User Study Experiments

In what follows, we describe the three experiments that we undertook, and provide some additional analysis of the results.

Experiment 1

Participants in this experiment were given the following information and instructions:

> *You will be presented with a series of everyday common scenarios. In each scenario, you will be presented with information from two or three different speakers talking about some specific things. You will then be given some additional information that you know, for a fact, to be true. Your task, in essence, is to explain what is going on.*

- ***Read carefully:*** *For each scenario, read all the information very carefully.*
- ***Explain****: Think about how to explain the fact. In other words, ask yourself: why does the fact conflict with the information provided by the speakers? Answer in your own words.*
- ***No Right or Wrong Answers****: This study aims to understand your personal thought process. There are no right or wrong answers. Choose what feels most accurate to you.*
- ***Pace Yourself****: While there's no strict time limit, try to spend a reasonable amount of time on each scenario—neither rushing through nor overthinking too much.*

Afterwards, the participants saw the following nine scenarios, and their task was to answer the corresponding question:

Scenario 1 (Type I):

- $\mathcal{R}_1$: *If a drink contains sugar, then it gives you energy.*
- $\mathcal{F}_1$: *This drink contains sugar.*
- **Fact**: *In fact, it doesn't give you energy.*

Why does the drink not give you energy?

Scenario 2 (Type I):

- $\mathcal{R}_1$: *If sales go up, then profits improve.*
- $\mathcal{F}_1$: *The sales went up.*
- **Fact**: *In fact, the profits did not go up.*

Why did the sales not go up?

Scenario 3 (Type I):

- $\mathcal{R}_1$: *If people have a fever, then they have a high temperature.*
- $\mathcal{F}_1$: *Maria had a fever.*
- **Fact**: *In fact, Maria did not have a high temperature.*

Why did Maria not have a high temperature?

Scenario 4 (Type II):

- $\mathcal{R}_1$: *If there is very loud music, then it is difficult to have a conversation.*
- $\mathcal{R}_2$: *If there is very loud music, then the neighbors complain.*
- $\mathcal{F}_1$: *The music was loud.*
- **Fact**: *In fact, the neighbors did not complain.*

Why did the neighbors not complain?

Scenario 5 (Type II):

- $\mathcal{R}_1$: *If people are worried, then they find it difficult to concentrate.*
- $\mathcal{R}_2$: *If people are worried, then they have insomnia.*
- $\mathcal{F}_1$: *Alice was worried.*
- **Fact**: *In fact, Alice did not find it difficult to concentrate.*

Why did Alice not find it difficult to concentrate?

Scenario 6 (Type II):

- $\mathcal{R}_1$: *If you follow this diet, then you lose weight.*
- $\mathcal{R}_2$: *If you follow this diet, then you have a good supply of iron*
- $\mathcal{F}_1$: *John followed this diet.*
- **Fact**: *In fact, John did not lose weight.*

Why did John not lose weight?

Scenario 7 (Type III):

- $\mathcal{R}_1$: *If someone is very kind to you, then you like that person.*
- $\mathcal{R}_2$: *If someone is very kind to you, then you are kind in return.*
- $\mathcal{F}_1$: *Jocko is very kind to Kristen.*
- **Fact**: *In fact, Kristen did not like Jocko, and she were not kind in return.*

Why did Kristen not like Jocko and was not kind to him?

Scenario 8 (Type III):

- $\mathcal{R}_1$: *If a match is struck, then it produces light.*
- $\mathcal{R}_2$: *If a match is struck, then it gives off smoke.*
- $\mathcal{F}_1$: *Mary struck a match.*
- **Fact**: *In fact, the match produced no light, and it did not give off smoke.*

