A Human-Centric Evaluation Platform for Explainable Knowledge Graph Completion

Zhao Xu[1], **Wiem Ben Rim**[1], **Kiril Gashteovski**[1,2], **Timo Sztyler**[1], **Carolin Lawrence**[1]
[1]NEC Laboratories Europe, Heidelberg, Germany
[2]CAIR, Ss. Cyril and Methodius University, Skopje, North Macedonia
`firstname.lastname@neclab.eu`

Abstract

Explanations for AI are expected to help human users understand AI-driven predictions. Evaluating plausibility, the helpfulness of the explanations, is therefore essential for developing eXplainable AI (XAI) that can really aid human users. Here we propose a human-centric evaluation platform[1] to measure plausibility of explanations in the context of eXplainable Knowledge Graph Completion (XKGC). The target audience of the platform are researchers and practitioners who want to 1) investigate real needs and interests of their target users in XKGC, 2) evaluate the plausibility of the XKGC methods. We showcase these two use cases in an experimental setting to illustrate what results can be achieved with our system.

1 Introduction

A Knowledge Graph (KG) is a structured representation of knowledge that captures the relationships between entities. It is composed of triples in the format *(subject, relation, object)*, denoted as $t = (s, r, o)$, where two entities are connected by a specified relation. For example, in the triple *(London, isCapitalOf, UK)*, *London* and *UK* are the entities, and *isCapitalOf* is the relation. These entities can be depicted as nodes in a knowledge graph, while the relation denotes a labeled link connecting the subject to the object. Knowledge graphs are beneficial for many NLP tasks, e.g., fact checking (Hu et al., 2021; Kim et al., 2023), question answering (Hu et al., 2022; Srivastava et al., 2021) and information extraction (Gashteovski et al., 2020).

The applicability of KGs in downstream tasks, however, is often limited by their incompleteness (Saxena et al., 2022): they do not contain exhaustive information about *all* relationships between the defined entities (Destandau and Fekete, 2021). To address this issue, researchers and practitioners have worked on Knowledge Graph Completion (KGC): the task of predicting new relationships between the entities in the knowledge graph. For this, two parts of a triple (i.e., *slots*) are given to a KGC system (Rossi et al. (2021); Lin et al. (2018), *inter alia*) and the third is inferred; e.g., answering the query $t = (s, r, ?)$. Such methods learn low-dimensional representations of entities and relations for predictive inference.

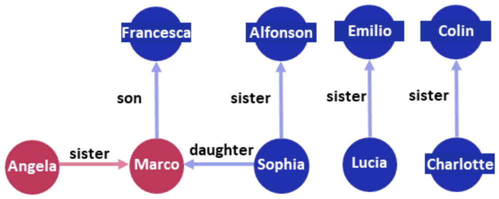


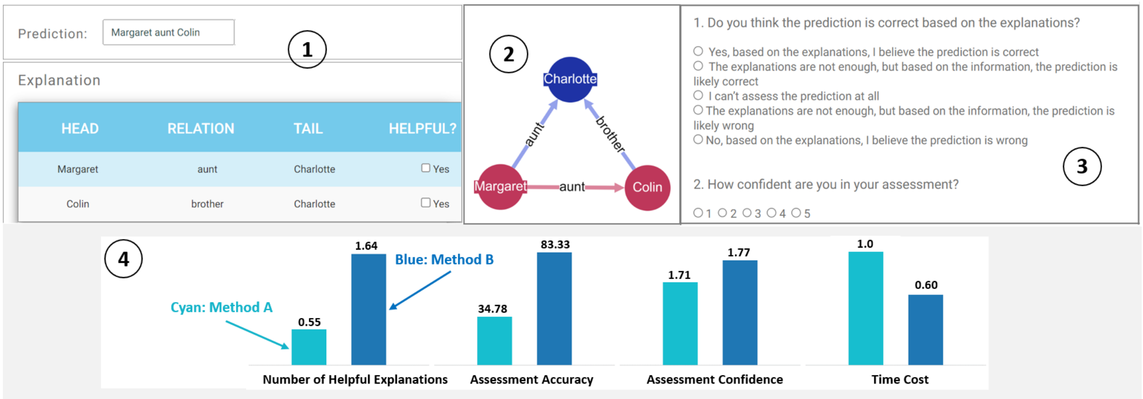
Figure 1: An example explaining a predicted triple (in red) with important training triples (in blue), learned according to gradients by Lawrence et al. (2021). They are faithful, yet not helpful for users to understand the prediction.

