Towards Unifying Evaluation of Counterfactual Explanations: Leveraging Large Language Models for Human-Centric Assessments

Marharyta Domnich[1], Julius Välja[1], Rasmus Moorits Veski[1,2], Giacomo Magnifico[1], Kadi Tulver[1], Eduard Barbu[1], Raul Vicente[1]*

[1] Institute of Computer Science, University of Tartu, Tartu, Estonia
[2] École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
marharyta.domnich@ut.ee

Abstract

As machine learning models evolve, maintaining transparency demands more human-centric explainable AI techniques. Counterfactual explanations, with roots in human reasoning, identify the minimal input changes needed to obtain a given output, and hence, are crucial for supporting decision-making. Despite their importance, the evaluation of these explanations often lacks grounding in user studies and remains fragmented, with existing metrics not fully capturing human perspectives. To address this challenge, we developed a diverse set of 30 counterfactual scenarios and collected ratings across 8 evaluation metrics from 206 respondents. Subsequently, we fine-tuned different Large Language Models (LLMs) to predict average or individual human judgment across these metrics. Our methodology allowed LLMs to achieve an accuracy of up to 63% in zero-shot evaluations and 85% (over a 3-classes prediction) with fine-tuning across all metrics. The fine-tuned models predicting human ratings offer better comparability and scalability in evaluating different counterfactual explanation frameworks.

Introduction

The rapid adoption of AI in various domains has significantly increased the urgency for explainable AI models. Counterfactual explanations, which address the question "How should the input be different in order to change the model's decision outcome?" (Wachter, Mittelstadt, and Russell 2017), not only clarify the machine's reasoning but also suggest potential changes that users might implement to achieve different results. These explanations enhance user trust and understanding by providing a richer mental representation compared to causal explanations (Warren, Byrne, and Keane 2023). Additionally, counterfactual explanations align closely with human cognitive processes (Miller 2019), as they provide alternative hypothetical realities that are pervasive in our natural reasoning (Byrne 2002).

Evaluating counterfactual explanations poses a significant challenge in the field. While various quantitative metrics, such as validity, proximity, sparsity, coherence, robustness, and diversity (Guidotti 2022; Karimi et al. 2022; Rasouli and Chieh Yu 2024) are currently used, they often fall short in capturing the human perspective, missing key explanatory virtues, and leading to inconsistent findings that complicate the development of a standardized evaluation framework. User studies are commonly recommended to assess the efficacy of counterfactual explanations, as "excellent computational explanations may not be good psychological explanations" (Keane et al. 2021). Despite this, such studies are rarely utilized for benchmarking counterfactual explanations (Longo et al. 2024). One of the reasons for this is the difficulty and expense of recruiting a sufficient number of experts capable of performing these evaluations. Even when executed, user studies do not guarantee consistent and reproducible results as perceptions of what constitutes a reasonable explanation can vary widely between individuals and user groups (Kenny et al. 2021). Furthermore, most studies only employ a few qualitative measures, such as satisfaction and trust, which fail to address the nuanced features influencing human preferences (Warren, Byrne, and Keane 2023). While human assessments of counterfactual explanations are invaluable, these issues of cost and scalability make it very challenging to make meaningful comparisons and generalizations between multiple frameworks or domains.

Recognizing the limitations of existing methodologies, this paper explores the potential of Large Language Models (LLMs) to serve as a benchmark for automating the evaluation of counterfactual explanations. Current LLMs have demonstrated remarkable capabilities in interacting with natural language data, ranging from extensive data summarisation (Liu et al. 2024) and pattern deduction (Jin et al. 2024) to idea generation (Girotra et al. 2023) and problem-solving through branching solutions (Yang et al. 2024). Based on these premises, LLMs are hypothesized to mimic human evaluative judgments effectively, offering a more accessible and cost-efficient alternative to traditional methods.

In light of these considerations, this paper addresses the following question: **Can the evaluation process of counterfactual explanations be effectively automated using LLMs?** To answer this question, we created a diverse set of 30 counterfactual scenarios that were designed to vary across multiple dimensions of explanatory qualities. The scenarios were evaluated by 206 human respondents in overall satisfaction, feasibility, consistency, completeness, trust, fairness, complexity, and understandability. Next, we divided data for fine-tuning several LLM models to assess ev-