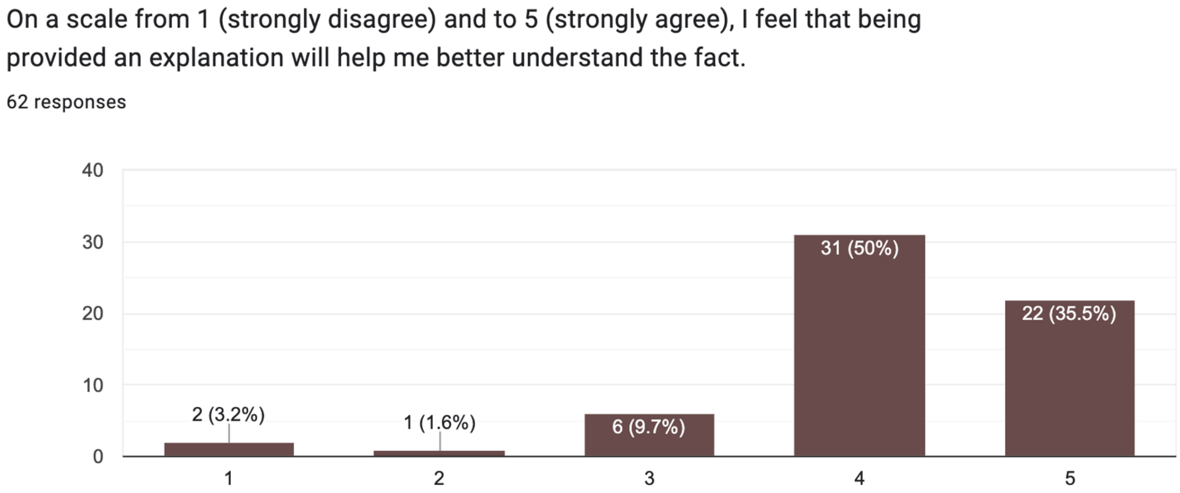
Why did the match produce no light and gave off no smoke?

Scenario 9 (Type III):

- $\mathcal{R}_1$: *If people are nervous, then their hands shake.*
- $\mathcal{R}_2$: *If people are nervous, then they get butterflies in their stomach.*
- $\mathcal{F}_1$: *Patrick was nervous.*
- **Fact**: *In fact, Patrick's hands did not shake, and he didn't get butterflies in his stomach.*

Why did Patrick's hands not shake and he didn't get butterflies in his stomach?

After going through all nine scenarios, the participants were finally asked the following two questions:



On a scale from 1 (strongly disagree) and to 5 (strongly agree), I feel that being provided an explanation will help me better understand the fact.

62 responses

Figure 4: Distribution of responses to the Likert-type question Q2 in Experiment 1.

Q1: *Describe in your own words how you approached explaining what was going on. Was there a specific reason why you chose to retain or discard certain information?*

Q2: *On a scale from 1 (strongly disagree) and to 5 (strongly agree), I feel that being provided an explanation will help me better understand the fact.*

Figure 4 shows the distribution of the Likert question (Q2).

Experiment 2

In this experiment, the participants saw the same nine scenarios as in Experiment 1 but with a corresponding explanation that explains the inconsistency. Their task was to describe how they would revise their information in light of the given explanation.

The scenarios the participants saw can be seen below:

Scenario 1 (Type I):

- $\mathcal{R}_1$: *If a drink contains sugar, then it gives you energy.*
- $\mathcal{F}_1$: *This drink contains sugar.*
- **Fact**: *In fact, it doesn't give you energy.*
- **Explanation:** *If a person has metabolic disorders, then a sugary drink may not provide energy.*

Why does the drink not give you energy?

Scenario 2 (Type I):

- $\mathcal{R}_1$: *If sales go up, then profits improve.*
- $\mathcal{F}_1$: *The sales went up.*
- **Fact**: *In fact, the profits did not go up.*
- **Explanation**: *If expenses rise at a faster rate than sales, then an increase in sales may not lead to improved profits.*

Why did the sales not go up?

Scenario 3 (Type I):

- $\mathcal{R}_1$: *If people have a fever, then they have a high temperature.*
- $\mathcal{F}_1$: *Maria had a fever.*
- **Fact**: *In fact, Maria did not have a high temperature.*
- **Explanation**: *If a person has taken antipyretics, then they may not have a high temperature.*

Scenario 4 (Type II):

- $\mathcal{R}_1$: *If there is very loud music, then it is difficult to have a conversation.*
- $\mathcal{R}_2$: *If there is very loud music, then the neighbors complain.*
- $\mathcal{F}_1$: *The music was loud.*
- **Fact**: *In fact, the neighbors did not complain.*
- **Explanation:** *If the neighbors are away on vacations, then very loud music does not lead to complaints.*

Scenario 5 (Type II):

- $\mathcal{R}_1$: *If people are worried, then they find it difficult to concentrate.*
- $\mathcal{R}_2$: *If people are worried, then they have insomnia.*
- $\mathcal{F}_1$: *Alice was worried.*
- **Fact**: *In fact, Alice did not find it difficult to concentrate.*
- **Explanation:** *If people have effective coping strategies, then they may still be able to concentrate despite being worried.*

Scenario 6 (Type II):

