TAKING ADVICE FROM CHATGPT *

Peter Zhang †
UC Berkeley
Berkeley, CA
`petez@berkeley.edu`

ABSTRACT

A growing literature studies how humans incorporate advice from algorithms. This study examines an algorithm with millions of daily users: ChatGPT. In a preregistered study, 118 student participants answer 2,828 multiple-choice questions across 25 academic subjects. Participants receive advice from a GPT model and can update their initial responses. The advisor's identity ("AI chatbot" versus a human "expert"), presence of a written justification, and advice correctness do not significantly affect weight on advice. Instead, participants weigh advice more heavily if they (1) are unfamiliar with the topic, (2) used ChatGPT in the past, or (3) received more accurate advice previously. The last two effects—algorithm familiarity and experience—are stronger with an AI chatbot as the advisor. Participants that receive written justifications are able to discern correct advice and update accordingly. Student participants are miscalibrated in their judgements of ChatGPT advice accuracy; one reason is that they significantly misjudge the accuracy of ChatGPT on 11/25 topics. Participants *under-weigh* advice by over 50% and can score better by trusting ChatGPT more.

Keywords ChatGPT · algorithm aversion · human computer interaction

1 Introduction

In late 2022, ChatGPT showed the world the power of large language models (LLMs) [1]. ChatGPT is a generative pretrained language model developed by OpenAI, an AI research lab. AI chatbots like ChatGPT and its cousins (BingChat, Bard, Jasper) achieve "surprisingly superior performance" [2] due to an instruction-tuning process that teaches them to do what humans want [3, 4]. Combined with pre-training at scale, LLMs are powerful interfaces for accessing knowledge [5, 6].

The most recent model GPT-4, which now underlies ChatGPT Plus, is much more powerful [7] and has been rigorously benchmarked on a variety of academic tests. According to OpenAI's internal testing, GPT-4 outperforms the median human test-taker on SATs, LSATs, GREs, and several AP exams [8]. Other researchers have found that ChatGPT can pass the bar [9], achieve medical certifications [10, 11, 12, 13], and even complete a college physics class [14].

The novel accessibility and broad capabilities of AI chatbots are likely to reshape education [5]. Many educators are scrambling to reconcile with ChatGPT with responses ranging from outright bans in school to welcome integration into curricula [15]. Some point towards risks to testing integrity [16] and plagiarism [17], while others argue that it provides personalized and immediate information [18, 19]. A recent meta-analysis finds that "the number of papers that see ChatGPT as a threat is almost equal to the number of those that view it as opportunity" [20]. Others still are rethinking traditional views of academic integrity and encouraging uses such as co-authorship [21, 22, 23].

The multiple choice (MC) exam is particularly vulnerable to a rethinking. MC questions continue to be a predominant format for assessing understanding, analysis and recall [24]. The strength of AI chatbots on multiple choice exams is worrying [25] because students most commonly cheat by consulting online sources [26]. While some have suggested workarounds [17], the fast-paced evolution of the underlying LLMs means that it "may not be long before [these] models become so intelligent that we can no longer exploit their weaknesses" [27].

This study seeks to document how students use information from ChatGPT on MC tests, contributing to a largely qualitative literature on how students empirically interact with AI chatbots [28]. While it takes MC tests as a starting point, the work has implications for broader research on algorithm aversion and appreciation, as well as on human-AI collaboration. The study is guided by two questions: First, what influences the weight humans place on chatbot advice? Second, are humans good at judging when AI chatbot advice is correct?

2 Literature Review

A rich literature examines how people take advice from algorithms. Two core competing findings are algorithm aversion [29] (a tendency to disproportionately punish algorithms when they err) and algorithm appreciation [30] (a tendency to prefer algorithm advice prima facie). Numerous studies have explored mediating mechanisms, including task objectivity [31], perceived competence [32], human input [33], learning [34, 35], and time pressure [36], among others [31]. One literature review categorizes these effects into algorithm characteristics (agency, performance, capabilities, and human involvement) and human characteristics (expertise and social distance) [37]. Another analyzes broad themes of expectations and expertise, decision autonomy, incentivization, cognitive compatibility, and divergent rationalities [38]. Five types of explanations are relevant to this study.

This study is a direct test of explanations about *social distance*. If algorithm aversion is truly a preference for humans, a natural remedy is to make algorithms more human-like [39]. Both adjacent literature [40] and experimental evidence suggests that people are more likely to accept advice from an anthropomorphized algorithm . In the business world, AI chatbots are now successful consumer-facing assistants [41], and the their perceived human-likeness is important to their success [42]. At the same time, other studies suggest that appearing too human can induce aversion if algorithms traverse into an uncanny valley [43]. ChatGPT's ability to provide natural language explanations comparable to humans [44] may cause humans to treat ChatGPT similarly to a human advisor and distinctly from other algorithms.

Three other explanations are commonly cited. The first, *task difficulty*, suggests that increasing task difficulty causes people to rely more heavily on (algorithmic) advice [45, 46] and is supported by real-world evidence on teachers [47]. In this study, familiarity in the question topic an approximate measure of task difficulty. A second explanation, *algorithm familiarity*, reasons that people who are more familiar with using algorithms for some task will be less averse to the advice [31, 48], an effect that was confirmed in a real-world medical context. This study measures algorithm familiarity by asking questions about past usage. The third explanation, *experience*, argues that participants are rationally updating their beliefs about algorithm competence and that presenting their performance can reduce aversion and develop trust over time [49, 50, 51], although some studies find that accuracy matters less than expected [52]. This study uses a simple model of participant beliefs about advice accuracy as a measure of experience.

Finally, a developing literature studies the effect of algorithm *interpretability* on aversion. Interpretability is theorized as allowing "the user to rapidly calibrate their trust in the system's outputs, spotting flaws in its reasoning or seeing when it is unsure" [53]. Studies have found mixed effects of output interpretability [54, 55, 56, 57] and model transparency [58, 59], although field experiments on physicians find that they benefit from explainable AI advice [60, 61]. In this study, providing GPT model's text reasoning enables a test of whether interpretability makes a difference.

Surprisingly few studies have examined the role of ChatGPT as an adviser. Some studies have explored potential problems with using ChatGPT for advice on health [62, 63, 64, 65], investing [66], and education [19]. Empirical studies have documented a corrupting effect of moral advice generated by GPT-2 [67], GPT-3[68], and ChatGPT [69]. On Twitter, GPT-3 generated texts appear to be more effective at convincing humans to believe (accurate and inaccurate) information [70]. Finally, humans appear to trust robots more with ChatGPT as an interface [71]. One recent study studies ChatGPT in the algorithm aversion context on a essay-writing task [72]. The authors find that while people may devalue the outputs of ChatGPT relative to a human author, they judge the content equally and are not deterred from sharing.

Little is known about how humans judge the accuracy of ChatGPT. A plethora of studies have shown that humans tend to overestimate their own abilities [73] and misjudge the abilities of others [74]. Similarly, studies suggest that LLMs calibration is good after pre-training [75, 76, 77] but degrades after learning from human feedback [78]. Yet studies have not evaluated whether humans accurately estimate the accuracy of LLM outputs. One study suggests that humans may become miscalibrated on AI feedback because of misallocation of blame [79], the this study and others fail to explicitly document the level and nature of (mis)calibration. This study seeks to fill that gap by evaluating human confidence in LLM outputs on a broad set of topic areas.
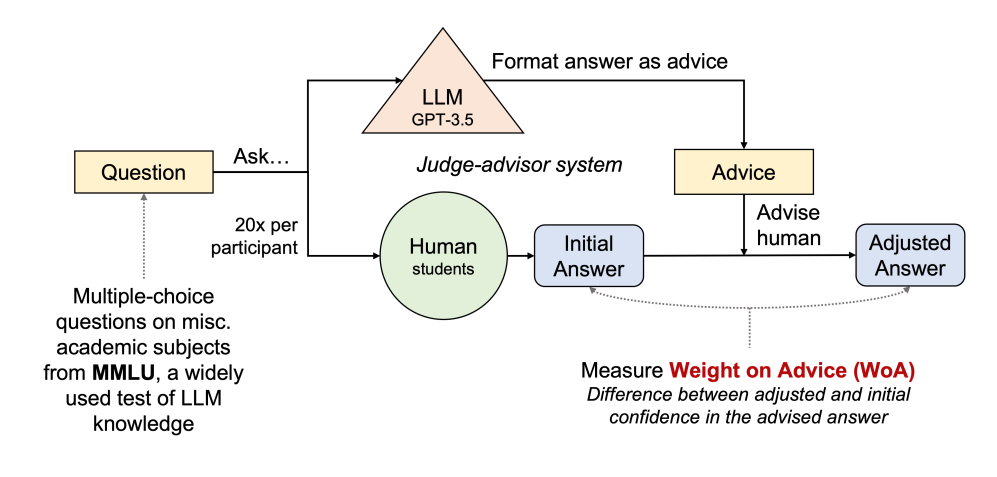
Figure 1: **Overview of study design.** Both LLMs and human participants answer questions. The study focuses on how humans take LLM advice.

Table 1: MMLU topics included in this study.

Supercategory	Topics	Example Question
STEM	Clinical Knowledge, Physics, Elementary Mathematics, Formal Logic, APs (Biology, Chemistry, Comp. Sci, Physics, Statistics), Human Aging	In which situation can the expression 64 + 8 be used?
Social Science	APs (Human Geo, Government, Macro/Micro, Psych), Sociology, U.S. Foreign Policy, Global Facts	What does Berger (1963) describe as a metaphor for social reality?
Humanities	APs (US/World/European History), Philosophy, Misc. topics	Descartes argues against trusting the senses on the grounds that _____.

3   Procedure

Overview   The study simulates an environment in which students receive aid from ChatGPT. MC questions are sourced from real academic tests and original outputs are obtain by quering GPT models. Participants attempt to make calibrated guesses before and after seeing the advice that is generated. An overview of the study design is displayed in Figure 1. All code and data except for survey responses are documented in the accompanying GitHub repository. The study methods are preregistered on AsPredicted under predictions #122800 and #126040.

Dataset   Answers are drawn from the Massive Multitask Language Understanding (MMLU) dataset [80], a widely used benchmark [8] of LLM knowledge understanding that broadly encompasses academics. The dataset consists entirely of MC questions and draws from real tests such as the Advanced Placement exams. Participants answer questions from only 25 of the original 57 topics, topics which college students are expected to have a reasonable chance to succeed. A total of 688 questions are sampled from the topics. See Appendix A.1 for descriptions of the 25 topics and selection procedure.

Model evaluation   The advice is generated by GPT-3.5, a LLM by OpenAI fine-tuned to follow human instructions, on the constructed dataset [81]. Specifically, calls are made to the Completions API with `text-davinci-003` as the engine [3] Models are prompted with standard and chain-of-thought (CoT) prompts [82]. CoT prompts yield the same accuracy but better explanations. The advice used in the survey is generated using a zero-shot CoT prompt. See Appendix A.2 for a comparison to standard prompting and illustration of the prompt text.
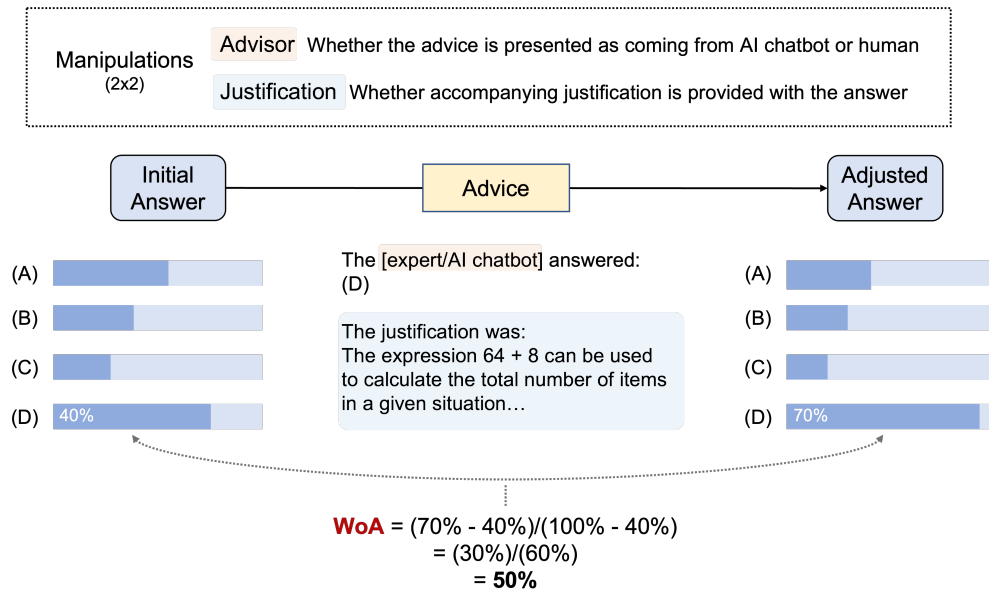
---

[3]See the API documentation.

Figure 2: **Judge-advisor system.** Participants provide judgements about the probability that each answer is true. WoA is a measure of how advice changes the probability allocated to the advised answer.

Lab experiment   Participant use this advice is a survey-based lab experiment.[4] The setup models the well-studied judge-advisor system [83]. Participants are shown randomly selected questions and report their confidence in each answer choice before and after receiving advice. The advice is manipulated by varying the advisor's identity and selectively providing justifications. Participants receive advice from an advisor randomly identified as a generic "expert" or an "AI chatbot". They are also randomly assigned to receive a justification in addition to the answer. The manipulations are displayed in Figure 2.