*corresponding author

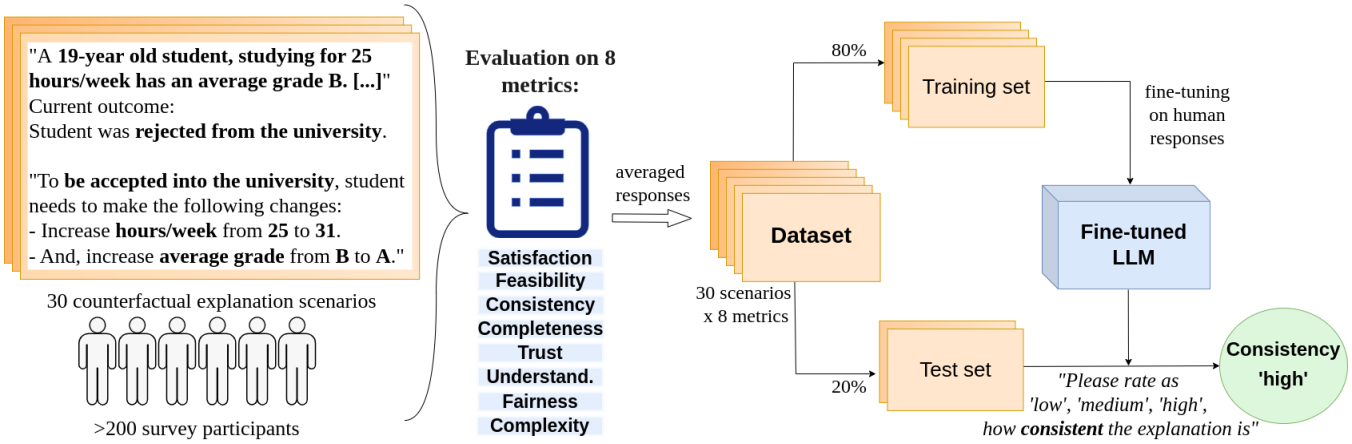


Figure 1: We created a diverse set of counterfactual scenarios where we varied feasibility, consistency, completeness, trust, fairness, complexity and understandability, resulting in 30 counterfactual questions which were evaluated by 206 human respondents on the 8 metrics. We subsequently divided data for fine-tuning several LLM models to assess every metric score and compared the results to human data on a reserved set.

ery metric score and compared the results to human data on a reserved test set. The pipeline can be seen in Figure 1.

Through systematic exploration, this study seeks to bridge the gap between algorithmic outputs and human-centric evaluations, advancing towards more reliable and universally accepted counterfactual explanations in AI systems.

The contributions of the paper are twofold:

- First, we present a diverse dataset of human-evaluated counterfactual explanations CounterEval, encompassing a variety of metrics and scenarios, which could serve both for benchmarking and for training better causal representations of data, as demonstrated in (Chen et al. 2023).
- Second, we introduce a fine-tuned LLM-based evaluator of counterfactual explanations that captures understanding of various explanatory virtues, such as Feasibility, Consistency, Trust, Completeness, Understandability, Fairness, Complexity and Overall Satisfaction.

Related Works

In the following section, we review user studies that focus on evaluating counterfactual explanations, and the potential of LLMs to simulate human responses.

User Studies for Evaluating Counterfactual Explanations

In addition to quantitative explanatory metrics like proximity, validity, or sparsity, most researchers agree that it is crucial to capture the subjective preferences of human users to achieve more human-centric AI explanations (Kirsch 2017; Keane et al. 2021; Longo et al. 2024). Yet, a survey found that only 21% of 100 studies on counterfactual methods included user evaluations (Keane et al. 2021). Furthermore, many of those studies test the use of counterfactual explanations vs no-explanations rather than comparing different methods, leaving only 7% of papers that report user evaluations for benchmarking different counterfactual algorithms.

In recent years, some user studies have been conducted with tabular counterfactual data. For instance, (Warren, Byrne, and Keane 2023) conducted a study with 127 participants to compare the effects of counterfactual and causal explanations on objective prediction accuracy and subjective judgments of satisfaction and trust. (Bove et al. 2023) explored the impact of plural counterfactual examples on objective understanding and a modified Explanation Satisfaction Scale (Hoffman et al. 2018) in a lab study with 112 participants. (Förster et al. 2021) conducted a study with 46 participants assessing explanation realism and typicality. Two user studies have benchmarked counterfactual methods for perceived practicality of users in a study with 135 participants (Ghazimatin et al. 2020), and an online study with 500 responders (Spreitzer, Haned, and van der Linden 2022). Additionally, (Akula et al. 2022) tested their approach on image data, evaluating justified trust as a quantitative metric and explanation satisfaction as qualitative metric.

Overall, user studies on explanation satisfaction often focus on a limited range of aspects (Mueller et al. 2019), typically measuring satisfaction and trust, while neglecting other essential qualities of the explanations themselves. Such studies may fail to capture human preferences, which are shaped by context, presentation, and cognitive biases, especially when preferences are ill-defined (Kliegr, Bahník, and Fürnkranz 2021; Tversky and Simonson 1993). As a result, this limited scope leads to inconsistent perspectives on explanatory qualities, leaving a significant gap in understanding which features define good explanations.

Potential of LLMs in Simulating Human Responses

Predicting human evaluation with machine learning has gained widespread acceptance in various domains, such as human-computer interaction (Kiseleva et al. 2016), recom-

mendation systems (Siro, Aliannejadi, and De Rijke 2023), speech quality assessment (Reddy, Gopal, and Cutler 2022), etc. The advancement of LLMs' causal reasoning abilities (Bhattacharjee et al. 2024) supports their use in explainability, as their natural language explanations exhibit qualities similar to human output (Castelnovo et al. 2024) and the explanatory process can be further enhanced through a post-output chat pipeline (Slack et al. 2023). LLMs have also been used to evaluate and model user satisfaction, providing insight into choices and preferences (Kim et al. 2024), and as artificial user-model tuning pairs (Gao et al. 2024).

At the time of this paper's submission, no prior work existed on simulating human assessments for evaluating counterfactual explanations using LLMs. However, a concurrent effort has since emerged, replicating an existing user study that compared counterfactual and causal explanations by replacing human participants with seven LLMs (Bona et al. 2024), demonstrating that LLMs can replicate some conclusions from the original study. Nonetheless, to the best of our knowledge, no work has yet focused on evaluating the intrinsic quality of counterfactual explanations.

Development and Human Evaluation of a Counterfactual Explanation Dataset

Training LLMs to evaluate the quality of counterfactual explanations in a human-like manner requires human-labeled data. Currently, no widely-used dataset of human-evaluated counterfactual explanations exists. To fill this gap, we created a varied dataset of 30 counterfactual explanation instances, which were graded on 8 different criteria by 206 people through an online survey.

