

Exploring Taxonomic Emotion Detection via Saliency Mapping and Large Language Model Context Generation

Tawfik Ali* Ala Omidfar* Sai Phalguna*
Raaneya Dehlavi* Kevin Zhu Aslihan Akalin Sean O'Brien

Abstract

Detecting human emotions is a complex process often restricted by predetermined emotion labels, limiting the ability of systems to recognize nuanced or novel emotional expressions. To overcome this constraint, we explore a generative approach that uses large language models (LLMs) context creation and analyze the results with the usage of saliency maps. Instead of relying solely on static sample texts, our method generates dynamic backstories that provide additional context for understanding emotions in text or speech, and these backstories improved our evaluations. This novel framework re-imagines emotion detection as a context-driven process, which in turn enables systems to "see the bigger picture" and make more informed and accurate predictions based on added adjectives, nouns, and punctuation.

1 Introduction

Human emotions are complex and multifaceted, often defying the boundaries of labels such as "happy" or "sad". This limitation has significant challenges for emotion detection systems, particularly in zero-shot scenarios where the system must identify emotions it has not been trained on yet (Singh et al., 2024). Current approaches often fail to capture the subtlety in emotional nuance, leaving room for improvement in how machines interpret human feelings. Emotion detection methods use supervised learning with set emotion categories, relying on deep language models trained on labeled data (Chutia and Baruah, 2024). In some situations, emotions are context-dependent and explaining Demeszky and colleagues' 0.40 precision compared to our 0.60 precision, and the reason why these approaches might struggle while recognizing subtle emotions and applying what they learn beyond their training because they often can't take into account context, which can lead to misunderstandings (Demeszky et al., 2020). However,

• Common Approach (Without Backstory)



• Enhanced Saliency Methods

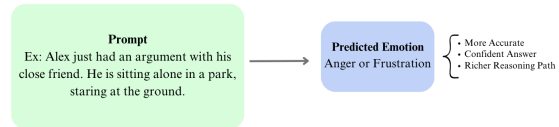


Figure 1: This diagram illustrates how providing backstories can improve the saliency and reasoning of the mode behind every decision.

emotions are rather context dependent which is why we started with backstory generation. For example, rather than having "The sun is shining" alone versus the addition of "It is my birthday today." provides some excitement. To address this challenge, we proposed a novel approach that reimagines emotion detection as a context-driven process. Using large language models (LLMs), we generate backstories to provide context for the emotional content. These backstories enable systems to "see the bigger picture" and make better predictions, bridging raw inputs and accurate emotional interpretation, as shown in Figure 1. We introduced a backstory generation framework to analyze and explore how factors influence emotion classification in language models; backstories permit us to extract insight through their corresponding saliency maps. By using Saliency mapping methods, we examine which contextual elements - such as nouns, adjectives, and punctuation - contribute most to emotion detection and how these factors vary across different emotions. Through our approach, we provide insights into the types of linguistic features such as nouns, adjectives, etc. Rather than improving classification accuracy, this work explores how language models interpret emotional reasoning and the effect of properties that drive their predictions.

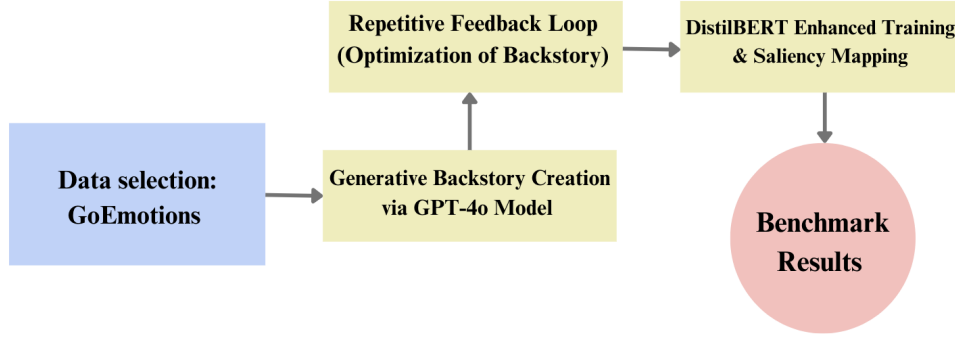


Figure 2: Overview of the proposed workflow for emotion detection

2 Related Work

2.1 Fine-Tuning BERT

A popular classification method involves fine-tuning pre-trained language models such as BERT, which has shown significant results in various natural language understanding tasks. The strategies for this model include preprocessing long texts, layer selection, and layer-wise learning rates (Sun, 2024). Although fine-tuning BERT has been successful, it is constrained by a fixed set of emotion labels, which impacts the recognition of emotional expressions. Considering these limitations, we use an LLM to generate context and allow for a more flexible approach to emotion labels.

