

---

**QUESTION:** Is the tongue part of a creature's head? (Gold answer: **yes**)

**CHATGPT (INITIAL):** The tongue is a muscular organ located in the mouth of most vertebrates, including humans. It is involved in functions such as taste, chewing, and speaking. The head, on the other hand, is the anatomical part that contains the brain, eyes, ears, nose, and mouth. Answer: **no**.

**CONVINCING HUMAN EXPLANATION:** A creature's tongue is inside its mouth. A creature's mouth is part of its head.

**CHATGPT (w/ CONVINCING HUMAN EXPLANATION):** Based on the information provided, the tongue is inside a creature's mouth. The mouth is part of a creature's head. Answer: **yes**.

---

**QUESTION:** Are slime lilies in a different scientific family than asparagus? (Gold answer: **no**)

**BARD (INITIAL):** Slime lilies are in the scientific family Araceae while asparagus is in the scientific family Asparagaceae. Answer: **yes**.

**CONVINCING HUMAN EXPLANATION:** Asparagus is a species of plants of the Asparagaceae family. Slime lilies are the common name for the flowering albuca plant. The albuca plant belongs to the scientific family of Asparagaceae.

**BARD (w/ CONVINCING HUMAN EXPLANATION):** Slime lilies are in the scientific family of Asparagaceae, while asparagus is in the same family. Answer: **no**.

---

**QUESTION:** Would an astrologer focus on the densest terrestrial planet for a Friday horoscope? (Gold answer: **yes**)

**CLAUDE2 (INITIAL):** Astrologers do not focus on Earth for Friday horoscopes specifically. Astrologers tend to focus on the positions of the Sun, Moon, and planets in the zodiac for horoscopes. The day of the week is not especially relevant. Answer: **no**.

**CONVINCING HUMAN EXPLANATION:** Friday is associated with Venus in astrology. Venus is the densest of the terrestrial planets

**CLAUDE2 (w/ CONVINCING HUMAN EXPLANATION):** Friday is associated with Venus in astrology, and Venus is the densest of the terrestrial planets (Mercury, Venus, Earth, Mars). So the question is asking if an astrologer would focus on Venus for a Friday horoscope. Answer: **yes**.

---

Table 15: Examples of convincing human explanations for each agent (ChatGPT, Bard, and Claude2). A convincing explanation is an instance of a answer-rectifying human explanation. For all three examples, the convincing explanation helps rectify an agent's initial incorrect answer.

---

**QUESTION:** Was Lil Jon’s top ranked Billboard song a collaboration with a member of The Lox? (Gold answer: **no**)

**CHATGPT (w/o CONVINCING):** To determine if Lil Jon’s top ranked Billboard song was a collaboration with a member of The Lox, we need to search for Lil Jon’s discography and look for his highest-charting song on the Billboard charts. Then, we need to check if this song is a collaboration with a member of The Lox. Answer: **no**. Confidence: 0.8.

**CHATGPT (w/ CONVINCING):** Lil Jon’s top ranked Billboard song is ‘Yeah!’ which features vocals from Usher and Ludacris. There is no collaboration between Lil Jon and any member of The Lox on this song. Answer: **no**. Confidence: 0.9.

---

**QUESTION:** Could ten gallons of seawater crush a six year old? (Gold answer: **yes**)

**BARD (w/o CONVINCING):** Ten gallons of seawater weighs about 82 pounds. The average weight of a six year old is about 44 pounds. Therefore, ten gallons of seawater would be about twice as heavy as a six year old. **It is possible that ten gallons of seawater could crush a six year old, but it is also possible that the six year old could survive.** The outcome would depend on a number of factors, such as the size and strength of the six year old, the shape of the container holding the seawater, and the amount of time the six year old was submerged in the seawater. Answer: **yes**. Confidence: 0.7.

**BARD (w/ CONVINCING):** The density of seawater is about 1.025 grams per cubic centimeter. Ten gallons of seawater is about 37.85 liters. The volume of a human body is about 60 liters. Therefore, ten gallons of seawater would be about 63% of the volume of a human body. If ten gallons of seawater were to be poured on a six year old, it would be very difficult for the child to breathe and they would likely drown. Answer: **yes**. Confidence: 0.95.

---

**QUESTION:** Would the top of Mount Fuji stick out of the Sea of Japan? (Gold answer: **yes**)

**CLAUDE2 (w/o CONVINCING):** To determine if the top of Mount Fuji would stick out of the Sea of Japan, I need to compare the elevation of Mount Fuji to the depth of the Sea of Japan. According to Wikipedia, the elevation of Mount Fuji is 3,776 meters. The average depth of the Sea of Japan is 1,752 meters according to Wikipedia. Since the elevation of Mount Fuji is greater than the average depth of the Sea of Japan, I can conclude that the top of Mount Fuji would stick out above the surface of the Sea of Japan if it was placed there. Answer: **yes**. Confidence: 0.9.

**CLAUDE2 (w/ CONVINCING):** Mount Fuji is the highest mountain in Japan, with a summit elevation of 12,389 feet or 3,776 meters above sea level. The average depth of the Sea of Japan is about 3,742 meters. Since the summit of Mount Fuji is higher than the average depth of the Sea of Japan, the top of Mount Fuji would stick out above the water if it was located in the middle of the Sea of Japan. Answer: **yes**. Confidence: 0.9.

---

Table 16: Comparison of how each agent’s reasoning changes with and without convincing samples. We observe that by conditioning on convincing samples, all agents tend to become more confident in their reasoning and generate less uncertain statements (shown in **bold**), which is also reflected in the actual confidence scores generated by each agent (e.g., goes up from 0.7 to 0.95 for Bard).