Dimensions of Explanatory Qualities

To select the dimensions for our study, we reviewed the literature on qualitative metrics that influence human judgments. Among the most frequently cited explanatory virtues are coherence and simplicity (Mackonis 2013), aligning with the understanding of human mental models and a preference for consistent and parsimonious information (Johnson-Laird 2010).

Coherence can be measured internally, representing consistency within the explanation, or externally, taking into account the prior knowledge of the rater (Zemla et al. 2017). Our work focuses on internal coherence, measuring consistency within different parts of the explanation, independent of an individual's prior experiences.

The virtue of simplicity, also referred to as (Desired) **Complexity** (Zemla et al. 2017) or Selection (Vilone and Longo 2021), assumes that people prefer simple explanations (Lombrozo 2007). However, evidence suggests humans sometimes favor complex explanations involving more causal links, or that moderate complexity and sufficient detail are preferred (Zemla et al. 2017; Hoffman et al. 2018). In this study, we include Complexity as a metric, with desired values lying in the middle, as explanations may be perceived as overly simple or complex.

A commonly assessed quality in user studies is **Trust**. Definitions often focus on trust in the system generating explanations (Perrig, Scharowski, and Brühlmann 2023). Trust in explanations is considered in terms of trustworthiness, or the perceived credibility of suggested changes (Stepin et al. 2022). We define Trust as the belief that following the explanation will lead to the desired outcome.

Feasibility is one of the most agreed-upon metrics when discussing counterfactual explanations, although discussed under different names: Controllability (Byrne 2019), Actionability (Rasouli and Chieh Yu 2024) and split into Actionability and Mutability (Karimi et al. 2022). While actionability has been employed as a quantitative measure (Guidotti 2022), feasibility refers to whether the proposed changes are perceived as achievable and realistic. Explanations that fail this criterion are rated poorly (Butz et al. 2024).

Understandability, also known as Readability (Stepin et al. 2022) or Comprehensibility (Vilone and Longo 2021), relates to how effectively an explanation conveys the model's decision process to the user and how easily it is grasped. Higher understandability is generally linked to greater user satisfaction, with clear explanations preferred, though complex answers may be favored in some contexts.

Completeness previously discussed as Incompleteness (Zemla et al. 2017) or Informativeness, the latter of which also includes the notion of extraneous information (Stepin et al. 2022), is tied to understanding causal relations and partially depends on domain knowledge (Keil 2006). Evaluating completeness is challenging, since people often fill logical gaps in explanations (Strickland and Keil 2011).

Finally, the dimension of **Fairness** in counterfactual explanations has been highlighted in recent work (Wang et al. 2024). Concerns about models that unintentionally encode or amplify biases in training data (Corbett-Davies et al. 2023) make it crucial to address potential unfairness and discrimination. Fairness has been viewed mainly as a quantitative metric (Ge et al. 2022), with limited understanding of its influence on the perceived quality of the explanation.

Generating Counterfactual Explanations Scenarios

Relying on previous work on human preferences and explanatory virtues, we selected 8 different criteria capturing a range of relevant dimensions (see the previous section for an overview) to guide the creation of diverse counterfactual scenarios. The scenarios in our study are grounded in actual outputs from counterfactual algorithms applied to commonly used datasets for counterfactual explanation evaluation, such as the Adult dataset and the Pima Indians Diabetes dataset. In some cases, we modified the algorithmic outputs to ensure a broader representation across explanatory quality metrics. Additionally, we tailored certain features to enhance clarity for human evaluators based on feedback from a pilot study. All counterfactual scenarios were designed from the perspective of improving the factual situation, as directionality has been shown to influence the way explanations are perceived (Kuhl, Artelt, and Hammer 2023).

We included examples of explanations that fulfilled the different qualities at varying levels to train LLM models to distinguish between good and bad explanations. Specific instances were created by varying metrics, with the exception

of **Understandability** and **Overall satisfaction**. We did not specifically vary the overall satisfaction of explanations, as this metric serves as a general indicator of the perceived quality of an explanation. Also, all explanations were designed to be as understandable as possible, and no instances with purposefully poor wording were included to ensure participants could reliably assess other metrics.

Our dataset contained examples of extreme changes in both categorical and continuous features, as people may evaluate these differently (Warren, Byrne, and Keane 2023). For example, we explored how humans perceive **Feasibility** by creating explanations which changed inactionable features (e.g. age); features by different margins (a 1000€ pay increase vs 10 000€); continuous features outside and within distribution, starting from the value 0; and ordinal features in infeasible directions (e.g. lowering education level). For **Consistency**, we changed features widely considered connected (e.g. hours studied and average grade) in both covarying and conflicting directions, using categorical and continuous features. Differences in **Completeness** were implemented with sufficiently detailed explanations or those containing obvious gaps. Furthermore, useful context was provided for some questions to ensure minimal domain knowledge was sufficient, the lack of which could influence perceptions of completeness. Variety in **Trust** was induced by having logical, solution-oriented explanations and others unlikely to bring about the desired change. Poor **Fairness** was represented by recommendations involving controversial features (e.g. gender, age). To vary **Complexity**, we included instances that might be perceived as too complex as well as too simple by having explanations with a different length and number of recommendations to similar problems. Here, we hypothesised that a desired level of Complexity lies in the middle, which is also reflected in the slightly different scale of measurement compared to other metrics. All selected metrics, along with their definitions and scales as presented in the questionnaire, are detailed in Table 1.

