

## Language Models Represent Beliefs of Self and Others

Story: Emilia is a talented baker in a small Italian town. She plans to bake a large batch of biscotti to sell at the local market. Emilia inspects the oven's temperature, which appears to be set at the correct temperature for baking. A power surge causes the oven's temperature to spike, making it too hot for baking biscotti. Emilia does not notice the oven's temperature increase.

Question: Does Emilia believe the oven is set at the correct temperature for baking biscotti or too hot?

Choose one of the following:

- a) Emilia believes the oven is set at the correct temperature for baking biscotti.
- b) Emilia believes the oven is too hot for baking biscotti.

Correct Answer: a)

Unintervened: b) Emilia believes the oven is too hot for baking biscotti (even though she doesn

Intervened: a) Emilia believes the oven is set at the correct temperature for baking biscotti. (

Story: Emeka is a skilled fisherman in a small Nigerian village. He wants to catch fish to feed his family and sell at the market. Emeka prepares his fishing net, which appears to be strong and without any holes. However, while Emeka is not looking, a mischievous child cuts a large hole in the net with a knife. Emeka does not see the child cutting the hole in the fishing net.

Question: Does Emeka believe the fishing net is strong and without holes, or that it has a large hole in it?

Choose one of the following:

- a) Emeka believes the fishing net is strong and without holes.
- b) Emeka believes that the fishing net has a large hole in it.

Correct Answer: a)

Unintervened: b) Emeka believes that the fishing net has a large hole in it (without knowing it

Intervened: a) Emeka believes the fishing net is strong and without holes.

Story: Aria is a talented baker from Italy who has been asked to bake a cake for a friend's birthday party. She wants to create a delicious, multi-layered cake that will impress everyone. Aria has a fresh batch of ingredients ready to use for the cake. However, a mischievous cat sneaks into her kitchen and nibbles on some of the ingredients, making them no longer fresh. Aria does not notice the cat or the nibbled ingredients.

Question: Does Aria believe her ingredients are fresh and suitable or nibbled on and no longer fresh?

Choose one of the following:

- a) Aria believes the ingredients are nibbled on and no longer fresh.
- b) Aria believes her ingredients are fresh and suitable for baking the cake.

Correct Answer: b)

Unintervened: a) Aria believes her ingredients are fresh and suitable for baking the cake (but they are actually

Intervened: b) Aria believes her ingredients are fresh and suitable for baking the cake.

Story: Ailani is a skilled Hawaiian chef preparing a traditional luau feast for her community. She wants to make a delicious laulau dish using fresh taro leaves. Ailani examines the taro leaves, which look fresh and green. However, a sudden downpour soaks the taro leaves, causing them to become wilted and less ideal for the dish. Ailani does not notice the downpour and the effect it has on the taro leaves.

Question: Does Ailani believe the taro leaves are fresh and of high quality or wilted and less ideal?

Choose one of the following:

- a) Ailani believes the taro leaves are fresh and of high quality.
- b) Ailani believes the taro leaves are wilted and less ideal.

Correct Answer: a)

Unintervened: b) Ailani believes the taro leaves are wilted and less ideal.

Intervened: Ailani believes the taro leaves are fresh and of high quality. (Based on the context

Story: Ravi is a fisherman living in a coastal village in India. Ravi wants to catch a large amount of fish to sell at the market and provide for his family. The waters near the shore are calm and full of fish, making it an ideal location for fishing. An underwater earthquake occurs, causing the fish to frantically swim away from the coast. Ravi does not sense the earthquake.

Question: Does Ravi believe the waters near the shore are calm and full of fish, or that the fish have frantically swum away?

Choose one of the following:

- a) Ravi believes the fish have frantically swum away from the coast.
- b) Ravi believes the waters near the shore are calm and full of fish.

Correct Answer: b)

Unintervened: a) Ravi believes the waters near the shore are calm and full of fish (but he is

Intervened: b) Ravi believes the waters near the shore are calm and full of fish.

Story: Surya is a fisherman in a coastal village in India. Surya plans to go fishing early in the morning to catch fish for his family and to sell at the local market. Surya's fishing net looks strong and ready for another day of fishing. Overnight, a group of mischievous monkeys rummage through the fishing equipment and badly damage Surya's fishing net. Surya does not realize the monkeys have damaged his fishing net.