The embedding based KGC models, however, are black boxes that do not (and cannot) provide explanations of why the model makes a certain prediction. The lack of transparency significantly hampers users' trust and engagement with KGC systems, especially in the high-risk domains, such as medicine (Han and Liu, 2022; Chaddad et al., 2023). To provide explanations for such embedding-based KGC systems, researchers have proposed explainable KGC (XKGC) methods (Betz et al., 2022; Lawrence et al., 2021; Pezeshkpour et al., 2019). However, it remains unexplored how helpful users find the explanations provided by these methods. For instance, Figure 1 shows an example explanation that would not be helpful for the end user.

We thus target to evaluate what kind of explanations are *helpful* for the users because ultimately,

[1] The video of the demo: `https://www.dropbox.com/scl/fi/p2sczcyvqk6zyr9omcf1e/eacl2024EvaXKGC.mp4?rlkey=j2pvz8alqihxmyiv5q7cxkx1z&dl=0`.
The live demo website of the human-evaluation platform: `https://xai.privacy.nlehd.de/start-evaluation`.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics System Demonstrations, pages 18–26
March 17-22, 2024 ©2024 Association for Computational Linguistics



Prediction: Margaret aunt Colin

Explanation

HEAD	RELATION	TAIL	HELPFUL?
Margaret	aunt	Charlotte	□ Yes
Colin	brother	Charlotte	□ Yes

1. Do you think the prediction is correct based on the explanations?

○ Yes, based on the explanations, I believe the prediction is correct
○ The explanations are not enough, but based on the information, the prediction is likely correct
○ I can't assess the prediction at all
○ The explanations are not enough, but based on the information, the prediction is likely wrong
○ No, based on the explanations, I believe the prediction is wrong

2. How confident are you in your assessment?

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 2: Our evaluation platform to measure plausibility of XKGCs with human-centric evaluation. (1) Shows the prediction and explanations to human testers. (2) Visualizes the prediction and explanations as a graph for the testers to easily comprehend and reason about the relationships. Users can evaluate the helpfulness of the explanations by clicking either the tick boxes in (1) or the edges in the visualized graph (2). (3) Asks the testers to assess the correctness of the prediction based on the explanations. (4) After collecting feedback from N (defined by researchers) testers, plausibility of XKGC is measured with: number of helpful explanations (helpExpl), accuracy (Acc) and confidence of testers assessment, time cost. More details can be found in Sec. 6.

the explanations should directly aid them. Therefore, it is important to measure the *plausibility* of the explanations: the extent to which an explanation generated by XAI is comprehensible and beneficial to human users (Jacovi and Goldberg, 2020; Lage et al., 2019). Thus, to evaluate the plausibility of the explanations, we present a human-centric platform illustrated in Figure 2.

Our evaluation platform offers the following novel contributions. First, it introduces a new evaluation paradigm that assesses how well explanations can assist users in judging the correctness of KGC predictions. In contrast to the prevalent human evaluation paradigm in the literature that requests annotators to simulate AI's behavior (Yin and Neubig, 2022; Hase and Bansal, 2020; Lage et al., 2019; Doshi-Velez and Kim, 2017), the new paradigm aligns better with real human-AI interaction systems, where AIs facilitate humans rather than the other way around. Furthermore, given the growing complexity of AI, it becomes increasingly challenging for annotators to imitate AI's behavior without comprehensive training, especially when utilizing crowdsourcing platforms like Amazon Mechanical Turk (AMT) (Clark et al., 2021). Another notable advantage of our system is its capability to quantify the helpfulness of explanations in an objective manner. Our system suggests metrics, such as the accuracy rate of annotators' judgments, which stems itself from well-defined ground truth to quantitatively measure human feedback.

