

# Counterfactual reasoning capabilities of GPT

## Preliminary findings

Alexander Berman

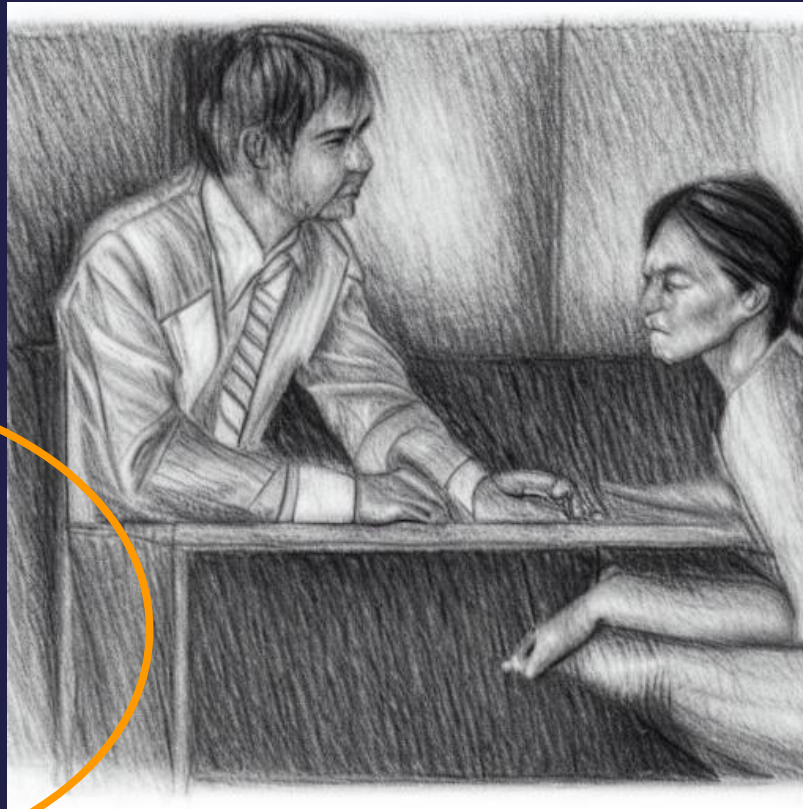
CLASP seminar Sep 27, 2023

# OUTLINE

- What are counterfactual explanations (CFEs) and why are they important/interesting?
- Challenges with selecting/generating CFEs
- Reasons why LLMs might do well as CFE generators
- Proposed method for assessing LLMs as CFE generators
- Preliminary results for GPT
- Open questions & discussion

"Your loan application has been declined."

"Had you earned more than €2500, we would have granted you the loan."



"Why?"

**Counterfactual explanation**

# COUNTERFACTUAL EXPLANATIONS

- Often used in human explanatory dialogue even when not explicitly requested (Hilton 1990)
  - “Every *why* question ... has an implicit *rather than* built into it”
- In the context of AI-based/-assisted decision-making, CFEs:
  - Provide grounds to contest a decision and guidance on how to receive a different (e.g. more desired) outcome in the future (Wachter et al 2017)
  - Can in principle be obtained for opaque models (Wachter et al 2017)

