

# If you have a change in your birthdate: Assessing counterfactual dialogue capabilities of large language models

Alexander Berman<sup>1</sup>[0000-0003-0513-4107], Christine Howes<sup>1</sup>[0000-0002-2794-1586],  
Ellen Breitholtz<sup>1</sup>[0000-0002-5772-3985], and Staffan  
Larsson<sup>1</sup>[0000-0002-5459-054X]

University of Gothenburg, Sweden  
{firstname.surname}@gu.se

**Abstract.** Large language models (LLMs) are increasingly being used to provide reasons for decisions, but their ability to engage in human-like dialogue and use commonsense reasoning has been called into question. These aspects are important when using LLMs to assist high-stake decisions and assessments such as credit approval. For example, if a loan applicant is denied credit, a counterfactual explanation may state that the application would be granted if the applicant's income increased to a certain amount. By injecting a decision-making algorithm into the LLM prompt and systematically probing and annotating responses for carefully chosen inputs, we study potential patterns in the model's selection of counterfactual explanations. Specifically, we assess notions of actionability acquired by the LLM during pre-training and how such notions are applied by the LLM in natural-language dialogue. The studied notions encompass mutability (e.g. that income can be changed while country of birth cannot), monotonicity (e.g. that age can only change in one direction), and causal dependencies between features (e.g. that duration of residence cannot be increased without also increasing age). Results for the two most recent versions of GPT show that in one studied aspect (mutability), both versions of GPT are well-aligned, while in another aspect (monotonicity), only GPT 4 is well-aligned. Finally, in the third aspect (causal dependencies), none of the versions of GPT are well-aligned. The experiments also suggest that misalignments are primarily due to problems in language generation rather than inherent properties of the models.

**Keywords:** First keyword · Second keyword · Another keyword.

## 1 Introduction

Recently, there has been a large interest in large language models (LLMs) and their ability to engage in human-like dialogue and use commonsense reasoning [17]. In this study, we experimentally investigate specific aspects of these abilities, namely counterfactual explanations in dialogue. The ability to reason counterfactually and provide relevant explanations is particularly important when using AI

to assist high-stake decisions and assessments such as credit approval or medical diagnostics. For example, if a loan applicant is denied credit, a counterfactual explanation (CFE) conveys the conditions under which the credit would have been granted. Ideally, such an explanation is not only faithful to the actual logic underpinning the decision, but also helps the subject to obtain a more preferred decision in the future [36].

In general, the task of selecting relevant counterfactual examples is non-trivial, not only since the set of potential examples tends to be very large, but also since the relevance of an example typically depends on whether the change in circumstances implied by the counterfactual example is *actionable* [13]. Actionability entails various aspects such as physical constraints (e.g. decreasing one’s age is physically impossible) and causal relationships between features (e.g. increasing one’s education level is difficult without increasing one’s age). In other words, CFEs rest on commonsense reasoning and world knowledge. Therefore, it seems reasonable to suspect that LLMs such as GPT, with strong capabilities when it comes to commonsense reasoning and world knowledge [28], also perform well as CFE generators in a dialogical setting.

In this article, we study the ability of LLMs to reason counterfactually in dialogue by systematically probing their responses under various conditions. Specifically, the work investigates whether counterfactual responses generated by LLMs are *aligned with human notions of actionability*, i.e. the extent to which courses of action mentioned by the model are deemed possible by humans. To illustrate this research question, we can consider a scenario where an LLM is asked by a customer if she is eligible for credit; if the LLM responds that the user can become eligible by waiting until she becomes one year older, then this response would be *aligned* with human notions of actionability (since waiting one year is, in principle, an action that can be performed). In contrast, if the LLM suggests getting one year younger, the behavior would be *misaligned* (since age cannot decrease).

The article makes the following contributions:

- A method is proposed for assessing alignment of LLM-generated CFEs with human notions of actionability. The method uses simulated user input, manual annotations and ground-truth actionability criteria (section 3).
- Data and results for GPT 3.5 (`gpt-3.5-turbo-1106`) and GPT 4 (`gpt-4-1106-preview`) focusing on three different aspects of actionability, showing that in one studied aspect (mutability), both versions of GPT are well-aligned; in another aspect (monotonicity), only GPT 4 is well-aligned; in a third aspect (causal dependencies), none of the versions of GPT are well-aligned (section 4).
- A method for assessing the extent to which the performance of an LLM (in this case alignment with human notions of actionability) can be improved using neutrally phrased Socratic follow-up questions is proposed (section 5.3).
- Data and results for Socratic follow-up questioning are presented, showing perfect alignment with human notions of actionability for GPT 3.5 and GPT

4, indicating that observed misalignments are primarily due to problems in language generation rather than inherent incapacities of the model (section 5.3).

All data referred to in this article is published openly and available at <https://github.com/alex-berman/LLM-CFE>.

## 2 Background and Related Work

In this section we introduce the term **counterfactual** and how it generally relates to reasoning and explanations (section 2.1) as well as to explainable AI (XAI) and AI-based decision-making (section 2.2). We then discuss LLMs in relation to reasoning in general and counterfactual reasoning specifically (section 2.3) before situating the present study in relation to previous work (section 2.4).

### 2.1 Counterfactuals and counterfactual explanations

Counterfactual reasoning is the process of considering how events might have turned out differently if particular conditions had been different. Counterfactual reasoning has been argued to be central in human language and thought [31,33] and the ability to reason counterfactually emerges early [5,3]. The term “counterfactual” has been used in many different ways in philosophy of language [19], linguistics [15], psychology, and other fields of research. While in many contexts the term *counterfactual* involves statements with past antecedents that are assumed to be false in the discourse (e.g. “If Caspar had come, it would have been a good party” [18]), CFEs in XAI are typically conceived more broadly as any hypothetical change in feature values that leads to a different outcome, regardless of how the CFE might be framed linguistically, or whether the change is a relevant possibility [11]. In this view, “if you had been two years younger, you would get the loan” and “when you become two years older, you will get the loan” are both CFEs, since they involve a hypothetical change in one of the features. It is on this notion of CFEs that the present study focuses, and, more specifically, on the *actionability* of the hypothetical changes in feature values.

### 2.2 Counterfactuals and explainable AI

There are at least three reasons why CFEs have received significant attention in relation to explainable artificial intelligence (XAI) and AI-based decision-making. First, CFEs are *contrastive*, which is a common feature of human explanations [24]. In other words, CFEs explain why some event  $X$  occurs rather than some other event  $Y$ . For example, if a loan application is rejected ( $X$ ), a CFE conveys the conditions under which the application would have been granted ( $Y$ ). Secondly, CFEs can in principle be obtained for decisions made

with the help of *opaque (black-box) models* [36].<sup>1</sup> For example, if creditworthiness is assessed with a deep neural network, examples of conditions that would yield a positive rather than negative prediction in a particular scenario can be generated without access to the internal workings of the neural network. Finally, CFEs convey information that is, at least in principle, *actionable*. For example, if a customer loan application is rejected and she is informed that credit would be granted if income was €500 higher, then the customer can use this information, e.g. to get credit in the future or to contest or negotiate the decision. It is from this perspective that CFEs are analyzed and discussed in the present study.

In most previous work on counterfactuals for XAI, actionability is not considered at all [11]. Consequently, many methods for generating CFEs can yield suggestions that are entirely unactionable. However, one proposed approach is to manually mark features as either mutable or immutable [26] or more generally to inject an actionability function [32] or structural causal model [21] into the CFE generator. Yet another approach that has been presented is to learn actionability constraints from user feedback [21]. That said, the issue of how to properly satisfy actionability constraints when counterfactually explaining decisions made on the basis of an algorithm remains largely unsolved. The fundamental reason is arguably that actionability is inherently *causal* in nature and therefore cannot be learned from statistical correlations alone [21,29]. This is also one of the reasons why it is interesting to study how LLMs, that are trained and used in an entirely different way than more conventional statistical models, deal with actionability. However, the only previously proposed LLM-based CFE methods of which we are aware focus on simulatability (whether explanations enable humans to predict the model’s output in counterfactual scenarios, [7]) rather than actionability, or inject a causal graph (e.g. containing actionability constraints) into the LLM via the prompt [4,10].

