GPT as a financial advisor

Paweł Niszczota*, Sami Abbas

Poznań University of Economics and Business
Institute of International Business and Economics
Humans & AI Laboratory (HAI Lab)

* Corresponding author: Paweł Niszczota, Poznań University of Economics and Business, al. Niepodległości 10, 61-875 Poznań, Poland, pawel.niszczota@ue.poznan.pl

Data

Data and the pre-registration document are available at:
https://osf.io/tnbya/?view_only=ff15ee5ed54148da9208012b64b52dbb

Declarations of interest

None.

Ethical approval

Informed consent was obtained from all participants and the experiment adhered to appropriate guidelines. The study was approved by an Ethics Committee.

Financial support

This research was supported by grant 2021/42/E/HS4/00289 from the National Science Centre, Poland, and grant 004/RID/2018/19 from the Regional Initiative for Excellence program of the Minister of Science and Higher Education of Poland (2019-2022).

Wordcount (after appropriate exclusions)

2,486

GPT as a financial advisor

Abstract

We assess the ability of GPT–a large language model–to serve as a financial robo-advisor for the masses, by combining a financial literacy test and an advice-utilization task (the Judge-Advisor System). Davinci and ChatGPT (variants of GPT) score 58% and 67% on the financial literacy test, respectively, compared to a baseline of 31%. However, people overestimated GPT's performance (79.3%), and in a savings dilemma, they relied heavily on advice from GPT (*WOA* = 0.65). Lower subjective financial knowledge increased advice-taking. We discuss the risk of overreliance on current large language models and how their utility to laypeople may change.

JEL codes: D14, G11, G53

Keywords: financial literacy, robo-advice, financial advice, advice utilization, GPT-3.5, ChatGPT

1. Introduction

The introduction of ChatGPT – a conversational variant of the third iteration of the Generative Pre-Trained Transformer (GPT) model (Brown et al., 2020) – in November 2022 has caused an enormous increase in public interest in large language models (e.g., Roose, 2022; Sample, 2023). Soon after, a number of papers investigated the model's performance in various non-trivial tasks. For example, Bommarito II and Katz (2022) measured the performance of GPT on the Bar exam, required in the US to practice law. Performance (50.3%) far exceeded the expected score one would obtain by guessing (25%). Kung et al. (2023) assess the performance of ChatGPT on parts of a US medical license exam to be in excess of 60%. However, assessments of GPT on parts of a test mimicking a CPA exam suggest that its performance varies heavily, being far in excess of the guessing rate of 25% for parts of the exam, yet being poor on parts requiring quantitative reasoning (Bommarito et al., 2023).

This paper focuses on *financial advice* from GPT. Our paper has four contributions. First, we measure how GPT performs on tests aimed at measuring financial literacy. We consider this as a basic test of

GPT's ability to provide free or low-cost financial advice to the public. In essence, GPT could be considered a cheap robo-advisor for the masses (e.g., D'Acunto et al., 2019; Kaya et al., 2017). However, it remains unclear how well GPT will perform on these tasks, given that financial advice usually has a strong numerical component and this is the Achilles heel of large language models, even the most recent ones (Floridi and Chiriatti, 2020; Srivastava et al., 2022). Second, we compare the actual financial literacy scores with laypeople's predicted scores. This is intended to serve as an early indication of how expected performance diverges from real performance, which could promote overreliance on such technologies. Third, we use a household financial problem (concerning savings) to assess the degree that people utilize advice coming from GPT in a hypothetical scenario. Finally, we compare the performance of the latest release (GPT-3.5) in two variants: the conventional variant (*text-davinci-003*; Davinci) and the publicly-available, conversational variant (ChatGPT OpenAI, 2022). In comparison to the initially released version of GPT-3 (Brown et al., 2020), it incorporates reinforcement learning from human feedback: outputs are extensively tested and rated to help reduce the likelihood of producing non-factual or toxic outputs (Ouyang et al., 2022).

There are yet no studies on GPT's utility in finance and economics, except for the work published by Dowling and Lucey (2023) and Korinek (2023), who assess whether ChatGPT can be used to aid financial and economic research.