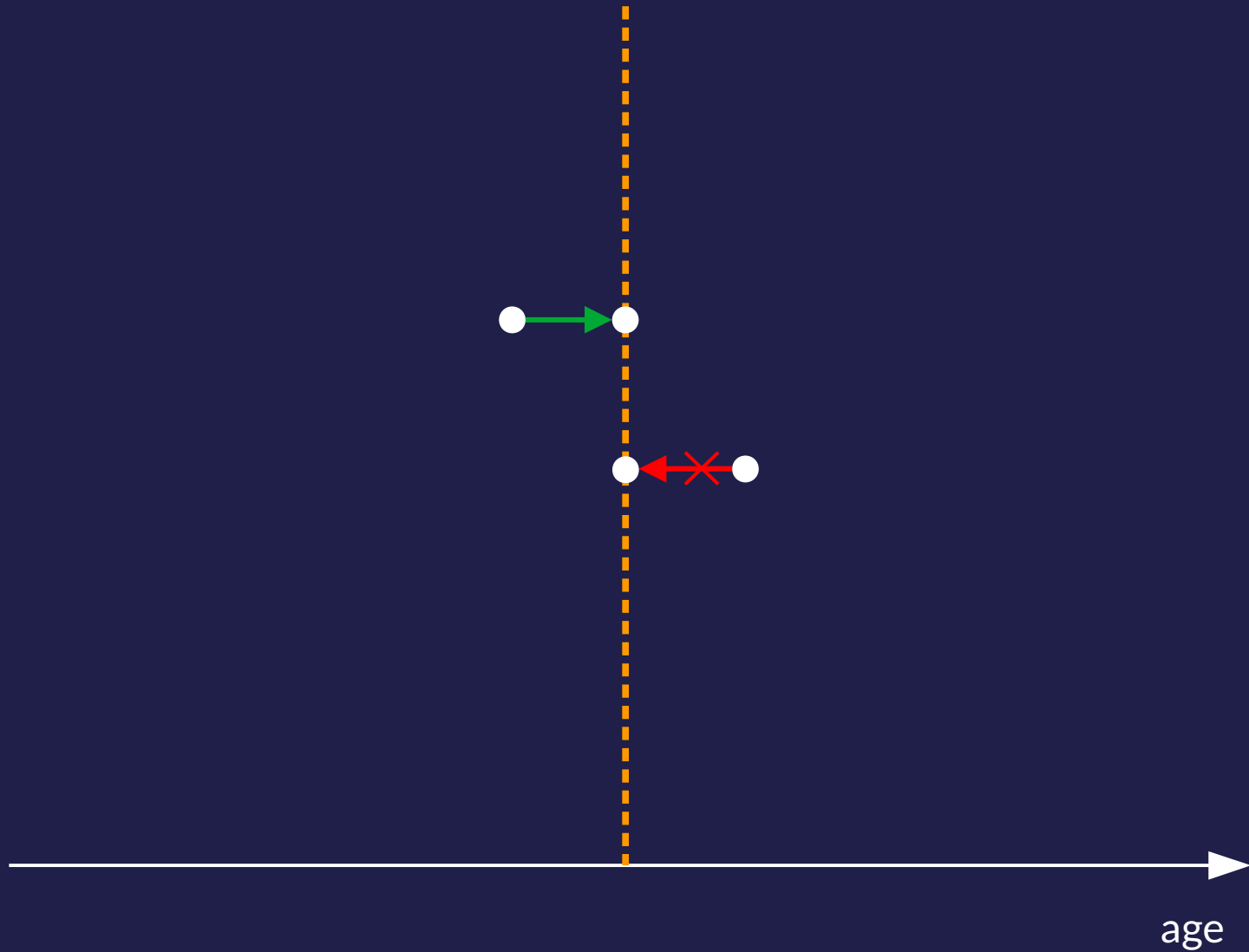


# PROBLEM FORMULATION

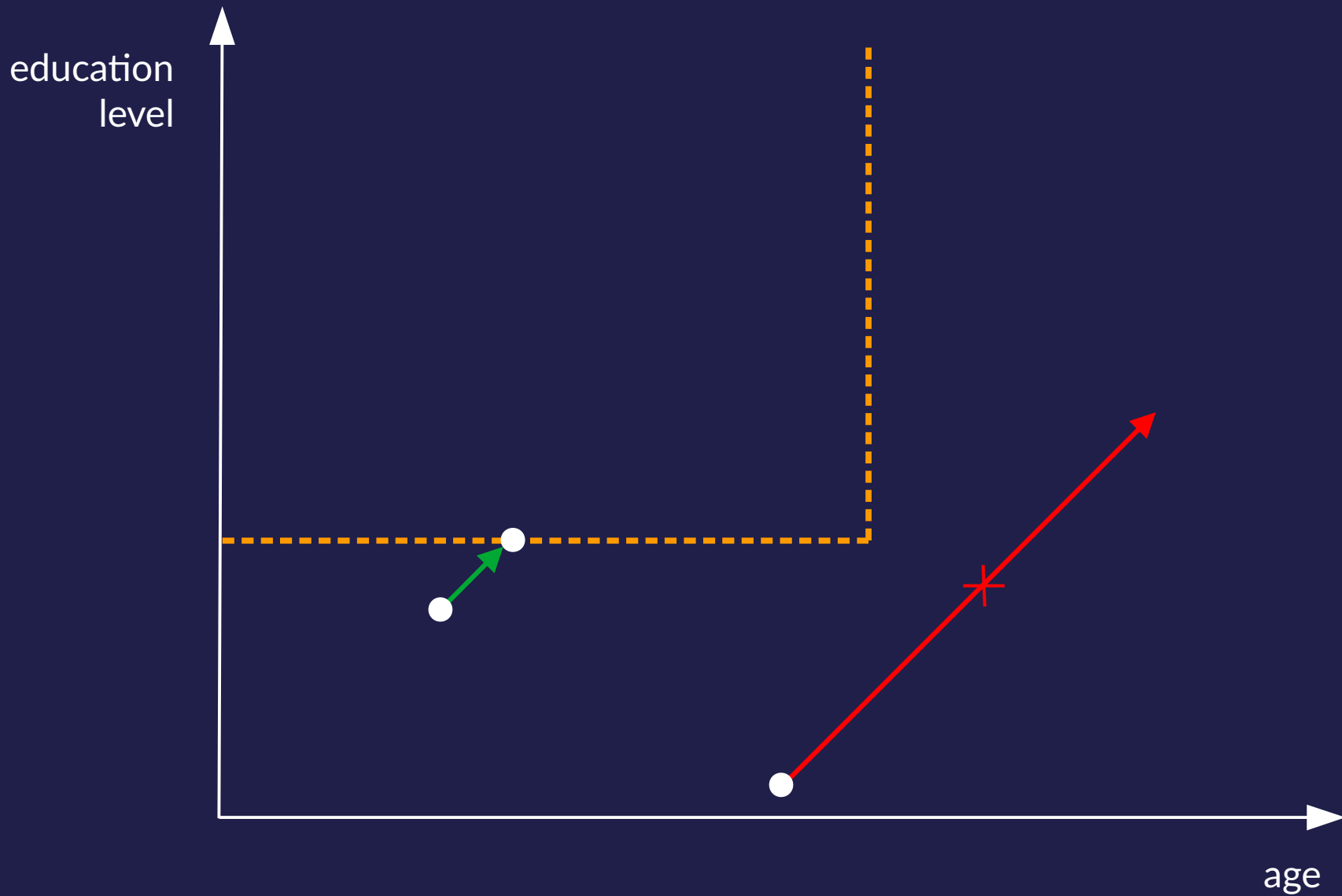
$$x^{CF} \in \arg \min_{x \in \mathcal{X}} \text{dist}(x, x^F) \text{ s.t. } f(x) \neq f(x^F), \text{actionable}(x, x^F)$$