The experiment is administered via a Qualtrics survey. A live link and full printout of the survey are available for readers. Participants

- are assigned to the conditions;
- must pass a simple attention check;
- provide their level of familiarity ("comfortable", "neutral", or "uncomfortable") with 8 topic areas that are constructed by grouping topics, as well as their major(s);
- complete an example that explains the judge-advisor setup, the concept of confidence, and identifies the advice format (advisor identity and presence of justification);
- pass a manipulation check that reinforces the advisor identity;
- complete at least 20 questions in which they:
    - are assigned a random question and provide an initial answer;
    - receive advice and update their answer;
    - discover the correct answer and the points they have earned; and
    - have the opportunity to out-opt once they have completed 20;
- fill out a questionnaire about their usage of ChatGPT; and finally
- exit the survey.

The survey flow is displayed in Figure 3.

Scoring and compensation   Participants are scored by a point system that rewards accurate and calibrated answers with cash prizes. The system is based on Brier scores, a widely used scoring rule for encouraging both accuracy and

---

[4]The experiment is approved under UC Berkeley CPHS Protocol #2023-03-16125.
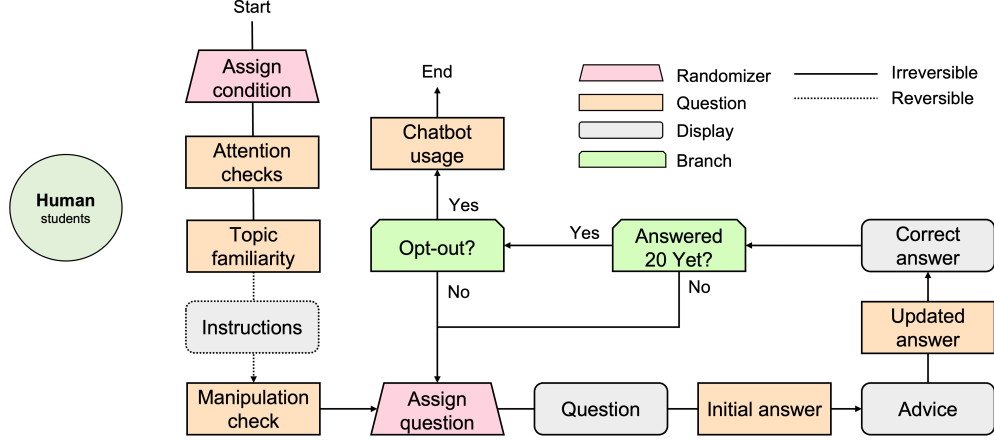
Figure 3: **Qualtrics survey flow.** Survey blocks are color coded by element type and line pattern corresponds to reversibility. The survey begins with a consent notice and ends with a debriefing. Note that participants are *required* to pass the attention and manipulation checks. Participants may return to previous instruction pages to pass the manipulation check.

calibration [84]. Let $f_{X,\text{init}}$ denote the initial confidence and $f_{X,\text{adj}}$ denote the adjusted confidence for each answer choice $X \in \{A, B, C, D\}$. Let o_X be an indicator for whether X is correct. Then, the score is:

$$\text{BS}(f) = \sum_{X \in \{A,B,C,D\}} \left[(f_X - o_X)^2 \right]$$

The Brier score is scaled to give 0 points to a uniform (25% across choices) distribution and 750 points for a full-confidence correct answer. The score is asymmetric insofar as it penalizes a full-confidence incorrect answer by -1250 points. The score is centered at 0 so that participants are not able to earn points by merely completing more questions with uniform distributions. The re-scaled scoring rule evenly weights the initial and adjusted forecast:

$$\text{Score} = \sum_{f \in \{f_{X,\text{init}}, f_{X,\text{adj}}\}} 750 - 1000 \cdot \text{BS}(f)$$

Participants can earn prizes by (1) placing among the top 5 scorers and earning 10 USD or (2) through a random drawing for 50 USD with tickets proportional to score. The former is designed to reward effort[5] while the latter ensures that payout remains somewhat proportionate to score [85].

Participants   A total of 142 undergraduate students at UC Berkeley are recruited through the Research Participant Pool (RPP) at the Haas School of Business. The participants are primarily business majors in their third and fourth years. Six small pilot sessions were conducted from 04/04/2023 to 04/11/2023 to debug the survey. The 12 sessions comprising the study dataset were administered from 04/13/2023 to 04/25/2023 and included 118 participants. All sessions were conducted at the Experimental Social Science Laboratory. Participants were compensated with course credit and performance-based monetary awards.

Data Processing   Letting $\hat{X}$ denote choice of the advisor, weight on advice WoA is computed as

$$\text{WoA} = \frac{f_{\hat{X},\text{adj}} - f_{\hat{X},\text{init}}}{1 - f_{\hat{X},\text{init}}}$$

The winsorization procedure replaces negative values with zeroes:

---

[5]Anecdotally, scoring dramatically improves participant engagement. Subjects reported feeling more interested and invested in the questions, particularly compared to a setup that does not reveal the correct answer. The effect appear to be even stronger when the reward is deterministic.
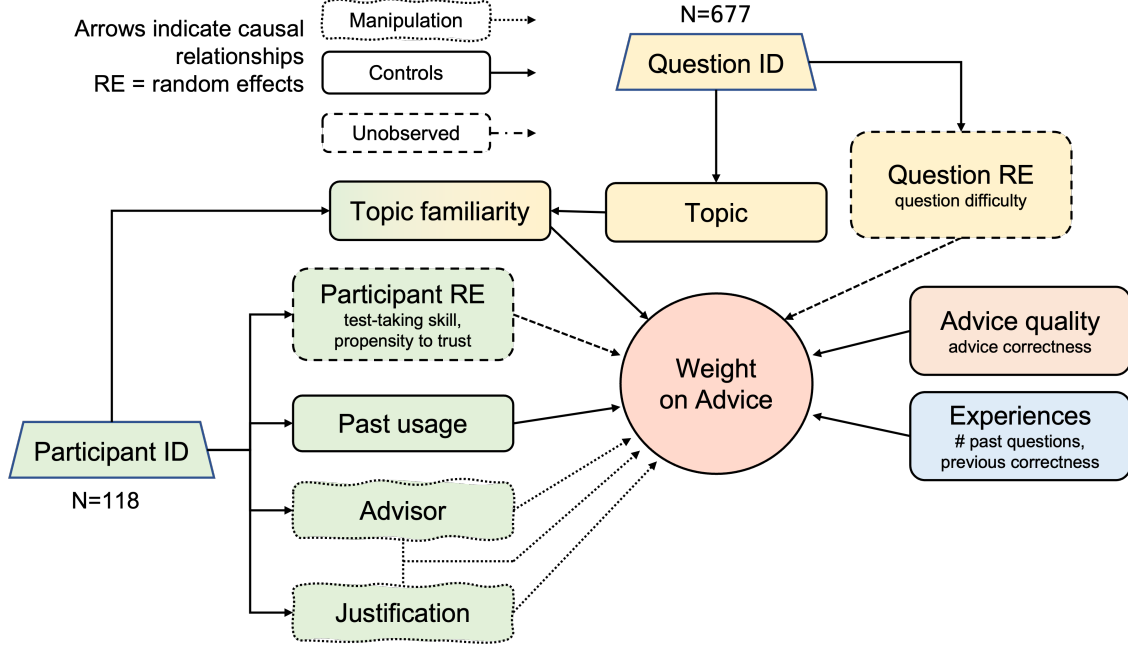
Figure 4: **Proposed causal mechanisms.** The study examines several predictors of weight on advice. The random assignment of **participant** / **question**, **experience** over several questions, and varying **advice quality** permits a decomposition of what predicts **weight on advice**.

$$\text{WoA} \leftarrow \max(0, \text{WoA})$$

The term "advice confidence" denotes $\text{AC} = f_{\hat{X},\text{adj}}$, the adjusted confidence in the advisor's answer. Categorical variables (topic familiarity, chatbot usage) are converted to integer values using basic rules. A Beta-Bernoulli process is used to model beliefs in advice correctness [86]. See Appendix A.3 for some limitations of the approach and details of these decisions.

4 Analysis

Participant answered 2,828 questions with an average of 23.97 questions per participant. After computing WoA, 166 questions (5.87%) with negative weight on advice are winsorized. Descriptive statistics are available in Appendix B.1. Qualitative findings about ChatGPT usage are presented in Appendix B.10.

4.1 Weight on Advice

Hypothesis   AsPredicted #126040) predicts that participants place greater weight on advice when the advisor is identified as an "AI chatbot."

Method   Weight on advice is progressively regressed on a broader set of variables in each specification (see Figure 4). Weight on advice is regressed on the advisor identity (`advisor`), justification (`give_justification`), and their interaction in **A**. Controls, including topic familiarity (`topic_familiarity`) are included in **B**, past usage (`usage_level`) in **C**, advice quality (`advice_is_correct`) in **D**, and experiences (`advice_accuracy_belief`, `question_num`) in **E**. (AsPredicted #126040) preregisters regressions **A**-**B** and conduct regressions **C**-**E** as a non-preregistered exploratory analysis. All regressions include random effects for participants (`participant_id`) to account for unobservable differences in participants such as test-taking skill, propensity to trust, calibration skill, etc. Random effects are included for questions (`question_id`) to account for within-topic differences in question difficulty. For concision, additional analyses are reserved for Appendix B.

Results   The results of the regression are displayed in Table 2. From specification **A**, there is no support for the initial hypothesis that WoA is greater if the advisor is identified as a chatbot. The coefficients are directionally correct but not

statistically significant. In this and other specifications, the random effects for participants and questions are highly significant. Appendix B.2 explores whether participant engagement might mediate the effect.

After including topic familiarity in specification **B**, there is a highly significant increase in weight of advice when the participant is uncomfortable with the topic. Compared to a baseline of comfort in the topic, weight on advice is 6.1%, 95% CI [2.19-10.08%] higher when a participant is uncomfortable in the topic. The effect is persistent and similar in size across specifications **C**-**E**. Robustness checks are conducted in Appendix B.3.

In specification **C**, past usage of chatbots has a marginally significant effect for participants in the "AI chatbot" advisor condition. The interpretation is that for each step increase in usage level (e.g. having used AI chatbots instead of merely hearing about them), weight on advice increases by 5.0 %, 95% CI [0.0%, 9.9%]. For participants in the "expert" advisor condition, the effect is not significant, suggesting that the result is driven by participant understanding of the advisor's capabilities. This effect is persistent and similar in size across specifications **D**-**E**. Appendix B.4 suggests that the result may mostly be driven by whether or not participants have used chatbots before.

On face, specification **D** reveals a surprising absence of a direct effect from the quality of advice, measured as whether the advice is actually correct. The coefficient becomes significant in specification **E**, suggesting that participants place 2.9% more weight on advice if it is true. Additional exploratory analyses are performed in Appendix B.5 controlling for initial confidence; the results reveal a strong effect that is mediated by giving justifications.

Finally, specification **E** identifies significant effect of experience that may be mediated by the advisor's identity. For every 10% increase in believed advice accuracy, participants with an AI chatbot advisor place 6.02% 95% CI [4.29%, 7.75%] greater weight on advice. If the advisor is a generic expert, the coefficient is 5.05% 95% CI [3.43%, 6.66%] per 10% increase in belief. Participants appear to place less weight on advice over time. For each additional question completed, participants place 0.4% 95% CI[0.362%, 0.6%] less weight on advice. Appendix B.6 shows that the deflated coefficient in the human expert condition is robust to beliefs. Moreover, there is a significant effect of the last advice's correctness that is mediated by advisor identity.

4.2 Advice Confidence

Hypothesis (AsPredicted #126040) predicts that (1) student's advice confidence will display overconfidence in language model accuracy and that (2) the overconfidence is mitigated by feedback.

Method For choice X and question j, let $t_{j,X}$ denote whether X is correct and $f_{j,X}$ denote the participant's confidence in the choice. Calibration curves are constructed by partitioning advice confidences over $[0, 1]$ into 10 equal-width bins and plotting the average advice confidence $e_i \equiv \mathbb{E}_{j \in I}[f_{j,X}]$ and accuracy $o_i \equiv \mathbb{E}_{j \in I}[t_{j,X}]$ for each bin i.

This section focuses on the advised choice $\hat{X}$ with the goal of evaluating how well the participants evaluate advice accuracy. Setting $e_i \equiv \mathbb{E}_{j \in I}[f_{j,\hat{X},\mathrm{adj}}]$ and accuracy $o_i \equiv \mathbb{E}_{j \in I}[t_{j,\hat{X}}]$, a calibration curve is constructed for participant's confidence in the advisor's answers. Miscalibration is measured by expected calibration error (ECE), the average deviation from ideal calibration weighted by sample size. Letting $P(i)$ denote the proportion of samples in bin i, ECE is defined as:

$$\mathrm{ECE} = \sum_{i=1}^{10} P(i) \cdot |o_i - e_i|$$

To evaluate whether participants become more calibrated over time, ECE is computed over groups of 5 question numbers. To calculate standard errors, ECE is bootstrapped on 1000 samples within each question group.

As preregistered, AC is measured for each topic and compared to the actual accuracy of the model on a larger evaluation set on the topic (discussed in Appendix A.2). Mistaken beliefs about per-topic confidence are identified by a Brunner-Munzel [87] comparisons between the mean accuracy $T = \{t_j\}$ and confidence $F = \{f_{j,X,\mathrm{adj}}\}$. The false discovery rate is controlled to 0.05 by using the Benjamin-Hochberg procedure [88].

Finally, participant performance is compared to a simple, proportionate update under different values of WoA. The optimal weight on advice $\mathrm{WoA}^* \in [0, 1]$ is the value that minimizes Brier Score.

Results Consistent with an extensive prior literature [73], participants are overconfident in their own answers. Figure 5a reveals (1) significant overconfidence in initial answers ($\mathrm{ECE}_{\mathrm{init}} = 0.183$) for confidence levels greater than 0.5 and (2) attenuated but persistent overconfidence in adjusted answers ($\mathrm{ECE}_{\mathrm{adj}} = 0.137$). Receiving advice appears to "lift" the actual accuracy of high-confidence answers and improve calibration.

