Generating and Evaluating Plausible Explanations for Knowledge Graph Completion

Antonio Di Mauro[1,3*], **Zhao Xu**[1], **Wiem Ben Rim**[1,2], **Timo Sztyler**[1], **Carolin Lawrence**[1]
[1]NEC Laboratories Europe, Germany; [2]University College London, UK
[1]firstname.lastname@neclab.eu, [2]wiem.rim.23@ucl.ac.uk, [3]antonio.dimauro@niuma.it

Abstract

Explanations for AI should aid human users, yet this ultimate goal remains under-explored. This paper aims to bridge this gap by investigating the specific explanatory needs of human users in the context of Knowledge Graph Completion (KGC) systems. In contrast to the prevailing approaches that primarily focus on mathematical theories, we recognize the potential limitations of explanations that may end up being overly complex or nonsensical for users. Through in-depth user interviews, we gain valuable insights into the types of KGC explanations users seek. Building upon these insights, we introduce GradPath,[1] a novel path-based explanation method designed to meet human-centric explainability constraints and enhance plausibility. Additionally, GradPath harnesses the gradients of the trained KGC model to maintain a certain level of faithfulness. We verify the effectiveness of GradPath through well-designed human-centric evaluations. The results confirm that our method provides explanations that users consider more plausible than previous ones.

1 Introduction

Explainability is an essential requirement of AI, especially in high-risk areas, such as healthcare, which directly influence the life and health of humans. Only if users understand why an AI system arrives at a particular result can they trust the prediction and engage with the AI-driven system. This makes eXplainable AI (XAI) an essential ingredient for the adoption of AI in high-risk areas (Han and Liu, 2022; Chaddad et al., 2023). However, most existing XAI approaches aim to learn algorithmic explanations, which are supported by mathematical theories (like other ML methods), but are often difficult for users to understand, especially for users without an AI background. Not only do such implausible explanations not improve user trust in AI, in a contrary manner, they amplify users' doubt and hesitation.

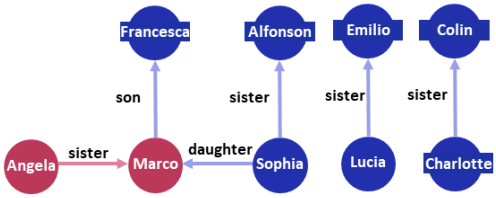


Figure 1: Explain a predicted triple (in red) with the top 5 most influential training triples (in blue), learned by Gradient Rollback (Lawrence et al., 2021). The algorithmic explanations are not plausible for users.

With this concern, we focus to improve plausibility of explanations, which refers to how convincing explanations are for human users (Jacovi and Goldberg, 2020). For the prediction task with knowledge graph data, a.k.a. Knowledge Graph Completion (KGC), the concept of plausibility is especially tricky, because non-expert users are not familiar with graph structures and the existing explanation methods do not aim to generate graph explanations in a human understandable manner (Betz et al., 2022; Lawrence et al., 2021; Pezeshkpour et al., 2019).

Fig. 1 illustrates the challenge with a KGC prediction (in red) and its explanations (in blue) generated by Gradient Rollback (GR) (Lawrence et al., 2021). In particular, the predicted triple *(Angela, is_sister_of, Marco)* is explained with the top 5 most influential training triples that are selected as per gradients of the learned KGC model. Although the provided explanations are faithful subject to a bound on the approximation error (Lawrence et al., 2021), it is difficult for users to understand. For example, how can the training triple *(Charlotte, is_sister_of, Colin)* explain the prediction? It is not plausible for users.

To better understand what kind of explanations

[*]Research work conducted during internship at NEC Laboratories Europe.

[1]More information is available at: https://github.com/nec-research/gradpath.

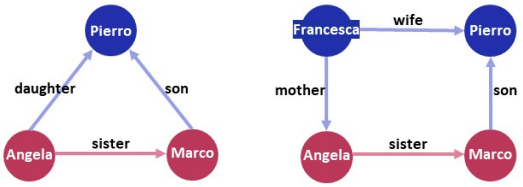
Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12106–12118

would be more plausible for users in the context of KGC, we started with a series of interviews. We provided users with predicted triples and GR-based explanations and interviewed them to understand their opinions about the explanations. This investigation led to an insightful finding: Interviewees remarked that the explanations linked to a path are the most plausible, see e.g. Fig. 2 (a). Here we learned that Angela is indeed the sister of Marco because Angela is the daughter of Pierre whereas Marco is the son of Pierro, therefore, explaining the prediction plausibly. The explanations are no longer scattered; instead, they form a connection between the entities in the predicted triples. This allows users to link them for reasoning.

Building upon this insight, we introduce GradPath, a novel method for generating human-centric explanations to enhance plausibility. The path based explanations intend to establish reasoning loops akin to human reasoning. Technically, the existence of the explanation path can increase the likelihood of the predicted triple, compared with the scenario wherein the path does not exist. To approximate the probabilities, GradPath utilizes gradients collected by GR during training, which ensures a certain level of faithfulness of the generated explanations. These explanation paths can vary in length, as illustrated in Fig. 2.

Evaluating plausibility is also challenging, as it inherently relies on human perception (Lyu et al., 2024; Wood-Doughty et al., 2021). There is not yet a standard way in the literature to evaluate plausibility of explanations (see also Appendix A.1). To draw solid conclusion for plausibility, we investigate diverse aspects of a human evaluation study, and introduce a comprehensive human-centric evaluation framework. We utilize human-understandable benchmark data and quantify plausibility with different metrics. The human evaluations validate that GradPath produces more plausible explanations for users in comparison to previous XAI KGC methods. The major contributions of the paper can be summarized as follows:

- We propose a novel method GradPath to improve plausibility of post-hoc explanations for KGC by extracting explanation paths.
- We suggest a human-centric evaluation framework to evaluate plausibility. It investigates diverse aspects of a human evaluation study for comprehensive assessment of plausibility.
- Experiments on the benchmark datasets demonstrate the proposed method is more plausible than previous ones.



