorm}(a) = \frac{a}{\sum_{T=1}^{t-1} \operatorname{AIEG}_T(x_t)},$$
(13)

where, a > 0, $\forall T AIEG_T(x_t) > 0$, ϵ is the minimum difference between two values of α , $(\alpha - \epsilon) \ge 0$ and $0 < \epsilon < 1$. Also $||F(x'_t + \alpha(x_t - x'_t))| + |F(x'_t + (\alpha - \epsilon)(x_t - x'_t))|| > 0$, so that the EF remains defined for all values of x_t, x'_t and α .

288 The PosNorm function normalizes the attribution of a token generated at time t' for the output token 289 at time t (t' < t) across all positive attributions of tokens generated from T = 1 to T = t - 1. We 290 focus solely on positive attributions as we are interested in identifying words that positively contribute to the output. In our approach, attention and gradients are weighted equally, as both contribute equally 291 to the generation of the next token. If the gradient value is high but the attention value of the input 292 token with respect to the output token is low, the overall attribution decreases, and the reverse is also 293 true. This has been depicted in figure 1 (a) with the text "The movie took me by surprise." and the output word with respect to which the gradients are calculated is "surprise". The algorithm for the 295 method has been discussed in Appendix A.3. Next, we will present two theorems, along with their 296 proofs, and three axioms to further explore the properties of the above equations. 297

Theorem1: Consider a function $F(x) : \mathbb{R} \to \mathbb{R}$ and the Emphasis Factor EF function, mentioned in equation 12, which is continuous over the entire range of F(x). We argue that $F(x) \times EF \leq F(x)$, $\forall F(x) \in \mathbb{R}$, keeping the sign of F(x) intact.

Proof: The Emphasis Factor (EF) can be expressed in a simple form as $EF = \frac{|m-n|}{|m|+|n|}$.

303 When $m \neq n$: $\forall m$ and n, where $m \neq n$, the following holds: $0 < EF \le 1$. This is true because 304 $|m - n| = |m + (-n)| \le |m| + |-n| = |m| + |n|$ by the triangle inequality. 305 When m = n: we have: EF = 0

Thus, in both cases, the product $F(x) \times EF$ satisfies the following condition: $F(x) \times EF \le F(x)$. This demonstrates that the Emphasis Factor ensures the product is always less than or equal to the original function F(x) and hence checks the exploding gradient that arises while calculating the gradients. Also, since EF is always greater than equal to zero, it keeps the sign of the product the same as F(x). Hence the contribution of the words remains same, that is, a positively contributing word does not change to negative because of $EF.\Box$

Theorem2: Consider a function $F(x) : \mathbb{R} \to \mathbb{R}$ and the Emphasis Factor EF function, We assert that the Emphasis Factor prioritizes gradients in regions where the model is making decisions, while disregarding gradients in areas where the output has already been predicted.

Proof: Consider an input token with an attention value greater than 0 with respect to the output token. When $F(x'_t + \alpha(x_t - x'_t)) \neq F(x'_t + (\alpha - \epsilon)(x_t - x'_t))$, the model is still in the decision-making phase, resulting in EF > 0, and thus $\frac{\partial F(x'_t + \alpha(x_t - x'_t))}{\partial x'_t} \times \text{EF} \, d\alpha > 0$. Conversely, when $F(x'_t + \alpha(x_t - x'_t)) = F(x'_t + (\alpha - \epsilon)(x_t - x'_t))$, the model has already made a decision, implying EF = 0, and $\frac{\partial F(x'_t + \alpha(x_t - x'_t))}{\partial x'_t} \times \text{EF} \, d\alpha = 0$ for a specific value of α . Hence, the Emphasis Factor (EF) selectively captures only high-quality gradients, filtering out low-quality ones. Figure 3 and Appendix A.7 shows this theorem through graphs for different examples.

³²³ Due to the properties outlined in Theorems 1 and 2, the use of the EF effectively mitigates the gradient explosion issue commonly observed in standard Integrated Gradients.

	0	GPT2-smal	1		GPT-nano			LLaMA	
Method	LO↓	Comp ↑	Suff↓	LO↓	Comp ↑	Suff↓	LO↓	Comp ↑	Suff↓
Grad*Inp	-0.245	0.173	0.322	-0.290	0.165	0.368	-0.360	0.148	0.445
IG	-0.527	0.338	0.260	-0.780	0.362	0.236	-1.180	0.310	0.415
IGCG	-0.480	0.278	0.174	-0.435	0.229	0.280	-1.040	0.295	0.418
DeepLIFT	-0.195	0.054	0.488	-0.299	0.079	0.433	-0.174	0.064	0.469
GradShap	-0.377	0.217	0.309	-0.522	0.167	0.346	-0.685	0.224	0.434
AttnOnly	-0.137	0.121	0.294	-0.144	0.133	0.308	-0.177	0.187	0.445
AIEG	-0.535	0.348	0.140	-0.860	0.368	0.258	-1.510	0.395	0.355

Table 1: Comparison of our proposed method with various feature attribution methods across three language models fine-tuned and tested on the SST-2 dataset. For \uparrow metrics, higher values indicate better performance, while for \downarrow metrics, lower values are preferred.

	G	PT2-smal	1		GPT-nano			LLaMA	
Method	LO↓	Comp ↑	Suff↓	LO↓	Comp ↑	Suff↓	LO↓	Comp ↑	Suff↓
Grad*Inp	-0.252	0.170	0.319	-0.115	0.163	0.370	-0.233	0.145	0.442
IG	-0.530	0.334	0.165	-0.792	0.358	0.254	-1.185	0.305	0.413
IGCG	-0.482	0.276	0.179	-0.429	0.227	0.284	-1.048	0.291	0.412
DeepLIFT	-0.196	0.073	0.487	-0.198	0.080	0.432	-0.175	0.065	0.470
GradShap	-0.478	0.216	0.310	-0.521	0.168	0.347	-0.684	0.225	0.365
AttnOnly	-0.138	0.111	0.295	-0.143	0.134	0.309	-0.176	0.186	0.446
AIEG	-0.542	0.345	0.137	-0.865	0.365	0.240	-1.515	0.393	0.351

Table 2: Comparison of our proposed method with various feature attribution methods across three language models fine-tuned and tested on the IMDB dataset. For \uparrow metrics, higher values indicate better performance, while for \downarrow metrics, lower values are preferred.

Axiom: Sensitivity: Consider an autoregressive neural network function F, which is continuous and differentiable with respect to α , ensuring that $\frac{\partial F}{\partial \alpha}$ is well-defined. Our attribution method at α along a given path is defined as: $\frac{\partial F}{\partial x} \times EF$. The term (x - x') is omitted here, as it is a post-processing factor. For an input token closely related to the output token NetAttention > 0.

When EF = 0, which implies the change in output is 0 and therefore, the attribution is naturally zero. Conversely, when $EF \neq 0$, at least one feature will have $\frac{\partial F}{\partial x} \neq 0$, resulting in a nonzero attribution. Therefore, by definition, our proposed attribution method AIEG satisfies the Sensitivity axiom.

Axiom: Implementation Invariance: Consider an autoregressive neural network F, where g is the input at time t', h represents the hidden layer, and f is the output generated at time t (with t > t'). In AIEG, the computation of $\frac{\partial f}{\partial g}$ is performed through the chain rule, such that $\frac{\partial f}{\partial g} = \frac{\partial f}{\partial h} \cdot \frac{\partial h}{\partial g}$. Given that the input word g contributes positively to the output token (NetAttention > 0), our proposed method adheres to the principle of Implementation Invariance. \Box

Axiom: Linearity: Assume we combine two autoregressive deep networks, represented by the functions f_1 and f_2 , to form a third network that models the function $a \times f_1 + b \times f_2$, i.e., a linear combination of the two networks. The attributions computed by the AIEG method for $a \times f_1 + b \times f_2$ result in a weighted sum of the attributions for f_1 and f_2 , with weights a and b, respectively. Therefore, our method satisfies the principle of linearity. \Box

EXPERIMENT AND EVALUATION

- EXPERIMENT DESIGN 5.1

We evaluate our proposed method against the following baseline models: Grad*Inp (Shrikumar et al. (2016)), Integrated Gradients (IG) (Sundararajan et al. (2017)), Integrated Gradients with Clipped Gradients (IGCG) as described in Appendix A.2, DeepLift (Shrikumar et al. (2017)), GradientShap





Figure 5: Impact of varying the top-k% on the log-odds, comprehensiveness, and sufficiency metrics for the GPT2-small model fine-tuned on the IMDb (Maas et al. (2011)) dataset.