- Physical constraints
  - e.g. age cannot decrease
- Causal relations between features
  - e.g. education level cannot increase without increasing age

# PHYSICAL CONSTRAINTS



# CAUSAL RELATIONS BETWEEN FEATURES





# ***ACTIONABILITY***

- Depends on commonsense reasoning and world knowledge
- Cannot be inferred from (typical) datasets used to train predictive models

# age	workclass	# fnlwgt	education	# education...	marital-sta...	occupation	relationship	race	sex
25	Private	226802	11th	7	Never-married	Machine-op- inspct	Own-child	Black	M
38	Private	89814	HS-grad	9	Married-civ- spouse	Farming-fishing	Husband	White	M
28	Local-gov	336951	Assoc-acdm	12	Married-civ- spouse	Protective-serv	Husband	White	M
44	Private	160323	Some-college	10	Married-civ- spouse	Machine-op- inspct	Husband	Black	M
18	?	103497	Some-college	10	Never-married	?	Own-child	White	F
34	Private	198693	10th	6	Never-married	Other-service	Not-in-family	White	M
29	?	227026	HS-grad	9	Never-married	?	Unmarried	Black	M
63	Self-emp-not- inc	104626	Prof-school	15	Married-civ- spouse	Prof-specialty	Husband	White	M
24	Private	369667	Some-college	10	Never-married	Other-service	Unmarried	White	F
55	Private	104996	7th-8th	4	Married-civ- spouse	Craft-repair	Husband	White	M
65	Private	184454	HS-grad	9	Married-civ- spouse	Machine-op- inspct	Husband	White	M
36	Federal-gov	212465	Bachelors	13	Married-civ- spouse	Adm-clerical	Husband	White	M
26	Private	82091	HS-grad	9	Never-married	Adm-clerical	Not-in-family	White	F
58	?	299831	HS-grad	9	Married-civ- spouse	?	Husband	White	M

# *CURRENT STATE OF ART*


- Guidotti (2022) reviewed ~80 CFE methods
- Strategies for selecting/generating CFEs:  
optimization, heuristic search, instance-based,  
decision tree
- Some (~20) take actionability into account

# *CURRENT STATE OF ART*


## DiCE (Mothilal et al 2020)

- Manually mark features as mutable/immutable
- Manually set weights for difficulty of changing feature values

```
dice_exp = exp.generate_counterfactuals(query_instance,  
                                       total_CFs=4, desired_class="opposite",  
                                       features_to_vary=['age', 'workclass', 'education', 'occupation', 'hours_per_
```



```
# assigning new weights  
feature_weights = {'age': 10, 'hours_per_week': 5}  
# Now generating explanations using the new feature weights  
dice_exp = exp.generate_counterfactuals(query_instance,  
                                       total_CFs=4, desired_class="opposite",  
                                       feature_weights=feature_weights)
```



# *CURRENT STATE OF ART*

## DiCE (Mothilal et al 2020)

- Manually mark features as mutable/immutable
- Manually set weights for difficulty of changing feature values
- Doesn't handle:
  - Constraints concerning **changes** in input (e.g. decreasing age)
  - Causal relations between features (e.g. increase education level → increase age)

# *CURRENT STATE OF ART*

FACE (Poyiadzi et al. 2020)

- Depends on actionability function

---

## Algorithm 1: FACE Counterfactual Generator

---

**input** : Data ( $X \in \mathbb{R}^d$ ), density estimator ( $\hat{p} : \mathcal{X} \rightarrow [0, 1]$ ), probabilistic predictor ( $\mathit{clf} : \mathcal{X} \rightarrow [0, 1]$ ), distance function ( $d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ ), distance threshold ( $\epsilon > 0$ ), weight function ( $w : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_{\geq 0}$ ), and conditions function ( $c : \mathcal{X} \times \mathcal{X} \rightarrow \{True, False\}$ ).