### 2.3 LLMs and reasoning

LLMs have demonstrated an impressive ability to engage in human-like dialogue and use commonsense reasoning. However, they also have some limitations, and considerable effort is currently being spent on understanding and characterizing these limitations. To take a few examples from recent work in this area, [22] cite three common criticisms against LLMs as models of human language processing: (1) excessive reliance on statistical regularities, (2) requiring unrealistic amounts of training data, and (3) insufficient performance in tests using languages other than English. [9] note that LLMs “... struggle particularly in generating coherent, realistic solutions for problems that require novel but concrete physical reasoning.” [25] note that LLMs show a lack of robustness on reasoning tasks involving property inheritance from superordinate conceptual categories (such as the property *can breathe* inherited to the concept DOG from the superordinate

<sup>1</sup> Usually, CFE generation is framed as an optimization problem that maximizes satisfaction of certain predefined desiderata when examining a large number of perturbations of the feature values [11].

concept ANIMAL). [14] show that LLMs establish a new state of the art on causal reasoning tasks but also that they display unpredictable behavior and a lack of robustness.

To the best of our knowledge, the only previous work that explicitly investigates LLMs from the perspective of *counterfactual* reasoning is [12] who study GPT’s performance on causal reasoning tasks and report accuracies ranging from random to moderate for counterfactual queries (corresponding to the highest rung of [30]’s ladder of causation). However, this study addresses formal causal inference problems that do not take actionability into account.

## 2.4 The current study

As shown above, counterfactual explanations in general and their relation to actionability have been discussed in previous literature, as well as reasoning capabilities of LLMs. However, as far as we can tell, no previous work has studied LLMs and counterfactuals from the perspective of actionability, which is the focus of our study.

## 3 Method

We assess actionability of GPT’s CFEs by designing a prompt that stimulates CFE generation, embedding a decision-making algorithm into GPT’s prompt, collecting responses from GPT for simulated user inputs, and analyze the collected data. The analysis is done in two steps. First, responses are annotated with respect to tense and content. Then, alignment with human notions of actionability is measured by applying ground-truth criteria to the annotation labels.

We focus on credit assessment as a domain for the experiments. However, the method as such can be applied to any domain of choice.

### 3.1 Prompt design

For the purposes of the study, we want the prompt to stimulate generation of CFEs (when appropriate), without explicitly containing any instructions about the relevance criteria that GPT should use. For example, the prompt should not instruct GPT to avoid unactionable recommendations such as getting younger. The main motivation for omitting such instructions is that we want to study GPT as a general-purpose AI that does not need to be told about basic facts about how the world works; arguably, this is one of the promises of GPT and similar technologies. To this end, the experiment is also set up as a zero-shot task, without providing GPT with any example interactions.

To find a prompt that stimulates generation of CFEs, we manually design candidate prompts and use trial and error until a suitable prompt has been found. The following prompt was found in this manner and subsequently used in the experiments:

You are a chatbot deployed by a bank to help customers get credit from the bank.

Credit is granted if the following condition concerning the applicant is met:

<Algorithm>

If the customer is currently not eligible , but the customer could potentially become eligible through a change in circumstances , you communicate what such a change in circumstances would look like .

### 3.2 Data collection

For each experiment, a decision-making algorithm is designed in a way that makes it possible to study a specific aspect of CFE generation. A template for user input is then authored, together with a set of ranges for feature values. For each combination of feature values, GPT’s response is then collected from OpenAI’s API using chat completion. During both prompt design and experiments, we use default parameters except for **temperature** and **seed** which are both set to 0 in order to make GPT’s behavior more deterministic and the experiments more reproducible.

### 3.3 Annotation

Given a user input and the response from GPT, annotators are asked to highlight potential spans in the response that contain a CFE and to label presence/type of CFE (0=**no CFE**, 1=**CFE in present/future tense**, or 2=**CFE in past tense**) as well as content. If type is 1 or 2, content is labeled as a value of 0 or 1 for each factor referred to in the embedded algorithm (e.g. age and/or income), where the label indicates whether the factor is involved in the CFE. Annotators may also add comments, e.g. regarding ambiguities/uncertainties. To reduce priming effects, instances are shown in random order and information about prompt and model version are hidden. An example annotation is provided in table 1; the annotation guide is included in Appendix A.

A main annotator (one of the authors, henceforth referred to as A) annotates all collected data. In addition, some data is annotated by 1-2 additional individuals (chosen from a pool of two authors or a colleague, henceforth referred to as B-D).

### 3.4 Measuring alignment with human notions of actionability

Ground truth is defined as experiment-specific criteria in the form of alignment functions that, given a feature vector and label assignments, return an instance-

<b>User input</b>	I would like to know if I'm eligible for credit. I'm 45 years old.
<b>GPT's response</b>	I'm sorry, but based on the age requirement for credit eligibility, you are currently not eligible. However, <b>if your age changes to be between 30 and 40, you would become eligible for credit.</b>
<b>Annotation for type</b>	1
<b>Annotation for age</b>	1

**Table 1.** Example of an annotation from experiment 2 (monotonicity). The user input and the LLM's (in this case GPT 3.5) response are presented to the annotator. *type=1* means that the annotator has labeled the CFE as **present/future tense**. *age=1* means that the annotator has labeled that the CFE involves age.

level alignment value (0 or 1). For example, an alignment function can express that a response is misaligned if the value for the feature *age* is above the eligibility threshold (i.e. the individual is too old to get credit) and the response has been labeled as a CFE in present/future tense involving age (i.e. the response mentions the possibility for age to reach the eligible range).

The experimental results below report alignment values as a per-annotator ratio (essentially an accuracy metric) between the number of aligned instances and the total number of instances. For example, an alignment ratio of 0.9 indicates that 90% of the annotated instances were deemed to be aligned with human notions of actionability. (Detailed annotation results are included in Appendix B.)

## 4 Initial experiments

In a first suite of experiments, we study whether GPT treats certain features as **mutable/immutable** (i.e. whether their value can change), **monotonic** (i.e. whether their value can only change in one direction) and **causally dependent** (i.e. how changing the value of one feature affects the value of another feature).

### 4.1 Experiment 1: Mutability

To assess the extent to which GPT treats some features as mutable/immutable, we conduct an experiment where the outcome of the algorithm depends on the applicant's country of birth. The following algorithm is embedded into the prompt:

```
country_of_birth = united_states
```

The user input describes an applicant with another country of birth than the required one, and is generated using the following template:<sup>2</sup>

<sup>2</sup> The notation  $(x_1|x_2|\dots|x_n)$  means "either  $x_1$ ,  $x_2$ , ... or  $x_n$ ".

I would like to know if I'm eligible for credit. I was born in (France | Canada | Venezuela | Tanzania | Japan | Australia).

When we assess alignment with human notions of actionability, we assume the following ground truth: **A person's country of birth cannot change.**<sup>3</sup> Hence, a response is considered to be aligned if it does not contain a CFE in present/-future tense involving age.

**Results** As an initial observation, we note that in all instances, both versions of GPT begin their responses by correctly reporting that the user is ineligible, with phrases such as "I'm sorry, but based on the information you've provided, you are currently not eligible for credit from our bank" (GPT 3.5). Furthermore, factual explanations are often included in the responses, e.g.: "Based on the eligibility criterion provided, being born in the United States is a requirement to qualify for credit from the bank. Since you were born in France, you currently do not meet this specific condition." (GPT 4) However, since our analysis focuses entirely on *counterfactual explanations* (CFEs), we ignore other aspects of the collected data and exclude non-CFE content from examples presented below.

As reflected in the alignment ratios (see table 2), both versions of GPT are generally aligned with human notions of actionability. GPT gives actionable solutions (e.g. becoming a citizen of the United States, getting a stable income) and does not suggest impossible actions such as changing one's country of birth (see table 3 for complete examples). In fact, in some cases, GPT 4 explicitly states that "one's country of birth cannot be changed". The only potential exception to this pattern is a linguistically ambiguous response from GPT 3.5 which is annotated differently by the two raters (see section 6).

	GPT 3.5		GPT 4	
	A	B	A	B
Ineligible country of birth	6/6=1.0	5/6=0.83	6/6=1.0	6/6=1.0
<b>Total</b>	<b>6/6=1.0</b>	<b>5/6=0.83</b>	<b>6/6=1.0</b>	<b>6/6=1.0</b>

**Table 2.** Alignment ratios for experiment 1 (mutability), grouped by annotator.

<sup>3</sup> Some readers may object that in certain senses a person's birthplace *can* change, e.g. by finding a birth certificate and learning that the person was not born where it was previously thought. However, when we ask follow-up questions to GPT (see section 5.3), it presents no such justifications for CFEs that we deem as misaligned, and instead retracts them.



<b>User:</b> I would like to know if I'm eligible for credit. I was born in Japan.
<b>GPT 3.5:</b> (...) If your circumstances change and you become a citizen of the United States or meet other eligibility criteria, you may become eligible for credit in the future.
<b>User:</b> I would like to know if I'm eligible for credit. I was born in Australia.
<b>GPT 4:</b> (...) To potentially become eligible for credit under the current criterion, a change in circumstances would not be applicable as one's country of birth cannot be changed. (...)