Metric and scale	Description
Overall satisfaction from 1 to 6	This scenario effectively explains how to reach a different outcome
Feasibility from 1 to 6	The actions suggested by the explanation are practical, realistic to implement and actionable
Consistency from 1 to 6	All parts of the explanation are logically coherent and do not contradict each other
Completeness from 1 to 6	The explanation is sufficient in explaining the outcome
Trust from 1 to 6	I believe that the suggested changes would bring about the desired outcome
Understand. from 1 to 6	I feel like I understood the phrasing of the explanation well
Fairness from 1 to 6	The explanation is unbiased towards different user groups and does not operate on sensitive features
Complexity from -2 to 2	The explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex

Table 1: Definitions of the evaluation criteria provided to the respondents in the questionnaire with ranking scale (Understand. stands for Understandability).

Questionnaire Results

To assess the suitability and comprehensibility of the compiled scenarios and evaluation metrics, a pilot study was conducted with 15 volunteers recruited from university students and colleagues. Feedback gathered during the pilot led to revisions in the wording of some metric descriptions. Additionally, the Coherency metric was renamed to Consistency and Bias was changed to Fairness to aid comprehension for the participants.

The final version containing 30 counterfactual scenarios was shared on the Prolific platform and evaluated by 206 respondents. On average, completing the questionnaire took 42 minutes. All metrics were rated on scales detailed in Table 1, with a 6-point ordinal scale from 1 (lowest) to 6 (highest) except for Complexity, rated on a 5-point scale from -2 (too simple) to 2 (too complex), where 0 corresponded to desired complexity. Counterfactual scenarios were presented to participants in random order, while the evaluation metrics remained in the same order. All respondents had to be at least 18 years of age and fluent in English to participate. CounterEval dataset with annotated human responses is available on HuggingFace[1], and an example question is shown in Appendix A, Table A.1 (refer to the extended version of this paper for appendix (Domnich et al. 2024)).

To detect fraudulent participants, a hidden attention check question was included in the questionnaire. Responses were also analyzed based on response time, average understandability score, respondence clustering, and the uniformity of response patterns. Additionally, individual answers to 3 indicator questions were reviewed. For example, if a participant rated an explanation recommending a change in place of birth as feasible, that respondent was flagged. Respondents failing the three aforementioned criteria were excluded from further analysis, with a total of 10 respondents removed.

The survey results indicated satisfactory variance in ratings of the metrics. The questionnaire contained examples of extreme ratings for all metrics with the mean usually balanced in the middle of the scale, as seen in Table 2.

The correlation diagram in Figure 2 shows that all explanatory qualities significantly correlate with each other (p-value $< 10^{-4}$, $\alpha = \frac{0.05}{28}$), except for Complexity and Fairness. An analysis of questions involving varied fairness revealed they did not include overly complex explanations. The intercorrelated responses are likely to reflect that humans grade the explanations as a whole, rating different metrics in the context of the entire scenario and other explanatory virtues. Notably, all metrics correlate positively with satisfaction, highlighting their importance for evaluating the

[1]https://huggingface.co/datasets/anitera/CounterEval

Metric	mean (±sdv)	min / max
Satisfaction	3.02 (±1.11)	1.4 / 5.21
Feasibility	3.27 (±1.15)	1.34 / 5.11
Consistency	3.69 (±1.14)	1.77 / 5.43
Completen.	3.38 (±0.92)	1.78 / 5.33
Trust	3.16 (±1.15)	1.42 / 5.32
Understand.	4.82 (±0.51)	3.92 / 5.58
Fairness	3.89 (±0.97)	1.61 / 5.42
Complexity	-0.26 (±0.39)	-1.03 / 0.84

Table 2: Metric statistics with values averaged per individual question. The table displays mean, standard deviation (sdv), minimum (min), and maximum (max) values.

overall quality of counterfactual explanations. Furthermore, reducing the 7 metrics' scores (excluding Overall Satisfaction) to a two-dimensional space using t-SNE, and coloring by Satisfaction, shows a distinct distribution correlating with overall satisfaction, as detailed in Appendix A, Figure A.1.

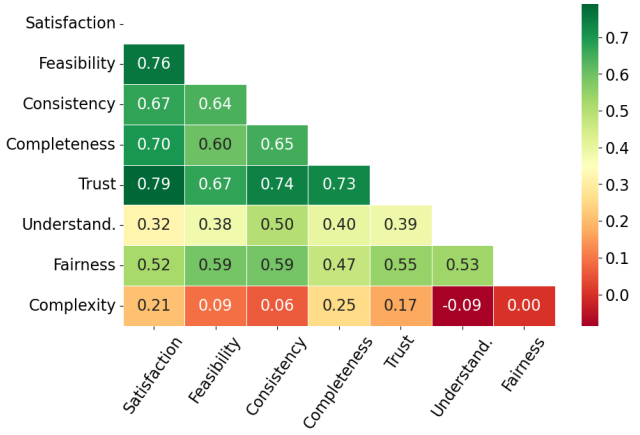


Figure 2: Spearman correlation table between metrics. The values for Complexity were mapped linearly from the original [-2,2] scale to [1,6] to be in line with the other metrics.

Modelling Human Assessment With LLMs

With the questionnaire data as the input dataset, we aimed to test and fine-tune LLMs for automated evaluation of counterfactual explanations. The models selected for this were Llama 3.1 Instruct, Llama 3 Instruct (Dubey et al. 2024) and GPT-4 (OpenAI 2023). Llama models were fine-tuned on HPC clusters equipped with NVIDIA Tesla A100 GPUs using the transformers library by Huggingface (Wolf et al. 2020). QLoRA, which relies on rank decomposition matrices and quantization, was used to reduce memory requirements during fine-tuning (Dettmers et al. 2023).

