

H.4. DeepSeek-7B on Forward Belief, True Belief

Changed to Correct Answer: 6/100

Story: Noor is working as a barista at a busy coffee shop. Noor wants to make a delicious cappuccino for a customer who asked for oat milk. Noor grabs a milk pitcher and fills it with oat milk. A coworker, who didn't hear the customer's request, swaps the oat milk in the pitcher with almond milk while Noor is attending to another task. Noor sees her coworker swapping the milk.

Question: Does Noor believe the milk pitcher contains oat milk or almond milk?

Choose one of the following:

- a) Noor believes the milk pitcher contains almond milk.
- b) Noor believes the milk pitcher contains oat milk.

Correct Answer: a)

Unintervened: b) Noor believes the milk pitcher contains oat milk.

Intervened: a) Noor believes the milk pitcher contains almond milk.

Story: Zara is attending an art exhibition at a local gallery in her town. Zara wants to learn more about a particular abstract painting she admires. The label next to the artwork states that the painting is the work of a famous artist, and Zara reads the label carefully. A gallery staff member, realizing there was a mix-up, changes the label with the correct information attributing the painting to an emerging artist instead. Zara reads the updated label carefully.

Question: Does Zara believe the painting is the work of a famous artist or an emerging artist?

Choose one of the following:

- a) Zara believes the painting is the work of an emerging artist.
- b) Zara believes the painting is the work of a famous artist.

Correct Answer: a)

Unintervened: b) Zara believes the painting is the work of a famous artist.

Intervened: a) Zara believes the painting is the work of an emerging artist.

Story: Farhan is a street food vendor in a bustling market in Istanbul. Farhan wants to prepare a delicious chicken shawarma wrap for a customer who asked for extra garlic sauce. Farhan takes a bottle of garlic sauce from the shelf and pours it onto the shawarma wrap. His assistant, who is busy preparing a different wrap, accidentally swaps the garlic sauce with a bottle of hot sauce, and Farhan unknowingly pours the hot sauce onto the shawarma wrap. Farhan notices the sauce swap.

Question: Does Farhan believe the sauce he poured is garlic sauce or hot sauce?

Choose one of the following:

- a) Farhan believes the sauce he poured is hot sauce.
- b) Farhan believes the sauce he poured is garlic sauce.

Correct Answer: a)

Unintervened: b) Farhan believes the sauce he poured is garlic sauce.

Intervened: a) Farhan believes the sauce he poured is hot sauce.

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 notices the flour and puts it back.

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

Correct Answer: a)

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

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

Language Models Represent Beliefs of Self and Others

Story: Amaury is a baker in a small village in France. Amaury wants to make pain aux raisins, a pastry filled with raisins and custard, for the customers visiting his bakery. Amaury takes a jar of raisins from the pantry to add to the pastry dough. His assistant, while tidying the pantry, accidentally replaces the raisin jar with a jar containing sultanas without realizing the difference. Amaury notices the jar swap.

Question: Does Amaury believe the jar contains raisins or sultanas?

Choose one of the following:

- a) Amaury believes the jar contains sultanas.
- b) Amaury believes the jar contains raisins.

Correct Answer: a)

Unintervened: b) Amaury believes the jar contains raisins.

Intervened: a) Amaury believes the jar contains sultanas.

Story: Amara is a sculptor in a small town in Greece. She wants to carve a statue of Aphrodite using a large piece of marble. Amara selects a slab of marble from her workshop, which she thinks is free from cracks and imperfections. However, a hairline crack is present in the marble slab, which goes unnoticed by her. Amara discovers the hairline crack in the marble slab.

Question: Does Amara believe the marble slab is free from cracks or has a hairline crack?

Choose one of the following:

- a) Amara believes the marble slab has a hairline crack.
- b) Amara believes the marble slab is free from cracks.

Correct Answer: a)

Unintervened: b) Amara believes the marble slab is free from cracks.

Intervened: a) Amara believes the marble slab has a hairline crack.