2.2 Emotion Detection: Diverse Approaches

Emotion detection is a critical challenge in NLP focused on assigning emotions to text. In traditional approaches, convolutional models, recurrent models, and attention mechanisms were used to learn text representations. However, facial expression analysis often falls short due to limited contextual understanding and factors like angles and obstructions that can reduce accuracy (Anwarul and Dahiya, 2020). Our approach not only can efficiently address such limitations by underscoring contextual understanding and promote saliency mapping to detect emotion in real-life scenarios by feeding videos to large models like GPT-4o which support video input, but can efficiently address such limitations by underscoring contextual understanding and promote saliency mapping to detect emotion in real-life scenarios by feeding videos to large models like GPT-4o which support video input. Combining these techniques will help overcome these constraints and lead to a more robust path for emotion recognition.

2.3 Saliency Mapping

Saliency mapping is commonly used to interpret models and visualize the necessity of features. Gradient-based methods highlight key inputs that influence the predictions; adaptations to NLP tasks help identify important phrases that can be helpful for tasks including emotion detection and classification (Simonyan et al., 2014).

3 Methodology

3.1 Addressing Limitations

Our approach addresses some limitations of traditional emotion detection systems by introducing a generative framework utilizing LLMs’ capabilities of generating a backstory so that we can have more to analyze about how and what LLMs understand about emotion detection. At the center of our approach is the GPT-4o model, which we used to create backstories for text samples provided in a GoEmotions dataset (Demszky et al., 2020). These backstories provide a contextual grounding for attaching emotional labels to text.

3.2 Cyclic Generative Process

This generative process includes a repetitive feedback loop to optimize backstories; the sample text must be aligned with the generated backstory with the prompt provided to the GPT-4o model, where we include the sample text along with asking it to question every sentence it generates; each single backstory is then referred back to GPT-4o five times in a feedback loop. With these backstories, we integrated saliency mapping to our pipeline to identify the words that the LLM finds integral for emotional prediction. Overall, through benchmarking emotion detection precision, F1-Score, and accuracy we compared the compatibility of our approach and how the integration of generative context along

with saliency mapping can provide insight to how these Large Language Models BERT in this situation analyze and detect emotion throughout text and context.

3.3 Backstory Generation

GPT-4o is used to produce backstories for each sample in the GoEmotions dataset. These contextual statements serve as a foundation for a better emotional classification. The process is cyclical, incorporating a feedback loop of providing the previous generated backstory to GPT and the original sample and asking questions such as "Provide insightful marks of punctuation that highlight emotion according to the sample text".

3.4 Saliency Mapping

To improve interoperability, we incorporate saliency mapping techniques. These methods identify words in the text that the model considers crucial for emotional prediction by providing insights into the model's decision-making process. This additionally gives insight into the field of how these transformation models work exactly and which words most resonate with the models for better predictions in the future. An example that could be looked at is Figure 4 which showcases the key of what each highlighting color means. High scoring saliencies according to our saliency maps indicate a high probability of that word or token to be of high priority in emotion classification. This would highlight specific words that demonstrate emotional alignment, and then identify tokens with the most accuracy to cross-validate against the markers reported in GoEmotions, reinforcing the reliability of our approach. Overall, the analysis of tokens highlights the emotional cues of certain parts of the text to help future models in predicting emotions beyond the GoEmotions dataset. In terms of saliency mapping, we utilized this technique to identify which parts of the text would most influence the output, which allows us to highlight specific words that demonstrate emotional alignment and then identify tokens with the highest accuracy to cross-validate against the markers reported in GoEmotions, reinforcing the reliability of our approach. Overall, token analysis highlights the emotional cues of certain text parts to help future models predict emotions beyond the GoEmotions dataset.

3.5 Model

To achieve our goal of understanding emotion prediction with textual passages, we utilized a fine-tuned BERT model using a strategy inspired by the GoEmotions dataset. With this in mind, we achieve a multi-label classification of emotions, we trained the DistilBERT model utilizing the GoEmotions dataset and the integration of saliency mapping, passage scoring, and backstory generation (via GPT-4o) to all be used to analyze emotion classification and create this fine-tuned model. The loss across the model was monitored to adjust depending on any overfitting or underfitting.

3.6 Data Preparation & Processing

To process the GoEmotions dataset, we transformed these labels into multi-hot vectors, as seen in Table 1. With this in mind, we then tokenized text with the DistilBERT tokenizer with a maximum of 512 tokens. Each multi-hot vector represents an emotion and the model assigns either a 0 or 1 to each one if it classifies it as that emotion.