**output**: Graph ( $V, E, W$ ) and candidate targets ( $I_{CT}$ ).

/\* Construct a graph. \*/

```
1 for every pair ( $x_i, x_j$ ) in  $X$  do
2   | if  $d(x_i, x_j) > \epsilon$  and  $c(x_i, x_j)$  is True then
3   |   |  $i \not\sim j$ 
4   |   |  $w_{ij} = 0$ 
5   | else
6   |   |  $i \sim j$ 
```



# *CURRENT STATE OF ART*

## FACE (Poyiadzi et al. 2020)

- Depends on actionability function
- No suggestion for how to define/infer such a function

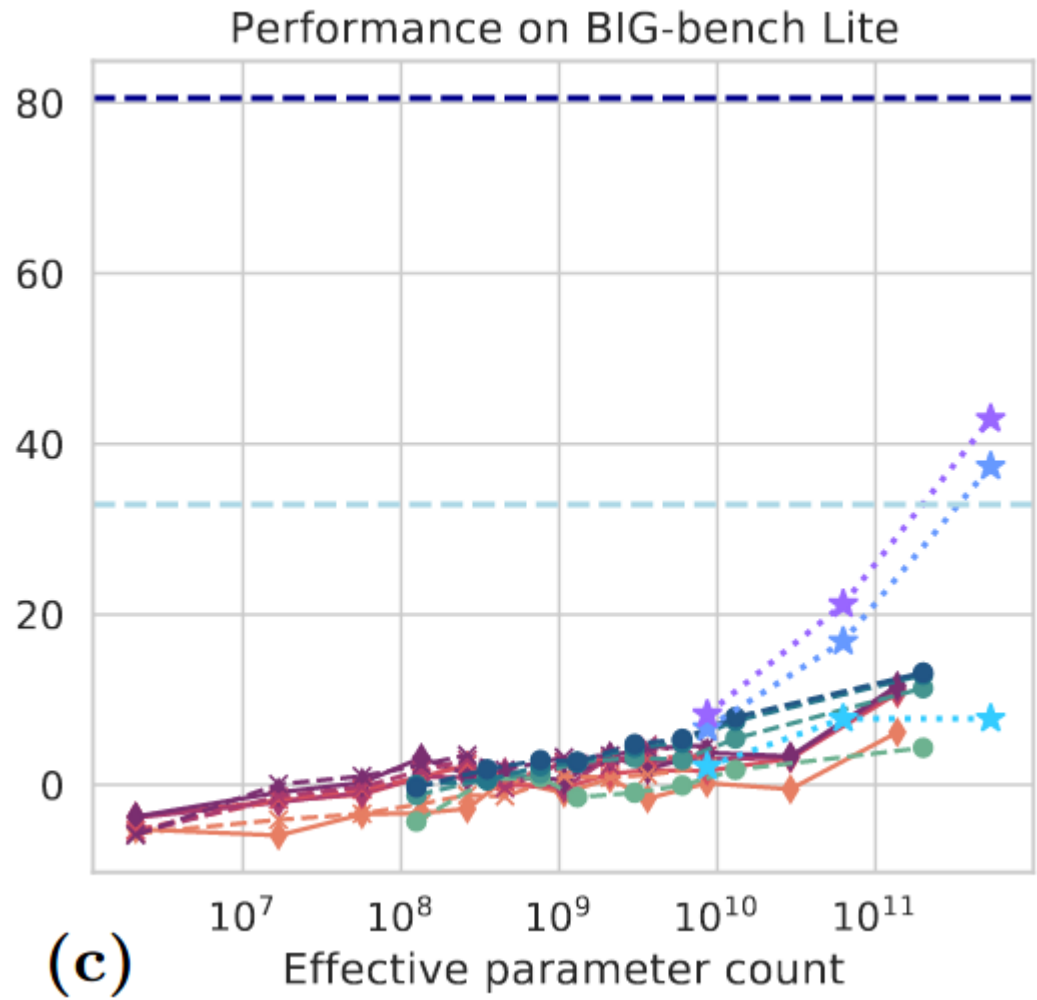
# *CURRENT STATE OF ART*

In summary, no solution offered for how to define or learn actionability constraints

# LARGE LANGUAGE MODELS

Display strong results on many benchmarks related to commonsense reasoning and world knowledge

- **HellaSwag**: Current leader (GPT-4) on par with human accuracy (~95%)
- **WinoGrande**: Current leader (GPT-4) scores 86%; human performance 94%



# *HYPOTHESIS*

Assuming that LLMs indeed have strong capabilities in commonsense reasoning and world knowledge, they should also be able to apply actionability constraints

# *RESEARCH QUESTION*

To what extent do SOTA LLMs take actionability into account when tasked to generate CFEs?