Table 2: Results of regression analyses examining weight on advice.

	A	B	C	D	E
Intercept	0.329*** (0.044)	0.302*** (0.045)	0.169** (0.081)	0.161** (0.082)	-0.159* (0.096)
advice_accuracy_belief					0.602*** (0.088)
advice_accuracy_belief:advisor[T.expert]					-0.097 (0.120)
advice_is_correct[T.True]				0.013 (0.015)	0.029** (0.014)
advisor[T.expert]	-0.027 (0.059)	-0.028 (0.059)	0.056 (0.118)	0.054 (0.118)	0.067 (0.137)
advisor[T.expert]:give_justification[T.yes]	-0.027 (0.079)	-0.021 (0.079)	-0.025 (0.079)	-0.025 (0.079)	-0.023 (0.076)
give_justification[T.yes]	0.057 (0.058)	0.056 (0.058)	0.064 (0.057)	0.064 (0.057)	0.074 (0.055)
participant_id Var	0.364*** (0.056)	0.362*** (0.055)	0.354*** (0.055)	0.353*** (0.055)	0.337*** (0.052)
question_id Var	0.044** (0.018)	0.040** (0.018)	0.040** (0.018)	0.040** (0.018)	0.045** (0.018)
question_num					-0.004*** (0.001)
topic_familiarity[T.Neutral]		0.026 (0.018)	0.027 (0.018)	0.027 (0.018)	0.026 (0.017)
topic_familiarity[T.Uncomfortable]		0.061*** (0.020)	0.062*** (0.020)	0.062*** (0.020)	0.063*** (0.020)
usage_level			0.050* (0.025)	0.049* (0.025)	0.045* (0.024)
usage_level:advisor[T.expert]			-0.033 (0.036)	-0.033 (0.036)	-0.019 (0.035)

Considering only adjusted confidence in advised choices, participants are significantly more miscalibrated ($\text{ECE}_{\text{advised}} = 0.201$), displaying both overconfidence and underconfidence at different confidence levels. Participants dramatically underestimate the accuracy of advised choices at confidence levels below 0.5. For example, when participants place 0 to 10% confidence in the advisor's answer, the answer is actually correct 42.4%, 95% CI [29.2%, 55.5%] of the time. Moreover, the effect applies across both advisor conditions (Figure 5b) and is persistent (see Appendix B.8).

The remainder of this section considers only participants in the AI chatbot condition. These participants are limited in their ability to predict differences in ChatGPT's advice quality across topics. Participants overestimate advice accuracy on Elementary Mathematics questions and underestimate accuracy for 10 topics displayed in Figure 6. Average beliefs in advice accuracy are only moderately correlated (Pearson's $r=0.339$, $p=0.097$) with actual advice accuracy when grouped by topic (Figure 7b).

These errors are significant for knowing when to place weight on advice. WoA and AC are highly correlated at the question level (Pearson's $r=0.639$, $p=1.69\text{e-}150$) and participants place significantly more (Human Sexuality) or less weight (Elementary Mathematics) on some topics compared to others (Figure 7a).

Overall, participants do not place enough weight on advice (Figure 8). The Brier score is minimized at $\text{WoA}^* = 0.61$, achieving an average score of $\overline{\text{BS}} = 0.556$. The optimal weight on advice is over 50% higher than the average participant weight on advice, $\overline{\text{WoA}} = 0.367$.

Notably, participants also score ($\text{BS} = 0.673$) significantly worse in practice if they uniformly applied the same weight on advice with a proportionate update ($\text{BS} = 0.588$). The poor performance compared to a uniform baseline is due to (1) misallocation, the suboptimal proportioning of confidence to other answer choices and (2) extremism, participants' tendency to place no weight on advice or too much weight on advice, leading to excessive overconfidence. Appendix B.9 analyzes both effects and concludes that extremism is a larger factor.

(a) All participant answers.
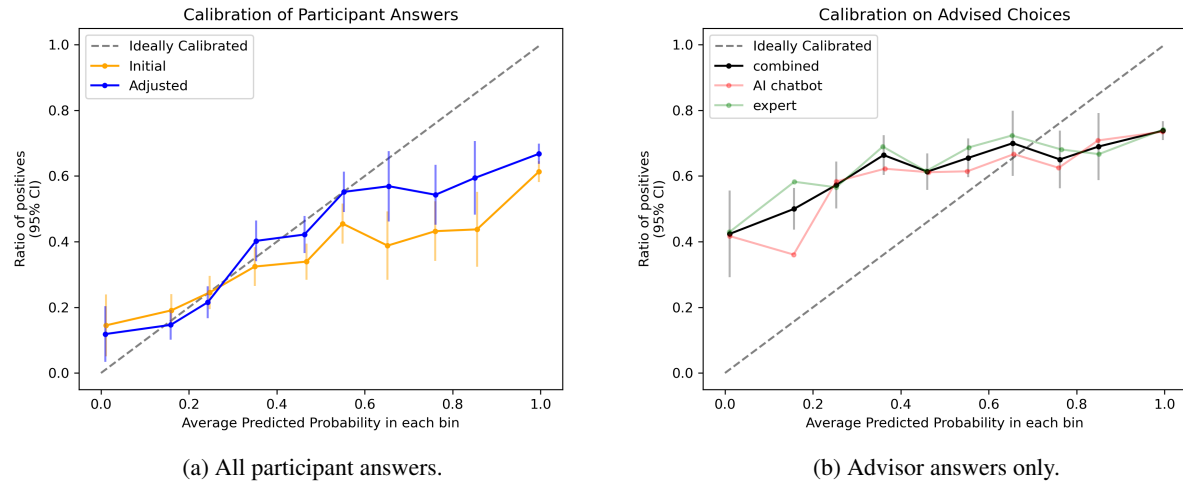
(b) Advisor answers only.

Figure 5: **Calibration curves.** Error bars are 95% confidence intervals. Dotted line corresponds to a theoretically ideal calibration curve in which average predicted probability exactly equals ratio of positives.

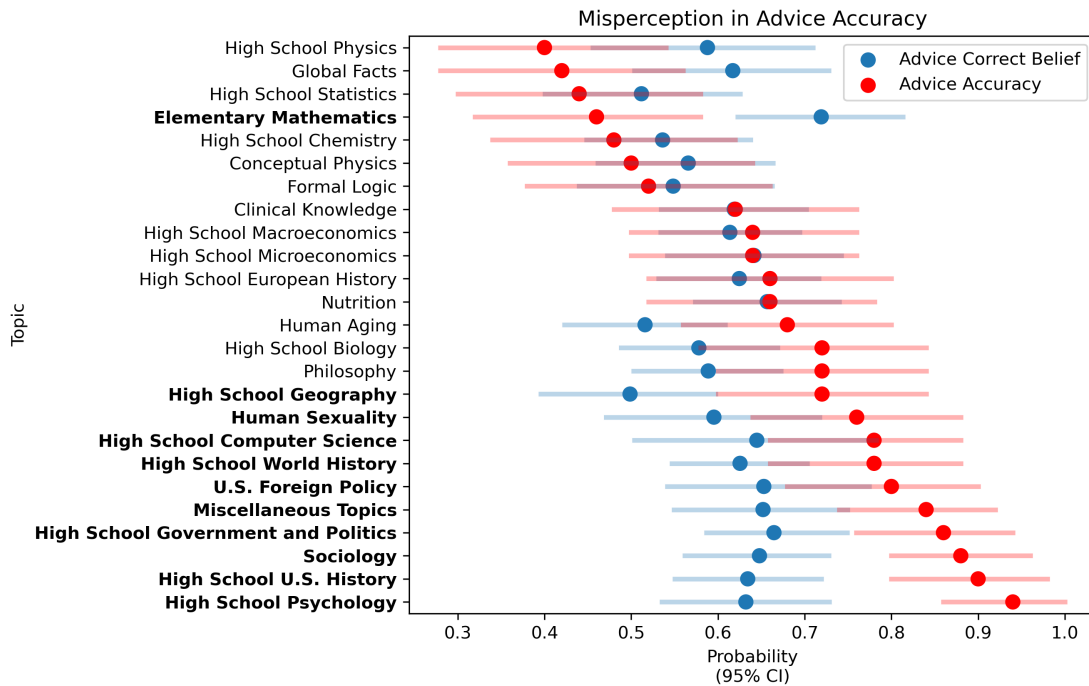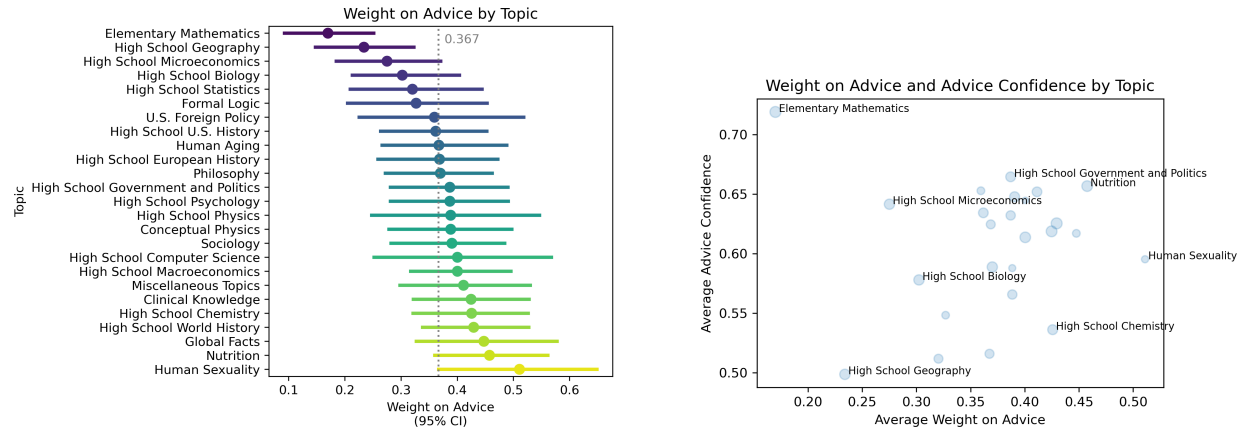Figure 6: **Believed and actual advice accuracy.** Error bars are 95% confidence intervals. Blue corresponds average participant confidence in the correctness advised answer. Red corresponds to the results from an evaluation on 50 questions in each subject. **Bolded** topics have significantly different advice correct beliefs (blue) and advice accuracy (red), suggesting a discrepancy between participant beliefs and reality.

(a) **Weight on advice by topic**. Error bars are 95% confidence intervals. The average weight on advice is plotted as a dotted line.

(b) **Weight on advice and correctness**. Point size corresponds to the number of samples in the topic.

Figure 7: Figure 7a examines weight on advice for each topic and Figure 7b displays its relationship with advice confidence.
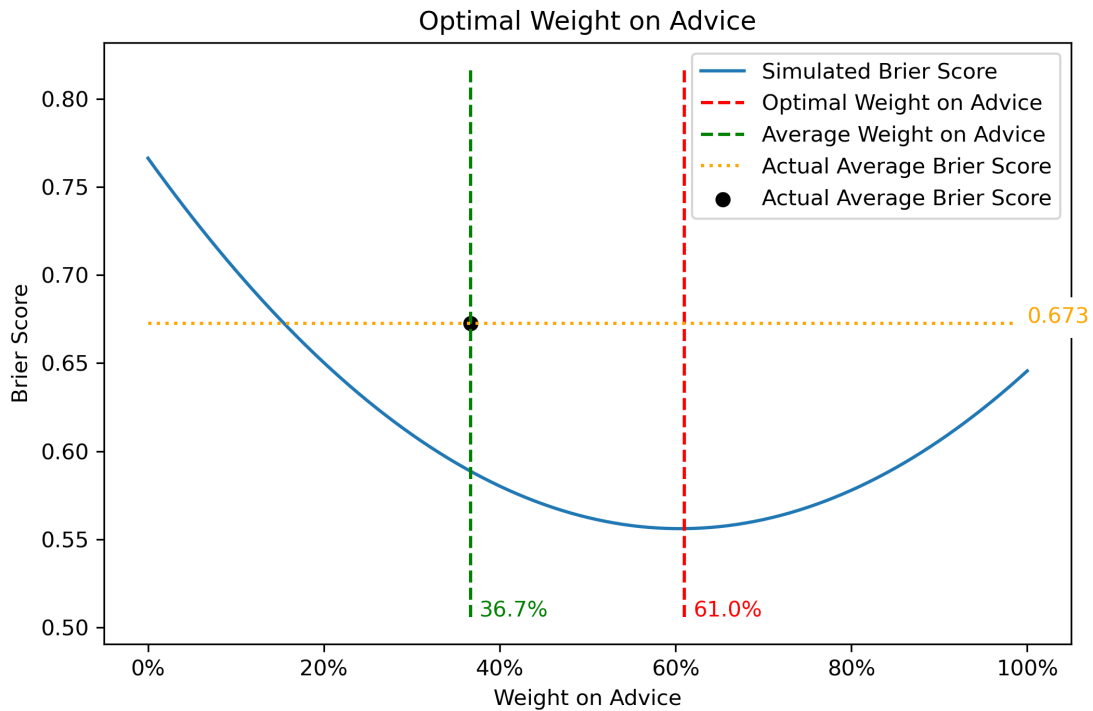

Figure 8: **Brier score-optimal weight on advice.** The dot marks actual participant weight on advice and Brier score. Participants significantly underweigh advice (optimal 61.0% versus actual 36.7%) and under-perform the baseline even at the sub-optimal weight on advice.

5 Discussion

This study contributes to literature on algorithm aversion and human computer interaction by performing a study of how students incorporate advice from ChatGPT on MC tests.