3.7 Fine-Tuned Model System

Our adjustments to the DistilBERT fine-tuned model was sufficient for the classification of emotions in the GoEmotions dataset. To be more specific, GoEmotions is composed of 58K Reddit comments that are each labeled for at least one of 28 emotion(s) found in the appendix (Demszky et al., 2020). The overall system is the DistilBERT model being trained by the GoEmotions dataset and then trained again based off of context-rich versions of the same samples; then, we ask it to attach an emotion to a text sample from a selection of 27 emotion(s) or neutral as seen in the benchmark results section of Figure 2. Our fine-grained emotion classification over GoEmotions was showcased the potential the GPT-4o model has in context generation if prompted to be aligned with a sample text. We utilized a promising architecture and employed the DistilBERT pre-trained model, specifically fine-tuned for multi-label classification. An output layer was added to the BERT's hidden layer to train the model to support the prediction of the 28 emotion categories. Our loss function was a binary cross-entropy loss to handle those multi-label predictions. Enhancing the data was done by including generated backstories. We trained this system through gradient scaling over batches of size 16 to max-

imize mixed-precision computation (Kashi et al., 2025). We utilized the OpenAI GPT-4o model to generate contextual backstories as training samples for the BERT-based model. This augmented training data with backstories to provide more contextual understanding and less probabilistic predictions for the model. Each backstory was created and optimized through a feedback loop, as seen in Table 3.

Text Sample	Emotion Labels	Multi-Hot Vector
"That game hurt."	Sadness	[0, 0, 0, 0, 0, . . .]
"I'm glad he's okay but I'm even gladder [...]"	Admiration	[0, 0, 0, 0, 1, . . .]
"Lol. Build your own or make some tweaks to the printer."	Anger	[0, 0, 1, 0, 0, . . .]

Table 1: Mapping Text Samples to Multi-Hot Vectors

3.8 Evaluation & Metrics

The metrics used were precision, F1 score, and recall for each emotional and sentimental category. Each model was separately evaluated and compared after each run, in general, several subsets.

4 Results & Analysis

The saliency maps illustrate that the model emphasizes changes depending on the context, even when the same word is used. For example, the word "is" is relevant in the anger context in Figure 5 but less so in Figure 14. This highlights the model's capability to understand context when determining which words contribute to a particular emotion. For **Anger**, which is shown in Figure 5, the words **"deleting," "stop," "reason," and "under"** are strongly associated with anger in the sentence, "People deleting and stop caring IS THE REASON why artifact is under 2k players". As mentioned before, the word "is" is also highlighted, indicating its importance in linking "reason" and "under." In Figure 5, the same sentence is shown with a different pattern, with **"deleting," "stop," and "caring"** being highlighted with the highest confidence. "reason" is highlighted with less intensity, whereas "under" is not. This shows that the model can concentrate on key elements of anger, and different words can be significant depending on the context, even within the same sentence. Additionally, the model can identify natural text by marking words without any particular emotional association, meaning it understands how neutrality manifests in language, where confident-scoring words almost equal non-confident-scoring words. Table 2 evokes the model's performance in terms of precision, recall, F1 score, and precision to detect

specific emotions. These metrics show how well the model captures the emotions related to specific taxonomies.

Sentiment	Precision	Recall	F1-score	Accuracy (%)
Neutral	0.66	0.23	0.34	79.10
Positive	0.74	0.76	0.75	79.10
Negative	0.62	0.66	0.64	79.10
Macro Avg	0.67	0.55	0.58	79.10
Std	0.05	0.22	0.17	N/A

Table 2: Fine-tuned Model Sentimental Group Analysis-GPT-4o Backstory Generation

4.1 Fine-Tuned BERT Model Results

According to Table 2, our fine-tuned model appears to have greater difficulty in identifying emotions that were 'neutral.' This increase in precision showcases the promise in backstory generation along with saliency mapping for that in our initial experimentation. However, just having an LLM predict the emotion performed worse than Demszky and colleague's benchmark which showcases that providing context to train any model permits for increased performance, even if it's minimal (Demszky et al., 2020).

4.2 Saliency Map Results

4.2.1 Emotion Analysis

The first emotion that should be looked at for its saliency map would most likely be that of remorse for its extremely high accuracy of 99.8% and precision of 0.60, according to Table 3. The saliency map in the appendix Figure 7 shows that the word of most significance was "meant," along with the exclamation mark having relatively high relevance. This suggests that these transformer (DistilBERT in this case) models most likely after long training focus on conjugated verbs, punctuation, nouns, and adverbs because those more significantly more common and differentiable than articles like "the", "an", and "a". This can further be seen in Figure 8 where there is the backstory of said emotion where any noun such as "fellow" is of high confidence to the model, along with adjectives like "collaborative," which showcases the great potential of these models to identify keywords that would signify emotion, similar to how any human would in a texting scenario.

