Probing Robustness of LLM-Generated AI Explanations with Zero-Shot and Few-Shot Prompting

Anonymous Author(s)
Affiliation
Address
email

Abstract

The emergence of Large Language Models (LLMs) in the Foundation Model Landscape of Artificial Intelligence (AI) has led to its increasing usage of Generative AI applications with the aim of achieving human comprehension quality. Explainable Artificial Intelligence (XAI) gives insight on what factors impacted the AI models' outcomes by attributing them to the features in the training data. A large number of XAI models like SHAP, LIME, gradient with respect to the input, and more motivate an important question: Can a LLM robustly generate its preferred XAI model outputs among multiple choices that can appropriately interpret an AI model satisfying user preferences? We answer this question with zero-shot and few-shot prompt engineering on LLMs to automatically generate which XAI model, SHAP or LIME, is preferred for the AI model, random forest classifier on two separate training datasets, evaluating the approach with human annotations. We probe the influence of our prompt tokens on the LLM-Generated Explanation using XAI and remove the words in our prompt most important to generate the explanation in order to investigate the robustness of zero-shot and few-shot prompting for LLM-Generated Explanations.

1 Introduction

In recent years, Large Language Models (LLMs) have become prominent in the field of foundation models for artificial intelligence, revolutionizing natural language understanding and generation tasks, mostly for high-resource languages. These models have displayed capabilities of human-like outputs in myriad applications, including text generation, machine translation, and interactive Generative AI models like ChatGPT. However, as LLMs continue to be incorporated in more cognitive applications with good quality generations in high-resource languages, their inherent opacity has raised questions regarding their ability to explain and reason about what they output. Some previous work has shown that LLMs perform poorly on reasoning tasks [1], while other work has shown by using methods like Chain-of-Thought, which use few-shot prompting methods, LLMs can effectively reason to answer questions [2].

This paper evaluates the performance of two different approaches, zero-shot and few-shot prompting, to determine if LLMs are capable of reasoning about another task: choosing between XAI outputs to explain the output of an AI model. We have a chain of instructions to give context about XAI model outputs, the AI model, the training data, predicted class label and the ground truth class label. Then we ask the LLM to generate its preferred XAI model output. We evaluate the performance of the LLM on this task by comparing its labels with human annotations preferring one XAI model output over another. We investigate which set of words are important to determinate robust LLM generation by using Gradient based XAI [3] method. Removing the top four tokens from the prompt with highest feature scores impact the output preference and observing changes in generated labels helps us to get insights on how effectively LLMs can reason about XAI model outputs.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

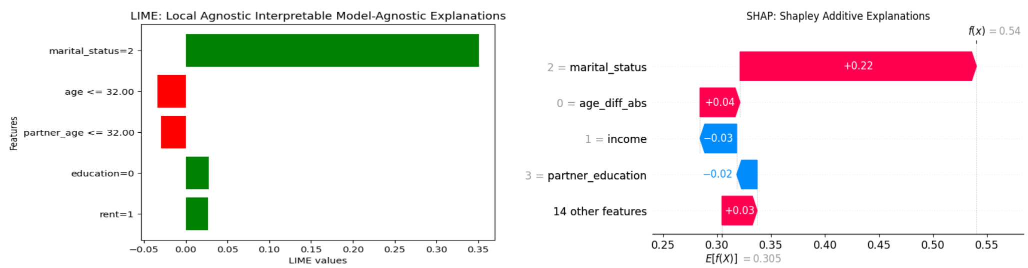


Figure 1: (a) LIME explanation and (b) SHAP explanation for a test sample in the Couples dataset

2 XAI Selection

Our goal is to have a LLM generate its preference for one of the two XAI models, LIME or SHAP to explain Random Forest Classification. SHAP (Shapley Additive Explanations), calculates a score for each feature used by an AI model. A positive score indicates a feature contributed more heavily to a model output while a negative score indicates a feature did not contribute as much to a model output. It calculates these weights by considering all collations of features in the model that can be provided to the model and then measuring the change in model output. [4]. LIME (Locally Interpretable Model-Agnostic Explanations) takes a different approach to calculate these scores by applying local perturbations to an input to a model and seeing how a model's predictions change. By measuring the way the predictions of the model change in relation to the perturbations, LIME determines which features contribute positively or negatively to an outcome. These XAI models change their outputs depending on which AI model and dataset is used as per user requirement [5].

