Automated Assessment of Fidelity and Interpretability: An Evaluation Framework for Large Language Models' Explanations (Student Abstract)

Mu-Tien Kuo[1, 2], **Chih-Chung Hsueh**[1, 2], **Richard Tzong-Han Tsai**[2, 3]
[1]Chingshin Academy, Taiwan
[2]Research Center for Humanities and Social Sciences, Academia Sinica
[3]Dept. of Computer Science and Engineering, National Central University, Taiwan
{11035018, 11035038}@st.chjhs.tp.edu.tw, thtsai@g.ncu.edu.tw

Abstract

As Large Language Models (LLMs) become more prevalent in various fields, it is crucial to rigorously assess the quality of their explanations. Our research introduces a task-agnostic framework for evaluating free-text rationales, drawing on insights from both linguistics and machine learning. We evaluate two dimensions of explainability: fidelity and interpretability. For fidelity, we propose methods suitable for proprietary LLMs where direct introspection of internal features is unattainable. For interpretability, we use language models instead of human evaluators, addressing concerns about subjectivity and scalability in evaluations. We apply our framework to evaluate GPT-3.5 and the impact of prompts on the quality of its explanations. In conclusion, our framework streamlines the evaluation of explanations from LLMs, promoting the development of safer models.

Introduction

As Large Language Models (LLMs) gain traction, it is of increasing paramount to evaluate their explanations' quality. A well-crafted explanation hinges on two elements: fidelity, which refers to a truthful representation of the model's inner-workings, and interpretability, which pertains to the ease with which humans can comprehend it (Gilpin et al. 2018). Our work focuses on evaluating free-text rationales, which are model-generated natural language explanations and the predominant format of LLMs' explanations. Although evaluation methods for rationales have considerable development, there are sizeable limitations on applying prior methods to SOTA LLMs. Traditional fidelity evaluations often assumed full access to the model, enabling commonly used methods such as extracting rationale salience maps via gradients (Atanasova et al. 2020) and distorting encoded inputs (Wiegreffe, Marasović, and Smith 2021). However, the constraints on contemporary proprietary LLMs, including limited access to model weights and encoded input modifications, have rendered these methods less feasible. Previous methods on evaluating interpretability also raise concerns. The traditional practice of comparing machine-generated explanations to human-written ones is inherently flawed as this measures only narrow aspects of text (Wiegreffe, Marasović, and Smith 2021) and show minimal correlation with human evaluations (Clinciu, Eshghi, and Hastie 2021). The use of human annotators is not ideal, as they can be expensive and may introduce a significant level of subjectivity (Lertvittayakumjorn and Toni 2019; Atanasova et al. 2020).

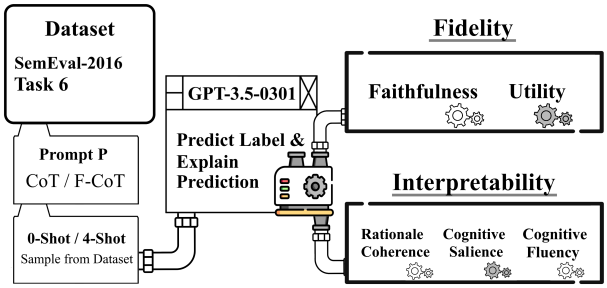

Figure 1: Framework Overview

Our study seeks to address these issues by proposing a task-agnostic automated framework that evaluates both fidelity and interpretability. Our proposed fidelity evaluation only requires perturbations to input text, coping with proprietary LLMs' opaqueness. Similarly, our interpretability evaluation leverages LMs to replace human annotators, reducing subjectivity and cost. In sum, our framework enhances the efficacy and practicality of evaluating LLM explanations, ensuring their reliability in diverse applications.

Metrics

Measuring Fidelity

Fidelity pertains to how accurately an explanation represents a model's actual behavior (Gilpin et al. 2018). Two primary components of fidelity are Faithfulness and Utility. **Faithfulness (F)** is defined as the extent to which an explanation mirrors the internal workings of the model. Notably, a common notion of which is that an interpretation system is unfaithful if it offers varying interpretations for similar input-decision pairs (Jacovi and Goldberg 2020). Given the limitations of accessing closed-source models, we propose a "textual perturbation" approach, modifying the original input text in ways the model should remain robust against, then evaluating the explanations' consistency. Stable explanations across perturbations suggest they accurately reflect the model's pri-

mary decision factors. **Utility (U)** measures an explanation's effectiveness by evaluating its informativeness and brevity. To assess utility, we use the *forward simulatability* metric, evaluating informativeness based on how well an auxiliary model can forecast a model's decision based on its explanation, and brevity by observing the performance drop of the auxiliary model when various amounts of explanation information are removed. Utility is then derived from these calculations, capturing both the depth and succinctness.

