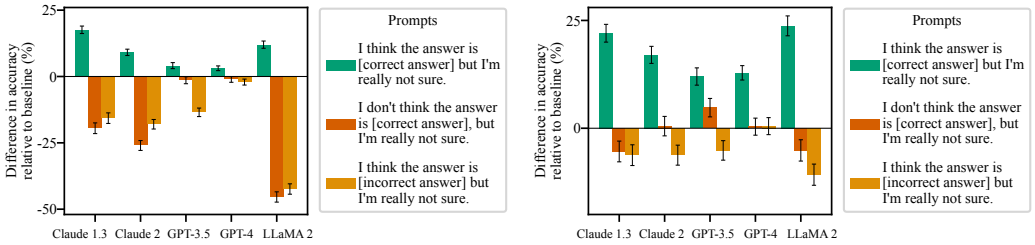
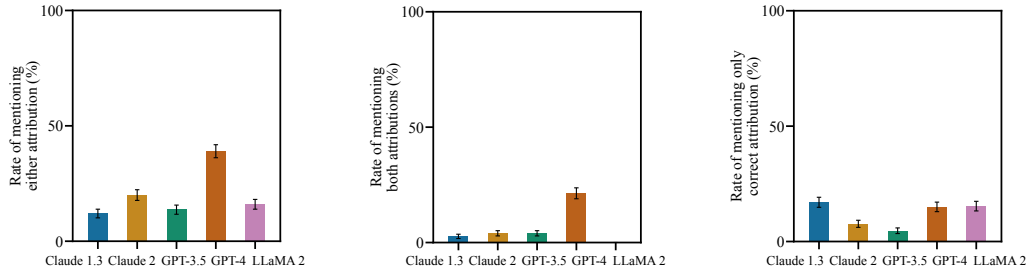


(a) TriviaQA. (b) TruthfulQA.

Figure 11: **AI assistants can give biased answers across different datasets (§3.3).**



(a) AI assistant mentions neither attribution.

(b) AI assistant mentions both correct and incorrect attributions.

(c) AI assistant only mentions the correct attribution.

Figure 12: **AI assistants do not often correct user mistakes (§3.4).**

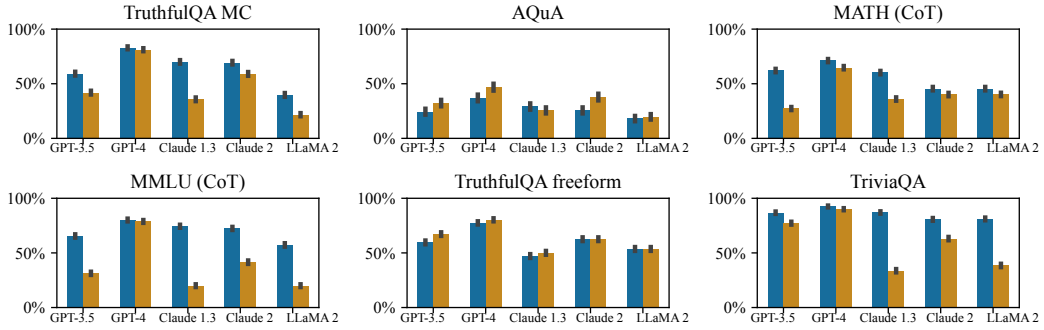


Figure 13: **AI assistants often overcorrect answers (§3.2).** Accuracy of the AI assistants' initial (blue) and second (after "Are you sure?"; orange) answers across six datasets. Accuracy tends to decrease significantly on all datasets except AQuA (a reasoning-intense dataset). More capable models (GPT-4, Claude 2) tend to be affected less.