2. Assessment of financial literacy of GPT

2.1. Methodology

To measure financial literacy, we used the 'Big Three' items (Mitchell and Lusardi, 2022), and items from the Financial Literacy Baseline Survey (Heinberg et al., 2014), altogether 19 multiple-choice test questions. These had 2–5 options after the removal of options that indicated a lack of knowledge of the answer (*"Don't know", "Refuse to answer"*), which ensured a more accurate assessment of the actual performance of GPT against an agent who would randomly answer all options, but only in a way that could be potentially correct. An agent guessing would correctly answer 20%–50% of the time, and overall, guessing would lead to an expected score of 31%.

We tested the financial literacy of GPT by asking it each item 20 times. Tests were conducted on the most deterministic setting for the sampling temperature (0), as recommended by OpenAI when eliciting factual answers (OpenAI, 2023). Additionally, to assess how well GPT-3.5 performs depending on whether it has been pre-prompted before the actual question to act as a financial advisor, we test both the performance with and without such role-playing. In cases where GPT did not provide the letter indicating the correct answer (*a*-*e*) but answered correctly, we manually marked the answer as correct.

2.2. Results

In **Table 1** we present the mean accuracy of both investigated variants of GPT-3.5 on financial literacy. Using questions with the absence of a pre-prompt as a benchmark, Davinci achieved a 58% financial literacy score, while ChatGPT achieved a score of 67%. The difference between Davinci and ChatGPT was largely due to the poor performance of the former on the *Employer Match* and *Risk Diversification* items. Interestingly, pre-prompting Davinci to "play" the role of a financial advisor did not improve financial literacy scores, but in fact, reduced them by 8 pp. For ChatGPT there was also a reduction in the score, albeit it was much less pronounced (2 pp).

Table 1. Performance of Davinci and ChatGPT on the financial literacy test (20 trials)

	Davinci			**ChatGPT**		
	Without pre-prompt	With pre-prompt	Overall	Without pre-prompt	With pre-prompt	Overall
Big Three (3 items)	100.00%	100.00%	100.00%	73.33%	66.67%	70.00%
Compound Interest (4 items)	75.00%	50.00%	62.50%	67.50%	70.00%	68.75%
Tax-favored assets (5 items)	60.00%	60.00%	60.00%	57.00%	65.00%	61.00%
Inflation (2 items)	50.00%	50.00%	50.00%	50.00%	10.00%	30.00%
Employer Match (3 items)	33.33%	0.00%	16.67%	66.67%	66.67%	66.67%
Risk diversification (2 items)	0.00%	0.00%	0.00%	100.00%	100.00%	100.00%
Financial literacy (all items)	**57.89%**	**47.37%**	**52.63%**	**67.11%**	**64.47%**	**65.79%**

Notes: The pre-prompt was "*You are a financial advisor.*"

3. Advice utilization

To assess laypeople's expectations of GPT performance and assess how knowledge about actual performance impacts advice utilization, we used the Judge-Advisor System (Sniezek and Buckley, 1995;

van Swol and Sniezek, 2005). While the Judge-Advisor System has been used in the past to assess the utilization of algorithmic advice (Logg et al., 2019), it has not yet – to the best of our knowledge – been used to assess advice-taking from GPT.

We tested two hypotheses concerning advice-taking from GPT. First, we expected that predicted scores on the financial literacy test will predict advice-taking behavior on a more specific problem, that would require actual computations. Second, we hypothesized that for people with lower subjective financial knowledge, advice utilization will be greater.

3.1. Methodology

In the advice-taking task participants rated Davinci. First, they saw the 19-item financial literacy test we used prior and were asked how well they thought GPT performed on this test. All participants were then given a dilemma concerning the appropriate monthly payment necessary to reach a goal (see **Appendix** for details).