Dataset Preparation

After gathering and filtering questionnaire responses, further data processing was needed. The preprocessing scripts and model weights are publicly available in the GitHub repository[2]. For each question-metric pair, the average response from 196 participants was used as the final value. Complexity, originally rated on a -2 to 2 scale, was linearly scaled to align with the 1 to 6 scale used for other metrics. To minimize scale effects and enhance generalizability, we consolidated all metric values into three distinct categories: "low," "medium," and "high". Data analysis suggested that the differences between scores of 1 and 2, 3 and 4, and 5 and 6 could be effectively compressed into these categories. This classification ensured a balanced distribution across the dataset. With 30 questions and 8 metrics per question, this resulted in 240 instances of metric evaluation in total.

Prompt Engineering

To achieve the best possible performance from an LLM, three prompt structures were tested and compared.

Importantly, the instruction part of the prompt was taken from the questionnaire directly to ensure that the task reflects the gathered data, and all changes were made in what is known as a "system prompt". The following system prompts were developed:

- A baseline prompt which contains an introduction to counterfactual explanations, the expected output format, and the definition of the metric being evaluated.
- A prompt that contains all the information present in the baseline prompt, but additionally provides definitions for all the metrics, not just the metric being evaluated.
- A prompt that additionally contains two examples of input and expected output, one with Consistency rated as "high" and the other with Feasibility rated as "low". These examples were crafted based on the examples provided for metrics in the questionnaire. The specific examples were chosen to contain different metrics and different output values. All the additional information present in previous prompts is contained in this prompt as well.

The instruction or "user prompt" was adapted from the questionnaire, meaning it contained a factual-counterfactual pair from the questionnaire, alongside a modified metric evaluation question, such as "Please rate as 'low' (very unfeasible), 'medium' or 'high' (completely feasible), how feasible is this explanation:". Consequently, each counterfactual explanation resulted in 8 instances, one for every metric under evaluation. The specific phrasing of all prompts can be found in Appendix B.

All of the prompts were tested using preliminary data from 100 participants and four LLMs, including Mistral-7B Instruct, Llama 2 7B Chat, and 8B and 70B versions of Llama 3 Instruct. Based on the results (Appendix B, Table B.1), the baseline prompt was selected for all further experiments.

Modelling Averaged Human Ratings

Two data splits were tested, with 20% of the dataset set aside for testing and 80% used for training LLMs. The first experiment used a metric-based split, ensuring the testing dataset

[2]https://github.com/anitera/CounterEval

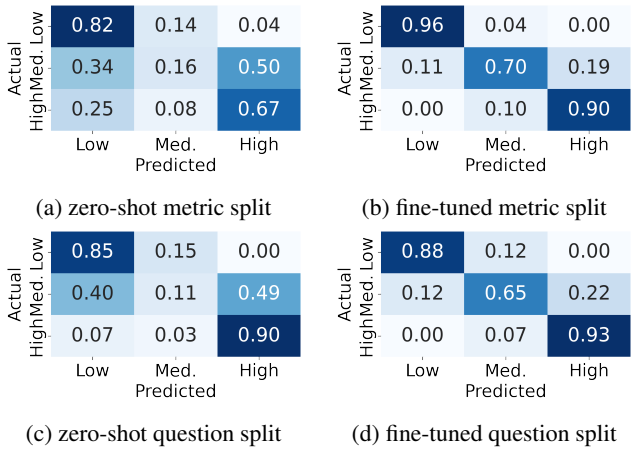
contained examples from all metrics in equal amounts, with 6 examples per metric. In addition, it provided at least one example with a 'high', 'medium' and 'low' answer for every metric. This split has the advantage of a bigger set of unique counterfactual explanation scenarios being present in the test set, leading to a more diverse range of metrics.

The second split, focused on counterfactual explanations, comprised 6 hand-picked questions for the test set. Each question was initially designed to assess a specific metric, typically aiming to elicit either a positive or negative evaluation of that metric. This design informed the selection of questions for the testing set, ensuring that each question covered a different metric with both positive and negative examples. This split accounts for correlations between metrics and ensures that none of the questions are shown in the training set associated with different metrics.

Model	Metric Split		Question Split	
	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
Llama 3 8B	0.48	0.80	0.45	0.77
Llama 3.1 8B	0.52	**0.85**	0.50	0.74
GPT-4	**0.63**	-	0.58	-
Llama 3 70B	0.57	**0.85**	**0.59**	**0.81**

Table 3: Accuracy for metric-based and question-based testing set across evaluated LLMs. Scores averaged over 4 runs, highest score for each column highlighted in bold.

The optimal fine-tuning hyperparameters for every model were discerned through extensive testing (see Appendix C, Table C.1). All models were fine-tuned using a completion-only data collator from Huggingface's *trl* library (Werra et al. 2020) to improve the predictive performance of the models. With a typical language modeling data collator, the model would have learned to predict the question text as well, but this was unnecessary for the task at hand. Due to its proprietary nature, GPT-4 was not used for fine-tuning.



(a) zero-shot metric split

(b) fine-tuned metric split

(c) zero-shot question split

(d) fine-tuned question split

Figure 3: Confusion matrices for Llama 3 70B Instruct for metric split: zero-shot model (a), fine-tuned (b); and question split: zero-shot (c) and fine-tuned (d).

Results in Table 3 show that LLMs possess some ability to evaluate counterfactual explanations even with zero-shot learning, with the GPT-4 reaching 63% accuracy on metric split and Llama 3 70B Instruct reaching 59% on question split. All models surpassed the 33% accuracy expected from random guessing in a three-class task. Fine-tuning significantly improved scores, with the Llama 3 70B Instruct reaching 85% accuracy on the metric split and the recent but significantly smaller Llama 3.1 8B Instruct matching the result. For question split, the highest-performing model was Llama 3 70B Instruct, which after fine-tuning achieved 81% accuracy for three class prediction across 8 metrics.

