UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Julius Välja

Assessing the Quality of Counterfactual Explanations with Large Language Models

Bachelor's Thesis (9 ECTS)

Supervisors: Marharyta Domnich
Raul Vicente
Eduard Barbu

Tartu 2024

Assessing the Quality of Counterfactual Explanations with Large Language Models

Abstract:

With the accelerating spread of machine learning models, the complexity and lack of transparency of current models has become a major source of concern. The field of Explainable AI focuses on finding methods that can uncover the inner logic of these models. One such method is counterfactual explanations, which seek to answer the question "How would the original situation need to be different to achieve a different prediction from the model?". However, the qualities that make a counterfactual explanation good are not fully understood and are difficult to quantify. In this thesis, a survey was used to gather a dataset of human-evaluated counterfactual explanations, with an array of qualities defined based on previous literature. This dataset was used to explore Large Language Models' (LLMs) ability to evaluate subjective qualities of counterfactual explanations with and without fine-tuning. The results showed that large LLMs exhibit 70% to 95% accuracy at this task, depending on the specific model and testing dataset. While smaller LLMs could be fine-tuned to achieve acceptable accuracy, they were generally significantly less capable. In addition, the effect of correlations between metrics was tested for and experiments performed to assess the feasibility of predicting user satisfaction as well as modelling individual preferences. These results pave the way for future research regarding the automatic evaluation of counterfactual explanations and the development of new search algorithms.

Keywords:
Articial Intelligence, XAI, Counterfactual explanations, LLM, Human evaluation

CERCS: P176, Artificial intelligence

Kontrafaktuaalsete seletuste kvaliteedi hindamine suurte keelemudelite abil

Lühikokkuvõte:

Masinõppe mudelite kiire leviku tõttu on nende mudelite keerulisus ja läbipaistmatus üheks levinud murekohaks. "Seletatav tehisintellekt"(*Explainable AI*) on informaatika ala, mis keskendub meetoditele, millega on võimalik masinõppe mudelite sisemist loogikat uurida. Üks selline meetod on kontrafaktuaalsed seletused, mis vastavad küsimusele "Kuidas peaks olukord erinema, et mudel ennustaks teistsugust tulemust?". Tunnuseid, mis teevad sellise selgituse heaks on vähe uuritud ning neid on raske arvutuslikult hinnata. Selles lõputöös loodi küsimustik, mille abil koguti andmestik kontrafaktuaalsetest

seletustest ning nendele antud hinnangutest erinevate kriteeriumite põhjal. Selle andmestiku abil uuriti suurte keelemudelite võimet neid kriteeriume automaatselt hinnata, muuhulgas kasutades siirdeõpet. Tulemusena saavutati suurte keelemudelite puhul täpsus 70% kuni 95%, sõltuvalt konkreetsest mudelist ja testimisandmestikust. Väiksemad keelemudelid olid oluliselt vähem võimekad, kuid siirdeõppe abil suutsid saavutada 70% täpsust. Lisaks uuriti kriteeriumivaheliste korrelatsioonide mõju tulemustele ning hinnati asjaolu, kuivõrd on inimeste üldist rahulolu seletusega võimalik automaatselt hinnata . Need tulemused võimaldavad edasisi uuringuid kontrafaktuaalsete seletuste automaatses hindamises ning uute kontrafaktuaalsete seletuste otsingualgoritmide arenduses.

Võtmesõnad:
Tehisintellekt, XAI, Kontrafaktuaalsed seletused, Keelemudelid, Inimhindamine

CERCS: P176, Tehisintellekt

Contents

Introduction

Artificial Intelligence (AI) has seen a major resurgence in the last decade, permeating most areas of life. With predictive AI models making their way into both everyday and high-risk areas, the need for users to be able to trust these systems is greater than ever. Despite the explosion of AI models vying for users' attention, it has been found that their uptake is limited by the users' lack of trust in these models [1, 2]. This lack is well-justified, since hidden biases and questionable decisions are not uncommon in such models [3]. To address this challenge, the field of Explainable Artificial Intelligence (XAI) has been gaining traction, with the goal of making these decisions explainable and understandable.

One of the cutting-edge methods used in XAI is Counterfactual Explanations (CEs) [4, 5, 6, 7]. They can be generated from any machine learning model, which they treat as a "black box" [4]. Counterfactual explanations seek to answer the question: "How would the original situation need to be different to achieve a different prediction from the model?". For example, let us say a 20-year-old person with no income applied for a loan, but was rejected by an automated system. A counterfactual explanation for this rejection might suggest that if the person earned 1000€ per month instead, they would have qualified for the loan. One of the main advantages of CEs is their ability to provide actions the person can take to improve their outcome [4]. However, there can be many counterfactual scenarios that theoretically improve the outcome, but not all of them are equally feasible and realistic. A different CE might suggest that the person instead change their country of birth from Germany to United States. Clearly, this is not something the person can wilfully change, but machine learning models lack any understanding of such human preferences.

This reflects the difficulty of finding "good" counterfactual explanations. There have been many attempts to create algorithms that find "better" counterfactual explanations [6, 8], but what makes a "good" CE is still an open question [9]. Additionally, even if we know what a good counterfactual explanation should include, that may not necessarily mean that we can generate them. Human evaluation of explanations is extremely costly and time-consuming, and therefore automatic evaluation of explanations is a major goal of XAI.

This thesis attempts to further our understanding of what makes a counterfactual explanation "good" and validate a potential solution for automatically evaluating CEs by use of Large Language Models (LLMs). LLMs have made unprecedented leaps in recent years and have shown marked ability to solve tasks requiring text comprehension. Additionally, evaluation by LLMs can be significantly cheaper and more accessible than human evaluation.

However, automatic evaluation requires "ground truth" data as a basis for learning.

As part of this thesis, a comprehensive survey was developed, which consisted of 30 counterfactual explanations. 100 participants were tasked with evaluating these explanations on 7 different metrics, which were chosen based on extensive analysis of previous literature, as well as general satisfaction. The results from this survey were then used to create a dataset for CE evaluation. The capabilities of several different Large Language Models were explored and compared based on this dataset. The effects of fine-tuning on evaluation accuracy were thoroughly investigated, as well as the differences between evaluating all metrics and evaluating general satisfaction. Additionally, the feasibility of evaluating based on specific participants preferences was tested.

Specifically, this thesis attempts to answer the following research questions:

- RQ: Can the process of evaluating counterfactual explanations be automated by LLMs?
  - Q1: Can LLMs already answer like humans without fine-tuning?
  - Q2: Is evaluating user satisfaction easier than evaluating other metrics?
  - Q3: Are there significant differences between different LLMs?

In Chapter 1 of the thesis, an extensive theoretical overview of related research will be presented. Chapter 2 will give an overview of the development of the questionnaire and the background of the metrics used for evaluation. Chapter 3 will present the methods used for answering the research questions. Chapter 4 contains the results of the experiments. Chapter 5 will provide an analysis and discussion of the results, including the implications and limitations of the research. Finally, the last chapter will summarise the results and contribution of this thesis.

The first Appendix presents examples of questionnaire questions at different stages of development. The second Appendix contains examples for metrics provided to participants in the questionnaire. The third Appendix contains examples of three different prompts used in experiments.

The Artificial Intelligence tool Microsoft Copilot[1] was occasionally used to improve formatting and phrasing in some sections of this thesis.

[1]Microsoft (2024). Microsoft Copilot: `https://copilot.microsoft.com`

1 Background

The field of explainable artificial intelligence, commonly abbreviated as XAI, has been around for decades, first gaining traction in the 1980s [10]. The prevalence of machine learning and AI has skyrocketed in the following 40 years, but with the rise of neural networks, their decisions have become even more opaque [11]. In recent years, the importance of explainability has been recognised by both industry [12, 13] and regulatory bodies [14, 15]. Explainability or interpretability in this context can be defined as the capacity to present an explanation of the processes leading a model to a specific decision [16]. The need for explainability in AI models stems from two main goals: transparency and trust [17]. With typical "black box" models, it is extremely challenging to analyse the direct relations between inputs and outputs, which leads to a high degree of uncertainty about the models' behaviour [18].

1.1 Current state of XAI

XAI methods can generally be divided into two possible approaches. The first approach is developing AI models which have explainability built into their structure. One of the most common examples of an inherently explainable model is the decision tree, since its structure reveals the reasoning behind its decision fully [5, 19]. The second approach is called *post-hoc* explanations, which consists of any methods applied after training the model. Research has shown that there tends to be an inverse relationship between the explainability of a model and its performance [20]. For this reason, *post-hoc* explanations have become the preferred area of research in the field.

Post-hoc explanations enable developers and users to extract information about the decision-making process of the model, which aids in ensuring that the model works reliably, predictably, and without propagating bias. Such explanations can also be helpful to justify or improve decisions made by autonomous agents, when there is need to do so [21]. For example, if an autonomous vehicle veers off the road, it can be important to investigate why it made such a decision and whether it was appropriate.

One of the most popular recent approaches in XAI is generating proxy explanations, also known as surrogation. With this solution, an inherently interpretable model is created as a proxy of the primary model. A commonly used surrogation method is LIME, or Local Interpretable Model-agnostic Explanations [22]. As the name suggests, this method investigates what happens in the near vicinity of the data point through perturbation and uses this data to create a local interpretable proxy model that estimates the linear decision boundary of the original model in this locality [5, 22]. Through interpreting this proxy model, users can assess which features were most relevant to the model's decision.

There have been attempts to utilise game theory in explaining model decisions, resulting

in the popular method created by Lundberg and Lee known as SHAP (SHapley Additive exPlanations). SHAP uses Shapley values, a concept from game theory, to estimate the contribution of each attribute to the final prediction. The relative sizes of these contributions can then be used to gauge which attributes had an outsized effect on the output in a specific instance [5, 23].

There are several ways to categorise *post-hoc* XAI methods. Methods can be model-agnostic, which means that the method can be used for explaining any machine learning model [19]. In the modern competitive landscape of AI, having methods that can explain different models is a highly coveted goal, since it enables not only comparing different models on their interpretability, but also verifying whether models base their decisions on sound reasoning. However, some methods can also be specific for some model type, such as neural networks. Such model-specific methods can be useful when explaining more complex models, but their utility is limited due to their inability to explain other models [19].

Another major distinction is whether methods generate local or global explanations, as specified by Molnar. According to him, global explanations attempt to explain the overall structure of a model, such as which input features it takes into account the most on average. In contrast, local explanations provide information about how a decision for a specific instance was reached, but this information may not be applicable to the model as a whole [19]. One local method which has gained popularity in recent years is known as counterfactual explanations, which is the main topic of this thesis.

1.2 Counterfactual explanations

Counterfactual Explanations (CEs) are a local model-agnostic *post-hoc* method for explaining AI decisions, proposed by Wachter et al. in 2017 [4]. The core idea behind CEs is proposing an alternative scenario, as close to the original scenario (the "factual") as possible, that would result in a different outcome. For example, if a person's loan application was rejected, they might receive a counterfactual explanation, such as "If your monthly income was 200€ higher, you would have been approved for the loan". These kinds of explanations are easy to understand by people without any technical knowledge and indicate which properties of the scenario had a major impact on the outcome [4, 24]. In addition, they provide concrete steps that the person could make to change their outcome to a more desirable one.

As local explanations, CEs are applied to explain every instance individually [4, 5]. A counterfactual explanation generates an alternative scenario that informs a specific user about their particular case and the corresponding model prediction, which may not be relevant to others and provides minimal information on the model as a whole. This counterfactual scenario is obtained by optimising an objective function, which usually

consists of a loss function to flip the outcome and a measure of distance from original data point to counterfactual data point [4]. The result of the optimisation process is an alternative set of feature values that leads to a model predicting the opposite class, while remaining reasonably similar to the original situation.

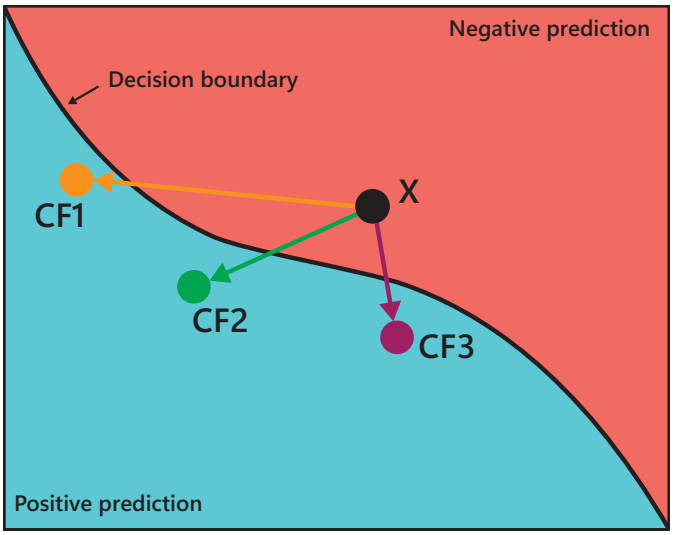

Figure 1. Illustration of counterfactual explanations. X refers to the original data point, while CF1, CF2 and CF3 are different possible counterfactual scenarios that flip the prediction.