In contrast to financial literacy scores, there were substantial differences between Davinci and ChatGPT on this particular task. We gave this task to Davinci at three different temperatures (the minimal (0), the default (0.7), and the midpoint between these (0.35)) and ChatGPT. To measure accuracy, we used the Mean Absolute Percentage Error (MAPE; Harvey and Fischer, 1997). Results, presented in **Table 2**, suggest that Davinci in the lowest (most deterministic) setting has a MAPE of 0.4, indicating that in 20 trials it deviated 0.4% on average from the correct answer. ChatGPT performed substantially worse, with a MAPE score of 14.8.

Table 2. Performance of Davinci and ChatGPT on task (20 trials)

Model	Temperature	Mean	SD	MAPE
Davinci	0	637.50	0	0.4
Davinci	0.35	639.63	42.01	3.7
Davinci	0.70	657.27	64.56	7.8
ChatGPT	-	571.27	138.99	14.8

Notes: MAPE = | (estimate – true) / true | × 100.

Participants provided their answer and later received the answer provided by GPT. Similarly to the procedure used in the financial literacy test, we asked GPT this question 20 times and gave participants the mean of its answers.

The advice utilization we used was a weight of advice index (Harvey and Fischer, 1997), defined as:

$$WOA = \frac{final\ answer - initial\ answer}{advice - initial\ answer} \quad (1)$$

WOA typically ranges between 0 – indicating that the judge has entirely disregarded the advice and 1 – indicating that the judge used precisely the answer indicated by the advisor. Values outside of this range were winsorized, in line with previous research on the utilization of advice from machines (Logg et al., 2019).

To assess advice utilization from GPT, we recruited 200 participants from Prolific (Palan and Schitter, 2018), that had a 98% or higher approval rating, were located and born in the US, and whose first language was English. As preregistered, 16 participants that didn't correctly answer both attention check questions were excluded, leaving a final sample size of 184. The mean age of participants was 47.1 years (*SD* = 13.8); 32% of them interacted with GPT in the past. We used one item to assess their subjective financial knowledge *("My knowledge concerning finance is good."*, rated on a scale of 1 (*fully disagree*) to 7 (*fully agree*), with a mean score of 3.86 (*SD* = 1.50). We also applied a six-item financial knowledge test to assess objective financial (investment) knowledge (*M* = 4.69, *SD* = 1.15).

To incentivize participants to put effort into the task, we informed them that the top 10% of participants will receive a bonus (0.50 GBP).

The study was pre-registered at https://aspredicted.org/2QC_LJH and approved by an Ethics Committee. Data and materials are available at https://osf.io/tnbya/?view_only=ff15ee5ed54148da9208012b64b52dbb.

3.2. Results

3.2.1. Predicted performance of GPT

Participants predicted that GPT will have a score of 79.32% (*SD* = 17.65%), which was significantly higher than the actual score (t = 16.47, p < .001). The distribution of predicted scores broken-down based on the median subjective financial knowledge is shown in **Fig. 1**. People with low and high subjective financial knowledge predicted similar performance from GPT (M_{low} = 77.2% vs M_{high} = 80.8%, t = 1.32, p = .19).

