Explainable Artificial Intelligence and its relations to Plausibility and Faithfulness

YINUO ZHAO, LUT University, Finland

ACM Reference Format:
Yinuo Zhao. 2025. Explainable Artificial Intelligence and its relations to Plausibility and Faithfulness. 1, 1 (February 2025), 5 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 OBJECTIVE

The objective of this study is to investigate the **plausibility** and **faithfulness** of **Explainable Artificial Intelligence (XAI)**, specifically in two types: **Feature Attribution-based Explanation** and **Generative Assistant Model-based Explanation**. It is likely that this will be linked to recommender systems and chatbots in the future.

Generally, two possible explanations are conducted for a specific model with several datasets and evaluated using two methods, including

(1) **Plausibility and Faithfulness metrics**
(2) **Humans recognition or liking**

Plausibility and Faithfulness metrics can be measured immediately right after the explanation generation process. Then we can know which explanation gets a higher score in the metrics.

Subsequently, we illustrate these explanations or corresponding conclusions to humans, expecting some **feedback or responses**. The feedback or responses should be simple selections or options, that **indicate which one the human users prefer from their perspective**. In this way, humans recognition or liking are collected, and we can know which explanation humans prefer.

We can then see whether the winning explanation in the metrics is consistent with the explanation preferred by humans. However, sometimes human users would prefer the generative assistant model-based explanation due to its powerful deception in natural language, even if it is partially incorrect or completely wrong.

By repeating this experiment on several datasets, we can verify the plausibility of the explainable artificial intelligence to a certain extent.

Author's address: Yinuo Zhao, Yinuo.Zhao@student.lut.fi, LUT University, Lahti, Finland.

Manuscript submitted to ACM

2 CURRENT ISSUES NEED TO BE DISCUSSED

(1) Only one original dataset has the annotated rationales that can be used to measure the overlap between it and the explanations to get some degree of plausibility metrics. Most of the other datasets do not have such annotations available. **Only the original dataset was calculated right now.**
(2) Are these explainers good enough? Should we choose other alternative explanation methods outside of "Feature Attribution-based"? e.g. Attention-based, Example-based or Counterfactual....
(3) The generative explanation way must be determined:
  (a) LLMs will do the prediction and explanation based on the entire text sentence
  (b) **LLMs only do the explanation based on the attributes and prediction already generated by the explainers(e.g. SHAP)**
(4) Human-centered study:
  (a) User Interface to collect responds, see example
  (b) Any scale? For example, Explanation Goodness Checklist + Explanation Satisfaction Scale and so on...
(5) Future works
  (a) Find more similar datasets that have annotated rationales.
  (b) Try out more explainers. However, different explainer will cost more time to adjust my code. Some of the plausibility metrics have been calculated, but there will be more in the future.
  (c) Is it really necessary to let the LLMs, such as GPT-4-mini to do the prediction tasks? Because the prediction tasks has already done by the prediction model. Then the **LLMs will be utilized as a explainer**.
  (d) Design the survey web page that can be used to collect human recognition.

3 THE ITEMS SHOULD BE SENT VIA EMAIL

(1) The model I have used
(2)

4 MODEL

The model selected should be Hate-speech-CNERG/bert-base-uncased-hatexplain, and we believe that this model can **classify a text as Hatespeech, Offensive, or Normal**.

5 DATASETS

Several datasets will be used as inputs to the classification model and then get the results as output.

The current dataset Hate-speech-CNERG/hatexplain is the one used to train the model.

6 EXPLAINERS

- SHAP (Shapley Additive Explanations)
- LIME (Local Interpretable Model-agnostic Explanations)
- DeepLIFT
- Integrated Gradients

7 PLAUSIBILITY

For classification metrics, we used Precision (P), Recall (R), and F1 scores

8 GENERATIVE LLMS

Design a template to prompt GPT or other LLMs, see example.

9 FEATURE ATTRIBUTION-BASED EXPLANATION

1. Hierarchy Tree

Traditional Fine-Tuning Paradigm
—-Local Explanation
———Feature Attribution-Based Explanation

This method will highlight the **specific features or inputs** that **influence the decisions or outputs** of the model.
2. Plausibility evaluating metrics

Typically five dimensions: grammar, semantics, knowledge, reasoning, and computation.