We generate explanations using the XAI models, LIME and SHAP, on test samples for random forest classifiers trained on two datasets. The HCMST (Couples) dataset [6] documents whether couples stay together or not and various attributes about the couple such as age, education, and income. The Diabetes dataset [7] documents whether patients have diabetes or not and other health information like insulin levels, glucose levels, and blood pressure. Figure 1 represents the LIME and SHAP XAI plots for a test sample explaining the goal of random forest classification on Couples dataset.

Dataset	Model	# Features	# Train	# Test
Diabetes[7]	Random Forest Classifier	18	537	231
Couples[6]	Random Forest Classifier	8	1030	442

Table 1: Statistics of Two Training & Testing Datasets for Random Forest Classification

3 LLM Prompts

Flan-T5 is a class of Large Language Model (LLM) that is instruction fine-tuned to increase performance. Flan-T5 has been shown to improve performance with instruction fine-tuning over many different model sizes from 80M parameters to 11B and can even outperform models that have not been fine-tuned of much larger size [8]. We prompt the Flan-T5-XL checkpoint to generate the LLM labels for XAI outputs.

3.1 Zero shot

Zero shot prompting is a method to prompt an LLM by providing it with instructions describing the task without context [9]. In this setting, we define zero shot prompting as asking the LLM to choose between SHAP and LIME without providing it context of human annotations for a similar task.

The prompt is structured so that the LLM is given context on the dataset and model being used. In the case of Figure 2, it is the Couples dataset and the random forest classifier. This includes a description of the type of model being used as well as the number of training and test samples used for training and evaluating the model. After that, the features and feature values for the test sample that the XAI models were run on is given. Next, the prediction by the random forest classifier as well as the ground truth label is added to the prompt. Finally, the LIME and SHAP scores and their corresponding features and the question being asked of the LLM: to choose between the provided explanations.

```
Pick an XAI model, Shapley Additive Explanation (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME), that does the best
job of explaining a prediction by an AI model.
The AI model is a random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various
sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.
The training data is on couples and the classes are whether the couple stayed together, denoted by 1, or the couple broke up, denoted
by 0.
There are 1472 samples, 1030 of which have been used for training and 442 for testing the model.
The feature values for the test sample are as follows:
age=29
partner_age=29
age_diff_abs=0
children=0.0
visits_relatives=4
education=bachelor's degree or higher
marital_status=married
partner_education=some college
gender=female
house=a one-family house detached from any other house
income=$100,000 to $124,999
msa=metro
rent=owned or being bought by you or someone in your household
political=democrat
religion=catholic
work=working - self-employed
gender_older=0
education_difference=1
The prediction for the test sample using Random Forest Classifier is 1.0
The ground truth label for the test sample is 1.0
LIME, which means Local Interpretable Model-Agnostic Explanations, has scores for the features marital_status, age, house,
partner_age, partner_education, as 0.3261, -0.0547, 0.051, -0.0462, -0.0294.
SHAP, which means Shapley Additive Explanation, has scores for the features marital_status, income, partner_education, age,
partner_age, as 0.153, -0.0478, -0.0436, -0.0249, -0.0245.
Keep in mind that LIME and SHAP are very comparable models.
Using only the information just provided, please make the following choice:
1.) The Local Interpretable Model-Agnostic Explanations XAI model's output is preferred
2.) The Shapley Additive Explanations XAI model's output is preferred
```