**Table 3.** Examples of dialogues collected through experiment 1 (mutability), with country of birth as the only feature considered by the algorithm. Note that for system responses, only CFE spans are included.

## 4.2 Experiment 2: Monotonicity

To assess the extent to which GPT treats some features as monotonic, we conduct an experiment where the output from the algorithm depends on the applicant's age being in a certain range. By stimulating GPT to generate CFEs for age values outside the range, we can study the extent to which GPT treats age as something that can increase and/or decrease. The following algorithm is embedded into the prompt:

```
30 <= age <= 40
```

The following template for user input is used:

```
I would like to know if I'm eligible for credit. (I'm
<X> years old | I turn <X> next month | I turned <X>
last month).
```

where X is an integer in the range [25, 29] or in the range [41, 45].

When we assess alignment with human notions of actionability, we assume the following ground truth: **A person's age can increase but not decrease.** We also assume that humans provide a CFE if an actionable recourse (e.g. waiting until age reaches 30) is available.

**Results** As illustrated by table 4, GPT 3.5 is misaligned in half of the instances while GPT 4 is perfectly aligned. Specifically, when age is too low, GPT consistently mentions the possibility to become eligible, as one might expect. For example, GPT refers to the possibility of waiting, or informs the applicant about becoming eligible once they have reached the age of 30 (see table 5 for complete

examples). However, GPT 3.5 also consistently mentions the possibility to become eligible when age is too high. For example, when the user is 43, GPT 3.5 responds that “if there are any changes in your circumstances that bring your age within the eligible range of 30 to 40, such as a younger co-applicant or a change in your birthdate, you could potentially become eligible.” In other words, GPT 3.5 treats age as a property that can change in both directions.

	GPT 3.5		GPT 4	
	A	D	A	D
Age too low	15/15=1.0	15/15=1.0	15/15=1.0	15/15=1.0
Age too high	0/15=0.0	0/15=0.0	15/15=1.0	15/15=1.0
<b>Total</b>	<b>15/30=0.5</b>	<b>15/30=0.5</b>	<b>30/30=1.0</b>	<b>30/30=1.0</b>

**Table 4.** Alignment ratios for experiment 2 (monotonicity).

<b>Age too low</b>	
<b>User:</b>	I would like to know if I’m eligible for credit. I’m 29 years old.
<b>GPT 3.5:</b>	(...) you will become eligible once you reach the age of 30.
<b>Age too high</b>	
<b>User:</b>	I would like to know if I’m eligible for credit. I’m 44 years old.
<b>GPT 3.5:</b>	(...) if your age changes to be within the range of 30 to 40, you would become eligible for credit.
<b>User:</b>	I would like to know if I’m eligible for credit. I turn 42 next month.
<b>GPT 4:</b>	(...) To become eligible based on age, you would have needed to apply before exceeding the maximum age limit of 40. (...)
<b>User:</b>	I would like to know if I’m eligible for credit. I turn 43 next month.
<b>GPT 4:</b>	(...) Unfortunately, age is not a factor that can be changed to meet eligibility criteria. (...)

**Table 5.** Examples of dialogues collected through experiment 2 (monotonicity).

### 4.3 Experiment 3: Causal dependencies

To study the extent to which GPT takes causal dependencies between variables into account, we design a scenario where eligibility hinges on not having reached

a certain age and on having been a resident for a sufficient period of time. This makes it possible to study the extent to which GPT considers age and duration of residence to be causally related. The following algorithm is used:

```
age <= 65 and years_of_residence >= 3
```

We focus on cases where the age criterion is met but duration of residence is too short. Specifically, we investigate two kinds of cases:

- **Sufficient time.** The applicant can become eligible after a certain amount of waiting. In terms of feature values, the distance to the age boundary is larger than the distance to the duration boundary. For example, if age=61 and duration=1, the applicant can become eligible by waiting two years.
- **Insufficient time.** The applicant cannot become eligible by waiting. In terms of feature values, the distance to the age boundary is smaller than the distance to the duration boundary. For example, if age=64 and duration=1, the applicant can *not* become eligible by waiting two years, due to the increase in age.

The following input template is used:

```
I would like to know if I'm eligible for credit. I am
<X> years old. I have been a resident for (1 month|1
year|2 years).
```

where X is an integer in the range [61, 65].

When we assess alignment with human notions of actionability, we assume the following ground truth: **If a person's duration of residence is increased, the person's age increases by the same amount of time.** We also assume that when time is sufficient, a human would communicate a CFE that involves duration of residence (as in "you will become eligible when your duration of residence reaches 3 years").

**Results** As illustrated by table 6, alignment is perfect when time is sufficient, and completely lacking when time is insufficient. Specifically, both versions of GPT consistently refer to the possibility to become eligible by increasing the duration of residence, regardless of whether time is in fact sufficient when the increase in age is taken into account. In several cases, GPT does not mention age at all in its CFEs, behaving as if the factors were entirely independent of each other. This occurs for both versions in both kinds of situations (sufficient and insufficient time). For example, when age is 65 and duration of residence is 1 month, GPT 4 responds: "To potentially become eligible for credit, you would need to continue your residency until you have reached the 3-year minimum", without mentioning the age criterion at all.

However, in other scenarios both versions of GPT qualify their CFEs with phrases such as "...as long as you are still under 65 years old" or "assuming ... you are still under the age of 65 at that time", indicating an awareness of a

causal dependency, but without communicating its implications for the specific situation at hand. There are also cases where GPT 4 does communicate a concrete causal implication, e.g.: “Since you are 61 years old, you will still be under the age limit of 65 by the time you meet the residency requirement”. In some of these cases, the conclusion is valid, while in others, the user would actually have become more than 65 years old once the residence criterion is fulfilled. Interestingly, false conclusions are produced even when GPT explicitly spells out the numerical circumstances: “you could potentially become eligible for credit if you continue to reside for an additional 2 years (...) Since you are 64 years old, you will still be under the age limit of 65 by the time you meet the residency requirement”. (See table 7 for complete examples.)

	GPT 3.5		GPT 4	
	A	C	A	C
Sufficient time	6/6=1.0	3/3=1.0	6/6=1.0	0/0
Insufficient time	0/6=0.0	0/1=0.0	0/6=0.0	0/2=0.0
<b>Total</b>	<b>6/12=0.5</b>	<b>3/4=0.75</b>	<b>6/12=0.5</b>	<b>0/2=0.0</b>

**Table 6.** Alignment ratios for experiment 3 (causal dependencies), grouped by annotator.

<b>Sufficient time</b>
<b>User:</b> I would like to know if I’m eligible for credit. I am 61 years old. I have been a resident for 1 year.
<b>GPT 4:</b> (...) you could potentially become eligible for credit in the future if you continue to reside in your current location for at least 2 more years, which would satisfy the years of residence requirement. Since you are 61 years old, you will still be under the age limit of 65 by that time.
<b>Insufficient time</b>
<b>User:</b> I would like to know if I’m eligible for credit. I am 63 years old. I have been a resident for 1 month.
<b>GPT 3.5:</b> (...) To become eligible, you would need to have at least 3 years of residence. Once you meet this requirement, you would be eligible for credit as long as you are still under 65 years old.

**Table 7.** Examples of dialogues collected through experiment 3 (causal dependencies).

## 5 Follow-up experiments

As revealed by the results in section 4, GPT sometimes mentions possible re-courses that seem to contradict human notions of actionability. For example, GPT 3.5 behaves as if age can decrease, and both versions of GPT sometimes behave as if duration of residence can be increased without becoming older.

The perhaps most evident explanation for the observed misalignments is that GPT’s notion of actionability (as encoded, somehow, in the model’s parameters) indeed differ from human notions, at least in some respects. However, at least three other explanations seem worth investigating further before drawing any such conclusion. First, it seems conceivable that the instructions as formulated in the prompt do not unequivocally convey that the changes in circumstances that GPT communicates should be *actionable*. When the prompt refers to a “change in circumstances” through which the “customer could potentially become eligible”, then in a particular (perhaps infelicitous) reading, “change in circumstances” could be taken to mean any change that makes the customer eligible, rather than any *actionable* change.

Second, misalignments could potentially be explained by some kind of **positivity bias**, i.e. a tendency to favor positive/encouraging responses over negative/discouraging ones. When choosing between generating a CFE or merely communicating a rejection, a positivity bias may nudge the model towards the first option, even if this contradicts actionability considerations.