4.2.2 Neutrality Identification

The model identified neutrality within the data points. As shown in Figure 9, if the model was

neutral about the text or lacked confidence in it, then it's likely a neutral text. This also asserts the non-confident scoring words almost equal to the confident scoring words and tokens. Neutrality to the model seems to be its attitude towards the text, once again providing insight into how an LLM like BERT can do natural language processing tasks efficiently. Indeed, according to Figure 9, it marked three with saliency scores higher than 0.7 and around 3 with saliency scores less than 0.4. What it means is that the amount of high-confidence tokens that it marked was usually equivalent to those that were of low-confidence, and as a result, it is likely that is why neutrality was picked. This is a pure coincidence, but this occurred with Figure 10, a much longer text due to our increased token length in generated backstories. There is only one word in it, which happens to be a noun with a saliency score of 1.00, indicating emotion is present due to that one word. However, the abundance of words in the phrase being less than 0.50 saliency like "road" means the model the neutrality of the emotion; as a result, the fine-tuned model can correctly identify neutrality that there is nothing that was too blatant indicating emotion.

4.2.3 Punctuation Saliency

Another underlying idea would be punctuation; in fact, Figure 11 with the exclamation mark has a saliency score of 1.00 9. By treating punctuation as words effectively helps the model identify emotions in text, supported by Figure 11. This emphasizes that containing punctuation is essential for fine-tuned NLP models in emotion recognition.

4.2.4 Adjective Saliency

A key insight from saliency maps is the monumental role of adjectives in the model's emotion classification. Figures 12 and 13 demonstrate how strongly adjectives can contribute to the model's confidence in identifying positive emotions, such as admiration. Words such as 'amazing' and 'fun' had high saliency scores ranging from 0.80 to 1.00 in Figure 12, conveying the critical role that adjectives can play. Similarly, in Figure 14, the adjective 'good' was indicated as the most salient, further reinforcing that adjective play a crucial role in indicating emotional sentiment. Suppose this pattern of high saliency scores for adjectives appears consistently across multiple samples. In that case, they are highly significant to the model's emotion prediction process, suggesting that it relies heavily on

descriptive language to assess emotional sentiment. In contrast, Figure 15 highlighted how adjectives can contribute to negative emotions, specifically disappointment, assigning high saliency scores to adjectives such as 'competitive' with a score of 0.80, encouraging the idea that descriptive terms drive emotional classification. This pattern in our results indicates the significance of adjectives as key decision-making factors for fine-tuned models, aligning with human intuition that descriptive words shape emotional interpretation.

5 Conclusion

This paper explored the emotion detection of large language models (LLMs), specifically DistilBERT, in detecting subtle emotions by generating background stories for given sample texts. This approach provided additional context, enabling the LLM to make more informed emotional predictions. Additionally, we incorporated saliency maps to visualize the key elements the model focuses on when analyzing both the text and its background story. Our findings indicate that nouns, adjectives, and even punctuation significantly influence emotion detection, highlighting the effectiveness of contextual enrichment and interpretability in exploring and understanding model performance.

6 Limitations

Although our approach improves emotion detection systems by integrating generative backstories and saliency mapping, several limitations may warrant further investigation and experimentation. We hope to better understand our findings and guide future research by understanding these limitations.

6.1 Saliency Maps

To confirm that our saliency maps were not coincidental, we tested the dataset again and the confidence score when removing the high-confidence tokens. This can be seen in Figure 6, where the saliency map was dull after removing the high-confidence saliencies and had near-zero confidence. The comparison between Figure 5 and 6 proves that our saliency maps are valid; this is because removing certain sections that are of the highest importance to the model and comparing the saliency map before and after indicates that there no longer is any saliency in the sample text. (Gomez et al., 2022).

6.2 Model Dependency

Our methodology and framework rely on GPT-4o to generate backstory contextual understanding, but this has some limitations. As noted in offshore alignment between pre-trained models (PLMs) and parameter-efficient methods, it often exhibits inconsistencies due to the objectives not being downstream fine-tuning tasks (Ma et al., 2023). In our case, this may lead to the generated backstories producing content that fails to encapsulate our standards of nuanced emotion (such as redundant or simplistic text). The challenge lies in prompt-based fine-tuning, where the quality of the generated prompts significantly influences the results (Sun, 2024).

6.3 Task Specificity

The methodologies developed to evaluate backstory alignment and saliency mapping were explicitly optimized for the GoEmotions dataset. Although this dataset is a strong and flexible resource for emotion detection, it contains only 27 categories. While said categories are robust, they are not complex enough to capture the true nuance of human emotions. The topic of cross-task generalization is also to be discussed. Using cross-task generalization through multitask learning (MTL) could significantly enhance the strength of our proposed framework (Yu et al., 2024).

6.4 Computational Constraints

The process of backstory generation imposes a significant computational overhead, which has been noted in previous research regarding parameter-efficient fine-tuning methods. The trade-offs often determine the deployability of these models in real-world applications so integrating better solutions such as adapters and prompt tuning may be able to address this (Huang et al., 2024).