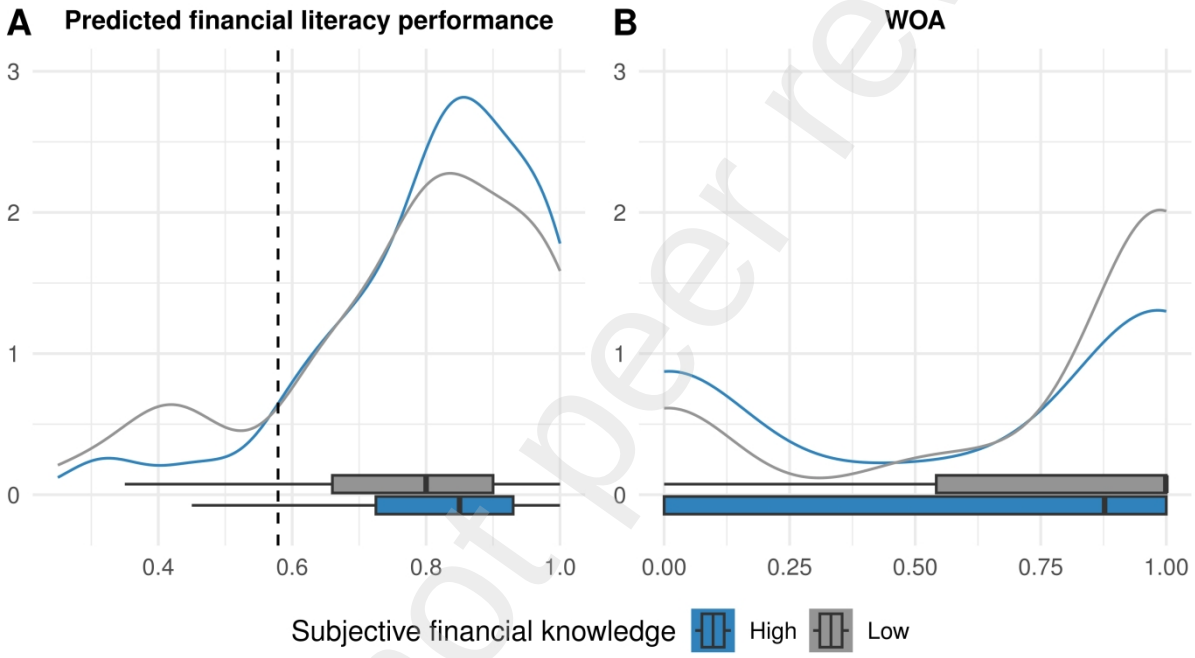

Fig. 1. The predicted financial literacy performance (A) and advice utilization (B) of GPT

Notes: This figure shows density plots and bar plots for high and low subjective financial knowledge (split based on the median). The dashed line corresponds to the actual performance of Davinci (*text-davinci-003*) on the 19-item financial literacy test without pre-prompts. Advice utilization is measured via WOA.

3.2.2. Advice utilization

The *WOA* score – after performing the pre-registered winsorization – was 0.650 (*SD* = 0.438). This indicated that participants relied more on GPT's estimates than their own. For reference, this is substantially higher than the *WOA* = 0.39 obtained in a recent meta-analysis (Bailey et al., 2022). As illustrated by **Fig. 1**, people with low subjective financial knowledge relied more heavily on advice from

GPT than people with high subjective financial knowledge (M_{low} = 0.740 vs M_{high} = 0.588, t = 2.39, p = .018).

In **Table 3**, we show regressions with WOA as the dependent variable. As alternative estimation methods, we used OLS and beta regression, the latter being suitable for data in the (0, 1) range, but also computable in the [0, 1] range after a transformation (Smithson and Verkuilen, 2006).

Table 3. Predictors of advice utilization from GPT (*WOA*)

	OLS	**Beta regression**
(Intercept)	1.16 ***	1.59
	(0.38)	(1.20)
Predicted score on financial literacy	-0.32 *	-0.66
	(0.19)	(0.60)
Subjective financial knowledge	-0.06 **	-0.15 **
	(0.02)	(0.07)
Objective financial knowledge	0.02	0.07
	(0.03)	(0.10)
Prior interaction with GPT	-0.14 *	-0.35
	(0.08)	(0.22)
Age	-0.03	-0.12
	(0.10)	(0.31)
Gender (baseline = *female*)	Yes	Yes
N	184	184
R^2 adjusted	0.035	0.078

Notes: Robust standard errors are shown in parentheses. * $p < 0.1$ ** $p < 0.05$ *** $p < 0.01$

Regressions weakly support the hypothesis positing that higher expected scores on the financial literacy tests lead to greater utilization of GPT's advice. However, subjective financial knowledge seems to show a robust link to advice-taking: for both estimation methods, people with lower subjective financial knowledge relied more heavily on advice from GPT.