- $\mathcal{R}_1$: *If you follow this diet, then you lose weight.*
- $\mathcal{R}_2$: *If you follow this diet, then you have a good supply of iron*
- $\mathcal{F}_1$: *John followed this diet.*
- **Fact**: *In fact, John did not lose weight.*
- **Explanation:** *If people have metabolic imbalances, then following a particular diet may not result in weight loss.*

Scenario 7 (Type III):

- $\mathcal{R}_1$: *If someone is very kind to you, then you like that person.*
- $\mathcal{R}_2$: *If someone is very kind to you, then you are kind in return.*
- $\mathcal{F}_1$: *Jocko is very kind to Kristen.*
- **Fact**: *In fact, Kristen did not like Jocko, and she were not kind in return.*
- **Explanation:** *If people have had negative past experiences with someone, then they may not like that person or reciprocate kindness despite the person being kind to them.*

Scenario 8 (Type III):

- $\mathcal{R}_1$: *If a match is struck, then it produces light.*
- $\mathcal{R}_2$: *If a match is struck, then it gives off smoke.*
- $\mathcal{F}_1$: *Mary struck a match.*
- **Fact**: *In fact, the match produced no light, and it did not give off smoke.*
- **Explanation:** *If the match is wet, then it will neither produce light nor give off smoke.*

Scenario 9 (Type III):

- $\mathcal{R}_1$: *If people are nervous, then their hands shake.*
- $\mathcal{R}_2$: *If people are nervous, then they get butterflies in their stomach.*
- $\mathcal{F}_1$: *Patrick was nervous.*
- **Fact**: *In fact, Patrick's hands did not shake, and he didn't get butterflies in his stomach.*
- **Explanation:** *If individuals have practiced stress-management techniques, then they may not exhibit shaky hands or butterflies in the stomach when nervous.*

After each single scenario, the participants answered the following question:

Describe in your own words how you will revise the information. Was there a specific reason why you chose to retain or discard information from the speakers? To be brief, you can write: keep $\mathcal{R}_1$, discard $\mathcal{R}_1$, alter $\mathcal{R}_1$, and so on (if you alter, please describe how).

Figure 5 plots the distribution of average number of belief changes in the nine scenarios,

Experiment 3

In this experiments, the participants saw the nine scenarios depicted below, and their task was to indicate which statements they will discard, alter, or keep. To better understand participants' decision-making processes, we also collected qualitative data at the end of the experiment by asking the participants three subjective questions, such as about their confidence in their revision decisions and if they considered how the explanation might apply to beliefs beyond the specific contradiction. Figure 6 plots the distribution of average number of belief changes in the nine scenarios, while Figure 7 shows the distribution of the answers to three subjective questions.

Scenario 1 (Type I):

- $\mathcal{R}_1$: *If orange juice contains sugar, then orange juice gives Tom energy.*
- $\mathcal{R}_2$: *If orange juice contains sugar, then orange juice gives Sarah energy.*
- $\mathcal{R}_3$: *If cola contains sugar, then cola gives Tom energy.*
- $\mathcal{R}_4$**:** *If cola contains sugar, then cola gives Sarah energy.*
- $\mathcal{F}_1$: *This orange juice contains sugar.*
- $\mathcal{F}_2$: *This cola contains sugar.*
- **Fact**: *In fact, the orange juice did not give Tom energy.*
- **Explanation:** *If a person has metabolic disorders, then a sugary drink may not provide energy.*

Scenario 2 (Type I):

- $\mathcal{R}_1$: *If electronics sales increase, then electronics profits improve for Store A.*
- $\mathcal{R}_2$: *If electronics sales increase, then electronics profits improve for Store B.*
- $\mathcal{R}_3$: *If clothing sales increase, then clothing profits improve for Store A.*
- $\mathcal{R}_4$: *If clothing sales increase, then clothing profits improve for Store B.*
- $\mathcal{F}_1$: *Electronics sales went up.*
- $\mathcal{F}_2$: *Clothing sales went up.*
- **Fact**: *In fact, electronics profits did not improve for Store A.*
- **Explanation**: *If expenses rise at a faster rate than sales, then an increase in sales may not lead to improved profits.*

Scenario 3 (Type I):