(a) score = 0.2891 (b) score = 0.2497

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

2 Background and Notation

Assume a knowledge graph $\mathcal{G}$, consisting of entities and relation types. Two entities and a relation are combined to form a triple in the format (subject, relation, object), shortened as $t = (s, r, o)$. The subject and object are entities that can be visualized as the nodes of the graph $\mathcal{G}$, while the relation specifies a labeled link from the subject entity to the object one.

Typically knowledge graphs are incomplete. Therefore, Knowledge Graph Completion (KGC) is the task of predicting unknown triples in the knowledge graph. For this, two parts of a triple are given to a prediction system and the third is inferred. For example, $t_* = (s_*, r_*, ?)$ would predict the object given the particular subject and relation.

We consider differential KGC methods, such as DistMult (Yang et al., 2015) and Complex (Trouillon et al., 2016). They learn embedding vectors x_e for each entity e and x_r for each relation type r. Given the vectors, the likelihood of a triple is defined via a score function $\phi(x_s, x_r, x_o)$. During training, the vectors are optimized, such that the likelihood of a training triple is larger than that of a randomly sampled unknown triple. These methods can often well predict an object o_* for a test triple $t_* = (s_*, r_*, ?)$, but they are black boxes without explanations for users' trust.

The task of our work is to explain why the KGC model trained on the graph $\mathcal{G}$ arrives at a particular prediction. Existing work achieved this by focusing on purely mathematical aspects. For instance, Criage (Pezeshkpour et al., 2019) utilized influence functions, and Gradient Rollback (Lawrence et al., 2021) tracked gradients during training to produce explanations.

In contrast to previous work, which focused on algorithmic faithfulness, we place our focus on the human factor: extending algorithmic XAI to also respect human needs and produce explanations that are plausible for users. With this, we aim to combine the best of both worlds and therefore move KGCs closer to real-world applications with human-centric explanations to facilitate trust.

3 GradPath: Generate Plausible Explanations for KGC

We propose the method GradPath to learn human-centric explanations for KGC methods, such that the explanations are plausible for human users, and facilitate them to assess whether the KGC predictions are reasonable or not.

3.1 Initial Human-Centric Survey

We started with an initial survey to assess what individuals think about KGC predictions and explanations learned with the recent XAI method GR. We first used the dataset Kinship (Hinton, 1990), which is about familial relations and easy for testers to understand. We asked three testers[2] to assess whether they think the kinship predictions are true or false based on the explanations. We further interviewed them by the following questions: *What will be a helpful explanation? Why do you think an explanation is helpful?* Next, we repeated the interview using real-world user preference data collected from a recommender system.[3] We interviewed two testers[4] with the same questions. Upon the collected feedback, we concluded the following two significant findings.

First, the interviews revealed that the testers often searched for "paths" that link the nodes of the predicted triple to the nodes of explanations. See for example the "triangle" explanation in Fig. 2(a), 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 do not connect to the predicted triples (e.g., the scattered explanations in Fig. 1), the testers considered the explanations nonsensical.

Second, the testers often found a small set of explanations plausible (2-3) and remarked that a large number of explanations (e.g., > 10) create confusion.

3.2 Defining Path-based Explanations

Inspired by the findings from the interviews, we suggest extracting *influential paths* as explanations for a prediction, where the paths connect the two entities of the prediction. The explanation paths intend to emulate the sequential reasoning process humans use to deduce over a knowledge graph, therefore making the resulting explanations more human-centric, namely more plausible for humans. Formally, we define:

Definition 1. *An explanation is an influential path of length ℓ, denoted as $\gamma = \{t_1, \ldots, t_\ell \mid t_j \in \mathcal{G}\}$, with the constraints:*

1. *All t_j are selected from training triples;*
2. *There is a joint entity between two adjacent triples t_j and t_{j+1};*
3. *The subject s_* and object o_* of the test triple are entities of the start triple t_1 and the end triple t_ℓ, respectively;*
4. *Sequentially, existence of each current triple has the potential to increase the likelihood of the subsequent triple;*
5. *Ultimately, at the end of the path, the predictive probability of the test triple is improved when compared with the scenario wherein the path of triples does not exist.*

Thus, an influential path is a series of training triples that connect the subject and object entities of the predicted triple and make the predicted triple more likely. Such influential paths are the explanations that the testers searched for in our initial interviews, thereby motivating the concept of influential paths from a human point of view. Technically, we look for a path that increases the likelihood of the predicted triple, compared with the scenario wherein the path does not exist and therefore decreases the likelihood of the predicted triple. Fig. 3 illustrates the key idea. Note that the influential paths are not supposed to be directed. Regardless of the direction of a triple, it provides the testers with the information for reasoning. For instance,

[2]The testers have a good ML background but did not know this particular task.

[3]The data is about user preferences regarding the products of a company. It consists of the attributes of the products and users, alongside information of user preferences for specific products. We transformed the data into a knowledge graph. Due to user privacy and commercial confidentiality of the company providing the dataset, we are unable to publicly disclose the dataset.

[4]The two interviewees are the engineers of the recommendation system. They are closely familiar with the end users and possess a strong understanding of users needs.

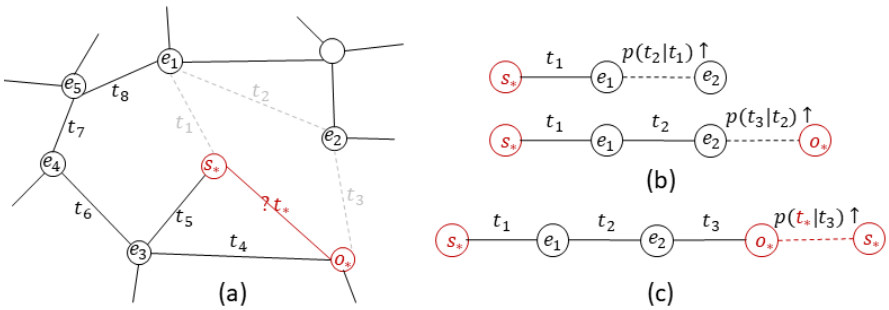


Figure 3: GradPath: (a) $\mathcal{G}' \equiv \mathcal{G} \setminus \gamma$ in which the path γ is removed from $\mathcal{G}$; (b) the likelihood of the subsequent triple increases if the current triple exists; (c) at the end of the path, the predictive probability of the test triple $t_* = (s_*, r_*, o_*)$ will be improved if the entire path exists.

the paths in Fig. 2 are not directed yet plausible. The testers can follow the path-based explanations to understand how a KGC method arrives at a prediction, and to judge the prediction by comparing the predictive procedure of KGC with the mental models in their minds.