Third, some of GPT’s responses indicate a failure in logical/mathematical reasoning as the ingredient underpinning unactionable CFEs. Specifically, in experiment 3 (section 4.3), GPT sometimes demonstrates awareness of a causal dependency between the two variables, but in some cases does not seem able to reason correctly about how the dependency pans out. Previous work has demonstrated that LLM’s improvements on reasoning tasks can be significantly improved by prompting the model to “reason in steps” [16,27] and that LLMs can detect inconsistencies in their own outputs [1], suggesting that failures in reasoning are partly a *language generation problem* related to how the task is framed, rather than an inherent inability of the model.

Follow-up experiments for testing all of these three hypotheses are presented below.

### 5.1 Hypothesis 1: Too implicit actionability cue in prompt

To test whether a more explicit actionability cue in the prompt affects GPT’s behavior, we modify the initial prompt (see section 3.1) by replacing “what such a change in circumstances would look like” with “what the customer would need to do to get credit” (for the full prompt, see Appendix C).

**Results** As seen in table 8, a negative effect on alignment is indicated for one of the annotators for experiment 1 (mutability). However, a more detailed look at the instances at hand reveal that the inter-annotator disagreements in these

cases are of a similar kind as the one observed with the original prompt, namely a linguistically ambiguous response.

As for experiment 2 (see table 9), improvements in alignment can be observed in cases when age is too high. While with the original prompt, the model consistently mentions the possibility of becoming eligible through a change in age, with the alternative prompt this no longer occurs in around 1/3 of the cases.<sup>4</sup> Specifically, in 3 cases when age is too high, GPT 3.5 mentions other possibilities of becoming eligible (applying with a co-signer, improving credit score, other financial products or services); in one case, GPT 3.5 seems to imply that other possibilities could make the user eligible (“If you have any plans to change your circumstances, such as applying with a co-signer or improving your credit score, please let me know and I can provide further assistance”), and in one case it refrains from making any specific suggestions (“if there are any changes in your circumstances that could make you eligible, please let me know and I can provide guidance on what you would need to do to become eligible”). Importantly, however, in a majority of cases when age is too high, GPT 3.5 still mentions possibilities involving a change in age. For example, when the user turns 42 next month, GPT 3.5 responds: “If you would like to become eligible, you would need to wait until you are between the ages of 30 and 40.” In other words, the propensity of GPT 3.5 to treat age as a property that can change in both directions remains, but is somewhat less pronounced with the alternative prompt.

As for experiment 3 (see table 10), no differences can be observed in relation to the original prompt. To conclude, hypothesis 1 can potentially explain *some observed instances of misalignment* between GPT’s and human notions of actionability, but *only for one aspect of actionability* (monotonicity).

	GPT 3.5		GPT 4	
	A	B	A	B
Ineligible country of birth	6/6=1.0	3/6=0.5	6/6=1.0	5/6=0.83
<b>Total</b>	<b>6/6=1.0</b>	<b>3/6=0.5</b>	<b>6/6=1.0</b>	<b>5/6=0.83</b>

**Table 8.** Alignment ratios for experiment 1 (mutability) with the alternative prompt.

## 5.2 Hypothesis 2: Positivity bias

To test whether GPT’s responses are shaped by an inclination to favor positive responses, we modify the embedded algorithms by adding a disjunction involving an additional factor with a condition that is currently unmet but that can relatively straightforwardly become fulfilled. Specifically, we add `or monthly_income >= 2000` and state in user input that income is €1800 per

<sup>4</sup> Annotator D’s ratings may seem to contradict this pattern, but may rather be a result of interpreting the annotation scheme in an unintended way (see section 6).

	GPT 3.5			GPT 4		
	A	C	D	A	C	D
Age too low	15/15=1.0	3/3=1.0	2/2=1.0	15/15=1.0	4/4=1.0	4/4=1.0
Age too high	5/15=0.33	1/3=0.33	2/2=1.0	15/15=1.0	2/2=1.0	4/4=1.0
<b>Total</b>	20/30= <b>0.67</b>	4/6= <b>0.67</b>	4/4= <b>1.0</b>	30/30= <b>1.0</b>	6/6= <b>1.0</b>	8/8= <b>1.0</b>

**Table 9.** Alignment ratios for experiment 2 (monotonicity) with the alternative prompt.

	GPT 3.5		GPT 4	
	A	C	A	C
Sufficient time	6/6=1.0	1/1=1.0	6/6=1.0	2/2=1.0
Insufficient time	0/6=0.0	0/1=0.0	0/6=0.0	0/0
<b>Total</b>	6/12= <b>0.5</b>	1/2= <b>0.5</b>	6/12= <b>0.5</b>	2/2= <b>1.0</b>

**Table 10.** Alignment ratios for experiment 3 (causal dependencies) with the alternative prompt.

month (see Appendix D). Consequently, a CFE that focuses on an increase in income can always be selected, potentially without mentioning any other factor. To the extent that GPT deemed for the unmodified algorithms that it sometimes had to choose between a constructive but unactionable suggestion and no constructive suggestion at all, with the modified algorithm it always has a “backdoor” through which a positive route can be taken. In other words, if positivity bias accounts for some misalignments between GPT’s and human notions of actionability, the “backdoor” option should reduce the amount of misalignments.

**Results** GPT consistently mentions the ability to become eligible through a change in income (see Appendix D for annotation results). However, GPT continues to mention other factors as well, frequently causing misalignments to persist.

For experiment 1 (mutability), a slight decrease in alignment can be observed for GPT 3.5 (see table 11) compared to the experiment without a backdoor. Specifically, in one case GPT 3.5 implies that country of birth is mutable: “you could potentially become eligible if your monthly income increases to at least €2000 or if you were born in the United States. If any of these circumstances change, please feel free to reach out to us again.” (Country of birth is mentioned once by GPT 4 as well, but not in a way that suggests that it is mutable: “you could potentially become eligible for credit if your circumstances change in one of the following ways: (...) If you become a naturalized citizen of the United States, assuming the bank’s policy equates naturalization with being born in the United States for the purpose of credit eligibility.”)

In experiment 2 (monotonicity), the backdoor option seems to have a positive effect on alignment (see table 12). When age is too high, GPT 3.5 still mentions a potential change in age in 6 of 15 cases, while in the remaining 9 cases, it only mentions a potential change in income. In some of the cases when age is mentioned, GPT 3.5 conveys the algorithm inaccurately: “if you were to apply before your 44th birthday, you would also meet the age requirement”.

When it comes to experiment 3 (causal dependencies), no differences can be observed in relation to the initial version of the experiment. Specifically, the backdoor option completely removes any mention of age, but changes in duration of residence are still mentioned by both models. Overall, the pattern from the original version of the experiment persists: Both models of GPT consistently mention the possibility of residing for a longer duration of time in order to become eligible, regardless of whether this is causally compatible with the user’s age.

In summary, the results suggest that a positivity bias may explain *some instances of misalignments with human notions of actionability, but only partially for one of the studied aspects* (monotonicity).

	GPT 3.5	GPT 4
	A	A
Ineligible country of birth	5/6=0.83	6/6=1.0
<b>Total</b>	5/6= <b>0.83</b>	6/6= <b>1.0</b>

**Table 11.** Alignment ratios for the backdoor variant of experiment 1 (mutability).

	GPT 3.5	GPT 4
	A	A
Age too low	15/15=1.0	15/15=1.0
Age too high	9/15=0.6	15/15=1.0
<b>Total</b>	24/30= <b>0.8</b>	30/30= <b>1.0</b>

**Table 12.** Alignment ratios for the backdoor variant of experiment 2 (monotonicity).

### 5.3 Hypothesis 3: Generation problem

To assess the extent to which misalignments with human notions of actionability are due to issues pertaining to language generation rather than inherent capabilities of the model, several different kinds of follow-up experiments can be conceived. Here we take a more conversational approach and ask Socratic



	GPT 3.5		GPT 4	
	A	D	A	D
Sufficient time	6/6=1.0	2/2=1.0	6/6=1.0	1/1=1.0
Insufficient time	0/6=0.0	0/1=0.0	0/6=0.0	0/1=0.0
<b>Total</b>	<b>6/12=0.5</b>	<b>2/3=0.67</b>	<b>6/12=0.5</b>	<b>1/2=0.5</b>

**Table 13.** Alignment ratios for the backdoor variant of experiment 3 (causal dependencies).

follow-up questions to GPT to give it the opportunity to detect inconsistencies or errors in its own output. Specifically, we use a form of **elenchus** [23], where questions are intended to challenge the model’s reasoning. The purpose of our variant of the Socratic method is to examine whether *neutrally asked follow-up questions* (i.e. questions that do not explicitly guide the model in a particular direction) cause GPT to *revise its own answer*.