Figure 2: An example of a zero-shot prompt for Flan-T5-XL

3.2 Few shot

Few shot prompting can be contrasted with zero-shot prompting in that some context is provided to the LLM of the task it must do in addition to instructions to complete the task [9]. We provide this context with one example each of both SHAP and LIME human annotations along with the SHAP and LIME feature scores that were given to the human annotator, as seen in. Additional specificity is also provided in the query to the LLM by asking it to choose the XAI model whose outputs are most similar to the ones provided in the few shot context, as seen in Figure 3.

```
Pick an XAI model, Shapley Additive Explanation (SHAP) or Local Interpretable Model-Agnostic Explanations (LIME), that does the best job of explaining a prediction by an AI model. The AI model is a
random forest classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy
and control over-fitting.
The training data is on couples and the classes are whether the couple stayed together, denoted by 1, or the couple broke up, denoted by 0.
There are 1472 samples, 1030 of which have been used for training and 442 for testing the model.
The features of this dataset are as follows: age, partner_age, age_diff_abs, children, visits_relatives, education, marital_status, partner_education, education, marital_status, partner_education, gender,
house, income, msa, rent, political, religion, work, gender_older, and education_difference.
To give some context for this task, this is an example of a choice made by a human on a different test sample:
When Local Interpretable Model-Agnostic Explanations had scores for the features marital_status, partner_age, education_difference, age, partner_education, as 0.3477, 0.0552, 0.033, 0.0313, -0.026
and Shapley Additive Explanations had scores for the features marital_status, age, partner_age, house, work, as 0.1311, 0.0447, 0.0291, 0.0188, 0.0169, the human chose Local Interpretable Model-
Agnostic Explanations (LIME) as having a preferable output
Another example of a choice made by a human on a different test sample:
When Local Interpretable Model-Agnostic Explanations has scores for the features marital_status, partner_education, msa, house, visits_relatives, as -0.3452, 0.0408, -0.0203, 0.0158, -0.0122 and
Shapley Additive Explanations had scores for the features marital_status, partner_education, education_difference, gender_older, partner_age, as -0.3333, 0.0248, -0.0217, 0.0192, 0.0168, the human
chose Shapley Additive Explanations (SHAP) as having a preferable output.
The feature values for the test sample you will be asked to classify are as follows:
age=29
partner_age=29
age_diff_abs=0
children=0.0
visits_relatives=4
education=bachelor's degree or higher
marital_status=married
partner_education=some college
gender=female
house=a one-family house detached from any other house
income=$100,000 to $124,999
msa=metro
rent=owned or being bought by you or someone in your household
political=democrat
religion=catholic
work=working - self-employed
gender_older=0
education_difference=1
The prediction for the test sample you will be asked to classify using Random Forest Classifier is 1.0
The ground truth label for the test sample you will be asked to classify is 1.0
LIME, which means Local Interpretable Model-Agnostic Explanations, has scores for the features marital_status, age, house, partner_age, partner_education, are, respectively, 0.3261, -0.0547, 0.051,
-0.0462, -0.0294.
SHAP, which means Shapley Additive Explanation, has scores for the features marital_status, income, partner_education, age, partner_age, as 0.153, -0.0478, -0.0436, -0.0249, -0.0245.
Choose the option that has scores most similar to the examples you've already seen.
1.) The SHAP XAI model's output is preferred
2.) The LIME XAI model's output is preferred
```

Figure 3: An example of a few-shot prompt for Flan-T5-XL

4 Results

We evaluate performance on the labelling task by comparing LLM labels with that of a human annotator. We randomly sampled 200 test samples, 100 from the Couples dataset and 100 from the Diabetes dataset, based on which we construct the prompts for the zero shot and few shot setting. A human annotator was provided the same prompts given to LLM and asked to choose between SHAP and LIME. Their annotations were validated by another human annotator.