448 predicted by the respective models. This approach will provide insight into the models' confidence and 449 facilitate a comparative performance analysis. For the Integrated Gradients with Clipped Gradients 450 (IGCG) method, we applied a threshold of 1,000,000 to clip extreme gradient values during the attribution calculation. Detailed explanations of the metrics has been discussed in Appendix A.4.

452 SST2 contains 11,855 individual sentences extracted from movie reviews, while the IMDB dataset 453 consists of 50,000 movie reviews. We randomly selected 5000 reviews from each dataset and 454 fine-tuned the models as masked language models. A smaller number of examples was chosen 455 for fine-tuning, as our objective is to understand the model's behaviour rather than to generate 456 high-quality, task-related outputs. Similarly for testing, we randomly selected around 2100 movie 457 reviews from each dataset and used a portion of the review to construct a paragraph of 50, 200, and 400 tokens, with each category having an equal amount of movie reviews (700). Since the model 458 outputs tokens, we convert them back to words before presenting the final output. For words that are 459 split during tokenization, the tokens are reassembled, and their individual attributions are summed to 460 compute the attribution of the entire word. From the generated text, we manually selected a token 461 of interest to calculate its positive attribution based on the preceding tokens. The attributions were 462 computed and compared across different models. Table 1 presents the attribution comparisons for 463 the SST2 dataset, while Table 2 compares the results for the IMDB dataset. 464

5.2 Results

445

446 447

451

465 466

467

Tables 1 and 2 compare the performance of our proposed algorithm against the other attribution 468 methods discussed above. Our results consistently outperform the other methods across the datasets 469 and language models. This suggests that the attention mechanism and the emphasis factor play a 470 crucial role in determining the attribution of each token towards the output token. In Figures 4, 471 we aim to identify the positive attribution of the input words toward the output word (the word of 472 interest) and compare the outputs generated by our proposed method (left examples) with those 473 from Integrated Gradients (right examples). The greener a word appears, the greater its positive 474 contribution to the word of interest. In Figures 4 (a) and (b), the word of interest is "galaxy." It 475 is clear that the words with the highest attributions in (a) are "land," "lived," "warrior," "aliens," 476 and "wars," which are coherent. In contrast, (b), generated by the IG method, highlights "In" along with the other words. Similarly, from the other examples, it is evident that our proposed method 477 outperforms IG. More visual examples has been shown in Appendix A.6, where we have compared 478 the attributions computed by all the above-mentioned methods. We tested our method with text 479 summarising and compared it with the IG method in Appendix A.5. In almost all the cases AIEG 480 gives more reasonable attributions than IG method. 481

482 Ablation Studies on the values of k in Evaluation Metrics: Figure 5 illustrates the impact of varying the top-k% on the log-odds, comprehensiveness, and sufficiency metrics for the GPT2-small 483 model fine-tuned on the IMDb dataset. We compare our AIEG method against the Integrated 484 Gradients (IG). Our results show that both variants outperform IG across all values of k. Notably, the 485 performance gap between AIEG and IG is minimal at lower k values but progressively widens as k increases, as depicted in figure 5 (a). In figure 5 (b) and figure 5 (c) the gap between the values remains almost the same but AIEG outperforms IG for all values of k.

488 489 490

491

492

493

494

495

496

497 498

499

6 CONCLUSION

In this paper, we demonstrated the limitations of the Integrated Gradients (IG) method in computing input token attributions toward the output token. Specifically, we highlighted the issue of exploding gradients when calculating the gradients of input tokens with respect to the output. To address this, we introduced the Attended Integrated and Emphasized Gradients (AIEG) method, which mitigates the exploding gradient problem by focusing on high-quality gradients. Our proposed method consistently outperforms other approaches in attribution calculation across multiple datasets and models.

References

- Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.408. URL https: //aclanthology.org/2020.acl-main.408.
- Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models. *arXiv preprint arXiv:2305.15853*, 2023.
- Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber.
 Pathologies of neural models make interpretations difficult. *arXiv preprint arXiv:1804.07781*, 2018.
- Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. Transformer feed-forward layers
 build predictions by promoting concepts in the vocabulary space. *arXiv preprint arXiv:2203.14680*, 2022.
- Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. *arXiv preprint arXiv:2310.15916*, 2023.
- John Hewitt and Christopher D Manning. A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4129–4138, 2019.
- Anupama Jha, Joseph K. Aicher, Matthew R. Gazzara, Deependra Singh, and Yoseph Barash. Enhanced integrated gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome biology*, 21:1–22, 2020.
- Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep
 bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1,
 pp. 2. Minneapolis, Minnesota, 2019.
- Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Analyzing feed-forward blocks in transformers through the lens of attention map. *arXiv preprint arXiv:2302.00456*, 2023.
- Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time
 intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36, 2024.
- Zachary C Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018.
- Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho.
 The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*, 2024.

553

554

561

567

574

- Scott Lundberg. A unified approach to interpreting model predictions. arXiv preprint arXiv:1705.07874, 2017.
- Andrew Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts.
 Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142–150, 2011.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372, 2022.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn.
 Memory-based model editing at scale. In *International Conference on Machine Learning*, pp. 15817–15831. PMLR, 2022.
 - W James Murdoch, Peter J Liu, and Bin Yu. Beyond word importance: Contextual decomposition to extract interactions from lstms. *arXiv preprint arXiv:1801.05453*, 2018.
- Shashi Narayan, Shay B Cohen, and Mirella Lapata. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*, 2018.
- Hao Peng, Xiaozhi Wang, Shengding Hu, Hailong Jin, Lei Hou, Juanzi Li, Zhiyuan Liu, and Qun
 Liu. Copen: Probing conceptual knowledge in pre-trained language models. *arXiv preprint arXiv:2211.04079*, 2022.
- Gandhi Prerak, Pramanik Vishal, and Bhattacharyya Pushpak. Kurosawa: A script writer's assistant. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pp. 540–550, 2023.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language
 models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.
- Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?" explaining the
 predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.
- Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black
 box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.
- Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through
 propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMIR, 2017.
- Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 865–878, 2021.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642, 2013.
- 587 Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In 588 *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.
- Eric Todd, Millicent L Li, Arnab Sen Sharma, Aaron Mueller, Byron C Wallace, and David Bau.
 Function vectors in large language models. *arXiv preprint arXiv:2310.15213*, 2023.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

594 595	A Vaswani. Attention is all you need. Advances in Neural Information Processing Systems, 2017.
596	Chase Walker Sumit Iba Kenny Chen and Rickard Ewetz Integrated decision gradients: Compute
507	your attributions where the model makes its decision. In <i>Proceedings of the AAAI Conference on</i>
598	Artificial Intelligence, volume 38, pp. 5289–5297, 2024.
599 600 601	Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. <i>arXiv</i> preprint arXiv:2211.00593, 2022.
602	Vershi Ver Dere Were Deshere Time Gimer Chare Zhenhe Li Chumin Dere Hering Char
603	and Ningyu Zhang. Editing large language models: Problems, methods, and opportunities, grViv
604	nrenrint arXiv:2305 13172 2023
605	proprim urAtv.2505.15172, 2025.
606	
607	
608	
609	
610	
611	
612	
613	
614	
615	
616	
617	
618	
619	
620	
621	
622	
623	
624	
625	
626	
627	
628	
629	
621	
620	
622	
63/	
635	
636	
637	
638	
639	
640	
641	
642	
643	
644	
645	
646	
647	

APPENDIX А 649

651

650 A.1 RELATED WORKS

652 Explainability in machine learning, particularly for large language models (LLMs), has become a 653 crucial area of research as these models grow in complexity. The "black-box" nature of models 654 like GPT and LLaMA-2 Touvron et al. (2023) poses significant challenges in understanding how 655 these models make predictions, leading to a demand for more transparent methods to interpret their 656 behaviour.