The confusion matrices show that after fine-tuning, the best-performing models for both splits (Figure 3b, Figure 3d) made no errors misclassifying 'low' as 'high' or vice versa, suggesting a high-level understanding of the metrics. Table 4 highlights accuracy improvements after fine-tuning, with notable gains in Completeness (improving from 33% to 83% and 75% for the metric and question split, respectively), Complexity (from 42% to 75% and 83%), and Understandability, which achieved perfect accuracy. Importantly, Satisfaction showed substantial improvements reaching 96% for metric split and 88% for question split. Feasibility and Trust remain challenging for prediction, largely because assessing the feasibility and outcomes of categorical changes is complex and often unclear as to whether it would bring the desired outcome.

Modelling Individual Preferences

Different people's preferences for explanations exhibit significant variability. To explore the effects of this, we conducted an experiment using a dataset based on specific participants' answers rather than sample averages. To ensure that these participants represent different subgroups of participants, t-SNE was used to reduce the dimensionality of the data, and DBSCAN clustering identified the largest clusters. random participant was selected from each of the four largest clusters. The results of clustering and participant selection can be viewed in Appendix D, Figure D.1.

The selected participants, each from different European

Metric	Metric Split		Question Split	
	Zero-shot	Fine-tuned	Zero-shot	Fine-tuned
Satisfaction	0.67	**0.96**	0.50	**0.88**
Consistency	0.58	0.83	0.83	0.88
Feasibility	0.79	0.96	0.54	0.67
Understand.	0.54	**1.0**	0.92	1.0
Fairness	0.50	**0.83**	0.67	**1.0**
Trust	0.50	0.67	0.50	0.50
Complexity	0.42	**0.75**	0.42	**0.83**
Completeness	0.33	**0.83**	0.33	**0.75**

Table 4: Evaluation of various metrics for Llama 3 70B Instruct model. The largest improvements are highlighted in bold. Each of the accuracy scores is the average score over 4 runs. (Understand. is an abbreviation for Understandability).

Participant	Zero-shot	Fine-tuned
A	0.67	0.87
B	0.58	0.66
C	0.69	0.90
D	0.69	0.90

Table 5: Evaluation accuracy over all metrics for four participants that were selected to represent different subgroups of participants.

countries with educational levels from high school to Master's degree, ensured a diverse range of viewpoints. One participant's experience in machine learning further enriched the variety of responses, detailed in Appendix D, Table D.1.

For each of these participants, zero-shot evaluation and fine-tuning was carried out using the same procedure as in the previous experiments, but using only the model Llama 3 70B Instruct, as it proved to be the most capable (for hyperparameters see Appendix C, Table C.2). The testing set contained the same question-metric pairs as in the first experiment, but with answers from the specific participant.

The results of this process varied, with accuracies ranging from 58% to 69% for zero-shot evaluation. Table 5 shows that the LLM's predictions improved significantly after fine-tuning, reaching accuracies over all metrics of $\sim$90% for 3 participants. One participant appeared to be less consistent, as the model managed to simulate their answers with an accuracy of only 66%. This leads to two conclusions: while LLM's biases and preferences can be tuned to match specific participants to a great extent, some participants' preferences prove significantly more difficult to mimic. However, since this comparison only contained 4 participants and 30 explanations, these conclusions should be considered tentative.

Discussion

The traditional assessment of counterfactual explanations often overlooks human aspects, relying either on inconsistent quantitative metrics (frequently used both within objective function optimization and for evaluation (Cheng, Ming, and Qu 2021)) or on user studies that focus on a specific subset of individuals, lacking comparability over time and methods. To address this, we developed a novel dataset of counterfactual explanations, evaluated by human participants, which demonstrated a diverse spread of evaluations across all metrics, highlighting its applicability in different contexts. Utilizing this dataset to fine-tune LLMs demonstrated promising results, achieving an 85% accuracy, suggesting they can be used to approximate human judgment across various metrics. Furthermore, the zero-shot LLM performance was already notable, achieving up to 63% accuracy. Our experiments also indicate the potential to fine-tune models to individual experts, to target specific expertise or individual preferences.

However, employing LLMs for evaluating counterfactual explanations introduces ethical considerations. There is a risk of reinforcing or introducing biases if the models are not continuously monitored and updated with diverse training data. Furthermore, optimizing explanations to align with model preferences might lead to "gaming" the system, skewing results towards what the model favours rather than enhancing the relevance of the explanations to human users.

A considerable limitation of our study is the dataset size, consisting of only 30 unique counterfactual explanations. A larger dataset would likely enhance model training capabilities. Future work should aim to generate larger datasets using recent counterfactual algorithms (Rasouli and Chieh Yu 2024; Domnich and Vicente 2024; Dandl et al. 2024). These should be presented in smaller subsets to participants for evaluation, given that a single participant can only assess a limited number of explanations thoroughly.

In the future, the main implication of this work is that a fine-tuned LLM should be applied to evaluate various counterfactual algorithms. Additionally, the model can be iteratively retrained with newer and larger architectures and datasets. With the continuously improving size and capabilities of LLMs, this is likely to lead to further improvements in mimicking human evaluation patterns.

Despite the potential, it is crucial to acknowledge that LLMs do not replace the nuanced insights provided by human evaluations. Instead, they can serve as a complementary tool, enhancing scalability and reducing the resources required for broad assessments across multiple frameworks. Moreover, we propose exploring the idea of integrating this model within a human-in-the-loop approach to produce a hybrid model that could refine the quality of counterfactual explanations during the generation process, effectively creating an LLM-in-the-loop approach instead of a human (Abrate et al. 2024), combining the strengths of automated and human evaluations.

