# EFEMO: The Explanation Faithfulness Evaluation Measures Ontology

Danielle Villa
Advisor: Dr. Deborah McGuinness

TWC Tetherless World Constellation

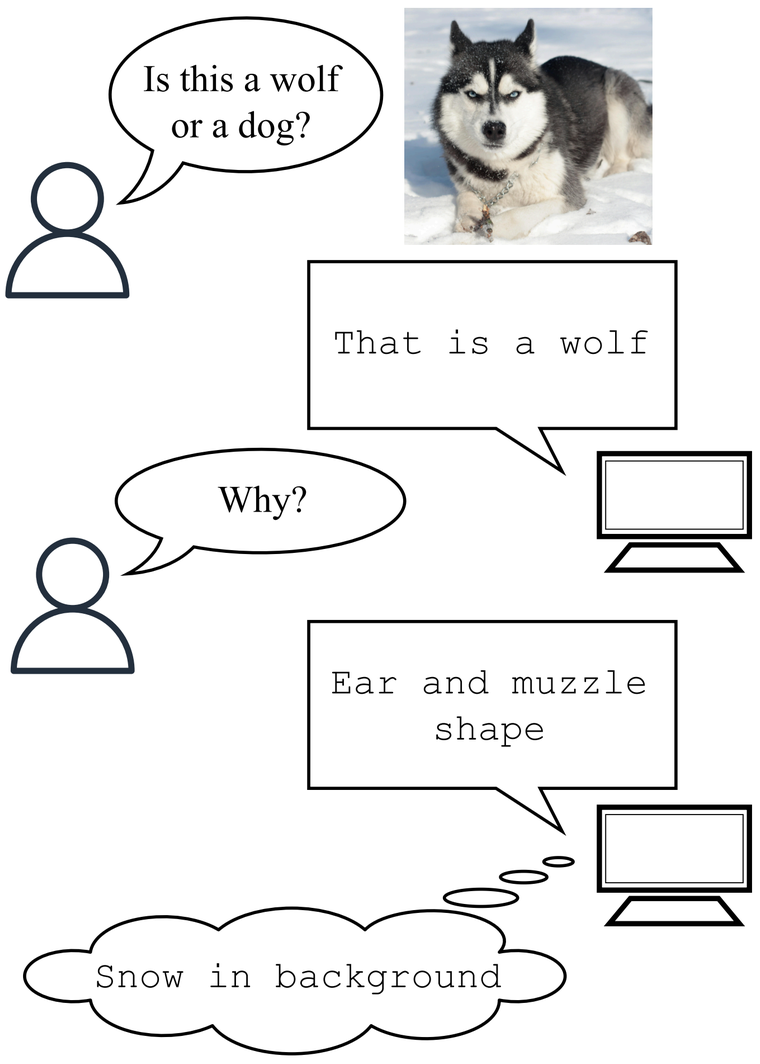
## Motivation

Artificial Intelligence (AI) has become very popular in the past few years with the rise of AI chatbots such as ChatGPT [1]. This has raised questions about when and how to apply this technology to sensitive fields such as finance, healthcare, and military applications.

One popular technique to increase user trust is to provide explanations. These give insight into how a model behaviors and why it made a decision, and are sometimes legally required [2]. Unfortunately, even among papers that focus on explainability, there's no consensus on how to judge those explanations [3]. Experts cannot agree on the best form of an explanation and which properties should be prioritized since it's very context dependent [4], and this had led to a lack of consistent evaluations.

As such, dozens of measures have proposed for evaluating explanations, specifically the faithfulness of explanations, in the past five years alone. It is unclear when a given measure is applicable, how it compares to other measures, and what the measure is actually evaluating.

To help solve this, we built the Explanation Faithfulness Evaluation Measures Ontology (EFEMO). This is a publicly available tool for organizing and recommending measures. It was build to be used by both experts and non-experts to encourage usage of these measures for explanation evaluation. We are currently focusing on measures for text-based models and tasks.