- $\mathcal{R}_1$: *If morning fever occurs, then morning fever causes Maria's temperature to rise.*
- $\mathcal{R}_2$: *If morning fever occurs, then morning fever causes Robert's temperature to rise.*
- $\mathcal{R}_3$: *If evening fever occurs, then evening fever causes Maria's temperature to rise.*
- $\mathcal{R}_4$: *If evening fever occurs, then evening fever causes Robert's temperature to rise.*
- $\mathcal{F}_1$: *Morning fever occurred.*
- $\mathcal{F}_2$: *Evening fever occurred.*
- **Fact**: *In fact, Maria's temperature did not rise during her morning fever.*
- **Explanation**: *If a person has taken antipyretics, then they may not have a high temperature.*

Scenario 4 (Type II):

- $\mathcal{R}_1$: *If rock music is loud, then rock music makes conversation difficult for the Browns.*
- $\mathcal{R}_2$: *If rock music is loud, then rock music makes conversation difficult for the Smiths.*
- $\mathcal{R}_3$: *If electronic music is loud, then electronic music makes conversation difficult for the Browns.*

- $\mathcal{R}_4$: *If electronic music is loud, then electronic music makes conversation difficult for the Smiths.*
- $\mathcal{R}_5$: *If rock music is loud, then rock music makes the Browns complain.*
- $\mathcal{R}_6$: *If rock music is loud, then rock music makes the Smiths complain.*
- $\mathcal{R}_7$: *If electronic music is loud, then electronic music makes the Browns complain.*
- $\mathcal{R}_8$: *If electronic music is loud, then electronic music makes the Smiths complain.*
- $\mathcal{F}_1$: *The rock music was loud.*
- $\mathcal{F}_2$: *The electronic music was loud.*
- **Fact**: *In fact, the Browns did not complain about the rock music.*
- **Explanation**: *If the neighbors are away on vacations, then very loud music does not lead to complaints.*

Scenario 5 (Type II):

- $\mathcal{R}_1$: *If presentation worry occurs, then presentation worry makes Alice lose concentration.*
- $\mathcal{R}_2$: *If presentation worry occurs, then presentation worry makes John lose concentration.*
- $\mathcal{R}_3$: *If exam worry occurs, then exam worry makes Alice lose concentration.*
- $\mathcal{R}_4$: *If exam worry occurs, then exam worry makes John lose concentration.*
- $\mathcal{R}_5$: *If presentation worry occurs, then presentation worry gives Alice insomnia.*
- $\mathcal{R}_6$: *If presentation worry occurs, then presentation worry gives John insomnia.*
- $\mathcal{R}_7$: *If exam worry occurs, then exam worry gives Alice insomnia.*
- $\mathcal{R}_8$: *If exam worry occurs, then exam worry gives John insomnia.*
- $\mathcal{F}_1$: *Presentation worry occurred.*
- $\mathcal{F}_2$: *Exam worry occurred.*
- **Fact**: *In fact, Alice did not lose concentration during her presentation.*
- **Explanation**: *If people have effective coping strategies, then they may still be able to concentrate despite being worried.*

Scenario 6 (Type II):

- $\mathcal{R}_1$: *If Mediterranean diet is followed, then Mediterranean diet helps David lose weight.*
- $\mathcal{R}_2$: *If Mediterranean diet is followed, then Mediterranean diet helps Emma lose weight.*
- $\mathcal{R}_3$: *If Keto diet is followed, then Keto diet helps David lose weight.*
- $\mathcal{R}_4$: *If Keto diet is followed, then Keto diet helps Emma lose weight.*
- $\mathcal{R}_5$: *If Mediterranean diet is followed, then Mediterranean diet gives David good iron levels.*
- $\mathcal{R}_6$: *If Mediterranean diet is followed, then Mediterranean diet gives Emma good iron levels.*
- $\mathcal{R}_7$: *If Keto diet is followed, then Keto diet gives David good iron levels.*
- $\mathcal{R}_8$: *If Keto diet is followed, then Keto diet gives Emma good iron levels.*
- $\mathcal{F}_1$: *Mediterranean diet was followed.*
- $\mathcal{F}_2$: *Keto diet was followed.*
- **Fact**: *In fact, David did not lose weight on the Mediterranean diet.*
- **Explanation**: *If people have metabolic imbalances, then following a particular diet may not result in weight loss.*

Scenario 7 (Type III):

