Appendices

Appendix A

Imagine you are in this scenario:
"You are a 31-year-old person, who has had one pregnancy, and who has the following medical readings: glucose level: 173 mg/dL, diastolic blood pressure: 82 mm Hg, skin thickness: 48 mm, Insulin: 160 μIU/ml, body mass index (BMI): 32.8."
Current outcome: You have an **increased risk of diabetes.**
Useful context: Normal glucose levels are 70-130, 140-200 is prediabetes, 200+ is diabetes. Healthy BMI is 18.5 - 25, 25-30 is considered overweight, 30+ is considered obese. Normal diastolic blood pressure is roughly below 80 mm Hg. Normal insulin is between 16 and 166 μIU/ml.

"To **not be at increased risk of diabetes** you would need to make the following changes:

- Decrease your **glucose level** from **173** to **130**.
- And, increase your **insulin level** from **160** to **181**."

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

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

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

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

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

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

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

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

Table A.1: Example of a questionnaire question from the final study

Participant id	Survey time	Failed attention check	PCA outlier	Similar answering pattern	Question 16	Question 21	Question 32	Avg. understandability below 3
12		x	x	x	x	x	x	
27	x	x	x	x		x		
28				x	x	x		
44		x		x	x	x		
75					x			x
83		x		x	x	x		x
84			x	x		x		
88				x	x	x	x	x
97				x	x	x	x	x
201			x		x		x	

Table A.2: Dropped participants and corresponding indicators of low-quality. X marks a check that was failed by the specific participant.

Demographic	Description	Number	Percentage (%)
Age Group	18-24 years old	64	31.07
	25-34 years old	95	46.12
	35-44 years old	25	12.14
	45-54 years old	15	7.28
	55-64 years old	7	3.40
Nationality	South African	32	15.53
	Polish	30	14.56
	British	27	13.11
	Mexican	20	9.71
	Portuguese	19	9.22
	Chilean	12	5.83
	Italian	8	3.88
	Greek	7	3.40
	Hungarian	7	3.40
	USA	6	2.91
	Other	38	18.45
Education Level	Primary school or lower	1	0.49
	High school	69	33.50
	Bachelor's degree or equivalent	93	45.15
	Master's degree or equivalent	39	18.93
	Doctoral degree or equivalent	4	1.94
English Proficiency	Native speaker / Fully proficient	166	80.58
	Moderately proficient	40	19.42
Experience in Machine Learning	No experience	100	48.54
	Some experience	89	43.20
	I am studying in a related field	12	5.83
	I work in the field / Extensive experience	5	2.43
Experience with Causality Frameworks	No	145	70.39
	I am familiar with the general concept	57	27.67
	I have previous experience with them	4	1.94
Medical Background	No	181	87.86
	I am currently studying medicine or a related field	10	4.85
	I work with medical data or in a field related to medicine	8	3.88
	I have a degree in medicine or a related field	4	1.94
	I work in the field of medicine	3	1.46

Table A.3: Demographics of the participants.