Importantly, any specific situation can have different paths to reach the desired outcome, hence there can be several different CEs with different proposed changes [4]. This is illustrated in Figure 1. Determining which of these is the optimal explanation can be challenging, since the relevance of an explanation depends significantly on the original situation, the properties of the variables, and the preferences of the user. Many researchers have suggested providing a set of several CEs as a solution to this problem [4, 7]. However, subsequent research has made strides in attempting to provide more useful explanations directly, without the need to generate many CEs.

Mothilal et al. proposed an improved method for generating CEs called DiCE (Diverse Counterfactual Explanations), which prioritises generating a diverse set of CEs through implementing a diversity measure into the objective function [7]. They attempt to tackle the problem of feasibility by letting a user provide constraints on which features can be changed. However, the utility of this approach is limited, since it relies on human input and only enables the user to exclude certain features completely, which ignores the possibility of infeasible changes made to otherwise feasible features.

An alternative was suggested by Poyiadzi et al. with their FACE (Feasible and Actionable

Counterfactual Explanations) framework. Their solution is twofold, taking into account both feasibility and actionability. Instead of simply searching for the closest possible data point, FACE considers both the density of the region where the data point lies and the existence of a feasible and high-density path between the original data point and the counterfactual data point. This aims to solve two crucial problems with previous methods: the suggested changes may consist of an unlikely or impossible set of feature values, meaning it is located in a low-density region of the data distribution, and there may not exist a realistically achievable way to go from the factual to the counterfactual. Unfortunately, this approach does not solve the question of actionability fully, since it is unable to account for the direction of changes [8].

One of the most comprehensive proposed frameworks for generating CEs is CARE (Coherent Actionable Recourse based on sound counterfactual Explanations), by Rasouli et al. It is based on the concept of actionable recourse, which refers to counterfactual explanations that suggest feasible changes based on the user's preferences. CARE consists of four modules, called "validity", "soundness", "coherency", and "actionability". Each module attempts to satisfy a particular desired property of the explanation, with the optimisation algorithm balancing all the modules to find the best overall explanations. The "actionability" module additionally enables users to set fixed rules which the algorithm must follow, such as not changing age. Furthermore, Rasouli et al. explore generating temporal sequences, which propose a sequence of steps to achieve an outcome, instead of a simple set of changes. This better reflects real-world situations, where changes are rarely instantaneous and independent from each other [6].

While all of these algorithms have attempted to take into account attributes that may make their explanations preferable to humans, a strong evidence-based justification for their choice and definition of attributes was not provided. This reflects a wider problem in XAI, where researchers often struggle to thoroughly demonstrate that their methods are genuinely superior for most people, rather than merely preferred based on their own assessments. Consequently, there is a growing need for research into how humans evaluate counterfactual explanations in practical, real-world settings.

1.3 Desirable qualities of explanations

Humans commonly create explanations to inform their decisions and opinions about the world [25]. True explanations aid in understanding complex processes and predicting future outcomes. Despite this, research has shown that humans exhibit many biases that affect how they create and evaluate explanations [26, 27].

It is important to note that explanations can serve different functions. Explanations can be used to facilitate learning by improving the user's understanding of why events occur [27] or to help the user avoid or achieve certain outcomes [4]. The desired function

of the explanation determines its structure and content, since a single explanation can rarely fulfil all the possible functions [25, 28].

The way humans explain unexpected phenomena often relies on a process called *inference to the best explanation* (IBE), which is a formalisation of the logical process of considering alternatives and choosing the most likely explanation as true [29]. IBE is commonly used when there are competing scientific theories aiming to explain the same phenomena [29]. This process relies on comparing different possible explanations on merits known as *explanatory virtues* [26, 29, 30]. However, the precise criteria making up these virtues is not agreed upon, although some, such as simplicity, seem to be widely regarded as important [31, 32]. Such criteria can be either intrinsic to the explanation, such as the interpretability of the explanation, or extrinsic, such as the perceived expertise of the person or system providing the explanation [26]. The latter becomes important when considering explanations in everyday contexts, since people are predisposed to cognitive biases and shortcuts [27]. Therefore, the criteria used to evaluate common explanations may not fully overlap with the objective criteria used for logical inference.

There exists a large body of research exploring which explanatory criteria are relevant to people in everyday situations, but so far consensus has not been reached. For example, while it is often assumed that simpler explanations that cite fewer causes are preferred [31, 33], some recent research has found the opposite effect [26]. Zemla et al. propose two possible reasons why people may prefer complex explanations: either people prefer to maximise the likelihood of the result, or people are naturally predisposed to prefer more complex explanations, since real-world phenomena are rarely explainable through a single causal mechanism [26]. Additionally, people may prefer different explanations, based on their previous knowledge and the context of the explanation. For example, an explanation for why a person is likely to have cancer needs to be far more detailed and convincing than an explanation for why it will rain tomorrow. Finally, a person's intuition can play a role in how complex an explanation needs to be. When an explanation confirms the person's intuitive expectation, they likely won't require further convincing, whereas when there is a conflict between intuition and explanation, the explanation needs to provide convincing evidence for why it is true.

Another explanatory criterion widely supported by research is coherence [26, 29, 31]. While it can refer to different concepts, it most often relates to two concepts: the explanation must not contradict prior beliefs and the explanation must not contradict itself [26]. While the latter is required for the explanation to be valid and useful, the former can depend on the specific situation. Beliefs are inherently malleable, and therefore there may be situations where an explanation can be a basis for revising previously held beliefs. For example, an explanation can bring to light relationships in the data that were either previously unknown or are an indicator of the low quality of the

underlying data.

The question of how humans evaluate explanations has garnered significantly more attention in recent years. This stems from the need for AI explanations to match humans' preferences and needs. An interpretable AI model is useless if the explanations it presents are not useful or understandable to its users. Tackling this problem has driven researchers to test AI explanations with real participants to gain a better understanding of the explanatory virtues that humans use in different contexts.

In 2017, Zemla et al. presented their seminal paper on evaluating everyday explanations. They gathered a dataset of eight explananda, or "thing being explained", and three distinct explanations per explanandum from the Reddit *Explain Like I'm Five* community. They then had 235 participants evaluate these explanations on 21 different attributes, such as Complexity or Quality. Each attribute was tested for correlation with Quality and with other attributes. Their findings contradicted previously held assumptions regarding the simplicity of explanations, since their study revealed that participants preferred more complex explanations. This finding was further tested with a second experiment designed to specifically target complexity and the number of causal mechanisms present in an explanation. This experiment supported their earlier findings, providing valuable new information about humans' preference for complex explanations [26].

Spreitzer et al. identified a set of attributes that need to be satisfied by counterfactual explanations for them to be practical and had survey participants evaluate CEs based on these attributes. Their set of attributes was developed based on previous research by Miller regarding what constitutes a good explanation [27]. Their experimental setup consisted of 2 datasets and 30 instances from either dataset. Each instance consisted of a factual, a counterfactual explanation generated using CARE, and a counterfactual explanation generated using the original algorithm proposed by Wachter et al. [4]. Each participant was randomly assigned one of the datasets and was asked to rate the instances on 9 questions that captured the previously mentioned set of attributes. Their results show that CARE was consistently rated as giving more practical counterfactual explanations. However, their experimental setup did not evaluate the authors' choice of attributes and questions. Therefore, it is possible that the questions they presented did not provide a comprehensive assessment of the practicality of counterfactual explanations [24].

Stepin et al. set out to assess the correlation between an objective metric called perceived explanation complexity (PEC) and a handful of qualitative metrics based on Gricean Maxims, as set out by H.P. Grice [34]. They defined an objective and quantitative way to evaluate explanation complexity based on previous research in linguistics. The four Gricean Maxims were divided into 5 evaluation metrics, such as Trustworthiness and Relevance. A small sample of 18 participants rated 15 counterfactual explanations on these evaluation metrics, and these results were used to gauge correlations between the

metrics and PEC. PEC was shown to have a strong negative correlation with readability and relevance and a moderate correlation with informativeness. In a subsequent survey, they simplified the evaluation metrics to include only Trustworthiness and Satisfaction. Despite a larger sample size of 60 participants, no statistically significant correlations were found [35].

Whereas the previous studies all used classification tasks, Yao et al. examined how humans judge counterfactual explanations in recommender systems and whether it can be estimated computationally. The recommender system they chose gave users recommendations on which movies to watch, with explanations on why a particular movie was recommended to them. 400 participants were asked to evaluate 12 explanations, with seven metrics per explanation. These results were then compared with the authors' proposed computational methods, to validate whether the methods are a reasonable proxy for human evaluation. They showed that the proposed methods, based on counterfactual logic, are a reasonable proxy for human evaluation in recommender systems [36].

In their 2012 paper, Mccloy and Byrne studied how humans approach controllability in counterfactual thinking. Their experiments showed that humans prefer to mutate events that are both controllable and inappropriate. In this context, inappropriate events are events that can be considered abnormal or exceptional based on societal norms of behaviour in a given context. Additionally, their research showed that when presented with several controllable events, humans prefer to mutate inappropriate ones regardless of whether the original result was positive or negative [37].

The inconsistent results and limited scope of previous research highlights the need for a comprehensive exploration of which CE qualities are important to users. To alleviate this problem, the next section of this thesis details the creation of a comprehensive survey to gather human evaluations of counterfactual explanations.

2 Survey design

Alongside Rasmus Moorits Veski [38] and with help from researchers at the Natural and Artificial Intelligence Lab of the University of Tartu[2], the author of this thesis developed a questionnaire to gauge how humans evaluate and rate counterfactual explanations. The questionnaire contained 30 examples of counterfactual explanations and was distributed on Prolific[3], an online research platform. Participants rated explanations on 8 different metrics, aimed at capturing a variety of qualities present in a counterfactual explanation.

2.1 Evaluation criteria

An event can have many explanations that are equally valid and true, yet people prefer some explanations over others based on certain cognitive biases [26]. In this study, participants were asked to rate counterfactual explanations on 8 metrics designed to capture as many different aspects and biases as possible. For rating the metrics, a five-point or six-point ordinal rating scale, also known as a Likert scale, was used [39]. The metrics were derived based on a wide range of previous research, from the field of social science as well as computer science [5, 26, 27, 36, 40]. Due to the user-centric nature of the research, the names and definitions of metrics were chosen to be as clear and understandable as possible. The result is presented in Table 1, while the subsequent paragraphs will delve into the metrics individually and the specific choices and reasoning behind them.

Table 1. Evaluation criteria used in the questionnaire

Metric	**Definition**	**Rating scale**	**Related qualities from literature**
Overall satisfaction	This scenario effectively explains how to reach a different outcome.	6-point Likert scale, from 1 to 6	Quality [26]
Feasibility	The actions suggested by the explanation are practical, realistic to implement and actionable.	6-point Likert scale, from 1 to 6	Controllability [41], Actionability [6]
Consistency	All parts of the explanation are logically coherent and do not contradict each other.	6-point Likert scale, from 1 to 6	Internal coherence [26], Coherence [29, 31], Coherency [6]

[2]See `https://nail.cs.ut.ee/`

[3]Available at `https://www.prolific.com`

Completeness	The explanation is sufficient in explaining the outcome.	6-point Likert scale, from 1 to 6	Informativeness [35], Incompleteness [26]
Trust	I believe that the suggested changes would bring about the desired outcome.	6-point Likert scale, from 1 to 6	Trustworthiness [35], Perceived truth [26], Truthfulness [24]
Understandability	I feel like I understood the phrasing of the explanation well.	6-point Likert scale, from 1 to 6	Readability [35], Comprehensibility [5, 11]
Fairness	The explanation is unbiased towards different user groups and does not operate on sensitive features.	6-point Likert scale, from 1 to 6	Fairness [5]
Complexity	The explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex.	5-point Likert scale, from -2 to 2	Complexity [26], Desired complexity [26], Selection [11]

The metric of **Overall satisfaction** captures how generally satisfied the respondents were with the presented explanation. Ideally, this metric would capture all the variance of the other metrics and provide a baseline or "ground truth" for explanation quality, which could later be used to compare the extent to which people value different metrics.