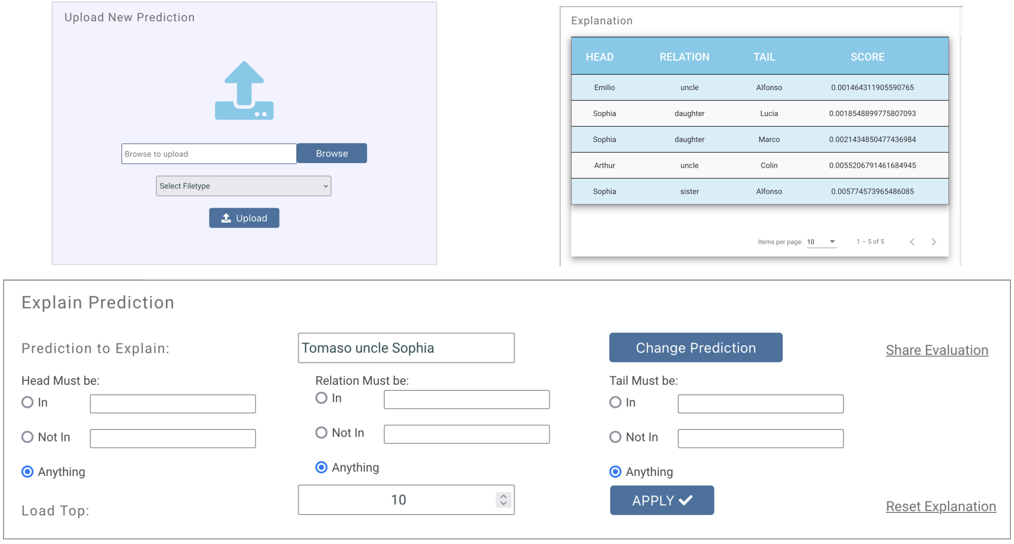
With these novel contributions, our evaluation platform can effectively measure plausibility of XKGC methods. Considering the diversity of humans, our system also provides various statistical tools to rigorously and comprehensively analyze the collected feedback for reliable conclusions. Additionally, our evaluation platform aids in identifying genuine requirements from users regarding explanations, thereby it can assist in developing and refining XKGC methods to generate explanations that are centered around human needs. Finally, we formulate our study on human-centric evaluation as practical guidelines, which can be replicated to design evaluations for other use cases in the future.

2 Human Centric Evaluation for XKGC

We build an online system to evaluate XKGCs in a human centric manner. Our system considers the real needs and interests of human users in collaboration with AI, allowing us to investigate: *can humans assess correctness of a KGC prediction based on its explanations? Which explanations are helpful for human users?* The answers to these questions provide hints for evaluating the ultimate goal of an XAI method: the generated explanations are expected to assist human users in understanding AI-driven predictions. To this end, our platform has two user views: one for researchers to set up a test and the other for testers to give feedback.

2.1 Researcher

Researchers can prepare the evaluation study by uploading a JSON file that contains both the predic-



HEAD	RELATION	TAIL	SCORE
Emilio	uncle	Alfonso	0.001464311905590765
Sophia	daughter	Lucia	0.0018548899775807093
Sophia	daughter	Marco	0.0021434850477436984
Arthur	uncle	Colin	0.0055206791461684945
Sophia	sister	Alfonso	0.005774573965486085

Items per page: 10 1 – 5 of 5

Prediction to Explain: Tomaso uncle Sophia

Head Must be: ○ In ○ Not In ● Anything

Relation Must be: ○ In ○ Not In ● Anything

Tail Must be: ○ In ○ Not In ● Anything

Load Top: 10

Figure 3: Top-left: the page to load the input JSON file about predictions and explanations to be evaluated. Top-right: the overview of the predictions in the input file for the researchers to check. Bottom: the filter for the researchers to select predictions and explanations to show to testers. The researchers can select any number of predictions they need for user evaluation.

tions and possible explanations. Here is an example JSON file including one prediction and its explanation. If the researchers want to evaluate multiple predictions, then they only need to add these predictions in the json file.

```
"Colin son James": {
    "correct": 1,
    "probability": 0.56884,
    "explanation": [
        [
            [
                "James",
                "father",
                "Colin"
            ],
            0.198171710573O8347
        ],
        [
            [
                "Charlotte",
                "sister",
                "Colin"
            ],
            0.217222705276661
        ],
}
```