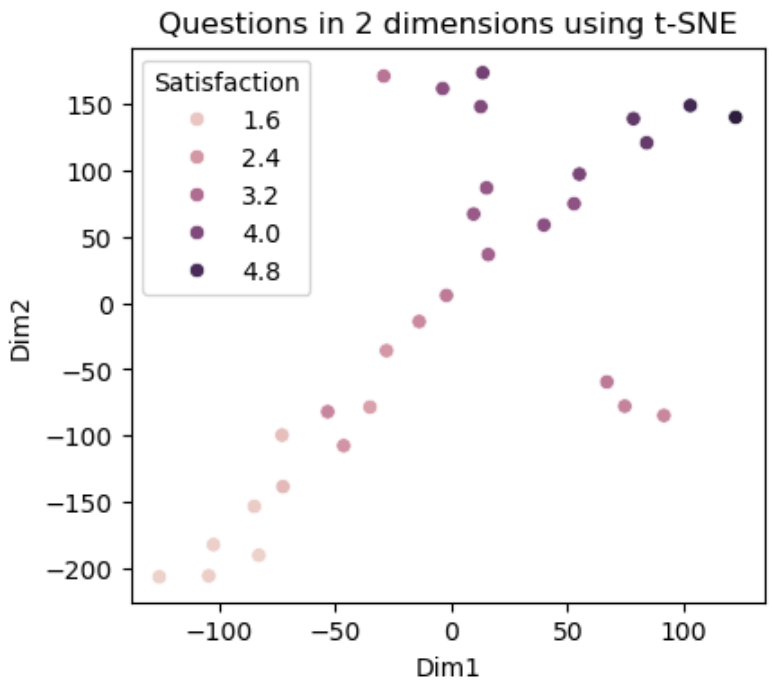


Figure A.1: Questionnaire questions' 7 average metric values (no Overall Satisfaction) reduced to 2 dimensions (t-SNE perplexity 3). Colored by average Overall Satisfaction

Appendix B

Baseline system prompt: You are evaluating counterfactual explanations generated by AI. Counterfactual explanations explain what parameters of a situation should have been different for the outcome to have been different. You are not expected to provide reasoning or explanation and should answer with the appropriate value from the set [”low”, ”medium”, ”high”]. The definition of completeness: the explanation is sufficient in explaining how to achieve the desired outcome. The following is the counterfactual explanation.

System prompt with all definitions: You are evaluating counterfactual explanations generated by AI. Counterfactual explanations explain what parameters of a situation should have been different for the outcome to have been different. You are not expected to provide reasoning or explanation and should answer with the appropriate value from the set [”low”, ”medium”, ”high”]. The definition of satisfaction: this scenario effectively explains how to reach a different outcome. The definition of feasibility: the actions suggested by the explanation are practical, realistic to implement and actionable. The definition of consistency: the parts of the explanation do not contradict each other. The definition of completeness: the explanation is sufficient in explaining how to achieve the desired outcome. The definition of trust: I believe that the suggested changes would bring about the desired outcome. The definition of understandability: I feel like I understood the phrasing of the explanation well. The definition of fairness: the explanation is unbiased towards different user groups and does not operate on sensitive features. The definition of complexity: the explanation has an appropriate level of detail and complexity - not too simple, yet not overly complex. The following is the counterfactual explanation.

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

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

Table B.1: Accuracies for different prompt-model combinations. The highest accuracy for each model is highlighted in bold.

Appendix C

Split	Metric-wise			Question-wise		
Model	Llama 3 70B Instruct	Llama 3 8B Instruct	Llama 3.1 8B Instruct	Llama 3 70B Instruct	Llama 3 8B Instruct	Llama 3.1 8B Instruct
Batch size	8	4	4	8	4	4
Learning rate	0.0002	0.00005	0.00005	0.0001	0.0002	0.0001
Epochs	5	5	6	5	4	5
Hardware	2x NVIDIA Tesla A100 80GB	1x NVIDIA Tesla A100 80GB	1x NVIDIA Tesla A100 80GB	2x NVIDIA Tesla A100 80GB	1x NVIDIA Tesla A100 80GB	1x NVIDIA Tesla A100 80GB