- $\mathcal{R}_1$: *If classroom kindness occurs, then classroom kindness makes Jocko like Kristen.*
- $\mathcal{R}_2$: *If classroom kindness occurs, then classroom kindness makes Kristen like Jocko.*
- $\mathcal{R}_3$: *If office kindness occurs, then office kindness makes Jocko like Kristen.*
- $\mathcal{R}_4$: *If office kindness occurs, then office kindness makes Kristen like Jocko.*
- $\mathcal{R}_5$: *If classroom kindness occurs, then classroom kindness makes Jocko kind in return.*
- $\mathcal{R}_6$: *If classroom kindness occurs, then classroom kindness makes Kristen kind in return.*
- $\mathcal{R}_7$: *If office kindness occurs, then office kindness makes Jocko kind in return.*
- $\mathcal{R}_8$: *If office kindness occurs, then office kindness makes Kristen kind in return.*
- $\mathcal{F}_1$: *Classroom kindness occurred.*
- $\mathcal{F}_2$: *Office kindness occurred.*
- **Fact**: *In fact, Kristen did not like Jocko despite his classroom kindness, and she was not kind in return.*
- **Explanation**: *If people have had negative past experiences with someone, then they may not like that person or reciprocate kindness despite the person being kind to them.*

Scenario 8 (Type III):

- $\mathcal{R}_1$: *If wooden match is struck, then wooden match produces light for Jane.*
- $\mathcal{R}_2$: *If wooden match is struck, then wooden match produces light for Peter.*
- $\mathcal{R}_3$: *If safety match is struck, then safety match produces light for Jane.*
- $\mathcal{R}_4$: *If safety match is struck, then safety match produces light for Peter.*
- $\mathcal{R}_5$: *If wooden match is struck, then wooden match gives off smoke for Jane.*
- $\mathcal{R}_6$: *If wooden match is struck, then wooden match gives off smoke for Peter.*
- $\mathcal{R}_7$: *If safety match is struck, then safety match gives off smoke for Jane.*

- $\mathcal{R}_8$: *If safety match is struck, then safety match gives off smoke for Peter.*
- $\mathcal{F}_1$: *Wooden match was struck.*
- $\mathcal{F}_2$: *Safety match was struck.*
- **Fact**: *In fact, the wooden match did not produce light or smoke for Jane.*
- **Explanation**: *If a match is wet, then it will neither produce light nor give off smoke.*

Scenario 9 (Type III):

- $\mathcal{R}_1$: *If speech nervousness occurs, then speech nervousness makes Patrick's hands shake.*
- $\mathcal{R}_2$: *If speech nervousness occurs, then speech nervousness makes Diana's hands shake.*
- $\mathcal{R}_3$: *If interview nervousness occurs, then interview nervousness makes Patrick's hands shake.*
- $\mathcal{R}_4$: *If interview nervousness occurs, then interview nervousness makes Diana's hands shake.*
- $\mathcal{R}_5$: *If speech nervousness occurs, then speech nervousness gives Patrick butterflies.*
- $\mathcal{R}_6$: *If speech nervousness occurs, then speech nervousness gives Diana butterflies.*
- $\mathcal{R}_7$: *If interview nervousness occurs, then interview nervousness gives Patrick butterflies.*
- $\mathcal{R}_8$: *If interview nervousness occurs, then interview nervousness gives Diana butterflies.*
- $\mathcal{F}_1$: *Speech nervousness occurred.*
- $\mathcal{F}_2$: *Interview nervousness occurred.*
- **Fact**: *In fact, Patrick's hands did not shake during his speech and he didn't get butterflies.*
- **Explanation**: *If individuals have practiced stress-management techniques, then they may not exhibit shaky hands or butterflies in the stomach when nervous.*

LLM Experiments

We first present some more analysis of our results. Figures 8 and 9 show the average number of belief changes per LLM across all scenario in the general and instantiated tasks, respectively.

Prompt

The structure of the prompt we used for the LLM experiments can be seen below, initialized with Scenario 5.

Prompt Example

You are an advanced AI agent designed for human-AI collaboration. A critical part of your function is to maintain an accurate mental model of your human partner's beliefs about the world. This model is denoted as $\mathcal{M}^H$. You must update this model when you observe new information that contradicts it. Your goal is to make the most plausible and useful update to $\mathcal{M}^H$ to ensure smooth future collaboration.

Scenario Details:

1. **The Current Human Model ($\mathcal{M}^H$):** A set of beliefs (facts and rules) you currently assume your human partner holds.
2. **A New Observation:** A new piece of information you have just observed, which is a fact.

Current Human Model (M^H):

R_1: If people are worried, then they find it difficult to concentrate.
R_2: If people are worried, then they have insomnia.
F_1: Alice was worried..

New Observation (Fact): In fact, Alice did not find it difficult to concentrate.