657 Local explainability techniques such as SHAP Lundberg (2017) and LIME Ribeiro et al. (2016) have 658 been widely adopted to provide insight into the contributions of individual input features. These 659 methods rely on perturbations and attribution techniques to assess the influence of tokens on model 660 outputs. However, they are often computationally intensive and assume feature independence, which 661 may not hold in real-world datasets Feng et al. (2018). Also, these methods may not be able to capture 662 the decision of why the model generated the token under consideration. Gradient-based methods such as Integrated Gradients Sundararajan et al. (2017) accumulate the gradients along the input feature 663 path, providing a smoother attribution but at the cost of higher computational demand and reduced 664 faithfulness (Sikdar et al. (2021); Shrikumar et al. (2017)). 665

666 Global explainability focuses on extracting and interpreting broader patterns within models. 667 Probing-based methods have been essential in identifying the syntactic and semantic representations encoded within LLMs (Hewitt & Manning (2019); Peng et al. (2022)). Studies by Geva et al. 668 (2022) and Kobayashi et al. (2023) delve into the internal mechanisms of models, showing that 669 feed-forward networks and attention heads capture complex linguistic knowledge. Mechanistic 670 interpretability has also become an essential field, aiming to reverse-engineer neural networks into 671 comprehensible circuits Wang et al. (2022), allowing for a deeper understanding of tasks like indirect 672 object identification. 673

674 Model editing has recently garnered attention as a way to directly alter specific knowledge within LLMs without extensive retraining. Techniques such as hypernetwork-based editing Mitchell et al. 675 (2022) and causal tracing Meng et al. (2022)) allow for targeted interventions in model behavior, 676 improving its responses to particular inputs. These techniques have shown potential in enabling 677 models to adapt without disrupting overall performance Yao et al. (2023). 678

679 Explainability has also been used to enhance task-specific capabilities. In-context learning (ICL), 680 for instance, has benefited from studies showing that specific attention heads play a pivotal role in transferring knowledge from prompt examples to downstream tasks (Hendel et al. (2023); Todd et al. 681 (2023)). Moreover, explainability methods like inference-time intervention (ITI) have been leveraged 682 to address issues of hallucination in text generation, where models generate outputs that deviate from 683 factual content. Li et al. (2024) demonstrated that truthful interventions in attention layers could 684 significantly enhance the factuality of model outputs, mitigating the impact of hallucinations. 685

Beyond improving factuality, explainability has also been used to tackle biases within models. 686 Techniques like integrated gradients (Sundararajan et al. (2017)) and its variations have been applied 687 to identify neurons responsible for social biases (Liu et al. (2024)), offering a pathway to fairer and 688 more ethically aligned language models. 689

690 Overall, the body of research highlights the importance of developing both local and global 691 explainability methods to improve trust and transparency in LLMs. These methods not only facilitate 692 understanding but also open new avenues for enhancing the performance and ethical alignment of models in diverse NLP applications. 693

694 695

696

697

WHY NOT USE GRADIENT CLIPPING BEYOND A THRESHOLD TO STOP THE GRADIENT A.2 EXPLOSION?

We are suggesting that instead of using an emphasis factor, when the gradient exceeds a predefined 699 positive threshold, further multiplication is halted. This approach prevents the gradients from becoming too small or too large, ensuring more stable and meaningful gradient calculations. 700 Specifically, we define positive and negative thresholds θ^+ and θ^- , respectively. The gradients 701 are modified as follows:

$$\frac{\partial h_t}{\partial h_{t-1}}^{\text{(capped)}} = \begin{cases} \frac{\partial h_t}{\partial h_{t-1}}, & \text{if } \frac{\partial h_t}{\partial h_{t-1}} \le \theta^+ \\ \theta^+, & \text{if } \frac{\partial h_t}{\partial h_{t-1}} > \theta^+ \end{cases}$$

Clipping the gradients has a significant impact on the calculation of attributions. This approach may
 hinder the proper accumulation of critical gradients, particularly in areas where the model makes key
 decisions. As a result, the method fails to accurately highlight the tokens that contribute the most
 to the output. The impact of gradient clipping on the attribution process is further illustrated in the
 accompanying figures in Appendix A.6.

 A.3

PROPOSED ALGORITHM

1. Encode input text to token IDs: x = tokenizer(text)2. Set baseline: $x_0 = 0$ (if not provided) 3. Initialize total gradients: $G_{\text{total}} = 0$ 4. For each $\alpha \in [0, 1]$ with steps N: $x_i(\alpha) = x_0 + \alpha(x - x_0)$ $y(\alpha) = \text{model}(x_i(\alpha))$ $\nabla x_i(\alpha) = \frac{\partial y(\alpha)}{\partial x_i(\alpha)}$ $\Delta G(\alpha) = \nabla x_i(\alpha) \cdot \frac{\left|y(\alpha) - y(\alpha - \frac{1}{N})\right|}{\left|y(\alpha)\right| + \left|y(\alpha - \frac{1}{N})\right|}$ $G_{\text{total}} + = \Delta G(\alpha)$ $G_{\rm avg} = \frac{G_{\rm total}}{N}$ 5. Compute IG scores: $IG(x) = (x - x_0) \cdot G_{avg}$ 6. Normalize IG scores: $IG_{\text{norm}}(x) = \frac{IG(x)}{\sum IG(x)}$ 7. Extract attention and Average: A = average attention from each layer l and head h 8. Compute contribution: $C(x) = IG_{norm}(x) \cdot A$ 9. Return token contributions: C(x) for each token A.4 **EVALUATION METRICS** • Log-odds (LO) score: Shrikumar et al. (2017), measures the average change in negative logarithmic probabilities for the predicted class when the top k% of features are masked using zero padding. To calculate this, the top k% of words are identified based on attribution scores from an explanation algorithm and are then replaced with zero padding. Specifically, for a dataset with N sentences, the LO score is defined as:

C

$$\log\text{-odds}(k) = \frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{p(\hat{y} \mid x_i^{(k)})}{p(\hat{y} \mid x_i)}\right)$$

where \hat{y} is the predicted class, x_i is the *i*-th sentence, and $x_i^{(k)}$ is the modified sentence with the top k% words replaced by zero padding. Lower scores indicate better performance.

• **Comprehensiveness (Comp) score:** DeYoung et al. (2020), quantifies the average change in predicted class probability resulting from the removal of the top k% of features. This score, similar to the Log-odds, assesses the impact of the most influential words on the model's prediction. It is defined as:

$$\text{Comp}(k) = \frac{1}{N} \sum_{i=1}^{N} \left[p(\hat{y} \mid x_i^{(k)}) - p(\hat{y} \mid x_i) \right]$$

where $x_i^{(k)}$ represents the modified sentence with the top k% of words removed. Higher scores indicate better performance.

- The Sufficiency (Suff) score: DeYoung et al. (2020), measures the average change in predicted class probability when only the top k% of features are retained. This score evaluates how well the top k% attributions alone account for the model's prediction. It is calculated similarly to the Comprehensiveness score, but here $x_i^{(k)}$ refers to the sentence containing only the top k% of words. Lower scores indicate better performance.

A.5 APPLICATIONS IN OTHER TASKS

We evaluated our model for text summarization. For this purpose, we employed the XSum dataset (Narayan et al. (2018)). The GPT-2 (small) model was used for summarization, with the string "TL;DR" appended to the end of the input to guide the summarization process. Following the generation of the summarized text, we computed attributions for each word and aggregated their contributions. For example, given the summarized text "The scientists discovered a new animal," we determined the contribution of each word, starting from "The" to "animal," and accumulated their respective contributions. Since it is an autoregressive model, the words generated in the summarised text depend on the previous words as well, and therefore, even the summarised text has its contributions in generating the next word. This process is illustrated in Figures 6 to 14 below.

We then compared the output of the Integrated Gradients (IG) method with that of the AIEG method.
Our results indicate that the attributions produced by AIEG are more reasonable and consistent compared to those from IG.

