Leveraging Large Language Models as Faithful Explainer for Text Classification

Le Xuan Tung [1, 2*] and Nguyen Cao Quoc [1, 2†]

[1]University of Information Technology, Ho Chi Minh city, Vietnam.
[2]Vietnam National University, Ho Chi Minh city, Vietnam.

*Corresponding author(s). E-mail(s): tunglx.19@grad.uit.edu.vn;
Contributing authors: quocnc.19@grad.uit.edu.vn;
†These authors contributed equally to this work.

Abstract

Text classification is a pivotal task in Natural Language Processing (NLP), underpinning a wide range of applications such as sentiment analysis, spam detection, and topic categorization. While deep learning based models have significantly advanced the state-of-the-art in this domain, their complex decision-making processes raise concerns regarding explainability and trust, particularly in high-stakes applications. This study investigates the use of Large Language Models (LLMs) as a post-hoc explainer for text classification tasks. We propose a novel, prompt-based framework that harnesses the semantic reasoning and generative capabilities of pretrained LLMs to identify influential input features and produce coherent, human-interpretable explanations. Furthermore, we conduct a comprehensive faithfulness evaluation, comparing our method to traditional explanation techniques. The results highlight the potential of LLMs not only as effective classification tools but also as explainable agents, offering a promising direction for explainability in deep learning-based NLP systems.

Keywords: Text Classification, Explainable AI, Large Language Model

1 Introduction

Text classification is a fundamental task in Natural Language Processing (NLP), where the goal is to assign predefined labels to textual inputs. It plays a critical role in various

real-world applications, including sentiment analysis, spam detection, topic categorization, and content moderation. With the advancement of deep learning, especially the emergence of deep learning based [1][2] architectures the performance of text classification systems has improved significantly. However, despite their success in accuracy and generalizability, these models often operate as black boxes, providing limited transparency into their internal decision-making processes. This lack of explainability raises challenges in areas where trust, accountability, or regulatory compliance are critical, such as healthcare, finance, and legal systems.

To address these issues, Explainable Artificial Intelligence (XAI) has attracted increasing attention[3][4] as a means of providing human-understandable justifications for model predictions. In the context of text classification, XAI methods aim to identify and highlight the most influential words or phrases that contributed to a given classification decision. These explanations can help users better understand the model's behavior, identify potential biases, and guide debugging or fine-tuning of the model. Many techniques such as simplification[5], gradient-based explanations[6], perturbation-based explanations[7][8] produce feature attributions or saliency scores over input words.

In this context, Large Language Models (LLMs) have emerged as a promising approach for post-hoc explanation in text classification. With their strong capabilities in semantic understanding and natural language generation, LLMs can serve as intelligent explanation engines that provide human-interpretable insights. Specifically, by leveraging appropriately designed prompts, LLMs can be guided to identify important words that influence classification results while generating coherent, user-friendly explanations in natural language. This capability allows LLM-based methods to overcome some of the limitations of traditional explanation methods, such as instability or lack of context understanding, thus opening a new research direction for explainability in deep learning models for natural language processing.

In this study, we aim to leverage the reasoning capabilities of Large Language Models (LLMs) and present an experimental study into their use as explainers for text classification tasks. Specifically, we explore how pretrained LLMs can be prompted to identify key words that contribute significantly to a model's classification decision and to generate user-friendly natural language explanations. This approach seeks to assess the feasibility of using LLMs not only as language understanding tools but also as interpretable agents that enhance the transparency of black-box text classifiers.

Our main contributions are summarized as follows:

1. **Prompt-Based Model-Agnostic Post-Hoc Explainer using LLMs**: We propose a novel framework that uses pretrained LLMs to generate textual explanations for text classification decisions.
2. **Faithfulness evaluation**: We evaluate the proposed method in terms of faithfulness, comparing it against established explanation approaches.

2 Related Work

A wide range of post-hoc explainability methods have been developed to interpret the decisions made by text classifiers. Model-agnostic techniques such as LIME[5] and

SHAP[7] are among the most popular. LIME constructs a local surrogate model to approximate the classifier's behavior near a given input, thereby identifying influential words. SHAP assigns Shapley values to each token based on a cooperative game theory formulation. These methods offer flexibility across different models and provide interpretable word-level importance scores. However, they often suffer from instability due to their reliance on random perturbations or local approximations, and they may yield inconsistent or conflicting explanations across runs. In contrast, gradient-based methods[6] (e.g., integrated Gradients, saliency maps) utilize the model's internal gradients to assess token importance. While these methods are typically more faithful to the model's behavior, they can be sensitive to noise and are less explainable for end-users. These approaches, though useful, mainly provide token-level saliency maps without capturing higher-level semantic reasoning, and they often lack the ability to convey explanations in natural language that are accessible to humans.