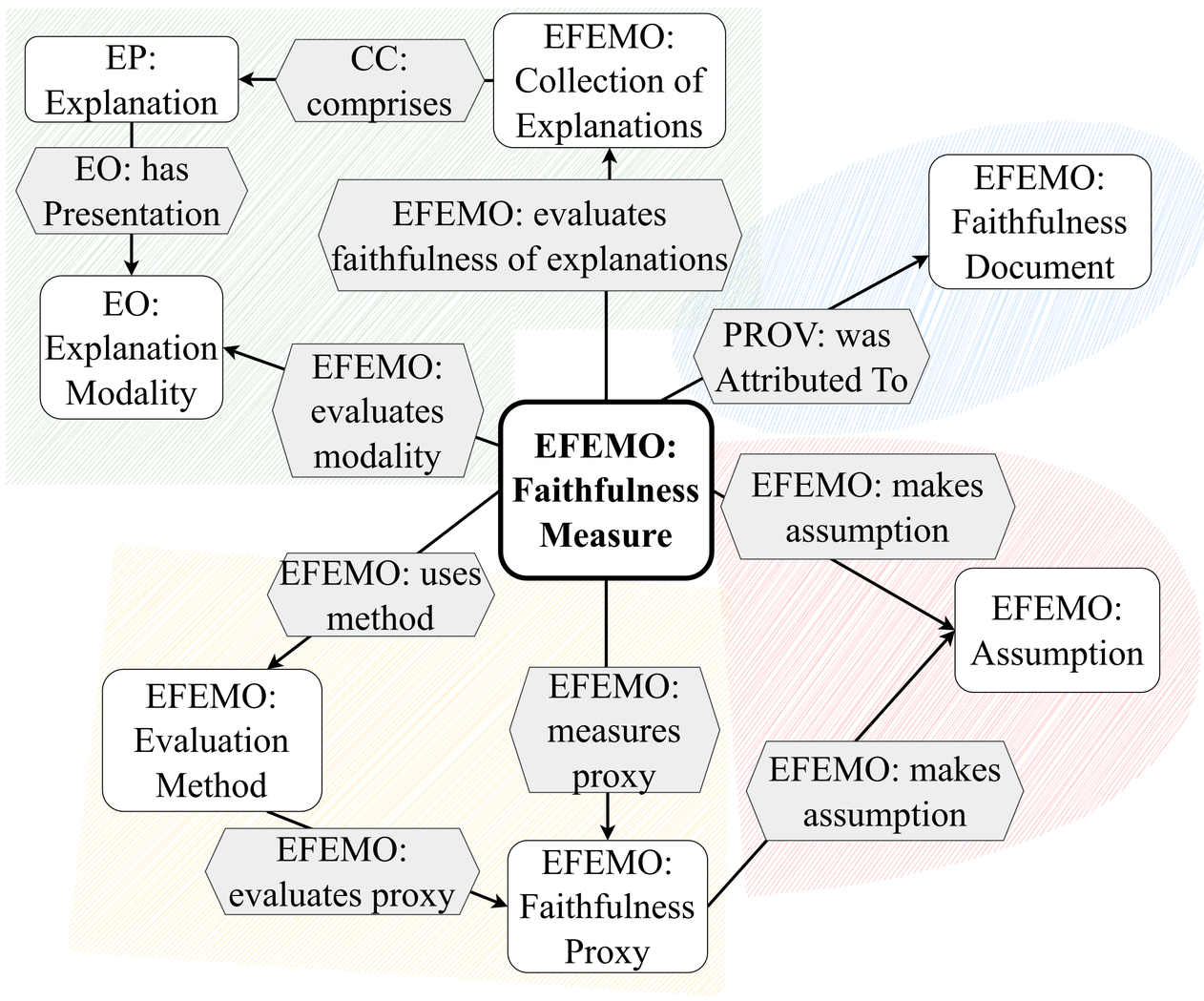

## Background

We consider an *explanation* to be anything that provides information about the system's workings and the knowledge used in its general reasoning processes or the processes behind a specific decision [5]. It may be in natural language or a heat-map of which parts of the input were the most important. An explanation is *faithful* if it accurately represents that reasoning process [6].

There is no gold-standard or generally accepted way to measure faithfulness. It may be impossible to guarantee complete faithfulness [7]. Instead, related properties and proxy characteristics are measured. These may include self-consistency [8], sufficiency [9], or robustness equivalence [10], and are usually considered necessary conditions for faithfulness.

Others have tried to organize explainability methods and evaluation measures into taxonomies [4,11]. Unfortunately these provided limited information about each measure, do not provide recommendations, and must make arbitrary decisions on the order of levels. An ontology is a rich schema, used organize a large set of information that includes formal definitions of all terms used [12]. Since it is more flexible than a taxonomy, we can avoid making arbitrary decisions. It also allows machines to explicitly reason over this data, making it very useful for recommendation systems.