A species of crustacean which lives in the gut of sea cucumbers has been discovered in the waters off Scotland. Scientists found the "furry" crab, which is about 3cm long, during a research project in the Inner Hebrides. The creature has been named after its thick coating of hair, which gives it a striking appearance. It is also known to feed on the mucus and faeces of sea cucumbers. The researchers described it as a "strange and fascinating" species. Summary A new species of furry crab has been found in Scottish waters. Output from AIEG A species of crustacean which lives in the gut of sea cucumbers has been discovered in the waters off Scotland . Scientists found the "furry" crab , which is about 3cm long, during a research project in the Inner Hebrides. The creature has been named after its thick coating of hair, which gives it a striking appearance. It is also known to feed on the mucus and faeces of sea cucumbers. The researchers ribed it as a "strange and fascinating" species Summary: A new species of furry crab has been found in Scottish waters. Output from IG A species of crustacean which lives in the gut of sea cucumbers has been discovered in the waters off Scotland. Scientists found the "furry" crab, which is about 3cm long, during a research project in the Inner Hebrides. The creature has been named after its thick coating of hair, which gives it a striking appearance. It also known to feed on the mucus and faeces of sea . The researchers described it as a "strange and fascinating" species Summary A new species of furry crab has been found in Scottish waters. Figure 6: Image 1

A new study has found that extreme weather conditions could lead to a spike in premature births. Researchers analysed data from more than 2.3 million births across 50 US states between 2000 and 2010. They found that an increase in the number of days with extremely hot temperatures was linked to a rise in premature deliveries. The research, published in the journal Environment International, calls for urgent action to address the impact of climate change on health. A study suggests extreme weather may cause more premature births in the US. Summary A study suggests extreme weather may cause more premature births in the US. Output from IG A new study has found that extreme weather conditions could lead to a spike in premature . Researchers analysed data from more than 23 million births across 50 US states between 2000 and 2010. They found that an increase in the number of days with extremely hot temperatures was linked to a rise in premature deliveries. The research, published in the journal Environment International, calls for urgent action to address the impact of climate change on health. Summary A study suggests extreme weather may cause more premature births in the US. Output from AIEG A new study has found that extreme weather conditions could lead to a spike in premature births . Researchers analysed data from more than 23 million births across 50 US states between 2000 and 2010. They found that an increase in the number of days with extremely hot temperatures was linked to a rise in premature deliveries. The research, published in the journal Environment International, calls for urgent action to address the impact of climate change on health. Summary: A study suggests extreme weather may cause more premature births in the US. Figure 7: Image 2

Parents who send their children to private schools may be wasting their money, according to a new report. Researchers claim that a state school education is just as likely to produce high-achieving pupils as private schooling. The study examined the academic achievements of more than 4,000 pupils at both types of school. It found that any advantage from private schooling "disappears" by the time pupils reach the age of 16. A report suggests state schools are just as good as private schools at producing high achievers.

Output from AIEG

Parents who send their children to private schools may be wasting their money, according a new report. Researchers claim that a state school education is just as likely to produce high-achieving pupils as private schooling. The study examined the academic achievements of more than 4000 pupils at both types of school. It found that any advantage from private schooling "disappears" by the time reach the age of 16.

A report suggests state schools are just as good as private schools at achievers. high

Output from IG

Parents who send their children to private schools may be wasting their money, according a new report. Researchers claim that a state school education is just as likely to produce high -achieving pupils as private schooling. The study examined the academic achievements of more than 4000 pupils at both types of school. It found that any advantage from private schooling "disappears" by the time reach the age of 16.

A report suggests state school are just as good as private schools at producing achievers

Figure 8: Image 3

A man who set up a fake hospital to offer illegal abortions has been sentenced to three years in prison. The man, known as Dr. Henry, had operated the clinic in Lagos for several years. Police said he lured women into the clinic by promising low-cost medical procedures. The court heard how he charged up to \$1,000 for each abortion and often left women with life-threatening complications. Summary: A man in Lagos has been jailed for running a fake clinic offering illegal abortions. **Output from AIEG** A man who set up a fake hospital to offer illegal abortions has been sentenced to three years in prison . The man, known as Dr Henry, had operated the clinic in Lagos for several years. Police said he lured women into the clinic by promising low -cost medical procedures. The court heard how he charged up to \$1000 for each abortion and often left women with life-threatening complications. Summary: A man in Lagos has been jailed for running a fake clinic offering illegal abortions **Output from IG** A man who set up a fake hospital to offer illegal abortions has been sentenced to three years in prison . The man, known as Dr Henry, had operated the clinic in Lagos for several years. Police said he lured women into the clinic by promising low -cost medical procedures. The court heard how he charged up to \$1000 each abortion and often left women with life threatening complications. Summary: A man in Lagos has been jailed for running a fake clinic offering illegal abortions Figure 9: Image 4

1026	
1027	
1028	
1029	
1030	
1031	
1032	
1033	
1034	
1035	
1036	
1037	
1038	The first images of the inside of a nuclear fusion experiment have been released by
1039	scientists. The pictures show the inside of the "tokamak" reactor at the UK Atomic
1040	Energy Authority's facility in Oxfordshire. The device is being used in research aimed at
1041	harnessing the power of nuclear fusion to generate electricity. It is hoped that this form of
1042	energy will one day provide a clean, virtually limitless source of power.
1043	Summary:
1044	Scientists have released the first images of a nuclear fusion experiment in the UK.
1045	
1046	Output from AIEG
1047	The first images of the incide of a nuclear fusion experiment have been released by
1048	scientists. The nictures show the inside of the "tokamak" reactor at the LIK Atomic
1049	Energy Authority 's facility in Oxfordshire. The device is being used in research singer at
1050	bernassing the newer of nuclear fusion to concrete electricity. It is being used that this provide
1051	of anorghy will and day provide a clean wirtually limitians equipe of neuron of neuron
1052	of energy will one day provide a clean , virtually limitless source of power.
1053	0
1054	Summary.
1055	Scientists have released the first images of a nuclear fusion experiment in the UK.
1056	538 L
1057	Output from IG
1058	The first mages of the inside of a nuclear fusion experiment have been released by
1059	scientists. The pictures show the inside of the "tokamak " reactor at the UK Atomic
1060	Energy Authority's facility in Oxfordshire. The device is being used in research aimed at
1061	harpessing the power of nuclear fusion to generate electricity. It is hoped that this form
1062	of energy will one day provide a clean, virtually limitless source of power
1063	or chergy will one day provide a clean, wradany innities source of power.
1065	Summary
1065	Scientists have released the first images of a nuclear fusion experiment in the LIK
1067	Scientists nave released the first images of a nuclear fusion experiment in the ort.
1062	
1069	
1070	Figure 10: Image 5
1071	
1072	
1073	
1074	
1075	
1076	
1077	

081	
082	
083	
084	
085	
1086	
1087	
1088	
1089	
1090	
1091	The number of homeless people in the UK has risen by 10% in the last year,
1092	according to government statistics. Charities have warned that the situation
1093	could worsen if measures aren't taken to provide more affordable housing.
1094	
1095	Summary:
1096	Homelessness in the UK has increased by 10% over the past year.
1097	
1098	
1100	
1100	The number of homeless people in the UK has risen by 10 % in the last year ,
1101	according to government statistics . Charities have warned that the
1102	situation could worsen if measures aren 't taken to provide more affordable
1104	housing
1105	nousing.
1106	
1107	Summary:
1108	Homelessness in the UK has increased by 10% over the past year
1109	
1110	Output from IG
1111	number of hemelage people in the LIK has risen by 10 % in the last
1112	Indifiber of nonneless people in the ok has risen by 10 % in the last
1113	year , according to government statistics . Charities have warned that
1114	situation could worsen if measures aren 't taken to provide more
1115	affordable housing.
1116	
1117	Summary:
1118	Homelessness in the UK has increased over the past year
1119	rendession in the orthan increased over the publication
1120	
1121	
1122	
1124	Figure 11: Image 6
1125	Figure 11: Image o
1126	
1127	
1128	
1129	
1130	
1131	