3.3 Generating Path-based Explanations

Based on the formulation above, we now turn to how the influential paths can be generated. The paths that satisfy the constraints 1-3 can be readily computed with arbitrary graph traversal algorithms, such as Depth-First Search (DFS) and Breadth-First Search (BFS). However, the computation of the constraints 4 and 5 is tricky. Mathematically, they can be written as

$$p(t_{j+1}|\mathcal{G}' \cup \{t_j, \ldots, t_1\}) > p(t_{j+1}|\mathcal{G}'), \quad (1)$$

$$p(t_*|\mathcal{G}) > p(t_*|\mathcal{G}'), \quad (2)$$

where $\mathcal{G}' \equiv \mathcal{G} \setminus \gamma$ denotes the rest of $\mathcal{G}$ in which the path γ is removed, $p(t_{j+1}|\mathcal{G}' \cup \{t_j, \ldots, t_1\})$ is the predictive probability of the next triple t_{j+1} given $\mathcal{G}'$ and the triples preceding t_{j+1} in the path γ. The other terms in (1) and (2) are defined in a similar manner. Intuitively, each term computes the probability of a triple (the predicted triple t_* or a triple that is part of the path γ) under the simulation of *what if* certain triples from the path γ had not been part of $\mathcal{G}$ during training. We consider both the influence of the path γ on the predictive triple t_* (Constraint 5, Eq. (2)), and the influence of other triples on the next one (Constraint 4, Eq. (1)).

As the conditions of each prediction is different (i.e. DFS/BFS returns different paths), the probability needs to be computed for each prediction and explanation path individually. To be fully faithful, we would have to retrain a new KGC model for each '*what if*' analysis. But the computational cost for this is prohibitively expensive. To address the problem, we introduce an approximate approach, which is based on gradients of training triples (Lawrence et al., 2021). In our work, we take a step further to identity influential paths, instead of single training triples, for a test triple. Concretely, we approximate the '*what if*' analysis as follows.

First we train the differentiable KGC model with entire training triples $\mathcal{G}$, and during training employ GR to compute changes of embedding vectors caused by each training triple. Formally, a triple $t = (s, r, o)$ causes a change in the embedding vector x_s of its subject, which is approximated by

$$\delta(x_s|t) = \sum_{i=1}^{N} \eta_i \nabla_{x_s} \mathcal{L}^i(t), \quad (3)$$

where i denotes the i'th training iteration, and η_i is the learning rate. $\mathcal{L}^i(t)$ specifies the loss of the triple t at the i'th iteration, while $\nabla_{x_s}\mathcal{L}^i(t)$ means the derivative of the loss over x_s. The changes $\delta(x_r|t)$ and $\delta(x_o|t)$ of the object and relation embedding vectors x_o and x_r are computed in an equivalent manner. The approximation error of the approach has been linked to stability theory (Hardt et al., 2016) and bound by Lawrence et al. (2021).

Second, given the influence of each training triple t, we can now approximate what would happen if a path had been removed from the training data without retraining the model. By removing the influence of an explanation path, we can simulate the probability of any triple that we would have had if the triple had not been part of the training data. Based on this, we can approximate the terms of Eqs. (1) and (2). In particular, for the predictive probability $p(t_{j+1}|\mathcal{G}')$ of the triple $t_{j+1} = (s_{j+1}, r_{j+1}, o_{j+1})$, we approximate the new embedding vectors $\bar{x}_{s_{j+1}}$ and $\bar{x}_{o_{j+1}}$ conditioned on the training set $\mathcal{G}' \equiv \mathcal{G} \setminus \gamma$ by

$$\bar{x}_{s_{j+1}} = x_{s_{j+1}} + \delta(x_{s_{j+1}}|t_j) \quad (4)$$

$$\bar{x}_{o_{j+1}} = x_{o_{j+1}} + \delta(x_{o_{j+1}}|t_{j+2}) \quad (5)$$

and the vector of the relation r_{j+1} is updated by

$$\bar{x}_{r_{j+1}} = x_{r_{j+1}} + \sum_{i=1}^{\ell} \delta(x_{r_i}|t_i)\mathbf{1}(r_{j+1} = r_i), \quad (6)$$

where the indicator function $\mathbf{1}(x = y)$ is one if the condition $x = y$ is true and zero otherwise. Intuitively, this means that we update subject s_{j+1} by removing the influence of the connected triple t_j (analogous for the object o_{j+1}). The relation r_{j+1} is updated only if the same relation appears again in the explanation path.

After the influence of the path has been updated, the predictive probability $p(t_{j+1}|\mathcal{G}')$ can be computed with the score function $\phi(\bar{x}_{s_{j+1}}, \bar{x}_{r_{j+1}}, \bar{x}_{o_{j+1}})$ of the KGC method. The other predictive probabilities in Eqs. (1) and (2) can be calculated in a similar manner.

Finally, putting the constraints 4 and 5 together, we score an explanation path γ for a test triple t_*:

$$\frac{1}{L}\left[\sum_{j=2}^{L} p(t_j|\mathcal{G}' \cup \{t_{j-1}, \ldots, t_1\}) + p(t_*|\mathcal{G}) - \sum_{j=2}^{L} p(t_j|\mathcal{G}') - p(t_*|\mathcal{G}')\right]. \quad (7)$$

The score function (7) measures the importance of the candidate paths between the subject and object of the test triple t_*, and identifies the ones which most likely increase the likelihood of t_* as explanation paths.[5] Our method generates path-based explanations using predictive probabilities of the involved triples. Along the path, the existence of each current triple has the potential to increase the likelihood of the next triple. When the entire path exists, the likelihood of the test triple is ultimately improved.[6]

3.4 Faithfulness vs. Plausibility

Faithfulness and plausibility represent two pivotal principles in XAI. Faithfulness demands explanations accurately reflect the model's reasoning process, while plausibility necessitates alignment with human reasoning. Our GradPath method aims to find a balance between faithfulness and plausibility.

