RESEARCH-ARTICLE

Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration

CRYSTAL QIAN, Massachusetts Institute of Technology, Cambridge, MA, United States

JAMES WEXLER, Google LLC, Mountain View, CA, United States

Open Access Support provided by:

Google LLC

Massachusetts Institute of Technology

Published: 18 March 2024

Citation in BibTeX format

IUI '24: 29th International Conference on Intelligent User Interfaces
March 18 - 21, 2024
SC, Greenville, USA

Conference Sponsors:
SIGCHI
SIGAI

IUI '24: Proceedings of the 29th International Conference on Intelligent User Interfaces (March 2024)
https://doi.org/10.1145/3640543.3645198
ISBN: 9798400705083

Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration

Crystal Qian
cjqian@google.com
Google Research
Cambridge, MA, USA
Massachusetts Institute of Technology
Cambridge, MA, USA

James Wexler
jwexler@google.com
Google Research
Cambridge, MA, USA

ABSTRACT

Although recent developments in generative AI have greatly enhanced the capabilities of conversational agents such as Google's Bard or OpenAI's ChatGPT, it's unclear whether the usage of these agents aids users across various contexts. To better understand how access to conversational AI affects productivity and trust, we conducted a mixed-methods, task-based user study, observing 76 software engineers (N=76) as they completed a programming exam with and without access to Bard. Effects on performance, efficiency, satisfaction, and trust vary depending on user expertise, question type (open-ended "solve" questions vs. definitive "search" questions), and measurement type (demonstrated vs. self-reported). Our findings include evidence of automation complacency, increased reliance on the AI over the course of the task, and increased performance for novices on "solve"-type questions when using the AI. We discuss common behaviors, design recommendations, and impact considerations to improve collaborations with conversational AI.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; **Natural language interfaces**; **Empirical studies in HCI**; • **Social and professional topics** → *Automation*; • **Information systems** → *Trust*.

ACM Reference Format:
Crystal Qian and James Wexler. 2024. Take It, Leave It, or Fix It: Measuring Productivity and Trust in Human-AI Collaboration. In *29th International Conference on Intelligent User Interfaces (IUI '24), March 18–21, 2024, Greenville, SC, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3640543.3645198

1 INTRODUCTION

Recent advancements in generative artificial intelligence (AI) have the potential to enhance productivity across domains such as medicine [37], research [38, 50], and technology [48]. In the context of software development, conversational AI such as Google's Bard and Open AI's ChatGPT can generate code, and Github's Copilot can autocomplete code. However, it's unclear whether these systems strictly improve productivity. State-of-the-art agents suffer from imperfect accuracy and biases [44, 59], and humans have demonstrated cognitive biases such as automation bias and effort substitution when using LLM-based systems such as ChatGPT and Copilot [2, 6, 42, 57]. Furthermore, behaviors and outcomes may depend on the expertise or literacy of the user; novices have been found to be less discerning of automation errors [44].

The opportunity remains that many developer tasks can benefit from an open-ended conversational AI interface, such as brainstorming ideas, answering questions [38], debugging errors, and addressing subjective topics such as translation [62] or coding conventions. To better understand developer interactions with conversational AI, we conducted a user study with 76 software engineers (N=76) at a large technology company as they completed an occupation-specific programming language exam with and without assistance from a conversational AI agent, Google's Bard.

We pose the following research questions in this setting:

- **RQ1, Effects on productivity**: How does usage of conversational AI affect productivity?
- **RQ2, Behaviors of trust**: How do users demonstrate trust in conversational AI?

To evaluate the value-add of conversational AI on productivity in **RQ1**, our experimental design randomizes ordering of access to conversational AI within-participant, comparing productivity when adding access to Bard to productivity when adding access to traditional resources. To evaluate **RQ2**, we construct an action space of trusting and distrusting behaviors. Across both productivity and trust constructs, we consider behavioral and self-reported measures. We also explore how effects may vary across levels of user expertise, constructing expertise rankings using a rich database of company-internal engineering statistics and self-reported survey responses.

We find that participants of all expertise levels increasingly depend on conversational AI over the course of the exam, despite mixed results on measured and perceived productivity. Novices are particularly influenced by these systems, which could imply opportunities for skill equalization. However, this can be a cause for concern, as generative models may be more likely to propagate misleading information compared to traditional forms of decision support tools with fixed outputs [59]. We also find evidence of incongruity between users' anticipated and demonstrated behaviors, suggesting that users are not fully cognizant of their interactions with these systems.

IUI '24, March 18–21, 2024, Greenville, SC, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0508-3/24/03
https://doi.org/10.1145/3640543.3645198

This paper contributes the following empirical evidence:

- **Access to AI can affect productivity and perceived productivity in different directions:** Users perceive increased productivity and efficiency when using the AI, despite spending more time on the task.
- **Expertise and context matter:** AI usage can improve performance, particularly for novices on open-ended tasks.
- **Users increasingly rely on AI over the course of the task:** Participants of all expertise levels increasingly depend on the AI, despite reporting less trust in the AI.
- **Experts distrust, and distrust imperfectly:** Relative to novices, experts are more likely to reject the AI. This can punish performance, as experts are less likely to use the AI to correct mistakes.
- **Usage of AI reduces cognitive load:** Participants substitute effort to the AI and report reduced cognitive load.

We also discuss common behaviors, design implications, and ethical considerations for more beneficial intelligent systems. This work aims to advance the understanding of productivity and trust formation in using conversational AI.

2 RELATED WORK

2.1 Usage of conversational agents

Since the release of ChatGPT in 2022, there has been increased interest in evaluating the performance of conversational agents. ChatGPT and GPT-4 can perform sufficiently well on a diverse range of analytical NLP tasks (e.g. sentiment analysis, question answering), but accuracy decreases as the difficulty of the task increases [28]. Despite this, these agents are increasingly adopted in professional [42] and academic [38, 48, 50, 54] settings, and are used to inform high-impact decision making around topics such as safety [44] and healthcare [37, 65].

2.2 Variation in interactions by user ability

Not all users interact equally with these systems; populations with lower literacy and education have been found to have higher risk of consuming unreliable content by conversational agents [44]. There are arguments as to why both experts and novices may be more or less amenable to automated assistance. Experts may reject systems due to a rational allocation strategy, deciding not to delegate tasks to an external system if the expert's internal trustworthiness is high [36, 41]. On the other hand, novices may exhibit an overestimated belief in their ability (Dunning-Krueger effect [31]) and also reject automated assistance [51]. Perceived expertise may be more relevant than expertise in trust formation; self-confidence can causally relate to automation usage [11, 23, 33]. Users of automated systems across all ability levels have exhibited self-reported, implicit, and explicit propensity to trust automation without prior interactions or adequate evidence about its capabilities [39], and exhibit higher tolerance for AI misfires [27].

2.3 Variation in interactions by context

On a writing task, ChatGPT usage increased performance for low performers and velocity for high performers [42]. However, on a programming task, Copilot did not increase the success rate of solving programming tasks or reduce task completion time, as developers spent more time validating generated outputs [57]. Writers substituted effort by directly copy-pasting outputs from ChatGPT without verification [42], and developers directed less visual attention to Copilot-generated code, despite its quality being comparable to human-written code [2]. Behaviors of automation bias and effort substitution have also been found in automation environments such as manufacturing and aviation technology [36, 45, 63, 68]. Effort substitution may be the optimal strategy if automated systems can perform perfectly. However, LLM-based systems can regurgitate or hallucinate potentially misleading information [5, 50, 61], and there is no one-size-fits-all solution on how to present information optimally in conversational systems to mitigate misinterpretation [12, 53, 66]. Despite that user trust is affected by the accuracy of ML systems [26, 67], users are willing to accept help from imperfect assistants [47, 61], and imperfect assistance can nonetheless improve performance [62].