Conclusion

This study aims to advance towards more standardized and human-centric evaluations of counterfactual explanations in AI systems. The development and application of our novel dataset, which captures a broad spectrum of human evaluations, reveals the significant potential of LLMs to mirror human judgment with a high degree of accuracy.

Ethical Statement

All data were collected without any personal identifiers. The study was approved by The University of Tartu Research Ethics Committee, and participants provided informed consent for their anonymized data to be used for educational and research purposes.

Acknowledgments

This research was supported by Estonian Research Council Grants PRG1604, the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 952060 (Trust AI), the Estonian Centre of Excellence in Artificial Intelligence (EXAI), the Estonian Ministry of Education and Research.

References

Abrate, C.; Siciliano, F.; Bonchi, F.; and Silvestri, F. 2024. Human-in-the-Loop Personalized Counterfactual Recourse.

In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*, 18–38. Cham: Springer Nature Switzerland. ISBN 978-3-031-63800-8.

Akula, A. R.; Wang, K.; Liu, C.; Saba-Sadiya, S.; Lu, H.; Todorovic, S.; Chai, J.; and Zhu, S.-C. 2022. CX-ToM: Counterfactual explanations with theory-of-mind for enhancing human trust in image recognition models. *iScience*, 25(1): 103581.

Bhattacharjee, A.; Moraffah, R.; Garland, J.; and Liu, H. 2024. Towards LLM-guided Causal Explainability for Black-box Text Classifiers. In *AAAI ReLM 2024*.

Bona, F. B. D.; Dominici, G.; Miller, T.; Langheinrich, M.; and Gjoreski, M. 2024. Evaluating Explanations Through LLMs: Beyond Traditional User Studies. arXiv:2410.17781.

Bove, C.; Lesot, M.-J.; Tijus, C. A.; and Detyniecki, M. 2023. Investigating the intelligibility of plural counterfactual examples for non-expert users: an explanation user interface proposition and user study. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 188–203.

Butz, R.; Hommersom, A.; Schulz, R.; and van Ditmarsch, H. 2024. Evaluating the Usefulness of Counterfactual Explanations from Bayesian Networks. *Human-Centric Intelligent Systems*, 4(2): 286–298.

Byrne, R. M. 2002. Mental models and counterfactual thoughts about what might have been. *Trends in cognitive sciences*, 6(10): 426–431.

Byrne, R. M. J. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, 6276–6282. International Joint Conferences on Artificial Intelligence Organization.

Castelnovo, A.; Depalmas, R.; Mercorio, F.; Mombelli, N.; Potertì, D.; Serino, A.; Seveso, A.; Sorrentino, S.; and Viola, L. 2024. Augmenting XAI with LLMs: A Case Study in Banking Marketing Recommendation. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*, 211–229. Cham: Springer Nature Switzerland. ISBN 978-3-031-63787-2.

Chen, Z.; Gao, Q.; Bosselut, A.; Sabharwal, A.; and Richardson, K. 2023. DISCO: Distilling Counterfactuals with Large Language Models. arXiv:2212.10534.

Cheng, F.; Ming, Y.; and Qu, H. 2021. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2): 1438–1447.

Corbett-Davies, S.; Gaebler, J. D.; Nilforoshan, H.; Shroff, R.; and Goel, S. 2023. The measure and mismeasure of fairness. *The Journal of Machine Learning Research*, 24(1): 14730–14846.

Dandl, S.; Blesch, K.; Freiesleben, T.; König, G.; Kapar, J.; Bischl, B.; and Wright, M. N. 2024. CountARFactuals – Generating Plausible Model-Agnostic Counterfactual Explanations with Adversarial Random Forests. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*, 85–107. Cham: Springer Nature Switzerland. ISBN 978-3-031-63800-8.

Dettmers, T.; Pagnoni, A.; Holtzman, A.; and Zettlemoyer, L. 2023. QLoRA: Efficient Finetuning of Quantized LLMs. arXiv:2305.14314.

Domnich, M.; Valja, J.; Veski, R. M.; Magnifico, G.; Tulver, K.; Barbu, E.; and Vicente, R. 2024. Towards Unifying Evaluation of Counterfactual Explanations: Leveraging Large Language Models for Human-Centric Assessments. arXiv:2410.21131.

Domnich, M.; and Vicente, R. 2024. Enhancing Counterfactual Explanation Search with Diffusion Distance and Directional Coherence. In Longo, L.; Lapuschkin, S.; and Seifert, C., eds., *Explainable Artificial Intelligence*, 60–84. Cham: Springer Nature Switzerland. ISBN 978-3-031-63800-8.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783.

Förster, M.; Hühn, P.; Klier, M.; and Kluge, K. 2021. Capturing Users' Reality: A Novel Approach to Generate Coherent Counterfactual Explanations. In *54th Hawaii International Conference on System Sciences, HICSS 2021, Kauai, Hawaii, USA, January 5, 2021*, 1–10. ScholarSpace.

Gao, G.; Taymanov, A.; Salinas, E.; Mineiro, P.; and Misra, D. 2024. Aligning LLM Agents by Learning Latent Preference from User Edits. arXiv:2404.15269.

Ge, Y.; Tan, J.; Zhu, Y.; Xia, Y.; Luo, J.; Liu, S.; Fu, Z.; Geng, S.; Li, Z.; and Zhang, Y. 2022. Explainable fairness in recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 681–691.