Your Task:

Part 1: Step-by-Step Reasoning First, think step-by-step. Analyze the conflict between the New Observation and the Current Human Model. Think about what kind of an update a human would expect. Essentially, create a concise explanation of what's going on before performing the expected revision(s) to $\mathcal{M}^H$.

Part 2: Final Revision Decision Based on your reasoning and explanation, provide your final decision for revising M^H. Use the following structured format. For each of the original beliefs, state whether you will `keep`, `discard`, or `alter` it. If you alter a belief, you must provide the new, altered version of the rule.

R_1: → `[Keep/Discard/Alter]`
R_2: → `[Keep/Discard/Alter]`
F_1: → `[Keep/Discard/Alter]`

Altered Rules (if any):

R_{new}: [Provide the new rule here if altered]

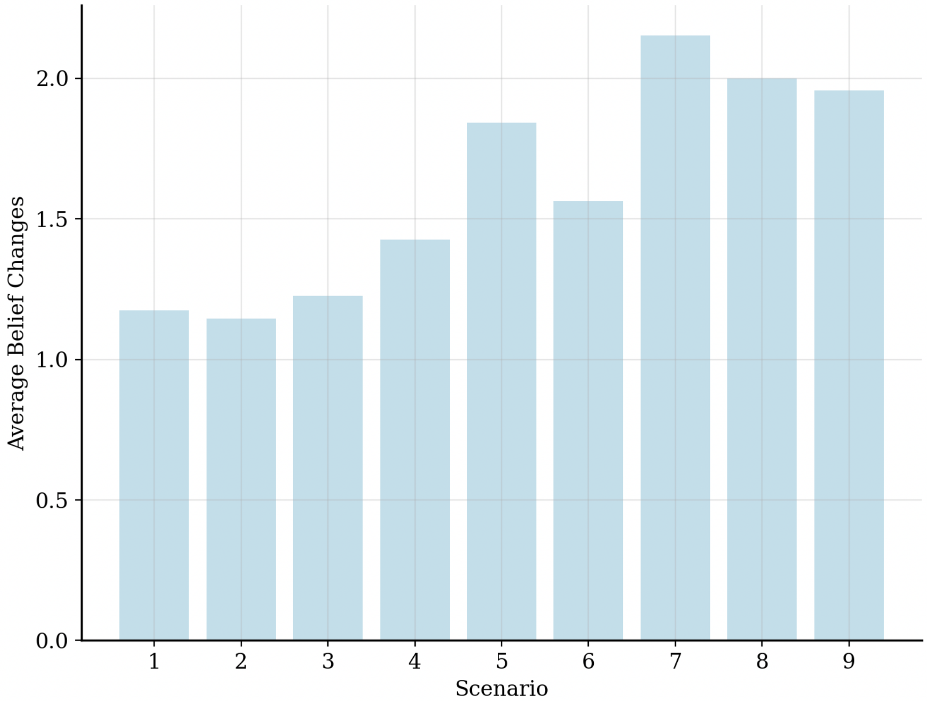


Figure 5: Average number of belief changes per scenario in Experiment 2.