3 EXPERIMENT DESIGN

3.1 Procedure

We invited a random sample of 1,400 US-based, full-time software engineers at a large technology company to participate in our study. Of the 220 who responded, 96 were accepted given our screening requirement that they had written and submitted code within the last 6 months. 79 respondents completed informed consent paperwork and scheduled study sessions, and 76 respondents completed the study (2 dropouts and 1 timeout, a 4% attrition rate).

Following a pre-task survey, each participant joined a 1-on-1, virtual, hour-long study with a moderator; they completed a 10-question multiple-choice online exam while sharing their screen and thinking aloud. Finally, they completed a post-task assessment.

3.2 Task design

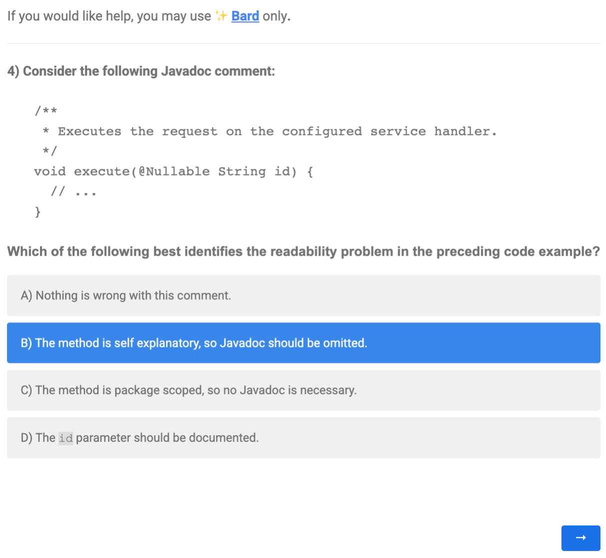
Exam: We chose a modified exam format for this task [19, 54] and calibrated the number of questions across 8 pilot sessions. The ten multiple-choice, single-answer questions on the exam come from a company-internal "readability" exam on the Java programming language, which is one component of the process undergone by software engineers to obtain a Java readability certification.[1] The exam tests understanding of the company's coding standards, which is documented in the company's internally- and externally- published Java Style Guide. Coding conventions can be subjective; the style guide is written to ensure consistency across the company. Often, multiple answers in the exam can compile and are technically valid, but one of the answers is more correct than others according to the company's Style Guide. This is stated in the introduction to the exam: *"There may be multiple correct answers.. Please choose the best one in line with the [Company] Java Readability Style Guide."* We verified that Bard could complete this task independently with reasonable performance; when directly given the exam questions

[1] This certification allows software engineers to submit code without requiring additional Java-specific review. There are also exams for other languages. We chose Java because it is a popular language [62] used commonly across our company, which yielded a broader distribution of expertise given our random sample of participants.

and corresponding multiple-choice options as inputs, Bard could answer 7.5 out of 10 questions correctly on average.[2,3]

Format: We divided the ten exam questions into two sections of five questions each: a "Bard-First" and a "Bard-Last" section. Section order and question order were randomized within-subject. Participants had two passes at each question. During the first pass per question in the "Bard-First" section, participants had access to Bard only. After selecting an answer, they had the option of modifying their answer with access to any non-LLM-based resources of their choice (e.g. documentation, search engines), loosely categorized as "Book" resources. During the first pass in the "Bard-Last" section, participants had access to Book resources. In the second pass, participants could *add Bard*, meaning they could modify their answer using Bard. Our intervention was *access* to these resources; participants did not have to use them. By allowing participants two passes at each question, we isolated the effect of adding access to Bard or Book in the second pass. To help users calibrate trust in the AI, we displayed a feedback screen after the second pass showing the correct answer and explanation [9].

Question types (search vs. solve): There are five "search"-type and five "solve"-type questions; participants are not explicitly told these classifications. Search-type questions are more straightforward and have answers that map directly to the style guide. For example, the correct answer to *"When would you declare a nested class as static?"* can be found near-verbatim in the chapter of the documentation about the "static" keyword. Solve-type questions are more open-ended and involve the critique of a provided code snippet; the answer cannot be found directly in the documentation. An example of a solve-type question is shown in *Fig.s 1* and *2*.



If you would like help, you may use ✨ Bard only.

4) Consider the following Javadoc comment:

```
/**
 * Executes the request on the configured service handler.
 */
void execute(@Nullable String id) {
  // ...
}
```

Which of the following best identifies the readability problem in the preceding code example?

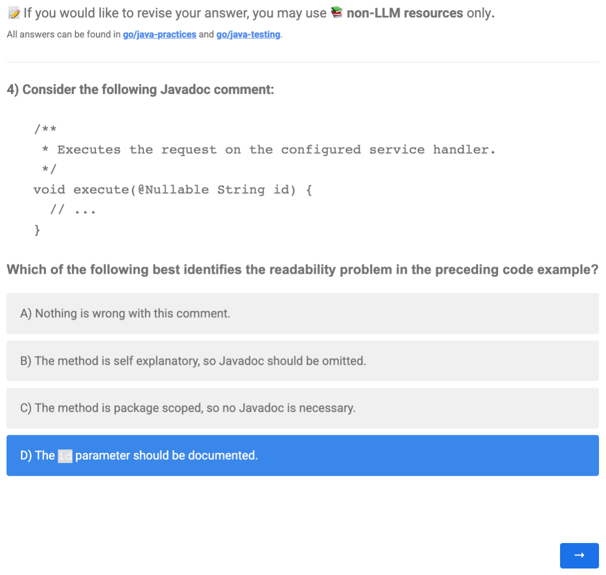
A) Nothing is wrong with this comment.

B) The method is self explanatory, so Javadoc should be omitted.

C) The method is package scoped, so no Javadoc is necessary.

D) The `id` parameter should be documented.

Figure 1: First pass on a Bard-first, solve-type question.



📝 If you would like to revise your answer, you may use 📚 **non-LLM resources** only.

All answers can be found in go/java-practices and go/java-testing.

4) Consider the following Javadoc comment:

```
/**
 * Executes the request on the configured service handler.
 */
void execute(@Nullable String id) {
  // ...
}
```

Which of the following best identifies the readability problem in the preceding code example?

A) Nothing is wrong with this comment.

B) The method is self explanatory, so Javadoc should be omitted.

C) The method is package scoped, so no Javadoc is necessary.

D) The `id` parameter should be documented.

Figure 2: Second pass on a Bard-first, solve-type question.

3.3 Expertise measurement

To evaluate variation in outcomes by expertise and perceived expertise, we constructed both an objective and self-reported expertise percentile rank within our sample. The objective expertise rank considers company-internal statistics such as amount of code written, tenure, Java experience, and previous "readability" exam experience. The perceived expertise rank considers participants' self-assessment of previous experience (Java, LLM-based tools, productivity) measured during the pre-task survey. *Appendix C* shows the data and calculations for these ranks. We refer to *experts* relative to *novices* as those with a higher objective expertise percentile rank.

3.4 Thematic analysis

In addition to evaluating descriptive data from the screening/pre-task surveys and quantitative data from the exam using statistical methods, we conducted a thematic analysis [7] on transcriptions of in-session think-aloud commentary and post-task survey responses. Initial codes for this analysis were generated from prior research, our 8 pilot studies, and the first batch of participant data. Themes about participant behaviors in this paper are saturated by codes that appeared in at least ten distinct user sessions [4, 21]. Representative quotes from this analysis supplement our findings.

4 RESULTS

4.1 Productivity

In *Fig. 3*, we show variation in productivity outcomes by expertise through a structural equation model (SEM) [56] relating our two percentile rank-transformed expertise measures (objective and self-reported) and three productivity constructs. Moving forward in the paper, we use the *objective* expertise percentile rank as the primary measurement of expertise in our discussion, as self-reported measures have low correlation with the outcomes of interest. Latent measures of productivity were constructed [17] as follows:

[2] We repeated trial runs and varied levels of priming and prompt engineering (e.g. providing context such as "*According to the [Company] Style Guide . . .*").
[3] Note that the task was built in the Qualtrics survey platform, which disabled copy-pasting of multiple-choice answers by default. This inadvertently required participants to exert more effort if they wanted to apply this direct copy-paste strategy.

