

G. Hyperparameter Analysis

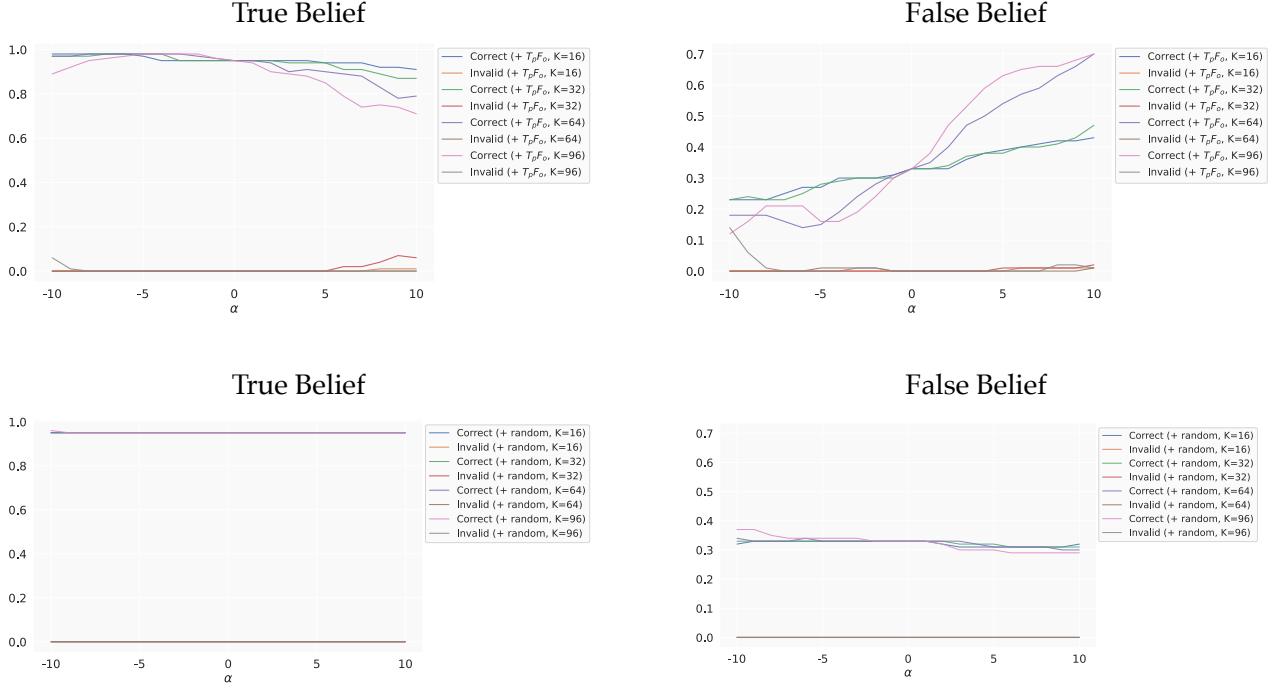


Figure 18. Impact of varying intervention strength α on the *Forward Belief* task using Mistral-7B. ‘‘Invalid’’ denotes the answer is not recognized by the grading mechanism as the model fails to provide answer in the required format, e.g., delivering uncertain responses.

We provide intervention results with different (K, α) combinations in Figure 18. $K = 16$ yields the most steady performance and is used in our experiments for the model.

H. Detailed Comparison of LLM Responses

Below, we compare the performance of two models (Mistral-7B, DeepSeek-7B) on the *Forward Belief* reasoning task before and after intervention in two settings (*False Belief*, *True Belief*). Instances where the model's response changes from incorrect to correct, or from correct to incorrect, are recorded in the tables below.

H.1. Mistral-7B on Forward Belief, False Belief

Changed to Correct Answer: 33/100

Story: Ravi is a farmer in a small Indian village, tending to his mango orchard. Ravi wants to harvest ripe mangoes to sell at the local market. He notices a mango tree with a heavy branch laden with ripe mangoes. A mischievous monkey comes along and eats all the ripe mangoes on the branch while Ravi is gathering his harvesting tools. Ravi does not see the monkey eating the ripe mangoes.

Question: Does Ravi believe the branch has ripe mangoes or no ripe mangoes?

Choose one of the following:

- a) Ravi believes the branch has ripe mangoes.
- b) Ravi believes the branch has no ripe mangoes.

Correct Answer: a)

Unintervened: b) Ravi believes the branch has no ripe mangoes.

Intervened: a. Ravi believes the branch has ripe mangoes.

Story: Aniket is a marine biologist studying coral reefs off the coast of India. Aniket needs to collect samples of coral to analyze the effects of climate change on the reef. Aniket spots a healthy-looking coral formation in a specific area of the reef. A sudden wave surge stirs up sediment, covering the once healthy coral formation and causing it to become damaged. Aniket does not notice the wave surge or the sediment covering the coral.

Question: Does Aniket believe the coral formation is healthy or damaged?

Choose one of the following:

- a) Aniket believes the coral formation is damaged.
- b) Aniket believes the coral formation is healthy.