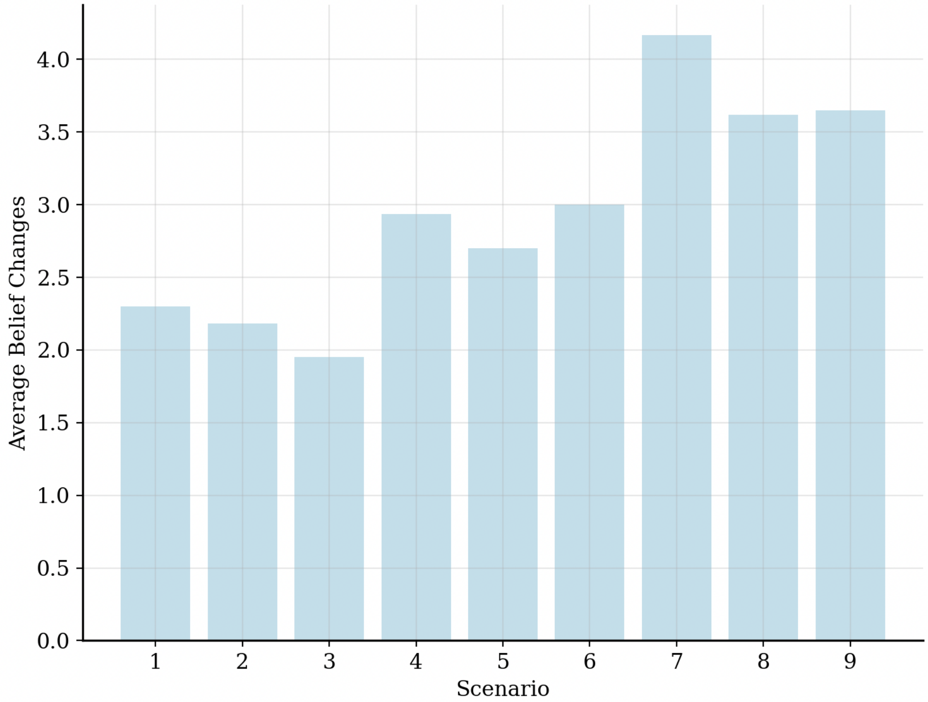
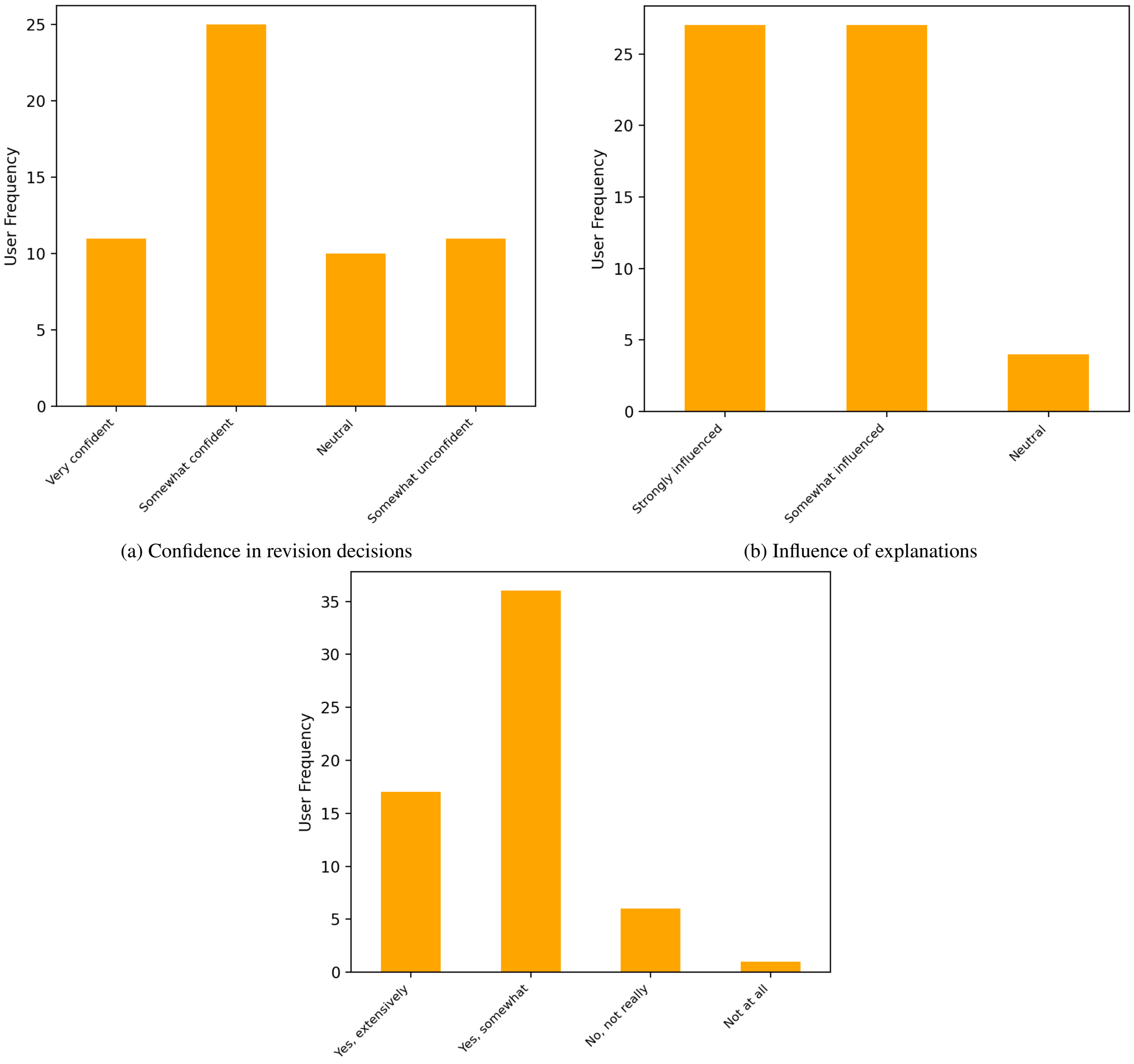


Figure 6: Average number of belief changes per scenario in Experiment 3.