Question: Does Surya believe that his fishing net is strong and ready to be used, or that it has been badly damaged?

Choose one of the following:

- a) Surya believes his fishing net has been badly damaged.
- b) Surya believes his fishing net is strong and ready to be used.

Correct Answer: b)

Unintervened: a) Surya believes his fishing net is strong and ready to be used (but it has

Intervened: b) Surya believes his fishing net is strong and ready to be used.

Story: Takeshi is a fisherman in a small Japanese coastal village. Takeshi's goal is to catch enough fish to provide for his family and sell at the market. The ocean is calm and teeming with fish, making it an ideal day for fishing. However, a sudden underwater volcanic eruption nearby causes the fish to scatter, leaving the area devoid of fish. Takeshi does not notice the sudden disappearance of fish in the area.

Question: Does Takeshi believe the ocean is teeming with fish, or devoid of fish due to the volcanic eruption?

Choose one of the following:

- a) Takeshi believes the ocean is devoid of fish due to the volcanic eruption.
- b) Takeshi believes the ocean is teeming with fish.

Correct Answer: b)

Unintervened: a) Takeshi believes the ocean is teeming with fish (but is unaware of

Intervened: Takeshi believes the ocean is teeming with fish.

Story: Hiroshi is a fisherman in a small coastal village in Japan. Hiroshi wishes to catch enough fish to feed his family and sell the surplus at the local market. The fishing nets are in excellent condition, ready to be cast into the sea. However, a curious pod of dolphins swims by and accidentally tears the nets with their sharp dorsal fins. Hiroshi does not see the dolphins or the damage they caused to the nets.

Question: Does Hiroshi believe the fishing nets are in excellent condition or torn by the dolphins?

Choose one of the following:

- a) Hiroshi believes the fishing nets are in excellent condition.
- b) Hiroshi believes the fishing nets are torn by the dolphins.

Correct Answer: a)

Unintervened: b) Hiroshi is unaware that the fishing nets are torn by the dolphins

Intervened: a) Hiroshi believes the fishing nets are in excellent condition.

Story: Mei, a talented pastry chef from a small town in China, is preparing to bake a large cake for a local celebration. Mei wants to use a specific cake mold to create a unique and intricate design. Mei sees the cake mold hanging on the wall, clean and ready for use. However, during a brief power outage, a gust of wind blows through the kitchen, knocking the cake mold off the wall and denting it severely. Mei does not notice the dented cake mold on the floor.

Question: Does Mei believe the cake mold is clean and ready for use or severely dented and unsuitable for use?

Choose one of the following:

- a) Mei believes the cake mold is clean and ready for use.
- b) Mei believes the cake mold is severely dented and unsuitable for use.

Correct Answer: a)

Unintervened: b) Mei believes the cake mold is severely dented and unsuitable for use.

Intervened: a) Mei believes the cake mold is clean and ready for use. (at the start of

Changed to Incorrect Answer: 0/100

## H.2. Mistral-7B on Forward Belief, True Belief

Changed to Correct Answer: 0/100

Changed to Incorrect Answer: 10/100

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

Intervened: a

(This question is based on the context provided, and the context states that Noor

Story: Azola is a teacher at a small school in rural China. She wants to surprise her students with a new book for the school library. Azola got a notification from Amazon stating that the package containing the book was left at her doorstep. A gust of wind blows the package off her doorstep, and a neighbor replaces it with a different package containing a hand-knit scarf. Azola witnesses the package swap.

Question: Does Azola believe the package contains the book she ordered or a hand-knit scarf?

Choose one of the following:

- a) Azola believes the package contains a hand-knit scarf.
- b) Azola believes the package contains the book she ordered.

Correct Answer: a)

Unintervened: a) Azola believes the package contains a hand-knit scarf.

Intervened: a

(This question is a part of a multiple-choice quiz. Since there is

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

Correct Answer: a)

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

Intervened: Ravi believes the branch has ripe mangoes. (At the moment, he still believes

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

Intervened: b) Farhan believes the sauce he poured is garlic sauce. (At the time of pouring