1135	
1136	
1137	
1138	
1139	
1140	
1141	
1149	
1143	
1144	
1145	
1146	A new study has shown that regular physical activity can help reduce the
1147	risk of heart disease. The research, conducted over a 10-year period, found
1148	that people who exercised for at least 30 minutes a day had a significantly
1149	lower risk of developing heart-related conditions.
1150	
1151	Summary:
1152	Regular exercise can lower the risk of heart disease according to a study
1153	Regular exercise our level the next of nour disouse, decording to a study.
1154	
1155	Output from AIEG
1156	A new study has shown that regular physical activity can help reduce the
1157	risk of heart disease. The research, conducted over a 10 -vear period
1158	found that needle who every include the st least 20 minutes a day had a
1159	found that people who exercised for at least 30 minutes a day had a
1160	significantly lower risk of developing heart-related conditions.
1161	
1162	Summary:
1163	Regular exercise can lower the risk of heart disease, according to a study
1164	Regular excluse carrier are not of near allocate, according to a staay
1165	
1166	Output from IG
1167	new study has shown that regular physical activity can help reduce the
1168	risk of heart disease. The research conducted over a 10 -vear period
1169	found that needle who eversiond for at least 20 minutes a day had
1170	found that people who exercised for at least 50 minutes a day had
1171	significantly lower risk of developing heart-related conditions.
1172	
1173	Summary:
1174	Regular exercise can lower the risk of heart disease, according to a study
1175	
1176	
1177	
1178	Figure 12: Image 7
1179	
1180	
1181	
1182	
1183	
1184	
1185	
1186	
1187	

1189	
1190	
1191	
1192	
1193	
1194	
1195	
1196	
1197	
1198	
1199	The local council has announced plans to build a new community park in the
1200	center of town. The park will feature playgrounds, walking paths, and areas for
1201	outdoor sports. Construction is expected to begin later this year.
1202	
1203	Summary:
1204	A new community park is set to be built in the town center, according to the local
1205	council.
1200	
1207	
1200	Output from AIEG
1200	The local council has announced plans to build a new community park in the
1210	center of town . The park will feature playgrounds , walking paths and areas for
1212	outdoor sports. Construction is expected to begin later this year.
1213	
1214	Summary
1215	A new community park is get to be built in the town center according to the local
1216	A new community park is set to be built in the town center, according to the local
1217	council
1218	
1219	Output from IG
1220	The local council has appounded plans to build a new community park in the
1221	conter of town. The park will feature playerounde, walking paths and areas for
1222	center of town. The park will reactive praygrounds, waiking parts and areas for
1223	outdoor sports. Construction is expected to begin later this year.
1224	
1225	Summary:
1226	A new community park is set be built in the town center, according to the local
1227	council.
1228	
1229	
1230	
1201	Einer 12. Incore 9
1232	Figure 15: Image 8
1234	
1235	
1236	
1237	
1238	
1239	
1240	
1241	

A fire broke out at a historic building in downtown London late last night. Firefighters worked through the night to contain the blaze, and no injuries have been reported. Investigators are now trying to determine the cause of the fire. Summary: A fire at a historic building in London has been brought under control. Output from AIEG A fire broke out at a historic building in downtown London late last night . Firefighters worked through the night to contain blaze , and no injuries have been reported. Investigators are now trying to determine the cause of the fire. Summary: A fire at a historic building in London has been brought under control Output from IG A fire broke out at a historic building in downtown London late last night . Firefighters worked through the night to contain blaze , and no injuries have been reported Investigators are now trying to determine the cause of the fire. Summary: A fire at a historic building in London has been brought under control Figure 14: Image 9

1296 A.6 VISUAL COMPARISONS AMONG THE METHODS

1298 In this section, we present additional visual examples comparing the attribution calculations for each method. We employed the GPT-2 (small) model, finetuned with the IMDb dataset Prerak et al. 1299 (2023) using 500 data points, and generated text based on small prompts. The generated sequences 1300 varied in length, randomly set to 50, 100, or 200 tokens. We then calculated the attributions for 1301 specific words of interest, as illustrated in Figures 15 to 29. The colour intensity of green reflects the 1302 magnitude of each word's contribution-darker green indicates a higher contribution. Across almost 1303 all examples, the AIEG method produced more interpretable and reasonable attributions compared to 1304 the IG method. 1305

1306	
1307	
1308	Input Prompt:
1309	Two Americans searching for work in Mexico convince an old prospector to help them mine for gold in the Sierre Madre Mountaine
1310	Sierra Maure Mountains.
1311	
1312	Generated Text:
1313	Two Americans searching for work in Mexico convince an old prospector to help them mine gold the Sierra
1314	Madre Mountains . The that is a single millimeter of their veins would be best deal they could get , but because. Word of Interest> deal
1315	
1316	
1317	
1318	Output from AIEG
1319	Americans searching for work in Mexico convince an old prospector to help them mine gold the Sierra
1320	Madre Mountains . The that 's a single millimeter of their veins would be best deal they could get , but because
1321	Output from IG
1322	Two Americans searching for work in Mexico convince an old prospector to help them mine gold the Sierra
1323	Madre Mountains . The that 's a single millimeter of their veins would be best deal they could get , but because
1324	Output from Inp*Grad
1325	Two Americans searching for work in Mexico convince an old prospector to help them mine gold the Sierra
1326	Madre Mountains . The that 's a single millimeter of their veins would be best deal they could get , but because
1327	Output from IGCG
1328	Two Americans searching for work in Mexico convince an old prospector to help them mine gold the Sierra
1329	Madre Mountains . The that 's a single millimeter of their veins would be best deal they could get , but because
1330	Output from DeepLift
1331	Two Americans searching for work in Mexico convince at the prospector to help them mine gold the Sierra
1332	Madre Mountains . The that 's a single millimeter of their veins would be best deal they could get , but because
1333	Output from GradShap
1334	Two Americans searching for work in Mayico convince an old prospector to help them mine gold the Sierra
1335	Madre Mountains. The that 's a single millimeter of their veins would be best deal they could get, but because
1336	
1337	Output from Attention-Only
1338	Americans searching for work in Mexico convince an old prospector to help them mine gold the Sierra
1339	Madre Mountains . The that 's a single millimeter of their veins would be best deal they could get , but because
1340	
1341	
1342	
1343	
1344	Figure 15: Image 1
1345	
1346	
1347	

	Input Prompt:
	A married daughter who longs for her mother's love is visited by the latter, a successful concert pianist.
	Generated Text:
	A married daughter who longs for her mother 's love is visited by the latter , a successful concert pianist . The
	couple have no children and all but four are raised his wife But their lives diverge sharply when they realize the
	mother.
	word of Interest> mother
	Output from AIEG
	A married doubter the lange for her mether's laws is visited by the latter a susceptible proved size in The
	A married daughter with longs for her mother's love is visited by the latter, a successful concert planist. The couple have no children and all but four are raised his wife But their lives diverge sharply when they realize th
	mother
	Output from IG
	A married daughter who longs for her active is love is visited by the latter, a successful concert planist. The couple have no children and all but four are raised his wife But their lives diverge sharply when they realize the
	mother
	Output from Inp*Grad
	A married doughter who longs for her water 's love is visited by the latter a successful concert nighter. The
	couple have no children and all but four are raised his wife But their lives diverge sharply when they realize the
	mother
	Output from IGCG
	A married daughter who longs for her mother s love is visited by the latter, a successful concert pianist. The
	couple have no children and all but four are raised his wife But their lives diverge sharply when they realize th
	mother
	Output from DeepLift
	A married daughter who longs for her nother love is visited by the latter , a successful concert pianist . The
	couple have no children and all but four are raised his wife But their lives diverge sharply when they realize the
	mother
	Output from GradShap
	A married daughter who longs for her mother ' love is visited by the latter , a successful concert pianist . The
	couple have no children and all but four are raised his wife But their lives diverge sharply when they realize th
	mother
	Output from Attention-Only
	married daughter who longs for her mother ' love is visited by the latter , a successful concert pianist . The
	couple have no children and all but four are raised his wife But their lives diverge sharply when they realize th
	mother
6	
	Figure 16: Image 2

lanu	t Prompt
A sir	nple act of kindness always sparks another, even in a frozen, faraway place.
	· · · · · · · · · · · · · · · · · · ·
Gene	Prated Text: male act of kindness always sparks another, even in a frazen faraway place. This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pain and
ange	ar.
Word	d of Interest> anger
Outp	ut from AIEG
A sir	nple 📰 of kindness always sparks another , even in a frozen faraway place This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pair and ar
anyo	A
Outr	but from IG
	nnia of kindness always snarks another, even in a frozen faraway nlace. This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pain and
ange	ar in the second se
Outr	out from Inp*Grad
Asi	to a group and of kindness always sparks another , even in a frozen faraway place . This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pain and
ange	ar an
Outp	ut from IGCG
A sir	nple er of kindness always sparks another , even in a frozen faraway place . This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pain and
ange	4
Outr	but from DeepLift
A	not find people and kind and a parke another alian in a frazer foreway alaba. This was are bably the series
for n	nost humans who encountered me too The fact that I would willingly let them go through my nain and
ange	۲ ا
Outr	out from GradShap
Asi	nple and of kindness always sparks another, even in a frozen faraway place. This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pain and
ange	er
Outp	ut from Attention-Only
sir	nple act of kindness always sparks another , even in a frozen faraway place . This was probably the cas
for n	nost humans who encountered me too The fact that I would willingly let them go through my pain and
ange	
	Figure 17: Image 3