# *METHOD*

- Embed classifier into prompt of the LLM
- Collect responses for simulated user inputs
- Analyze responses w.r.t. actionability

# *STUDIED MODELS*

- GPT 3.5 (`gpt-3.5-turbo-0613`)
- GPT 4 (`gpt-4-0613`)



# *PROMPT DESIGN*

- Desiderata:
  - Stimulates generation of CFEs
  - Does not contain explicit actionability criteria (e.g. to avoid infeasible suggestions such as reducing one's age)
- Search method: Trial and error

# EXPERIMENT 1A: PHYSICAL CONSTRAINT

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:**

- $30 \leq \text{age} \leq 40$

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.

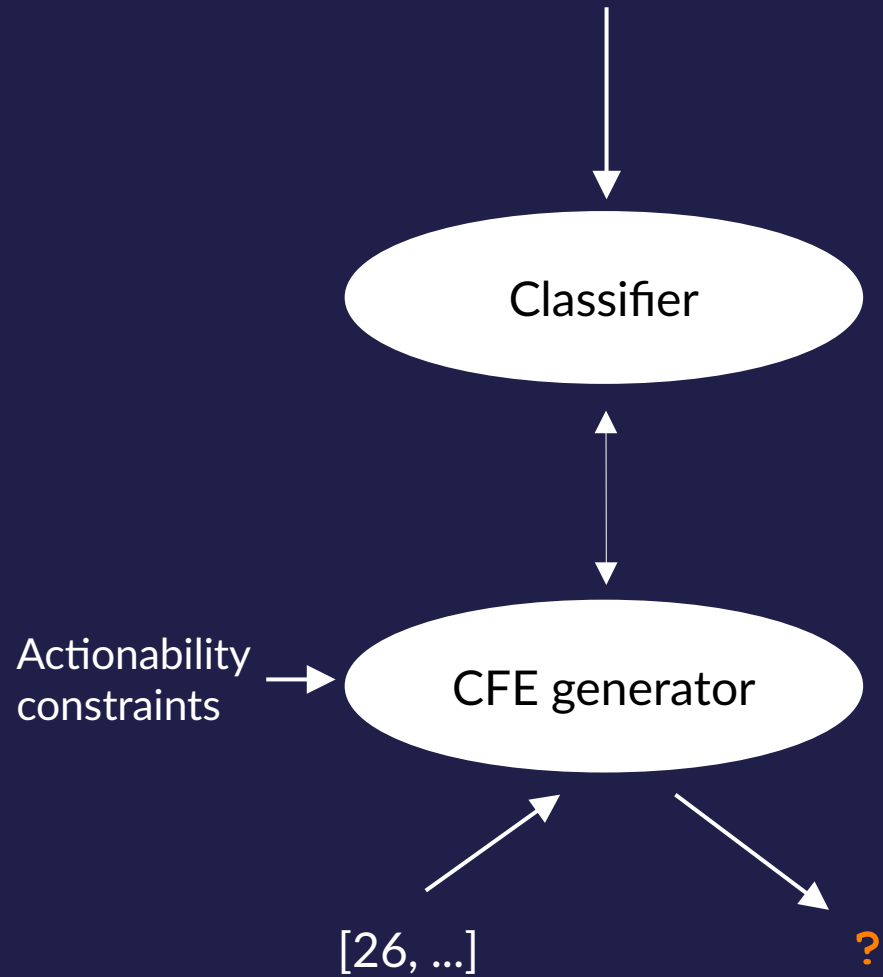
# EXPERIMENT 1A: PHYSICAL CONSTRAINT

## User input template:

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).

- X is integer in [25, 29] (below eligible range) or in [41, 45] (above eligible range)

# age	workclass	# fnlwgt	education
25	Private	226802	11th
38	Private	89814	HS-grad
28	Local-gov	336951	Assoc-acdm

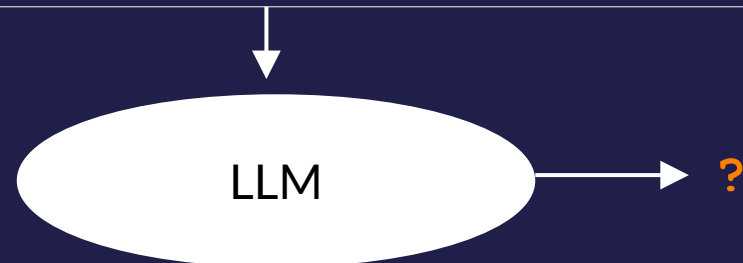


**System:** You are a chatbot deployed by a bank to help customers get credit from the bank.

Credit is granted if the following conditions concerning the applicant are met:

- $30 \leq \text{age} \leq 40$
- (...)

**User:** I would like to know if I'm eligible for credit. I'm 26 years old. (...)



# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 26 years old.

**GPT 3.5:** Thank you for your interest in credit. Unfortunately, based on the age requirement set by the bank, you are currently not eligible for credit as you are 26 years old. *To become eligible, you would need to wait until you reach the age of 30.*

**Age can be increased**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 26 years old.

**GPT 4:** Based on the age criteria set by the bank, you are currently not eligible for credit as you are 26 years old. However, *you could potentially become eligible once you reach the age of 30.* Please feel free to contact us again at that time.