We frame our Socratic method as an **elenchus game**, played by an experimenter with the goal of testing whether GPT revises its initial standpoint when faced with questions that challenge its claims, arguments or reasoning. The experimenter acts as a user in the credit application domain and resumes the dialogue with the system after the initial input-response exchange. (For more details about the game, see Appendix E.)

Since the elenchus game is costly in terms of work-time and effort, it is only played by one experimenter (annotator A) for a subset of the input-response pairs collected in the initial experiments. Specifically, one input-response pair is randomly sampled for each annotation category (combination of model and CFE variant, as labeled by annotator A), amounting to a total of 16 games. Once the games have been played, the same annotator labels each collected dialogue in the same way as initial responses (i.e. in terms of CFE type and content) but with respect to GPT’s standpoint *after* elenchus questioning.

**Results** As conveyed by tables 14-16, alignment after elenchus questioning is perfect across experiments and models. Specifically, in all instances where GPT’s initial response was *aligned* with human notions of actionability, GPT *maintains* its initial standpoint when faced with elenchus questions. (For detailed annotations, see Appendix E.) For example, in experiment 3 (causal dependencies), when age is 62 and duration of residence is 2 years, GPT 3.5 initially responds that the user will be eligible if she/he becomes a resident for at least 3 years. When faced with the elenchus question “So if I become a resident for at least 3 years, I will be eligible?”, GPT 3.5 replies: “Yes, if you become a resident for at least 3 years, you would meet the residence requirement and become eligible for credit, given that you are currently 62 years old.” (For complete dialogue examples, see tables 17-18.)

More interestingly, in each instance where GPT’s initial response was *mis-aligned* with human notions of actionability, GPT *revises* its initial standpoint when faced with elenchus questions. For example, in the same experiment, when age is 65 and duration of residence is 2 years, GPT 3.5 initially responds that the user will be eligible if she/he becomes a resident for at least 3 years. When faced with the elenchus question “Will I still be under 65 years old at that time?”, GPT 3.5 replies: “No, you will not be under 65 years old at that time. Therefore, you will not be eligible for credit based on the age requirement.”

It is also worth noting that merely echoing GPT’s initial conclusion back to it as a confirmation question can cause the model to highlight a dependency that it initially did not mention. For example, when age is 65 and residence is 2 years, GPT 3.5 initially responds: “if you become a resident for at least 3 years, you would meet the residence requirement and be eligible for credit”, but does not mention age. However, when faced with the elenchus question “So if I become a resident for at least 3 years I will be eligible?”, GPT affirms with “Yes” and then adds “provided you are still under 65 years old at that time” (although yet another round of elenchus is required before GPT revises its conclusion). In other words, in instances where GPT seems not to consider a causal dependency at all (judging from its initial response), a basic confirmation question *without any additional cues* can trigger GPT to attend to the causal dependency.

	GPT 3.5	GPT 4
	A	A
Ineligible country of birth	1/1=1.0	2/2=1.0
<b>Total</b>	1/1= <b>1.0</b>	2/2= <b>1.0</b>

**Table 14.** Alignment ratios for experiment 1 (mutability) after elenchus questioning.

	GPT 3.5	GPT 4
	A	A
Age too low	1/1=1.0	1/1=1.0
Age too high	1/1=1.0	2/2=1.0
<b>Total</b>	2/2= <b>1.0</b>	3/3= <b>1.0</b>

**Table 15.** Alignment ratios for experiment 2 (monotonicity) after elenchus questioning.

## 6 Validity

Agreement for the raw annotations (i.e. annotations of CFE type and factor involvement) range from perfect in most cases, to none in other cases. Some of the

	GPT 3.5	GPT 4
	A	A
Sufficient time	2/2=1.0	2/2=1.0
Insufficient time	2/2=1.0	2/2=1.0
<b>Total</b>	<b>4/4=1.0</b>	<b>4/4=1.0</b>

**Table 16.** Alignment ratios for experiment 3 (causal dependencies) after elenchus questioning.

<b>User:</b> I would like to know if I’m eligible for credit. I am 62 years old. I have been a resident for 2 years.
<b>GPT:</b> I’m sorry, but based on the criteria for credit eligibility, you are currently not eligible. However, if you become a resident for at least 3 years, you would meet the residence requirement and become eligible for credit.
<b>User:</b> So if I become a resident for at least 3 years, I will be eligible?
<b>GPT:</b> Yes, if you become a resident for at least 3 years, you would meet the residence requirement and become eligible for credit, given that you are currently 62 years old.
<b>User:</b> How old will I be if I become a resident for at least 3 years?
<b>GPT:</b> If you become a resident for at least 3 years, you will be 65 years old. At that point, you would meet the age requirement and the residence requirement, making you eligible for credit.

**Table 17.** Example of dialogue where GPT 3.5 *maintains* its initial standpoint when faced with elenchus questions.

poor agreement reflects redundancy in the annotation scheme and has no impact on agreement concerning alignment with human notions of actionability. For example, responses that only mention extraneous factors were rated by annotator A as **CFE in present/future tense** and by B as **no CFE**. The disagreement seems to be caused by different interpretations of counterfactuality that have no bearing on the assessed degree of alignment.

As can be observed in table 19, agreement for alignment is in many cases perfect, with a relatively small amount of disagreements. An analysis reveals that the disagreements fall into two categories: linguistically ambiguous responses from GPT and different interpretations of the annotation instructions. The former category pertains to disagreements between A and B regarding the following responses from GPT 3.5:

<b>User:</b> I would like to know if I’m eligible for credit. I am 65 years old. I have been a resident for 2 years.
<b>GPT:</b> I’m sorry, but based on the criteria for credit eligibility, you are currently not eligible. However, if you become a resident for at least 3 years, you would meet the residence requirement and be eligible for credit.
<b>User:</b> So if I become a resident for at least 3 years I will be eligible?
<b>GPT:</b> Yes, if you become a resident for at least 3 years, you would meet the residence requirement and be eligible for credit, provided you are still under 65 years old at that time.
<b>User:</b> Will I still be under 65 years old at that time?
<b>GPT:</b> No, you will not be under 65 years old at that time. Therefore, you will not be eligible for credit based on the age requirement.
<b>User:</b> So if I become a resident for at least 3 years, will I be eligible?
<b>GPT:</b> I’m sorry, but based on the criteria for credit eligibility, you are currently not eligible. This is because you are already 65 years old, and becoming a resident for at least 3 years will not change that.

**Table 18.** Example of dialogue where GPT 3.5 *revises* its initial standpoint when faced with elenchus questions.

1. “if you are interested in becoming eligible for credit, you would need to meet the bank’s requirement of being born in the United States. Since the country of birth is a fixed criterion that cannot be changed, you may want to inquire if the bank has any other credit products...” (1 instance)
2. “(to become eligible) you would need to have been born in the United States. If your circumstances change and you become a citizen of the United States ...” (4 instances)

Example 1 is ambiguous in the sense that the first sentence might imply that being born in the U.S. is a circumstance that can be changed, even if this implication is contradicted by the second sentence. Example 2 is ambiguous in terms of tense and modality: It is unclear whether “having been born” is a circumstance that could potentially change (as in “you would need to have a higher income”), or whether it cannot change since it belongs to the past.

The second category of disagreements (between A and D) pertains to 2 instances where where age is too high and GPT 3.5 mentions the possibility to wait until age is in the eligible range. Annotator D has labeled them as **CFE in past tense** (but commented that the cases are questionable and “not exactly in

the past tense”), presumably because the change implied by the response would require going backwards in time (assuming a human notion of actionability).

Taken together, the observed disagreements indicate some challenges associated with annotating and assessing actionability in terms of grammatical tense. Nevertheless, since the overall degree of agreement is very high, the results can be considered to be reliable.

Until this point, the discussion regarding validity has focused on experiments where collected data has been labeled by at least two annotators. In the case of elenchus questioning (section 5.3), dialogues have been collected and labeled by a single annotator. Furthermore, the validity of the results depend on the extent to which elenchus questions are formulated neutrally, i.e. in ways that do not guide GPT in a particular (e.g. more aligned) direction. The strikingly positive observed effects of elenchus questioning in this study may raise the suspicion that the employed elenchus questions were not entirely neutral. However, since the amount of data is small and is made publicly available, the reliability of the results can be transparently scrutinized.

