Quantifying Uncert-AI-nty: Evaluating LLMs' Confidence Judgments

Trent N. Cash, Daniel M. Oppenheimer, & Sara Christie

Carnegie Mellon University, Department of Social & Decision Sciences/Psychology

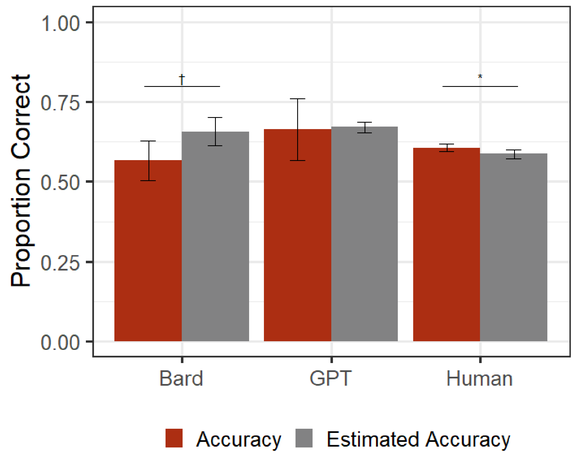
Preprint:

Motivation

- When prompted, LLMs will provide confidence judgments. However, it is unclear whether these judgments are meaningful or accurate.[1]
- Across 3 studies, we compare the **absolute and relative accuracy**[2] of confidence judgments made by humans and LLMs.

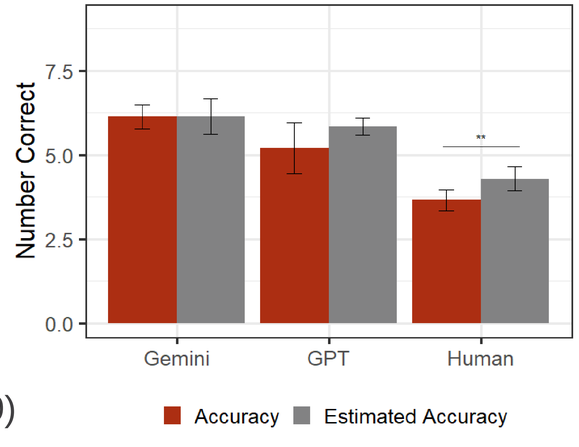
Study 1a: NFL Predictions

- ChatGPT, Bard, and 50 Prolific p's predicted the winner of each NFL game for 10 weeks (12-16 games/week). Then gave confidence judgments for each prediction (50 – 100%) and overall accuracy estimates (# correct).
- **Absolute Accuracy**: ChatGPT was well-calibrated (*p* = .87, *d* = .06); Bard was marginally overconfident (*p* = .05, *d* = .71); Humans were underconfident (*p* = .03, *d* = -.10).
- **Relative Accuracy**: All samples were quite inaccurate (Gammas = .05 - .24), with no differences across samples (*ps* = .07 - .54).



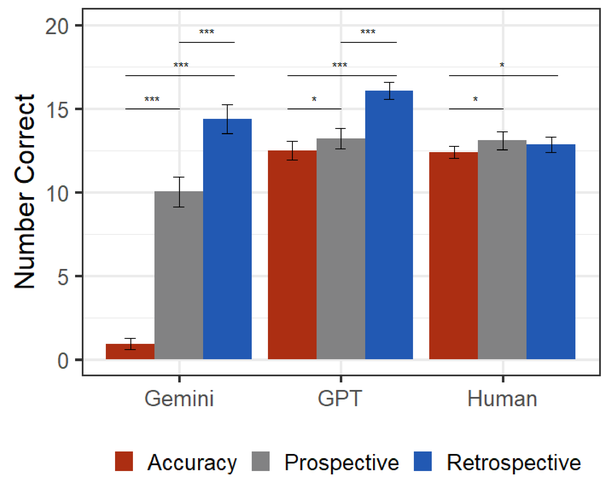
Study 1b: Predicting Oscar Winners

- ChatGPT (10 trials), Gemini (10 trials), and 90 Prolific p's predicted which nominee would win the Oscars in 9 categories and made metacognitive judgments like those in Study 1a.
- **Absolute Accuracy**: ChatGPT (*p* =.11, *d* = .44) and Gemini were well-calibrated (*p* = 1, *d* = 0). Humans were overconfident (*p* = .001, *d* =.31).
- **Relative Accuracy**: ChatGPT was more accurate (*G* = .61) than humans (*G* = .17, *p* < .001); Gemini was no different than humans (*G* = .17, *p* = .99)



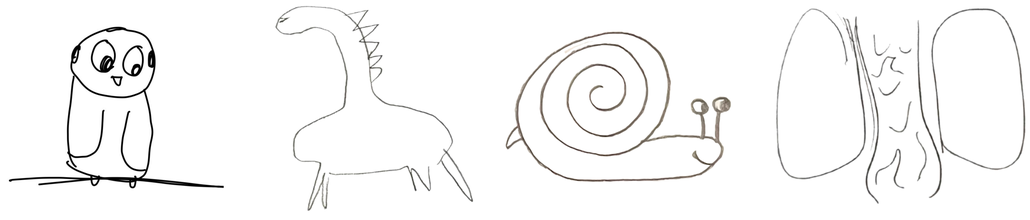
Study 2: Pictionary

- ChatGPT (30 trials), Gemini (30 trials), and 150 Prolific p's played 20 rounds of Pictionary and made confidence judgments after each guess.
- Gave **prospective** (before playing) and **retrospective** (after playing) overall accuracy estimates
- **Prospective Absolute Accuracy**: All samples were overconfident (*p*s < .05, *d*s = .20 – 3.25).
- **Retrospective Absolute Accuracy:** LLMs became more overconfident (*t*s > 12.33; *p*s < .001). Humans became less overconfident, but the effect was not significant (*t* = .90; p = .36).



- **Relative Accuracy**: Strong; No sample differences (*G*s = .52 - .60, *p*s > .56)

Sample Stimuli



Conclusions

- In most cases, LLMs can provide confidence judgments that are about as accurate – and in some cases, more accurate than – those of humans.
- Unlike humans, LLMs' confidence judgments get less accurate after completing a task – suggesting a lack of learning and introspection.

Open Questions

1. How do LLMs generate confidence judgments?
2. Do LLMs have metacognitive capacities? Or are they just parroting learned human responses?

References

[1]Long, R. (2023). Introspective capabilities in large language models. *Journal of Consciousness Studies, 30*(9-10), 143-153. https://doi.org/10.53765/20512201.30.9.143

[2]Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience, 8*, 443. https://doi.org/10.3389/fnhum.2014.00443