1512	
1513	
1514	
1515	
1516	
1517	
1518	
1519	
1520	
1521	Input Prompt: A pragmatic U.S. Marine observes the dehumanizing effects the Vietnam War has on his fellow recruits from their brutal boot camp
1522	training to the bloody street fighting in Hue.
1523	Generated Text
1524	A pragmatic U .S Marine observes the dehumanizing effects Vietnam War has on his fellow recruits from their brutal boot camp
1525	training to bloody street fighting in Hue "The people who are still 'black wars ,' that is kind of war for , they 're ," he says The Americane a countrue bundred different cultures atheicities religions and languages. You see these folks have been sent here with
1526	no.
1527	Word of Interest> languages
1528	
1529	Output from AIEG
1530	pragmatic U.S. Marine observes the dehumanizing effects Vietnam War has on his fellow recruits from their brutal boot camp training to bloody street fighting in Hue 'The people who are still 'black wars I that is for a for . they 're .' he save The
1531	Americant a countrya hundred different cultures ethnicities religions and languages You see these folks have been sent here with no
1532	Output from IG
1533	A pragmatic 🛛 🔜 Marine observes the dehumanizing effects Vietnam War has on his fellow recruits from their brutal boot camp
1534	training to bloody street fighting in Hue 'The people who are still 'black wars ,' that is kind of war for theyre ," he says The Americans
1535	a countrya hundred different cultures ethnicities religions and languages fou see these folks have been sent here with no .
1536	Output from Inp*Grad
1537	pragmatic U.S Marine observes the dehumanizing effects Vietnam War has on his fellow recruits from their brutal boot camp
1538 1539	training to bloody street fighting in Hue "The people who are still 'black wars ,' that is kind of war for , they 're ," he says The Americans a countrya hundred different cultures ethnicities religions and languages You see these folks have been sent here with no
1540	Output from IGCG
1541	A pragmatic U Marine observes the dehumanizing effects Vietnam War has on his fellow recruits from their brutal boot camp
1542	training to bloody street fighting in Hue "The people who are still 'black wars ,' that is kind of war for theyre ," he says The Americans
1543	a countrya hundred different cultures ethnicities religions and languages. You see these folks have been sent here with no .
1544	Output from DeeoLift
1545	A pragmatic U Marine observes the dehumanizing effects Vietnam War has on his fellow recruits from their brutal hoot camp
1546	training to bloody street fighting in Hue 'The people who are still 'black wars ,' that is kind of war for theyre," he says The Americans
1547	a countrya hundred different pultures ethnicities religions and languages You see these folks have been sent here with no .
1548	Output from GradShap
1549	A promotion and a Marine alternation the debumanising offerse Westerner Washer as his follow service from their burget based on the
1550	raining to bloody street fighting in Hue 'The people who are still 'black wars,' that is kind of war for theyre,' he says The Americans
1551	a countrya hundred different cultures ethnicities religions and languages You see these folks have been sent here with no .
1552	Output from America Only
1553	
1554	pragmatic U.S Marine observes the penumanizing jettects Vietnam War has on his fellow recruits from their brutal boot camp training to bloody street fighting in Hue "The people who are still "black wars." that is kind of war for thevre." he savs The server
1555	a pountrya hundred different cultures ethnicities religions and languages You see these folks have been sent here with no .
1556	
1557	
1558	Figure 10. Image 5
1559	rigute 19. mage 5
1560	
1561	
1562	

Input Prompt: An executive of a shoe company becomes a victim of extortion when his chauffeur's son is kidnapped and held for ransom
Generated Text: An executive of a shoe company becomes victim extortion when his chauffeur 's son is kidnapped and held for ransom . A
school basketball coach , agent vice chancellor state are involved in double -murder their hometown which two clients kille
t know why they them but I imagine being hunted down like dogs pack wolves with sharpened teeth. Word of Interest> teeth
Output from AIEG
An executive of a shoe company becomes victim extortion when his chauffeur's son is kidnapped and held for ransom . A
school basketball coach, agent vice chancellor state are involved in double -murder their hometown which two clients kille
t know why they them but himagine being numee down like dogs pack wolves with sharpened teem
Output from IG
executive of a shoe company becomes victim extortion when his chauffeur is son is kidnapped and held for ransom . A
school basketball coach , agent vice chancellor state are involved in double -murder their hometown which two clients invo
don't know why they them but I imaging being hunted down like dogs pack wolves with sharpened teeth
Output from Inp*Grad
executive of a shoe company becomes victim extortion when his chauffeur 's son is kidnapped and held for ransom . A
school basketball coach , agent vice chancellor state are involved in double -murder their hometown which two clients kille
't know why they them but I imagine being hunted down like dogs pack wolves with sharpened teeth
Output from IGCG
An eventure of a shoe company becomes victim extortion when his chauffeur is son is kidnanned and held for ransom. A
school basketball coach, agent vice chancellor state are involved in double -murder their hometown which two clients kille
dont know why they them but I imagine being hunted down like dogs pack wolves with sharpened teeth
Output from Doord ift
output from DeepLint
An executive of a snoe company becomes victim extortion when his chauffeur's son is kidnapped and herd for ransom . A school basketball coach, agent vice chancellor state are involved in double -murder their hometown which two clients kille
dont know why they them but I imagine being hunted down like dogs pack wolves with sharpened teeth
Output from GradShap
An executive of a shoe company becomes victim extortion when his chauffeur's son is kidnapped and held for ransom. A school haskethall coach, agent vice chancellor state are involved in double amurder their hometown which two clients kilk
dont know why they them but I imagine being hunted down like dogs pack wolves with sharpened teeth
Output from Attention-Only
executive of a shoe company becomes victim extortion when his chauffeur's son is kidnapped and held for ransom . A
school basketball coach, agent vice chancellor state are involved in double -murder their hometown which two clients kille dont know why they them but Limanine being hunted down like doos pack wolves with sharpened teeth

1620	
1621	
1622	
1623	
1624	
1625	
1626	
1627	
1628	
1629	
1630	Input Prompt:
1631	After being kidnapped and imprisoned for fifteen years, Oh Dae-Su is released, only to find that he must find his captor in five days.
1632	Generated Text:
1633	After being kidnapped and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days . He
1634	then by YeonMin a man with no experience kidnapping YoonKol appears Season One where was present atoJ is funeral askedofang about her She told him she nurse which makes sense since patient but now.
1635	Word of Interest> patient
1636	
1637	Output from AIEC
1638	
1639	After being admappen and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days ; He then by YeonMin a man with no experience kidnapping YoonKol appears Season One where was present ato J is funeral askedofang
1640	about her She told him she nurse which makes sense since patient but now
1641	Output from IG
1642	After being adhapper and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days . He
1643	then by YeonMin a man with no experience kidnapping YoonKol appears Season One where was present atoJ is funeral askedofang
1644	about her she told him she nurse which makes sense since patient but now
16/5	Output from Inp*Grad
1646	After being actuappen and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days ; He then by YeonMin a man with no experience kidnanning YoonKol appears Season One where was present atoul is funeral askedofang
1647	about her She told him she nurse which makes sense since patient but now
16/18	Output from IGCG
16/10	After being adresses and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days . He
1650	then by YeonMin a man with no experience kidnapping YoonKol appears Season One where was present atoJ is funeral askedofang
1651	about her She told him she nurse which makes sense since patient but now
1652	Output from DeepLift
1653	After being advanced and imprisoned for fifteen years. Oh Dae -Su is released only to find that he must his captor in five days. He
1654	then by YeonMin a man with no experience kidnapping YoonKol appears Season One where was present atoJ is funeral askedofang
1655	about her She told him she nurse which makes sense since patient but now
1656	
1657	
1658	After being non-pres and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days ; He then by YeonMin a man with no experience kidnapping YoonKo l appears Season One where was present atoJ 's funeral askedofang
1650	about her She told him she nurse which makes sense since patient but now
1660	
1661	Output from Attention-Only
1662	being kidnapped and imprisoned for fifteen years , Oh Dae -Su is released only to find that he must his captor in five days . He
1663	about her She told him she nurse which makes sense since patient but now
166/	
1665	
1666	
1667	Figure 21: Image 7
1669	
1660	
1670	
1070	
1670	
1072	
10/3	