Feasibility is one of the most common and agreed-upon metrics when discussing counterfactual explanations [8, 41, 42]. It refers to whether the proposed changes are theoretically achievable and importantly, whether they are practical. Previous research has shown that explanations that do not fulfil this criterion are consistently rated worse [37, 43]. This metric encompasses several commonly discussed metrics, such as "actionability", which refers to whether an attribute change is reasonably achievable [8], and "controllability", which refers to whether an attribute change corresponds to a possible decision that the subject could have made [41, 44]. For the purposes of this survey, these two metrics were combined into the overarching metric of Feasibility, since they are closely linked and often inextricable.

To sufficiently cover different aspects of Feasibility, the differences between continuous and categorical variables were considered and subsequently, 8 distinct subcases with suspected low Feasibility were defined. These consisted of:

- changing a continuous value to an extreme, out-of-distribution value;

- changing a continuous value from 0 to a different value;
- changing a continuous value to the same extent as in the previous case, but not beginning from zero;
- a large change in a continuous value, but with both values in the distribution;
- a large jump in an ordered categorical feature;
- a change in the opposite direction for an ordered categorical feature;
- changing an unordered categorical feature;
- changing a feature that cannot usually be changed.

For each of these subcases, a separate counterfactual explanation contradicting this aspect of Feasibility was created. Using these subcases enables more in-depth analysis on which aspects of feasibility are crucial for humans and which are irrelevant.

Previous research has been somewhat inconsistent in its approach to "coherence" or **Consistency**, leading to a lack of consensus regarding the precise definition, composition, and importance of the metric [8, 26, 45]. Coherence in the context of explanations was first laid out by Thagard in his seminal 1989 book, "Explanatory Coherence" [31]. In some recent research, the metric is subdivided into Internal Coherence ("The parts of the explanation fit together coherently") and External Coherence ("This explanation fits with what I already know") [26]. Consistency, as defined in this thesis, is equivalent to Internal Coherence. Different aspects of External Coherence are covered by other metrics, such as Feasibility and Trust, and do not warrant adding a separate metric. This approach is supported by the strong correlation between External Coherence and Perceived Truth, which corresponds to our defined metric of Trust, found by Zemla et al. [26]. Consistency in counterfactual explanations has been explored by Rasouli et al, who integrate it into their proposed framework for generating CEs by analysing correlations between features [6] and by Guidotti, who refers to it as Causality [46]. In addition, Consistency, a more widely understood term, was chosen to describe this metric, since Coherence was found to be confusing to many participants in the pilot study.

The changes that are evaluated for Consistency can be either independent, as when changing income and BMI, or have a causal relationship, such as when increasing weight also increases BMI. To properly evaluate consistency, "good" and "bad" explanations were created for both cases, with different combinations of continuous and categorical variables.

Completeness is one of the more difficult metrics to evaluate due to two main reasons: people have a tendency to automatically fill any logical gaps or leaps in an explanation [47,

48] and evaluating completeness can require significant domain knowledge [40, 49]. The former is an inherent function of human cognition [47, 48], but the latter can be mitigated to some extent. To achieve this, useful context was added to relevant questions to guarantee a similar baseline knowledge of the presented topic. Evaluating Completeness relies on grasping the causal relations between the original situation and the result, which is a core process when attempting to explain a result [49]. It is important to note that the definition of Completeness used in this thesis is distinct from the definitions presented in some literature, such as by Vilone and Longo [11], who use it to refer to how fully an explanation represents the underlying inferential mechanisms of the model.

One of the main goals of any XAI method is to increase users' **Trust** in a model [27, 50]. However, in the context of this study, we are measuring the user's trust in the effectiveness of the proposed changes, not in the model that is being explained. Since one of the core goals of an explanation is to inform future behaviour, it is necessary for users to believe that the proposed changes are valid and would bring about a change in outcome. While the truthfulness of an explanation is a necessary attribute, some research suggests it is not the most important basis on which people choose explanations [51]. This is easy to understand in the context of counterfactual explanations, since an explanation that is true but unfeasible is far less useful and therefore less valuable than an explanation that is both true and feasible.

In relevant literature, **Understandability** can refer to either the capacity of an explanation to explain the model's decision process to the user or to the ease with which the user understands the presented explanation. The latter definition, also labelled as Comprehensibility [40] or Readability [35] in some studies, is used in this questionnaire, due to a need to distinguish low satisfaction caused by aspects of the explanation itself, as opposed to low satisfaction due to difficulties understanding the presented explanation. The latter problem is not meant to be solved in the scope of this thesis and to minimise its effect on the results, a pilot study was conducted to evaluate how well our presentation of the counterfactual explanations is understood by users. Based on the results of the pilot study, necessary modifications and simplifications were made to our question structure to assist users in understanding the explanations, which can be seen in Appendix I and are further discussed in the following sections.

Discriminatory bias is one of the most frequently discussed issues of machine learning models and often arises from real-world biases contained in the training data [4, 17, 52]. Mitigating bias in machine learning models is an important goal in current and future AI development to ensure fair treatment for people of different backgrounds by automated systems. Counterfactual explanations and other XAI techniques can be powerful tools to uncover any biases present in the model. However, the goals of discovering biases in models and presenting good explanations are contradictory, since a biased explanation is unlikely to be useful or acceptable for most people. Since this study aims to explore the

qualities of a good counterfactual explanation, the capacity of an explanation to discover bias in models is not considered relevant. To evaluate the importance of **Fairness** in explanations, we included some explanations that changed features that are commonly viewed as conducive to bias, such as gender and relationship status.

Low **Complexity** is widely agreed upon to be a desired feature of explanations [27, 31]. Several studies have shown that humans prefer simple explanations when possible [32], although there is some conflicting evidence [26]. Low complexity of an explanation suggests a minimal number of causes used to explain the outcome. However, the minimal number of causes may not be equivalent to the optimal number to sufficiently explain the outcome. Therefore, we used a different scale for Complexity, with options for both overly simple and overly complex explanations, to capture the cases where more causes would improve the quality of the explanation. Additionally, instead of a 6-point scale, we used a 5-point scale for Complexity, since that enables having a middle or "ideal" value. Complexity is closely related to the technical metric of sparsity, which refers to the number of features changed in a counterfactual explanation [7].

2.2 Developing explanations

Once the evaluation criteria had been developed, a set of counterfactual explanations was created based on these metrics. For each metric besides Understandability, at least three explanations were created to cover extreme cases as well as a neutral case based on the evaluations of the researchers, in addition to the sub-cases mentioned above. Afterwards, a pilot study was used to evaluate whether all cases of all the metrics had been sufficiently covered, and additional explanations were created where necessary.

The explanations were originally based on a handful of common datasets, such as the Adult dataset [53] and the Pima Indians diabetes dataset [54]. These provided a set of features to manipulate and predictions to provide to the participants. This was used as a starting point and some subsequent explanations were generated using artificial data, since the extremes of some metrics could not be captured fully using only the features from these datasets. The features and predictions of each explanation were chosen based on the metric that was to be covered, since no single dataset was capable of providing sufficiently clear and unambiguous examples for all metrics. Additionally, using a single dataset would have limited the generalisability of the results. The wording of the explanations was standardised to help the user focus on the counterfactual scenarios and to avoid any confusion around the phrasing. A pilot study was used to simplify and improve the phrasings of the explanations. In total, 30 explanations were chosen to be a part of the final questionnaire. An example of the formatting and content of a question can be found in Appendix I.

2.3 Questionnaire structure

The environment used for creating the questionnaire was the survey creation tool LimeSurvey[4], provided by the University of Tartu (UT). This was chosen due to several factors, such as being able to store the participants data securely on UT servers and the availability of extensive customisation options.

Before taking the survey, participants were first presented with a consent form. This ensured that participants were aware of the goal and ethical ramifications of the study. If participants chose to take part in the study, they were asked to provide minimal demographic information, such as age and level of English proficiency, which could later be used for data analysis. As an introduction to the topic and the survey, the participants were then shown an example of a counterfactual explanation, along with definitions and simple examples for each of the metrics they would be asked to rate, as shown in Table 1 and Appendix II. The examples for the metrics consisted of one "good" and one "bad" explanation, and these were designed to be simple and unambiguous to avoid biasing the users' answers. The main part of the survey consisted of 30 counterfactual explanations, provided in a random order to ensure reliable results, since it was suspected that users might begin to pay less attention after seeing too many explanations. Each counterfactual explanation was shown on a separate page, with 8 rating scales below the question for the metrics. To ensure that participants can refer to them at any point in the questionnaire, the above-mentioned metric definitions and examples were present under the rating questions as well. Additionally, one of the questions contained an attention check, with instructions in the question text to type "here" in a textbox at the bottom of the page. The goal of the attention check was to test whether participants were paying attention and reading through the text carefully.

2.4 Pilot survey

In January of 2024, the Research Ethics Committee of the University of Tartu approved a small-scale survey with 29 questions, which was carried out among researchers of the University of Tartu. The goal of the pilot survey was to gather detailed feedback and preliminary data before carrying out the survey with a larger participant pool. In total, 15 participants filled out the entire survey and provided their feedback. Some of the participants suggested that the current phrasing of the questions was overly verbose and diverted attention from the explanations themselves. Based on this feedback, the text surrounding the explanations was significantly cut down and formatting was improved to help emphasise the explanations. Both the original and improved question structure can be seen in Appendix I. Additionally, some participants were confused by the metric Coherency, which led to renaming it to Consistency and making the definition clearer.

[4]Available at `https://survey.ut.ee`

Similarly, Bias was renamed to Fairness, to make it more in line with the rest of the metrics, which were defined as positive qualities.

Preliminary data analysis was carried out on the data, which confirmed that our explanations were able to cover the different metrics well. The only metric that required further examples was Complexity, where our current explanations failed to give an example of an overly complex explanation. For this, an additional explanation was created, which included 6 suggested changes. Finally, data analysis showed that there were significant correlations between some metrics, which fit with our predictions. Most metrics were correlated with satisfaction as well, which validated our choice of metrics. However, the sample size was far too small to make any conclusions based on the data analysis.

2.5 Conducting the survey

Due to changes made in the content of the survey, the survey needed to be reapproved by the Research Ethics Committee of the University of Tartu. Approval from the Ethics Committee enabled the survey to be distributed to a larger audience, while minimising any related risks to privacy.

An online research platform named Prolific[5] was used to distribute the questionnaire. Prolific is an online platform that aims to bring together researchers and a pool of diverse and active people willing to participate in studies, who get compensated for their efforts. Our goal was to obtain data from 100 participants, which would result in 3000 answers in total, with 24000 instances of metric evaluation. The survey was conducted in March 2024, for a total cost of 980 British pounds, approximately equivalent to 1150 euros.

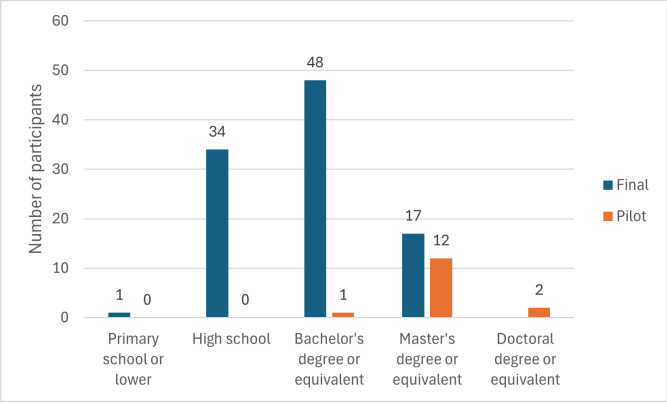

Figure 2. Level of education of survey participants

[5]Available at www.prolific.com

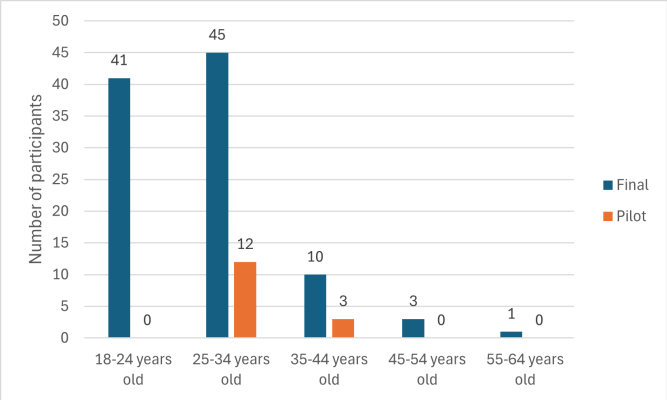

Figure 3. Age of survey participants

The survey was filled out by a diverse set of participants from 22 countries. Compared to the pilot study, the level of education of participants is well-distributed and contains many people both with and without higher education, as can be seen in Figure 2. The ages of the participants were somewhat less varied as a large majority of participants were younger than 35 years old, with just 14 participants older than 35 years old, as seen in Figure 3.