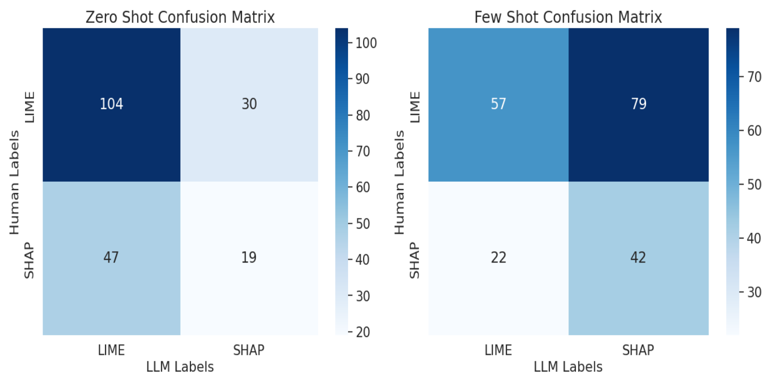


Figure 4: Confusion Matrices for the Couples dataset comparing zero shot and few shot prompting

Dataset	Number of Samples	Prompting Strategy	Accuracy
Diabetes	200	Zero Shot	54%
Couples	200	Zero Shot	61.5%
Diabetes	200	Few Shot	44.5%
Couples	200	Few Shot	49.5%

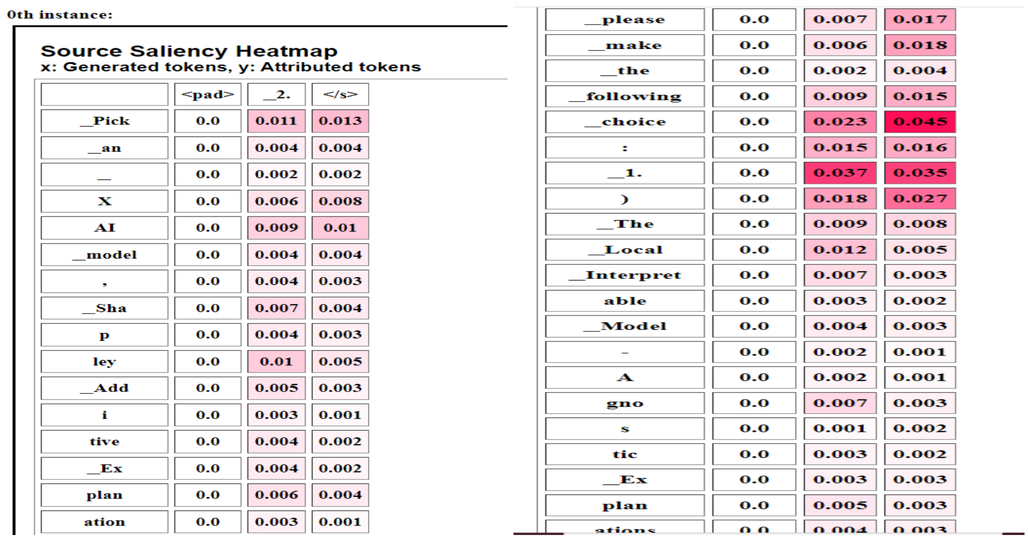
Table 2: Accuracy of LLM labelling with different prompting strategies

4.1 Analysis

A decrease in performance is observed for LLM-Generated automatic explanation labelling with the few-shot prompting strategy, for both the Couples and Diabetes datasets. The performance on the Couples dataset drops 12% while the performance on the Diabetes dataset drops 9.5%, as seen in Table 2 and Figure 4. This shows the increased context length does not necessarily help the LLM gain understanding of how a task should be performed. Due to the $O(n^2)$ computational complexity of self-attention [10] used in transformers, LLMs can get exponentially worse at attending to larger contexts. This can provide an explanation for why adding more context hurts performance more than it helps it.