6.5 Future Research

While our approach represents an innovative advancement in emotion detection, critical improvements to enhance applicability are still needed. Future endeavors should address these limitations by incorporating diverse datasets, exploring parameter-efficient fine-tuning strategies, and experimenting with PLMs being GPT-4o to provide a more robust and effective approach.

References

- Shahina Anwarul and Susheela Dahiya. 2020. *A Comprehensive Review on Face Recognition Methods and Factors Affecting Facial Recognition Accuracy*, pages 495–514.
- Tulika Chutia and Nomi Baruah. 2024. *A review on emotion detection by using deep learning techniques*. *Artificial Intelligence Review*, 57.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. *GoEmotions: A dataset of fine-grained emotions*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- Tristan Gomez, Thomas Fréour, and Harold Mouchère. 2022. *Metrics for saliency map evaluation of deep learning explanation methods*. *Preprint*, arXiv:2201.13291.
- Lianghua Huang, Wei Wang, Zhi-Fan Wu, Huanzhang Dou, Yupeng Shi, Yutong Feng, Chen Liang, Yu Liu, and Jingren Zhou. 2024. *Group diffusion transformers are unsupervised multitask learners*. *Preprint*, arXiv:2410.15027.
- Aditya Kashi, Hao Lu, Wesley Brewer, David Rogers, Michael Matheson, Mallikarjun Shankar, and Feiyi Wang. 2025. *Mixed-precision numerics in scientific applications: survey and perspectives*. *Preprint*, arXiv:2412.19322.
- Bolei Ma, Ercong Nie, Helmut Schmid, and Hinrich Schütze. 2023. *Is prompt-based finetuning always better than vanilla finetuning? insights from cross-lingual language understanding*. *Preprint*, arXiv:2307.07880.
- Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. *Deep inside convolutional networks: Visualising image classification models and saliency maps*. *Preprint*, arXiv:1312.6034.
- Gopendra Vikram Singh, Mauajama Firdaus, Dushyant Singh Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2024. *Zero-shot multitask intent and emotion prediction from multimodal data: A benchmark study*. *Neurocomputing*, 569:127128.
- Pingwei Sun. 2024. *Fine-tuning vs prompting, can language models understand human values?* *Preprint*, arXiv:2403.09720.
- Jun Yu, Yutong Dai, Xiaokang Liu, Jin Huang, Yishan Shen, Ke Zhang, Rong Zhou, Eashan Adhikarla, Wenxuan Ye, Yixin Liu, Zhaoming Kong, Kai Zhang, Yilong Yin, Vinod Nambodiri, Brian D. Davison, Jason H. Moore, and Yong Chen. 2024. *Unleashing the power of multi-task learning: A comprehensive survey spanning traditional, deep, and pretrained foundation model eras*. *Preprint*, arXiv:2404.18961.