2.6 Data analysis and filtering

The raw data from the survey was first filtered to isolate any low-quality responses by participants. The methods used for this included the attention check found in the questionnaire, a handful of indicator questions for which a reasonable range of answers could be assumed, and clustering methods to detect outliers.

To enable the clustering of participants, dimensionality reduction methods were used, specifically Principal Component Analysis (PCA) [55] and t-distributed Stochastic Neighbour Embedding (t-SNE) [56]. PCA is a linear method that finds linear combinations of variables that can account for most variance in the data. This enables a meaningful distribution of data in lower dimensions than the original data. Subsequently, DBSCAN clustering [57] was performed on the lower-dimensional data to detect outliers, meaning participants whose answers differed significantly from the rest of the sample. A similar process was carried out using t-SNE, a nonlinear dimensionality reduction method, instead of PCA. A core goal of t-SNE is to map similar data points closely together, while more dissimilar data points are distributed further. This is highly useful for both clustering and visualisation, as isolated data points can be considered outliers. Any outliers that were found using these methods were investigated more thoroughly using a set of indicators, including but not limited to how long they took to complete the

survey and whether they successfully completed an attention check question present in the questionnaire. In addition, there were a handful of questions in the questionnaire where a reasonable range of answers could be assumed. For example, one counterfactual explanation suggested changing the country you were born in, which can reasonably be considered an infeasible change. Therefore, if a participant rated that explanation as highly feasible, their answers were considered more suspicious. Combining all of the aforementioned indicators, a total of 9 out of 100 participants' answers were considered low-quality data and excluded from the dataset.

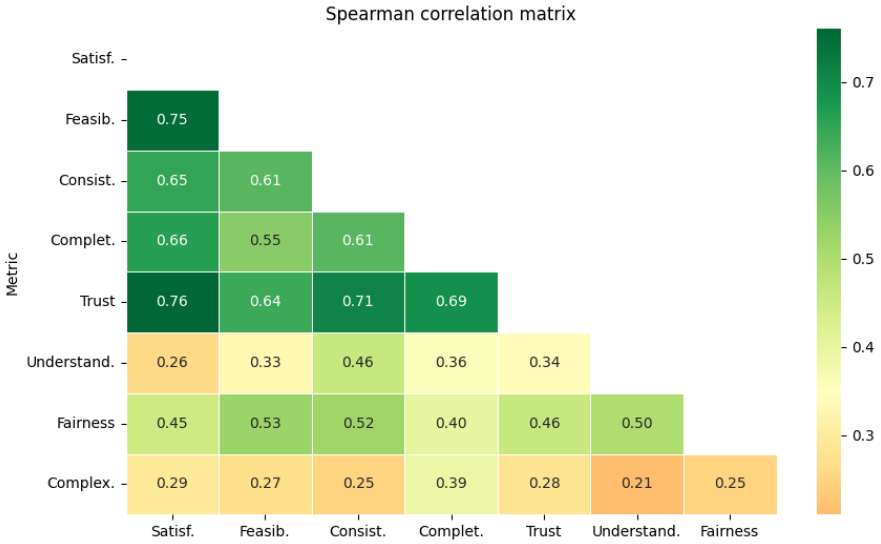

Figure 4. Pairwise correlations between metrics. All of the correlations had p-values below the significance threshold, which was determined using the Bonferroni correction method. Adapted from the work of Veski [38].

As seen in Figure 4, most of the metrics correlated strongly with Overall satisfaction. This is expected, since the metrics are designed to capture aspects of an explanation that render it preferable for users. The three metrics that had considerably weaker correlations were Understandability, Fairness and Complexity. The correlation between Understandability and Overall satisfaction is expected to be low, due to the fact that most bad explanations are still understandable. Fairness and Overall satisfaction may have been weakly correlated due to the scarcity of examples with low Fairness in the questionnaire and participants' varied opinions on what is fair. Zemla et al. [26] have previously shown that people desire explanations that are sufficiently complex, whereas other research has suggested simplicity as a core value of good explanations [31]. This is reflected in the low correlation between Complexity and Overall satisfaction observed in our survey, since it suggests that neither overly complex nor overly simple explanations

were preferred. For further analysis of the correlations between other metrics, see the work of Veski [38].

3 Methods for Large Language Model experiments

With the questionnaire data as the input dataset, this thesis aimed to test and fine-tune Large Language Models to evaluate counterfactual explanations automatically. The models selected for this were Llama 2[6], Llama 3[7], Mistral-7B[8] and GPT-4[9]. GPT-4 was accessed using the OpenAI API[10] and the rest of the models were accessed and fine-tuned using the University of Tartu High Performance Computing (UTHPC) Centre's Rocket cluster [58] and the *transformers* library by Huggingface [59].

3.1 Large Language Models

Language models are a type of machine learning model that can process and generate text. Large Language Models (LLMs) are a new paradigm in this field, as they are many times larger than traditional models and are able to generate fluent and coherent text in many languages. Current Large Language Models use an architecture known as a transformer. Transformers are neural networks that learn to generate sequential data with an attention mechanism [60].

Llama 2 is an open-access Large Language Model (LLM) family developed by Meta in 2023 as a successor to its previous Llama 1 model. It consists of models ranging from 7 billion parameters to 70 billion parameters. Parameters refer to the tunable weights present in a neural network and are often used as a rough estimate of a model performance and potential. Llama 2 is one of the most widely used LLMs, since it is free for both research and commercial use with the Apache 2.0 licence, and strikes a good balance between performance and size [61]. In April of 2024, Meta released the successor to Llama 2, Llama 3 [62]. Similarly to Llama 2, Llama 3 contained models with 8 billion parameters and 70 billion parameters. A significantly larger 400 billion parameter model was announced, but as of the writing of this thesis, this model is yet to be released. Meta claimed that despite the Llama 3 models' similar sizes to Llama 2, the new models are trained on 7 times more data than Llama 2 [62]. According to Meta, this has led to significant performance gains in a wide variety of tasks [62]. In this thesis, both the older Llama 2 7 billion parameter model (7B) and the newly released Llama 3 8 billion parameter model (8B) were used and compared, in addition to the significantly larger 70 billion parameter Llama 3 model (70B). For both Llama 2 and Llama 3 models, a version that is fine-tuned for instructions was used instead of the base model, as specified by the suffix "Chat" or "Instruct".

[6]Meta, 2023. Llama 2: `https://llama.meta.com/llama2/`

[7]Meta, 2024. Llama 3: `https://llama.meta.com/llama3/`

[8]Mistral AI, 2023. Mistral 7B: `https://mistral.ai/technology/#models`

[9]OpenAI, 2023. GPT-4: `https://openai.com/research/gpt-4`

[10]More information available at `https://openai.com/blog/openai-api`

Mistral 7B is a popular open-source LLM released by Mistral.ai in 2023, which at release claimed to be the most powerful 7 billion parameter model to date. According to testing carried out by Mistral.ai, it outperformed Llama 2 13B in all benchmarks, despite the latter having almost twice as many parameters [63]. A version of Mistral 7B fine-tuned for instruction tasks was released as well, called Mistral 7B Instruct, which was used in this thesis. Both models are free for research and commercial use based on the Apache 2.0 licence [63].

As of May 2024, GPT-4 is the most powerful LLM from OpenAI, which is largely responsible for the new wave of AI research and public interest due to their public release of ChatGPT in 2022. GPT-4 is widely considered to be the leading LLM in the world as of the writing of this thesis [64]. However, it is a proprietary model and therefore, the details of its structure, such as the number of parameters, are not publicly available. Regardless, it has shown industry-leading performance on a wide range of benchmarks [65].

For the experiments in this thesis, all of the abovementioned models, besides GPT-4, were sourced from Huggingface's repositories through the *transformers* library. GPT-4 was accessed using OpenAI's API service.

3.2 Efficient fine-tuning of LLMs

As models based on neural networks, LLMs' primary structure consists of billions of tunable parameters, also known as weights. Consequently, the training process for such models tunes these parameters to achieve a set of values with the maximum predictive power on the training data. Fine-tuning refers to a process of using an already trained model and tuning the parameters further to optimise their values for specific tasks.

Fine-tuning LLMs can be a computationally expensive process due to the need to alter billions of parameters at a time. To achieve this, all of the original parameters need to be stored in memory at the same time as the alterations to every parameter. For a model with 65 billion parameters, this process can require more than 780 GB of GPU memory and significant processing time [66]. In this thesis, resource constraints were alleviated in two ways: by using the University of Tartu's High Performance Computing (UTHPC) [58] clusters to fine-tune the models using several Nvidia V100 GPUs, and by reducing the load on GPU memory using a technique known as QLoRA. The general process carried out in this thesis is illustrated in Figure 5.

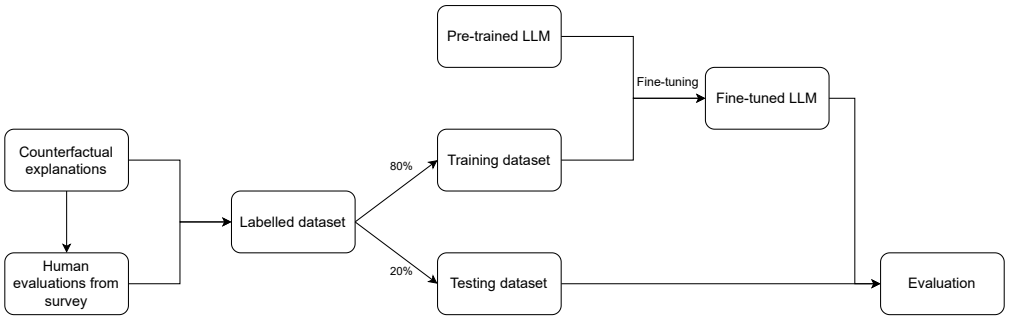

Figure 5. Illustration of the fine-tuning process in this thesis.

LoRA or Low-Rank Adaptation is a commonly used method for significantly reducing the memory requirements for fine-tuning LLMs [67]. The primary way this is achieved is through "freezing" the original weights and adding a separate set of lower-rank weights that are fine-tuned on the specific task at hand. For example, instead of fine-tuning a full $M \times N$ weights matrix (lets call it X), LoRA instead creates two significantly smaller $M \times K$ and $K \times N$ matrices A and B, which approximate the full matrix such that $AB \approx X$, where the resulting matrix has a shape of $M \times N$. K refers to the "rank" used by LoRA, which is a tunable hyperparameter. This relies on previous research that has shown that large pre-trained models often have low "intrinsic dimension" and therefore can be effectively fine-tuned with matrices of lower dimensions [68]. LoRA is a widely used method that is supported by common machine learning libraries, such as Huggingface's *transformers* [59].

A more recent extension of LoRA is called QLoRA [66]. This method retains all the benefits of LoRA and adds additional optimisations to further decrease GPU memory use with minimal performance loss. These mainly rely on the process of quantisation, which discretises an input from a higher precision format, usually one with more bits per parameter, to a lower precision format, while preserving as much information as possible. QLoRA achieves this by introducing a novel data type called NF4, which is theoretically optimal for normally distributed data and only uses 4 bits per parameter, instead of the more common 16 or 32 bits. Furthermore, QLoRA enables the use of Double Quantisation, which quantises the variables ("quantisation constants") used to quantise the original parameters, reducing memory use further. This purports to reduce the GPU memory required to fine-tune an LLM by a factor of 10 or more, compared to typical full-parameter fine-tuning, while retaining comparable performance [66].

Despite these techniques, fine-tuning an LLM still requires considerable GPU resources. This problem was solved by using computing clusters from the High Performance Computing Centre (UTHPC), a consortium of the University of Tartu [58]. Its primary function is to maintain and develop the infrastructure needed for scientific computing.

Among other projects, it provides the Rocket HPC cluster for running complex scientific computations. The Rocket cluster contains different partitions for different tasks, some of which are meant for GPU-intensive tasks, such as those relating to LLMs. One of these partitions, named “falcon”, is composed of 6 nodes with 24 NVIDIA Tesla V100 32GB GPUs in total. For the experiments in this thesis, a “falcon” node with two NVIDIA V100 32GB GPUs was used to run and fine-tune the smaller models, while the larger Llama 3 70B Instruct model warranted the use of four NVIDIA V100 32GB GPUs.