## Model




| Prefix | Ontology |
|---|---|
| EFEMO | Explanation Faithfulness Evaluation Measures Ontology |
| EO | Explanation Ontology |
| EP | Explanation Patterns Ontology |
| PROV | Provenance Ontology |
| CC | Commons Collections Ontology |

## Competency Questions

An explainability researcher is looking for faithfulness measures to evaluate an existing explanation method. They want to use a state-of-the-art model that produces chain-of-thought explanations, but it is only available through API inference. They are specifically interested in robustness-based evaluation methods since they have a large token budget and can let the code run over the weekend. The researcher would prefer a local measure to determine the faithfulness of each explanation individually.

```
PREFIX efemo:
<http://www.semanticweb.org/villad4/ontologies/efemo#>
PREFIX indv: <http://www.semanticweb.org/villad4/ontologies/
2025/11/efemo_indv#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX av:
<https://www.omg.org/spec/Commons/AnnotationVocabulary/>
SELECT ?measure ?source
WHERE {
        ?measure a efemo:Faithfulness_Measure .
        ?measure efemo:evaluates_modality
efemo:cot_explanation_modality .
        ?measure efemo:has_granularity efemo:local_scope .
        ?measure efemo:uses_method ?method .
        ?method a efemo:Robustness_Evaluation_Method .
        ?measure prov:wasAttributedTo ?doc .
        ?doc av:directSource ?source .
        ?measure efemo:requires_access_level ?access .
        FILTER NOT EXISTS{?access efemo:higher_access_level
efemo:inference_access_level .}
}
```

*The system first infers that the measure must only require inference-level access or lower and must be able to evaluate chain-of-thought style of explanation. The system also knows the user prefers measures that have a local granularity and use robustness-based methodologies. The system will first query for measures that have no access requirements higher than inference-level access, evaluate chain-of-thought explanation modality, are local measures, and use robustness-based methods. If any are found, the system will present the name of the measures and the paper the measures are from, unordered. If nothing is found, the system notifies the user and removes the restrictions on granularity and methods, presenting any newly found results.*

A financial advisor has been using an LLM when analyzing applications for home loans. The LLM provides natural language explanations along with its suggestions, but the advisor isn't sure if the explanations are actually good, since they're not an expert on explainability. A coworker suggested Dasgupta et al.'s (2022) local sufficiency measure, but the advisor isn't sure if they can use it, since they only have inference access to the model through the API.

```
PREFIX efemo:
<http://www.semanticweb.org/villad4/ontologies/efemo#>
PREFIX indv: <http://www.semanticweb.org/villad4/ontologies/
2025/11/efemo_indv#>
SELECT ?access
WHERE {
        indv:local_sufficiency efemo:requires_access_level ?
access
}
```

*The system finds the method called "local sufficiency" and returns any required access levels.*

## Future Work

- The ontology is currently publicly available, we plan on publishing the ontology in an official ontology repository, such as purl.org
- Submit a poster to the Knowledge Graph Conference and a resource paper to the International Semantic Web Conference
- Add a formal recommendation module
- Develop a more user-friendly interface for adding new measures and searching for existing ones
- Develop and release v2.0 with support for image-based and tabular-based explanation faithfulness measures
- Test the ontology for usability using human subjects, including both explainability and semantic experts as well as non-experts

## References

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv:2303.08774.
[2] Article 9 of the Regulation (EU) 2016/679 of the European Parliament and of the Council [2016] OJ L119/1
[3] Agarwal, C., Tanneru, S. H., & Lakkaraju, H. (2024). Faithfulness vs. plausibility: On the (un) reliability of explanations from large language models. arXiv preprint arXiv:2402.04614.
[4] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, October). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA) (pp. 80-89). IEEE.
[5] Chari, S., Gruen, D. M., Seneviratne, O., & McGuinness, D. L. (2020). Foundations of explainable knowledge-enabled systems. In Knowledge Graphs for eXplainable Artificial Intelligence: Foundations, Applications and Challenges (pp. 23-48). IOS Press.
[6] Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. arXiv preprint arXiv:2004.03685.
[7] Ju, Y., Zhang, Y., Yang, Z., Jiang, Z., Liu, K., & Zhao, J. (2022, May). Logic traps in evaluating attribution scores. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 5911-5922).
[8] Parcalabescu, L., & Frank, A. (2023). On measuring faithfulness of natural language explanations. CoRR.
[9] Dasgupta, S., Frost, N., & Moshkovitz, M. (2022, June). Framework for evaluating faithfulness of local explanations. In International Conference on Machine Learning (pp. 4794-4815). PMLR.
[10] Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in nlp: A survey. Computational Linguistics, 50(2), 657-723.
[11] Alangari, N., El Bachir Menai, M., Mathkour, H., & Almosallam, I. (2023). Exploring evaluation methods for interpretable machine learning: A survey. Information, 14(8), 469.
[12] Kendall, E. F., & McGuinness, D. L. (2019). Ontology engineering. Morgan & Claypool Publishers.

Visit our website for more information: