EFEMO-KG: A Knowledge Graph for Explanation Faithfulness Measures

Danielle Villa[1,*], Deborah L. McGuinness[1]

[1]Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180, United States of America

Abstract

As AI systems advance, there is a growing need to ensure that the explanations of those systems are faithful to the true decision-making process of the system. A large number of techniques have been proposed to evaluate these explanations, necessitating a significant literature review to determine the most appropriate measure for an experiment. The Explanation Faithfulness Evaluation Measures Ontology-based Knowledge Graph (EFEMO-KG) provides a structured representation of these measures in a comparable format. Our pipeline of prompt templates, document-grounded question answering, and symbolic reasoning helps accurately build this knowledge graph. Further development and evaluation will lead to the publication of EFEMO-KG as an available online resource.

Keywords

Knowledge Graph, Explanation Evaluation, Knowledge Graph Generation

1. Introduction

The use of large language models (LLMs) has rapidly increased in recent years and has led to the development of more advanced artificial intelligence (AI) systems. Explanations of an AI system's decisions are often generated to help improve a user's trust in that system. However these explanations may not accurately reflect the true decision-making process. This aspect of an explanation is known as its *faithfulness* [1]. It has been shown that explanations are not inherently faithful [2, 3] and therefore they must be evaluated before conclusions about trustworthiness can be made. Currently there is no conclusive ground-truth method for determining faithfulness.

This gap has led to the development of a large number of techniques for evaluating explanation faithfulness. These measures vary across the explanations they evaluate, the mechanisms they use as proxies for faithfulness, and the methods they use in the evaluations [4]. This volume and variety of measures makes it difficult for researchers to identify a measure appropriate for their work. The Explanation Faithfulness Evaluation Measures Ontology (EFEMO) provides a schema for comparing measures, but is not a resource for finding those measures.

A centralized resource would benefit researchers looking for measures. A knowledge graph (KG) based on EFEMO would allow for the collection and comparison of these measures. However, manually building a knowledge graph is time intensive and requires subject matter expertise. It requires both identifying scientific papers that propose faithfulness measures and extracting the appropriate triples from each paper.

This paper is preliminary work for the EFEMO-based Knowledge Graph (EFEMO-KG). It introduces a pipeline for knowledge graph construction based on EFEMO and grounded in scientific publications. This pipeline is based on custom prompt templates and document-grounded AI workspaces. Initial evaluation is based on a dataset of manually created KGs and shows that many aspects of these papers can be reliably extracted. Further development and evaluation is ongoing.

XAI-KG@ISWC2026 - 2nd International Workshop on eXplainable AI and Knowledge Graphs, co-located with ISWC 2026, October 25-29, 2026, Bari, Italy

*Corresponding author.

villad4@rpi.edu (D. Villa); dlm@cs.rpi.edu (D. L. McGuinness)

https://www.cs.rpi.edu/~dlm/ (D. L. McGuinness)

0000-0001-7037-4567 (D. L. McGuinness)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-07

CEUR-WS.org/Vol-4318/paper5.pdf

2. Related Work

One issue in explainable AI is that explanations are not always evaluated, especially the faithfulness of those explanations [1]. This is due to the imprecise definition of faithfulness. This has led to a wide variety of proposed measures, but there is currently no centralized resource to find and compare those measures. Similar resources have been successful for other aspects of evaluation, such as fairness [5].

The number of faithfulness measures proposed creates a problem for scaling any manual creation methods. Automatic knowledge graph construction methods have been proposed to convert free-text into structured triples based on an ontology. Most of these systems take one or a few sentences as input and create a triple by extracting entities and their relations [6, 7]. The Prompting as Probing approach [8] takes advantage of the capabilities of LLMs for knowledge base construction. It uses manual prompt curation, including true/false questions and a dictionary of entity aliases, to increase precision. While this system is not built to process full scientific papers, these techniques are critical to our construction pipeline.

Recently, there have been proposed systems that can process an entire paper. The IDEA project using the SUDO ontology[1] builds a knowledge graph by extracting the arguments, artifacts, and claims from a paper. However it is difficult to map the resulting knowledge graph from the SUDO ontology to a more specific ontology such as EFEMO. Additionally the process of converting a PDF to a text document for use in this system risks losing information. We have not found such issues when using the document-grounded workspaces in our pipeline.

Our work takes advantage of the rise of agentic AI systems and document-grounded workspaces. These systems, such as Gemini Notebook[2], allow users to ask questions about specific documents and receive well-supported answers. Compared to traditional chatbot-style language models, this setup allows us to easily use PDFs as sources and receive precise quotes from those sources as citations. Previous work often uses traditional LLMs that require users to extract the text from their document and may have higher rates of citation hallucination [9].

3. The Explanation Faithfulness Evaluation Measures Ontology

The Explanation Faithfulness Evaluation Measures Ontology[3] (EFEMO) [4] is an ontology built to facilitate the organization and comparison of explanation faithfulness measures. To design the ontology, we began with a literature analysis centered around [1] to help ensure a consistent definition of explanation faithfulness. The papers found in this literature analysis that introduced a faithfulness measure were collected to form the sources of a knowledge graph based on EFEMO.

EFEMO models many different aspects of each faithfulness measure. For the EFEMO-KG, these are:

- Proxy Characteristic Measured - What concrete property the measure evaluates
- Explanation Modality Evaluated - What type of explanation the measure evaluates
- Required Model Access - How much access to the model is needed to run the measure
- Output Range - Whether the measure makes a binary decision or measures the degree of faithfulness
- Granularity - Whether the measure is local to a single explanation or determines the faithfulness of a whole model
- Model Specificity - Whether the measure is specific to a model architecture
- Evaluation Method Category - What type(s) of techniques are used by the measure, according to the taxonomies of [10, 11, 12]

EFEMO also tracks the assumptions made by each faithfulness measure by inferring them from the measure's evaluation method [10, 11, 12] and proxy characteristic [1, 10, 13]. The provenance of each measure is tracked using the Provenance Ontology (PROV-O) [14].

[1] Available at: https://tetherless-world.github.io/sudo-ontology/ontology/releases/3.3.0/docs/index-en.html
[2] Formerly NotebookLM (renamed July 2026), found at: https://notebooklm.google/
[3] Available at our project website: https://tetherless-world.github.io/faithfulness-measures-ontology/

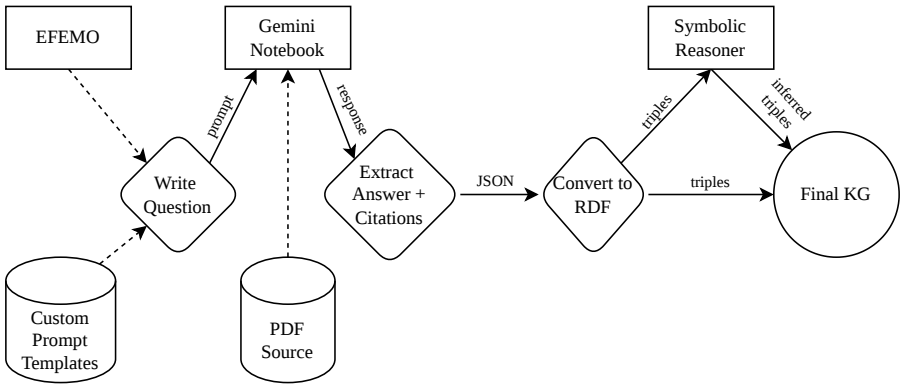


Figure 1: The pipeline for constructing an EFEMO-based knowledge graph from a source document.

Table 1
Example prompt template, using EFEMO and SKOS (the Simple Knowledge Organization System ontology [15]).

Prompt Template
Does document [DOC NAME] introduce any faithfulness measures? A faithfulness measure is {definition[Faithfulness_Measure]}. Present the name and a description of each faithfulness measure introduced in JSON format. Follow this example: {example[Faithfulness_Measure]}
Triples Searched
(efemo:Faithfulness_Measure, skos:definition, *xsd:string*) (efemo:Faithfulness_Measure, skos:example, *xsd:string*)

In addition to characteristics of the measures themselves, EFEMO also records the datasets and explanation generation methods used in the experiments demonstrating each measure.

4. Knowledge Graph Construction

We used the literature analysis performed in [4] to develop EFEMO to identify papers that present explanation faithfulness measures. 93 such papers were identified. In order to create a knowledge graph of these measures based on EFEMO, we developed an LLM-based construction pipeline. We began by selecting a subset of these papers and manually creating a knowledge graph of their measures. This was used to evaluate the quality of the construction pipeline.

4.1. Dataset Creation

Of the 93 measure-introducing papers, 33 were randomly selected as part of the evaluation dataset. The average year of release of these papers was 2021. The average length of the papers (excluding references and appendices) was 10 pages.

Each paper was read to determine the introduced measures and their relevant properties, as described in Sec. 3. The sources of each measure's properties were also noted. This was done by one team member familiar with EFEMO. On average, it took 15.2 minutes to evaluate each paper. Basic statistics for the cumulative KG are in Appendix A.