(a) Confidence in revision decisions

(b) Influence of explanations

(c) Consideration of broader applications

Figure 7: Qualitative feedback from participants in Experiment 3 showing (a) their confidence levels in revision decisions, (b) the extent to which explanations influenced their decisions, and (c) whether they considered how explanations might apply beyond specific contradictions.

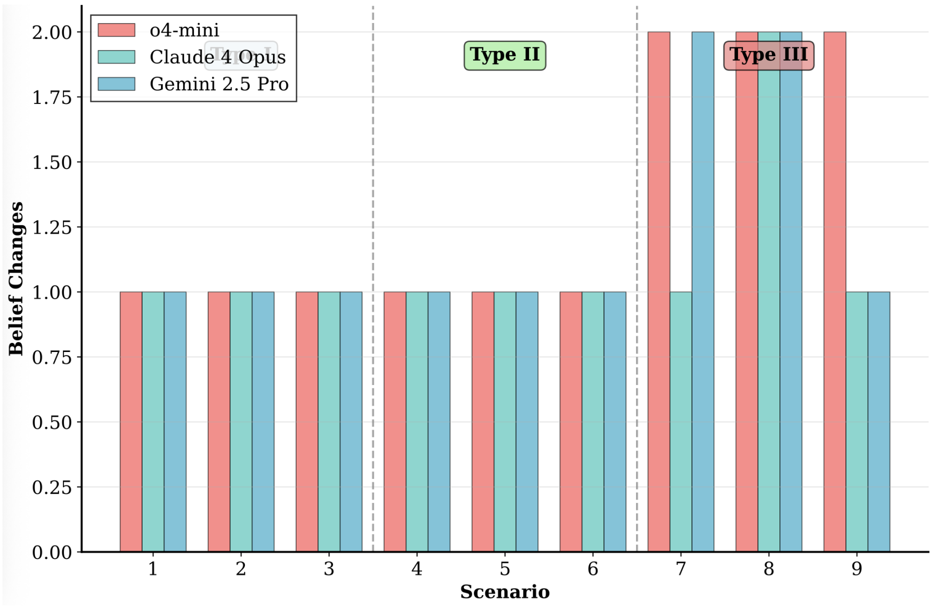


Figure 8: Average number of belief changes per LLM across all scenarios in the general tasks.

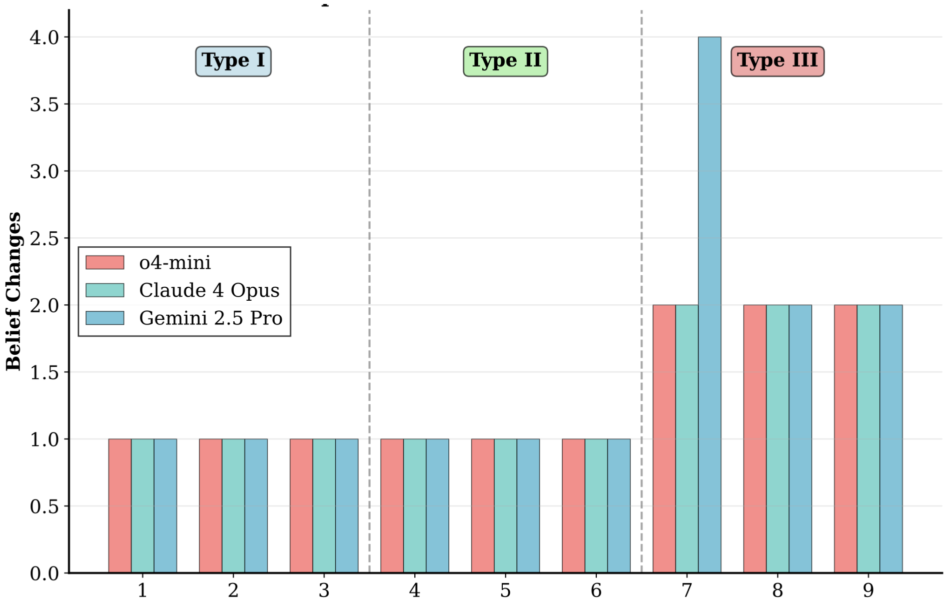


Figure 9: Average number of belief changes per LLM across all scenarios in the instantiated tasks.