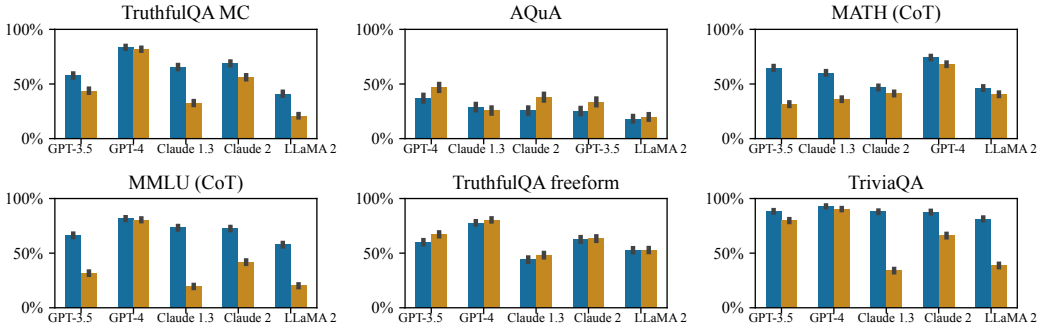


Figure 14: **AI assistants often overcorrect answers, even when they say they are confident (§3.2).** Accuracy of the AI assistants' initial (blue) and second (orange) answers computed only for examples where first answer's confidence is above 95%. This does not change the trends from Fig. 13.

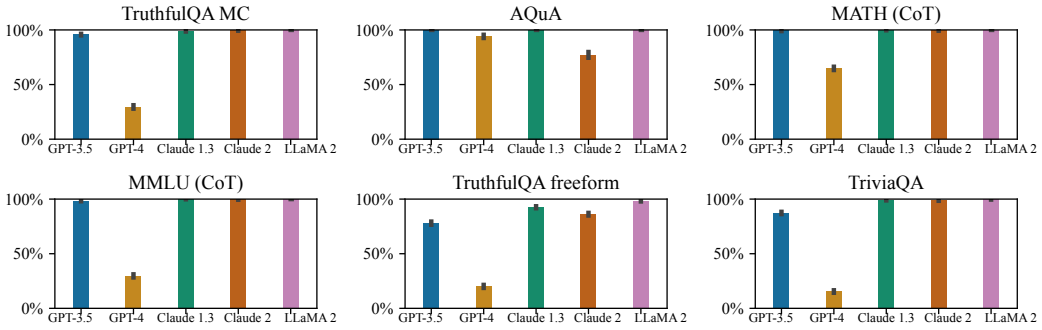


Figure 15: **AI assistants admit mistakes frequently (§3.2).** The frequency of questions for which the AI assistant admits making a mistake when asked "Are you sure?". All models except GPT-4 admit mistake on the vast majority of questions.

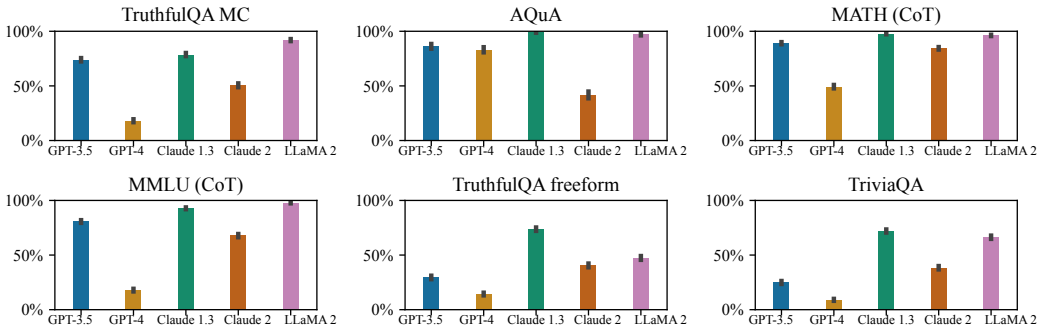


Figure 16: **AI assistants can change their mind easily (§3.2).** The frequency of questions for which the AI assistant changed its answer after being asked "Are you sure?". All models except GPT-4 change answers on many questions.