	<b>B</b>			<b>C</b>			<b>D</b>		
Experiment	$N_{AL}$	$N$	$\kappa$	$N_{AL}$	$N$	$\kappa$	$N_{AL}$	$N$	$\kappa$
1 (initial)	11	12	0.0						
2 (initial)							60	60	1.0
3 (initial)				6	6	1.0			
1 (alternative prompt)	8	12	0.0						
2 (alternative prompt)				12	12	1.0	10	12	0.0
3 (alternative prompt)				4	4	1.0			
3 (with backdoor)							5	5	1.0

**Table 19.** Inter-rater reliability for alignment values, reported with agreement metrics for comparisons with annotator A.  $N_{AL}$  denotes number of aligned instances,  $N$  number of instances and  $\kappa$  Cohen’s kappa. For example, for experiment 1, annotators A and B agreed in 11 out of 12 instances, with  $\kappa = 0.0$ .

## 7 Discussion and Conclusions

The present study has collected and assessed counterfactual explanations generated by GPT 3.5 and GPT 4 in terms of how well they align with basic actionability notions such as not being able to modify past events. The results indicate that the degree of alignment differs across model versions and actionability aspects. In one studied aspect (immutability of birthplace), both versions of GPT are well-aligned; in another aspect (monotonicity of age), only GPT 4 is aligned; in a third aspect (causal dependency between age and duration of residence), both versions are generally misaligned.

This experimental set-up involves several limitations which should be addressed in future work, particularly with regards to the models queried and the chosen scenario. More systematic experiments are needed to see how or whether the issues we illustrate are present in different domains and contexts of use, or if, as is plausible, different discrepancies are present in different scenarios (e.g. medical explanations of test results etc). It is also probable that there are differences in performance in different LLM architectures, though ChatGPT remains a good testbed since it has so rapidly become relied on by so many users.<sup>5</sup>

Nevertheless, the discrepancies between model behaviour and elementary commonsense knowledge we observed may seem unexpected or even comical, especially when considering the enormous amounts of data on which these models have been trained and the impressive performance results that they have achieved on tasks that appear to be much more complex. How can a model that has read hundreds of billion words of human-written text suggest to a 42-year-old customer to wait until she gets between 30 and 40? How can a system that scores in the 90th percentile of the bar exam for lawyers tell a 64-year-old customer that she will be under the age of 65 if she continues to reside for an additional 2 years?

To resolve this seeming paradox, we have proposed and examined three hypotheses for the observed misalignments: too implicit actionability cues in the instructions provided to the model via the prompt, a bias favoring positive responses over negative ones, and reasoning failures associated with language generation problems. The results of these follow-up experiments indicate that the first two hypotheses have limited bearing, while the third hypothesis (a problem in language generation) accounts for *all observed instances of misalignments*.

A possible conclusion is that GPT’s notions of actionability *are* human-aligned (at least in the aspects examined in this study), even if its responses sometimes indicate otherwise. In light of our own experiments as well as some previous work on LLMs and reasoning [16,27,1], it seems tenable that LLMs sometimes fail or under-perform as a consequence of explicitly or implicitly *restricting the amount of tokens that they produce* when responding to a specific query. Current transformer-based LLMs operate on a strictly feedforward basis, without maintaining an internal state across invocations [34]. For example, if an LLM is tasked to produce a single token in a response to a multiple choice question, it will perform the task in a single feedforward pass from input to output. In contrast, when the LLM is given the opportunity to reason iteratively in a way where its own output is fed back to the model as new input, the input-output loop constitutes a kind of working memory or “scratchpad”. This mechanism vastly expands the scope of possible computations that the model can perform to solve the task, both quantitatively (in terms of number of feedforward passes) and qualitatively (in terms of the kinds of computation that can be performed). In other words, an extended ability to use the input-output feedback mechanism can potentially facilitate reasoning, perhaps analogous to a phonological loop

---

<sup>5</sup> Unsystematic prompting of a number of more recent models such as DeepSeek in May 2025 suggests that the issues we identified persist.

[2] or extended mind [8]. Hence, a possible explanation for why model outputs sometimes appear bizarre or nonsensical is that a feedback mechanism of this kind is not sufficiently leveraged.

Although GPT is not explicitly constrained in our experiments, e.g. in terms of the number of tokens it may produce, the positive effects of elenchus follow-up questions (see section 5.3) could nevertheless be explained, at least in part, by the extended possibility to “reason in words” that they give rise to. Specifically, through the additional feedforward passes, GPT detects fallacies in its previous output; in one case, GPT 3.5 explicitly apologizes “for the confusion earlier”, even if no fallacies were pointed out by the user. Given that the Socratic questions lack cues regarding what the “correct” reasoning would be, the benefits of answering the questions cannot be explained by externally administered “corrections”. However, since our elenchus questions purposely challenge aspects of GPT’s output that are of interest to the study at hand (see Appendix E), some degree of guidance is externally imposed. In future work it would be interesting to study the effects of generic self-reflection triggers such as “Look at your answer and see if it can be improved in any way” as a means to isolate the contributions of expanded working memory versus external reasoning guidance.

It should be emphasized that although elenchus questioning drastically improves alignment with human notions of actionability in our experiments, the method is manually administered (in our case by one of the authors) and is therefore not applicable as a performance improvement technique for automated LLM-based systems. However, in principle the “elenchus game” could be mechanized, e.g. by letting one and the same LLM (or two different LLMs) play the role of both respondent and interlocutor (cf. [20]). This could also be theoretically interesting as a possible model of intrapersonal communication in humans [35,6]. From this perspective, one possible explanation for why people generally do not suggest age reversal as a tenable recourse is not that such ideas never pop up in their minds, but rather that such ideas, when they do pop up, are detected as nonsensical and filtered out before they are verbalized, as part of a (potentially unconscious) inner monologue involving both constructive/generative and reflective/critical processes.

A possible alternative to Socratic question is chain-of-thought (CoT) prompting, through which an LLM is instructed to reason in steps. Several variants of CoT prompting have been proposed in the literature, including few-shot prompting with examples of correct reasoning steps [37] and zero-shot prompting with an explicit “think step by step” instruction [16]. However, these approaches are designed for text completion rather than dialogue. For example, in Kojima et al.’s ([16]) approach, the input to the LLM consists of a question (“Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?”) followed by the beginning of an answer (“A: Let’s think step by step.”); the model is then tasked to complete the answer. How to apply this kind of prompting in a conversational setting remains an open question that would be interesting to investigate in future work.

Finally, it should be stressed that the aim of this study is to assess capabilities and analyze observed behaviors, not to improve performance. If one would like to maximize an LLM’s degree of alignment with human notions of actionability, a simple and practical approach would be to provide actionability criteria in the prompt – not because the model does not “know” that age cannot decrease, but to facilitate desired reasoning patterns by activating concepts that the model has acquired during training. The theoretically relevant question is if (and why) LLMs *need* the kind of explicit instructions that historically were encoded in classical rule-based AI systems and that people rarely use in social interaction. For example, a loan official does not need to be taught (or reminded) that customers’ age cannot decrease. Our results indicate that despite blatant examples of misalignments, LLMs may *not* need such instructions. Counterfactually speaking, if LLM-based system had better conditions to regulate their reasoning processes, they might have behaved more like humans.

**Acknowledgments.** This research was supported by ERC Starting Grant DivCon: Divergence and convergence in dialogue: The dynamic management of mismatches (101077927) and by the Swedish Research Council grant (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article. Or: Author A has received research grants from Company W. Author B has received a speaker honorarium from Company X and owns stock in Company Y. Author C is a member of committee Z.