Metrics measuring two **token-level rationales** include

- **Intersection-Over-Union (IOU)**
- **Precision**
- **Recall**

Metrics that measure **overall plausibility** include

- **F1 score** for discrete cases
- **Area under the precision recall curve (AUPRC)** for continuous or soft token selection cases.

Our additional evaluating component: Evaluating plausibility based on **human recognition or liking** of the explanation, i.e. how much the human agree or like the explanation.

10 GENERATIVE ASSISTANT MODEL-BASED EXPLANATION

1. Hierarchy Tree

Prompting Paradigm
—-Assistant Model (such as GPT-4)

2. Plausibility evaluating metrics

Evaluating plausibility of this explanation should be considered as a little bit complex, including

- **whether explanations satisfy human expectations**

- proposes to evaluate the counterfactual simulatability of natural language explanations

Explanations satisfy human expectations ?= Human recognition or liking of the explanation

This explanation may be completely wrong because not attributed to the model.

11 RELATED PAPERS (STARTING POINTS)

Click to access papers.

(1) Benchmarking eXplainable AI - A Survey on Available Toolkits and Open Challenges
(2) A Diagnostic Study of Explainability Techniques for Text Classification
(3) Diagnostics-Guided Explanation Generation
(4) Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods
(5) "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction
(6) The Impact of Imperfect XAI on Human-AI Decision-Making
(7) On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations
(8) Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making
(9) Explainability for Large Language Models: A Survey
(10) Generating Fact Checking Explanations

12 DESIRABLE CONFERENCE

Click to access related links below.

- Preparing Your Article with LaTeX
- ACM Conference on Intelligent User Interfaces (ACM IUI) 2025
- LATEX Class for the Association for Computing Machinery

ACM Conference on Intelligent User Interfaces (ACM IUI) 2025 is the 30th annual premiere venue, where researchers and practitioners meet and discuss state-of-the-art advances at the intersection of Artificial Intelligence (AI) and Human-Computer Interaction (HCI). Ideal IUI submissions should address practical HCI challenges using machine intelligence and discuss both computational and human-centric aspects of such methodologies, techniques and systems.

In addition to traditional IUI themes, we also consider **Large Language Models (LLMS)** for the people. As LLMs become more powerful and more accessible, end users can interact with them in various ways and a wide variety of new applications become feasible. This raises new research questions regarding the interaction between users and generative AI models, including the design of new interactions and intelligent systems, **when to trust systems powered with LLMs, fairness in LLMs**, studying the effect of such models on people's work and aligning user expectations with model capabilities.

REFERENCES

[1] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. Generating Fact Checking Explanations. *arXiv.org* (2020).
[2] Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2021. Diagnostics-Guided Explanation Generation. *arXiv.org* (2021).

[3] Astrid Bertrand, Tiphaine Viard, Rafik Belloum, James R. Eagan, and Winston Maxwell. 2023. On Selective, Mutable and Dialogic XAI: a Review of What Users Say about Different Types of Interactive Explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21.

[4] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–17.

[5] Sara Vera Marjanović, Isabelle Augenstein, and Christina Lioma. 2024. Investigating the Impact of Model Instability on Explanations and Uncertainty. *arXiv.org* (2024).

[6] Katelyn Morrison, Philipp Spitzer, Violet Turri, Michelle Feng, Niklas Kühl, and Adam Perer. 2024. The Impact of Imperfect XAI on Human-AI Decision-Making. *Proceedings of the ACM on human-computer interaction* 8, CSCW1 (2024), 1–39.

[7] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2022. On Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. *arXiv.org* (2022).

[8] Haeun Yu, Pepa Atanasova, and Isabelle Augenstein. 2024. Revealing the Parametric Knowledge of Language Models: A Unified Framework for Attribution Methods. *arXiv.org* (2024).

[9] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. 2024. Explainability for Large Language Models: A Survey. *ACM transactions on intelligent systems and technology* 15, 2 (2024), 1–38.