4. Discussion

Our analysis suggests that extant large language models possess limited financial literacy, obtaining a 58-67% score against a 31% score that would be expected from random responses. Given that these were tests designed to study laypeople, it is reasonable to posit that large language models do not yet possess the capabilities to serve as robo-advisors (Chak et al., 2022; D'Acunto et al., 2019) for the masses. This is consistent with a weakness in large language models in quantitative reasoning

(Bommarito et al., 2023; Floridi and Chiriatti, 2020; Srivastava et al., 2022). However, people do not fully expect such underperformance, expecting GPT to score 79% on this test.

Our study provides preliminary evidence on how much people might rely on information from GPT concerning financial matters. Using a simple savings problem and the Judge-Advisor System, we estimate advice utilization to be substantial (*WOA* = 0.650), and even higher in people with low subjective financial knowledge (*WOA* = 0.740). The latter remain at a greater risk of overreliance on large language models. However, it is not yet known to what extent laypeople will use GPT for financial advice: some earlier research suggests that people with less financial knowledge might be less likely to use robo-advice (Isaia and Oggero, 2022; Niszczota and Kaszás, 2020).

4.1. The future of large language models in finance

4.1.1. Risk of overreliance

A comparison of the performance of Davinci and ChatGPT points to a divergence between the performance on the financial literacy test – in which the latter performs better – and the accuracy of estimates – in which the former performs better. This relationship is problematic, as someone who judges the performance of ChatGPT to be acceptable – based on answers to questions of limited difficulty, as the ones that were used in the financial literacy test – might build trust towards this model, which could lead to misuse. Large language models are not eager to acknowledge incompetence. In fact, large language models such as GPT-3.5 are known to "hallucinate" answers, i.e. answer prompts counterfactually, but deliver it in a confident fashion that would appear to be correct to a layperson. Although more recent iterations of these models are less prone to such hallucinations (Ouyang et al., 2022), it remains a significant weakness of these models. Overconfidence of ChatGPT would be of course paradoxical, as advice should not contain the same flaws that it is meant to address (Lewis, 2018).

4.1.2. Increase in utility

Future large language models should improve their performance and utility to laypeople. An optimistic account would be that they might soon gain new emergent abilities, such as excellent quantitative reasoning skills (Wei et al., 2022). A more realistic way of addressing the weaknesses of large language

models – even current ones such as Davinci – is to fine-tune them by certified professionals (e.g., experienced investment fund managers, CFAs, or PhDs in finance) that are experts in the field. Knowing how users with different levels of experience interact with large language models and see the responses that are typically produced by these models, they could better calibrate them, to – at the very least – refrain from providing an answer when there is a non-trivial probability that it is misleading. However, providing feedback from experts – in finance or other domains – will be more human and capital-intensive, and thus effectively more challenging to fine-tune.

4.1.3. Possibility of disutility

Most studies discuss artificial intelligence as intended to be socially beneficial, and much effort is put into aligning AI with societal goals. However, sophisticated natural language processing models are – like any technology – a double-edged sword, with the potential to be used by bad actors ("Tackling the perils of dual use in AI," 2022). For example, GPT can be asked to provide advice concerning tax evasion. While much effort is put into addressing these issues – either via fine-tuning on a massive scale (Ouyang et al., 2022) or via manually injecting norms into the models – there are already ways to overcome them and to use large language models in a socially-harmful way.

4.2. Limitations

There are several limitations to our work. First, we used proprietary language models from one source (OpenAI). Since the introduction of GPT-3.5, several companies have developed alternative models (e.g., Google Bard or LLaMA; "LLaMA," n.d.; Pichai, 2023). Second, responses from large language models are sensitive to prompting, and perhaps more accurate responses could be achieved using different prompts. Third, we used a specific combination of financial literacy items and only one hypothetical financial dilemma. Alternative items and problems might produce different responses. It is crucial not to extrapolate GPT's good performance on our financial dilemma to other financial dilemmas as, overall, the quantitative reasoning of large language models remains mediocre.