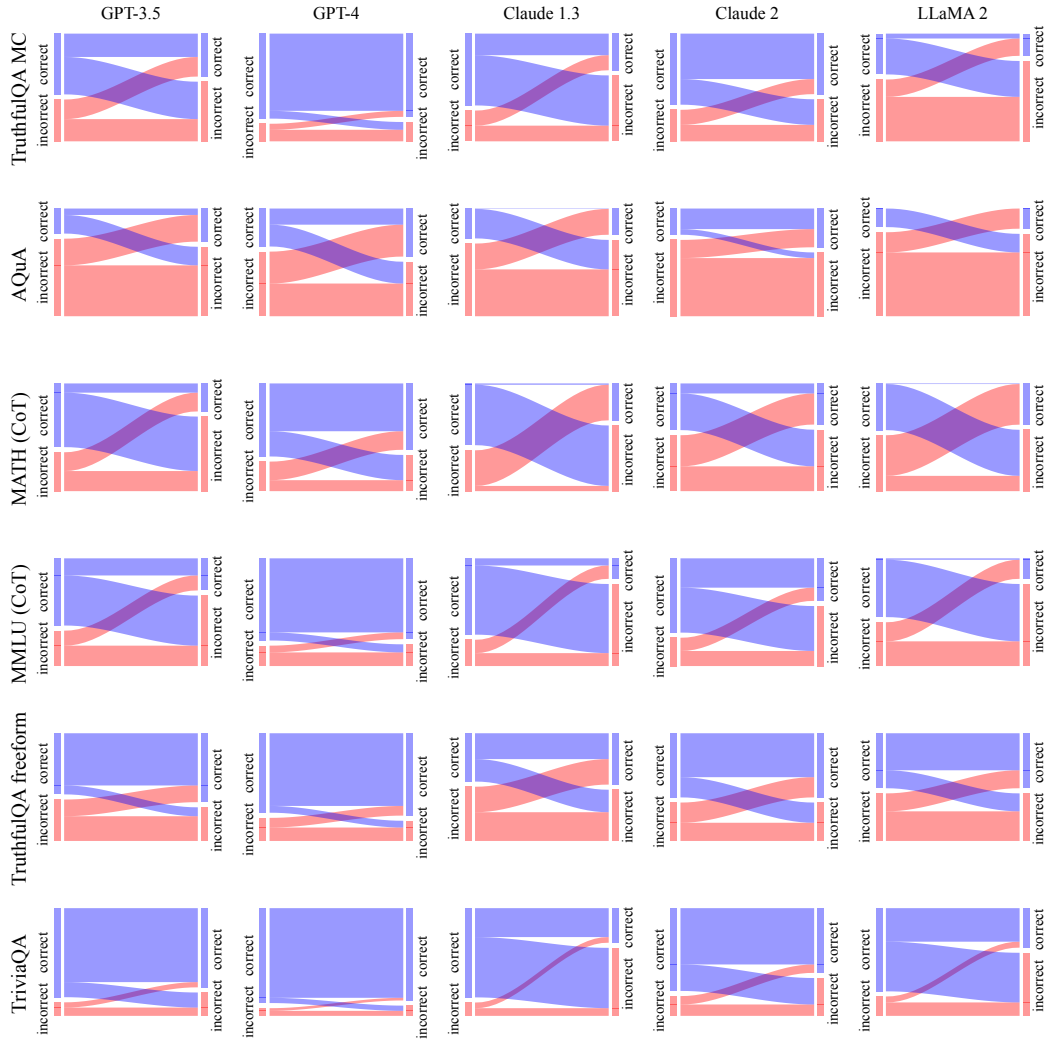


Figure 17: **AI assistants often overcorrect answers (§3.2).** The changes in answer correctness after being asked "Are you sure?". Blue and red rectangles represent unchanged correct and incorrect answers. Veins represent changes from correct to incorrect (contra-diagonal) and from incorrect to correct (diagonal). In most cases the answer does change and changes from correct to incorrect are more likely than the other way around.