4.2. Construction Pipeline

An overview of the automated construction pipeline is shown in Figure 1. We begin the knowledge graph construction process with a set of custom prompt templates[4] (e.g. Table 1). Similar to [8], we find that prompts that provide examples of variable length answers with specific formatting instructions help ensure that the AI model's responses are correct and usable. Simple true/false or other binary choice questions are also used. When asking more complex questions, our prompts ask the model to provide its answer in JSON format for easy extraction.

Our prompt templates include slots for definitions, examples, and other properties of concepts in the ontology. These slots are filled by the information present in the EFEMO ontology, retrieved from the online resource. This allows us to use the most up-to-date version of these properties in our questions without changing the prompt templates.

Once filled in, these prompts become questions across three categories: the datasets used by the paper, the methods used to generate the explanations evaluated by the paper, and the measures used to evaluate those explanations. We have one question about datasets and one question about explanation methods. One question is asked to determine the number of measures introduced in the paper, and for each measure we ask twelve additional questions to gather specific details.

After uploading the source to the notebook, we prompt[5] the model, Gemini 3.5 Flash, with the questions[6]. The responses to these questions are then extracted as JSON format. All citations from responses are also saved, including the document quotes the model claims to have used. These are then transformed into a knowledge graph for each paper. The Pellet reasoner [16] is then used to determine all inferences that can be made and these are added to each knowledge graph. These final graphs are used in the evaluation. Basic statistics for the cumulative extracted KG are in Appendix A.

4.3. Evaluation

The knowledge graphs generated by the pipeline were manually compared to those in the ground-truth dataset. For each paper, the evaluation was done in four steps:

1. Comparison of Datasets used: A dataset is considered correct if it appears in the ground truth with the same label and source document.
2. Comparison of Explanation Generation Methods used: An explanation generation method is considered correct if it appears in the ground truth with the same label and source document.
3. Comparison of Faithfulness Measures: A faithfulness measure is considered correctly identified if it can be matched to one in the ground truth. The definition of the measure is used to determine if slightly different names refer to the same measure.
4. Comparison of Individual Faithfulness Measure Details: For each measure that appears in both the ground truth and the generated knowledge graphs, we compare six properties for correctness: explanation modality evaluated, required model access, output range, measure granularity, model specificity, and evaluation method category.

The specific faithfulness measure details to evaluate were chosen because they result in triples in the knowledge graph. Some of the information extracted in the construction process becomes annotations and, while they are used to help resolve entities, they are not compared to those in the ground-truth KG. Additional evaluation details are given in Appendix B.

Our comparison is based on aspects of the paper and its proposed measures that are extracted from the original source paper. Other aspects with more abstract values, such as the description of the evaluation and the proxy characteristic, were not included in our evaluation.

Table 2
Results of the evaluation, divided into the four evaluation categories.

	Accuracy	Precision	Recall	F1 Score
Dataset	0.94	0.96	0.97	0.96
Explanation Generation Method	0.61	0.61	1.0	0.76
Measure Identification	0.58	0.61	0.92	0.73
Measure Details	0.49	0.54	0.83	0.65

Table 3
Results of the evaluation of the measure details, per aspect of faithfulness measure evaluated.

	Explanation Modality	Model Access	Output Range	Granularity	Model Specificity	Evaluation Method
Accuracy	0.48	0.42	0.69	0.57	0.41	0.47
Precision	0.48	0.53	0.82	0.73	0.47	0.53
Recall	0.96	0.76	0.82	0.73	0.78	0.80

4.4. Results

The KG construction pipeline was used to generate knowledge graphs on the measures for 10 papers from the dataset[7]. The limited evaluation was due to time constraints. Overall, the knowledge graphs were 57% accurate to the ground-truth dataset. The results of the evaluation are shown in Tables 2 and 3. The full set of data counts are in Appendix C.

The knowledge graph generation process is the most accurate at extracting the datasets used in experiments. This is likely because many papers have an explicit *Dataset* section or subsection. Explanation generation method extraction was also done well. The extraction of the measures themselves was not as consistent. The precision of both the measure identification and measure details are low. We suspect this is because the terms used in EFEMO are infrequently used explicitly in the wider literature, and as such the AI model often had to infer the properties based on the definitions in the ontology and the descriptions in the source documents.

The specific measure details of output range and granularity were the most accurate. This is expected since these properties each have only two options. Other properties have up to 14 options that a measure can have multiple of. We find that the precision of these details is also much lower than the recall.

While we do hope to increase accuracy, our primary concern is the precision of the results. EFEMO-KG uses the open world assumption, so the lack of a property is not assumed to mean that the property is not there. Decreasing the number of incorrect assertions in the graph is therefore the primary goal.