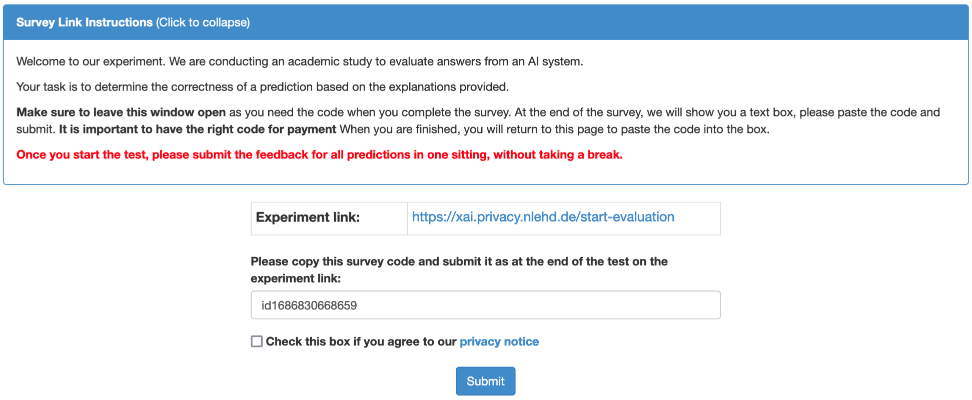
Each predicted triple is associated with a set of *explanation* triples. Each explanation triple has a score that indicates their importance, which can be used for filtering and ranking the explanations. This score can e.g. come from the XKGC method. In addition, each prediction has the *correct* attribute which indicates whether this prediction is correct or not. The false prediction can be viewed as a control setup, which allows us to test whether users can determine if a prediction is correct based on the given explanations. Additionally, it allows us to assess the engagement of testers (see Sec. 5 for details). The *probability* attribute specifies the likelihood of the predicted triple by the KGC method.

After the JSON is uploaded (top-left panel of Figure 3), the system lists all triples for the researchers to check (top-right panel). Next, the researchers can click on a particular triple to see its explanations as well as a filtering options (bottom panel). With the filtering options, the researcher can choose which predicted triples and explanations they would like to keep for the human evaluation. Finally, the system shows a preview page where the researchers can check the evaluation test that will be displayed for the testers.

2.2 Tester

After the evaluation test has been setup, the researchers can share the link of the online system with the testers to evaluate. The system can work with crowdsourcing websites, e.g. Amazon Mechanical Turk (AMT), to employ testers for human evaluations. Figure 4 illustrates how an evaluation task can be set up with our system on AMT.

The top panels of Figure 2 showcase the interface for testers. For each prediction, the tester can inspect the explanations, which are displayed in two formats (table and graph). Panel (1) shows the prediction and explanations in a format of table. For the testers to easily comprehend and reason about the relationships, the prediction and explanations are also visualized as a graph, shown as Panel (2). Based on the explanations, they can decide on whether they believe a prediction to be correct



Survey Link Instructions (Click to collapse)

Welcome to our experiment. We are conducting an academic study to evaluate answers from an AI system.

Your task is to determine the correctness of a prediction based on the explanations provided.

Make sure to leave this window open as you need the code when you complete the survey. At the end of the survey, we will show you a text box, please paste the code and submit. **It is important to have the right code for payment** When you are finished, you will return to this page to paste the code into the box.

Once you start the test, please submit the feedback for all predictions in one sitting, without taking a break.

Experiment link:	https://xai.privacy.nlehd.de/start-evaluation

Please copy this survey code and submit it as at the end of the test on the experiment link:

id1686830668659

□ **Check this box if you agree to our privacy notice**

Figure 4: Launch an human evaluation study based on our system in Amazon Mechanical Turk.

or not on a scale from 1 to 5. In addition, they need to specify whether an explanation is *Helpful*. This can either be done by clicking a tick box in the explanation table (see Panel (1)) or by clicking on an edge in the graph to mark the corresponding triple as helpful (see Panel (2)). The selections of the user will be synchronized in both formats.

Once done, the tester can submit the feedback and move on to the next prediction. After the last prediction, we offer the tester an additional form to share any feedback with us. This page can also be used, e.g. to share an identifying code that allows us to utilize the evaluation system with AMT, where the code is used to check completeness and engagement for payment.

3 Architecture of the Evaluation System

The system is as a web application consisting of frontend (HTML5/JavaScript) and backend (Python). We will describe the respective components and the data flow (shown in Figure 5) in detail.

3.1 Backend

The backend is a Python-based software framework (Flask[2]) providing multiple HTTP REST interfaces to enable human-centric data management, evaluation, feedback collection, and analysis. Combining these interfaces essentially leads to an all-in-one solution for conducting a human-centric evaluation. When it comes to data modeling, we ensure flexibility and scalability by using a key-value database (MongoDB[3]). While our solution also encompasses a frontend component, the versatility of the backend allows it to seamlessly integrate with any other application or system.