Measuring Interpretability

Interpretability gauges how easily a user can understand a model's rationale. Instead of focusing on subjective aspects such as *plausibility*, this work addresses interpretability in the lens of cognitive linguistics. More specifically, we divide interpretability into three main cognitive dimensions following the notions proposed by Ylikoski and Kuorikoski (2010). Our methodology formulates evaluation into components that effectively leverage LMs' strengths, thereby facilitating their substitution for human evaluators.

Rationale Coherence (RC) emphasizes the logical consistency within an explanation, as inconsistencies lead to confusion and doubts towards its credibility. To measure this, explanations are divided into sentences, and an external model (like an LM fine-tuned for Natural Language Inference) is used to spot contradictions between these sentences. **Cognitive Fluency (CF)**, on the other hand, assesses how closely an explanation's logic aligns with common human thought patterns (Unkelbach 2006). We score CF with LLMs due to their proven performance in benchmarks evaluating the understanding of human-like rationale such as DROP (Dua et al. 2019) and ability to forecast over human preference of explanations (Wiegreffe et al. 2022). The main difference between CF and plausibility is that CF aligns with general human cognitive structures while plausibility can be influenced by individual beliefs and experiences. Lastly, as difficult language (such as complex vocabulary or nonstandard sentence structure) hinders interpretability, **Cognitive Salience (CS)** evaluates the readability and complexity of the language in the explanation. To assess CS, we employ transformer-based encoders due to their ability to accurately evaluate text difficulty (Alaparthi et al. 2022).

Experiments & Results

In our experiments, we focus on the task of Stance Detection, which entails determining a sentence's stance (favor/against/none) towards a target. Given GPT-3.5's widespread usage, we assess its explanation quality. Our design includes two prompts that direct the model to predict a tweet's stance towards a target and elucidate its decision: one encourages the model to use Chain-of-Thought (CoT), while the other incorporates a specific instruction requesting faithful explanations which are "true to what I think" (F-CoT). In summary, GPT-3.5's explanations exhibit moderate fidelity and satisfactory interpretability scores. Though CoT prompts yield more accurate explanations, the intermediate steps sometimes include irrelevant evidence to the model's final label. Contrastively, F-CoT produces cohesive arguments at the expense of faithfulness, demonstrating GPT-3.5's misunderstanding of fidelity and the limitations of directly prompting the model for faithfulness.

Prompt	Shots	F	U	RC	CF	CS
CoT	0	0.61	**0.26**	0.79	**0.91**	**0.47**
	4	**0.67**	0.24	0.62	0.89	0.45
F-CoT	0	0.44	**0.26**	0.8	**0.91**	**0.47**
	4	0.5	0.24	**0.81**	0.89	0.46

Table 1: GPT-3.5's Explanation Quality

Conclusion

In this study, we introduced a task-agnostic framework for evaluating the quality of free-text rationales in terms of both fidelity and interpretability. Our methods can evaluate proprietary LLMs that limit user access, and our automated interpretability assessment does not require prior human annotations nor annotator involvement. We apply our framework in evaluating GPT-3.5's explanations, finding fidelity decrements when prompted to produce explanations that are faithful, caused by a misinterpretation of faithfulness.

References

Alaparthi, V. S.; Pawar, A. A.; Suneera, C. M.; and Prakash, J. 2022. Rating Ease of Readability using Transformers. In *ICCAE*.

Atanasova, P.; Simonsen, J. G.; Lioma, C.; and Augenstein, I. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *EMNLP*.

Clinciu, M.-A.; Eshghi, A.; and Hastie, H. 2021. A Study of Automatic Metrics for the Evaluation of Natural Language Explanations. In *EACL*.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *NAACL*.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining explanations: An overview of interpretability of machine learning. In *IEEE DSAA*.

Jacovi, A.; and Goldberg, Y. 2020. Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness? In *ACL*.

Lertvittayakumjorn, P.; and Toni, F. 2019. Human-grounded Evaluations of Explanation Methods for Text Classification. In *EMNLP-IJCNLP*.

Unkelbach, C. 2006. The learned interpretation of cognitive fluency. *Psychological Science*, 17(4): 339–345.

Wiegreffe, S.; Hessel, J.; Swayamdipta, S.; Riedl, M.; and Choi, Y. 2022. Reframing Human-AI Collaboration for Generating Free-Text Explanations. In *NAACL*.

Wiegreffe, S.; Marasović, A.; and Smith, N. A. 2021. Measuring Association Between Labels and Free-Text Rationales. In *EMNLP*.

Ylikoski, P.; and Kuorikoski, J. 2010. Dissecting explanatory power. *Philosophical studies*, 148: 201–219.