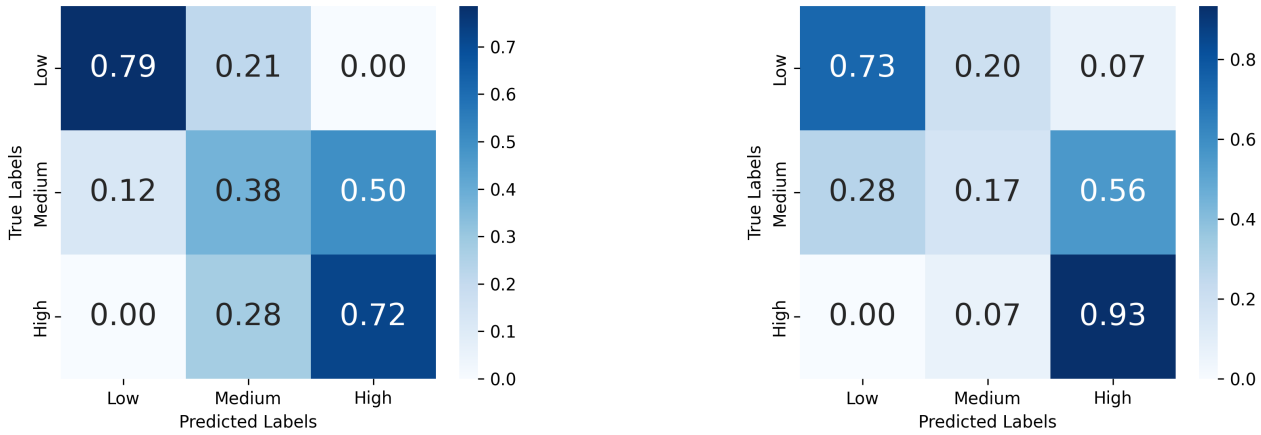
Table C.1: Hyperparameters and hardware used for the fine-tuning of LLMs on averaged human ratings.

Model	Llama 3 70B Instruct
Batch size	8
Learning rate	0.0001
Epochs	3
Hardware	2x NVIDIA Tesla A100 80GB

Table C.2: Hyperparameters and hardware used for the fine-tuning of LLMs on specific participants answers.

r	32
alpha	64
Data type	NF4
Format	4bit

Table C.3: QLoRA parameters used for fine-tuning LLMs.



(a) (b)

Figure C.1: Confusion matrices for GPT4 for metric split (a) and question split (b).

Appendix D

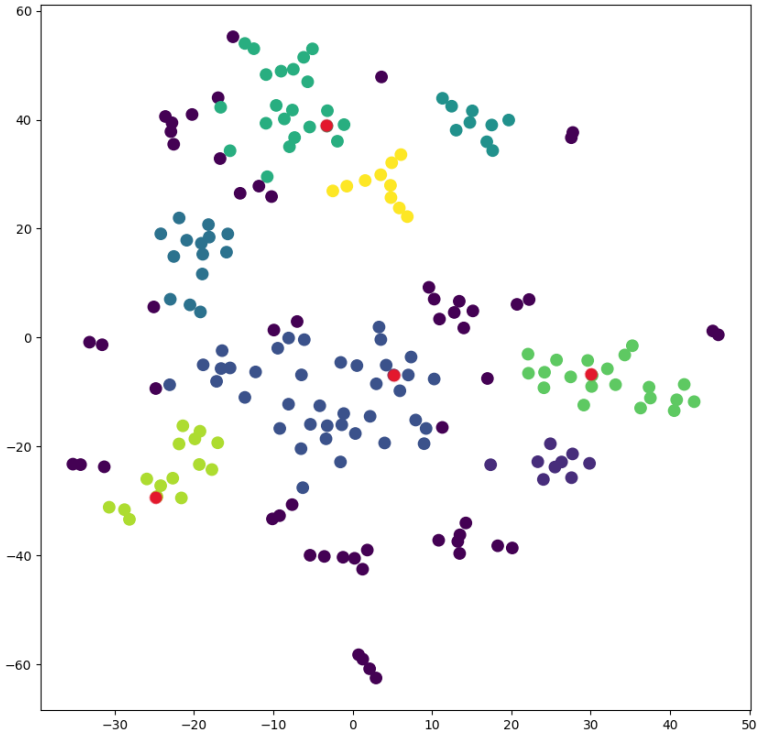


Figure D.1: DBSCAN clustering of participants. The 4 participants chosen for LLM modelling are marked in red.

Participant	A	B	C	D
Age	35-44 years old	35-44 years old	25-34 years old	25-34 years old
Citizenship	Italy	Portugal	Poland	Hungary
English proficiency	Native speaker / Fully proficient	Native speaker / Fully proficient	Native speaker / Fully proficient	Native speaker / Fully proficient
Education	High school	Bachelor's degree or equivalent	Master's degree or equivalent	Master's degree or equivalent
Experience with machine learning	Some experience	No experience	No experience	No experience

Table D.1: Demographic information of individual participants.