4 Experiments

Using the methods described in the previous section, five experiments were conducted. As a prerequisite, dataset preparation and prompt engineering was carried out. Subsequently, the goal of the first experiment was to compare the effects of different LLM prompts on the accuracy of evaluating counterfactual explanations. The second experiment aimed to study the application of fine-tuning LLMs to counterfactual explanation evaluation. The third experiment was used to validate the results of the second experiment. The fourth experiment explored using specific participants' answers instead of sample means. Finally, the results were compared to GPT-4, a cutting-edge LLM from OpenAI. The code used for the experiments in this thesis can be found in Github[11].

4.1 Dataset preparation

After the questionnaire responses had been gathered and filtered, further data processing was needed to create a useful dataset. For each of the question-metric pairs, the average of the 91 remaining participants' answers was used as the final value. Importantly, as Complexity used a different scale of -2 to 2 instead of 1 to 6, the values for Complexity were scaled to match the scale of the other criteria. This was done using the formula $x * 1.25 + 3.5$, where x is the value for Complexity. To minimise the effects of the scale used to rate metrics and make the results more generalisable, all values were then mapped to a string, with values below 3 as "low", values between 3 and 4 as "medium" and values above 4 as "high". With 30 questions and 8 metrics per question, this resulted in 240 instances in total.

For evaluating the models, a subset of 48 instances was set aside as a test dataset and excluded from fine-tuning. This dataset was designed to include examples of all metrics in equal amounts, 6 instances per metric, and to provide at least one instance with a "high", "medium", and "low" answer for every metric. Importantly, different metric evaluations for the same question were sometimes divided between the training and the testing datasets, since it was hypothesised that different metrics could be evaluated independently.

In order to fine-tune Large Language Models, the data needs to be presented in a format specific to the LLM in question. Generally, an instance of training data contains three things: a system prompt, an instruction, and an output. A system prompt gives the LLM general information on how to approach any tasks it is presented with. For example, a system prompt may contain information such as "you are an AI language model" or "avoid giving false or misleading information". An instruction usually contains the task that the model is meant to perform. The output is the most important part

[11] Available at `https://github.com/JuliusValja/CF_Thesis_2024`

when fine-tuning a model, since it contains the response that the LLM should learn to provide.

There are two possible ways to use LLMs to evaluate counterfactual explanations. Firstly, text generation could be used to have the LLM answer the question, similar to how a human would. Alternatively, the problem could be considered as a classification task, and the LLM could be used to classify text into different classes, with different classes corresponding to high, medium or low values in this case. Both approaches have their merits and problems, but due to the small dataset, text generation was considered the superior option and used for this thesis.

4.2 Prompt engineering

To achieve the best possible performance from an LLM, it is crucial to find the right prompt, a process known as "prompt engineering". For the task of predicting metrics based on explanations, 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. All changes were made in what is known as a "system prompt", or a general prompt that provides the LLM necessary context to carry out any instructions. For chatbots, a common system prompt might contain phrases like "You are a friendly chatbot" or "Do not answer any questions that may propagate bias or harm". For this task, 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 the metrics in the questionnaire, as seen in Appendix II. 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 one of the metric evaluation questions, such as "Please rate as 'low' (very unfeasible), 'medium' or 'high' (completely feasible), how feasible is this explanation:". Consequently, each

factual-counterfactual pair resulted in 8 instances, one for every metric under evaluation. Examples of all three prompts can be found in Appendix III.

4.3 Experiment I: Prompt comparison without fine-tuning

The first experiment consisted of comparing the three formulated prompts on their accuracy. To comprehensively compare the prompts, Llama 3 70B Instruct model and three smaller models were used, namely Mistral 7B Instruct, Llama 2 7B Chat, and Llama 3 8B Instruct. The models were given the testing dataset as input, and their provided responses were compared with the "true" value, which was based on the responses from the survey. Each prompt-model combination was evaluated 4 times and the average accuracy of the 4 runs is presented in Table 2.

Table 2. Accuracies for different prompt-model combinations in the first experiment. The highest accuracy for each model is highlighted in bold.

Model	Base prompt	With all definitions	With examples
Mistral 7B Instruct	0.40	**0.41**	0.36
Llama 2 7B Chat	**0.46**	0.44	0.37
Llama 3 8B Instruct	0.56	**0.63**	0.55
Llama 3 70B Instruct	0.72	0.70	**0.75**
Average	**0.54**	**0.54**	0.51

The results of the experiment showed no consistent differences between the prompts, as each of the three prompts was the best performing choice for at least one of the models. The lowest average accuracy of 51% was seen for the prompt with examples, with the other two prompts achieving similar average accuracy scores of 54%. This result is somewhat unexpected, since it appears to suggest that “zero-shot” learning achieves better results than “few-shot” learning for this task, which contradicts the common view that "few-shot" learning is superior.

This discrepancy between the results of the experiment and generally accepted truths likely stems from one of four possible causes. The first possible cause is that the examples chosen for the prompt were simply not representative of the overall data and therefore did not provide useful new information to the model. The second possibility is that while the examples contained two metrics, the data contains eight metrics. Hence, it may be that the model would have performed better if the prompt provided examples for every metric, not just some of them. Thirdly, the 4 runs that the evaluation was based on may have been insufficient to fully account for the model’s variability, as the differences between the average accuracies as small and could still be an artifact from random variability. Finally, the definitions of every metric presented alongside the examples may have primed the model to overvalue certain metrics, which led to a lower overall accuracy across all

metrics. However, this possibility is the least likely, since the prompt that also contained all the definitions, but without examples, performed better.

Overall, larger LLMs seem to be significantly more capable of evaluating the quality of counterfactual explanations by default. The highest accuracy achieved by the smaller models was 63%, whereas the lowest accuracy of the larger model was 70%. This suggests that the size of the model plays a significant role in the model's capacity to understand and evaluate counterfactual explanations. This is an expected result, since LLMs' performance on a wide range of benchmarks improves greatly with increased scale [64]. It is important to note that the 63% accuracy mentioned above was achieved by Llama 3 8B Instruct, a cutting-edge model released in April 2024. Both of the similar older models failed to reach 50% accuracy with any of the prompts. This suggests that in addition to scale, advancements in model architecture and training data can have a significant impact as well, making small and fast models increasingly more viable in the near future.

4.4 Experiment II: Effects of fine-tuning LLMs on evaluation accuracy

The second experiment was designed to test the efficacy and expedience of fine-tuning LLMs to evaluate counterfactual explanations. For this, the set of four models from the previous experiment was used. Based on the results of the previous experiment, this experiment was run using the "base prompt", without examples and extra definitions, since none of the additions provided measurable accuracy improvements.

The models were fine-tuned using the UTHPC Centre's Rocket cluster [58]. The optimal hyperparameters for every model were discerned through extensive testing and can be viewed in Table 3. All models were fine-tuned using a completion-only data collator from Hugginface's *trl* library [69]. This means that the models were only fine-tuned to predict the answers to the questions, not the text of the questions themselves. This type of data collator was chosen to focus on improving the predictive performance of the models. With a typical language modelling data collator, the model would have learned to predict the question text as well, but this was unnecessary for the task at hand.

Table 3. Hyperparameter values used for fine-tuning models.

Model	General			LORA	
	Batch size	**Learning rate**	**Epochs**	**r**	**alpha**
Mistral 7B Instruct	4	0.0002	2	16	32
Llama 2 7B Chat	4	0.0002	3	32	64
Llama 3 8B Instruct	4	0.0002	4	32	64
Llama 3 70B Instruct	8	0.00005	5	32	64

Table 4. Evaluation accuracy for different models in Experiment II, with and without fine-tuning. The highest accuracy for each column is highlighted in bold.

Model	All metrics		Overall satisfaction	
	Baseline	Fine-tuned	Baseline	Fine-tuned
Mistral 7B Instruct	0.40	0.74	0.42	0.50
Llama 2 7B Chat	0.46	0.69	0.33	0.63
Llama 3 8B Instruct	0.56	0.78	0.33	0.58
Llama 3 70B Instruct	**0.72**	**0.91**	**0.5**	**0.96**

The accuracies of both baseline and fine-tuned models over all metrics can be seen in Table 4. Similarly to the first experiment, the accuracy of a model was calculated based on the average of 4 evaluation runs. This ensured that the natural variability and probabilistic nature of LLMs does not skew results significantly.

Additionally, accuracy when evaluating Overall satisfaction (or for brevity, Satisfaction) was tested separately as well, the results of which can be seen in Table 4. This was done using the same fine-tuned models as when measuring the overall accuracy, which were trained on the entire training dataset. The main reason this approach was chosen was the very limited size of the dataset when only including Satisfaction. In that case, the training dataset would have contained 24 instances, which proved to be insufficient for any meaningful learning. However, the models trained on the full dataset learned to evaluate satisfaction to some extent as well. This problem could be solved by gathering a dataset with significantly more explanations, but in this thesis that option was not feasible.

The second experiment convincingly showed that relatively small LLMs can be trained to be as capable as larger pre-trained models, even with a very limited dataset, such as the one used in this thesis. The most accurate small model was Llama 3 8B Instruct, which reached 78% accuracy after fine-tuning, narrowly beating out the larger model from the same family, which achieved 72% accuracy without any fine-tuning. This is a

promising result, since smaller LLMs are far superior in terms of both speed and resource requirements, enabling them to be used in wider contexts.

However, the gains seen through fine-tuning a larger LLM were equally significant, with the larger Llama 3 70B Instruct model reaching an impressive 91% accuracy on the testing dataset after 5 epochs of fine-tuning. However, it is important to note that this may not be purely a reflection of the real-world performance of the model. As presented in Figure 4, the analysis of questionnaire data revealed significant correlations among some metrics, which could assist a model in predicting one metric based on the value of another.

The largest improvement from fine-tuning was seen with the Mistral 7B Instruct model, which soared from an average overall accuracy of 40% to an overall accuracy of 74%, a 34% jump. The other smaller models showed improvements on a similar scale, with Llama 2 7B Chat and Llama 3 8B Instruct improving by 23% and 22%, respectively.

The observed Satisfaction accuracy of baseline models was significantly lower than the overall accuracy. Several of the models only reached accuracies around 33%, which is equivalent to the accuracy one would achieve by randomly guessing. Fine-tuning had a moderate positive effect on the smaller models, all of which managed to reach 50 to 60 percent accuracy. While the largest model, Llama 3 70B Instruct only managed a 50% accuracy for satisfaction at baseline, through fine-tuning it reached a remarkable 96% accuracy. An increase of this magnitude was unexpected, and may be partially related to the correlations between different metrics and satisfaction. To test this hypothesis, an additional experiment was carried out.

4.5 Experiment III: Validation using a question-based dataset

To assess whether the gains seen when fine-tuning LLMs were due to meaningful learning or learning correlations between metrics, two of the highest scoring LLMs, Llama 3 8B Instruct and Llama 3 70B Instruct, were fine-tuned again, with the hyperparameters shown in Table 3. Once again, the testing accuracies were averaged over 4 runs. This time, the training and testing dataset were constructed so they contained entire counterfactual-metric sets. This means that there were no instances where one metric from a specific question was in the training dataset and another was in the testing dataset. Therefore, the effects of metric correlations on accuracy were effectively removed.

The testing dataset contained 6 hand-picked questions. The questions were picked to enable the dataset to satisfy certain attributes. Most importantly, the dataset needed to be reasonably representative of the underlying data. In the initial design of the questionnaire, each question was crafted to assess a specific metric, typically to elicit either a positive or negative evaluation for the metric. This information was now used to select questions for

the testing dataset, so that each question covers a different metric and that some of them are positive examples and some negative. In conclusion, while the original testing dataset contained a representative set of question-metric pairs, the new dataset was designed to contain a set of questions, with all of the corresponding metrics.

Table 5. Evaluation accuracy for question-based testing set in Experiment III

Model	Baseline	Fine-tuned	Fine-tuned, testing Satisfaction only
Llama 3 8B Instruct	0.51	0.67	0.67
Llama 3 70B Instruct	0.64	0.76	0.83

As seen in Table 5, the baseline results on this dataset were somewhat lower, particularly for the Llama 3 70B Instruct model. This suggests that this testing dataset was inherently more difficult to evaluate than the one used in the previous two experiments. As such, the expected accuracies after fine-tuning were lower as well. This was confirmed by the results, where both models scored at least 10% lower than in the previous experiment. However, the gains in accuracy for the 8B and 70B models are 16% and 12%, respectively, which is a significant increase. This suggests that the improvements in accuracy in the previous experiment cannot be fully explained by correlations between metrics, and that meaningful learning was achieved by the LLMs.