Weight on Advice   Results are summarized in Table 3. Advisor identity, presence of justification, and their interaction do not significantly affect weight on advice. The effect continues to be insignificant at a 95% level after including various controls. This result joins several other studies in finding a no effect of the advisor's identity on weight on advice [37] and agrees with another study on ChatGPT [72]. At the participant level, the sample is powered to detect a median effect size comparable found previously [30]. At the measured effect size, identifying an effect for advisor identity or presence of justification requires a sample size about four times larger (Appendix B.11). This might suggest that algorithm aversion or appreciation is less significant for ChatGPT than for other algorithms. Alternatively, the null result may be an artifact of the description of the non-algorithm advisor, which is vaguely introduced as an "expert" [32]. Moreover, in Appendix B.2, the effects are recovered by including an interaction term for optional questions, suggesting that participant engagement might be adding noise to the results. Limitations to generalizing the result are discussed the limitations section below.

The analysis of controls confirms several prior results. Task difficulty in the form of topic familiarity mediates advice usage [45]. Harder tasks increase weight on advice more so for AI than human advice, but not significantly so (Appendix B.3).

In exploratory analyses, prior usage of chatbots predicts greater weight on advice in the chatbot condition. The result agrees with prior studies showing that familiarity with the algorithm predicts greater usage of algorithms [31]. Given the relatively recent release of ChatGPT, students and educators still have different levels of familiarity with the technology [15]. These results could explain for why prior usage and ChatGPT are highly correlated [89]: new users underestimate the performance of the tool and gradually learn to trust and use it more. This reinforcing dynamic may predict how different people will use ChatGPT in the classroom and beyond.

In a further exploratory analysis, performance on previous questions and corresponding beliefs about advice accuracy are predictive of WoA, agreeing with prior work [49, 50, 51]. The result is stronger for AI chatbots under different models of beliefs, particularly when examining the effect of the last piece of advice (see Appendix B.6). If the result holds, it suggests that people may be more perceptive or critical of AI performance.

Finally, there is initially a small effect of advice correctness on weight on advice. Appendix B.5 finds that the effect is much larger after correcting for the participant's initial answer. Moreover, there is a satisfying explanation for how participants identify correct advice: the effect is only significant for participants in the condition that receives justifications. These results the theory that interpretability improves human adoption of algorithms and suggests that natural language reasoning is an effect medium for interpretability [53].

Table 3: Summary of weight on advice results.

Explanation		Analysis			Result	
Explanation type	Study metric	Prereg?	Spec.	Appendix	WOA Effect	Interaction?
Social distance	Advisor identity	✓	A	—	—	—
Task difficulty	Topic familiarity	✓	B	B.3	+	—
Algorithm familiarity	ChatGPT usage	✗	C	B.4	+	+ Advisor identity
Experience	Past accuracy	✗	E	B.6	+	+ Advisor identity
Interpretability	Justification	✓	A	B.5	—	+ Advice quality

Advice Confidence   While participants are indeed overconfident in their own answers, they err even worse in judging the correctness of AI advice. The calibration error is significant and persistent across 40 rounds of feedback.

Exploring one potential explanation, participants misjudge advisor accuracy across topics and underestimate accuracy on 10 out of 25 topics. These misjudgements contribute to calibration error and affect weight on advice. Participants broadly misunderstand which tasks the AI advisor is good at. Participants generally overestimate accuracy on procedural topics such as mathematics and physics, while underestimating accuracy on social science topics such history, government, geography, and so on.

Furthermore, by increasing the average weight of advice by over 50% and uniformly adjusting their answers, participants could have improved their score significantly. Participants don't use advice enough and are inefficient when they do,

providing an important caveat to the results of [90]: in order study, participants were not able to perform better than ChatGPT alone when they were unable to interact with the model (see Appendix B.3). Overall, students do not place enough trust in algorithms like ChatGPT.

Limitations   There are several limitations to this study. First, the study setup may be inefficient for the attempted regressions. Sampling from hundreds of questions introduces substantial variance to the data collection process that limits the efficiency of the estimators. Reducing randomization, for example by fixing a set of equally-difficult questions, may lead to better estimates. Otherwise, a study with a similar setup may require a larger sample size to identify the same effect.

Simultaneously, the lengthy survey required by the design might limit participant engagement. Appendix B.2 suggests the possibility that the effects are real, but only for participants willing to do optional questions.

The advice could have been improved. The generated advice may not be representative of "ChatGPT's output," as suggested to participants. Model output is quite sensitive to prompting format, particularly across topics (see Appendix A.2). The study design was constrained to using a similarly powerful InstructGPT model instead of OpenAI's ChatGPT API, which was released after collecting generations.

Moreover, the advisor identity could benefit from clarification. Previous work shows that algorithm aversion and appreciation effects are sensitive to the description of the advisor [32]. It may have been worthwhile to more explicitly clarify the identity of the human expert, although Appendix B.7 addresses this criticism.

The incentives could also be improved. By providing an additional payout to the top performers, the rewards could encourage excessive confidence for the sake of scoring higher on the leaderboard, a phenomenon documented in forecasting tournaments [91].

Finally, the study population limits the external validity of the results. These findings apply to a set of business students at UC Berkeley who might be much more informed on tests and familiar with ChatGPT compared to even the average student.

Future Work   Extensions of this study could begin by addressing its limitations. First, introducing more advisor identity conditions could help contrast how people judge ChatGPT's advice compared to other algorithms or other human advisors. For example, identifying an advisor as "a previous participant" or a generic "statistical model" [30] might cause participants to weight the advice less than that of an expert or ChatGPT.

The advice can be improved with better prompting techniques such as iterative bootstrapping [92] or self-consistency [93] for chain-of-thought. To further study mediators of weight on advice, the advice might include model output probabilities over answer choices [76] (with varying levels of calibration [94, 53]), enable live user interaction [90], or support modification of prompts [33]. In addition to using the ChatGPT API, a comparison of ChatGPT output and ChatGPT Plus outputs could illuminate the relative impact of different AI feedback.

Researchers could study a much wider variety of tasks. Aside from the other 30+ topics in the MMLU benchmark, participants could provide quantitative answers, answer free response problems, or complete other natural language tasks (e.g. from BIG-Bench [95]). Less structured responses would require new and untested metrics of distance between answers such BLEU scores [96]. Moreover, multi-modal models are able to perform natural language reasoning over images [97, 98] and videos [99, 100], creating new opportunities for study.

Further research might document how different populations take advice from LLM-based tools. For example, students from various grade levels or courses of study might take advice in different ways. Beyond education, many occupations will likely involve increasing collaboration with AI tools [101, 102]. Previous studies have documented differences in algorithm aversion across populations [48]. As LLMs influence a greater number of human decisions, a nuanced understanding of how different people take their advice will be increasingly important.

Conflicts of Interest

The authors declare that they have no conflict of interest.

Acknowledgments

led laboratory sessions. Members of the Moore Accuracy Lab tested and provided feedback on the survey. Critical support for running experiments was provided by the Research Participant Program and Experimental Social Science Laboratory. Funding for prizes was provided by the Michael and Chris Boskin Scholarship.

References

[1] Chaoning Zhang, Chenshuang Zhang, Chenghao Li, Yu Qiao, Sheng Zheng, Sumit Kumar Dam, Mengchun Zhang, Jung Uk Kim, Seong Tae Kim, Jinwoo Choi, et al. One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era. *arXiv preprint arXiv:2304.06488*, 2023.

[2] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*, 2023.

[3] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021, 2020.

[4] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[5] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[6] Simon Razniewski, Andrew Yates, Nora Kassner, and Gerhard Weikum. Language models as or for knowledge bases. *arXiv preprint arXiv:2110.04888*, 2021.

[7] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[8] OpenAI. Gpt-4 technical report. *arXiv:2303.08774*, 2023.

[9] Jonathan H Choi, Kristin E Hickman, Amy Monahan, and Daniel Schwarcz. Chatgpt goes to law school. *Available at SSRN*, 2023.

[10] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination? the implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*, 9(1):e45312, 2023.

[11] Amarachi B Mbakwe, Ismini Lourentzou, Leo Anthony Celi, Oren J Mechanic, and Alon Dagan. Chatgpt passing usmle shines a spotlight on the flaws of medical education, 2023.

[12] Nino Fijačko, Lucija Gosak, Gregor Štiglic, Christopher T Picard, and Matthew John Douma. Can chatgpt pass the life support exams without entering the american heart association course? *Resuscitation*, 185, 2023.

[13] Matthew W Kemp, Susan JS Logan, Pooja Sharma Dimri, Navkaran Singh, Citra NZ Mattar, Pradip Dashraath, Harshaana Ramlal, Aniza P Mahyuddin, Suren Kanayan, Sean WD Carter, et al. Chatgpt outscored human candidates in a virtual objective structured clinical examination (osce) in obstetrics and gynecology. *American Journal of Obstetrics and Gynecology*, 2023.

[14] Gerd Kortemeyer. Could an artificial-intelligence agent pass an introductory physics course? *arXiv preprint arXiv:2301.12127*, 2023.

[15] Ahmed Tlili, Boulus Shehata, Michael Agyemang Adarkwah, Aras Bozkurt, Daniel T Hickey, Ronghuai Huang, and Brighter Agyemang. What if the devil is my guardian angel: Chatgpt as a case study of using chatbots in education. *Smart Learning Environments*, 10(1):15, 2023.

[16] Teo Susnjak. Chatgpt: The end of online exam integrity? *arXiv preprint arXiv:2212.09292*, 2022.

[17] Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in Education and Teaching International*, pages 1–12, 2023.

[18] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI, 2023.

[19] Mehmet Firat. How chat gpt can transform autodidactic experiences and open education. *Department of Distance Education, Open Education Faculty, Anadolu Unive*, 2023.

[20] Christoph Leiter, Ran Zhang, Yanran Chen, Jonas Belouadi, Daniil Larionov, Vivian Fresen, and Steffen Eger. Chatgpt: A meta-analysis after 2.5 months. *arXiv preprint arXiv:2302.13795*, 2023.

[21] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[22] Michael Jay Polonsky and Jeffrey D Rotman. Should artificial intelligent agents be your co-author? arguments in favour, informed by chatgpt, 2023.

[23] Brent A Anders. Is using chatgpt cheating, plagiarism, both, neither, or forward thinking? *Patterns*, 4(3), 2023.

[24] Archana Praveen Kumar, Ashalatha Nayak, Manjula Shenoy, Shashank Goyal, et al. A novel approach to generate distractors for multiple choice questions. *Expert Systems with Applications*, page 120022, 2023.

[25] Philip Mark Newton. Chatgpt performance on mcq-based exams. 2023.

[26] Kyle A Burgason, Ophir Sefiha, and Lisa Briggs. Cheating is in the eye of the beholder: An evolving understanding of academic misconduct. *Innovative Higher Education*, 44:203–218, 2019.

[27] Chahna Gonsalves. On chatgpt: what promise remains for multiple choice assessment? *Journal of Learning Development in Higher Education*, (27), 2023.

[28] Anna-Carolina Haensch, Sarah Ball, Markus Herklotz, and Frauke Kreuter. Seeing chatgpt through students' eyes: An analysis of tiktok data. *arXiv preprint arXiv:2303.05349*, 2023.

[29] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

[30] Jennifer M Logg, Julia A Minson, and Don A Moore. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes*, 151:90–103, 2019.

[31] Noah Castelo, Maarten W Bos, and Donald R Lehmann. Task-dependent algorithm aversion. *Journal of Marketing Research*, 56(5):809–825, 2019.

[32] Yoyo Tsung-Yu Hou and Malte F Jung. Who is the expert? reconciling algorithm aversion and algorithm appreciation in ai-supported decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–25, 2021.

[33] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management science*, 64(3):1155–1170, 2018.

[34] Benedikt Berger, Martin Adam, Alexander Rühr, and Alexander Benlian. Watch me improve—algorithm aversion and demonstrating the ability to learn. *Business & Information Systems Engineering*, 63(1):55–68, 2021.

[35] Taly Reich, Alex Kaju, and Sam J Maglio. How to overcome algorithm aversion: Learning from mistakes. *Journal of Consumer Psychology*, 2022.

[36] Markus Jung and Mischa Seiter. Towards a better understanding on mitigating algorithm aversion in forecasting: An experimental study. *Journal of Management Control*, 32(4):495–516, 2021.

[37] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? a comprehensive literature review on algorithm aversion. 2020.

[38] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making*, 33(2):220–239, 2020.

[39] Carey K Morewedge. Preference for human, not algorithm aversion. *Trends in Cognitive Sciences*, 2022.

[40] Pascal Oliver Heßler, Jella Pfeiffer, and Sebastian Hafenbrädl. When self-humanization leads to algorithm aversion: what users want from decision support systems on prosocial microlending platforms. *Business & Information Systems Engineering*, 64(3):275–292, 2022.

[41] Xueming Luo, Siliang Tong, Zheng Fang, and Zhe Qu. Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science*, 38(6):937–947, 2019.

[42] Scott Schanke, Gordon Burtch, and Gautam Ray. Estimating the impact of "humanizing" customer service chatbots. *Information Systems Research*, 32(3):736–751, 2021.

[43] Megan Strait, Lara Vujovic, Victoria Floerke, Matthias Scheutz, and Heather Urry. Too much humanness for human-robot interaction: exposure to highly humanlike robots elicits aversive responding in observers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3593–3602, 2015.

[44] Swarnadeep Saha, Peter Hase, Nazneen Rajani, and Mohit Bansal. Are hard examples also harder to explain? a study with human and model-generated explanations. *arXiv preprint arXiv:2211.07517*, 2022.

[45] Francesca Gino and Don A Moore. Effects of task difficulty on use of advice. *Journal of Behavioral Decision Making*, 20(1):21–35, 2007.