A Appendix

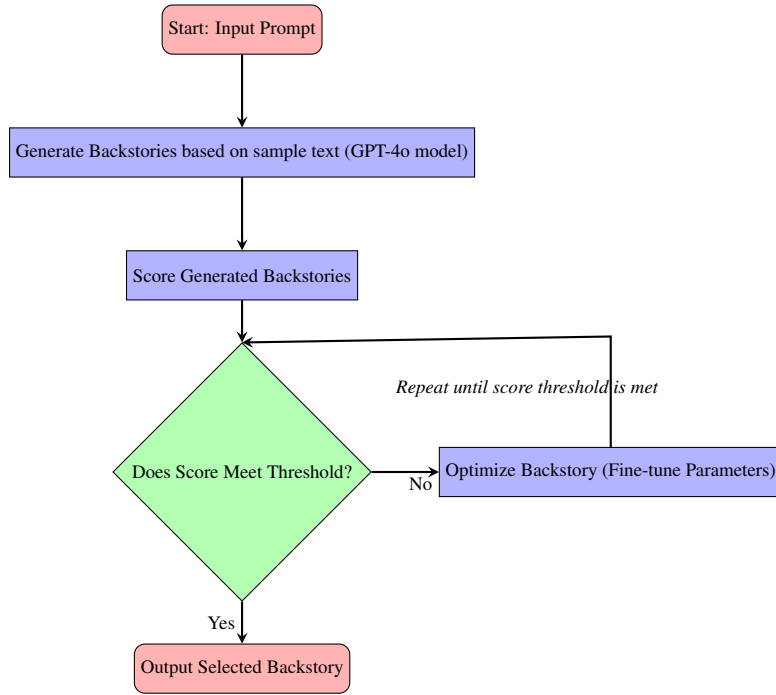


Figure 3: Flowchart showcasing the context generation process

Emotion	Precision	Recall	F1-score	Accuracy (%)
admiration	0.61	0.80	0.69	93.35
amusement	0.75	0.89	0.82	98.07
anger	0.53	0.45	0.49	96.54
annoyance	0.47	0.27	0.34	93.94
approval	0.48	0.35	0.40	93.31
caring	0.48	0.36	0.41	97.42
confusion	0.47	0.42	0.44	97.03
curiosity	0.48	0.67	0.56	94.49
desire	0.60	0.37	0.46	98.65
disappointment	0.47	0.22	0.30	97.14
disapproval	0.43	0.43	0.43	94.38
disgust	0.54	0.46	0.50	97.88
embarrassment	0.71	0.41	0.52	99.48
excitement	0.46	0.44	0.45	97.95
fear	0.58	0.68	0.62	98.82
gratitude	0.92	0.92	0.92	98.91
grief	0.00	0.00	0.00	99.89
joy	0.58	0.65	0.61	97.57
love	0.74	0.86	0.79	98.05
nervousness	0.33	0.39	0.36	99.41
optimism	0.62	0.53	0.57	97.27
pride	0.71	0.31	0.43	99.76
realization	0.35	0.19	0.25	96.90
relief	0.00	0.00	0.00	99.80
remorse	0.60	0.88	0.71	99.26
sadness	0.59	0.53	0.56	97.60
surprise	0.51	0.62	0.56	97.49
neutral	0.63	0.74	0.68	77.22
Micro Average	0.60	0.62	0.61	–
Macro Average	0.52	0.49	0.50	–
Weighted Average	0.59	0.62	0.59	–
Samples Average	0.61	0.64	0.61	–

Table 3: Model Performance - GPT-4o Backstory Generation and Saliency Mapping

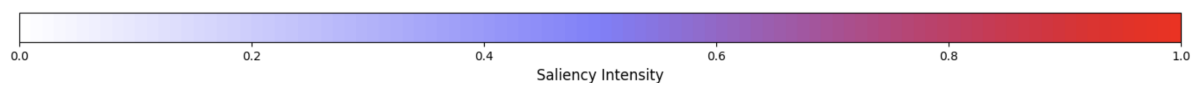


Figure 4: Saliency Intensity where red indicates locations of highest confidence and importance, while white is of the least confidence.

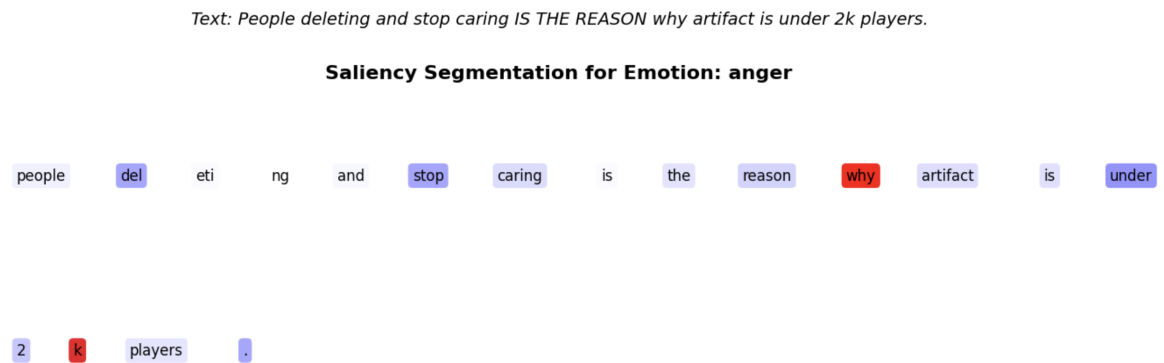


Figure 5: Anger saliency map with text sample from the GoEmotions dataset.

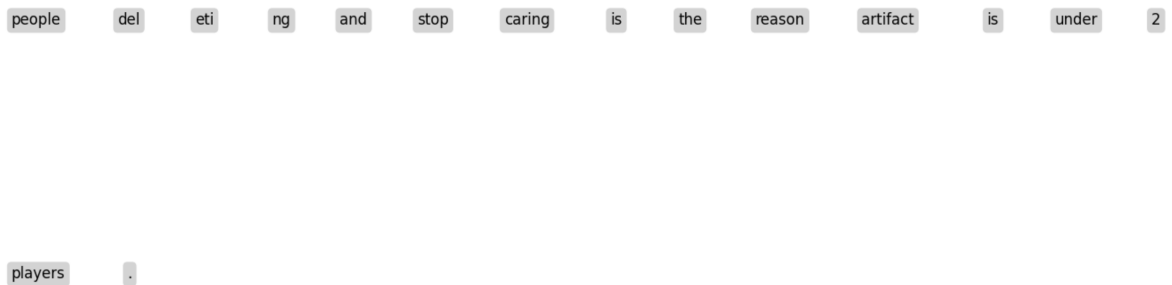


Figure 6: Removal of any words of high confidence to prove saliency accuracy.