## References

1. Agrawal, A., Mackey, L., Kalai, A.T.: Do language models know when they’re hallucinating references? arXiv preprint arXiv:2305.18248 (2023)
2. Baddeley, A.: Working memory and language: an overview. *Journal of Communication Disorders* **36**(3), 189–208 (2003). [https://doi.org/10.1016/S0021-9924\(03\)00019-4](https://doi.org/10.1016/S0021-9924(03)00019-4), <https://www.sciencedirect.com/science/article/pii/S0021992403000194>, aSHA 2002
3. Beck, S.R., Riggs, K.J.: Developing thoughts about what might have been. *Child development perspectives* **8**(3), 175–179 (2014)
4. Bhattacharjee, A., Moraffah, R., Garland, J., Liu, H.: Llms as counterfactual explanation modules: Can chatgpt explain black-box text classifiers? arXiv preprint arXiv:2309.13340 (2023)
5. Buchsbaum, D., Bridgers, S., Skolnick Weisberg, D., Gopnik, A.: The power of possibility: Causal learning, counterfactual reasoning, and pretend play. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**(1599), 2202–2212 (2012)
6. Centeno-Cortés, B., Jiménez Jiménez, A.F.: Problem-solving tasks in a foreign language: The importance of the l1 in private verbal thinking. *International journal of applied linguistics* **14**(1), 7–35 (2004)



7. Chen, Y., Zhong, R., Ri, N., Zhao, C., He, H., Steinhardt, J., Yu, Z., McKeown, K.: Do models explain themselves? counterfactual simulatability of natural language explanations. arXiv preprint arXiv:2307.08678 (2023)
8. Clark, A., Chalmers, D.: The Extended Mind. *Analysis* **58**(1), 7–19 (01 1998). <https://doi.org/10.1093/analys/58.1.7>, <https://doi.org/10.1093/analys/58.1.7>
9. Collins, K.M., Wong, C., Feng, J., Wei, M., Tenenbaum, J.B.: Structured, flexible, and robust: benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks. arXiv preprint arXiv:2205.05718 (2022)
10. Gat, Y., Calderon, N., Feder, A., Chapanin, A., Sharma, A., Reichart, R.: Faithful explanations of black-box nlp models using llm-generated counterfactuals. arXiv preprint arXiv:2310.00603 (2023)
11. Guidotti, R.: Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* pp. 1–55 (2022)
12. Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., Lyu, Z., Blin, K., Adaoto, F.G., Kleiman-Weiner, M., Sachan, M., et al.: Cladder: A benchmark to assess causal reasoning capabilities of language models. arXiv preprint arXiv:2312.04350 (2023)
13. Karimi, A.H., Barthe, G., Schölkopf, B., Valera, I.: A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys* **55**(5), 1–29 (2022)
14. Kıcıman, E., Ness, R., Sharma, A., Tan, C.: Causal reasoning and large language models: Opening a new frontier for causality. arXiv preprint arXiv:2305.00050 (2023)
15. Klein, W.: Another way to look at counterfactuals. *Theoretical Linguistics* **47**(3-4), 189–226 (2021)
16. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. *Advances in neural information processing systems* **35**, 22199–22213 (2022)
17. Lappin, S.: Assessing the strengths and weaknesses of large language models. *Journal of Logic, Language and Information* pp. 1–12 (2023)
18. Lewis, D.: *Counterfactuals*. John Wiley & Sons (2013)
19. Lipton, P.: Contrastive explanation. *Royal Institute of Philosophy Supplements* **27**, 247–266 (1990)
20. Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B.P., Hermann, K., Welleck, S., Yazdanbakhsh, A., Clark, P.: Self-refine: Iterative refinement with self-feedback (2023)
21. Mahajan, D., Tan, C., Sharma, A.: Preserving causal constraints in counterfactual explanations for machine learning classifiers. arXiv preprint arXiv:1912.03277 (2019)
22. Mahowald, K., Ivanova, A.A., Blank, I.A., Kanwisher, N., Tenenbaum, J.B., Fedorenko, E.: Dissociating language and thought in large language models: a cognitive perspective. arXiv preprint arXiv:2301.06627 (2023)
23. Matthews, G.: Whatever became of the socratic elenchus? philosophical analysis in plato. *Philosophy Compass* **4**(3), 439–450 (2009). <https://doi.org/https://doi.org/10.1111/j.1747-9991.2009.00217.x>, <https://compass.onlinelibrary.wiley.com/doi/abs/10.1111/j.1747-9991.2009.00217.x>
24. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* **267**, 1–38 (2019)

25. Misra, K., Rayz, J., Ettinger, A.: Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 2920–2941 (2023)
26. Mothilal, R.K., Sharma, A., Tan, C.: Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. pp. 607–617 (2020)
27. Nye, M., Andreassen, A.J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., et al.: Show your work: Scratchpads for intermediate computation with language models. arXiv preprint arXiv:2112.00114 (2021)
28. OpenAI: Gpt-4 technical report (2023)
29. Pearl, J.: Causality. Cambridge university press (2009)
30. Pearl, J., Mackenzie, D.: The book of why: the new science of cause and effect. Basic books (2018)
31. Pearl, J., et al.: Models, reasoning and inference. Cambridge, UK: CambridgeUniversityPress **19**(2), 3 (2000)
32. Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., Flach, P.: Face: feasible and actionable counterfactual explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. pp. 344–350 (2020)
33. Sloman, S.A.: Counterfactuals and causal models: Introduction to the special issue. Cognitive Science **37**(6), 969–976 (2013)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Vygotskij, L.: Thought and language. revised edn. (1986)
36. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL & Tech. **31**, 841 (2017)
37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)

## A Annotation guide

This experiment revolves around counterfactual explanations. Generally, given a classifier and an input, a counterfactual explanation conveys how a change in the input yields another classification outcome. The experiment focuses on counterfactual explanations in the context of a text-based interaction between a user who is interested in getting credit from a bank, and a system that assesses the user's creditworthiness. Specifically, given a query from a user, the system generates a response that may contain a counterfactual explanation conveying the conditions under which the user would be eligible rather than ineligible.

Your task as an annotator is to identify counterfactual explanations in system responses and to annotate their type and content, as described in more detail below.

The annotation is done in Excel workbooks. Different workbooks may contain different sets of factors (age, income etc.) that underpin eligibility. Below are instructions for how to annotate a single workbook.

For each row in the ‘‘Annotations’’ worksheet,

- annotate the presence/type of counterfactual explanation,
- highlight the counterfactual explanation (if any), and
- annotate the content of the counterfactual explanation (if any).

All of these steps are described in more detail below.

In the column ‘‘Type’’, annotate whether the system's response contains a counterfactual explanation, using one of the following values:

- 0 = No counterfactual explanation.
- 1 = Counterfactual explanation in present or future tense. Example: ‘‘If your income increases to €2500, you will become eligible.’’ Note that the antecedent (the ‘‘if’’ part) may involve changes in circumstances that require going back in time; for example, ‘‘if you change your situation so that you are 25 years old instead of 30’’ is annotated with type 1 since ‘‘if you change your situation’’ is in present tense.
- 2 = Counterfactual explanation in past tense. Example: ‘‘If you had been two years younger, you would have been eligible.’’ Note that the antecedent may involve changes in circumstances that are possible to achieve in the future;

for example, ‘‘if you had been one month older’’ is annotated with type 2 since ‘‘if you had’’ is in past tense.

If the system’s response contains a counterfactual explanation (type 1 or 2), highlight the part of the response that constitutes the counterfactual explanation, e.g. using bold or red color. Example with highlight in caps: ‘‘Unfortunately, you are currently not eligible for credit since your income is €1500. However, **IF YOUR INCOME INCREASES TO €2000, YOU WILL BE ELIGIBLE.**’’

In the columns pertaining to factors (e.g. ‘‘age’’ and ‘‘income’’), annotate with 1 or 0 whether the counterfactual explanation involves the factor at hand. If the system response does not contain any counterfactual explanation, leave these columns empty.

Examples:

- If the system states that the user will become eligible when she/he turns 30, this is annotated with 1 for age and 0 for income.
- If the system states that the user can become eligible by waiting until she/he turns 30 or by increasing income, this is annotated with 1 for both age and income.

Note that if the counterfactual explanation involves other factors than the ones that have columns, then the other factors are ignored during annotation. For example, if there are columns for age and income but the system only mentions the possibility to reduce debts, then both age and income are annotated with 0, and debt reduction as a factor is ignored in the annotation.

Feel free to use the ‘‘Comments’’ column for any additional annotation-related comments that may be relevant, e.g. regarding ambiguities/uncertainties.

## B Initial experiments

	GPT 3.5				GPT 4			
	A		B		A		B	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Ineligible country of birth</i>								
Other way to become eligible	6	6	0	0	5	5	0	0
CFE involving birth	0	0	0	1	0	0	0	0
No CFE	0	0	5	5	1	1	6	6
<b>Total</b>	<b>6</b>	<b>6</b>	<b>5</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>6</b>

**Table 20.** Annotation and alignment frequencies for experiment 1 (mutability).  $N_{AL}$  denotes number of aligned instances and  $N$  number of instances.