On one hand, GradPath leverages the gradients of training triples to identify influential paths. These gradients, collected during training using GR (Lawrence et al., 2021), reflect the true reasoning process of KGC models, thereby guaranteeing some level of faithfulness in the explanations. On the other hand, GradPath's path-based explanations are designed to mimic human reasoning over a knowledge graph, enhancing plausibility compared to explanations based solely on individual training triples (see Fig. 2 vs. Fig. 1). It is worth noting that this improved plausibility may come with a potential reduction in faithfulness.

The reason is as follows: Given the computational complexity of generating influential paths (see Sec. 3.3 for details), approximations are utilized. While these approximations stem from GR, the bounds of GR's approximation errors do not inherently apply to GradPath. GradPath removes a sequence of training triples around the predicted one, which may alter the embedding space of the predicted entities due to potentially accumulated approximation errors. We acknowledge the potential decrease in faithfulness, prioritizing explanations that are plausible for human users while maintaining a certain level of faithfulness.

4 Evaluation of Plausibility

GradPath is designed to improve plausibility of explanations, which necessarily leads to a question: *How do we evaluate plausibility?* To address this, we present a human-centric evaluation framework to assess plausibility in a rigorous manner.

Plausibility refers to how convincing explanations are for users (Jacovi and Goldberg, 2020). In analogy to, for instance, accuracy in classification, there is lack of solid and deterministic ground truth to automatically compute a single score as a measurement of plausibility. The measurement of plausibility is intrinsically based on human perception. Thus, we design a comprehensive human evaluation study to quantify plausibility.

4.1 Purpose Factor

In the human evaluation of XAI, a commonly used method is *human simulatability*, which measures the extent to which explanations can assist users in predicting the output of the model (Yin and Neubig, 2022; Hase and Bansal, 2020; Lage et al., 2019; Doshi-Velez and Kim, 2017). However, human simulatability is an artificial construct and does not gauge the actual use cases of explanations, such as what a user would do with an explanation in a real-world setting.

[5] Particularly it is the explanation path where its removal during the 'what if' analysis caused the largest drop in likelihood of t_* we want to explain - i.e. the explanation path had the largest contribution to raising the likelihood of t_*.

[6] If the new likelihood is instead lowered, then the explanation path could be served as an 'anti-explanation'.

Here, we aim to shift the focus to the *purpose factor* of human users with XAI, which is the specific objective why explanations are needed by users. Imagine that a doctor interacts with an AI-driven diagnosis system. The doctor wants to know if an AI prediction is true or false, such that she can use it for her task of prescribing treatments for a patient. Therefore the role of the explanations is to help the doctor to solve her question: *Is the AI-driven prediction correct?* If the explanations match the mental model of the doctor and/or the doctor can easily follow the logical reasoning of the explanations, then she likely trusts the prediction. Considering this ultimate goal, evaluations where humans simulate AI behavior do not align with the intended purpose of the users.

We therefore investigate in the human evaluation: *Can testers assess correctness of predicted triples with help of explanations?* In particular, we ask the following questions to each tester:

- Is the predicted triple correct? (Two scales, see below)
- How confident is the assessment? (Scale: 5 point Likert)
- Is an explanation helpful? (Scale: yes and no)

For the first question, we employ two distinct scales. The first scale comprises three options: yes, no, and not-sure, used for in-house evaluations conducted by testers possessing a ML background or with access to support in case of queries. The second scale offers a more detailed range of 5 options (shown in Fig. 4), used for crowdsourcing evaluations to mitigate potential misunderstanding. The testers' feedback about the question is processed as follows. If a tester's judgement coincide with the ground truth then the feedback is recorded as 1 otherwise 0. For the feedback of not-sure, it is recorded as 0.

The suggested human evaluation method is related to the decision-making-based approach, such as the one proposed by Alufaisan et al. (2021), but differs slightly. While the decision-making-based evaluation focuses on the success of human users in making correct predictions based on explanations, our method shifts to a more purpose-driven evaluation. It asks testers to judge the correctness of predictions by comparing the explanations with their prior knowledge and reasoning.

4.2 Human-Related Concerns

Human evaluation collects subjective ratings from individual testers to assess the plausibility of the explanations. To have reliable feedback, the following concerns need to be addressed (Hase and Bansal, 2020; Gajos and Mamykina, 2022).

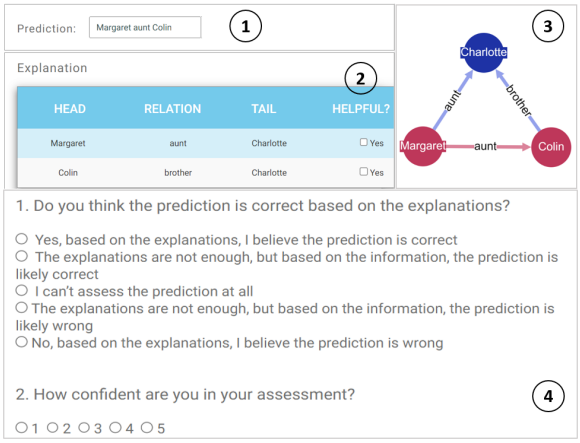


Figure 4: The online evaluation platform developed to collect feedback from testers regarding the explanations. (1) Displays the prediction. (2) Shows the explanations and prompts testers to evaluate their helpfulness. (3) Visualizes the prediction and explanations in a graph format for easier comprehension. (4) Asks testers to assess the correctness of the prediction and their confidence in it (refer to Appendix A.5 and A.6 for more information).

Balance: Predictions are balanced. Namely, the correctly and erroneously predicted triples should be of similar number, and randomly ordered, such that testers cannot simply guess their correctness.

Diversity: Testers may remember information or knowledge of previous test triples, which means the earlier predictive triples judged by a tester may influence their judgment on later ones. Thus, we propose that the predictions should be distinct from each other.

Furthermore, **confounders** need to be investigated, which are other factors potentially influencing user feedback besides the XAI method itself. In addition, **human-understandability of benchmark data** used in an evaluation is also essential. The details are reported in Appendix A.3 and A.4.