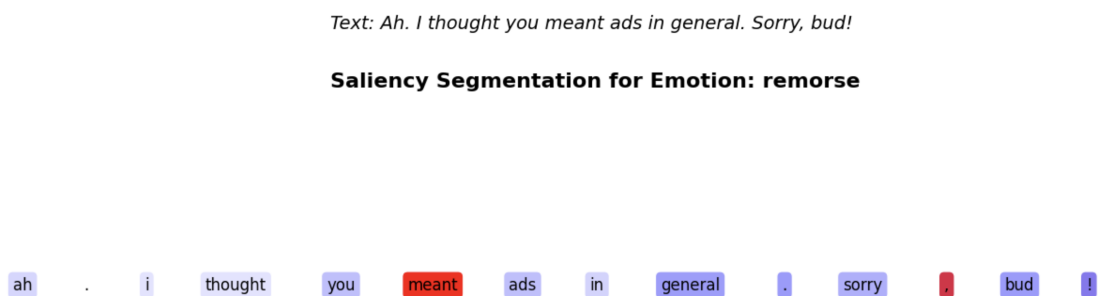


Figure 7: 'remorse' emotion from the GoEmotions dataset which highlights saliency map example.

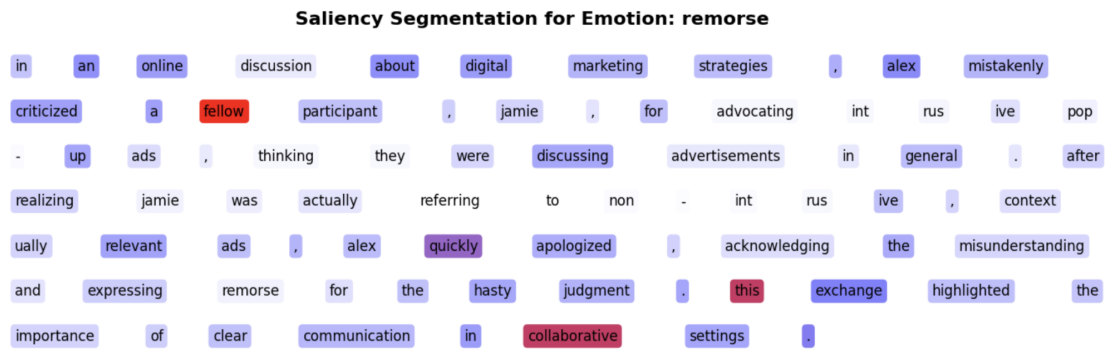


Figure 8: Saliency map of GPT-4o feedback-looped backstory generation of a sample text regarding the 'remorse' emotion category for the previous Figure 4 example.

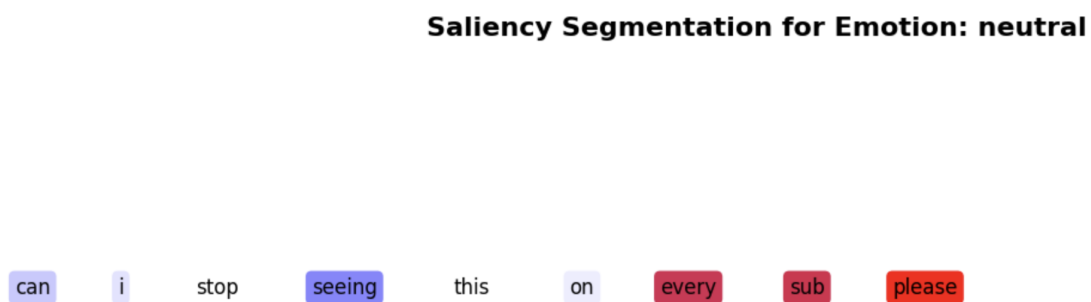


Figure 9: Neutral saliency map which highlights a sample text from the GoEmotions data set.

Text: I was well expecting the glee with which much of the userbase here would celebrate roadblocks being put up in front of "the other".

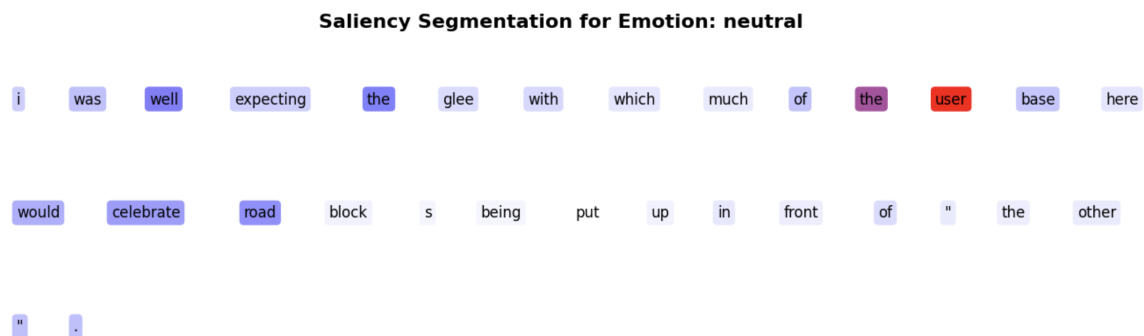


Figure 10: This is a saliency map on a neutral sample text with context generated by GPT-4o.