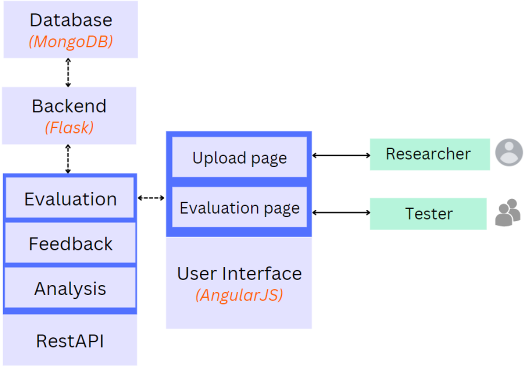


Figure 5: System architecture: the interaction of the researchers and the testers with their respective user interfaces and the overview of the backend and RestAPI.

3.2 Frontend

The frontend is implemented in JavaScript (AngularJS[4]), HTML5 and CSS, and provides user-friendly access to the functionalities provided by the backend. It consists of two environments, separating the evaluation and the data management. The data management includes uploading data, but also to specify filters and related settings to configure the evaluation.

3.3 Data Flow

Figure 5 illustrates an application example of our solution: First, a researcher interacts with the "Upload page" to upload the data (i.e., predictions and explanations) to be used in the evaluation. Then, she is redirected to the configuration page to, e.g., apply filters to the predictions (see Figure 3). Finally, the researcher can generate and share the URL to access the evaluation. When a tester visits the URL, the evaluation page presents the predictions to her one after the other and in a random

[2] https://palletsprojects.com/p/flask/

[3] https://www.mongodb.com/

[4] https://angularjs.org/

order. The tester submit her feedback on the predictions and the corresponding explanations. The evaluation results will be stored by the backend in our key-value database and can be downloaded as a JSON file.

Our system is deployed on a powerful server with 48 Threads (24 cores), 256 GB memory and 1GB Full-Duplex Internet connection. In theory, it can support more than one thousand testers to visit the evaluation platform.

4 Statistical Analysis of Human Feedback

Due to the complexity and costliness of human evaluation, as well as the diversity among human testers, the collected feedback tends to be both limited in quantity and diverse in quality. Consequently, statistical analysis assumes a critical role to draw reliable conclusions from human feedback. We include the following statistical analysis tools in our platform.

Power analysis. In human evaluation, there is often an important question: How many testers are necessary to draw a solid conclusion? There is no a universally applicable minimum sample size for obtaining statistically significant results (Hogg et al., 2015). Power analysis is commonly used in e.g., social-science and clinical research literature (Cohen, 1988), which determines an adequate sample size for a human evaluation based on a stated effect size that defines the difference level of the compared methods. As the effect size used in power analysis is prospectively anticipated before the evaluation by the researchers, it is good to analyze the post hoc power of an observed effect size derived from the collected human feedback, especially if the findings are non-significant (Onwuegbuzie and Leech, 2004).

Hypothesis testing. Are the observed results in human feedback statistically significant or simply due to chance? Hypothesis testing, e.g. t-tests, Wilcoxon Signed-Rank test, Mann Whitney test and Brunner-Munzel test, can be employed to measure them. With hypothesis tests, we can distinguish between real effects and random variations in a rigorous manner.

Mixed effect analysis. Human feedback is often subject to variability of individual differences, engagement levels, and other random variation. (Linear) mixed effect analysis (Bates et al., 2015) can thus be used to quantify and assess the variability within testers' responses. Specifically, it can measure both fixed effects (differences caused by the compared methods) and random effects (differences due to variation of individuals) quantitatively.

Correlation Analysis. In addition, correlation analysis can also be applicable to analyze the relations among different metrics. For example, we suggest multiple metrics to quantify plausibility, including: accuracy rate of tester's assessment, confidence of testers, number of helpful explanations, and time cost. Correlation analysis can explore relationships between metrics, and may provide insights into the reliability and validity of the results.

5 Guidelines for Human-Centric Evaluation

Human evaluation can be subject to various biases that may affect the reliability of the conclusions (Hase and Bansal, 2020; Chandrasekaran et al., 2018; Gajos and Mamykina, 2022). The following concerns need to be addressed.