4.3 Quantification of Plausibility

To quantify the plausibility of explanations, we suggest the following metrics, based on Zhou et al. (2021) and interviews with testers in the initial survey:

- Accuracy rate of assessment
- Number of helpful explanations for a triple
- Time cost for a tester to assess a prediction
- Confidence level of assessment

These metrics are abbreviated as Acc, helpExpl, Time, and Confidence, respectively, in the experi-

ments. The *accuracy rate* metric appears to be a good measure of plausibility since it is based on well-defined ground truth (the correctness of predictions). This metric could be less influenced by the diversity of testers (e.g., fast/slow thinking modes). The empirical analysis in Sec. 5 also demonstrates its effectiveness.

Regarding the metric *number of helpful explanations*,[7] there are two divergent opinions. On one side, more explanations may not be necessary if a single explanation explicitly explains a prediction. On the other side, more explanations are always beneficial as predictions can be explained from different perspectives, and users have different needs and interests. There would not be one single perfect explanation that satisfies everyone. We believe that combining this metric with others can provide more insights. For example, together with the accuracy rate of assessment, we can demonstrate if more explanations may improve users' assessments.

Another disputable metric is the time cost of assessment. Ideally, testers use less time for assessment if explanations are plausible. However, in practice, we found after interviewing the testers that they often skip a prediction when the explanations (e.g., the explanation in Fig. 1) are meaningless. Conversely, only when the explanations (e.g., Fig. 2) appear reasonable do they explore them carefully. Therefore, both too little and too much time can indicate bad explanations, whereas good explanations occupy the space in between.

To assess whether the observed differences in feedback are genuine or merely due to random variation, we employ the non-parametric significance test Brunner-Munzel (Brunner and Munzel, 2000). This test is preferred over the commonly used t-test, as it does not assume normality of the data (see Appendix A.2 for further discussion).

5 Experiment Settings and Results

Following the human-centric evaluation framework outlined in Sec. 4, we investigated plausibility of GradPath with two human-understandable benchmark datasets: Kinship (Hinton, 1990) and CoDEx (Safavi and Koutra, 2020) (see details of the datasets in Appendix A.4). An online evaluation platform was developed to collect feedback from testers, as illustrated in Fig. 4. We conducted a comparison between our human-centric method GradPath and the algorithmic explanation method Gradient Rollback (GR) (Lawrence et al., 2021) to investigate whether GradPath can offer more plausible explanations.

5.1 Experiment Settings

Kinship. We first utilized the kinship dataset (Hinton, 1990) because of its easy human understandability. Although the dataset is small in size, it presents key challenges found in commonly used knowledge graphs, such as multiple relations and 1:n relations between entities. We employed Complex (Trouillon et al., 2016) as the knowledge graph completion (KGC) method to be explained due to its popularity and strong performance. The parameters were set as follows: embedding dimension is 10, learning rate is 0.001, number of negative samples is 13, batch size is 1, epochs is 100, and optimizer is Adam. Gradient Rollback (GR) presented the top 5 important training triples as explanations. Our method utilized the top-1 path-based explanation (ranked by our score), where a path is of length less than four. Five-fold cross-validation was employed to compute predictions and explanations for all triples. We invited 30 coworkers[8] for in-house evaluation and received feedback from 23 of them. In addition, we recruited 105 lay testers[9] via Amazon Mechanical Turk (AMT) for crowdsourcing evaluation. The details of the settings are discussed in Appendix A.6 and A.7. For the question "Is the predicted triple correct?" we utilized a scale of 3 options for the in-house evaluation and a scale of 5 options for the crowdsourcing evaluation, as detailed in Sec. 4.1.

We conducted both in-house and crowdsourcing evaluations for the following benefits: (1) The coworkers intended to be well-engaged, providing high-quality feedback and can be further interviewed for detailed comments. (2) The crowdsourcing evaluation helped us obtain feedback from general users without a background in ML. The results shed light on their understanding of explanations. (3) By comparing the two evaluations, we could investigate the influence of tester diversity and explore the reliability of the metrics.

CoDEx. We further evaluated GradPath with CoDEx (Safavi and Koutra, 2020), which is a more complex and recent KGC dataset extracted from

[7]The metric *number of helpful explanations* (shortened as helpExpl) is collected based on the tester's answer to the question, "Is an explanation helpful?" Fig. 4 illustrates how the evaluation platform collects feedback from the testers.

[8]They have a good background in machine learning, but are not familiar with XAI and know nothing about our work.

[9]Each tester is an individual without ML background.

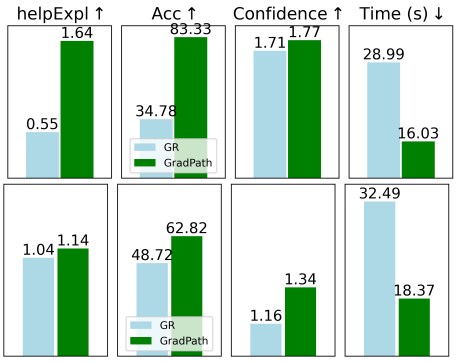


Figure 5: Human evaluation results of the Kinship data with in-house testers possessing a good ML background (top) and crowdsourced lay testers (bottom). GradPath provided more plausible explanations in both cases for testers to assess the correctness of the predictions.

Wikidata and Wikipedia. We employed DistMult (Yang et al., 2015) as another popular and effective KGC method to be explained. The parameters were set as follows: embedding dimension is 1024, learning rate is exponential decay with an initial value of 0.3, number of negative samples is 20, batch size is 1, epochs is 500, and optimizer is Adam. GR reported the top 10 training triples as explanations. Our method selected the top N paths as explanations, where N varied dynamically to ensure that the cumulative number of training triples is not larger than 10. Here, we followed Miller's law, which suggests that people can absorb at most 9 pieces of information (Miller, 1956). We recruited 105 lay testers via Amazon Mechanical Turk (AMT) (details in Appendix A.6). More information about the settings of the human evaluation can be found in Appendix A.8.