Importantly, the accuracy when predicting Overall satisfaction was equal to or higher than the overall accuracy. Whereas the 70B model displayed somewhat lower accuracy than in the previous experiment, the 8B model was considerably more successful in this experiment, achieving similar accuracy on Overall satisfaction as on the rest of the metrics. Importantly, the 83% accuracy of the 70B when evaluating Overall accuracy is still remarkably high and significantly higher than that of any other model. This confirms the result that large models are superior for evaluating counterfactual explanations. However, this result does suggest that the unexpectedly high 96% accuracy reached in the previous experiment may have been affected by metric correlations.

4.6 Experiment IV: Evaluation on specific people

Different people's preferences for explanations can exhibit significant variability. To explore the effects of this, an experiment was carried out with a dataset based on specific participants' answers, instead of the sample averages. To ensure that these participants represent different subgroups of participants, t-SNE [56] was used to reduce the dimensionality of the data and DBSCAN clustering [57] to cluster the results. DBSCAN was considered the optimal clustering algorithm, due to its ability to work with an unknown number of clusters and its robustness towards noise [57]. The goal of clustering was to discern the largest clusters present in the data, which can be seen in Figure 6. A random

participant was chosen from each of the four largest clusters, marked in Figure 6 with red.

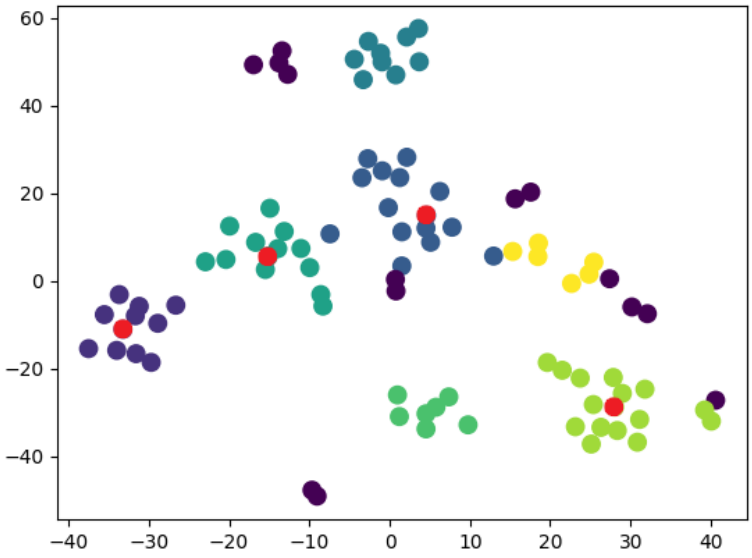

Figure 6. DBSCAN clustering used on t-SNE dimensions

As seen in Table 6, the chosen participants have distinct backgrounds. Each of the participants was from a different country, three of which are in Europe and one in South America. To protect the privacy of the participants, the specific countries are not displayed in the table. Additionally, different levels of English proficiency and education are represented. Some of the participants also have experience with the field of machine learning, whereas others have none. This increases the likelihood that the responses these participants provided were sufficiently different and came from different viewpoints.

Table 6. Demographic information of chosen participants

Participant	Age	English proficiency	Education	Experience with machine learning
A	18-24 years	Moderately proficient	High school	Some experience
B	18-24 years	Moderately proficient	Bachelor's degree or equivalent	No experience
C	18-24 years	Native speaker / Fully proficient	Bachelor's degree or equivalent	No experience
D	25-34 years	Native speaker / Fully proficient	Master's degree or equivalent	I am studying in a related field

Table 7. Evaluation accuracy for different participants in Experiment IV

Participant	All metrics		Overall satisfaction	
	Baseline	Fine-tuned	Baseline	Fine-tuned
A	0.71	0.68	0.67	0.42
B	0.64	0.74	0.33	0.42
C	0.77	0.84	1.0	0.88
D	0.66	0.68	0.66	0.66

For each of these participants, baseline evaluation and fine-tuning was carried out using the same procedure as in Experiment II, but using only the model Llama 3 70B Instruct, as it proved to be the most capable in previous experiments. The final hyperparameter values used in this experiment can be viewed in Table 3. Alternative hyperparameter values were tested, but these showed no further accuracy improvements. The testing dataset contained the same question-metric pairs as in Experiments I and II, but with answers from the specific participant.

The results of this process were varied. At baseline, it was clear that the LLMs answers did not match different participants' answers equally. As seen in Table 7, over all metrics, the LLMs predictions were reasonably in accordance with different participants answers. However, there were major differences when evaluating Overall satisfaction, with accuracies ranging from 33% to 100%. This leads to two important conclusions.

For one, it appears the LLM's biases and preferences match with certain participants significantly more than others, which has important implications when attempting to use LLMs in the real world. Secondly, it indicates that participants have highly varied opinions on which explanations are good or good enough, which may significantly impair attempts to generate reliably good explanations without user input. However, since this comparison only contained 4 participants and 30 explanations, these conclusions should be considered tentative.

The effects of fine-tuning in this experiment were inconsistent, compared to previous experiments. For all participants except Participant A, the accuracy over all metrics showed minor improvements. However, the gains ranged from 2% to 10%, which is in conflict with the significantly larger gains seen in Experiments II and III. For Participant A, the accuracy over all metrics decreased, and did so for all hyperparameter values tested. Learning to predict Overall satisfaction proved even less effective, with 3 out of 4 participants' scores decreasing after fine-tuning. This suggests that specific humans are not necessarily consistent in their evaluations, but over larger groups, the average values seem to be fairly consistent.

4.7 Experiment V: Comparison with GPT-4

To further supplement the results obtained in this thesis, the leading LLM as of May 2024, OpenAI's GPT-4, was tested [65, 70]. Compared to the models tested previously, GPT-4 has a significantly larger context length of 32 thousand tokens, which refers to the length of the input it can process at one time. To take full advantage of this, the model was tested using few-shot learning, by giving the model 30 examples from the training dataset, followed by evaluation on the testing dataset. The sets used in this experiment were identical to those used in Experiments I and II.

Evaluation over all metrics using the GPT-4 model resulted in an accuracy of **73%**. This is largely in line with the baseline accuracy provided by Llama 3 70B Instruct. This suggests that the gains seen when increasing model complexity provide diminishing returns. Going from an 8 billion parameter model to a 70 billion parameter model provided tangible benefits, but upgrading to GPT-4, which is estimated to consist of more than a trillion parameters [71], provided no further accuracy improvements. It is likely that the accuracy of models is limited by the size of the dataset, both in the amount of training data available and the lack of variety in testing data. Additionally, it supports the result from Experiment I, which suggested that few-shot learning was not significantly superior to zero-shot learning for the task of evaluating counterfactual explanations.

5 Discussion

The experiments of this thesis yielded many interesting results. The experiments aimed to explore the potential of LLMs in assessing different qualities in counterfactual explanations. The effects of fine-tuning and metric correlations were tested, and different prompts and models were compared.

5.1 Interpreting the results

For the main research question of this thesis, whether LLMs can be used to evaluate counterfactual explanations similarly to humans, strong evidence was found that such automation is effective and accurate. This was supported by the significant performance improvements seen in Experiment II, which explored the effects of fine-tuning on evaluation accuracy. The improvements seen suggest that fine-tuning LLMs is a viable path towards automatic evaluation of counterfactual explanations. Experiment III validated this result further, showing that the performance gains achieved through fine-tuning cannot be solely explained by the correlations between metrics. Inevitably, when interpreting the results of the experiments, it is necessary to take into account the inherent variability of human preferences. As a result of this, LLMs aiming to simulate human evaluations cannot reasonably achieve 100% accuracy, and should be approached differently from models based purely on objective data. This was further explored in Experiment IV, where an LLM attempted to evaluate explanations similarly to specific respondents. The results of this were varied, with signs that the LLM's inherent biases agreed with some participants more than others. However, larger experiments with more participants are needed to make any major conclusions.

Experiment I provided evidence that large LLMs can already evaluate counterfactual explanations to some extent, without any further fine-tuning, which helps answer research question Q1, whether LLMs already answer like humans without fine-tuning. According to the results, larger LLMs of 70B parameters or more are quite capable at evaluating counterfactual explanations using zero-shot learning, whereas smaller models are generally insufficient without further fine-tuning.

The second experiment, which explored fine-tuning LLMs, illustrated a surprising result, namely that evaluating Overall satisfaction is more difficult than the rest of the metrics. This was unexpected, since Overall satisfaction is, as its name suggests, a measure of the general quality of the explanation, whereas the rest of the metrics were more specific and nuanced. This result appears to contradict common views on human cognition, where it is often easier to give an indication of general quality, rather than specific positive or negative aspects [72]. However, Experiment III, which used a different testing dataset, showed that the previous results were more likely a result of the specific dataset and not a general trend. Here both models achieved similar or higher accuracies for Overall

Satisfaction than for other metrics. Therefore, research question Q2, whether evaluating user satisfaction is easier than other metrics, requires further research to fully answer, although the results of this thesis suggest that general satisfaction is not significantly easier to evaluate than the other metrics.

The scale and architecture of different LLMs had a clear impact on results throughout experiments. The Llama 3 70B Instruct model, which is significantly larger than the rest, consistently outperformed the smaller models. This suggests that the increased ability to understand context and meaning displayed by large LLMs aids in evaluating counterfactual explanations. However, the results from GPT-4 show that model size can only improve results up to a certain point. This suggests an optimal size of model for this task, which is large enough to accurately assess explanations, yet small enough to run in reasonable time and with reasonable hardware requirements. Alternatively, Experiment II showed that even smaller models, such as Llama 3 8B Instruct, can be fine-tuned to achieve performance close to much larger models. However, Llama 2 7B Chat failed to match its newer counterpart, scoring nearly 10% lower on accuracy, even after fine-tuning. This suggests that architectural and training differences have a measurable impact on accuracy as well. Overall, it can be said that the differences between different LLMs were consistent and significant, thus answering research question Q3.

5.2 Limitations

Due to the limited financial resources and time available in the creation of this thesis, several compromises needed to be made.

Firstly, the dataset that was created for this thesis is limited in both scope and scale. To achieve reliable results, each counterfactual explanation was hand-crafted to collectively cover all of the metrics. This was a time-intensive process, which limited the total number of questions to just 30. As a result of this, despite gathering data from 100 participants, the dataset amounted to 240 instances, with only 30 different explanations to process. This is clearly a small dataset to fine-tune a model with billions of parameters. This problem was especially evident when focusing solely on satisfaction, since the dataset only contains a total of 30 examples of satisfaction evaluation. This may have played a part in the exceptional accuracy Llama 3 70B Instruct reached when evaluating satisfaction. Although the testing was run for 4 iterations, which resulted in different accuracies, it may be that the 6 examples of satisfaction evaluation in the testing dataset were not enough to provide reliable and representative results.

Secondly, the selection of Large Language Models used for fine-tuning consisted of only two sizes of models, ones with 7 to 8 billion parameters and one with 70 billion parameters. Even the 70 billion parameter model is fairly small in 2024, and with the emergent capabilities demonstrated by the largest current models, using these might

provide novel data and greater strides in automating counterfactual explanation evaluation. However, fine-tuning larger models requires significant computational resources, and not all proprietary models are available for fine-tuning.

Thirdly, it is important to note that the metrics used in this thesis were not all equally difficult to rate. For example, Understandability can reasonably be evaluated without understanding the more nuanced meaning and relationship between the proposed changes, and Fairness can often be evaluated simply by checking whether a certain small subset of features, such as gender, has been modified. In contrast, evaluating Feasibility requires a significantly more in-depth analysis of the proposed changes in relation to the original values. Therefore, the ease of evaluating certain metrics may have contributed to deceptively high accuracy values at times. However, most of the metrics are not trivial to evaluate, therefore this issue does not significantly undermine the conclusions presented in this thesis.

Despite these limitations, the experiments performed in this thesis showed potential and suggested that fine-tuning Large Language Models may be a viable strategy for automating the evaluation of counterfactual explanations.