[46] Eric Bogert, Aaron Schecter, and Richard T Watson. Humans rely more on algorithms than social influence as a task becomes more difficult. *Scientific reports*, 11(1):1–9, 2021.

[47] Esther Kaufmann. Algorithm appreciation or aversion? comparing in-service and pre-service teachers' acceptance of computerized expert models. *Computers and Education: Artificial Intelligence*, 2:100028, 2021.

[48] Nicole Tsz Yeung Liu, Samuel N. Kirshner, and Eric T.K. Lim. Is algorithm aversion weird? a cross-country comparison of individual-differences and algorithm aversion. *Journal of Retailing and Consumer Services*, 72:103259, 2023.

[49] Sangseok You, Cathy Liu Yang, and Xitong Li. Algorithmic versus human advice: Does presenting prediction performance matter for algorithm appreciation? *Journal of Management Information Systems*, 39(2):336–365, 2022.

[50] Ibrahim Filiz, Jan René Judek, Marco Lorenz, and Markus Spiwoks. Reducing algorithm aversion through experience. *Journal of Behavioral and Experimental Finance*, 31:100524, 2021.

[51] Francesca Cabiddu, Ludovica Moi, Gerardo Patriotta, and David G Allen. Why do users trust algorithms? a review and conceptualization of initial trust and trust over time. *European Management Journal*, 2022.

[52] Veronika Alexander, Collin Blinder, and Paul J Zak. Why trust an algorithm? performance, cognition, and neurophysiology. *Computers in Human Behavior*, 89:279–288, 2018.

[53] Richard Tomsett, Alun Preece, Dave Braines, Federico Cerutti, Supriyo Chakraborty, Mani Srivastava, Gavin Pearson, and Lance Kaplan. Rapid trust calibration through interpretable and uncertainty-aware ai. *Patterns*, 1(4):100049, 2020.

[54] Timothy DeStefano, Katherine Kellogg, Michael Menietti, and Luca Vendraminelli. Why providing humans with interpretable algorithms may, counterintuitively, lead to lower decision-making performance. 2022.

[55] Onur Altintas, Abraham Seidmann, and Bin Gu. The effect of interpretable artificial intelligence on repeated managerial decision-making under uncertainty. *Available at SSRN 4331145*, 2023.

[56] Daehwan Ahn, Abdullah Almaatouq, Monisha Gulabani, and Kartik Hosanagar. Will we trust what we don't understand? impact of model interpretability and outcome feedback on trust in ai. *arXiv preprint arXiv:2111.08222*, 2021.

[57] Daniel Ben David, Yehezkel S Resheff, and Talia Tron. Explainable ai and adoption of financial algorithmic advisors: an experimental study. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 390–400, 2021.

[58] Philipp Schmidt and Felix Biessmann. Calibrating human-ai collaboration: Impact of risk, ambiguity and transparency on algorithmic bias. In *Machine Learning and Knowledge Extraction: 4th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2020, Dublin, Ireland, August 25–28, 2020, Proceedings 4*, pages 431–449. Springer, 2020.

[59] Cedric A Lehmann, Christiane B Haubitz, Andreas Fügener, and Ulrich W Thonemann. The risk of algorithm transparency: How algorithm complexity drives the effects on the use of advice. *Production and Operations Management*, 31(9):3419–3434, 2022.

[60] Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias FC Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, et al. Non-task expert physicians benefit from correct explainable ai advice when reviewing x-rays. *Scientific reports*, 13(1):1383, 2023.

[61] Cecilia Panigutti, Andrea Beretta, Fosca Giannotti, and Dino Pedreschi. Understanding the impact of explanations on advice-taking: a user study for ai-based clinical decision support systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2022.

[62] Oscar Oviedo-Trespalacios, Amy E Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, Timothy Gallagher, et al. The risks of using chatgpt to obtain common safety-related information and advice. *Available at SSRN 4346827*, 2023.

[63] Alex Howard, William Hope, and Alessandro Gerada. Chatgpt and antimicrobial advice: the end of the consulting infection doctor? *The Lancet Infectious Diseases*, 23(4):405–406, 2023.

[64] Yi Xie, Ishith Seth, David J Hunter-Smith, Warren M Rozen, Richard Ross, and Matthew Lee. Aesthetic surgery advice and counseling from artificial intelligence: A rhinoplasty consultation with chatgpt. *Aesthetic Plastic Surgery*, pages 1–9, 2023.

[65] Anthony J Nastasi, Katherine R Courtright, Scott D Halpern, and Gary E Weissman. Does chatgpt provide appropriate and equitable medical advice?: A vignette-based, clinical evaluation across care contexts. *medRxiv*, pages 2023–02, 2023.

[66] A Shaji George and AS Hovan George. A review of chatgpt ai's impact on several business sectors. *Partners Universal International Innovation Journal*, 1(1):9–23, 2023.

[67] Margarita Leib, Nils Köbis, Rainer Michael Rilke, Marloes Hagens, and Bernd Irlenbusch. Corrupted by algorithms? how ai-generated and human-written advice shape (dis) honesty. *arXiv preprint arXiv:2301.01954*, 2023.

[68] Ali Momen, Ewart J de Visser, Kyle Wolsten, Katrina Cooley, James Walliser, and Chad C Tossell. Trusting the moral judgments of a robot: Perceived moral competence and humanlikeness of a gpt-3 enabled ai.

[69] Sebastian Krügel, Andreas Ostermaier, and Matthias Uhl. Chatgpt's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13(1):4569, 2023.

[70] Giovanni Spitale, Nikola Biller-Andorno, and Federico Germani. Ai model gpt-3 (dis) informs us better than humans. *arXiv preprint arXiv:2301.11924*, 2023.

[71] Yang Ye, Hengxu You, and Jing Du. Improved trust in human-robot collaboration with chatgpt. *arXiv preprint arXiv:2304.12529*, 2023.

[72] Robert Böhm, Moritz Jörling, Leonhard Reiter, and Christoph Fuchs. Content beats competence: People devalue chatgpt's perceived competence but not its recommendations. 2023.

[73] Don A Moore and Paul J Healy. The trouble with overconfidence. *Psychological review*, 115(2):502, 2008.

[74] Don A Moore and Daylian M Cain. Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, 103(2):197–213, 2007.

[75] Yangyi Chen, Lifan Yuan, Ganqu Cui, Zhiyuan Liu, and Heng Ji. A close look into the calibration of pre-trained language models. *arXiv preprint arXiv:2211.00151*, 2022.

[76] Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

[77] Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977, 2021.

[78] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[79] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of ai advice. *Computers in Human Behavior*, 127:107018, 2022.

[80] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[81] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[82] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

[83] Janet A Sniezek and Lyn M Van Swol. Trust, confidence, and expertise in a judge-advisor system. *Organizational behavior and human decision processes*, 84(2):288–307, 2001.

[84] Kaspar Rufibach. Use of brier score to assess binary predictions. *Journal of clinical epidemiology*, 63(8):938–939, 2010.

[85] Mark S Roulston. Performance targets and the brier score. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, 14(2):185–194, 2007.

[86] Taha Gunes et al. *Strategic and Adaptive Behaviours in Trust Systems*. PhD thesis, University of Southampton, 2021.

[87] Edgar Brunner and Ullrich Munzel. The nonparametric behrens-fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 42(1):17–25, 2000.

[88] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[89] Avishek Choudhury and Hamid Shamszare. Investigating the impact of user trust on adoption and use of chatgpt: A survey analysis. *Tellus*, 2023.

[90] Samuel R Bowman, Jeeyoon Hyun, Ethan Perez, Edwin Chen, Craig Pettit, Scott Heiner, Kamile Lukosuite, Amanda Askell, Andy Jones, Anna Chen, et al. Measuring progress on scalable oversight for large language models. *arXiv preprint arXiv:2211.03540*, 2022.

[91] Nuño Sempere and Alex Lawsen. Alignment problems with current forecasting platforms. *arXiv preprint arXiv:2106.11248*, 2021.

[92] Jiashuo Sun, Yi Luo, Yeyun Gong, Chen Lin, Yelong Shen, Jian Guo, and Nan Duan. Enhancing chain-of-thoughts prompting with iterative bootstrapping in large language models. *arXiv preprint arXiv:2304.11657*, 2023.

[93] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[94] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016, 2022.

[95] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

[96] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[97] Shalini Ghosh, Giedrius Burachas, Arijit Ray, and Avi Ziskind. Generating natural language explanations for visual question answering using scene graphs and visual attention. *arXiv preprint arXiv:1902.05715*, 2019.

[98] Leonard Salewski, A Sophia Koepke, Hendrik PA Lensch, and Zeynep Akata. Clevr-x: A visual reasoning dataset for natural language explanations. In *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*, pages 69–88. Springer, 2022.

[99] Aman Chadha and Vinija Jain. ireason: Multimodal commonsense reasoning using videos and natural language with interpretability. *arXiv preprint arXiv:2107.10300*, 2021.

[100] Ting Yu, Jun Yu, Zhou Yu, Qingming Huang, and Qi Tian. Long-term video question answering via multimodal hierarchical memory attentive networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(3):931–944, 2020.

[101] Ali Zarifhonarvar. Economics of chatgpt: A labor market view on the occupational impact of artificial intelligence. *Available at SSRN 4350925*, 2023.

[102] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*, 2023.

[103] Phoebe E Bailey, Tarren Leon, Natalie C Ebner, Ahmed A Moustafa, and Gabrielle Weidemann. A meta-analysis of the weight of advice in decision-making. *Current Psychology*, pages 1–26, 2022.

[104] Mirta Galesic and Michael Bosnjak. Effects of questionnaire length on participation and indicators of response quality in a web survey. *Public opinion quarterly*, 73(2):349–360, 2009.

A   Procedure Appendix

A.1   Dataset

Question selection   To select topics, the authors attempted questions within each topic and used intuition to downsize the benchmark to 25 topics. The `dev` and `train` splits are used to sample questions. All 5 questions in each `dev` set and up to 30 questions from the `train` set are included. A smaller set of questions is preferred in order to better capture question-specific random effects. The much larger `test` set is reserved for model evaluation and follow-up studies. For model evaluation, 50 questions are sampled uniformly from all splits for the topics.

Table 4: Question counts and descriptions [80] by topic.

Topic	Questions	Description
Clinical Knowledge	34	Spot diagnosis, joints, abdominal examination, ...
Conceptual Physics	31	Electromagnetism, thermodynamics, special relativity, ...
Elementary Mathematics	35	Word problems, multiplication, remainders, rounding, ...
Formal Logic	19	Propositions, predicate logic, first-order logic, ...
Global Facts	15	Extreme poverty, literacy rates, life expectancy, ...
High School Biology	35	Cellular structure, molecular biology, ecology, ...
High School Chemistry	27	Analytical, organic, inorganic, physical, ...
High School Computer Science	14	Algorithms, systems, graphs, recursion, ...
High School European History	23	Renaissance, reformation, industrialization, ...
High School Geography	27	Population migration, rural land-use, urban processes, ...
High School Government and Politics	26	Branches of government, civil liberties, political ideologies, ...
High School Macroeconomics	35	Economic indicators, national income, international trade, ...
High School Microeconomics	31	Supply and demand, imperfect competition, market failure, ...
High School Physics	22	Kinematics, energy, torque, fluid pressure, ...
High School Psychology	35	Behavior, personality, emotions, learning, ...
High School Statistics	28	Random variables, sampling distributions, chi-square tests, ...
High School U.S. History	27	Civil War, the Great Depression, The Great Society, ...
High School World History	31	Ottoman empire, economic imperialism, World War I, ...
Human Aging	28	Senescence, dementia, longevity, personality changes, ...
Human Sexuality	17	Pregnancy, sexual differentiation, sexual orientation, ...
Miscellaneous Topics	35	Agriculture, Fermi estimation, pop culture, ...
Nutrition	35	Metabolism, water-soluble vitamins, diabetes, ...
Philosophy	35	Skepticism, phronesis, skepticism, Singer's Drowning Child, ..
Sociology	27	Socialization, cities and community, inequality and wealth, ...
U.S. Foreign Policy	16	Soft power, Cold War foreign policy, isolationism, ...

Question order   The dataset is reordered twice during survey administration to overcome a limitation in Qualtrics that presents questions in increasing order of index. In the first reordering, the first and second halves of the questions are switched. The second reordering, the questions are shuffled entirely. As a result, the distribution of question appearances is roughly normal with no apparent bias (Figure 9). The procedure for shuffling questions is documented in the accompanying repository.

Figure 9: **Question distribution.** Every question is presented at least once. The modal number of appearances is 4 with a maximum of 8. The distribution is consistent with the result of a uniform selection process.

A.2 Model Evaluation

All model evaluations use the Completions API with `engine=''text-davinci-003''`, `temperature=0`, and control the completion length. The CoT prompt (Figure 10) is compared against the standard prompt (Figure 11). In a reversal of the CoT approach, the model is first forced to output a response and then provide a justification.

Figure 10: **Chain-of-thought prompt.** The model is prompted to think "step-by-step" and provide a reasoning. The final answer is treated as the model's answer.

Figure 11: **Standard prompt.** The model is first prompted to immediately answer the question. Then it is prompted to provide a justification.

Both prompts achieved an accuracy of $62.9\% \pm 3.4\%$ but their responses are only moderately correlated (Spearman's r=0.573, p<1e-68). Prompt accuracy differs significantly across topics (Figure 12a) and categories (Figure 12b). Generally, CoT achieves improvements on procedural topics (Conceptual Physics, Elementary Mathematics) at the expense of performance losses on fact-based humanities topics (history topics, Philosophy).

(a) Accuracy by topic.

(b) Accuracy by category.

Figure 12: **Accuracy by topic and category.** Error bars are 95% confidence intervals. Topics and categories where prompts diverged significantly (at a 90% level) are **bolded**.