5.2 Results with Kinship Data

Fig. 5 presented the Kinship results regarding the metrics defined in Sec. 4.3. GradPath outperformed GR in all four metrics in both in-house and crowdsourcing evaluations. The in-house evaluation revealed GradPath provided more plausible explanations (helpExpl: 1.64 vs. 0.55), allowing testers to more accurately assess the correctness of predictions (Acc: 83% vs. 35%).[10] The crowdsourcing AMT evaluation reported similar trends. Interestingly, the in-house testers used less time than the crowdsourced lay testers. It might be because of good ML background of the in-house testers. The two human evaluations demonstrated that GradPath prioritizes user needs in its design, resulting in explanations that are more plausible for users to assess the correctness of the predicted triples.

[10]To gain further insights into the plausibility of the generated explanations, we analyzed testers' assessments in more details (see Appendix A.9).

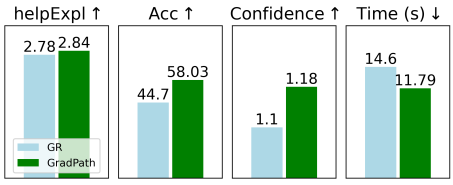


Figure 6: Human evaluation results of the CoDEx data with crowdsourced lay testers. GradPath provided more plausible explanations for testers to assess the correctness of the predictions.

To demonstrate the observed differences are statistically significant and not due to chance, we conducted the Brunner-Munzel (BM) test and reported the results in Table 1. Overall, the differences between GradPath and GR were significant in both in-house and crowdsourcing evaluations. The p-values for the metrics Acc, time, and helpExpl were much lower than the threshold $\alpha = 0.05$ in both evaluations, revealing strong statistical significance of the differences between GradPath and GR.

The results also indicated that the testers were confident, with a confidence metric of 1.71 for GR and 1.77 for GradPath (the upper bound being 2), showing no significant differences (p-values above 0.4). Upon analyzing the testers' responses, we observed that they primarily evaluated predictions when confident in their assessments; otherwise, they responded with "*not sure*" to the question "*Is the predicted triple correct?*"

5.3 Results with CoDEx Data

The results for CoDEx are reported in Fig. 6. Again, GradPath provided more plausible explanations for testers to assess the correctness of the predictions. With the path-based explanations generated by GradPath, the testers can judge the correctness of the predicted triples more accurately (58.0% vs. 44.7%) with less time (11.79 vs. 14.6 sec.)

To investigate the significance of the observed differences, the results of the BM test are reported in Table 1. The p-values 0.0000 for Acc and 0.0391 for helpExpl showed strong statistical significance of the differences between GradPath and GR. The metric Time is not statistically significant for this dataset, which could be due to the more complex nature of the dataset.

The significance test also revealed that the metric Acc, i.e. Accuracy rate of users assessing the correctness of predictions, can be a good measurement for human evaluation of plausibility. As discussed

	helpExpl	Acc	Confidence	Time
Kinship In-House	0.0000	0.0000	0.4346	0.0000
Kinship AMT	0.0299	0.0121	0.4335	0.0007
CoDEx AMT	0.0391	0.0000	0.9189	0.1648

Table 1: P-values of the BM test: the smaller the value, the more significant the difference between the human feedback on the two explanation methods.

in Sec. 4.3, the metric Acc is based on well-defined ground truth, correctness of predictions, and thus can be more resistant against the diversity of the testers (e.g. fast or slow thinking mode). Moreover, the experimental results also show the metric confidence might not be a good measurement in the current question design. As there is an option of "*not sure*" for the question of "*Is the predicted triple correct?*", the testers intended to assess when they are confident of the answer. For future studies, we would recommend excluding the "*not sure*" option.

Overall, the human evaluation results lead us to conclude that GradPath can help testers more accurately assess the correctness of predictions.

6 Conclusion

Explanations for AI should be plausible for humans. To this end, we propose a novel explanation method, GradPath, to improve plausibility for KGC. Moreover, a comprehensive human-centric evaluation framework is introduced to evaluate plausibility reliably. We evaluate our GradPath method with three human evaluations using two benchmark datasets. The evaluation results confirm that GradPath indeed provides more plausible explanations than previous methods. One future direction can be to collect a new KGC dataset with the goal of exploring XAI KGC for users, as we realize in the human evaluations that datasets should exhibit a high level of human understandability to enable meaningful human evaluations, which is an area where existing KGC datasets may have limitations (see Sec. 7 and Appendix A.4).

7 Limitations

Explanation methods involve a trade-off between faithfulness and plausibility. Faithfulness refers to the degree to which an explanation method truly represents the reasoning process of an AI model. Plausibility considers how well a human user can understand and use an explanation. There is a natural trade-off because a fully faithful explanation is not human-understandable (e.g., providing the weights and architecture of a neural network as an explanation is fully faithful, but humans cannot understand it). Here, we specifically focus on increasing plausibility. Therefore, our method may lose some faithfulness compared to other KGC explanation methods. Users of our method for real-world applications should be aware of this and determine whether this is an acceptable trade-off based on the given use case.

In our current human evaluation studies, we have employed two benchmark datasets to assess the performance of the proposed method. We acknowledge the importance of experimenting with a broader range of datasets. However, our research questions impose conditions on the suitability of these datasets. Specifically, KGC benchmarks used in human evaluations need to be understandable to humans. Without this attribute, testers are unable to effectively assess predictions and explanations, which unfortunately excludes the majority of currently available KGC benchmarks. The two critical criteria that need to be fulfilled are: (1) Entities and relations must be readily understandable to humans. If the entities or relations are only numerical values or lack a concise definition, humans cannot judge whether a prediction is correct or an explanation is meaningful. (2) Relations among entities should allow human testers to engage in logical deductions, facilitating reasoning. For example, the relation types within the Kinship dataset include strong logical connections, enabling testers to infer kinship between entities based on the provided training triples. Without such logical connections, there might not be any suitable explanation that is understandable to humans for that particular knowledge graph. Based on these criteria, our review of popular benchmarks revealed only two suitable KGC benchmarks, both of which we employed in this study. To generalize our results, it would be important to extend our evaluation to additional human-centric KGC benchmarks.

8 Ethics Statement

In light of the limitations mentioned above, users of our explanation method should be aware that the method may not be fully faithful due to approximations, especially when the explanation path is long. Therefore, a per-use-case estimation is required to determine if the approximation is acceptable. All

interactions with human testers and data usage comply with the General Data Protection Regulation (GDPR) of the European Union.