1674	
1675	
1676	
1677	
1678	
1679	
1680	
1681	
1682	
1683	
1684	Input Prompt
1685	A Phoenix secretary embezzles \$40,000 from her employer's client, goes on the run, and checks into a remote motel run by a young man under the domination of his mother.
1686	
1687	Generated Text:
1688	A Phoenix secretary embezzles \$40,000 from her employer's client goes on the run and checks into a remote motel by young man under domination
1680	of his mother . When FBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel). Meanwhile DEA investigation American woman allocady taking manage consistent threatening him evolution. One these consistence
1690	Word of Interest -> threatening
1691	
1692	Output from AIEG
1693	A Entry serveran embezzlas \$40,000 from her employer's client opes on the run and charks into a remote motel by young man under Reministran
1694	of his mother. When FBI enters building for more evidence she is suddenly arrested charged with a tender lide as an employee hotel). Meanwhile
1695	DEA investigating American woman allegedly taking money gangater threatening him execution One these conspirators
1696	Output from IG
1697	A Phoenix secretary embezzles \$40,000 from her employer's client goes on the run and checks into a remote motel by young man under domination
1698	of his mother . When FBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel). Meanwhile DEA investigating American woman allegedly taking money arrested threatening him execution One these conspirators
1699	Output from Inp®Grad
1700	M Phoenix secretary embezzles \$40,000 from her employer's client goes on the run and checks into a remote motel by young man under domination
1701	or his mother, when PBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel). Meanwhile DEA investigating American woman allegedly taking money gangster threatening him execution One these conspirators
1702	Output from IGCG
1703	A Process secretary embezzles \$40,000 from her employer's client goes on the run and checks into a remote motel by young man under domination
1704	of his mother . When FBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel). Meanwhile DEA investigating American woman allegedly taking money gangster threatening him execution One these conspirators
1705	Output from DeepLift
1706	A Property embezzles \$40,000 from her employer's client goes on the run and checks into a remote motel by young man under domination
1707	of his mother. When FBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel).
1708	Meanwhile UEA Investigating American woman allegedly taking money gangster threatening him execution Une these conspirators
1709	Output from GradShap
1710	A Phoenix secretary embezzies \$40,000 from her employer 's client goes on the run and checks into a remote motel by young man under domination
1711	of his mother. When FBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel).
1712	meaning out interaugeung American women anegedig taking money angene uneatening nim execution One trese conspirators
1713	Output from Attention-Only
1714	Phoenix secretary embezzles \$40,000 from her employer 's client goes on the run and checks into a remote motel by young man under domination
1715	of his mother . When FBI enters building for more evidence she is suddenly arrested charged with murder (because was an employee hotel). Meanwhile DEA investigating American woman allegedly taking money gangster threatening him execution One these conspirators
1716	meeting and an analysing remained memory analy menty gargers incorporation one meet complicately
1717	
1718	
1719	Figure 22: Image 8
1720	
1721	
1722	

lagut Dramati
A Polish Jewish musician struggles to survive the destruction of the Warsaw ghetto of World War II.
Generated Text:
A Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II . She also explores he
childhood , relationship with brother and struggle find home Her most recent tour dates begin on 20 March 14 18 M
2017 Tickets are available from www.zippospozier.com.
Output from AIEG
A Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II . She also explores he
2017 Tickets are available from the struggle find nome fier most recent tour pates begin on 20 March 14 18 M
Output from IG
Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II . She also explores he
childhood, relationship with brother and struggle find home Her most recent tour dates begin on 20 March 14 18 M
2017 Tickets are available from www.zippospozier.com
Output from Inp*Grad
Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II . She also explores he
chinghood, relationship with prother and struggle find home Her most recent tour dates begin on 20 March 14 18 M 2017 Tickets are available from www.zipposbozier.com
Output from IGCG
A Daliah la Mathematican struggles to supply the destruction of Warson abotto World War II. Che also evaluate he
childhood, relationship with brother and struggle find home Her most recent tour dates begin on 20 March 14 18 M
2017 Tickets are available from www.zippospozier.com
Output from DeepLift
A Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II . She also explores here
cniianooa , relationship with brother and struggle find home Her most recent tour dates begin on 20 March 14 18 M 2017 recent Tickets are available from www.zippospozier.com
Output from GradShap
A Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II . She also explores he
childhood , relationship with brother and struggle find home Her most recent tour dates begin on 20 March 14 18 M
2017 Tickets are available from www.zippospozier.com
Output from Attention-Only
Polish Jewish musician struggles to survive the destruction of Warsaw ghetto World War II. She also explores he
2017 Tickets are available from www.zippospozier com
Economic Contraction Contraction
Figure 23: Image 9

•	
	land Bronzet
	uput Frompt. During her family's move to the suburbs, a sullen 10-year-old girl wanders into a world ruled by gods, witches, and spirits, and wher
	humans are changed into beasts.
	Generated Text:
	During her family 's move to the suburbail a sullar 10 years of aid wanders into world suid by gode withher and epicite where
	humans are changed beasts . From start it difficult envision that any of has happened for these characters have been born throug
	magic from very early age But when things told them we see there is place find strength fight or die.
	Word of Interest> strength
	Output from AIEG
	During her family 's move to the suburbs , a sullen 10 -yearold girl wanders into world ruled by gods witches and spirits where
	maging from very early age But when things told them we see there is place the strength fight or die
	Output from IG
	During her family's move to the suburbs , a sullen 10 -yearold girl wanders into world ruled by gods witches and spirits where
	humans are changed beasts. From start it difficult envision that any of has happened for these characters have been born throug
	magio nom very early age par when mings told men we see mere is place ma sublight light of the
	output nom mp or du
	humans are changed beasts. From start it difficult envision that any of has happened for these characters have been born throug
	magic from very early age But when things told them we see there is place find strength fight or die
	Output from IGCG
	During her family move to the suburbs, a sullen 10 -yearold girl wanders into world ruled by gods witches and spirits where
	magic from very early age But when things told them we see there is place find strength fight or die
	Output from DeepLift
	During her family move to the suburbs, a sullen 10 yearold girl wanders into world nuled by gods witches and enirits where
	humans are changed beasts . From start it difficult envision that any of has happened for these characters have been born throug
	magic from very early age But when things told them we see there is place find strength fight or die
	Output from GradShap
	During her family move to the suburbs , a sullen 10 -yearold girl wanders into world ruled by gods witches and spirits where
	humans are changed beasts . From start it difficult envision that any of has happened for these characters have been born throug
	megre nom vary very age out when nings told men we see mele is piece ind sublight light of the
	Output from Attention-Only
	ber family's move to the suburbs, a sullen 10 yearold nirk wanders into world ruled hu wanders and entrine where
	humans are changed beasts. From start it difficult envision that any of has happened for these characters have been born through
	magic from very early age But when things told them we see there is place find strength fight or die
	Figure 24: Image 10
	-