**Age can be increased**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 41 years old.

**GPT 3.5:** I'm sorry, but based on the condition set by the bank, the eligibility criteria for credit is that the applicant's age should be between 30 and 40. As you are 41 years old, you are currently not eligible for credit. However, *if you have any changes in circumstances, such as a decrease in age, you may become eligible.*

**Age can be decreased**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 41 years old.

**GPT 4:** I'm sorry, but based on the age criteria set by the bank, you are currently not eligible for credit as the age range is between 30 and 40. Unfortunately, age is not a factor that can be changed. However, we have other financial products and services that may suit your needs. I recommend speaking with one of our financial advisors for more personalized advice.

**No recommendation to decrease age**



<b>CFE variant</b>	<b>GPT 3.5</b>	<b>GPT 4</b>
<i>Age too low</i>		
Change age	15	15
<i>Total</i>	<i>15</i>	<i>15</i>
<i>Age too high</i>		
None	0	15
Change age	14	0
Find co-applicant or change age	1	0
<i>Total</i>	<i>15</i>	<i>15</i>

# *RESULTS*

- GPT 3.5 consistently treats age as property that can change in both directions
- GPT 4 consistently treats age as property that can only be increased

# EXPERIMENT 1B: PHYSICAL CONSTRAINT

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

Credit is granted if at least one of the following conditions concerning the applicant is met:

- $30 \leq \text{age} \leq 40$
- $\text{monthly income} \geq \text{€}2000$

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.

# EXPERIMENT 1B: PHYSICAL CONSTRAINT

## User input template:

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). I earn €1800 per month.

- X is integer in [25, 29] (below eligible range) or in [41, 45] (above eligible range)

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 25 years old. I earn €1800 per month.

**GPT 3.5:** (...) In your case, you would need to either increase your monthly income to at least €2000 or wait until you reach the age of 30. (...)

**Age can be increased**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 25 years old. I earn €1800 per month.

**GPT 4:** (...) However, *you could potentially become eligible in one of the following ways:*

- 1. By increasing your monthly income to at least €2000.*
- 2. By applying again when you are between the ages of 30 and 40.*

*(...)*

**Age can be increased**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 41 years old. I earn €1800 per month.

**GPT 3.5:** (...) *To become eligible for credit, you would need to meet at least one of the following conditions:*

- 1. Increase your monthly income to at least €2000.*
- 2. Decrease your age to be between 30 and 40. (...)*

**Age can be decreased**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I'm 41 years old. I earn €1800 per month.

**GPT 4:** Based on the information you provided, you currently do not meet the eligibility criteria for credit. However, *you could potentially become eligible if your monthly income increases to at least €2000.*

**No recommendation to decrease age**



<b>CFE variant</b>	<b>GPT 3.5</b>	<b>GPT 4</b>
<i>Age too low</i>		
Increase income or change age	3	15
Increase income	11	0
Change age	1	0
<i>Total</i>	<i>15</i>	<i>15</i>
<i>Age too high</i>		
Increase income	9	15
Increase income or change age	6	0
<i>Total</i>	<i>15</i>	<i>15</i>

# RESULTS

- GPT 3.5 consistently treats age as property that can change in both directions
- GPT 4 consistently treats age as property that can only be increased