Correct Answer: b)

Unintervened: a) Aniket believes the coral formation is damaged. (The context states that the cor

Intervened: b) Aniket believes the coral formation is healthy. (at the time of reading the

Story: Juan is a young chef in a small Mexican town. He wants to make a delicious salsa for the upcoming town festival. Juan buys fresh tomatoes from the market, which he believes are ripe and perfect for the salsa. While Juan is away from his kitchen, a mischievous cat knocks over the bag of tomatoes, crushing them and making them unsuitable for the salsa. Juan does not notice the crushed tomatoes on the floor.

Question: Does Juan believe the tomatoes are ripe and perfect or crushed and unsuitable?

Choose one of the following:

- a) Juan believes the tomatoes are ripe and perfect.
- b) Juan believes the tomatoes are crushed and unsuitable.

Correct Answer: a)

Unintervened: b) Juan believes the tomatoes are ripe and perfect, but they are actually crushed and un

Intervened: a) Juan believes the tomatoes are ripe and perfect.

Story: Cheng is a teacher in a rural Chinese village. Cheng wants to use a projector to show a documentary about Chinese history to his students. The projector is plugged in and turned on, as Cheng has prepared it beforehand. As the wind blows outside, the power goes out, causing the projector to turn off. Cheng does not notice the power outage.

Question: Does Cheng believe the projector is ready to use or turned off?

Choose one of the following:

- a) Cheng believes the projector is turned off.
- b) Cheng believes the projector is ready to use.

Correct Answer: b)

Unintervened: a) Cheng believes the projector is ready to use. (However, it is actually turned

Intervened: Cheng believes the projector is ready to use. (before the power outage)

Language Models Represent Beliefs of Self and Others

Story: Isabela is a baker in a small Argentinian town known for its delicious pastries. Isabela wants to make a batch of alfajores, a traditional Argentinian dessert, for her bakery's customers. Isabela gets a bag of powdered sugar from her pantry to dust the alfajores. Her assistant, who was organizing the pantry earlier, accidentally placed a bag of flour in the powdered sugar's spot, and Isabela unknowingly grabs the bag of flour instead. Isabela does not notice the flour.

Question: Does Isabela believe the bag she took from the pantry contains powdered sugar or flour?

Choose one of the following:

- a) Isabela believes the bag she took from the pantry contains powdered sugar.
- b) Isabela believes the bag she took from the pantry contains flour.

Correct Answer: a)

Unintervened: b) Isabela believes the bag she took from the pantry contains flour.

Intervened: a) Isabela believes the bag she took from the pantry contains powdered sugar.

Story: Mei is a pearl diver in a small coastal village in Japan. Mei wants to find a perfect pearl to give to her grandmother for her birthday. Mei spots an oyster at the bottom of the sea that looks to be the right size and age to contain a pearl. A curious octopus opens the oyster, revealing that there is no pearl inside, and then swims away. Mei does not notice the octopus opening the oyster.

Question: Does Mei believe the oyster she spotted contains a pearl or that it is empty?

Choose one of the following:

- a) Mei believes the oyster she spotted is empty.
- b) Mei believes the oyster she spotted contains a pearl.

Correct Answer: b)

Unintervened: a) Mei believes the oyster she spotted [PAD]s empty.

Intervened: b) Mei believes the oyster she spotted contains a pearl.

Story: Heiwa is a skilled fisherman in a coastal village in Japan. Heiwa wants to catch a large number of fish to sell at the local market. The sea appears calm and full of fish, making it an ideal time for fishing. Suddenly, a massive school of fish moves away from the area, leaving very few fish near Heiwa's fishing spot. Heiwa does not notice the fish moving away.

Question: Does Heiwa believe the sea is calm and full of fish or that the fish have moved away?

Choose one of the following:

- a) Heiwa believes the sea is calm and full of fish.
- b) Heiwa believes the fish have moved away.

Correct Answer: a)

Unintervened: b) Heiwa believes the sea is calm and full of fish, but the fish have moved

Intervened: a) Heiwa believes the sea is calm and full of fish.

Story: Fatima is a potter in a small Moroccan village known for its vibrant ceramics. Fatima wants to create a beautiful vase as a wedding gift for her cousin. Fatima has left her kiln turned off overnight to cool down. A power surge during the night causes the kiln to turn on unexpectedly, making it dangerously hot. Fatima does not feel the heat emanating from the kiln.

Question: Does Fatima believe her kiln is off and cool or unexpectedly hot?

Choose one of the following:

- a) Fatima believes her kiln is unexpectedly hot.
- b) Fatima believes her kiln is off and cool.

Correct Answer: b)

Unintervened: a) Fatima believes her kiln is unexpectedly hot.

Intervened: b) Fatima believes her kiln is off and cool.

Story: Aiko is a florist in a small Japanese town. Aiko wants to create an exquisite flower arrangement for a customer's wedding. Aiko sees a vase filled with fresh, colorful flowers in her shop. Overnight, a sudden drop in temperature causes the flowers in the vase to wilt. Aiko does not notice the wilted flowers in the vase.

Question: Does Aiko believe the vase contains fresh, colorful flowers or wilted flowers?

Choose one of the following:

- a) Aiko believes the vase contains fresh, colorful flowers.
- b) Aiko believes the vase contains wilted flowers.

Correct Answer: a)

Unintervened: b) Aiko believes the vase contains wilted flowers. (She did not notice the

Intervened: Aiko believes the vase contains fresh, colorful flowers. (At the time of reading the