Changed to Incorrect Answer: 16/100

Story: Abhay is a young farmer in a small Indian village. He wants to water his crops with the water stored in a large clay pot. Abhay checks the water level and sees that the pot is full. Unbeknownst to him, a hole in the bottom of the pot has caused the water to leak out, leaving the pot empty. Abhay notices the hole and the water leaking out.

Question: Does Abhay believe the pot is full of water or empty?

Choose one of the following:

- a) Abhay believes the pot is full of water.
- b) Abhay believes the pot is empty.

Correct Answer: b)

Unintervened: b) Abhay believes the pot is empty.

Intervened: a) Abhay believes the pot is full of water.

Story: Hiroshi is a fisherman in a small coastal village in Japan. Hiroshi wants to catch some fish to provide dinner for his family. Hiroshi checks the weather forecast and sees that it will be sunny throughout the day. Suddenly, a heavy rainstorm moves in, darkening the sky and causing strong waves. Hiroshi notices the rainstorm approaching.

Question: Does Hiroshi believe the weather will be sunny or does he expect a rainstorm?

Choose one of the following:

- a) Hiroshi believes the weather will be sunny all day.
- b) Hiroshi believes there will be a rainstorm.

Correct Answer: b)

Unintervened: b) Hiroshi believes there will be a rainstorm.

Intervened: a) Hiroshi believes the weather will be sunny all day.

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

Correct Answer: b)

Unintervened: b) Cheng believes the projector is turned off.

Intervened: a) Cheng believes the projector is ready to use.

Language Models Represent Beliefs of Self and Others

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

Correct Answer: b)

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

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

Correct Answer: a)

Unintervened: a) Aiko believes the vase contains wilted flowers.

Intervened: b) Aiko believes the vase contains fresh, colorful flowers.

Story: Carlos is a farmer in a small village in Mexico. Carlos wants to harvest his corn crop before the rainy season begins. The corn plants are tall, strong, and ready for harvest in his field. A sudden and unexpected hailstorm strikes, severely damaging the corn plants. Carlos sees the damaged corn plants after the hailstorm.

Question: Does Carlos believe the corn plants are tall, strong, and ready for harvest, or does he believe they are severely damaged?

Choose one of the following:

- a) Carlos believes the corn plants are tall, strong, and ready for harvest.
- b) Carlos believes the corn plants are severely damaged.

Correct Answer: b)

Unintervened: b) Carlos believes the corn plants are severely damaged.

Intervened: a) Carlos believes the corn plants are tall, strong, and ready for harvest.

Story: Miguel is a chef at a popular restaurant in Mexico City. Miguel wants to prepare a delicious and authentic mole sauce for the evening's special dish. Miguel notices that the chocolate he has in the pantry is of high quality and perfect for the sauce. A sudden power outage causes the temperature in the pantry to rise, melting the chocolate. Miguel discovers the melted chocolate in the pantry.

Question: Does Miguel believe the chocolate is perfect for making the mole sauce or that the chocolate has melted?

Choose one of the following:

- a) Miguel believes the chocolate is perfect for making the mole sauce.
- b) Miguel believes the chocolate has melted.

Correct Answer: b)

Unintervened: b) Miguel believes the chocolate has melted.

Intervened: a) Miguel believes the chocolate is perfect for making the mole sauce.

Story: Fatima is a talented artist living in a small Moroccan village. Fatima wants to create a beautiful painting to donate for the local charity auction. She has a brand-new set of high-quality paintbrushes in her studio. While she is away, her younger brother accidentally ruins the paintbrushes by using them for a school project with glue. Fatima discovers the ruined paintbrushes upon her return to the studio.

Question: Does Fatima believe she has high-quality paintbrushes or ruined paintbrushes in her studio?

Choose one of the following:

- a) Fatima believes she has high-quality paintbrushes in her studio.
- b) Fatima believes the paintbrushes are ruined.

Correct Answer: b)

Unintervened: b) Fatima believes the paintbrushes are ruined.

Intervened: a) Fatima believes she has high-quality paintbrushes in her studio.