1836	
1837	
1838	
1839	
1840	
1841	
1842	
1843	
1844	
1845	
1846	
1847	Input Prompt:
1848	A bounty hunting scam joins two men in an uneasy alliance against a third in a race to find a fortune in gold buried in a remote competery
1849	a remote centerery.
1850	Generated Text:
1851	A bounty hunting scam joins two men in an uneasy alliance against a third race to find fortune gold buried remote
1852	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight.
1853	Word of Interest> fight
1854	
1855	Output from AIEG
1856	A housty suffice scam joins two men in an upgasy alliance against a third race to find fortune gold burried remote
1857	cemetery, man and woman their way the right fork of busy river search rare gems - unterest start off as one long fight
1858	
1859	Output from IG
1860	A bounty sunting scam joins two men in an uneasy alliance against a third race to find fortune gold buried remote
1861	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight
1862	Output from Inp*Grad
1863	A bounty sutting scam joins two men in an uneasy alliance against a third race to find fortune gold buried remote
1864	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight
1865	Output from IGCG
1866	A bounty building scam joins two men in an uneasy alliance against a third race to find fortune gold buried remote
1867	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight
1868	Output from DeepLift
1869	A bounty suming scam joins two men in an uneasy alliance against a third race to find fortune gold buried remote
1870	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight
1871	
1872	Output from GradShap
1873	A bounty scam joins two men in an uneasy alliance against a thirdrace to find fortune gold buried remote
1874	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight
1875	
1876	Output from Attention-Only
1877	bounty hunting scam joins two men in an uneasy alliance against a third race to find fortune gold buried remote
1878	cemetery . man and woman their way the right fork of busy river search rare gems - hunters start off as one long fight
1879	
1880	
1881	Figure 25: Image 11
1882	
1883	
1884	
1000	
1886	
1887	

1890	
1891	
1892	
1893	
1894	
1895	
1896	
1897	
1898	
1899	Prompt: When the menace known as the Joker wreaks havoc and chaos on the people of Gotham, Batman must accept one of the
1900	greatest psychological and physical tests of his ability to fight injustice Generated Text:
1901	When the menace known as the Joker wreaks havoc and chaos on the people of Gotham, Batman must accept one of the greatest
1902	psychological and physical tests of his ability to fight injustice. In a small New York neighbourhood to protest the recent government decree prohibiting crime. Base and the city band meet on the street and
1903	decree promoting crime, bare and the dity band meet on the affect and
1904	Word of Interest: crime
1905	Output from AIPC
1906	
1907	when the memory hown as Joker wreaks navoo and chaos on the people of Gotham, Batman must accept one greatest psychological and physical tests of his ability to fight contine . In a small New York neighbourhood to protest the recent government
1908	decree prohibiting crime, Bane and the city band meet on the street and
1909	
1910	Output from IG
1911	When the menage known as Joker wreaks havoc and chaos on the people of Gotham, Batman must accept one greatest
1912	decree prohibiting crime, Bane and the city band meet on the street and
1913	
1914	Output from Inp*Grad
1915	When the menage known as Joker wreaks havoc and chaos on people of Gotham , Batman must accept one greatest psychological
1916	physical tests his ability to fight injustice . In a small New York neighborhood protest recent government decree prohibiting crime Bane city hand meet street
1917	
1918	Output from IGCG
1919	When the menage known as Joker wreaks havoc and chaos on people of Gotham , Batman must accept one greatest psychological
1920	physical tests his ability to fight injustice . In a small New York neighborhood protest recent government decree prohibiting crime
1921	Bane city band meet street
1922	Output from DeepLift
1923	When the menaue known as Joker wreaks havoc and chaos on people of Gotham , Batman must accept one greatest psychological
1924	physical tests his ability to fight injustice . In a small New York neighborhood protest recent government decree prohibiting crime
1925	Bane city band meet street
1926	Output from GradChan
1927	When the manual second
1928	physical tests his ability to fight injustice . In a small New York neighborhood protest recent government decree prohibiting crime
1929	Bane city band meet street
1930	
1931	Output from Attention-Only
1932	the menace known as Joker wreaks havoc and chaos on people of Gotham , Batman must accept one greatest psychological physical tests his ability to proteiniustice . In a small New York neighbourhood protest recent government decree prohibiting crime
1933	Bane city band meet street
1934	
1935	
1936	Figure 26: Image 12
1937	Figure 20: Image 12
1938	
1939	
1940	

1944	
1945	
1946	
1947	
1948	
1949	
1950	
1951	
1952	
1953	
1954	
1955	Input Prompt:
1956	A successful entrepreneur accused of murder and a witness preparation
1957	
1958	Generated Text:
1050	Word of Interest> murder
1060	A successful entrepreneur accused of murder and a witness preparation Alwere charged with . For the first time,
1061	police have launched an investigation for murder into two members Ravi family This it was second those to.
1062	
1062	Output from AIEG
1903	
1904	A successful entrepreneur accused or murder and a witness preparation Arwere charged with . For the first time,
1900	Output from IC
1900	
1967	A successful intropreneur accused of murder and a witness preparation Alwere charged with . For the first time,
1968	police have launched an investigation for more more more skaw family this it was second mose to
1969	Output from Inp*Grad
1970	A successful entrepreneur accused of murder and a witness preparation Alwere charged with . For the first time,
1072	police have launched an investigation for murder into two members kay family this it was second those to
1072	
107/	A successful entrepreneur accused of murder and a witness preparation Alwere charged with . For the first time ,
1075	police have launched an investigation for murder into two members kavi family This it was second those to
1975	Output from DeepLift
1077	A successful entrepreneur accused of nurder and a witness preparation Âlwere charged with . For the first time,
1079	police have launched an investigation for murder into two members Ravi family This it was second those to
1970	Output from GradShap
1979	A successful entrepreneur accused of murder and a witness preparation Aflwere charged with . For the first time ,
1001	police have launched an investigation for into two members Ravi family This it was second those to
1000	Output from Attention-Only
1002	successful entrepreneur accused of murder and a witness preparation ÅĤwere charged with . For the first time ,
1000	police have launched an investigation for into two members Ravi family This it was second those to
1005	
1006	
1007	
1000	
1000	Figure 27: Image 13
1909	rigute 27. inluge 15
1990	
1991	
1992	
1993	
1994	



2052	
2053	
2054	
2055	
2056	
2057	
2058	
2059	
2060	
2061	Input Prompt
2062	Paul Rusesabagina, a hotel manager houses over a thousand Tutsi refugees during their struggle against the
2063	Hutu militia in Rwanda, Africa.
2064	
2065	Concentral Tauto
2066	Generated Text: Paul Rusesabanina - a hotel manager houses over thousand Tutsi refugees during their struggle against the
2067	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US.
2068	Word of Interest> army
2069	
2070	
2071	Output from AIEG
2072	Paul Rusesabagina, a hotel manager houses over thousand Tutsi refugees during their struggle against the
2073	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US
2074	Output from IG
2075	Paul Rusesabagina , a hotel manager houses over thousand Tutsi refugees during their struggle against the
2076	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US
2077	Output from Inp*Grad
2078	Paul Rusesabagina, a hotel manager houses over thousand Tutsi refugees during their struggle against the
2079	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US
2080	Output from IGCG
2081	Paul Rusesabagina, a hote manager houses over thousand Tutsi refugees during their struggle against the
2082	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US
2083	Output from DeepLift
2084	Paul Rusesabagina, a hotel manager houses over thousand Tutsi refugees during their struggle against the
2085	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US
2086	
2087	Output from GradShap
2088	Paul Rusesabagina, a hotel manager houses over thousand Tutsi refugees during their struggle against the
2089	Hutu militia in Rwanda Africa . She said Rwandan government supported Tut army US
2090	
2091	
2092	Rusesabagina, a hotel manager houses over thousand Tutsi refugees during their struggle against the
2093	nutu militia ni kwanda Amca . She salu kwandan government supported Tut army US
2094	
2095	
2096	
2097	$\Gamma'_{2} = 20$ Let 16
2098	Figure 29: Image 15
2099	
2100	
2101	
2102	

2106 A.7 GRAPHICAL VISUALISATION OF THE EF AND AIEG VALUES

In this section, Figures 30 to 34 present graphical representations of various metrics: IG values vs.
Tokens, EF (Exponential Factor) vs. Alpha, Output vs. Alpha, (Output x EF) vs. Tokens, and AIEG
values vs. Tokens for short sentences. These visualizations provide valuable insights into the behavior
of our proposed method and highlight the key differences compared to the standard IG approach.