The advent of Large Language Models (LLMs), such as GPT or T5, has opened up many new directions in the fields of NLP. Recent research has demonstrated the potential of LLMs to serve as a text classifier[9] by leveraging their strong semantic understanding and generative capabilities. Or in the field of XAI, studies have shown that LLMs can act as "judges" or "reviewers" to assess or explain decisions made by other models[10].

Among these LLMs, OpenAI's **Generative Pretrained Transformers (GPT)** constitute a prominent family of large-scale autoregressive models based on the transformer[1] architecture and trained on vast corpora of text data using unsupervised learning objectives. These models are designed to predict the next token in a sequence, enabling them to generate coherent, contextually relevant language across a wide range of tasks. Among the most notable versions, *GPT-3.5* contains approximately 175 billion parameters and is fine-tuned to follow instructions more effectively than its predecessor, achieving strong zero-shot and few-shot performance across diverse NLP benchmarks. More recently, *GPT-4o* contains around 1.8 trillion parameters, extends the capabilities of the GPT-4 family by incorporating native support for text, image, and audio inputs. While only the text modality is used in our work, *GPT-4o* demonstrates GPT-4-level reasoning performance at significantly lower latency and cost, making it a compelling choice for scalable, real-time explanation tasks.

Despite the growing interest in LLM-based applications, systematic exploration of using LLMs as faithful, post-hoc explainers for text classification tasks remains limited. There is no research has focused on explaining the decision process of black box models. In this work, we propose a prompt-based explanation framework in which large language models (LLMs) are used to extract important input tokens and generate natural language rationales for the predictions of text classifiers. We compare the explanations generated by LLMs with those produced by traditional methods such as LIME and SHAP, evaluating them in terms of both faithfulness and human interpretability. Our approach aims to bridge the gap between faithful representations of model decision processes and user-friendly explanations, thereby offering a novel contribution to post-hoc explainability in natural language processing.

3 Methodology

3.1 Task Formulation

Let $\mathcal{D} = \{(x^1, y^1), ..., (x^n, y^n)\}$ be a training dataset where each instance consists of a text input $x^i \in \mathcal{X}$ and its corresponding label $y^i \in \mathcal{Y}$. Here, $\mathcal{X}$ denotes the space of textual documents (e.g., sentences, paragraphs, or entire articles), and a set of label $\mathcal{Y} = c_1, c_2, ..., c_k$ corresponding to predefined categories. The goal of text classification is to learn a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, such that, given an unseen input text $x^i \in \mathcal{X}$, the classifier $f(x)$ predicts the most appropriate label $\hat{y} \in \mathcal{Y}$ corresponding to the semantic content of x.

To explain the classification model, the main goal is to find saliency score map ω for the tokens of an instance x, with each $\omega_i \in \{0, 1\}$. Saliency score is used to evaluate the importance of each token in a sentence or text. The larger the saliency score of a word or token, the more important it is to the semantics of the text, which also affects the classification result of the model for that text. From the saliency score map ω we can derive a list of the most important words $\mathcal{I}$ by arranging them in order of importance. In this study we consider a token equivalent to a word. Our explanation method is *post-hoc*, meaning it does not require access to the model's internal parameters or training process. Furthermore, it is *model-agnostic*, and thus can be applied to any classifier regardless of its underlying architecture or learning mechanism.

3.2 Prompt Engineering

In contexts where direct saliency maps, as produced by traditional XAI methods, are not accessible, we leverage the natural language interpretation capabilities of large language models (LLMs) to explain text classification decisions. Instead of requiring the model to output a detailed saliency heatmap, we design a prompt, presented in table 1, that instructs the LLM to return a list of the most important words (tokens) influencing the classification outcome. This strategy exploits the LLM's internal knowledge to clarify the reasoning behind predictions without direct access to model parameters or gradients. Recent studies have shown that LLMs can effectively serve as post-hoc explainers; for example, [11] demonstrate that prompt-based attribution mechanisms can accurately extract salient input features aligned with model decisions.