References

Bailey, P.E., Leon, T., Ebner, N.C., Moustafa, A.A., Weidemann, G., 2022. A meta-analysis of the weight of advice in decision-making. Curr. Psychol. https://doi.org/10.1007/s12144-022-03573-2

Bommarito II, M., Katz, D.M., 2022. GPT Takes the Bar Exam. https://doi.org/10.48550/arXiv.2212.14402

Bommarito, J., Bommarito, M., Katz, D.M., Katz, J., 2023. GPT as Knowledge Worker: A Zero-Shot Evaluation of (AI)CPA Capabilities. https://doi.org/10.48550/arXiv.2301.04408

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D., 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs].

Chak, I., Croxson, K., D'Acunto, F., Reuter, J., Rossi, A.G., Shaw, J.M., 2022. Improving Household Debt Management with Robo-Advice. Working Paper Series. https://doi.org/10.3386/w30616

D'Acunto, F., Prabhala, N., Rossi, A.G., 2019. The Promises and Pitfalls of Robo-Advising. Rev. Financ. Stud. 32, 1983–2020. https://doi.org/10.1093/rfs/hhz014

Dowling, M., Lucey, B., 2023. ChatGPT for (Finance) research: The Bananarama Conjecture. Financ. Res. Lett. 103662. https://doi.org/10.1016/j.frl.2023.103662

Floridi, L., Chiriatti, M., 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. Minds Mach. 30, 681–694. https://doi.org/10.1007/s11023-020-09548-1

Harvey, N., Fischer, I., 1997. Taking Advice: Accepting Help, Improving Judgment, and Sharing Responsibility. Organ. Behav. Hum. Decis. Process. 70, 117–133. https://doi.org/10.1006/obhd.1997.2697

Heinberg, A., Hung, A., Kapteyn, A., Lusardi, A., Samek, A.S., Yoong, J., 2014. Five steps to planning success: experimental evidence from US households. Oxf. Rev. Econ. Policy 30, 697–724. https://doi.org/10.1093/oxrep/gru036

Isaia, E., Oggero, N., 2022. The potential use of robo-advisors among the young generation: Evidence from Italy. Financ. Res. Lett. 48, 103046. https://doi.org/10.1016/j.frl.2022.103046

Kaya, O., Schildbach, J., AG, D.B., Schneider, S., 2017. Robo-advice–a true innovation in asset management. Deutsche Bank Research, August, available at https://www. dbresearch. com/PROD/DBR_INTERNET_EN-PROD/PROD0000000000449010/Robo-advice_-_a_true_innovation_in_asset_managemen. pdf.

Korinek, A., 2023. Language Models and Cognitive Automation for Economic Research. Working Paper Series. https://doi.org/10.3386/w30957

Kung, T.H., Cheatham, M., Medenilla, A., Sillos, C., Leon, L.D., Elepaño, C., Madriaga, M., Aggabao, R., Diaz-Candido, G., Maningo, J., Tseng, V., 2023. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health 2, e0000198. https://doi.org/10.1371/journal.pdig.0000198

Lewis, D.R., 2018. The perils of overconfidence: Why many consumers fail to seek advice when they really should. J Financ Serv Mark 23, 104–111. https://doi.org/10.1057/s41264-018-0048-7

LLaMA: Open and Efficient Foundation Language Models - Meta Research [WWW Document], n.d. . Meta Research. URL https://research.facebook.com/publications/llama-open-and-efficient-foundation-language-models/ (accessed 2.24.23).

Logg, J.M., Minson, J.A., Moore, D.A., 2019. Algorithm appreciation: People prefer algorithmic to human judgment. Organ. Behav. Hum. Decis. Process. 151, 90–103. https://doi.org/10.1016/j.obhdp.2018.12.005

Mitchell, O.S., Lusardi, A., 2022. Financial Literacy and Financial Behavior at Older Ages. https://doi.org/10.2139/ssrn.4006687

Niszczota, P., Kaszás, D., 2020. Robo-investment aversion. PLOS ONE 15, e0239277. https://doi.org/10.1371/journal.pone.0239277

OpenAI, 2023. OpenAI API [WWW Document]. URL https://beta.openai.com (accessed 1.30.23).