The CoT prompt is preferred due to the quality of justifications. Whereas the standard prompt often leads models to adopt awkward justifications, CoT generates a coherent stream of reasoning that resolves to an answer. One limitation of this choice is the possibility of invalid outputs. For five questions (< 1% of the dataset), the model justification rejects all of the answer choices and answers `"None"` or `"E"`. In the evaluation, these answers are as treated as incorrect with a weight on advice of 0. These "errors" are preserved in order to approximate a real-world analog in which an AI chatbot might similarly reject all options.

A.3 Data Processing

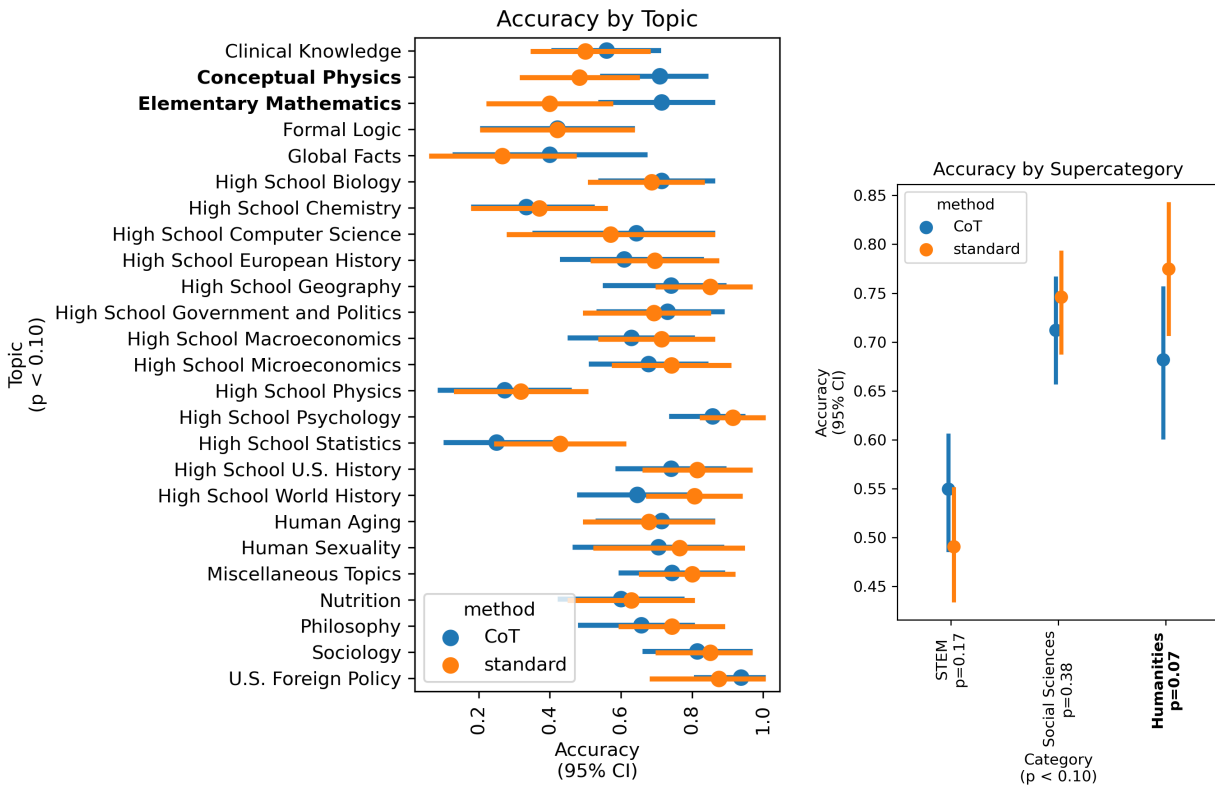**Weight on advice**   Most studies [103] compute WoA as the the difference between the initial and adjusted judgment divided by the difference between the initial judgment and the advice. The advice is assumed to correspond to 100% belief in the advised answer. In the context of probabilities, this approach might conflict with intuitions about scale in the context of probability sums. If a participant is initially 99% confident in the advisor's answer, then an adjustment to 100% is not a large change. Yet whereas the absolute change is small ($WoA = 0.01$), proportionate change is the largest possible value ($WoA = 1$). This property can lead to noisy results for WoA when many participants opt to "round" high confidence levels (80-99%) up to full confidence (100%).

Topic familiarity   Participants are asked "How comfortable do you feel with the following topic areas?" and provided the options "Uncomfortable", "Neutral", and "Comfortable." These are mapped to -1, 0, and 1 respectively in summary statistics and use indicators in the regression, as preregistered. to familiarity areas using the mapping in Table 5. These areas are created to roughly capture the skills and knowledge required in each topic area as an alternative to tedious topic-by-topic ratings of comfort level.

Table 5: Linking topics to familiarity areas.

Topic/Task	Questions	Familiarity Area
Clinical Knowledge	34	Biological Sciences
Conceptual Physics	31	Physics
Elementary Mathematics	35	Mathematics
Formal Logic	19	Mathematics
Global Facts	15	Trivia
High School Biology	35	Biological Sciences
High School Chemistry	27	Biological Sciences
High School Computer Science	14	Computer Science
High School European History	23	History
High School Geography	27	Trivia
High School Government and Politics	26	Economics
High School Macroeconomics	35	Economics
High School Microeconomics	31	Economics
High School Physics	22	Physics
High School Psychology	35	Biological Sciences
High School Statistics	28	Mathematics
High School U.S. History	27	History
High School World History	31	History
Human Aging	28	Biological Sciences
Human Sexuality	17	Biological Sciences
Miscellaneous Topics	35	Trivia
Nutrition	35	Biological Sciences
Philosophy	35	Literature
Sociology	27	Literature
U.S. Foreign Policy	16	History

Past usage   Participants answer four progressive questions about usage of AI chatbots. Each question is only displayed if they respond "yes" to the previous:

1. Have you heard of AI chatbots before?

2. Have you used AI chatbots before?

3. Have you used AI chatbots in a classroom setting before? (homework, quiz, studying, etc.)

4. Have you used AI chatbots to answer multiple choice questions before?

The number of questions to which they respond "yes", which may vary from 0 to 4, is summed and used to measure past usage.

Experience   Participant beliefs about the accuracy of advice are modeled with a Beta-Bernoulli. A fairly weak prior of $\text{Beta}(\alpha = 0.5, \beta = 0.5)$ is chosen. Upon receiving feedback on a question, it is assumed that participants update to $\text{Beta}(\alpha + 1, \beta)$ if the advice is correct and $\text{Beta}(\alpha, \beta + 1)$ if the advice is wrong. This is a fairly strong assumption since it assumes that participants can accurately track and evenly weight past performance.

B   Analysis Appendix

B.1   Descriptive Statistics

Summary statistics   Summary statistics are displayed in Table 6.

Table 6: Summary statistics for key variables.

	weight_on_advice	init_advice_confidence	advice_confidence	advice_is_correct
count	2828.000	2828.000	2828.000	2828.000
mean	0.337	0.359	0.587	0.639
std	0.395	0.309	0.357	0.480
min	0.000	0.000	0.000	0.000
25%	0.000	0.156	0.278	0.000
50%	0.149	0.250	0.571	1.000
75%	0.672	0.500	1.000	1.000
max	1.000	1.000	1.000	1.000

	net_familiarity	uncomfortable	neutral	comfortable
count	2828.000	2828.000	2828.000	2828.000
mean	0.056	0.269	0.405	0.326
std	0.770	0.444	0.491	0.469
min	-1.000	0.000	0.000	0.000
25%	-1.000	0.000	0.000	0.000
50%	0.000	0.000	0.000	0.000
75%	1.000	1.000	1.000	1.000
max	1.000	1.000	1.000	1.000

	usage_level	heard_of	used	used_in_class	answered_mc
count	2828.000	2828.000	2828.000	2828.000	2828.000
mean	2.678	0.986	0.832	0.595	0.264
std	1.068	0.120	0.374	0.491	0.441
min	0.000	0.000	0.000	0.000	0.000
25%	2.000	1.000	1.000	0.000	0.000
50%	3.000	1.000	1.000	1.000	0.000
75%	4.000	1.000	1.000	1.000	1.000
max	4.000	1.000	1.000	1.000	1.000

	question_num	correct_advice_count	incorrect_advice_count	init_time	adjusted_time
count	2828.000	2828.000	2828.000	2828.000	2828.000
mean	13.099	7.826	4.273	28.292	10.728
std	8.151	5.482	3.322	20.390	8.608
min	1.000	0.000	0.000	5.458	5.115
25%	6.000	3.000	2.000	13.497	6.126
50%	12.000	7.000	4.000	22.524	7.536
75%	18.000	11.000	6.000	36.145	11.388
max	40.000	27.000	18.000	90.141	90.038

Conditions   Participants ($n = 188$) were assigned randomly to a 2x2 condition (see Figure 7). Notably, significantly more participants (66) were assigned to receive a justification than not (52). The ability to evenly assign participants was limited by the survey platform and administration method, which led to many "in-progress" surveys that were not completed.

Table 7: Conditions assignments.

| Advisor | Justification | |
	Yes	No
AI chatbot	32	24
expert	34	28

Weight on advice   Average WoA was 0.237 with a standard deviation of 0.300 (Figure 13). In the sample, 40.45% of answers placed no weight on advice (WoA ≤ 0).

Figure 13: **Distribution of weight on advice.** The distribution is bimodal at 0 and 1. Weight on advice is depicted post-winsorization. Notice the absence of values between 80% and 95%, suggesting a tendency to "round up" high confidence to full confidence.

Topic familiarity   Familiarity varies significantly across topics (Figure 14). Participants are most familiar with economics and mathematics, which is sensible for a business degree program. Participants rated themselves least familiar with computer science, for which there is also the most variation. This seems to be due to the substantial number of participants majoring in Computer Science ($n = 10$), EECS ($n = 9$), or data science ($n = 19$).

Figure 14: **Familiarity in topic areas.** Net familiarity is annotated above each topic label. Error bars are 95% confidence intervals.

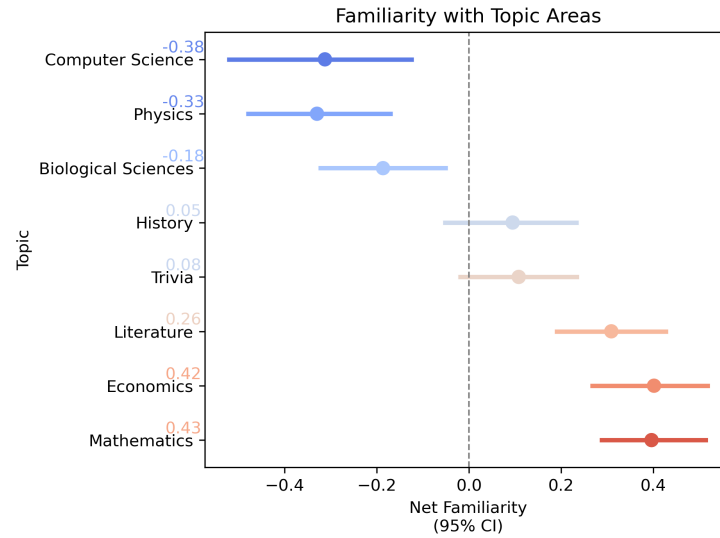**Questions answered.** Most participants (92/118) answered less than 25 questions. There is a significant dropoff after around questions 20-22 once participants notice that they have the option to opt-out (Figure 15a.

Weight on advice appears to decline throughout the experiment (Figure 15b). Conducting an type-2 ANOVA analysis of regressing weight on advice question groups yields a slightly significant result (p=0.0731). In the section 4, that question number is a significant predictor of weight on advice. Note that because answering questions 21-40 is voluntary, the effect might be confounded by participant attributes (e.g. diligence or competitiveness).

(a) Number of questions answered.

(b) Weight on advice by question group.

Figure 15: **Questions answered.** Error bars are 95% confidence intervals. Error bars widen past 20 questions because fewer participants voluntarily answered additional questions.

All but one participant heard of ChatGPT and other AI chatbots. Among participants in the AI chatbot condition, participants appear to place much greater weight on advice if they have used ChatGPT, and specifically if they have used it to answer multiple choice questions. Note that Figure X is plotted without accounting for random effects so the standard errors may not be reflective.

Past usage   All but one participant heard of ChatGPT and other AI chatbots. Among participants in the AI chatbot condition, participants appear to place much greater weight on advice if they have used ChatGPT, and specifically if they have used it to answer multiple choice questions. Note that Figure 16a is plotted without accounting for random effects so the standard errors may not be reflective.

These effects are much smaller but directionally similar in the expert condition (Figure 16b).

(a) AI chatbot condition.   (b) Expert condition.

Figure 16: **Weight on advice by usage level.** Error bars are 95% confidence intervals.

Advice correctness   The advice is correct for 63.9% of samples. Participants are somewhat able to discern the correctness of advice (Figure 17). In general, judgements about advice correctness are bimodal near 0 and 1. There is a noticeable absence of beliefs in the 60% to 90% confidence range, suggesting that when participants are "pretty sure" that the advice is correct, they round up to 100%.

Figure 17: **Advice correctness beliefs.** Participants are apparently capable of discerning correct and incorrect advice. When advice is correct, a greater proportion of participants judge the advised answer as very likely to be correct, and vice versa.

Experiences   Beliefs about advice accuracy are roughly normally distributed with a mode at uninformative prior, 0.5 (Figure 18. The mean belief in accuracy is 60.5% with a standard deviation of 13.1%.

Figure 18: **Model advice accuracy beliefs** Beliefs have much greater spread in the first 10 questions as expected due to the greater probability of receiving mostly correct or incorrect answers. After, beliefs tend to be more centered around the true accuracy of 64.9%.