In designing the prompt, we apply several prompt engineering techniques: the **Persona pattern** (adopting an expert role), **Chain-of-Thought prompting** (requiring step-by-step reasoning), and structured output formatting (akin to an output automator). For instance, the Persona pattern instructs the model to behave as an XAI specialist, helping it to prioritize contextually relevant features [12]. Chain-of-Thought prompting, as analyzed in [13], guides the model to generate intermediate reasoning steps, from which a logical top-k list of influential tokens can be extracted.

Moreover, by explicitly specifying the output format, such as in JSON or a ranked list, we ensure that the LLM generates consistent and parseable results. Prior experiments also show that when LLMs like GPT-style models are instructed to verbalize saliency attributions, they produce coherent and human-interpretable rationales. This

approach complements traditional attribution methods by converting model rationale into natural language, improving transparency and accessibility.

Our prompt construction borrows from methods that extract top-k important tokens by ranking vocabulary terms based on their learned attribution weights, as explored in recent works that combine prompt-based and fine-tuning techniques for text classification. The integration of explanation style instructions, output constraints, and reasoning steps into a unified prompt makes LLMs a practical tool for post-hoc interpretability in NLP classification tasks.

Table 1 Prompt template for model-agnostic post-hoc LLM explanations

`Task Description`	You are an explainer of the behavior of a text classification model. Your task is to sort the words in the given text in order of decreasing importance and explain the model's prediction result. A word's importance is determined by how much its removal affects the model's classification. You will be given the text, the model's classification result, and the label set.
`Output Formatter`	Your output must strictly be in the following JSON format and contain nothing else: {{`"list": [], "explanation": ""`}} Where 'list' is a list of the 10 most important words in the text sorted from most to least important, and "explanation" is a concise explanation of why the model produced that classification. Only return the JSON result. Do not include any other explanation or text.
`Classification Information`	Here is the input: Text: {*sample*} Classify result: {*predicted_result*} Label set: {*label_set*}
`Chain of Thought`	Instruction: 1. Only use words that strongly influenced the model's output. 2. Justify the influence using a step-by-step reasoning. 3. Do not return duplicate words.

3.3 Faithfulness

To evaluate the effectiveness of XAI techniques, we use a faithfulness measure to assess how meaningful explanations influence the model's prediction. Faithfulness can be measured in many ways, in this study we use the idea of [14] and [15] to test the comprehensiveness of explanations. Specifically, given an explanation of instance x as a list of important words $\mathcal{I}$, we construct a contrast example $\tilde{x}$ compared to the original data point by removing one or the k most important words $(\mathcal{I}_1, .., \mathcal{I}_k)$. Let $p[f(x)]_j$ be the initial prediction probability provided by model f for the predicted class j. Then, we consider the probability of the model's prediction for class j after the important words are removed. Intuitively, the model will be less confident in its

prediction after the arguments are removed from x. The formula for faithfulness is calculated as follows:

$$Faithfulness(x) = p[f(x)]_j - p[f(\tilde{x})]_j = p[f(x)]_j - p[f(x \setminus \mathcal{I}_k)]_j \quad (1)$$

3.4 Human Interpretability

With current XAI techniques for Text classification, the returned result is usually a saliency score map showing the importance of tokens in the text. This is not enough for users to better understand the inference process of the classification model. In the proposed method, we designed a prompt to produce results including a *list* of important words and an *explanation* of why the model predicted that way, illustrated in the table 2. Based on the explanation information, users can partly understand the decision-making process of the model. In the future we plan to conduct a human-centered study to evaluate the interpretability and usefulness of the explanations generated by LLM in a real decision-making context.

Table 2 Example of results of a sample

`List of important words`	['runner', 'base', 'ball', 'tag', 'scored', 'run', 'infield', 'fly', 'batter', 'runners']
`Explanation`	The model classified the text as 'rec.sport.baseball' because it contains specific terminology and concepts related to baseball. Words like 'runner', 'base', 'ball', 'tag', and 'scored' are directly associated with baseball gameplay. The mention of 'infield fly', 'batter', and 'runners' further reinforces the context of baseball, as these are terms commonly used in the sport. The presence of these words strongly influenced the model to categorize the text under the 'rec.sport.baseball' label.

