ARTIFICIAL INTELLIGENCE-GENERATED ADVICE: HARD TO IDENTIFY AND PERCEIVED TO BE BETTER THAN HUMAN ADVICE

Otto J. B. Kuosmanen, & Tove I. Dahl
Department of Psychology, UiT The Arctic University of Norway (Norway)

Abstract

Background: The rapid advancement of Artificial intelligence (AI), exemplified by models like Chat-GPT, has increasingly made AI a go-to source for advice, raising urgent questions about the quality of AI advice, and whether we can distinguish it from human resources.
Aim: We conducted a study investigating participants' ability to differentiate between human and AI advice and two studies exploring the advice-giving capabilities of GPT models in general life advice.
Method: A web-scraping script was made with Python. With this, we created a database of quality advice requests-answer pairs extracted from the website reddit.com/r/advice. We then prompted AI models to answer these advice requests. This resulted in a dataset of 50 advice requests, each paired with four advice answers (Human, GPT3, GPT3.5 GPT4). In Study 1, participants were presented with advice requests along with advice. Their task was to indicate whether they believed the advice originated from a human or AI. In Study 2, participants were presented with advice requests accompanied by two pieces of advice. One piece of advice was always the human response, the other was randomly chosen from the AI models. Participants had to indicate which advice they found the most helpful. In Study 3 participants were presented with an advice request and advice from a random source. They were asked to rate the advice quality on four 1-item- scales (Helpfulness, Effectiveness, Appropriateness, Sensitivity).
Results: Study 1: Participants could only correctly identify above the chance level (p<.0001) the human advice. Study 2: Participants preferred AI advice over human advice (p<.0001). Study 3: Participants rated the AI advice quality significantly higher than human advice (Advice Quality, p<.00001; Helpfulness p < .01, Appropriateness p < .01, Sensitivity p<.0001, Effectiveness p=.079).

Keywords: *Advice, helpfulness, perception, artificial intelligence, AI detection.*

1. Introduction

The emergence of generative artificial intelligence (AI) has brought about the release of large language models to the public. A prominent example is Chat-GPT, which gained 100 million users within its first month. Meanwhile, like, the early days of the internet, people are increasingly turning to AI for information and advice (Carpenter, McDade, & Childers 2018; Golia, 2021). While we know this is happening, little is known about the quality of the AI output. This trend raises urgent questions regarding the quality of the content generated by AI systems if its content is different, and if we can even distinguish it from human resources.

Zellers et al. (2021), found that humans did not prefer AI advice produced by GPT-3, preferring it over human advice in only 4 % of cases. The best-performing model in the study T5-11B was preferred over human advice in only 14 % of cases, and this was a model that was specifically trained with advice data. Our study examines if this preference has changed with the new advances in AI technology.

2. Objectives

We investigate the advice-giving capabilities of humans, GPT-3, GPT-3.5, and GPT-4, in the domain of general life advice. Can people distinguish AI from human advice, do they have a preference for the advice given, and are there any differences in the advice quality between humans and AI across the GPT models?

3. Method: Web scraping and data preparation

We decided to follow in the footsteps of Zellers et al.'s (2021) approach of using reddit.com/r/advice. A Python script was created to identify reddit.com/r/advice posts that had more than 20 likes. In total 27000 advice posts were processed by the script between the dates: 03.05.2023 - 04.19.2023 of which 325 posts matched the criteria set in the script. The questions and the best-rated advice of all these posts were saved, of these, 50 were randomly selected and new advice replies were generated by the three AI models. In total that gave us 50 advice requests with 4 matching advice each.

The AI advice was longer than the human, especially GPT4; the average length for AI was 230 words, and humans 88 words. Also, the AI replies often used numbered lists. The AIs were therefore prompted to rewrite the advice without numbered lists. From that, a Python script selected 15 questions that had the least difference in length between AI and human advice, reducing (human advice mean = 82 words, and AI mean = 135 words). Using these 15 questions we created studies 1 and 3. For study 1 we additionally removed answers that revealed the source in the answer, leaving 9 questions.