Ghazimatin, A.; Balalau, O.; Saha Roy, R.; and Weikum, G. 2020. Prince: Provider-side interpretability with counterfactual explanations in recommender systems. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 196–204.

Girotra, K.; Meincke, L.; Terwiesch, C.; and Ulrich, K. T. 2023. Ideas are Dimes a Dozen: Large Language Models for Idea Generation in Innovation. *SSRN Electronic Journal*.

Guidotti, R. 2022. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608*.

Jin, M.; Wang, S.; Ma, L.; Chu, Z.; Zhang, J. Y.; Shi, X.; Chen, P.-Y.; Liang, Y.; Li, Y.-F.; Pan, S.; and Wen, Q. 2024. Time-LLM: Time Series Forecasting by Reprogramming Large Language Models. arXiv:2310.01728.

Johnson-Laird, P. N. 2010. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43): 18243–18250.

Karimi, A.-H.; Barthe, G.; Schölkopf, B.; and Valera, I. 2022. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5): 1–29.

Keane, M. T.; Kenny, E. M.; Delaney, E.; and Smyth, B. 2021. If Only We Had Better Counterfactual Explanations: Five Key Deficits to Rectify in the Evaluation of Counterfactual XAI Techniques. arXiv:2103.01035.

Keil, F. C. 2006. Explanation and understanding. *Annu. Rev. Psychol.*, 57(1): 227–254.

Kenny, E. M.; Ford, C.; Quinn, M.; and Keane, M. T. 2021. Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, 294: 103459.

Kim, H.; Choi, Y.; Yang, T.; Lee, H.; Park, C.; Lee, Y.; Kim, J. Y.; and Kim, J. 2024. Using LLMs to Investigate Correlations of Conversational Follow-up Queries with User Satisfaction. arXiv:2407.13166.

Kirsch, A. 2017. Explain to whom? Putting the User in the Center of Explainable AI. In *CEx@AI*IA*.

Kiseleva, J.; Williams, K.; Hassan Awadallah, A.; Crook, A. C.; Zitouni, I.; and Anastasakos, T. 2016. Predicting User Satisfaction with Intelligent Assistants. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, 45–54. New York, NY, USA: Association for Computing Machinery. ISBN 9781450340694.

Kliegr, T.; Bahník, Š.; and Fürnkranz, J. 2021. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 295: 103458.

Kuhl, U.; Artelt, A.; and Hammer, B. 2023. For Better or Worse: The Impact of Counterfactual Explanations' Directionality on User Behavior in xAI. In *World Conference on Explainable Artificial Intelligence*, 280–300. Springer.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374.

Lombrozo, T. 2007. Simplicity and probability in causal explanation. *Cognitive psychology*, 55(3): 232–257.

Longo, L.; Brcic, M.; Cabitza, F.; Choi, J.; Confalonieri, R.; Ser, J. D.; Guidotti, R.; Hayashi, Y.; and et al. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106: 102301.

Mackonis, A. 2013. Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, 190(6): 975–995.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267: 1–38.

Mueller, S. T.; Hoffman, R. R.; Clancey, W.; Emrey, A.; and Klein, G. 2019. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. *arXiv preprint arXiv:1902.01876*.

OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774.

Perrig, S. A.; Scharowski, N.; and Brühlmann, F. 2023. Trust issues with trust scales: examining the psychometric quality of trust measures in the context of AI. In *Extended abstracts of the 2023 CHI Conference on human factors in computing systems*, 1–7.

Rasouli, P.; and Chieh Yu, I. 2024. CARE: Coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, 17(1): 13–38.

Reddy, C. K.; Gopal, V.; and Cutler, R. 2022. DNSMOS P. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 886–890. IEEE.

Siro, C.; Aliannejadi, M.; and De Rijke, M. 2023. Understanding and Predicting User Satisfaction with Conversational Recommender Systems. *ACM Trans. Inf. Syst.*, 42(2).

Slack, D.; Krishna, S.; Lakkaraju, H.; and Singh, S. 2023. Explaining machine learning models with interactive natural language conversations using TalkToModel. *Nature Machine Intelligence*.

Spreitzer, N.; Haned, H.; and van der Linden, I. 2022. Evaluating the Practicality of Counterfactual Explanations. In *XAI. it@ AI* IA*, 31–50.

Stepin, I.; Alonso-Moral, J. M.; Catala, A.; and Pereira-Fariña, M. 2022. An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences*, 618: 379–399.

Strickland, B.; and Keil, F. 2011. Event completion: Event based inferences distort memory in a matter of seconds. *Cognition*, 121(3): 409–415.

Tversky, A.; and Simonson, I. 1993. Context-dependent preferences. *Management science*, 39(10): 1179–1189.

Vilone, G.; and Longo, L. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76: 89–106.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wang, X.; Li, Q.; Yu, D.; Li, Q.; and Xu, G. 2024. Counterfactual explanation for fairness in recommendation. *ACM Transactions on Information Systems*, 42(4): 1–30.

Warren, G.; Byrne, R. M. J.; and Keane, M. T. 2023. Categorical and Continuous Features in Counterfactual Explanations of AI Systems. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23, 171–187. New York, NY, USA: Association for Computing Machinery. ISBN 9798400701061.

Werra, L. v.; Belkada, Y.; Tunstall, L.; Beeching, E.; Thrush, T.; Lambert, N.; and Huang, S. 2020. TRL: Transformer Reinforcement Learning.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2024. Large Language Models as Optimizers. arXiv:2309.03409.

Zemla, J. C.; Sloman, S.; Bechlivanidis, C.; and Lagnado, D. A. 2017. Evaluating everyday explanations. *Psychonomic bulletin & review*, 24: 1488–1500.