Evaluation of additional papers is ongoing.

5. Future Work

As we continue this work, in addition to creating explanation faithfulness evaluation measure knowledge graphs for more papers, we plan on expanding our evaluation. We will begin by including established knowledge graph quality metrics, such as [17]. We will also do a full comparison to a baseline, such as the IDEA project pipeline. Additionally, we will investigate an automated way to compare the generated knowledge graphs to the ground truth knowledge graphs without punishing insubstantial differences in free-text properties such as measure description and proxy characteristic. We would also like to evaluate the quality of the document quotes cited by the model.

[4]The full set of prompt templates is at: https://tetherless-world.github.io/faithfulness-measures-ontology/kg_prompts
[5]The notebook was access programatically using the NotebookLM-py package: https://github.com/teng-lin/notebooklm-py/
[6]The data for this paper was generated from July-August 2026.
[7]The set of papers used is at: https://tetherless-world.github.io/faithfulness-measures-ontology/kg_papers

The primary goal our future work is to improve the extracted knowledge graph. Specifically, we aim to increase the precision of the process by mitigating the errors and not allowing them to propagate. We will begin by analyzing the current errors and investigating if there are patterns. For the low precision, we will evaluate whether the model cited relevant passages and interpreted them correctly. For the categories with low recall, we will investigate the citations for the ground-truth KG and see if there is a difference between the citations extracted by the model and those missed.

We will then work to resolve those errors. We will investigate techniques such as relying on the text citations, using a second model to judge the first, and iterating on the prompts with specific few-shot examples. In addition to improving the correctness of the knowledge graphs, we plan on doing entity resolution to identify when multiple papers use the same datasets and explanation generation methods.

A goal we continue to work towards is providing access to EFEMO-KG as an online resource. Both the ground-truth and generated KG will be made available under the MIT license[8]. This would include a SPARQL endpoint and an LLM-based tool for exploring the data and providing recommendations.

6. Conclusion

As complex AI systems continue to advance in complexity and capability, ensuring that their decisions can be explained faithfully becomes more important. This requires researchers to evaluate their explanations using faithfulness measures, and a resource like EFEMO-KG could help researchers identify appropriate measures. Our work is a start to developing the tools necessary to build and maintain such a resource. We will continue to improve the process of automatically constructing this knowledge graph directly from scientific papers and further evaluate its correctness and efficiency.

Acknowledgments

Thanks to the members of the RPI Tetherless World Constellation for their advice and guidance with this work. Special thanks to Dr. Maria Chang for her mentorship.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Jacovi, Y. Goldberg, Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?, arXiv preprint arXiv:2004.03685 (2020).

[2] M. Turpin, J. Michael, E. Perez, S. Bowman, Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting, Advances in Neural Information Processing Systems 36 (2024).

[3] J. Lindsey, W. Gurnee, E. Ameisen, B. Chen, A. Pearce, N. L. Turner, C. Citro, D. Abrahams, S. Carter, B. Hosmer, J. Marcus, M. Sklar, A. Templeton, T. Bricken, C. McDougall, H. Cunningham, T. Henighan, A. Jermyn, A. Jones, A. Persic, Z. Qi, T. B. Thompson, S. Zimmerman, K. Rivoire, T. Conerly, C. Olah, J. Batson, On the biology of a large language model, Transformer Circuits Thread (2025). URL: https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

[4] D. Villa, D. L. McGuinness, Modeling measures for evaluating llm explanation faithfulness, 2026. The Knowledge Graph Conference 2026: Poster Track.

[5] J. S. Franklin, K. Bhanot, M. Ghalwash, K. P. Bennett, J. McCusker, D. L. McGuinness, An ontology for fairness metrics, in: Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society,