4 Experiment

4.1 Benchmark

In this study, we utilize the 20 Newsgroups and IMDB dataset as benchmark datasets. We utilize the 20 Newsgroups [16] from the *scikit-learn* library, a widely-used benchmark corpus for text classification and topic modeling. The dataset comprises approximately 18,000 documents categorized into 20 distinct classes, each corresponding to a specific newsgroup topic from the Usenet discussion forums. We use a version that removes titles, signatures, and citations to reduce noise during training. As part of the preprocessing process, the following steps are applied: all text is converted to lowercase; special characters are removed, leaving only letters, digits, periods, commas, and spaces; and excess spaces are normalized. We also utilize The IMDB dataset. The IMDB Large Movie Review dataset[17] is a benchmark corpus widely used for

Table 3 Faithfulness evaluation scores across different values of k (number of top salient words removed) on 20 Newsgroups dataset

Method	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 9$
LIME	**0.2030**	**0.3328**	**0.3747**	0.4013	0.4228
SHAP	0.1638	0.3018	0.3690	**0.4202**	**0.4520**
GPT-3.5	0.1366	0.2334	0.2944	0.3391	0.3658
GPT-4o	0.1555	0.2757	0.3543	0.4131	0.4511

binary sentiment classification. It consists of 50,000 movie reviews labeled as either positive or negative, with an even split of 25,000 reviews for training and 25,000 for testing. Each review is presented in plain text format and represents the opinion of a user about a particular movie.

We use the BERT[2] (Bidirectional Encoder Representations from Transformers) model for training, which is considered as black-box model and is also the object that needs to be explained in this study. Training is performed for 20 epochs with 20 Newsgroups data and 2 epochs with IMDB data after which the best models on the two validation sets are saved. The classification result of the model on the 20 Newsgroups dataset is 75.08% and on the IMDB dataset is 92.89%.

LLM-based Explanation (Prompt-based): We utilize **GPT-3.5** and **GPT-4o** to explain the black-box classifier's predictions. Given an input document and the classifier's output label, we construct prompts to elicit natural language rationales from the LLMs. These prompts are designed to guide the LLMs to simulate the black-box model's decision process in human-readable language. To benchmark the quality and faithfulness of these explanations, we conduct a comparative analysis with existing explainability methods, specifically **LIME** and **SHAP**. This setup allows us to explore the potential of modern LLMs as general-purpose explainers for NLP classifiers.

4.2 Experiment result

Tables 3 and 4 present the faithfulness scores of four explanation methods **LIME**, **SHAP**, **GPT-3.5**, and **GPT-4o** which measured by the impact on model output after removing the top-k most salient words. The higher the score, the more important the k-words that the methods give are and the more the model's prediction probability on a predicted result class decreases.

On the 20 Newsgroups dataset, all evaluated methods exhibit an increasing trend in faithfulness scores as the value of k grows. This observation confirms that removing a larger number of salient tokens results in more significant changes to the model's prediction confidence. LIME performs best when k is small (k = 1, 3, 5) and SHAP performs best when k is large (k = 7, 9). GPT-4o has a fidelity score that approaches SHAP and outperforms LIME when k is large (k = 7, 9) like SHAP. GPT-3.5 is the model with the lowest result.

Similarly to the results observed on the 20 Newsgroups dataset, experiments conducted on the IMDB dataset also demonstrate a positive correlation between

Table 4 Faithfulness evaluation scores across different values of k (number of top salient words removed) on IMDB dataset

Method	$k = 1$	$k = 3$	$k = 5$	$k = 7$	$k = 9$
LIME	0.0510	0.0990	0.1322	0.1617	0.1839
SHAP	**0.0532**	0.1024	0.1455	0.1695	0.1860
GPT-3.5	0.0366	0.0794	0.1013	0.1167	0.1294
GPT-4o	0.0511	**0.1039**	**0.1480**	**0.1823**	**0.2091**

faithfulness scores and the number of removed salient tokens (k). Notably, GPT-4o consistently outperforms traditional explanation methods across most k settings (except when $k = 1$) indicating its stronger ability to identify influential tokens. GPT-3.5 remains the least effective among all evaluated methods. Among traditional approaches, SHAP achieves higher faithfulness scores than LIME across all settings.