3.1. Measures

In study 3: we used a perceived advice quality measure with 4 items (helpfulness, effectiveness, appropriateness and sensitivity) rated on a 7-point Likert type scale (Goldsmith & MacGeorge, 2000; Jones & Burleson 1997; MacGeorge et al., 2004). Overall perceived advice reply quality was estimated using the mean of the 4 items.

3.2. Design

Study 1: We explored how well people can distinguish between AI and Human advice. We had 74 participants (female=39, male=28, other=6), yielding 579 observations. In random order, we presented 9 advice requests with a randomly chosen advice. Participants were asked to guess if the advice was AI or human.

Study 2: We explored the helpfulness of AI advice. This was in essence a replication of Zellers et al 2021. We presented a total of 15 advice requests to 66 participants (female=37, male=26, other=3), yielding 327 observations. To these requests, we had advice from 3 AI models (GPT3, GPT3.5, and GPT4) and 1 human advice (best-rated advice from Reddit). In the survey participants randomly received 6 requests, accompanied with human advice and advice from one of the AI models (in random order). The participants were asked to indicate the more helpful advice. Participants were blind to the AI involvement.

Study 3: We explored the quality of AI advice. Sixty-five participants (female=39, male=23, other=3), yielding 1332 observations, 333 when equating one observation as the full 4 item scale rating. From the sample of 15 questions, participants were randomly presented with 6 advice requests with a randomly chosen advice. Participants were asked to rate the advice on a 7-point-Likert scale on 4 scales (Helpfulness, Effectiveness, Appropriateness, and Sensitivity). Participants were blind to the AI involvement.

4. Results

We chose the significance level of $p<.05$, but for each study we used an adjusted significance threshold (Adjusted Alpha), according to the Bonferroni principle of multiple comparisons (0.05 / comparisons).

Study 1 (Adjusted Alpha = $p<.01$): Participant observations revealed that people can identify human advice above chance level (Chi-Square, $p<.01$). In AI advice there was a non-significant trend towards correct identification, AI total ($p=.021$), GPT3 ($p=.676$), GPT3.5($p=.33$), GPT4 ($p=.017$)

Study 2 (Adjusted Alpha = $p<.0125$): Participant observations revealed a preference for AI advice over human advice (Chi-Square, $p<.0125$). When comparing each model separately against the human advice, we found a preference of AI advice in models, GPT 3 ($p<.0125$) and GPT3.5($p<.0125$). Although GPT4 ($p=.029$) showed a trend towards significance it did not pass the adjusted significance threshold.

Study 3 (Adjusted Alpha = $p<.0038$): Participant observations quality ratings revealed that AI advice was rated significantly higher than human advice (Mann Whitney, $p<.0038$). We found this to be true for Helpfulness ($p<.0038$), Appropriateness ($p<.0038$), and Sensitivity ($p<.0038$), but not Effectiveness ($p=.079$). Direct comparisons of each AI model against human advice, revealed significant differences in GPT3.5 and GPT4 ($p< .0038$), while GPT3 ($p=.0089$) was not significantly different in quality. Additionally, testing revealed that there are significant differences between the AI models in Advice Quality (Kruskal Wallis, $p<.0038$). Specifically in the subscale, Helpfulness ($p<.0038$). Non-significant trends were found in Effectiveness ($p=.0179$), Appropriateness ($p=.0268$), and Sensitivity ($p=.039$).

Figure 1.

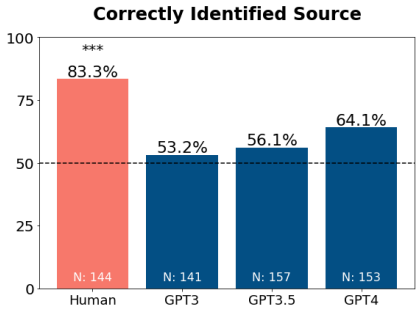


Figure 2.