[8] http://opensource.org/licenses/MIT

AIES '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 265–275. URL: https://doi.org/10.1145/3514094.3534137. doi:10.1145/3514094.3534137.
[6] Z. Bi, J. Chen, Y. Jiang, F. Xiong, W. Guo, H. Chen, N. Zhang, Codekgc: Code language model for generative knowledge graph construction, ACM Transactions on Asian and Low-Resource Language Information Processing 23 (2024) 1–16.
[7] N. Mihindukulasooriya, S. Tiwari, C. F. Enguix, K. Lata, Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text, in: International semantic web conference, Springer, 2023, pp. 247–265.
[8] D. Alivanistos, S. B. Santamaría, M. Cochez, J.-C. Kalo, E. van Krieken, T. Thanapalasingam, Prompting as probing: Using language models for knowledge base construction, arXiv preprint arXiv:2208.11057 (2022).
[9] H. Khorashadizadeh, N. Mihindukulasooriya, S. Tiwari, J. Groppe, S. Groppe, Exploring in-context learning capabilities of foundation models for generating knowledge graphs from text, arXiv preprint arXiv:2305.08804 (2023).
[10] Q. Lyu, M. Apidianaki, C. Callison-Burch, Towards faithful model explanation in nlp: A survey, Computational Linguistics (2024) 1–70.
[11] N. Alangari, M. El Bachir Menai, H. Mathkour, I. Almosallam, Exploring evaluation methods for interpretable machine learning: A survey, Information 14 (2023) 469.
[12] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques, arXiv preprint arXiv:1909.03012 (2019).
[13] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, L. Kagal, Explaining explanations: An overview of interpretability of machine learning, in: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA), IEEE, 2018, pp. 80–89.
[14] Belhajjame, Khalid and Cheney, James and Corsar, David and Garijo, Daniel and Soiland-Reyes, Stian and Zednik, Stephan and Zhao, Jun, Prov-o: The prov ontology, https://www.w3.org/TR/prov-o/, 2013. Accessed: 2025-12-10.
[15] A. Miles, J. R. Pérez-Agüera, Skos: Simple knowledge organisation for the web, Cataloging & Classification Quarterly 43 (2007) 69–83.
[16] E. Sirin, B. Parsia, B. C. Grau, A. Kalyanpur, Y. Katz, Pellet: A practical owl-dl reasoner, Journal of Web Semantics 5 (2007) 51–53.
[17] M. J. Luthfi, F. Darari, A. C. Ashardian, Sock: Shacl on completeness knowledge., in: WOP@ ISWC, 2022.

A. Exploratory Data Analysis of Ground-Truth Dataset

In the ground-truth KG there were 12 measures presented across the 10 papers initially evaluated. Each measure is included in 34.6 triples on average. The entire ground-truth KG has 168 entities and 416 triples.

In the AI extracted KG there were 18 measures found across the same 10 papers. Each measure is included in 46.7 triples on average. The entire extracted KG has 841 entities and 329 triples.

B. Evaluation Details

Each measure was evaluated individually and compared to the ground-truth KG.

Measure Identification was evaluated by comparing the names and descriptions of the extracted measure(s) to the ground-truth measure(s) from the same paper. The text cited with the measure was also used to determine if the extracted measure was one of the ground-truth measures. Note that if a measure was not found, the measure detail evaluation is skipped. Additionally, false negatives are not counted if they are details for a measure that was not extracted.

Table 4
Counts of the evaluation, divided into the four main evaluation categories.

	True Positive	False Positive	False Negative
Dataset	31	1	1
Explanation Method	27	17	0
Measure Identification	11	7	1
Measure Details	81	69	17

Table 5
Counts of the evaluation of the measure details, per aspect of faithfulness measure evaluated.

	Explanation Modality	Model Access	Output Range	Granularity	Model Specificity	Evaluation Method
True Pos.	29	8	9	8	7	20
False Pos.	31	7	2	3	8	18
False Neg.	1	4	2	3	2	5

Model Specificity was evaluated similarly, by comparing to the ground-truth KG and using the cited text to determine matches.

The **Datasets Used** and **Explanation Generation Method** related to the measure used were evaluated by comparing to the extracted set to the ground-truth set. These were only considered correct if they appeared in the ground-truth set and were linked to the correct citation.

Many of the characteristics we evaluated were categorical and the AI model was instructed to select the applicable categories from a list. These were compared to the set of categories that measure belonged to in the ground-truth KG. The characteristics evaluated this way were: **Explanation Modality**, **Model Access**, and **Evaluation Method Category**.

There are two characteristics that are binary: **Output Range** and **Measure Granularity**. These were evaluated based on whether the extracted KG related the correct option compared to the ground-truth KG.

For each category we evaluated, we noted whether how many true positive, false positive, and false negative triples occurred. True positive triples appeared in both the extracted KG and the ground-truth KG. False positive triples appeared in the extracted KG but not the ground-truth KG. False negative triples appeared in the ground-truth KG but not the extracted KG. For example if the generated KG connected a measure to the wrong output range (a binary characteristic), both a false positive and a false negative would be recorded during the evaluation. While the datasets used, measure identification, and explanation generation method were all evaluated and counted, we did not count the triples relating to a measure that did not appear in both KGs were not counted.

C. Raw Counts

Tables 4 and 5 have the number of true positive, false positive, and false negative instances per measure characteristic evaluated. Since the task is extracting characteristics, there are no true negatives.