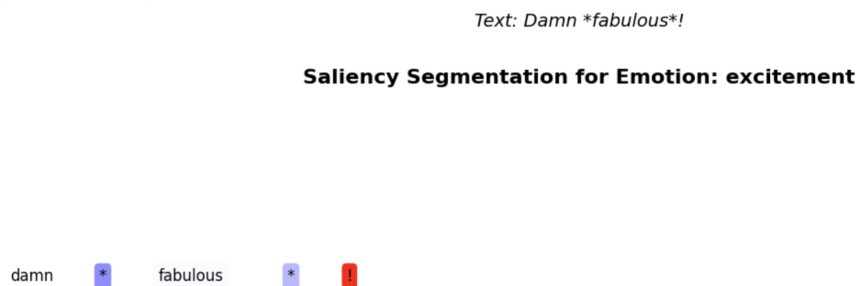


Figure 11: A sample text indicating excitement saliency map with a high scoring of emotion in punctuation.

Text: "If you don't wear BROWN AND ORANGE...YOU DON'T MATTER!" We need a tshirt with that on it asap!

Saliency Segmentation for Emotion: annoyance

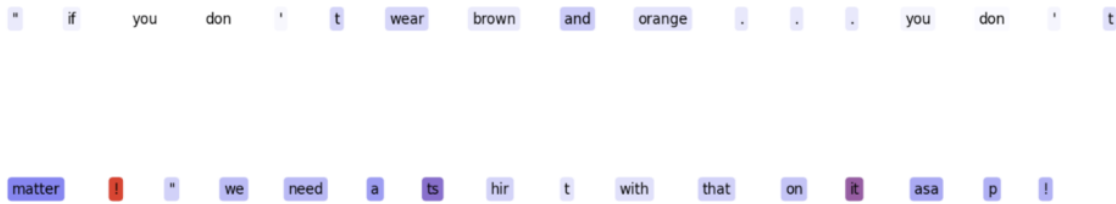


Figure 12: Another example of punctuation (an exclamation mark) indicating emotion (high saliency scores

Text: I absolutely love that idea. I went on an anniversary trip with a couple once and it was amazing! We had so much fun.

Saliency Segmentation for Emotion: admiration

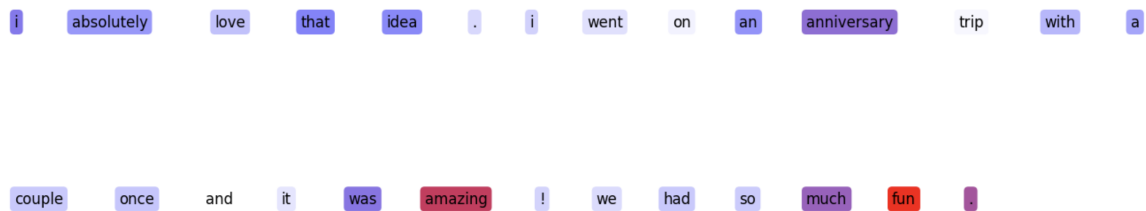


Figure 13: Adjective Saliency Analysis which gives insight to how adjectives make up a large portion of the model's confidence in selecting an emotion that corresponds to a text sample.

Text: This is good

Saliency Segmentation for Emotion: admiration



Figure 14: Another adjective saliency analysis provides evidence of adjectives being a major deciding factor for positive emotions identification for these fine-tuned models.

Saliency Segmentation for Emotion: disappointment

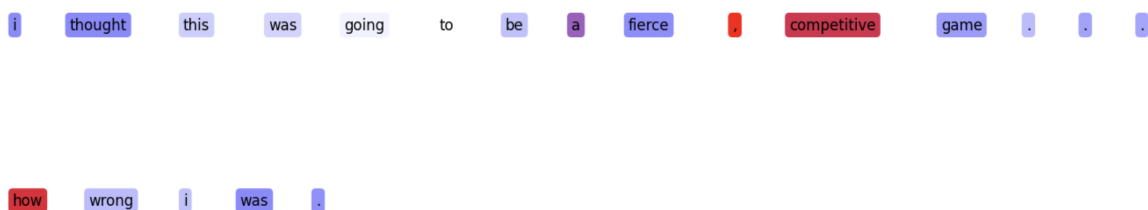


Figure 15: Negative Emotion Saliency map which showcases adjectives being a high indication of emotion.