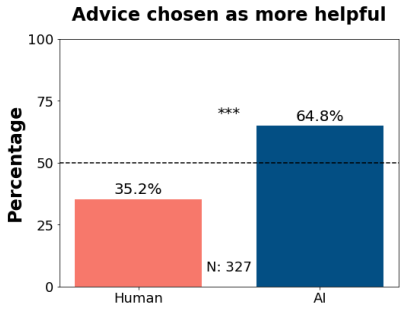


Figure 3.

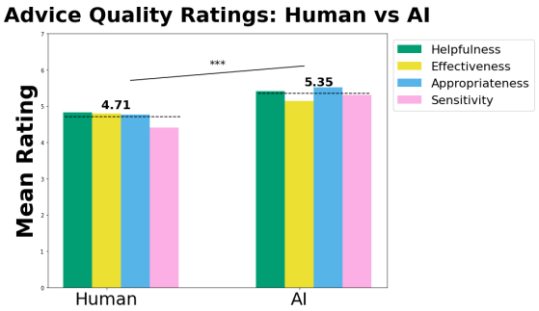
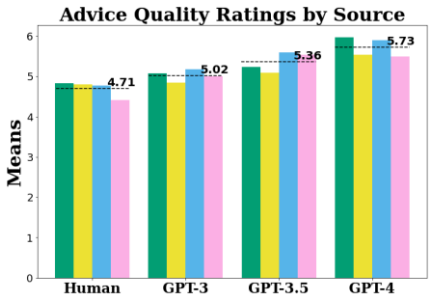


Figure 4.



5. Discussion

Our research shows that the quality of AI advice has dramatically improved in recent years, to the point where people find it hard to distinguish from human advice and even perceive AI advice to be of better quality. These results show an evolution since Zellers et al. (2021) and were recently confirmed in another recent study (Howe et al. 2023), that was published while our investigations were in motion. Our results carry with them big implications. It's still unclear if following AI advice will lead to better outcomes. This fact is alarming since the outcomes of following AI advice, though perhaps preferred, are still unknown. This might increase the impact of the technology in everyday life. What's our next step?

References

Carpenter, J., McDade, C., & Childers, S. (2018). Advice Seeking and Giving in the Reddit r/Teachers Online Space. In E. Langran & J. Borup (Eds.), *Proceedings of Society for Information Technology & Teacher Education International Conference* (pp. 2207-2215). Washington, D.C., USA: Association for the Advancement of Computing in Education (AACE).

Goldsmith, D. J., & MacGeorge, E. L. (2000). The Impact of Politeness and Relationship on Perceived Quality of Advice about a Problem. *Human Communication Research, 26*(2), 234-263. doi: 10.1111/j.1468-2958.2000.tb00757.x

Golia, J. (2021). *Newspaper Confessions: A History of Advice Columns in a Pre-Internet Age* (1st ed.). Oxford University Press.

Howe, P. D. L., Fay, N., Saletta, M., & Hovy, E. (2023). ChatGPT's Advice Is Perceived as Better than That of Professional Advice Columnists. *Frontiers in Psychology, 14*, 1281255. doi: 10.3389/fpsyg.2023.1281255

Jones, S. M., & Burleson, B. R. (1997). The Impact of Situational Variables on Helpers' Perceptions of Comforting Messages: An Attributional Analysis. *Communication Research, 24*(5). 530-555. doi: 10.1177/009365097024005004

MacGeorge, E. L., Feng, B., Butler, G. L., & Budarz, S. K. (2004). Understanding Advice in Supportive Interactions.: Beyond the Facework and Message Evaluation Paradigm. *Human Communication Research, 30*(1), 42-70. doi: 10.1111/j.1468-2958.2004.tb00724.x

Zellers, R., Holtzman, A., Clark, E., Qin, L., Farhadi, A., & Choi, Y. (2021). TuringAdvice: A Generative and Dynamic Evaluation of Language Use. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4856-4880. https://doi.org/10.18653/v1/2021.naacl-main.386