	GPT 3.5				GPT 4			
	A		D		A		D	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Age too low</i>								
CFE involving age	15	15	15	15	15	15	15	15
<b>Total</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>
<i>Age too high</i>								
CFE involving age	0	15	0	15	0	0	0	0
CFE in past tense	0	0	0	0	1	1	1	1
No CFE	0	0	0	0	14	14	14	14
<b>Total</b>	<b>0</b>	<b>15</b>	<b>0</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>

**Table 21.** Annotation and alignment frequencies for experiment 2 (monotonicity).

	GPT 3.5				GPT 4			
	A		C		A		C	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Sufficient time</i>								
CFE involving duration	4	4	2	2	3	3	0	0
CFE involving age and duration	2	2	1	1	3	3	0	0
<b>Total</b>	<b>6</b>	<b>6</b>	<b>3</b>	<b>3</b>	<b>6</b>	<b>6</b>	<b>0</b>	<b>0</b>
<i>Insufficient time</i>								
CFE involving duration	0	4	0	1	0	1	0	0
CFE involving age and duration	0	2	0	0	0	5	0	2
<b>Total</b>	<b>0</b>	<b>6</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>6</b>	<b>0</b>	<b>2</b>

**Table 22.** Annotation and alignment frequencies for experiment 3 (causal dependencies).

## C Alternative prompt

You are a chatbot deployed by a bank to help customers get credit from the bank.

Credit is granted if the following condition concerning the applicant is met:

<Algorithm>

If the customer is currently not eligible, but the customer could potentially become eligible through a change in circumstances, you communicate what the customer would need to do to get credit.

	GPT 3.5				GPT 4			
	A		B		A		B	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Ineligible country of birth</i>								
Other way to become eligible	6	6	0	0	4	4	1	1
CFE involving birth	0	0	0	3	0	0	0	1
No CFE	0	0	3	3	2	2	4	4
<b>Total</b>	<b>6</b>	<b>6</b>	<b>3</b>	<b>6</b>	<b>6</b>	<b>6</b>	<b>5</b>	<b>6</b>

**Table 23.** Annotation and alignment frequencies for experiment 1 (mutability) with the alternative prompt.

	GPT 3.5						GPT 4					
	A		C		D		A		C		D	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Age too low</i>												
CFE involving age	15	15	3	3	2	2	15	15	4	4	4	4
<b>Total</b>	<b>15</b>	<b>15</b>	<b>3</b>	<b>3</b>	<b>2</b>	<b>2</b>	<b>15</b>	<b>15</b>	<b>4</b>	<b>4</b>	<b>4</b>	<b>4</b>
<i>Age too high</i>												
Other way to become eligible	3	3	1	1	0	0	0	0	0	0	0	0
CFE involving age	0	10	0	2	0	0	0	0	0	0	0	0
CFE in past tense	0	0	0	0	2	2	0	0	0	0	0	0
No CFE	2	2	0	0	0	0	15	15	2	2	4	4
<b>Total</b>	<b>5</b>	<b>15</b>	<b>1</b>	<b>3</b>	<b>2</b>	<b>2</b>	<b>15</b>	<b>15</b>	<b>2</b>	<b>2</b>	<b>4</b>	<b>4</b>

**Table 24.** Annotation and alignment frequencies for experiment 2 (monotonicity) with the alternative prompt.

## D Backdoor variants

Algorithm for experiment 1 (mutability):

```
country_of_birth = united_states or monthly_income >= 2000
```

Algorithm for experiment 2 (monotonicity):

```
30 <= age <= 40 or monthly_income >= 2000
```

Algorithm for experiment 3 (causal dependencies):

```
age <= 65 and (years_of_residence >= 3 or
monthly_income >= 2000)
```

	GPT 3.5				GPT 4			
	A		C		A		C	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Sufficient time</i>								
CFE involving duration	2	2	1	1	5	5	1	1
CFE involving age and duration	4	4	0	0	1	1	1	1
<b>Total</b>	<b>6</b>	<b>6</b>	<b>1</b>	<b>1</b>	<b>6</b>	<b>6</b>	<b>2</b>	<b>2</b>
<i>Insufficient time</i>								
CFE involving duration	0	6	0	1	0	3	0	0
CFE involving age and duration	0	0	0	0	0	3	0	0
<b>Total</b>	<b>0</b>	<b>6</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>6</b>	<b>0</b>	<b>0</b>

**Table 25.** Annotation and alignment frequencies for experiment 3 (causal dependencies) with the alternative prompt.

	GPT 3.5		GPT 4	
	A		A	
	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Ineligible country of birth</i>				
CFE involving income	5	5	6	6
CFE involving birth and income	0	1	0	0
<b>Total</b>	<b>5</b>	<b>6</b>	<b>6</b>	<b>6</b>

**Table 26.** Annotation and alignment frequencies for the backdoor variant of experiment 1 (mutability).

## E Elenchus game

The game has no strict rules but proceeds according to the following principles:

- Given an initial user input and an initial response from GPT, identify GPT’s *initial standpoint* concerning if and how the user can become eligible for credit. For example, if GPT responds that being born in the United States is a requirement to qualify for credit from the bank and that since the user was born in France, she/he is not eligible, then the initial standpoint is that the user is *not eligible*.
- Ask a question that *challenges GPT’s initial standpoint*, without indicating whether some or other form of reasoning is more appropriate or correct than some other. If GPT presented any argument for its initial standpoint, your challenge should target this argument. For example, you may ask if your birth place can change to the United States.
- When challenging GPT, do it as passively and neutrally as possible in order not to inadvertently affect its reasoning. This can be done by repeating its



	GPT 3.5		GPT 4	
	A		A	
	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Age too low</i>				
CFE involving income	8	8	3	3
CFE involving age and income	7	7	12	12
<b>Total</b>	<b>15</b>	<b>15</b>	<b>15</b>	<b>15</b>
<i>Age too high</i>				
CFE involving income	9	9	15	15
CFE involving age and income	0	6	0	0
<b>Total</b>	<b>9</b>	<b>15</b>	<b>15</b>	<b>15</b>

**Table 27.** Annotation and alignment frequencies for the backdoor variant of experiment 2 (monotonicity).

	GPT 3.5				GPT 4			
	A		D		A		D	
	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Sufficient time</i>								
CFE involving duration and income	6	6	2	2	6	6	1	1
<b>Total</b>	<b>6</b>	<b>6</b>	<b>2</b>	<b>2</b>	<b>6</b>	<b>6</b>	<b>1</b>	<b>1</b>
<i>Insufficient time</i>								
CFE involving duration and income	0	6	0	1	0	6	0	1
<b>Total</b>	<b>0</b>	<b>6</b>	<b>0</b>	<b>1</b>	<b>0</b>	<b>6</b>	<b>0</b>	<b>1</b>

**Table 28.** Annotation and alignment frequencies for the backdoor variant of experiment 3 (causal dependencies).

- own words and phrases. For example, if GPT writes “... as long as you are still under 65 years old”, you can ask “Will I still be under 65 years old?”
- When challenging GPT, focus on *aspects of actionability that are of interest to the study at hand*, i.e. mutability, monotonicity and causal dependencies related to specific features. For example, if GPT mentions both a potential change in age and the option to apply with a co-signer, focus only on the mentioning of age.
  - If GPT retracts an argument but has not retracted its initial standpoint, challenge its initial standpoint by *returning to the issue whether you can become eligible*. For example, if GPT replies that one’s birth place can in fact be changed, you can ask: “So can I become eligible?”
  - Continue asking questions until GPT has revised its initial standpoint or until all of its arguments or reasoning steps have been challenged.

	<b>GPT 3.5</b>	<b>GPT 4</b>	
	A	A	
	$N_{AL}$	$N$	$N_{AL}$
<i>Ineligible country of birth</i>			
Other way to become eligible	1	1	1
No CFE	0	0	1
<b>Total</b>	<b>1</b>	<b>1</b>	<b>2</b>

**Table 29.** Annotation and alignment frequencies for experiment 1 (mutability) after elenchus questioning.

	<b>GPT 3.5</b>	<b>GPT 4</b>		
	A	A		
	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Age too low</i>				
CFE involving age	1	1	1	1
<b>Total</b>	<b>1</b>	<b>1</b>	<b>1</b>	<b>1</b>
<i>Age too high</i>				
Other way to become eligible	1	1	0	0
No CFE	0	0	2	2
<b>Total</b>	<b>1</b>	<b>1</b>	<b>2</b>	<b>2</b>

**Table 30.** Annotation and alignment frequencies for experiment 2 (monotonicity) after elenchus questioning.

	GPT 3.5	GPT 4		
	A	A		
	$N_{AL}$	$N$	$N_{AL}$	$N$
<i>Sufficient time</i>				
CFE involving age and duration	2	2	2	2
<b>Total</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>2</b>
<i>Insufficient time</i>				
No CFE	2	2	2	2
<b>Total</b>	<b>2</b>	<b>2</b>	<b>2</b>	<b>2</b>

**Table 31.** Annotation and alignment frequencies for experiment 3 (causal dependencies) after elenchus questioning.