B.2   Optionality
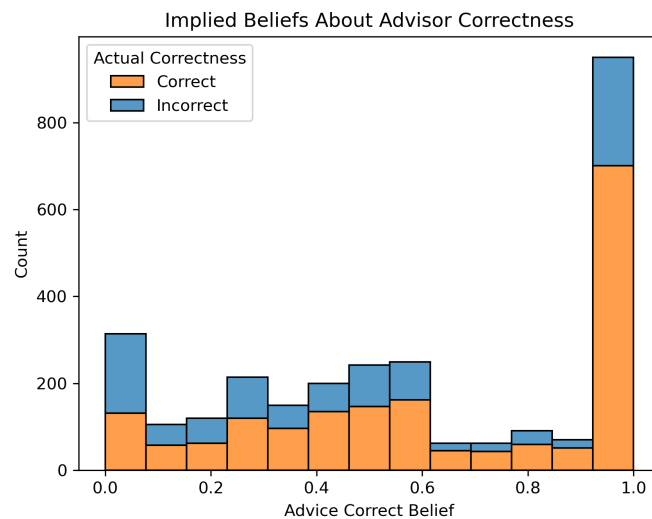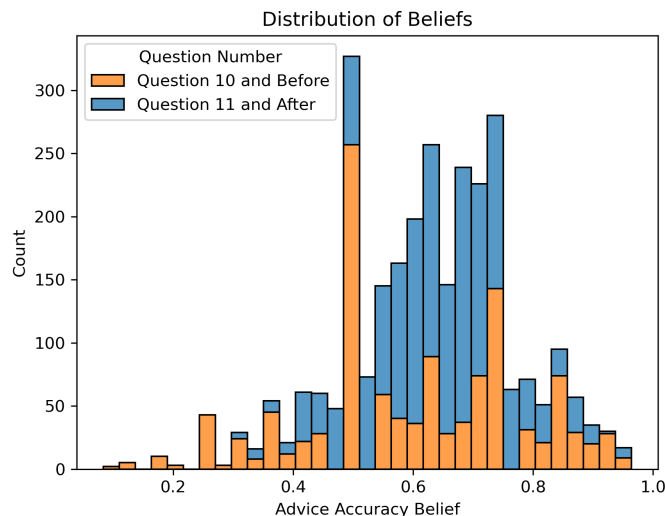
The survey is fairly long, taking most particpants at least 20 minutes. Longer questionnaires are known to suffer from reduced response quality, particularly for later questions [104]. This may, for example, explain and confound the significant decline in weight on advice over time we find in Table 2. But one unique feature of our survey is the opportunity to complete *optional* questions for additional points. If a participant voluntarily answers additional questions, it might suggest greater engagement with the survey and improved response quality.

The regressions in Table 8 consider whether the optionality of a question interacts with the interventions. In regression **A2**, we regress on both the question number and on `is\_optional`, whether the question number is greater than 20. Interestingly, the interaction terms recover the hypothesized effects. For optional questions (compared to non-optional questions):

- Advice with no justification is given 14.0% 95 CI[1.0%, 26.9%] greater weight if it comes from an advisor.

- When justifications are provided, advice from the AI advisor is given an insignificant 2.1% greater weight on advice.

These effects largely persist after adding all of the other controls in **A3**. Taken together, this exploratory analysis suggests that engagement mediates the effects of interest.

Table 8: Results of additional regressions on question optionality.

	A	A2	A3
Intercept	0.329***	0.323***	-0.174*
	(0.044)	(0.044)	(0.096)
advice_accuracy_belief			0.619***
			(0.089)
usage_level			0.046*
			(0.025)
topic_familiarity[T.Uncomfortable]			0.065***
			(0.020)
topic_familiarity[T.Neutral]			0.027
			(0.017)
question_num			-0.005***
			(0.001)
question_id Var	0.044**	0.046**	0.048***
	(0.018)	(0.018)	(0.018)
participant_id Var	0.364***	0.367***	0.340***
	(0.056)	(0.056)	(0.053)
is_optional[T.True]		0.080	0.166***
		(0.056)	(0.057)
give_justification[T.yes]:is_optional[T.True]		-0.147**	-0.134**
		(0.067)	(0.066)
give_justification[T.yes]	0.057	0.072	0.084
	(0.058)	(0.058)	(0.056)
advisor[T.expert]:is_optional[T.True]		-0.140**	-0.140**
		(0.066)	(0.065)
advisor[T.expert]:give_justification[T.yes]:is_optional[T.True]		0.161*	0.115
		(0.083)	(0.082)
advisor[T.expert]:give_justification[T.yes]	-0.027	-0.046	-0.029
	(0.079)	(0.080)	(0.077)
advisor[T.expert]	-0.027	-0.011	0.080
	(0.059)	(0.060)	(0.138)
advice_is_correct			0.030**
			(0.015)
advice_accuracy_belief:advisor[T.expert]			-0.101
			(0.120)
usage_level:advisor[T.expert]			-0.020
			(0.035)

B.3 Topic Familiarity

Topic familiarity tangibly affects participant performance. Participants are assumed to pick their highest confidence answer and randomly pick when there is a tie. Across all questions, participants achieve an accuracy of 42.1% before advice that improves to 57% after receiving advice. Both significantly underperform GPT's 63.9% accuracy. When divided by topic familiarity, participant initial accuracy on uncomfortable topics (37.7%) and comfortable topics (46.4%) is significantly worse and better than the combined baseline, respectively (Figure 19). These results suggest that participant's subjective judgements of their familiarity are predictive.
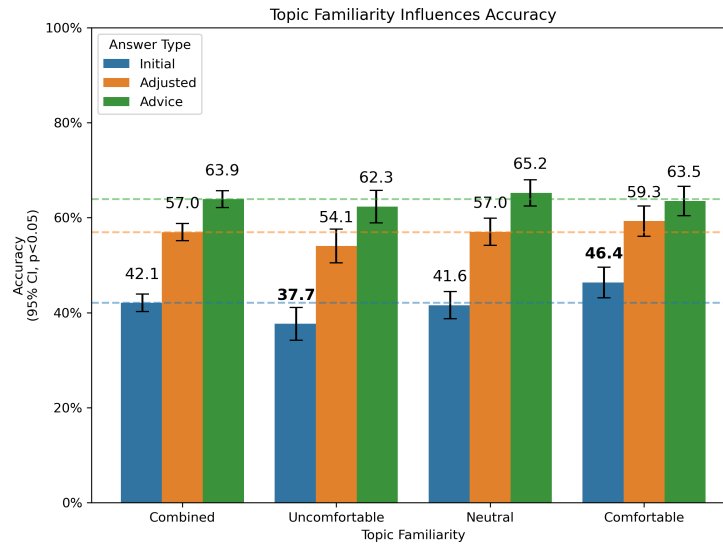
Figure 19: **Effect of topic familiarity on accuracy.** Dotted line corresponds to combined accuracy. **Bolded** values are significantly different at a 95% significance level than the combined accuracy in a t-test.

Two additional regressions explore the effect of topic familiarity (Table 9). specification **B2** tests whether the effect of topic familiarity is captured by differences in weight on advice for each topic. The effect is attenuated but continues to be significant. Specification **B3** further examines whether being uncomfortable with a topic causes greater weight on AI chatbot advice. Topic familiarity drives greater weight in advice more in the AI chatbot condition than in the human expert condition, but not significantly so.

Table 9: Results of additional regressions on topic familiarity.

	B	B2	B3
Intercept	0.302***	0.311***	0.299***
	(0.045)	(0.052)	(0.048)
advisor[T.expert]	-0.028	-0.028	-0.012
	(0.059)	(0.066)	(0.064)
advisor[T.expert]:give_justification[T.yes]	-0.021	-0.024	-0.025
	(0.079)	(0.088)	(0.080)
give_justification[T.yes]	0.056	0.060	0.058
	(0.058)	(0.064)	(0.058)
participant_id Var	0.362***	0.479***	0.367***
	(0.055)	(0.091)	(0.056)
question_id Var	0.040**	0.055***	0.026
	(0.018)	(0.021)	(0.017)
topic Var		0.051	0.014**
		(nan)	(0.007)
topic_familiarity[T.Neutral]	0.026	0.020	0.028
	(0.018)	(0.018)	(0.026)
topic_familiarity[T.Neutral]:advisor[T.expert]			-0.010
			(0.035)
topic_familiarity[T.Uncomfortable]	0.061***	0.041*	0.069**
	(0.020)	(0.021)	(0.029)
topic_familiarity[T.Uncomfortable]:advisor[T.expert]			-0.040
			(0.040)

B.4 Past Usage

Several additional regressions test the strength of the effect of past usage (Table 10). Specification **C2** removes the interaction term with the advisor. The effect is still significant but is diminished in size. In specification **C3**, indicators are added for each usage level. While the coefficients are directionally sensible, none are significant. Specification **C4** tests for the effect of having used AI chatbots before (`used`). There is a large and significant effect on weight on advice, suggesting that using chatbots before leads participants in the AI chatbot condition to place 0.149 greater weight on advice.

Table 10: Results of additional regressions on past usage.

	C	C2	C3	C4
C(usage_level)[T.1]			0.117	
			(0.160)	
C(usage_level)[T.2]			0.230	
			(0.157)	
C(usage_level)[T.3]			0.189	
			(0.156)	
C(usage_level)[T.4]			0.241	
			(0.156)	
Intercept	0.169**	0.211***	0.104	0.183***
	(0.081)	(0.066)	(0.159)	(0.071)
advisor[T.expert]	0.056	-0.036	-0.036	0.036
	(0.118)	(0.059)	(0.060)	(0.107)
advisor[T.expert]:give_justification[T.yes]	-0.025	-0.019	-0.020	-0.020
	(0.079)	(0.078)	(0.079)	(0.078)
give_justification[T.yes]	0.064	0.061	0.062	0.057
	(0.057)	(0.057)	(0.057)	(0.057)
participant_id Var	0.354***	0.353***	0.354***	0.350***
	(0.055)	(0.054)	(0.055)	(0.054)
question_id Var	0.040**	0.040**	0.040**	0.040**
	(0.018)	(0.018)	(0.018)	(0.018)
topic_familiarity[T.Neutral]	0.027	0.027	0.026	0.026
	(0.018)	(0.018)	(0.018)	(0.018)
topic_familiarity[T.Uncomfortable]	0.062***	0.062***	0.061***	0.063***
	(0.020)	(0.020)	(0.020)	(0.020)
usage_level	0.050*	0.034*		
	(0.025)	(0.018)		
usage_level:advisor[T.expert]	-0.033			
	(0.036)			
used[T.True]				0.149**
				(0.069)
used[T.True]:advisor[T.expert]				-0.088
				(0.105)

B.5 Advice Quality

The effect of advice quality/correctness may masked by misplaced trust in the advice answer. Suppose participants are modest and tend to put between 80% confidence in the advisor's answer regardless of their initial answer. Further suppose that participants are capable and tend to pick the same answer as the advisor. When the advice is correct, they are likely to place greater weight on the advice answer initially and under-adjust proportionately. When the advice is incorrect, they are likely to move more in the direction of the advice answer. Perversely, this would suggest that advice correctness is negatively correlated with weight on advice.

This is controlled for by directly regressing on initial confidence in the advice answer (`init_advice_confidence`) in a series of regressions (Table 11). Compared to the original specification **D**, specification **D2** adds a term that controls for initial advice confidence. Question random effects are no longer significant after controlling for initial confidence, suggesting that initial confidence indeed controls for question difficulty. Next, adding a quadratic term is justified, as

the coefficients in D3 remain significant. An interaction term with giving justifications in D4 suggests that the effect of advice correctness depends on giving justifications.

These results suggest a natural explanation in which correct advice is more convincing because the justifications are more coherent. If true, one would expect people to spend more time reading when there is a justification. Indeed, participants that receive justifications spend 2.5 95% CI[1.24%, 3.76%] more seconds on adjusting answers, 'adjusted_time'. Specification D5 regresses on 'adjusted_time' with an interaction term with 'advice_is_correct'. Consistent with the hypothesis, the interaction of advice being correct and time spent incorporating advice explains most of the effect. Every additional 10 seconds spent incorporating advice is associated with a 3.95% 95% CI[1.16%, 6.74%] increase in weight on advice.

Table 11: Results of additional regressions on advice correctness.

	D	D2	D3	D4	D5
I(init_advice_confidence ** 2)			-1.260***	-1.261***	-1.250***
			(0.071)	(0.071)	(0.071)
Intercept	0.161**	0.234***	0.098	0.118	0.134
	(0.082)	(0.084)	(0.094)	(0.092)	(0.092)
adjusted_time					-0.002
					(0.001)
advice_is_correct	0.013	0.053***	0.047***	0.018	-0.020
	(0.015)	(0.014)	(0.013)	(0.019)	(0.024)
advice_is_correct:adjusted_time					0.004***
					(0.001)
advice_is_correct:give_justification[T.yes]				0.051**	0.041
				(0.025)	(0.025)
advisor[T.expert]	0.054	0.051	0.044	0.044	0.043
	(0.118)	(0.120)	(0.136)	(0.131)	(0.131)
advisor[T.expert]:give_justification[T.yes]	-0.025	-0.031	-0.058	-0.059	-0.058
	(0.079)	(0.080)	(0.091)	(0.088)	(0.088)
give_justification[T.yes]	0.064	0.065	0.071	0.039	0.043
	(0.057)	(0.058)	(0.066)	(0.066)	(0.066)
init_advice_confidence		-0.274***	0.984***	0.984***	0.981***
		(0.022)	(0.074)	(0.074)	(0.074)
participant_id Var	0.353***	0.389***	0.577***	0.528***	0.531***
	(0.055)	(0.060)	(0.093)	(0.080)	(0.080)
question_id Var	0.040**	0.027	0.021	0.005	0.007
	(0.018)	(0.017)	(0.018)	(0.016)	(0.016)
topic_familiarity[T.Neutral]	0.027	0.025	0.012	0.011	0.010
	(0.018)	(0.017)	(0.016)	(0.016)	(0.016)
topic_familiarity[T.Uncomfortable]	0.062***	0.050**	0.019	0.019	0.019
	(0.020)	(0.020)	(0.019)	(0.019)	(0.019)
usage_level	0.049*	0.051*	0.045	0.045	0.045
	(0.025)	(0.026)	(0.029)	(0.028)	(0.028)
usage_level:advisor[T.expert]	-0.033	-0.031	-0.025	-0.025	-0.025
	(0.036)	(0.037)	(0.042)	(0.041)	(0.041)

B.6 Experience

Several additional regressions are conducted to assess how sensitive the finding is to the model of participant beliefs (Table 12). The original specification E uses a uninformative prior Beta($\alpha = 0.5, \beta = 0.5$) to model experiences. Specification E2 tests a term that indicates whether the last piece of advice given is correct (`last_advice_is_correct`). The advice accuracy belief term is reintroduced in E3. This specification tests whether recency bias leads participants to overweight recent advice. Specifications E4-E6 vary the strength of the uninformative prior. Specifications E7 and E8 modify the prior's estimate to correspond to accuracy beliefs of 25% and 75%, respectively.