OpenAI, 2022. ChatGPT: Optimizing Language Models for Dialogue [WWW Document]. OpenAI. URL https://openai.com/blog/chatgpt/ (accessed 1.30.23).
Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C.L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., Lowe, R., 2022. Training language models to follow instructions with human feedback. https://doi.org/10.48550/arXiv.2203.02155
Palan, S., Schitter, C., 2018. Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance 17, 22–27. https://doi.org/10.1016/j.jbef.2017.12.004
Pichai, S., 2023. An important next step on our AI journey [WWW Document]. Google. URL https://blog.google/technology/ai/bard-google-ai-search-updates/ (accessed 2.7.23).
Roose, K., 2022. The Brilliance and Weirdness of ChatGPT. The New York Times.
Sample, I., 2023. ChatGPT: what can the extraordinary artificial intelligence chatbot do? The Guardian.
Smithson, M., Verkuilen, J., 2006. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. Psychol. Methods 11, 54–71. https://doi.org/10.1037/1082-989X.11.1.54
Sniezek, J.A., Buckley, T., 1995. Cueing and Cognitive Conflict in Judge-Advisor Decision Making. Organ. Behav. Hum. Decis. Process. 62, 159–174. https://doi.org/10.1006/obhd.1995.1040
Srivastava, A., Rastogi, A., Rao, A., Shoeb, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, Aditya, Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A.W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A.S., Andreassen, A., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, Animesh, Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B.R., Loe, B.S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B.Y., Howald, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ramírez, C.F., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C.D., Potts, C., Ramirez, C., Rivera, C.E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, D., Khashabi, D., Levy, D., González, D.M., Perszyk, D., Hernandez, D., Chen, Danqi, Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, Derek, Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D.C., Yang, D., Lee, D.-H., Shutova, E., Cubuk, E.D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodola, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E.A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E.E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G.I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G., Jaimovitch-López, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H., Schütze, H., Yakura, H., Zhang, H., Wong, H.M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J.F., Simon, J.B., Koppel, J., Zheng, J., Zou, J., Kocoń, J., Thompson, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J.U., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Jones, J., Tenenbaum, J.B., Rule, J.S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K.D., Gimpel, K., Omondi, K., Mathewson, K., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Colón, L.O., Metz, L., Şenel, L.K., Bosma, M., Sap, M., ter Hoeve, M., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Quintana, M.J.R., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M.L., Hagen, M., Schubert, M., Baitemirova, M.O., Arnaud, M., McElrath, M., Yee, M.A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M.,

Xu, M., Suzgun, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M.V., Peng, N., Chi, N., Lee, N., Krakover, N.G.-A., Cameron, N., Roberts, N., Doiron, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N.S., Iyer, N.S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares, P.A.M., Doshi, P., Fung, P., Liang, P.P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P., Eckersley, P., Htut, P.M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R.E., Gabriel, R., Habacker, R., Delgado, R.R., Millière, R., Garg, R., Barnes, R., Saurous, R.A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., LeBras, R., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R., Lee, R., Stovall, R., Teehan, R., Yang, R., Singh, Sahib, Mohammad, S.M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S.R., Schoenholz, S.S., Han, S., Kwatra, S., Rous, S.A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, Shikhar, Asaadi, S., Gu, S.S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Shyamolima, Debnath, Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S.P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S.T., Shieber, S.M., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Telleen-Lawton, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V., Prabhu, V.U., Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Yu, Hou, Yufang, Bai, Y., Seid, Z., Zhao, Z., Wang, Zijian, Wang, Z.J., Wang, Zirui, Wu, Z., 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. https://doi.org/10.48550/arXiv.2206.04615

Tackling the perils of dual use in AI, 2022. . Nat. Mach. Intell. 4, 313–313. https://doi.org/10.1038/s42256-022-00484-6

van Swol, L.M., Sniezek, J.A., 2005. Factors affecting the acceptance of expert advice. Br. J. Soc. Psychol. 44, 443–461. https://doi.org/10.1348/014466604X17092

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E.H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., Fedus, W., 2022. Emergent Abilities of Large Language Models. https://doi.org/10.48550/arXiv.2206.07682