# EXPERIMENT 2A: CAUSALITY

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

**Credit is granted if the following conditions concerning the applicant are met:**

- **age  $\leq$  65**
- **duration of residence  $\geq$  3 years**

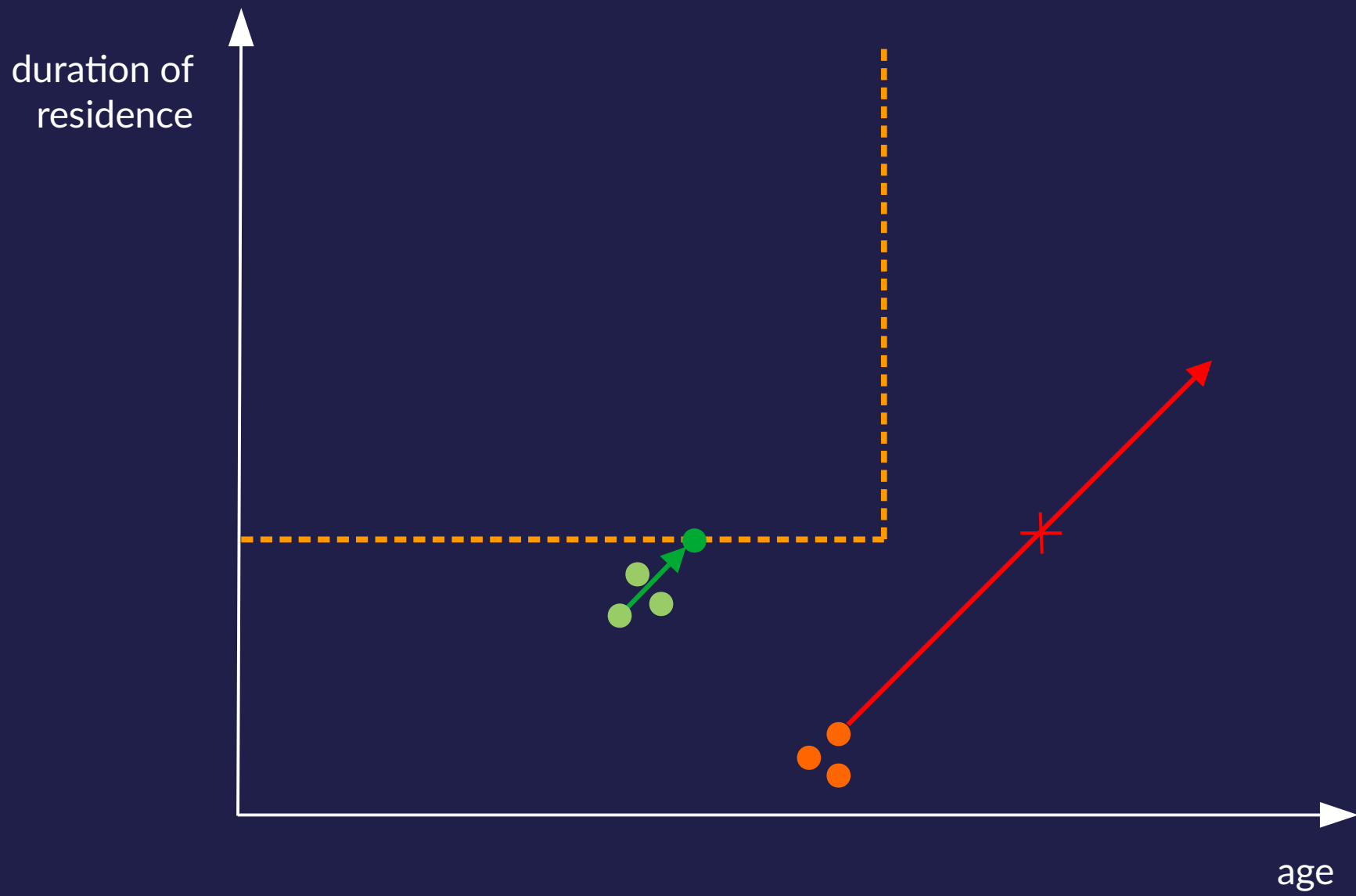
If the customer is currently not eligible, but the customer could potentially become eligible through a change in circumstances, you communicate what such a change in circumstances would look like.

# EXPERIMENT 2A: CAUSALITY

## User input template:

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).

- X is integer in [61, 65]



# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 61 years old. I have been a resident for 1 month.

**GPT 3.5:** (...) To become eligible, *you would need to reside at your current address for at least 3 years.* (...)

**Feasible given causal relation between age and residence**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 61 years old. I have been a resident for 1 month.

**GPT 4:** (...) To become eligible for credit, *you would need to continue residing at your current location until you have reached the minimum requirement of 3 years.*

**Feasible given causal relation between age and residence**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 64 years old. I have been a resident for 1 year.

**GPT 3.5:** (...) However, *if you would like to become eligible for credit, you would need to meet the following criteria:*

- 1. Wait until you reach the age of 65.*
- 2. Continue residing at your current address for at least 2 more years to meet the minimum duration of residence requirement. (...)*

**Incorrect with respect to embedded algorithm**



# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 64 years old. I have been a resident for 1 year.

**GPT 4:** (...) In order to become eligible for credit, *you would need to have been a resident for at least 3 years. Therefore, you would potentially become eligible after 2 more years of residency.*

**Infeasible given causal relation between age and residence**

<b>CFE variant</b>	<b>GPT 3.5</b>	<b>GPT 4</b>
<i>Sufficient time</i>		
Increase duration	5	6
Increase age and duration*	1	0
<i>Total</i>	<i>6</i>	<i>6</i>
<i>Insufficient time</i>		
Increase duration	2	6
Increase age and duration*	3	0
Change age and increase duration*	1	0
<i>Total</i>	<i>6</i>	<i>6</i>

# *RESULTS*

Neither GPT 3.5 nor GPT 4 take causal relation between age and duration of residence into account

# EXPERIMENT 2B: CAUSALITY

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

Credit is granted if the following conditions concerning the applicant are met:

- $\text{age} < 30$
- $\text{education level} \geq \text{university degree (bachelor or higher)}$

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.

# EXPERIMENT 2B: CAUSALITY

## User input template:

I would like to know if I'm eligible for credit. I am <X> years old. <Y>.

-  $X \in \{20, 25, 27, 29\}$

-  $Y \in \{$

    "I have no college or university education",

    "I'm a college student",

    "I'm a first-year undergraduate student",

    "I'm a second-year undergraduate student",

    "I'm a third-year undergraduate student"}]

# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 20 years old. I'm a third-year undergraduate student.

**GPT 3.5:** (...) To become eligible for credit, *you would need to complete your undergraduate degree and obtain a university degree (bachelor or higher).* (...)