5.3 Implications and potential ways forward

While this thesis has provided several promising results, it can be considered mainly exploratory research. To further test and confirm these results, a more expansive dataset is necessary. A possible way to achieve this is now proposed. Firstly, a set of explanations that cover all of the metrics must be constructed, similarly to the process carried out in this thesis. Additionally, machine learning models must be trained for a diverse set of datasets, which can then be used to generate counterfactual explanations. Several different cutting-edge CE algorithms should be used for this to ensure diversity in proposed changes and avoid the specific biases inherent to any single algorithm. Finally, these CE-s can be presented in a questionnaire in small random subsets to participants, since a single participant can only evaluate a limited amount of explanations. This process requires significant time and funding, which this thesis did not have access to, but could result in more useful and generalisable data.

If the results of this thesis are replicable, it will open up numerous potential avenues for further research. A major limitation of algorithms that seek to generate CE-s is their inability to fully take into account human preferences and biases. Since LLMs have proven to be capable of simulating human preferences, they could prove an invaluable tool. The potential utility of LLMs is twofold. Firstly, LLM evaluation could be integrated into a system for providing counterfactual explanations. As part of such a system, it could evaluate a number of proposed counterfactual explanations and seek out the ones most in line with user expectations. This could result in automated explanations that are

useful and comprehensible for users, which is a long sought-after goal. Alternatively, LLMs could be used to evaluate and compare different algorithms for CE generation. Currently, it is difficult to compare and rank different algorithms, but LLMs could solve this problem. For example, a standardised test set of instances and models could be used to generate an extensive set of counterfactual explanations using both algorithms. LLMs could then evaluate the explanations from either algorithm and the results could be compared to analyse the strengths and weaknesses of each algorithm.

Either of these paths would aid in developing systems that could be implemented in real-world settings to provide reliable and practical explanations to end-users. This would increase the transparency and trustworthiness of any model, especially in high-risk fields, such as medicine, criminal justice, etc. Additionally, this would help minimise the legal and ethical risks that coincide with the use of "black box" models.

Conclusion

In this thesis, the author set out to explore the possibilities of automatically evaluating counterfactual explanations. To this end, a survey was created and distributed to gather a novel dataset of human-provided ratings of counterfactual explanations. The survey contained 30 counterfactual explanations, each of which was rated on 8 qualitative metrics. This dataset was further used to explore the possibility of automating counterfactual explanation evaluation through the use of LLMs.

The results showed that large LLMs are capable of evaluating counterfactual explanations, while smaller models can be fine-tuned to achieve comparable performance. Fine-tuning large models results in accuracies of 80% to 90%, which is sufficient for many real-world applications. In addition, it was shown that evaluating overall satisfaction is not significantly easier than evaluating other metrics and that matching the preferences of specific people is more difficult than matching the average preferences of humans. These results lay the groundwork for further experiments and research, both in further improving automated evaluation of counterfactual explanations and in improving existing explanation generation algorithms with LLM integration.

List of references

[1] K. Stubbs, P. Hinds and D. Wettergreen. Autonomy and Common Ground in Human-Robot Interaction: A Field Study. *Intelligent Systems, IEEE*, vol. 22, pp. 42–50, 2007. DOI: 10.1109/MIS.2007.21.

[2] T. Loftus, B. Shickel, M. Ruppert, J. Balch, T. Ozrazgat Baslanti, P. Tighe, P. Efron, W. Hogan, P. Rashidi, G. Upchurch and A. Bihorac. Uncertainty-aware deep learning in healthcare: A scoping review. *PLOS Digital Health*, vol. 1, 2022. DOI: 10.1371/journal.pdig.0000085.

[3] A. Caliskan, J. J. Bryson and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, vol. 356, no. 6334, pp. 183–186, 2017. DOI: 10.1126/science.aal4230.

[4] S. Wachter, B. Mittelstadt and C. Russell. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 2018. arXiv: 1711.00399. (visited on 13/03/2024).

[5] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez and F. Herrera. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, vol. 99, p. 101 805, 2023. DOI: 10.1016/j.inffus.2023.101805.

[6] P. Rasouli and I. Chieh Yu. CARE: coherent actionable recourse based on sound counterfactual explanations. *International Journal of Data Science and Analytics*, vol. 17, no. 1, pp. 13–38, 2024. DOI: 10.1007/s41060-022-00365-6.

[7] R. K. Mothilal, A. Sharma and C. Tan. Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 607–617. DOI: 10.1145/3351095.3372850.

[8] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie and P. Flach. FACE: Feasible and Actionable Counterfactual Explanations. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20, 2020, pp. 344–350. DOI: 10.1145/3375627.3375850.

[9] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson and C. Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. 2022. arXiv: 2010.10596. (visited on 20/04/2024).

[10] W. R. Swartout. Explaining and Justifying Expert Consulting Programs. *Computer-Assisted Medical Decision Making*, New York, NY, USA: Springer, 1985, pp. 254–271. Available: https://doi.org/10.1007/978-1-4612-5108-8_15.

[11] G. Vilone and L. Longo. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, vol. 76, pp. 89–106, 2021. DOI: 10.1016/j.inffus.2021.05.009.

[12] Microsoft Research. Explainability. 2022. Available: https://www.microsoft.com/en-us/research/group/dynamics-insights-apps-artificial-intelligence-machine-learning/articles/explainability/ (visited on 25/03/2024).

[13] IBM. What is Explainable AI (XAI)? Available: https://www.ibm.com/topics/explainable-ai (visited on 25/03/2024).

[14] B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, vol. 38, no. 3, pp. 50–57, 2017. arXiv: 1606.08813.

[15] Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. 2021. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206 (visited on 22/03/2024).

[16] E. Tjoa and C. Guan. A Survey on Explainable Artificial Intelligence (XAI): Towards Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021. arXiv: 1907.07374.

[17] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. Specter and L. Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 80–89, 2018. DOI: 10.1109/DSAA.2018.00018.

[18] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud and A. Hussain. Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. *Cognitive Computation*, vol. 16, pp. 45–74, 2023. DOI: 10.1007/s12559-023-10179-8.

[19] C. Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2nd ed. 2022. Available: https://christophm.github.io/interpretable-ml-book.

[20] D. Gunning and D. Aha. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019. DOI: 10.1609/aimag.v40i2.2850.

[21] S. Atakishiyev, M. Salameh, H. Yao and R. Goebel. Explainable Artificial Intelligence for Autonomous Driving: A Comprehensive Overview and Field Guide for Future Research Directions. 2023. arXiv: 2112.11561. (visited on 25/03/2024).

[22] M. T. Ribeiro, S. Singh and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 2016. arXiv: 1602.04938. (visited on 12/04/2024).

[23] S. M. Lundberg and S.-I. Lee. A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, vol. 30, Curran Associ-

ates, Inc., 2017. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf.

[24] N. Spreitzer, H. Haned and I. v. d. Linden. Evaluating the Practicality of Counterfactual Explanations. *Workshop on Trustworthy and Socially Responsible Machine Learning*, 2022. Available: https://openreview.net/forum?id=gi2UZ9mRkUv (visited on 12/04/2024).

[25] T. Lombrozo. The structure and function of explanations. *Trends in Cognitive Sciences*, vol. 10, no. 10, pp. 464–470, 2006. DOI: 10.1016/j.tics.2006.08.004.

[26] J. C. Zemla, S. Sloman, C. Bechlivanidis and D. A. Lagnado. Evaluating everyday explanations. *Psychonomic Bulletin & Review*, vol. 24, no. 5, pp. 1488–1500, 2017. DOI: 10.3758/s13423-017-1258-z.

[27] T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. DOI: 10.1016/j.artint.2018.07.007.

[28] C. Tan. On the Diversity and Limits of Human Explanations. 2022. arXiv: 2106.11988.

[29] A. Mackonis. Inference to the best explanation, coherence and other explanatory virtues. *Synthese*, vol. 190, no. 6, pp. 975–995, 2013. DOI: 10.1007/s11229-011-0054-y.

[30] P. R. Thagard. The Best Explanation: Criteria for Theory Choice. *Journal of Philosophy*, vol. 75, no. 2, pp. 76–92, 1978, Publisher: Journal of Philosophy Inc. DOI: 10.2307/2025686.

[31] P. Thagard. Explanatory Coherence (Plus Commentary). *Behavioral and Brain Sciences*, vol. 12, no. 3, pp. 435–467, 1989. DOI: 10.1017/s0140525x00057046.

[32] T. Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, vol. 55, no. 3, pp. 232–257, 2007. DOI: 10.1016/j.cogpsych.2006.09.006.

[33] S. J. Read and A. Marcus-Newhall. Explanatory coherence in social explanations: A parallel distributed processing account. *Journal of Personality and Social Psychology*, vol. 65, no. 3, pp. 429–447, 1993. DOI: 10.1037/0022-3514.65.3.429.

[34] H. Grice. Logic and conversation. New York, NY, USA: Academic Press, 1975.

[35] I. Stepin, J. M. Alonso-Moral, A. Catala and M. Pereira-Fariña. An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences*, vol. 618, pp. 379–399, 2022. DOI: 10.1016/j.ins.2022.10.098.

[36] Y. Yao, C. Wang and H. Li. Counterfactually Evaluating Explanations in Recommender Systems. 2022. arXiv: 2203.01310. (visited on 07/03/2024).

[37] R. Mccloy and R. Byrne. Counterfactual Thinking about Controllable Events. *Mem. Cognit.*, vol. 28, pp. 1071–1078, 2012. DOI: 10.3758/BF03209355.

[38] R. M. Veski. Measuring Human Preferences in Counterfactual Explanations. University of Tartu Institute of Computer Science Bachelor's thesis, 2024.

[39] R. Likert. A technique for the measurement of attitudes. *Archives of Psychology*, vol. 22, pp. 5–55, 1932.

[40] S. McCarthy. Development of an Explainability Scale to Evaluate Explainable Artificial Intelligence (XAI) Methods. M.S. thesis, Technological University Dublin, 2022. DOI: 10.21427/w4vv-p113.

[41] R. M. J. Byrne. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*, pp. 6276–6282, 2019. DOI: 10.24963/ijcai.2019/876.

[42] N. Van Hoeck, P. D. Watson and A. K. Barbey. Cognitive neuroscience of human counterfactual reasoning. *Frontiers in Human Neuroscience*, vol. 9, 2015. DOI: 10.3389/fnhum.2015.00420. (visited on 25/03/2024).

[43] R. Butz, A. Hommersom, R. Schulz and H. van Ditmarsch. Evaluating the Usefulness of Counterfactual Explanations from Bayesian Networks. *Human-Centric Intelligent Systems*, 2024. DOI: 10.1007/s44230-024-00066-2.

[44] V. Girotto, P. Legrenzi and A. Rizzo. Event controllability in counterfactual thinking. *Acta Psychologica*, vol. 78, no. 1, pp. 111–133, 1991. DOI: 10.1016/0001-6918(91)90007-M.

[45] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi and F. Giannotti. Stable and actionable explanations of black-box models through factual and counterfactual rules. *Data Mining and Knowledge Discovery*, 2022. DOI: 10.1007/s10618-022-00878-5.

[46] R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, 2022. DOI: 10.1007/s10618-022-00831-6.

[47] F. Papenmeier, A. Brockhoff and M. Huff. Filling the gap despite full attention: the role of fast backward inferences for event completion. *Cognitive Research: Principles and Implications*, vol. 4, p. 3, 2019. DOI: 10.1186/s41235-018-0151-2.

[48] F. C. Keil. The Problem of Partial Understanding. *Linguistic Insights*, vol. 144, pp. 251–276, 2011.

[49] F. C. Keil. Explanation and Understanding. *Annual Review of Psychology*, vol. 57, pp. 227–254, 2006. DOI: 10.1146/annurev.psych.57.102904.190100.

[50] T. Miller. Are we measuring trust correctly in explainability, interpretability, and transparency research?, 2022. arXiv: 2209.00651. (visited on 14/03/2024).

[51] D. J. Hilton. Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning*, vol. 2, no. 4, pp. 273–308, 1996. DOI: 10.1080/135467896394447.

[52] M. J. Kusner, J. R. Loftus, C. Russell and R. Silva. Counterfactual Fairness. 2018. arXiv: 1703.06856. (visited on 14/03/2024).

[53] B. Becker and R. Kohavi. Adult. 1996. DOI: 10.24432/C5XW20. (visited on 22/03/2024).

[54] Pima Indians Diabetes Database. Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (visited on 22/03/2024).

[55] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901. DOI: 10.1080/14786440109462720. (visited on 24/04/2024).

[56] L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[57] M. Ester, H.-P. Kriegel, J. Sander and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland, Oregon, USA: AAAI Press, 1996, pp. 226–231. DOI: 10.5555/3001460.3001507.