Engagement. Testers often exhibit varying levels of engagement and various thinking modes. To mitigate the impact of tester bias, we propose that each tester assesses ≥ 2 XKGC methods, analyzing the feedback with paired tests, especially when the number of available testers is limited. Additionally, testers' engagement tends to decrease over time. Therefore, it is crucial to impose a constraint on the total evaluation time (e.g. one hour per session). Furthermore, to ensure the testers' proper engagement during the evaluation process, we can randomly assign some straightforward predictions as checkpoints for validation.

Equivalency. All testers should evaluate similar set of predictions in a similar order. This is to reduce deviations caused by individual predictions.

Diversity. Testers may have the tendency to retain information from previous predictions, which can result in the earlier assessments influencing the later ones. Consequently, we recommend selecting predictions that are as distinct from each other as possible to mitigate this concern.

Balance. Predictions should be balanced. Specifically, numbers of correct and erroneous predictions should be similar, and the order of predictions should be random, such that testers cannot simply guess prediction results.

Human-understandable benchmark data. The data used in a human evaluation needs to be human understandable, otherwise testers have no clue how to assess predictions and explanations.

While a seemingly obvious statement, in practice we found it difficult to find KGC data that satisfies this constraint. In addition, testers recruited for a human evaluation are often lay people, not professionals of an area, thus plain datasets without domain-specific knowledge (such as biology and healthcare) would be better. If the evaluated XKGCs are domain specific, e.g., disease diagnosis, then specialists should accordingly be employed.

6 Experimental Study

To demonstrate what results and findings can be acquired with the proposed system, we conducted two evaluations.

6.1 Interview Users for Needs on XKGC

XAI is human-centric in nature. There is no one-for-all solution to meet all users' expectations. Our human-centric evaluation platform can help the researchers and practitioners interview their users to find: (1) what the users really need for understanding the KGC predictions in their applications, and (2) whether the generated explanations by their methods make sense for their users.

We conducted a series of interviews with the evaluation system. A human-understandable KGC dataset was selected as benchmark data. We used the kinship dataset (Kok and Domingos, 2007) because it is easily human understandable. Although the dataset is of small size, it involves key challenges of knowledge graphs, such as multiple relations and 1:n relations between entities. We randomly selected a set of KGC predictions and explained them with an XKGC method (Lawrence et al., 2021), denoted as *Method A*. Figure 1 illustrated an example prediction and its explanations.

With the evaluation system, we visualized the predictions and their explanations to the testers and interviewed: *what will be a helpful explanations for them?* and *why do they think an explanation helpful?* The interview is summarized in Table 1. Based on the collected feedback in the interview, we have made the following significant findings.

First, the interviews revealed that the testers often search for "paths" that link the nodes of the predicted triple to the nodes of explanations. See for example the "triangle" explanation in left panel of Figure 6, where two triples as the explanations can connect the two nodes of the predictions with another node in a triangle relationship. In situations where explanations don't connect to the predicted triples users consider the explanations are nonsensical for them.

Purpose	Investigate needs of humans on explanations of AI-driven predictions in the context of knowledge graph completion.
Interviewees	5 interviewees: 3 with machine learning background, 2 with good understanding about users of their AI system.
Guide	A guide is created, including text- and video-introduction to the evaluation platform.
Questions	1. What will be a helpful explanations for users? 2. Why do users think an explanation helpful?

Table 1: User interview for their needs on XKGC.

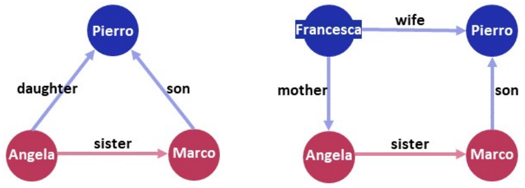


Figure 6: Explanations (in blue) learned with *Method B* for a predicted triple (in red): (a) explanation path of length $\ell = 2$, and (b) length $\ell = 3$. The path based explanations are more meaningful for human users because they create a connection between the entities in the predicted triple.

Second, testers often find a rather small set of explanations helpful (2-3) and remark that a large number of explanations (e.g. >10) create confusion.

Third, often it would be helpful for testers to have additional information from the knowledge graph - but this additional information was not identified by *Method A*. For example, *Method A* cannot create an explanation linking four entities, such as in right panel of Figure 6.