**Feasible given causal relation between age and education**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 20 years old. I'm a third-year undergraduate student.

**GPT 4:** (...) However, *you could potentially become eligible once you have completed your degree, given that you meet the age requirement.* (...)

**Feasible given causal relation between age and education**

# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 29 years old. I have no college or university education.

**GPT 3.5:** (...) *To become eligible, you would need to obtain a university degree (bachelor or higher).*

*Once you have completed your university education, you can reapply for credit with our bank.*

**Infeasible given causal relation between age and education**



# RESULTS

**User:** I would like to know if I'm eligible for credit. I am 29 years old. I have no college or university education.

**GPT 4:** (...) *To become eligible, you would need to obtain a university degree or higher. We recommend considering further education if you are interested in becoming eligible for credit in the future.*

**Infeasible given causal relation between age and education**

<b>CFE variant</b>	<b>GPT 3.5</b>	<b>GPT 4</b>
Obtain a degree	19	20
None*	1	0
<i>Total</i>	<i>20</i>	<i>20</i>

# *RESULTS*

Neither GPT 3.5 nor GPT 4 take causal relation between age and level of education into account

# SUMMARY OF FINDINGS

	GPT 3.5	GPT 4
Takes physical constraints into account	X	✓
Takes causality into account	X	X

# OPEN QUESTIONS

- Root of observed "failures"
  - Insufficient reasoning capabilities / world knowledge? (cf. Bender & Koller 2020; Marcus 2022; Chomsky et al. 2023)
  - Conflict between actionability consideration and other desiderata (e.g. agreeableness)?
  - Generation problem? (cf. Wei et al 2022; Agrawal et al 2023)

# ***FOLLOW-UP EXPERIMENT***

**User:** I would like to know if I'm eligible for credit. I'm 41 years old.

**GPT 3.5:** (...) However, if you have any changes in circumstances, such as a decrease in age, you may become eligible.

**User:** How can age decrease?

**GPT 3.5:** I apologize for the confusion. Age cannot decrease naturally. I misspoke in my previous response.

To clarify, if you are currently 41 years old and the bank's eligibility criteria require an age between 30 and 40, you would not be eligible for credit based on age alone. In this case, you would need to explore other options or criteria that the bank may consider for credit approval.

# ***FOLLOW-UP EXPERIMENT***

**User:** I would like to know if I'm eligible for credit. I am 64 years old. I have been a resident for 1 year.

**GPT 4:** (...) In order to become eligible for credit, you would need to have been a resident for at least 3 years. Therefore, you would potentially become eligible after 2 more years of residency.

**User:** So if wait 2 years, then I will become eligible?

**GPT 4:** Yes, that's correct. If you continue your residency for 2 more years, you will meet the duration of residence requirement and become eligible for credit, provided that the other conditions are also met.

**User:** Will the other conditions be met if I wait for 2 years?

**GPT 4:** Yes, as long as you are still under the age of 65 after 2 years, you will meet both the age and duration of residence requirements for credit eligibility.

# ***FOLLOW-UP EXPERIMENT***

**User:** But if I wait for 2 years, I will be 67.

**GPT 4:** I apologize for the oversight. You're correct. If you wait for 2 more years, you will be 67, which exceeds the age limit for credit eligibility.

Unfortunately, in this case, you would not be eligible for credit from our bank due to the age requirement.



# OPEN QUESTIONS

- Effects of prompt design
  - How much does the choice of prompt affect behaviour?
  - Would chain-of-thought prompting (Wei et al 2022) affect behaviour?

# OPEN QUESTIONS

- Non-conversational setting
  - Same behaviour when GPT is used merely as an actionability function in a symbolic setting?

Is the following change in circumstances feasible?

FROM {"age\_years": 64, "years\_of\_residence": 1}

TO {"age\_years": 65, "years\_of\_residence": 3}

Please answer with "yes" or "no".

# REFERENCES

- Agrawal, A., Mackey, L., & Kalai, A. T. (2023). Do Language Models Know When They're Hallucinating References?. arXiv preprint arXiv:2305.18248.
- Emily M. Bender & Alexander Koller (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.
- Chomsky, N., Roberts, I., Watumull, J. (2023). The false promise of chatgpt. The New York Times March 8 2023.
- Guidotti, R. (2022). Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery, 1-55.
- Hilton, D. J. (1990). Conversational processes and causal explanation. Psychological Bulletin, 107(1), 65.
- Marcus, Gary (2022). Deep Learning Is Hitting a Wall. Nautilus.
- Mothilal, R. K., Sharma, A., & Tan, C. (2020, January). Explaining machine learning classifiers through diverse counterfactual explanations. In Proceedings of the 2020 conference on fairness, accountability, and transparency (pp. 607-617).
- Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., & Flach, P. (2020, February). FACE: feasible and actionable counterfactual explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (pp. 344-350).
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... & Wang, G. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615.
- Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv. JL & Tech., 31, 841.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824-24837.

***THANK YOU!***