[58] University of Tartu. UT Rocket. 2018. DOI: 10.23673/PH6N-0144.

[59] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest and A. M. Rush. Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Association for Computational Linguistics, 2020, pp. 38–45. DOI: 10.18653/v1/2020.emnlp-demos.6. (visited on 15/04/2024).

[60] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin. Attention Is All You Need. 2017. arXiv: 1706.03762.

[61] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov and T. Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models. 2023. arXiv: 2307.09288. (visited on 15/04/2024).

[62] Meta AI. Introducing Meta Llama 3: The most capable openly available LLM to date. 2024. Available: https://ai.meta.com/blog/meta-llama-3/ (visited on 22/04/2024).

[63] Mistral.ai. Announcing Mistral 7B. 2023. Available: https://mistral.ai/news/announcing-mistral-7b (visited on 05/05/2024).
[64] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes and A. Mian. A Comprehensive Overview of Large Language Models. 2024. arXiv: 2307.06435.
[65] OpenAI. GPT-4 Technical Report. 2023. arXiv: 2303.08774.
[66] T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. 2023. arXiv: 2305.14314. (visited on 13/04/2024).
[67] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen. LoRA: Low-Rank Adaptation of Large Language Models. 2021. arXiv: 2106.09685. (visited on 13/04/2024).
[68] A. Aghajanyan, L. Zettlemoyer and S. Gupta. Intrinsic Dimensionality Explains the Effectiveness of Language Model Fine-Tuning. 2020. arXiv: 2012.13255. (visited on 05/04/2024).
[69] L. v. Werra, Y. Belkada, L. Tunstall, E. Beeching, T. Thrush, N. Lambert and S. Huang. TRL: Transformer Reinforcement Learning. 2020. Available: https://github.com/huggingface/trl.
[70] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, H. Zhang, B. Zhu, M. Jordan, J. E. Gonzalez and I. Stoica. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. 2024. arXiv: 2403.04132.
[71] S. M. Walker II. GPT-4: How OpenAI Built an AI Model 10x Larger Than GPT-3. *Medium*, 2023. Available: https://medium.com/@smwii/gpt-4-how-openai-built-an-ai-model-10x-larger-than-gpt-3-3f5eacaad69a (visited on 04/05/2024).
[72] P. Todd, J. Ortega, J. Davis, G. Gigerenzer, D. Goldstein, A. Goodie, R. Hertwig, U. Hoffrage, K. Laskey, L. Martignon and G. Miller. Simple Heuristics That Make Us Smart. Oxford University Press, Jan. 1999.

Appendix

I. Survey question examples

Table I.1. Example of a questionnaire question before the pilot study

Factual Scenario:

In the original scenario, we have the following information about an individual:
"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week."

Outcome: Under this setting, you are earning less than the average salary.

Useful context: the standard full-time workload is 40 hours per week.

Counterfactual scenario:

Now, we will present you with a counterfactual scenario that presents what changes you would need to make to earn more than the average salary. The rest of the values will remain constant. The explanation you are presented with is:

"To earn more than the average salary, You would need to increase your education level from high-school to Bachelor's degree."

Please rate on a scale from 1 (very unsatisfied) to 6 (very satisfied), how satisfied would you be with such an explanation: Please rate on a scale from 1 (very infeasible) to 6 (very easy to do), how feasible is this explanation: Please rate on a scale from 1 (very incoherent) to 6 (very coherent), how coherent is this explanation: Please rate on a scale from 1 (very incomplete) to 6 (very complete), how complete is this explanation: Please rate on a scale from 1 (not at all) to 6 (very much), how much do you trust this explanation: Please rate on a scale from 1 (incomprehensible) to 6 (very understandable), how understandable is this explanation: Please rate on a scale from 1 (very biased) to 6 (completely unbiased), how unbiased is this explanation: Please rate on a scale from -2 (too simple) to 0 (ideal complexity) to 2 (too complex), how complex is this explanation:

Table I.2. Example of a questionnaire question from the final study

Imagine you are in this scenario: **"You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week."** Current outcome: You are earning **less than the average salary.** Useful context: the standard full-time workload is 40 hours per week. "To **earn more than the average salary**, you would need to make the following changes: • Increase your education level from high-school to Bachelor's degree."

On a scale from 1 (very unsatisfied) to 6 (very satisfied), how **satisfied** would you be with such an explanation:

On a scale from 1 (very infeasible) to 6 (very easy to do), how **feasible** is this explanation:
Feasibility - the actions suggested by the explanation are practical, realistic to implement and actionable. (click to see examples)

On a scale from 1 (very inconsistent) to 6 (very consistent), how **consistent** is this explanation:
Consistency - all parts of the explanation are logically coherent and do not contradict each other. (click to see examples)

On a scale from 1 (very incomplete) to 6 (very complete), how **complete** is this explanation:
Completeness - the explanation is sufficient in explaining how to achieve the desired outcome. (click to see examples)

On a scale from 1 (not at all) to 6 (very much), how much do you **trust** this explanation:
Trust - I believe that the suggested changes would bring about the desired outcome. (click to see examples)

On a scale from 1 (incomprehensible) to 6 (very understandable), how **understandable** is this explanation:
Understandability - I feel like I understood the phrasing of the explanation well. (click to see examples)

On a scale from 1 (very biased) to 6 (completely fair), how **fair** is this explanation:
Fairness - the explanation is unbiased towards different user groups and does not operate on sensitive features. (click to see examples)

On a scale from -2 (too simple) to 0 (ideal complexity) to 2 (too complex), how **complex** is this explanation:
Complexity - the explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex. (click to see examples)

II. Examples for metrics in questionnaire

Table II.1. All example explanations for metrics, as presented in the introduction of the questionnaire

Feasibility	Good: "If you decrease your mobile phone use from 6 hours a day to 5.5 hours a day, you'll likely experience less symptoms of anxiety." Bad: Changes an attribute that cannot be changed. "If you decrease your age from 42 to 26, you will reduce your risk of a heart attack."
Consistency	Good: "You drove to a concert at an average of 98 km/h and got pulled over by the police 2 times. You left for the concert at 17:30. As a result, you were late for the concert. If you had left for the concert at 17:15, you would have made it to the concert on time." Bad: The suggested changes are contradictory. "You drove to a concert at an average of 98 km/h and got pulled over by the police 2 times. As a result, you were late for the concert. If you had driven at an average of 104 km/h and got pulled over by the police 0 times, you would have made it to the concert on time."
Completeness	Scenario: "You are a 20 year old woman, who has an average grade of 81%. You applied to university without the required motivation letter and meeting the average grade threshold of 90%. You were not accepted into the university." Good explanation: "If you had an average grade of 91% and had written a motivation letter, you would have been accepted into the university." Bad explanation: The suggested changes skip important steps for achieving the desired outcome. "If you had written a motivation letter, you would have been accepted into the university."
Trust	Good: "If you had left home 15 minutes earlier, you would have caught the earlier train, as per the schedule." Bad: The suggested change is unlikely to bring about the desired change. "If you called the train conductor to wait an extra 15 minutes, you wouldn't have missed the train."
Understandability	Good: "If the patient had arrived ten minutes earlier, immediate treatment could have prevented the cardiac arrest." Bad: The explanation uses overly specific and confusing phrases. "The patient's arrival time juxtaposed with the chronological treatment window delineates an alternate outcome scenario."

Fairness	Good: "Applicants with a minimum of 3 years of experience have a higher chance of getting hired for this role." Bad: The explanation suggests changes that would be commonly viewed as unfair or illegal. "People who are over 35 are not usually accepted for this position."
Complexity	Good: "Reducing your debt by 10% could improve your credit rating to meet our loan approval criteria." Too complex: The explanation is unnecessarily long and complex."To improve your credit score to a level that matches our loan approval benchmarks, you should decrease your debt by 4%, increase your monthly savings by 41$, increase your monthly income by 23$, increase your weekly work hours from 40 to 42, work at your current company for 1 more year and receive a letter of recommendation from your current employer.

III. Examples of prompts tested in Experiment I

Table III.1. Example of a baseline prompt

System prompt	You are evaluating counterfactual explanations generated by AI. Counterfactual explanations explain what parameters of a situation should have been different for the outcome to have been different. You are not expected to provide reasoning or explanation and should answer with the appropriate value from the set ["low", "medium", "high"]. The definition of completeness: the explanation is sufficient in explaining how to achieve the desired outcome. The following is the counterfactual explanation.
User prompt / Instruction	Imagine you are in this scenario: "You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week." Current outcome: You are earning less than the average salary. Useful context: the standard full-time workload is 40 hours per week. “To earn more than the average salary, you would need to make the following changes: Increase your education level from high-school to Bachelor’s degree.” The rest of the values will remain constant. Please rate as "low" (very incomplete),"medium" or "high" (fully complete), how complete is this explanation:
Output	medium

Table III.2. Example of a prompt with all definitions

System prompt	You are evaluating counterfactual explanations generated by AI. Counterfactual explanations explain what parameters of a situation should have been different for the outcome to have been different. You are not expected to provide reasoning or explanation and should answer with the appropriate value from the set ["low", "medium", "high"]. The definition of satisfaction: this scenario effectively explains how to reach a different outcome. The definition of feasibility: the actions suggested by the explanation are practical, realistic to implement and actionable. The definition of consistency: the parts of the explanation do not contradict each other. The definition of completeness: the explanation is sufficient in explaining how to achieve the desired outcome. The definition of trust: I believe that the suggested changes would bring about the desired outcome. The definition of understandability: I feel like I understood the phrasing of the explanation well. The definition of fairness: the explanation is unbiased towards different user groups and does not operate on sensitive features. The definition of complexity: the explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex. The following is the counterfactual explanation.
User prompt / Instruction	Imagine you are in this scenario: "You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week." Current outcome: You are earning less than the average salary. Useful context: the standard full-time workload is 40 hours per week. “To earn more than the average salary, you would need to make the following changes: Increase your education level from high-school to Bachelor’s degree.” The rest of the values will remain constant. Please rate as "low" (very incomplete),"medium" or "high" (fully complete), how complete is this explanation:
Output	medium

Table III.3. Example of a prompt with all definitions and examples

System prompt	You are evaluating counterfactual explanations generated by AI. Counterfactual explanations explain what parameters of a situation should have been different for the outcome to have been different. You are not expected to provide reasoning or explanation and should answer with the appropriate value from the set ["low", "medium", "high"]. The definition of satisfaction: this scenario effectively explains how to reach a different outcome. The definition of feasibility: the actions suggested by the explanation are practical, realistic to implement and actionable. The definition of consistency: the parts of the explanation do not contradict each other. The definition of completeness: the explanation is sufficient in explaining how to achieve the desired outcome. The definition of trust: I believe that the suggested changes would bring about the desired outcome. The definition of understandability: I feel like I understood the phrasing of the explanation well. The definition of fairness: the explanation is unbiased towards different user groups and does not operate on sensitive features. The definition of complexity: the explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex. Here are two examples of a prompt and the output. Example prompt 1: "Imagine you are in this scenario: 'You are a 21-year-old person who has an average grade of B. You work part-time for 20 hours per week.' Current outcome: Your university application was rejected. 'To have your application approved, you would need to make the following changes: Improve your average grade from B to A.' The rest of the values will remain constant. Please rate as 'low','medium' or 'high', how consistent is this explanation: " Example output 1: "high". Example prompt 2: "Imagine you are in this scenario: 'You are a 21-year-old person who has an average grade of B. You work part-time for 20 hours per week.' Current outcome: Your university application was rejected. 'To have your application approved, you would need to make the following changes: Increase your hours worked per week from 20 to 80.' The rest of the values will remain constant. Please rate as 'low','medium' or 'high', how feasible is this explanation: " Example output 2: "low". Please answer questions in a similar format. The following is the counterfactual explanation.

User prompt / Instruction	Imagine you are in this scenario: "You are a 31-year-old divorced woman. You have a high-school education and you work 20 hours per week." Current outcome: You are earning less than the average salary. Useful context: the standard full-time workload is 40 hours per week. “To earn more than the average salary, you would need to make the following changes: Increase your education level from high-school to Bachelor’s degree.” The rest of the values will remain constant. Please rate as "low" (very incomplete),"medium" or "high" (fully complete), how complete is this explanation:
Output	medium

IV. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Julius Välja**,

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   Assessing the Quality of Counterfactual Explanations with Large Language Models,

   supervised by Marharyta Domnich, Raul Vicente and Eduard Barbu.

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 4.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

Julius Välja
15/05/2024