6.2 Compare Plausibility of XKGCs

We also used the evaluation platform to compare two XKGC methods: which would be more helpful for users. The kinship dataset (Kok and Domingos, 2007) is selected again due to human understandability for lay testers. Figure 1 and Figure 6 illustrate the explanations of the two methods, *method A* and *method B*, respectively. In order to mitigate potential biases introduced by individual testers, we select the predicted triples based on the guidelines in Section 5. The details are presented in Table 2.

1	Each tester evaluates 14 predicted triples to keep their engagement.
2	The first two triples serve as practice to facilitate testers understanding and comfort with the system and the questions. The feedback is not included in statistical analysis.
3	The rest of the triples are different from each other. Each is randomly drawn from a unique relation (12 relation types in total in the dataset).
4	Half of triples are correctly or incorrectly predicted to avoid dummy feedback.
5	Paired test is employed. Half of triples are randomly selected for either XAI method.
6	The predicted triples are randomly shuffled.
7	All testers evaluate the same set of predicted triples in the same order for fairness.

Table 2: Selecting predictions for a human-centric evaluation with the Kinship data.

30 testers are invited to evaluate the predictions, following the steps illustrated in Section 2. We received the feedback from 23 of them. For each tester (anonymous) and each prediction, our platform collected the metrics: accuracy of assessment (denoted as Acc), confidence of assessment, number of helpful explanations (denoted as helpExpl), and time cost. Our platform also provides diverse statistical tools (see Section 4) to analyze the measurements, e.g. the results shown in the bottom panel of Figure 2. One can find that *Method B* outperforms *Method A* in all four metrics. Most notably, *Method B* is attributed more helpful explanations (1.64 vs. 0.55) and leads to enhanced accuracy in testers' assessments (35% vs. 83%). From this we conclude that *Method B* indeed generates more helpful explanations for human testers in the context of kinship predictions.

7 Related Work

Human evaluation has attracted increasing attention in XAI research due to its ultimate goal of aiding human to understand AI predictions. Many evaluation benchmarks are based on simulatability (Doshi-Velez and Kim, 2017): how well human can simulate AI with help of explanations. For instance, Nguyen (2018) employed forward simulation to evaluate attribute-based XAI methods for text classification. Hase and Bansal (2020) extended the simulation test with counterfactual simulation to compare different types of explanations for text and tabular data. Arora et al. (2021) executed in-depth analysis of simulation tests for explanations of review classification. In addition, there are other human evaluations for XAI outside of NLP. For example, Alufaisan et al. (2021) proposed a decision-making based evaluation to measure human performance on decisions given predictions and explanations. More human evaluation tests can be found in the surveys e.g. Zhou et al. (2021). However the literature lacks a human evaluation tool to facilitate researchers on human-centric evaluation of KGC explanations.

Existing KGC evaluation platforms focus on measurement of prediction performance. For instance, Zhou et al. (2022) proposed a reconsideration of the used metrics by creating a "complete" judgement set inspired by evaluation of information retrieval. Rim et al. (2021) proposed the use of unit tests in order to evaluate models in a fine-grained manner by considering different capabilities. Widjaja et al. (2022) provided refined performance evaluation by bucketizing the test set into user-specified chunks. To bridge the gap, our platform provides a tool to measure plausibility of KGC explanations with human evaluation.

8 Conclusion

AI explanations only achieve their goal if the explanation is helpful to the human user. To measure this, we present a human-centric evaluation platform in the context of explainable knowledge graph completion. Distinguishing from the simulatability-based evaluation, our system assesses how well explanations assist users in judging the correctness of KGC predictions, and thus aligns better with human-AI interaction systems, where AI facilities humans rather than the other way around. To alleviate possible biases, we provide a set of guidelines in experiment design, and diverse analysis tools for reliable conclusions. The experiments demonstrate the findings and results that can be acquired with the proposed system.

References

Yasmeen Alufaisan, Laura R. Marusich, Jonathan Z. Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making? In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6618–6626.

Siddhant Arora, Danish Pruthi, Norman M. Sadeh, William W. Cohen, Zachary C. Lipton, and Graham

Neubig. 2021. Explain, edit, and understand: Rethinking user study design for evaluating model explanations. *CoRR*, abs/2112.09669.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1).

Patrick Betz, Christian Meilicke, and Heiner Stuckenschmidt. 2022. Adversarial explanations for knowledge graph embeddings. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI-22)*.

Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2).

Arjun Chandrasekaran, Viraj Prabhu, Deshraj Yadav, Prithvijit Chattopadhyay, and Devi Parikh. 2018. Do explanations make vqa models more predictable to a human? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1036–1042.

Elizabeth Clark, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7282–7296, Online. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Marie Destandau and Jean-Daniel Fekete. 2021. The missing path: Analysing incompleteness in knowledge graphs. *Information Visualization*, 20(1):66–82.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint*, arXiv:1702.08608.

Krzysztof Z. Gajos and Lena Mamykina. 2022. Do people engage cognitively with ai? impact of ai assistance on incidental learning. In *Proceedings of the 27th Annual Conference on Intelligent User Interfaces*, pages 794–806.

Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling, and Christian Meilicke. 2020. On Aligning OpenIE Extractions with Knowledge Bases: A Case Study. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 143–154, Online. Association for Computational Linguistics.

Henry Han and Xiangrong Liu. 2022. The challenges of explainable ai in biomedical data science. *BMC Bioinformatics*, 22.

Peter Hase and Mohit Bansal. 2020. Evaluating explainable ai: Which algorithmic explanations help users predict model behavior? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5540–5552.

Robert V. Hogg, Elliot A. Tanis, and Dale L. Zimmerman. 2015. *Probability and Statistical Inference*. Pearson.

Linmei Hu, Tianchi Yang, Luhao Zhang, Wanjun Zhong, Duyu Tang, Chuan Shi, Nan Duan, and Ming Zhou. 2021. Compare to the knowledge: Graph neural fake news detection with external knowledge. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 754–763, Online. Association for Computational Linguistics.

Ziniu Hu, Yichong Xu, Wenhao Yu, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Kai-Wei Chang, and Yizhou Sun. 2022. Empowering language models with knowledge graph reasoning for open-domain question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9562–9581, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, page 4198–4205.

Jiho Kim, Sungjin Park, Yeonsu Kwon, Yohan Jo, James Thorne, and Edward Choi. 2023. FactKG: Fact verification via reasoning on knowledge graphs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16190–16206, Toronto, Canada. Association for Computational Linguistics.

Stanley Kok and Pedro Domingos. 2007. Statistical predicate invention. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J. Gershman, and Finale Doshi-Velez. 2019. Human evaluation of models built for interpretability. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 7(1):59–67.

Carolin Lawrence, Timo Sztyler, and Mathias Niepert. 2021. Explaining neural matrix factorization with gradient rollback. In *35th AAAI Conference on Artificial Intelligence*.

Yankai Lin, Xu Han, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2018. Knowledge representation learning: A quantitative review.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification.

In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078, New Orleans, Louisiana. Association for Computational Linguistics.

A. J. Onwuegbuzie and N. L. Leech. 2004. Post hoc power: A concept whose time has come. *Understanding Statistics*, 3(4):201–230.

Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating robustness and interpretability of link prediction via adversarial modifications. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 3336–3347.

Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert, and Naoaki Okazaki. 2021. Behavioral testing of knowledge graph embedding models for link prediction. In *3rd Conference on Automated Knowledge Base Construction*.

Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Trans. Knowl. Discov. Data*, 15(2).

Apoorv Saxena, Adrian Kochsiek, and Rainer Gemulla. 2022. Sequence-to-Sequence Knowledge Graph Completion and Question Answering. *ArXiv*, abs/2203.10321.

Saurabh Srivastava, Mayur Patidar, Sudip Chowdhury, Puneet Agarwal, Indrajit Bhattacharya, and Gautam Shroff. 2021. Complex question answering on knowledge graphs using machine translation and multi-task learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3428–3439, Online. Association for Computational Linguistics.

Haris Widjaja, Kiril Gashteovski, Wiem Ben Rim, Pengfei Liu, Christopher Malon, Daniel Ruffinelli, Carolin (Haas) Lawrence, and Graham Neubig. 2022. KGxBoard: Explainable and Interactive Leaderboard for Evaluation of Knowledge Graph Completion Models. In *Conference on Empirical Methods in Natural Language Processing*.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 184–198.

Jianlong Zhou, Amir H. Gandomi, Fang Chen, and Andreas Holzinger. 2021. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5).

Ying Zhou, Xuanang Chen, Ben He, Zheng Ye, and Le Sun. 2022. Re-thinking knowledge graph completion evaluation from an information retrieval perspective. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.