Table 12: Specifications for additional regressions on participant beliefs.

	Specification							
	E	**E2**	**E3**	**E4**	**E5**	**E6**	**E7**	**E8**
α prior	0.5	—	0.5	0.05	0.2	1	0.25	0.75
β prior	0.5	—	0.5	0.05	0.2	1	0.75	0.25
Last question?	✗	✓	✓	✗	✗	✗	✗	✗
Beliefs?	✓	✗	✓	✓	✓	✓	✓	✓

The results of the regressions are displayed in Table 13. If the last advice given was correct, participants place 8.7% 95 CI[4.8%, 12.4%] greater weight on advice in the AI advisor condition, but only 4.2% 95 CI[0.6%, 7.8%] in the human expert advisor condition. These results are subsumed after reintroducing the beliefs term. The results are robust to last advice correctness, prior strength, and prior estimate. Regardless of the prior, the coefficient on beliefs is lower for the human expert condition, corroborating the previous finding.

Table 13: Specifications for additional regressions on participant beliefs.

	E	E2	E3	E4	E5	E6	E7	E8
Intercept	-0.159*	0.108	-0.174*	-0.097	-0.139	-0.252**	-0.050	-0.247**
	(0.096)	(0.082)	(0.097)	(0.089)	(0.092)	(0.101)	(0.090)	(0.098)
advice_accuracy_belief	0.602***		0.502***	0.405***	0.475***	0.674***	0.361***	0.616***
	(0.088)		(0.099)	(0.067)	(0.075)	(0.104)	(0.073)	(0.089)
advice_accuracy_belief:advisor[T.expert]	-0.097		-0.043	-0.047	-0.077	-0.201	-0.146	-0.033
	(0.120)		(0.134)	(0.091)	(0.102)	(0.141)	(0.099)	(0.121)
advice_is_correct[T.True]	0.029**	0.015	0.029**	0.028*	0.029**	0.030**	0.023	0.032**
	(0.014)	(0.014)	(0.015)	(0.015)	(0.015)	(0.015)	(0.015)	(0.014)
advisor[T.expert]	0.067	0.079	0.063	0.041	0.058	0.136	0.116	0.022
	(0.137)	(0.118)	(0.139)	(0.129)	(0.132)	(0.144)	(0.129)	(0.141)
advisor[T.expert]:give_justification[T.yes]	-0.023	-0.026	-0.019	-0.017	-0.017	-0.019	-0.022	-0.016
	(0.076)	(0.078)	(0.076)	(0.076)	(0.076)	(0.076)	(0.076)	(0.076)
give_justification[T.yes]	0.074	0.067	0.070	0.067	0.068	0.070	0.067	0.070
	(0.055)	(0.057)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)	(0.055)
last_advice_is_correct[T.True]		0.084***	0.035					
		(0.020)	(0.022)					
last_advice_is_correct[T.True]:advisor[T.expert]		-0.046*	-0.042					
		(0.027)	(0.030)					
participant_id Var	0.337***	0.348***	0.333***	0.334***	0.333***	0.331***	0.332***	0.339***
	(0.052)	(0.054)	(0.052)	(0.052)	(0.052)	(0.052)	(0.052)	(0.053)
question_id Var	0.045**	0.042**	0.043**	0.044**	0.043**	0.042**	0.042**	0.043**
	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)	(0.018)
question_num	-0.004***							
	(0.001)							
topic_familiarity[T.Neutral]	0.026	0.029*	0.029	0.028	0.028	0.028	0.029*	0.025
	(0.017)	(0.018)	(0.017)	(0.017)	(0.017)	(0.017)	(0.018)	(0.017)
topic_familiarity[T.Uncomfortable]	0.063***	0.062***	0.063***	0.064***	0.063***	0.064***	0.063***	0.063***
	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)	(0.020)
usage_level	0.045*	0.049*	0.047*	0.047*	0.047*	0.046*	0.048*	0.046*
	(0.024)	(0.025)	(0.024)	(0.024)	(0.024)	(0.024)	(0.025)	(0.025)
usage_level:advisor[T.expert]	-0.019	-0.031	-0.022	-0.022	-0.022	-0.023	-0.027	-0.019
	(0.035)	(0.036)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)	(0.035)

B.7 Manipulation Efficacy

One plausible criticism of the results is that participants may not have remembered the manipulation. The survey design attempts to mitigate this in two ways. First, participants must pass a manipulation check that confirms the identity of the advisor. Second, each feedback page explicitly repeats the advisor identity (Figure 20).

The AI chatbot answered:

(B) identifying and eliminating the causes of the consultee's difficulties in handling a problem

Now you have the chance to update your previous answer.
Question 2. According to Caplan`s model of consultee-centered case consultation, the consultant is primarily interested in

| 0 | 10 | 20 | 30 | 40 | 50 | 60 | 70 | 80 | 90 | 100 |

(A) identifying the causes and solutions of the client`s presenting problems

Figure 20: **Advice format.** The screenshot is taken from a Qualtrics survey and reflects what participants would've seen.

There is additional anecdotal evidence that the manipulation has the desired effect. One participant in the expert condition was confused about the manipulation. They emailed to say that they "*personally assumed that a different expert was asked for each question for their guess at the correct answer (for example, professors from different departments from Berkeley [...])*", suggesting that "*the respondent will typically assume (at least that was the case for me), that the expert is human.*" Other anecdotal comments confirm this assumption.

B.8 Persistence of Miscalibration

The miscalibration for advised choices does not change significantly over time. Figure 21 suggests a small decrease in $\text{ECE}_{\text{advised}} = 0.201$ from questions 6-10 to questions 11-15. For the optional questions, participants are more miscalibrated, but the standard errors are much larger. None of these effects are significant at a 95% level.
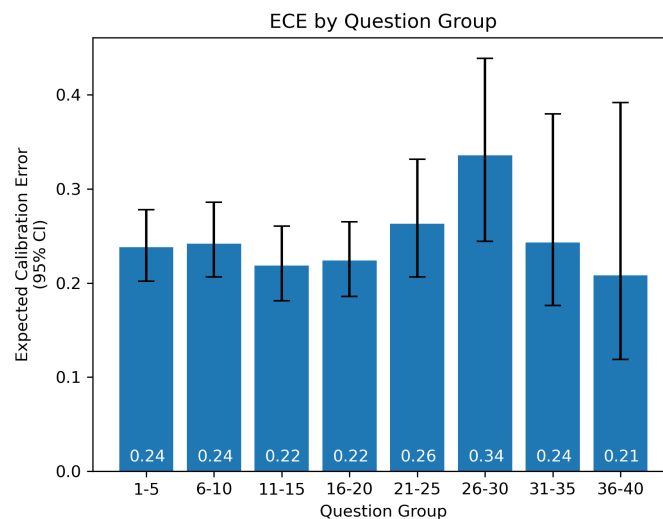

Figure 21: Expected calibration error over questions.

B.9 Sources of Inefficiency

Several habits of participants worsen their actual average score ($\overline{\text{BS}} = 0.674$) compared to a uniform baseline ($\text{BS} \approx 0.588$) for the same WoA. There are two principal mistakes.

Misallocation Misallocation is the tendency to poorly adjust other answers after receiving advice. After extending additional weight to one choice, participants often rescale the other answers in inefficient ways. For example, participants might assign zero weight to an option they initially assigned a little weight. Consider an alternative in which participants proportionately decrease their confidence in each of the other choices. Compared to the actual average Brier score $\overline{\text{BS}} = 0.674$, participants would have earned $\hat{\text{BS}} = 0.634$, a modest improvement.

Extremism Extremism is the tendency for participants put both too much and too little weight on advice. What if—holding the proportion of weight on the other adjusted answers the same—participants moved more consistently in the direction of the advice? This behavior is parameterized by a shrinking parameter s which proportionately "shrinks" each question's weight on advice closer to the sample average, $\overline{\text{WoA}}$. Specifically, for each question i:

$$\text{WoA} \leftarrow \text{WoA}_i + (1 - s)\overline{\text{WoA}}$$

Otherwise, the relative ratio of the other adjusted choices is preserved. When originally there is full confidence in the advised answer, confidence is uniformly allocated over answers. The optimal shrinking parameter is $s^*=0.62$, achieving a score superior to uniform allocation at $\tilde{\text{BS}}=0.578$. The relationship between scaling and Brier Score is quadratic and displayed in Figure 22.
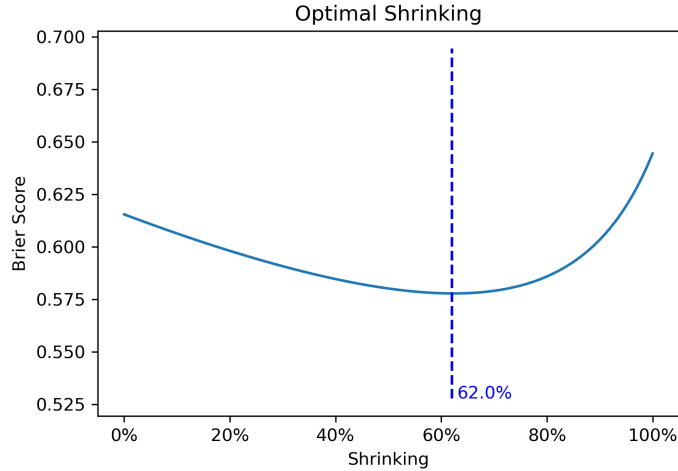

Figure 22: **Optimal shrinking.**

If the adjustment for misallocation is taken after depolarization, then the new optimal shrinking factor is $s^* = 0$. Taken together, these results suggest that extremism drives more of the inefficiency than malapportionment.

B.10 Qualitative Findings

Participants who said they used ChatGPT in the classroom were prompted to clarify how they used the tool. Although they were assured that the information would not be identifiable, these may not be honest representations of their usage. Nonetheless, they offer a glimpse into how students use ChatGPT in practice. Among other applications, students used ChatGPT to:

- Explain concepts and develop understanding, for homework, projects, and studying.
- Brainstorm, outline, and draft essays.
- Complete coding assignments or write code snippets.
- Complete multiple choice quizzes.

- Summarizes lengthy content.
- Preliminary research for essays.
- Validate answers.

The range of responses suggests a broad potential for ChatGPT in the classroom. A visualization of commonly used terms is displayed in Figure 23.
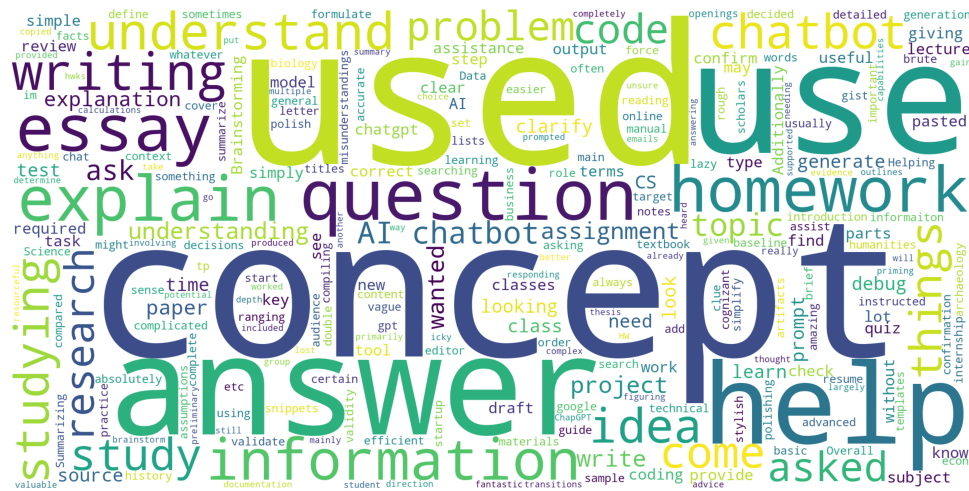

Figure 23: **Word cloud of ChatGPT usages.** The size of the phrase roughly corresponds to frequency. The most commonly mentioned use was learning "concepts."

B.11 Power Analysis

A power analysis is conducted for an unpaired t-test on each participant's average weight on advice across questions. Setting α=0.05 and power $1 - \beta = 0.8$, the sample is able to detect an effect size of Cohen's d=0.336. For the advisor condition, the effect size in the sample is d=0.195 requiring a sample size of n=412 to detect an effect. For the give justification conditions, the effect size is d=1.90 and it requires a sample of n=436 participants.