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.

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*, pages 2820–2826.

Edgar Brunner and Ullrich Munzel. 2000. The nonparametric Behrens-Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, 42(1):17–25.

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

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. 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.

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

Moritz Hardt, Benjamin Recht, and Yoram Singer. 2016. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning*, pages 1225–1234.

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.

Geoff Hinton. 1990. *Kinship data*. UCI Machine Learning Repository.

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

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*, page 4198–4205.

Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2019. An evaluation of the human-interpretability of explanation. ArXiv:1902.00006.

Carolin Lawrence, Timo Sztyler, and Mathias Niepert. 2021. Explaining neural matrix factorization with gradient rollback. In *35th AAAI Conference on Artificial Intelligence*, pages 4987–4995.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, pages 1–70.

G. A. Miller. 1956. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81–97.

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*, pages 3336–3347.

P. Jonathon Phillips, Carina Hahn, Peter Fontana, Amy Yates, Kristen K. Greene, David Broniatowski, and Mark A. Przybocki. 2021. *Four Principles of Explainable Artificial Intelligence*. NIST Interagency/Internal Report, National Institute of Standards and Technology.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you? Explaining the predictions of any classifier. ArXiv:1602.04938.

Tara Safavi and Danai Koutra. 2020. CoDEx: A comprehensive knowledge graph completion benchmark. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 8328–8350.

Kacper Sokol and Peter Flach. 2020. Explainability fact sheets: A framework for systematic assessment of explainable approaches. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 56–67.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

Zach Wood-Doughty, Isabel Cachola, and Mark Dredze. 2021. Faithful and plausible explanations of medical code predictions. ArXiv:2104.07894.

Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. ArXiv:1412.6575.

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

A Appendix

A.1 Discussion of Faithfulness and Plausibility

The XAI community has introduced and explored desired properties of explanations for AI. For example, Lage et al. (2019) introduced human-interpretability. Jacovi and Goldberg (2020) highlighted faithfulness. (Sokol and Flach, 2020) suggested fact sheets to characterise XAI systems. Phillips et al. (2021) proposed XAI principles, including accuracy, meaningfulness and knowledge limits. Although there is not yet a consensus regarding the desired properties of XAI, the two aspects, faithfulness and plausibility, have attracted significant attention in the literature. Faithfulness refers to how accurately explanations reflect a model's prediction process (Ribeiro et al., 2016; Jacovi and Goldberg, 2020). Plausibility is how convincing explanations are for human users (Jacovi and Goldberg, 2020).

From the descriptions, it is clear that faithfulness demands correct reflection of the model's reasoning process, while plausibility needs to align with human reasoning. Faithfulness does not guarantee plausibility, and vice versa (Lyu et al., 2024; Jacovi and Goldberg, 2020). Especially, when the model's reasoning process deviates from human's, XAI approaches must navigate a trade-off between faithfulness and plausibility.

A.2 Significance Testing for Human Evaluation

Given that human evaluations often recruit a limited number of testers for each setting due to high costs, it becomes imperative to use significance testing to demonstrate that observed differences attribute to fixed effects rather than random variations. The commonly used significance tests include the t-test and its variants, the Wilcoxon test, the Mann-Whitney U test, and the Brunner-Munzel test (Hogg et al., 2015; Brunner and Munzel, 2000). When choosing statistical tests for human feedback data, it is crucial to consider the limitations and assumptions inherent in these tests to derive robust conclusions. For instance, the assumption of normally distributed data is standard in various t-tests, but this assumption may not hold in Likert scale data. Additionally, the two-sample independent t-test assumes equal population variances, whereas Welch's t-test does not require such equality. Therefore, it is essential to evaluate the applicability of testing methods to ensure the significance of observed differences in the data. We selected the Brunner-Munzel test in our experiments because it does not assume normality or equal population variances, making it robust and applicable to a broader range of conditions.

A.3 Confounders

Besides the XAI methods themselves, other factors can potentially influence human evaluation. First, the engagement of testers will impact the evaluation. Some testers may be highly engaged and think slowly, while others might not exhibit the same level of attentiveness. During experiments, we found that checkpoint questions are an effective way to identify well-engaged testers, particularly when using crowdsourcing platforms like Amazon Mechanical Turk (AMT).

Moreover, the engagement level of a tester changes over time and often gradually decreases. Therefore, it is important to limit the number of predictions in the evaluation to maintain high tester engagement.

Additionally, some predictions and explanations can be more challenging, requiring testers to take more time for assessment. To minimize deviations caused by varying difficulty levels of predictions, all testers should evaluate a similar set of predictions in a similar order.

A.4 Human-Understandability of KGC Benchmark Data

The KGC benchmark data used in human evaluations must be understandable to testers. Otherwise, they would have no means to assess predictions and explanations. Additionally, testers recruited for human evaluations are often laypeople, not professionals in a particular field. Therefore, plain datasets without domain-specific knowledge (such as biology, chemistry, or healthcare) are preferred.

Unfortunately, many publicly available KGC datasets are not designed for human understanding. All datasets that we believe human understandable can be found in Table 2. Here, our focus primarily lies in ensuring that lay testers can understand the triples. However, after experimental analysis with the Kinship and CoDEx datasets, we realized that merely understanding individual triples is insuffi-

Dataset	Description	Entities	Relations	Triples	Example Triples
Kinship	Relations among family members	24	12	112	(Charlotte, niece, Arthur) (Christopher, father, Victoria)
Countries	Geographical relations of countries	271	2	1,159	(western_africa, locatedin, africa) (slovakia, neighbor, austria)
Movie-Lens	User-movies networks	2625	5	100,000	(User757, 3, Transformers) (User943, 1, Star Trek IV: The Voyage Home)
CoDEx	Relations from wiki	2034	42	36543	(David Evans, instrument, guitar) (Gustav Struve, citizenship, Germany)
YAGO3-10	Person relations from wiki	123,143	37	1,089,040	(Glencore, isLocatedIn, Rotterdam) (Ambareesh, isPoliticianOf, India)

Table 2: Datasets with human understandable triples for human evaluation of explainable KGC.

cient. Datasets should exhibit a high level of human understandability to enable meaningful human evaluations. To advance explainable KGC research, it is imperative to curate new KGC datasets in a human-centric manner (see also Sec. 7).

A.5 Human-centric Online Evaluation Platform

Following the human evaluation study outlined in Sec. 4, we developed an online system to evaluate XAI KGCs in a human centric manner. Our system considers the real needs and interests of human users in collaboration with AI, enabling us to investigate the following questions: *can testers assess the correctness of a KGC prediction based on its explanations? Which explanations are helpful for testers?* Fig. 4 illustrates how the evaluation platform collects feedback from testers. Our system works as an online website and can easily collaborate with crowdsourcing platforms such as AMT to recruit testers and distribute evaluation tasks.

A.6 Crowdworkers for Human Evaluation

To get feedback from general users about their evaluations on explanations, we recruited crowdworkers from AMT. Fig. 7 presents the instructions to the testers. Privacy and data usage strictly comply with the General Data Protection Regulation (GDPR) of the European Union, and testers are informed before they participate in the evaluation.

A.7 Settings of Kinship Evaluation

Based on the evaluation framework designed in Sec. 4, the settings of the test with the Kinship data are presented in Table 3. Among the randomly selected test predictions, there are two simple predictions, such as a prediction *(Colin, is_son_of, James)* associated with an explanation *(James, is_father_of, Colin)*, which serve as checkpoints to gauge tester engagement. After excluding testers who simply rejected or accepted all predictions or answered the checkpoint questions incorrectly, there were 26 well-engaged crowdsourced testers remained.

1	Each tester evaluates 14 predicted triples to maintain good engagement.
2	The first two triples serve as practice to help testers understand the system and the questions. The feedback is not included in statistical analysis.
3	Half of the triples are correctly or incorrectly predicted to prevent dummy feedback.
4	Half of the triples are randomly selected for either XAI method.
5	The predicted triples are randomly shuffled.
6	The same set of predicted triples is presented to testers in the same order.

Table 3: Settings of human-centric evaluation for the Kinship data.

A.8 Settings of CoDEx Evaluation

During the test, sensitive relation types related to the person entities, such as *medical conditions, religion, and ethnic group*, are removed from the data to comply with the General Data Protection Regulation (GDPR) of the European Union. The settings of the human evaluation with the CoDEx data are presented in Table 4. To investigate the

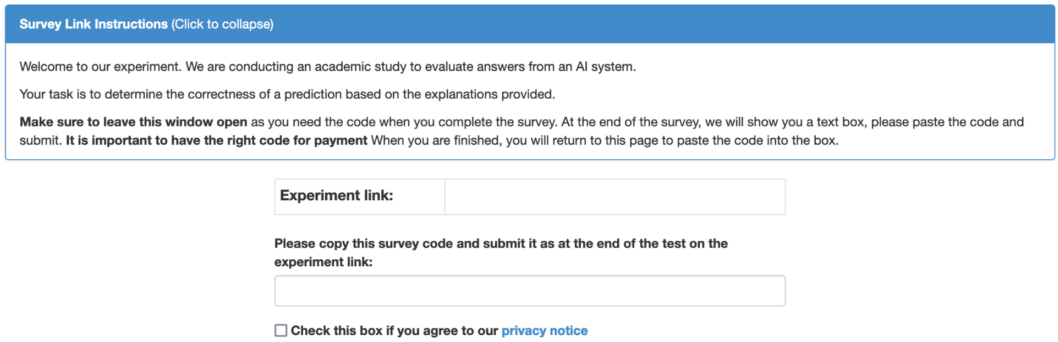


Figure 7: Instructions for crowdworkers recruited from Amazon Mechanical Turk (AMT) with a privacy notice that strictly adheres to General Data Protection Regulation (GDPR) of the European Union.

1	Each tester evaluate 26 predicted triples to maintain good engagement.
2	The first two triples serve as practice to help testers understand the system and questions. The feedback is not included in the statistical analysis.
3	Half of the triples are correctly or incorrectly predicted to avoid dummy feedback.
4	Half of the triples are randomly selected for either XAI method.
5	The order of the predicted triples are randomly shuffled.
6	The same set of predicted triples are presented to testers in the same order.

Table 4: Settings of human-centric evaluation for the CoDEx data.

engagement of the testers, we set checkpoints with three simple predictions, such as a prediction *(Belgium, has_diplomatic_relation_with, Malaysia)* associated with an explanation *(Malaysia, has_diplomatic_relation_with, Belgium)*. After excluding the testers who simply rejected or accepted all predictions or answered the checkpoint questions erroneously, there were 55 well-engaged crowdsourced testers left.

A.9 Further Analysis of Testers' Assessments

During the crowdsourcing evaluation conducted with AMT, we detailed the options to Question 1 (*Is the predicted triple correct?*) to mitigate potential misunderstandings of the testers. As illustrated in Fig. 4, these options include:

1. Yes, based on the explanations, I believe the prediction is correct.
2. The explanations are not enough, but based on the information, the prediction is likely correct.
3. I can't assess the prediction at all.
4. The explanations are not enough, but based on the information, the prediction is likely wrong.
5. No, based on the explanations, I believe the prediction is wrong.

Dataset	GR	GradPath
Kinship	0.7756	0.6571
CoDEx	0.9773	0.8023

Table 5: Error Magnitudes in tester assessments: Measured by MAE (mean absolute error), ranging [0, 2].

To gain further insights into the plausibility of the generated explanations, we conducted a more detailed analysis of testers' assessments, besides accuracy rate. We considered nuances between the options, such as distinguishing between Options 1 and 2, and Options 4 and 5. We set a tester's answer as 1, 0.5, 0, -0.5, -1 for the options 1 – 5, respectively. The ground truth of a prediction is set to be 1 if the prediction is correct, and -1 otherwise. We computed mean absolute error (MAE) between ground truth and testers' answers. The results were reported in Table 5. The experiments revealed that the explanations generated by our GradPath method resulted in smaller error magnitudes in tester assessments compared to those generated by GR. The MAE-based analysis provided additional insights into correct and incorrect assessments, besides the accuracy rate reported in Sec. 5. Taken together, these experimental results highlighted the effectiveness of our explanations in facilitating correct assessments by testers.