5 Robust Interpretation of LLM-Generated Explanations

To determine the important words in zero-shot and few-shot prompting that decide the robust generation by LLMs, we use a XAI model, computing gradients of outputs with input features [3] method to assign scores for each of the tokens generated by the model's tokenizer, as seen in Figure 5. The higher these scores are, the more the model rely on them to generate its output. We then test robustness by removing the top four highest scored tokens. We observed that out of the ten randomly sampled prompts, the LLM's generated XAI model output preference flipped for seven of the prompts with respect to their previous generation. This gives us insight on the set of important words that ensure the robustness of LLM generation.



0th instance:

Source Saliency Heatmap
x: Generated tokens, y: Attributed tokens

	<pad>	_2.	</s>
_Pick	0.0	0.011	0.013
_an	0.0	0.004	0.004
_	0.0	0.002	0.002
X	0.0	0.006	0.008
AI	0.0	0.009	0.01
_model	0.0	0.004	0.004
,	0.0	0.004	0.003
_Sha	0.0	0.007	0.004
p	0.0	0.004	0.003
ley	0.0	0.01	0.005
_Add	0.0	0.005	0.003
i	0.0	0.003	0.001
tive	0.0	0.004	0.002
_Ex	0.0	0.004	0.002
plan	0.0	0.006	0.004
ation	0.0	0.003	0.001
_please	0.0	0.007	0.017
_make	0.0	0.006	0.018
_the	0.0	0.002	0.004
_following	0.0	0.009	0.015
_choice	0.0	0.023	0.045
:	0.0	0.015	0.016
_1.	0.0	0.037	0.035
)	0.0	0.018	0.027
_The	0.0	0.009	0.008
_Local	0.0	0.012	0.005
_Interpret	0.0	0.007	0.003
able	0.0	0.003	0.002
_Model	0.0	0.004	0.003
-	0.0	0.002	0.001
A	0.0	0.002	0.001
gno	0.0	0.007	0.003
s	0.0	0.001	0.002
tic	0.0	0.003	0.002
_Ex	0.0	0.003	0.003
plan	0.0	0.005	0.003
ations	0.0	0.004	0.003

Figure 5: The rows of a table of the most important tokens in the prompt impacting the generation as scored by the Input X Gradient method [3]

6 Conclusion

LLMs like Flan-T5-XL can be used to generate AI explanations automatically for AI models like Random Forest Classification with corresponding training data. Corresponding to manually annotated preferences of SHAP or LIME XAI models, our experiments lead to a 61.5% accuracy and 54% accuracy for the Couples and the Diabetes datasets with zero-shot prompting respectively and 49.5% accuracy and 44.5% accuracy for the above mentioned datasets respectively with few-shot prompting. For our human annotation on randomly sampled data items, zero-shot prompting is closer to human preferences vis-a-vis few-shot prompting. The important words in the prompt impacting the generation of explanation models are identified using an XAI model computing gradients of output scores with respect to input prompt. When these important words are removed, the explanation preference generated by the LLM flips for most of the prompts giving insights into robustness. XAI models give us an understanding of which important words in the prompt control the robustness of LLM generated AI Explanations.

References

[1] Karthik Valmeekam, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati. Large language models still can't plan (a benchmark for llms on planning and reasoning about change). *arXiv preprint arXiv:2206.10498*, 2022.

[2] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[3] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[4] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[5] Ahmed Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E Petersen, Gloria Menegaz, and Karim Lekadir. Commentary on explainable artificial intelligence methods: Shap and lime. *arXiv preprint arXiv:2305.02012*, 2023.

[6] Reuben J. Thomas Michael J. Rosenfeld and Maja Falcon. How Couples Meet and Stay Together (HCMST). `https://data.stanford.edu/hcmst`, 2018. [Accessed 04-10-2023].

[7] Mehmet Akturk. Diabetes dataset. `https://www.kaggle.com/datasets/mathchi/diabetes-data-set`, 2020. [Accessed 04-10-2023].

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[9] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.