Overall, the experimental results on both the 20 Newsgroups and IMDB datasets reveal a consistent pattern: faithfulness scores increase with the number of removed salient tokens, reaffirming the reliability of top-k token attribution as a measure of explanation quality. Among traditional XAI methods, SHAP generally outperforms LIME, especially at larger k values. While GPT-3.5 consistently performs worse than all baseline methods, GPT-4o demonstrates strong explanatory capabilities, often rivaling or even surpassing SHAP, especially in higher-k settings. These findings suggest that more advanced LLMs, such as GPT-4o, offer competitive and potentially superior alternatives for post-hoc explanation in text classification tasks, provided that the explanation is carefully elicited through well-designed prompts.

5 Conclusion

This work explores the use of Large Language Models (LLMs) as model-agnostic post-hoc explainers for text classification tasks, leveraging prompt-based techniques to generate both explainable rationales and importance-based word rankings. Our results on the 20 Newsgroups dataset show that LLMs, particularly GPT-4o, can produce explanations that are competitive with established methods such as LIME and SHAP, especially at higher values of k. Notably, on the IMDB dataset, the LLM-based approach even surpasses traditional methods in terms of explanatory fidelity across most settings. The incorporation of structured prompting strategies, including persona-based cues and chain-of-thought reasoning, contributes to generating coherent and human-aligned outputs. Despite these promising results, the approach has key limitations. The explanations reflect the LLM's inferred reasoning rather than a faithful representation of the black-box classifier's internal mechanisms. As such, they may miss subtle decision-making cues embedded in the model's learned parameters. In the future, we aim to more directly link explanations to the classifier's internal computations, potentially through gradient-based or hybrid techniques. Additionally, integrating human-centered evaluations would further validate the practical interpretability of LLM-generated explanations. Overall, this study demonstrates that

LLMs offer a flexible and effective foundation for explainability in NLP, marking a step toward more transparent and user-aligned AI systems.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 5998–6008 (2017)

[2] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) 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. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423 . https://aclanthology.org/N19-1423/

[3] Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I.: A diagnostic study of explainability techniques for text classification. In: Webber, B., Cohn, T., He, Y., Liu, Y. (eds.) Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 3256–3274. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.263

[4] Cesarini, M., Malandri, L., Pallucchini, F., Seveso, A., Xing, F.: Explainable AI for text classification: Lessons from a comprehensive evaluation of post hoc methods. Cogn. Comput. **16**(6), 3077–3095 (2024) https://doi.org/10.1007/S12559-024-10325-W

[5] Ribeiro, M.T., Singh, S., Guestrin, C.: Model-agnostic interpretability of machine learning. CoRR **abs/1606.05386** (2016) 1606.05386

[6] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR, Sydney, NSW, Australia (2017)

[7] Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pp. 4765–4774 (2017)

[8] Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014, pp. 818–833. Springer, Cham (2014)

[9] Sun, X., Li, X., Li, J., Wu, F., Guo, S., Zhang, T., Wang, G.: Text classification via large language models. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 8990–9005. Association for Computational Linguistics, Singapore (2023)

[10] Wiegreffe, S., Hessel, J., Swayamdipta, S., Riedl, M., Choi, Y.: Reframing human-AI collaboration for generating free-text explanations. In: Carpuat, M., Marneffe, M.-C., Meza Ruiz, I.V. (eds.) Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 632–658. Association for Computational Linguistics, Seattle, United States (2022)

[11] Feng, Z., Zhou, H., Zhu, Z., Mao, K.: Promptexplainer: Explaining language models through prompt-based learning. In: Findings of the Association for Computational Linguistics: EACL 2024 (2024)

[12] Hattab, G., Anžel, A., Dubey, A., Ezekannagha, C., Yang, Z., Ilgen, B.: Persona adaptable strategies make large language models tractable. In: Proc. of NLPIR 2024 (2024)

[13] Yeo, W.J., Satapathy, R., Goh, R.S.M., Cambria, E.: How interpretable are reasoning explanations from prompting large language models? In: Findings of NAACL 2024 (2024)

[14] Yu, M., Chang, S., Zhang, Y., Jaakkola, T.: Rethinking cooperative rationalization: Introspective extraction and complement control. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4094–4103. Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1420

[15] DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized NLP models. In: Jurafsky, D., Chai, J., Schluter, N., Tetreault, J. (eds.) Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.408

[16] Mitchell, T.: Twenty Newsgroups. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5C323 (1997)

[17] Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: 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. Association for Computational Linguistics, Portland, Oregon, USA (2011). http://www.aclweb.org/anthology/P11-1015