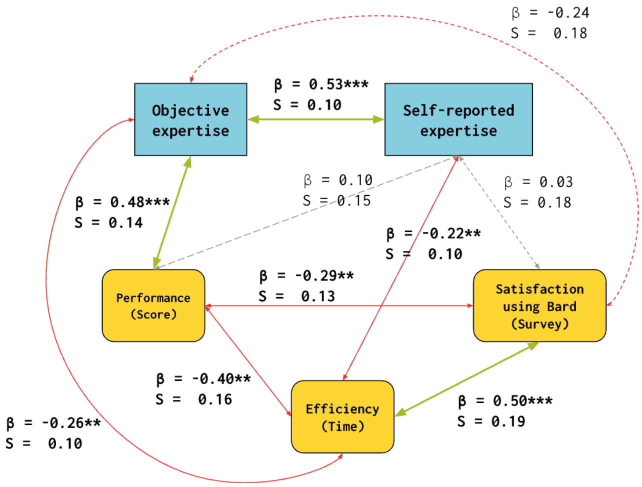


Figure 3: A structural equation model showing correlations between expertise measures and prodcutivity outcomes; β is the normalized effect size in standard deviations, and S denotes standard error.[4]

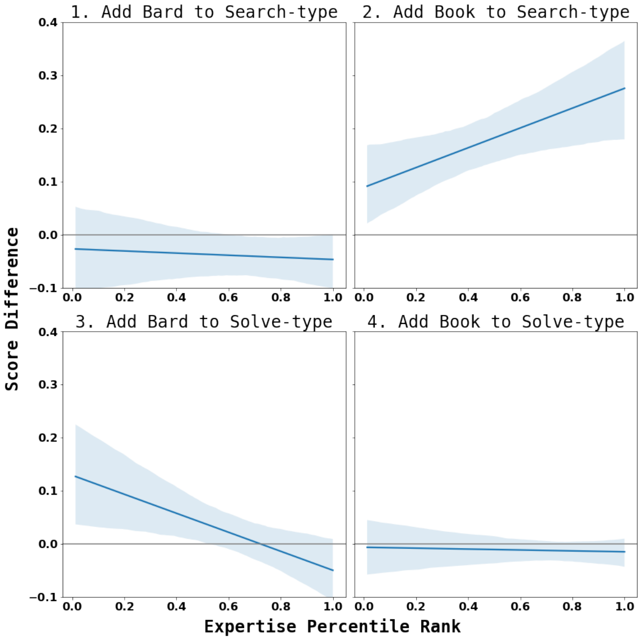


Figure 4: An OLS regression of the difference in scores between passes on expertise, with 95% confidence intervals.

- **Performance**: Total exam score (0–10). One point per correct answer following the second pass. No partial credit.
- **Efficiency**: Total time spent on the exam.
- **Satisfaction**: Summed satisfaction score aggregated from the post-task survey (*Table 2*).

4.1.1 *Performance.*

On average, participants scored 4.89 out of 10 points (σ = 1.7 points). There was no significant change in scores over time or between the passes of a question. Participants scored significantly higher on search-type questions. Measured expertise and final exam score have a significant positive relationship (*Fig. 3*).

Next, we calculate the effect of adding access to a resource across expertise by regressing the score difference between passes on expertise, adding fixed effects for treatment order (Bard-First vs. Bard-Last) and question type (search vs. solve). For example, a score difference of -1 indicates that the user changed from a correct to an incorrect answer between passes. These effect sizes are visualized in *Fig. 4* and described below.

(1) **Adding Bard to a search-type question** had no significant effect. Participants reported feeling confident in the answers they found using Book resources in the first pass, and would often skip their second pass with Bard.

> *"Sometimes [the docs] covered the exact topic."* —P12
>
> *"I trusted answers from the documentations more.. they were often more concrete. . . for example, 'make all nested classes static.' When I found answers in the documentation, I was much [more confident] in them."* —P77

(2) **Adding Book to a search-type question** significantly improved the score, especially for experts. Experts demonstrated more familiarity with navigating and finding relevant sections of documentation:

> *"I know we have documentation on this.. I've used it before. . . "* —P62
>
> *"Documentation was faster.. especially when I already had an idea of what the right answer was. "* —P64

Novices demonstrated less familiarity and more difficulty in identifying and interpreting relevant documentation:

> *"Because I don't use Java, none of the [documentation] means much to me. . . "* —P75
>
> *"I searched through the docs for what I lacked knowledge on, but . . . I wasn't sure what to search for."* —P63

(3) **Adding Bard to a solve-type question** improved performance for novices, but had no effect for experts. The Book did not contain answers to solve-type questions, so participants were largely reliant on their expertise. Experts were more likely to answer correctly in the first pass, benefiting less from Bard in the second pass.

(4) **Adding Book to a solve-type question** had no significant effect. Book resources had less specific guidance for critiquing a code sample:

> *"This is the kind of task that Bard does really well on. . . I don't know how to look in the docs for this."* —P24
>
> *"I don't think the documentation would be useful for this case."* —P69
>
> *"How would I even search for this?"* —P71

4.1.2 *Efficiency.*

On average, participants completed the exam in 39.6 minutes (σ = 9). Controlling for question number, there was a slight speedup per question of 20 seconds. There's no significant correlation between time per question and accuracy, but more time spent on the entire exam correlates with lower final score (*Fig. 3*).

[4] Across all tables and diagrams, we use the following significance notation: *: p<.10, **: p<.05, ***: p<.01.

	First Pass			*Second Pass*		
	Bard Only	*Book Only*	*Ind.*	*Add Book*	*Add Bard*	*Ind.*
All questions	15.11 (4.54)	12.92 (4.78)	***	4.93 (3.43)	6.66 (3.58)	***
Solve-type questions	7.30 (3.44)	6.03 (3.16)	**	2.30 (1.74)	3.45 (2.56)	***
Search-type questions	7.64 (3.58)	7.06 (4.02)		2.53 (2.33)	3.31 (2.41)	**

	First Pass		*Second Pass*	
	Bard Only	*Book Only*	*Add Book*	*Add Bard*
All questions	-0.92 (1.78)	**-4.95***** **(1.83)**	1.19 (1.18)	**-3.95**** **(1.54)**
Solve-type questions	-1.87 (1.42)	0.45 (1.25)	0.79 (0.71)	-0.76 (1.08)
Search-type questions	1.93 (1.46)	**-6.38***** **(1.46)**	0.97 (0.81)	**-3.75***** **(0.89)**

Table 1: *Left, a*) Descriptive statistics on efficiency. Mean time spent per section in minutes, with sd. in parentheses. The *Ind.* column displays significance values from a two-sample independence t-test between Bard and Book times. *Right, b*) Regression coefficients and standard errors from a regression of time spent (in minutes) on expertise percentile rank.[5]

Participants spent more time using Bard: Participants spent significantly more time using Bard, both in the first pass (Bard only) and in the second pass (adding Bard). This difference is more significant for solve-type questions (*Table 1a*). Here are some factors as to why participants spent more time with Bard:

- **Slower response times:** Participants noticed and expressed frustration at more latency in AI response times compared to search query response times. This may be a transient issue as generative AI technology matures.
- **Less specific visual direction:** The documentation had clear headers that led the eye to the correct answer; participants would stop reading after they identified the appropriate passage for a search-type question. Bard’s generated outputs were paragraph-like and verbose; the correct guidance was more obscured within the text.
- **Verbose interactions:** Participants wrote in longer sentences and had more interactions with Bard following their initial query, in contrast to having brief keyword interactions when using other resources. For example, P77 queried for *"IllegalArgumentException"* and P78 queried for *“best practice autovalue java”* in a search engine. When they used Bard to answer the same questions, they wrote the following:

  > *“Heya, could you please write me a Java function to assert that removing an item from a list throws an IllegalArgumentException? Thanks.”* —P77
  >
  > *“Give me code examples to show the best way to test for an expected exception. . . Do so in a unit test . . . I don’t feel like this is correct?”* —P78, in a back-and-forth conversation

Experts and novices spend similar time using Bard: *Table 1b* regresses time spent per question on expertise. When experts have access to Book resources in the first pass for a search-type question, they spend less time than novices on both passes. However, when experts have access to Bard first, they do not interact with or interpret Bard outputs any faster than novices.

Participants felt more efficient using Bard: Despite spending more time using Bard, participants significantly agreed with the following in their post-survey assessment (*Table 2*): *2. I complete tasks faster when using Bard, 3. I spend less mental effort when using Bard*, and *4. I spend less time searching for information or examples when using Bard.*

> *“For someone who knows very little about Java, Bard would speed up my workflow a lot. I would have to read a lot of the style guide and Bard makes things much faster.”* —P11
>
> *“I definitely found [Bard] easier than searching the documentation . . . I found that using Bard was surprisingly effective... It seemed faster to use Bard because I could ask it more stream of consciousness questions.”* —P50

4.1.3 Satisfaction. Participants felt significantly more productive when using Bard compared to Book (*Table 2, Statement 1*). Novices agreed with this more so than experts, perhaps because experts are more likely to identify flaws with the AI [43]. However, there is no significant change in satisfaction (*Statement 5*) or frustration (*Statement 6*) following the task.

Compare how you felt when completing tasks with Bard assistance, as opposed to without Bard assistance.	μ	β
1. I am more productive when using Bard.	**.24**** **(0.92)**	**-0.77*** **(0.40)**
2. I complete tasks faster when using Bard.	**.04***** **(1.11)**	-0.66 (0.48)
3. I spend less mental effort when using Bard.	**0.46***** **(1.19)**	-0.03 (0.49)
4. I spend less time searching for information or examples when using Bard.	**0.51***** **(1.06)**	-0.35 (0.44)
5. I feel more satisfied with my work completing this task when using Bard.	-0.11 (0.96)	-0.57 (0.41)
6. I find myself less frustrated when using Bard.	0.07 (0.99)	0.26 (0.43)

Table 2: Comparative satisfaction survey [69] administered post-task.[6]

[5] These regressions are performed at the per-question level with relevant controls as stated (N=190 per question type, N=380 across all questions). Data appears normally distributed, and a t-test for independence is sufficient at this sample size [15].

[6] Mean values μ indicate the degree of agreement with the statement, calculated using the method described in *Appendix D*. Significance values are calculated from a one-sample t-test with the null hypothesis that $\mu = 0$, and regression coefficients β demonstrate the relationship between response values and expertise. Highlighted cells indicate significant findings.

	First pass	Second pass	Behavior/Implication
1	Skip	Skip	Distrusted Both
2	Skip	Bard	Distrusted Book, trusted Bard
3	Skip	Book	Distrusted Bard, trusted Book
4	Bard	Skip	Distrusted Book, trusted Bard
5	Book	Skip	Distrusted Bard, trusted Book
6	Bard	Book	If the answer changed: trusted Book
7	Book	Bard	If the answer changed: trusted Bard

Table 3: Action space and implications. A participant *uses* a resource if they click into it during the pass. A participant *skips* if they do not click into the resource during the pass.

4.2 Trust

We measure trust using both demonstrated measures (actions taken) and perceived measures (self-assessment) [29].

4.2.1 Demonstrated measures of trust. Users trust and depend upon a resource when they delegate and rely on it [35, 65] and distrust when they reject it [46].[7] These actions in our study are described in *Table 3*. Participants can trust a resource either correctly or incorrectly;[8] this attribution depends on the resulting score.

Fig. 5 shows the percentage of answer-changing behaviors between passes; participants often do not change their answer across passes. When the correctness of a participant's answer does not change between passes, the trust implication is ambiguous: for example, if a participant got the correct answer during both passes, they could have already known the correct answer, relied upon either or both resources, or used but dismissed those resources.[9]

How does trust change over time? *Table 4a* shows a regression of trust actions on question order, with expertise and question number fixed effects.[10] Experts significantly depended on Book resources more so than novices, and dependence on Book resources did not change over time. All participants, particularly novices, increased dependence on Bard over the course of the exam, despite reporting that they wanted to decrease Bard usage:

> *"So I got it right, and then I got it wrong with Bard.. maybe I shouldn't trust Bard then. "* —P13

> *"I don't think what Bard is saying is true... I'll probably stop using Bard, just because it's incorrect."* —P35

Who correctly and incorrectly trusts? *Table 4b* regresses the likelihood of taking a correctly trusting or incorrectly trusting action on expertise. Participants of all expertise are equally likely to be led astray and incorrectly trust Bard. Novices are slightly more likely to correctly trust Bard across both types of questions.

Who exhibits distrust? *Table 5.1* regresses the likelihood of expressing distrust on expertise, conditional on question effects (question number and question order). As expertise increases, the likelihood of distrusting both resources and distrusting Bard increases, and the likelihood of distrusting Book resources decreases.

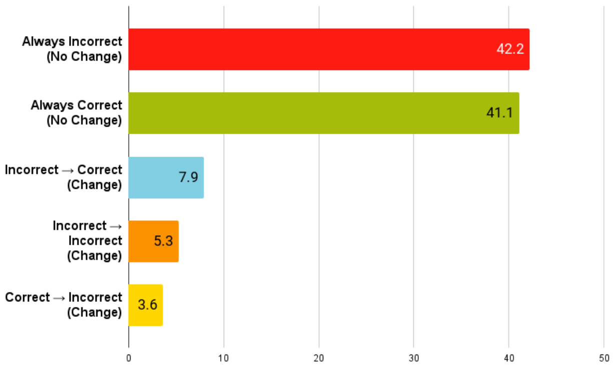


Figure 5: Answer-changing percentages between passes.

That is, novices were more likely than experts to distrust Book resources, and experts were more likely to distrust Bard.

Is it a good strategy to distrust? *Table 5.2* regresses score per question on likelihood of distrusting, conditional on question effects and expertise. For solve-type questions, distrusting either resource had no effect on the score. For search-type questions where the answer could be found in the Book, relying on the Book yielded better scores. Better performance by experts may be partially attributed to experts knowing when to appropriately distrust Bard and rely on the Book for Search-type questions. However, this distrust behavior could also punish performance; as expertise increased, the likelihood of using Bard to correct an incorrect answer decreased.

4.2.2 Perceived measures of trust. We administered Jian's Trust in Automated Systems survey [25] both before and after the task. Survey questions and responses are shown in *Appendix D*.

Participants calibrated trust post-task: Before the task, participants' sentiments towards Bard were mostly neutral. Following the task, participants expressed significantly less trusting sentiments towards Bard. This change is likely due to participants *appropriately calibrating trust* rather than *losing trust*, given that users tend to replace dispositional, pre-exposure trust in automation following exposure to that system with feedback [16, 40].

Most of our sample did not have significant prior experience interacting with the AI; 60.7% reported in the pre-survey that they had rarely or never used LLM-assisted tools in development tasks at work, and many reiterated this during the task.[11]

5 DISCUSSION

In this section, we discuss the key insights from our results, supported by literature review and patterns from our thematic analysis. Sections 5.1 through 5.3 touch on **RQ1: Effects on productivity**, digging into why we observed mixed results across different dimensions of productivity. Sections 5.4 and 5.5 touch on **RQ2: Behaviors of trust**, offering more context on participants' decision making behavior. Finally, in 5.6, we synthesize our observations into design implications, intended to aid developers in improving the design of intelligent conversational systems.

[7] With respect to our paper title, participants can use the AI (*Take it*), reject the AI (*Leave it*), or change their answer following usage of the AI (*Fix it*).

[8] Incorrectly trusting is also referred to as *mistrust* or *overtrust* in literature [34].

[9] In our experimental design, we considered asking participants which of these scenarios were the case after each question, but this added considerable time to each user session. Pilot participants demonstrated fatigue after the hour-mark.

[10] Tables 4 and 5 show regression coefficients with standard errors in parentheses.

[11] This may be due to a company-internal policy to avoid input of business-sensitive information and code into conversational AI agents.

	First Pass			*Second Pass*		
	Bard Only	***Book Only***	***Ind.***	***Add Book***	***Add Bard***	***Ind.***
All questions	15.11 (4.54)	12.92 (4.78)	***	4.93 (3.43)	6.66 (3.58)	***
Solve-type questions	7.30 (3.44)	6.03 (3.16)	**	2.30 (1.74)	3.45 (2.56)	***
Search-type questions	7.64 (3.58)	7.06 (4.02)		2.53 (2.33)	3.31 (2.41)	**

	First Pass		*Second Pass*	
	Bard Only	***Book Only***	***Add Book***	***Add Bard***
All questions	-0.92 (1.78)	**-4.95***** **(1.83)**	1.19 (1.18)	**-3.95**** **(1.54)**
Solve-type questions	-1.87 (1.42)	0.45 (1.25)	0.79 (0.71)	-0.76 (1.08)
Search-type questions	1.93 (1.46)	**-6.38***** **(1.46)**	0.97 (0.81)	**-3.75***** **(0.89)**

Table 4: *Left, a*) Regression of trust actions on question order (over time). *Right, b)* Regression of the likelihood of taking a correctly trusting or incorrectly trusting action (*Table 3*) on expertise.

	Distrusted both (Self-trust)	**Distrusted Bard**	**Distrusted Book**
1. Effect of expertise on the likelihood of expressing distrust			
All questions	**0.12***** **(0.04)**	**0.10***** **(0.04)**	**-0.20***** **(0.05)**
Solve-type questions	**0.16***** **(0.06)**	0.03 (0.04)	**-0.24***** **(0.08)**
Search-type questions	**0.09*** **(0.5)**	**0.17***** **(0.06)**	**-0.17**** **(0.07)**
2. Effect of expressing distrust on performance			
All questions	0.01 (0.06)	**0.24***** **(0.06)**	**-0.17***** **(0.04)**
Solve-type questions	0.09 (0.08)	0.11 (0.11)	-0.01 (0.06)
Search-type questions	-0.12 (0.08)	**0.16**** **(0.07)**	**-0.30***** **(0.06)**

Table 5: Regression coefficients on distrust behavior.

5.1 Why might using the AI hurt performance?

If users were rational, processed information optimally, and had perfect information, receiving more information would be better than receiving less information [52]. However, receiving more information at times hurt performance during the task: adding Bard did not strictly improve score (*Fig. 4*) and participants would at times change their answer from a correct to an incorrect answer after consulting additional resources (*Fig. 5*). This behavior penalized experts more, as experts were likely to get the correct answer in the first pass. Here are some commonly observed behaviors that may indicate why participants might switch from a correct to an incorrect answer following a consultation with the AI:

(1) **Users may not perceive the downsides of eliciting a second opinion.** Experts and novices were equally susceptible to changing an answer from correct to incorrect. Because this task was objectively scored based on specific correct answers, adding resources after already deriving the correct answer could only lead participants astray, not make them more correct. Participants did not seem to realize this pitfall:

> *"I can use Bard, since it's available.. why not?"* —P66

(2) **Participants exhibit confirmation bias.** Because participants could not copy multiple-choice answers easily due to infrastructure limitations[12], they would instead ask pointed questions to Bard such as *"Is A)... the right answer?"* Participants would seek out agreement and end their line of inquiry after receiving an affirming response, consistent with behavior exhibited in similar studies [12, 41].

> *"Let's go with Bard, [because] this time, Bard and I agree on the answer."* —P37

> *"[My strategy was that] I would ask Bard if they agreed that my answer was correct."* —P28

However, when participants prompted Bard without giving sufficient context, Bard could make a case for affirming any of the provided answers.

> *"[Bard] is not being very helpful because it's just validating everything I'm saying."* —P60

> *"In multiple cases, we had scenarios where [Bard] would confirm all of the answers..."* —P12

Participants found better success in asking comparative questions, e.g. *"Which is the bigger problem, X or Y?"*

> *""[I'd] either ask Bard for free-form improvements to the code sample (not very reliable), or gave Bard the code and a couple answers I'm undecided between and ask Bard to pick between them (more reliable)..."* —P60

> *"Maybe Bard is overeager to please- we got a yes on every single one of these when asked individually. Maybe I'd get a more discriminating answer if I put in multiple options."* —P12

5.2 Why do participants perceive that they are more productive and efficient with the AI?

Despite the mixed measured effects on productivity, participants still perceived increased productivity and efficiency when using the AI (*Table 2*), perhaps due to reduced cognitive load.

Using the AI is easy: When using Book resources, participants would spend time *actively*: searching for answers, skimming the text, and thinking about how to phrase keywords. When using Bard, participants spent time more *passively*, waiting for responses and reading outputs [41]. Participants across expertise levels exhibited effort substitution [64] by blindly copy-pasting questions.

[12] Often, participants' first instinct was to attempt to copy all answers into Bard. Several participants expressed frustration upon discovering that copying answers was disabled in the Qualtrics platform. From P21: *I can't copy the answers, so it would be too much effort to ask [Bard].* Four participants even inspected the elements within the web browser to copy-paste the source code as a workaround. One participant painstakingly typed out each multiple-choice answer for each question, but abandoned this strategy due to time constraints.

> *"I feel a little tired, so I will just start copy-pasting the question."* —P19
>
> *"I would use Bard to reduce the cognitive load."* —P31
>
> *"I liked Bard because you didn't have to do too much, you could just copy-paste and it would potentially find the answer that I'm looking for."* —P47
>
> *"I guess I'll actually read the question now while Bard is thinking."* —P74

Users exhibit automation complacency: If the AI produced optimal responses and users substituted effort appropriately, effort substitution would not be detrimental. However, given that we do not find strictly positive effects of access to the AI on productivity, this scenario meets Parasuraman's three requirements for automation complacency [46][13]: (1) A human operator is monitoring an automated system. (2) The frequency of such monitoring is lower than optimal. (3) There is a directly observable (negative) effect on performance.

5.3 Why is there no change in satisfaction or frustration?

Despite feeling more productive and efficient, participants were not more satisfied or less frustrated when using the AI (*Table 2*), perhaps because negative emotional reactions and inappropriate sentiments of trust offset the perceived gains.

Participants attribute blame asymmetrically: When participants missed a question using Book resources, they were less likely to offer an explanation or attribute the fault to the resource directly. When they missed a question using Bard, they were more likely to display defensive behavior, justify efforts, and blame Bard or their ability to interpret and prompt Bard.

> *"Maybe I interpreted Bard's outputs wrong."* —P29
>
> *"I'm not the master at prompting LLMs yet."* —P47
>
> *"Bard is leading me astray! But I don't know how to tell if it's telling me stuff incorrectly."* —P49

Participants perceive Bard as a collaborator: *"Emotional reactions may be a key element of trust and the decision to rely on automation ... "* [34] and can be activated through perceived collaborations with automation [32, 62].

> *"Looks like Bard and I were wrong for this one."* —P35
>
> *"[Bard] is like working with a pretty well-informed tutor. It's highlighting problems that are deeper than the multiple-choice questions and answers we were looking for."* —P39
>
> *"We don't have many colleagues in the office these days... sometimes, it's much faster to ask colleagues since they have context. But given they're not there, I always go with Bard. "* —P66

Further evidence of perceived collaboration was found as participants described Bard using language such as *"It probably doesn't like me"* -P10 and *"Bard is probably overwhelmed"* -P12. However, *"using speech to create a conversational partner.. may lead people to attribute human characteristics to the automation in such a way as to induce false expectations that could lead to inappropriate trust."* [34] This inappropriate trust, in turn, can induce less satisfaction with the AI [64].

5.4 How do participants use resources?

The overwhelmingly most common behavior was copy-pasting questions directly into Bard. Participants would also *prime* the agent by adding context (e.g. *"You are reviewing this code according to the [company] style guide"*). When using Book resources, participants would often search for keywords found in the question, either by manually skimming the documentation or querying external or company-internal search engines. If the question was *"When would you declare a nested class as static?"*, participants might search for tokens such as *"nested class static"*. They tended to not copy-paste a question directly into a search engine, especially when it was a solve-type question with code snippets.

5.5 How do participants pick what to use?

The decision to use either resource was significantly correlated with question type and user expertise. For both search-type and solve-type questions, participants, particularly novices, preferred to consult the AI first. Participants reported wanting to use both resources concurrently, specifically by using Bard, then Book, then Bard as a sanity check.

The AI reduces search frictions for novices: Consulting the AI as a jumping off point helped novices to identify where to look in the documentation based on its recommendation.

> *"When I really did not have an idea... Bard was helpful, because it looks broadly whereas searches had to be very precise."* —P64
>
> *"I use [Bard] for more open-ended/general questions... When I have a specific question, I go to sources that are written and more concrete. I use Bard when there's something I don't know."* —P60
>
> *"I would prefer to use Bard first, so I can ask a general question to Bard. It's fast and gives me an answer quickly... and if I am not happy with Bard's answer, I can do research on my own."* —P72

Users assume that the AI has limitations: Many participants assumed that because Bard is trained on publicly available data, it would not be familiar with company-internal coding conventions. None checked to see whether the style guide was publicly available or tried to validate if Bard was familiar with the company's Style Guide. Instead, they rejected the AI based on their assumption, perhaps because there is no way for users to concretely verify whether the AI is trained on any specific data source.

> *"I feel like Bard is an external tool, so for a task like readability [which is internal], it might not know the answers. For myself, if I want to find the answers, I'd just use [internal search], because I feel like external tools don't apply to our standards."* —P6

[13] We cannot make a similar claim about *automation bias*, which is defined as evidence of omission and commission errors when decision aids are imperfect [45]. We observe evidence of both errors; participants fail to notice omissions by the agent, and can be actively misled by the agent. However, this definition may be more appropriate for traditional automation; users have much more control over the outputs of conversational AI, so errors could be the result of either automation bias or imperfect usage.

> *"Because this is readability for [our company], and maybe what Bard gives me is general readability advice, so maybe it's not so useful for [our] readability questions... "* —P78

5.6 Design implications for more effective conversational AI

Despite Bard having the capability to perform better than the average user on this task when questions and answers were directly copy-pasted as inputs, participants did not employ this strategy, perhaps due in part to the burden of increased effort exertion.[14] Furthermore, the optimal strategy is dependent on both the context of the task and the expertise of the user relative to the capabilities of the agent. Feedback is given ex-post, which can make determining the optimal strategy intractable for the user. This suggests that there is room for improvement on behalf of the AI system to improve productivity; we recommend the following ideas for developers of conversational AI systems.

5.6.1 Design for appropriate trust. Users appropriately trust systems when they reject incorrect advice and accept correct advice [53, 61]. Designing for *appropriate* trust, not *greater* trust [34] can help mitigate to overreliance [24].

(1) **Lower the degree of confidence.** Generative models can be perceived as overconfident, which can cause users of all expertise levels to display inappropriate trust. Communicating uncertainty on behalf of the agent can reduce undue overreliance [47, 49]. However, this may also force a user's cognitive effort, which can decrease satisfaction [10].

> *"I feel like [Bard] confused me... it's so confident when it's wrong, so it's hard to follow your own barometer. The confidence scares me because it's just as confident when it's wrong compared to when it's right, and I'm bad at refuting someone who's confident. "* —P10

> *"I was not sure if I should trust the Bard result... but it sounds so smart, so it must be correct!"* —P63

(2) **Be cautious about creating a conversational partner.** Participants attributed human characteristics to the AI (*Section 5.3*); anthropomorphizing can raise user expectations [20] and cause overreliance [64].

(3) **Consider user customization.** Our findings suggest that access to the AI affect users of different expertise levels differently. Through longitudinal exposure, the AI could build models of its human users and customize output by expertise and form a mutual theory of mind [49, 60].

5.6.2 Cite sources. Participants preferred that Book resources were curated, deliberately written, peer-reviewed, and tried-and-true. When they navigated external search engines, they used websites that they were familiar with and looked for evidence of peer-review, such as highly ranked StackOverflow responses [1]. In contrast, there was a sense of skepticism when using Bard; its generated output felt less intentional. Improving source attribution in LLMs and integrating appropriate citations into conversational output could bridge the perceived disconnect between generated and human-written output, which could lead to more appropriate trust [64].

> *"There's a level of intentionality with the docs, like someone actually wrote this and put this together... not knowing Bard, [its output] could be anything."* —P75

> *"[I] tend to be skeptical about the answers people give without giving the source.'* —P78

6 LIMITATIONS AND FUTURE WORK

These findings may be limited to our particular context. Our study sample is limited to software engineers at a US-based technology company. They may have different attitudes towards AI and higher machine learning literacy as compared to laypeople, which may affect performance [14] and behaviors [22]. They've also been given direction to refrain from putting company-specific data in conversational AI systems, which may limit their familiarity.

Although empirical studies with other AI-based systems have produced similar findings on productivity and trust [2, 19, 42, 53, 57], it's possible that our results are idiosyncratic to Bard. Reproduction of this task with other agents or a standardization of a behavioral task suite can help to generalize this work to other AI-based systems.

Finally, the task design may be scrutinized. Users interacted with the system for less than an hour. Perhaps trust formation takes longer time, more exposure, and more feedback; we may benefit from a longitudinal study and extended follow-up [68]. This study was moderated, which may induce surveyor bias. Much of the recent, similar work on behavioral interactions with automated systems have been performed in unmoderated settings with online populations, who have been shown to behave differently [18]. Participants were given a flat thank-you gift regardless of performance on the exam; perhaps a higher-stakes setting or a piecewise incentive structure would induce more effort.

7 IMPACT CONSIDERATIONS

We demonstrate that users are willing to take up conversational AI to complete a workplace task, and that this assistance has the potential to improve user productivity. If users appropriately calibrate trust in these systems and use them in applicable settings, these systems can potentially increase productivity[23].

We also find that usage of these systems vary depending on user expertise. In this task, experts tended to distrust automated assistance. Novices, who were more likely to adopt and rely on these systems, were more susceptible to be influenced. This increased adoption of conversational AI by novices has the potential to equalize productivity across expertise. However, our findings suggest that adoption is not always beneficial: all participants exhibit automation complacency, and access to AI can be potentially detrimental in specific contexts. This could propagate inequitable outcomes [38, 58]; learning differences may be exacerbated if conversational AI is applied in an academic setting, and malicious actors may employ these systems to disseminate disinformation to populations with lower literacy [59].

Furthermore, the outputs of these systems hold weight. In our study, participants of all expertise levels could be convinced to change correct answers to incorrect answers following exposure

[14] Some participants also suggested that having agency in the task felt important. From P47: *"Copying things directly into Bard is a little silly."*

Results
1. *Performance:* Access to the AI can improve performance on certain task types, and benefits novices more than experts.
2. *Efficiency:* Users may spend more time using the AI, yet perceive increased efficiency when using it.
3. *Satisfaction:* Users may feel more productive using the AI, yet not more satisfied.
4.*Trust:* Users may increase dependence on the AI over time. Users of all expertise levels are susceptible to mistrust.
5. *Distrust:* Relative to novices, experts are more likely to distrust the AI. This can punish their performance, as experts are less likely to use the AI to recover from mistakes.
Behaviors
1. Using the AI reduces search frictions, particularly for novices.
2. Users may reject the AI based on assumptions about its limitations.
3. Users do not perceive potential downsides of eliciting a second opinion.
4. Users exhibit confirmation bias and seek out agreement from the AI.
5. Users substitute effort to the AI, which reduces cognitive load.
6. Users exhibit automation complacency.
7. Users attribute blame asymmetrically.
8. Users perceive the AI as a collaborator.
Recommendations
1. Design for appropriate trust.
2. Display the appropriate degree of confidence.
3. Be cautious about creating a conversational partner.
4. Consider user customization.
5. Cite sources.

Table 6: Summary of findings.

to the AI. The impact of missing a few questions on a company-internal coding exam is fairly minimal. However, conversational AI has the capability to inform more consequential actions such as obtaining a medical license [19], consulting in clinical consultations [37], or informing pandemic responses [64]. Furthermore, these LLM-based systems can mislead, hallucinate, and regurgitate incorrect information; for example, they can perpetuate unfair biases in the context of gender, race and religion [8] and cite non-existent research [50]. As we work to improve fairness and representation in these systems, we should concurrently improve model interpretability and user literacy to mitigate the potential of these systems to mislead.

8 CONCLUSION

In this study, we evaluated how access to conversational AI affects user productivity and trust formation through a user study of 76 software engineers as they completed an occupation-specific exam with and without access to a conversational AI agent. Broadly, we find that the effects on productivity and trust depend on the context and the user, and that having access to AI is not strictly better than not having access to AI. This evidence suggests that while these generative AI systems have the capability to affect and potentially augment worker productivity, they are not yet used in a way that can completely replace human effort or traditional resources.

We employed a mixed-methods approach of qualitative thematic analysis and quantitative statistical methods. This would at times yield seemingly inconsistent results; for example, participants perceived efficiency gains with AI assistance despite objectively taking more time on the task with AI assistance, and reported being less trustful of the AI despite increasingly depending on its outputs. We invite extensions of this work to continue exploring mixed-method approaches to capture a more holistic interpretation of behaviors.

As generative AI becomes more powerful and accessible, it becomes increasingly important for researchers and developers to understand the effect of these systems on users. We need to design these systems to account for human behaviors such as confirmation bias and automation complacency. System design should also prioritize minimizing the propagation of potentially misleading information, especially as our findings suggest that users, particularly novices, increasingly depend on these systems over time.

This paper contributes empirical evidence from a real-world scenario in the field of human-AI interaction [3]. We hope that this contribution will motivate more work in building more productive and trustworthy systems based on conversational AI.

ACKNOWLEDGMENTS

We thank our pilot and study participants, Google's People and AI Research (PAIR) team, and the MIT Economics department for their valuable advice and support, notably Michael Terry, Lucas Dixon, Carrie Cai, Tobias Salz, Nikhil Agarwal, Ryan Mullins, Robert Upton, Martin Wattenberg, Fernanda Viégas, Hal Ableson, Mike Schaekermann, Key Lee, Andrea Knight Dolan, Mahima Pushkarna, and Hansa Srinivasan.

REFERENCES

[1] Rabe Abdalkareem, Emad Shihab, and Juergen Rilling. 2017. What Do Developers Use the Crowd For? A Study Using Stack Overflow. *IEEE Software* 34, 2 (2017), 53–60. https://doi.org/10.1109/MS.2017.31

[2] Naser Al Madi. 2023. How Readable is Model-Generated Code? Examining Readability and Visual Inspection of GitHub Copilot. In *Proceedings of the 37th IEEE/ACM International Conference on Automated Software Engineering* (Rochester, MI, USA) *(ASE '22)*. Association for Computing Machinery, New York, NY, USA, Article 205, 5 pages. https://doi.org/10.1145/3551349.3560438

[3] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[4] Hikari Ando, Rosanna Cousins, and Carolyn Young. 2014. Achieving Saturation in Thematic Analysis: Development and Refinement of a Codebook,. *Comprehensive Psychology* 3 (2014), 03.CP.3.4. https://doi.org/10.2466/03.CP.3.4 arXiv:https://doi.org/10.2466/03.CP.3.4

[5] Razvan Azamfirei, Sapna R Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care* 27, 1 (2023), 1–2.

[6] Shraddha Barke, Michael B. James, and Nadia Polikarpova. 2023. Grounded Copilot: How Programmers Interact with Code-Generating Models. *Proc. ACM Program. Lang.* 7, OOPSLA1, Article 78 (apr 2023), 27 pages. https://doi.org/10.1145/3586030

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., Virtual, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[9] Emily Brunsen, Imani Murph, Anne C. McLaughlin, and Richard B. Wagner. 2021. The Influence of Feedback Types on the Use of Automation During Learning. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 65, 1 (2021), 143–147. https://doi.org/10.1177/1071181321651228 arXiv:https://doi.org/10.1177/1071181321651228

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 188 (apr 2021), 21 pages. https://doi.org/10.1145/3449287

[11] Wanling Cai, Yucheng Jin, and Li Chen. 2022. Impacts of Personal Characteristics on User Trust in Conversational Recommender Systems. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 489, 14 pages. https://doi.org/10.1145/3491102.3517471

[12] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. 2023. Supporting High-Uncertainty Decisions through AI and Logic-Style Explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 251–263. https://doi.org/10.1145/3581641.3584080

[13] Raj Chetty, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014. Where is the land of Opportunity? The Geography of Intergenerational Mobility in the United States *. *The Quarterly Journal of Economics* 129, 4 (09 2014), 1553–1623. https://doi.org/10.1093/qje/qju022 arXiv:https://academic.oup.com/qje/article-pdf/129/4/1553/30631636/qju022.pdf

[14] Chun-Wei Chiang and Ming Yin. 2022. Exploring the Effects of Machine Learning Literacy Interventions on Laypeople's Reliance on Machine Learning Models. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 148–161. https://doi.org/10.1145/3490099.3511121

[15] Morten W Fagerland. 2012. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC medical research methodology* 12, 1 (2012), 1–7.

[16] K. J. Kevin Feng and David W. Mcdonald. 2023. Addressing UX Practitioners' Challenges in Designing ML Applications: An Interactive Machine Learning Approach. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 337–352. https://doi.org/10.1145/3581641.3584064

[17] Nicole Forsgren, Margaret-Anne Storey, Chandra Maddila, Thomas Zimmermann, Brian Houck, and Jenna Butler. 2021. The SPACE of Developer Productivity: There's More to It than You Think. *Queue* 19, 1 (mar 2021), 20–48. https://doi.org/10.1145/3454122.3454124

[18] Guillaume R. Fréchette, Kim Sarnoff, and Leeat Yariv. 2022. Experimental Economics: Past and Future. *Annual Review of Economics* 14, 1 (2022), 777–794. https://doi.org/10.1146/annurev-economics-081621-124424 arXiv:https://doi.org/10.1146/annurev-economics-081621-124424

[19] Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, and David Chartash. 2023. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Med Educ* 9 (8 Feb 2023), e45312. https://doi.org/10.2196/45312

[20] Eun Go and S. Shyam Sundar. 2019. Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* 97 (2019), 304–316. https://doi.org/10.1016/j.chb.2019.01.020

[21] Greg Guest, Arwen Bunce, and Laura Johnson. 2006. How Many Interviews Are Enough?: An Experiment with Data Saturation and Variability. *Field Methods* 18, 1 (2006), 59–82. https://doi.org/10.1177/1525822X05279903 arXiv:https://doi.org/10.1177/1525822X05279903

[22] Neeraja Gupta, Luca Rigotti, and Alistair Wilson. 2021. The Experimenters' Dilemma: Inferential Preferences over Populations. arXiv:2107.05064 [econ.GN]

[23] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 453–463. https://doi.org/10.1145/3581641.3584052

[24] Makoto Itoh. 2011. A model of trust in automation: Why humans over-trust?. In *SICE Annual Conference 2011*. IEEE, Tokyo, Japan, 198–201.

[25] Jiun-Yin Jian, Ann M. Bisantz, and Colin G. Drury. 2000. Foundations for an Empirically Determined Scale of Trust in Automated Systems. *International Journal of Cognitive Ergonomics* 4, 1 (2000), 53–71. https://doi.org/10.1207/S15327566IJCE0401_04 arXiv:https://doi.org/10.1207/S15327566IJCE0401_04

[26] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C.P. Snijders. 2023. It Seems Smart, but It Acts Stupid: Development of Trust in AI Advice in a Repeated Legal Decision-Making Task. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 528–539. https://doi.org/10.1145/3581641.3584058

[27] Shivani Kapania, Oliver Siy, Gabe Clapper, Azhagu Meena SP, and Nithya Sambasivan. 2022. "Because AI is 100% Right and Safe": User Attitudes and Sources of AI Authority in India. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 158, 18 pages. https://doi.org/10.1145/3491102.3517533

[28] Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, Anna Kocoń, Bartłomiej Koptyra, Wiktoria Mieleszczenko-Kowszewicz, Piotr Miłkowski, Marcin Oleksy, Maciej Piasecki, Łukasz Radliński, Konrad Wojtasik, Stanisław Woźniak, and Przemysław Kazienko. 2023. ChatGPT: Jack of all trades, master of none. *Information Fusion* 99 (2023), 101861. https://doi.org/10.1016/j.inffus.2023.101861

[29] Spencer C Kohn, Ewart J de Visser, Eva Wiese, Yi-Ching Lee, and Tyler H Shaw. 2021. Measurement of trust in automation: A narrative review and reference guide. *Frontiers in psychology* 12 (2021), 604977.

[30] Vijay Krishna and John Morgan. 2001. A Model of Expertise*. *The Quarterly Journal of Economics* 116, 2 (05 2001), 747–775. https://doi.org/10.1162/00335530151144159 arXiv:https://academic.oup.com/qje/article-pdf/116/2/747/5375310/116-2-747.pdf

[31] Justin Kruger and David Dunning. 1999. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology* 77, 6 (1999), 1121.

[32] Sandeep Kaur Kuttal, Bali Ong, Kate Kwasny, and Peter Robe. 2021. Trade-Offs for Substituting a Human with an Agent in a Pair Programming Context: The Good, the Bad, and the Ugly. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 243, 20 pages. https://doi.org/10.1145/3411764.3445659

[33] John D. Lee and Neville Moray. 1994. Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies* 40, 1 (1994), 153–184. https://doi.org/10.1006/ijhc.1994.1007

[34] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors* 46, 1 (2004), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 arXiv:https://doi.org/10.1518/hfes.46.1.50_30392 PMID: 15151155.

[35] Stephan J Lemmer, Anhong Guo, and Jason J Corso. 2023. Human-Centered Deferred Inference: Measuring User Interactions and Setting Deferral Criteria for

Human-AI Teams. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 681–694.

[36] Stephan Lewandowsky, Michael Mundy, and Gerard Tan. 2000. The dynamics of trust: comparing humans to automation. *Journal of Experimental Psychology: Applied* 6, 2 (2000), 104.

[37] Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. 2024. ChatGPT in healthcare: A taxonomy and systematic review. *Computer Methods and Programs in Biomedicine* 245 (2024), 108013. https://doi.org/10.1016/j.cmpb.2024.108013

[38] Brady D Lund and Ting Wang. 2023. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Library Hi Tech News* 40, 3 (2023), 26–29.

[39] Stephanie M Merritt, Alicia Ako-Brew, William J Bryant, Amy Staley, Michael McKenna, Austin Leone, and Lei Shirase. 2019. Automation-induced complacency potential: Development and validation of a new scale. *Frontiers in psychology* 10 (2019), 225.

[40] Stephanie M. Merritt and Daniel R. Ilgen. 2008. Not All Trust Is Created Equal: Dispositional and History-Based Trust in Human-Automation Interactions. *Human Factors* 50, 2 (2008), 194–210. https://doi.org/10.1518/001872008X288574 arXiv:https://doi.org/10.1518/001872008X288574 PMID: 18516832.

[41] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2023. Reading Between the Lines: Modeling User Behavior and Costs in AI-Assisted Programming. arXiv:2210.14306 [cs.SE]

[42] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (2023), 187–192. https://doi.org/10.1126/science.adh2586 arXiv:https://www.science.org/doi/pdf/10.1126/science.adh2586

[43] Changkun Ou, Sven Mayer, and Andreas Martin Butz. 2023. The Impact of Expertise in the Loop for Exploring Machine Rationality. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 307–321. https://doi.org/10.1145/3581641.3584040

[44] Oscar Oviedo-Trespalacios, Amy E Peden, Thomas Cole-Hunter, Arianna Costantini, Milad Haghani, J.E. Rod, Sage Kelly, Helma Torkamaan, Amina Tariq, James David Albert Newton, Timothy Gallagher, Steffen Steinert, Ashleigh J. Filtness, and Genserik Reniers. 2023. The risks of using ChatGPT to obtain common safety-related information and advice. *Safety Science* 167 (2023), 106244. https://doi.org/10.1016/j.ssci.2023.106244

[45] Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human factors* 52, 3 (2010), 381–410.

[46] Raja Parasuraman and Victor Riley. 1997. Humans and Automation: Use, Misuse, Disuse, Abuse. *Human Factors* 39, 2 (1997), 230–253. https://doi.org/10.1518/001872097778543886 arXiv:https://doi.org/10.1518/001872097778543886

[47] Snehal Prabhudesai, Leyao Yang, Sumit Asthana, Xun Huan, Q. Vera Liao, and Nikola Banovic. 2023. Understanding Uncertainty: How Lay Decision-Makers Perceive and Interpret Uncertainty in Human-AI Decision Making. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 379–396. https://doi.org/10.1145/3581641.3584033

[48] Stephen Rice, Sean R. Crouse, Scott R. Winter, and Connor Rice. 2024. The advantages and limitations of using ChatGPT to enhance technological research. *Technology in Society* 76 (2024), 102426. https://doi.org/10.1016/j.techsoc.2023.102426

[49] Steven I. Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D. Weisz. 2023. The Programmer's Assistant: Conversational Interaction with a Large Language Model for Software Development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 491–514. https://doi.org/10.1145/3581641.3584037

[50] Michele Salvagno, Fabio Silvio Taccone, Alberto Giovanni Gerli, et al. 2023. Can artificial intelligence help for scientific writing? *Critical care* 27, 1 (2023), 1–5.

[51] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (Marina del Ray, California) *(IUI '19)*. Association for Computing Machinery, New York, NY, USA, 240–251. https://doi.org/10.1145/3301275.3302308

[52] Herbert A. Simon. 1986. Rationality in Psychology and Economics. *The Journal of Business* 59, 4 (1986), S209–S224. http://www.jstor.org/stable/2352757

[53] Jiao Sun, Q. Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D. Weisz. 2022. Investigating Explainability of Generative AI for Code through Scenario-Based Design. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 212–228. https://doi.org/10.1145/3490099.3511119

[54] Teo Susnjak. 2022. ChatGPT: The End of Online Exam Integrity? arXiv:2212.09292 [cs.AI]

[55] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53.

[56] Jodie B. Ullman and Peter M. Bentler. 2003. Structural Equation Modeling. In *Handbook of psychology: Research methods in psychology*, Vol. 2. John Wiley and Sons, Inc., New York, NY, USA, 607–634. https://api.semanticscholar.org/CorpusID:53619206

[57] Priyan Vaithilingam, Tianyi Zhang, and Elena L. Glassman. 2022. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 332, 7 pages. https://doi.org/10.1145/3491101.3519665

[58] Tiffany C Veinot, Hannah Mitchell, and Jessica S Ancker. 2018. Good intentions are not enough: how informatics interventions can worsen inequality. *Journal of the American Medical Informatics Association* 25, 8 (05 2018), 1080–1088. https://doi.org/10.1093/jamia/ocy052 arXiv:https://academic.oup.com/jamia/article-pdf/25/8/1080/34150998/ocy052.pdf

[59] Krzysztof Wach, Cong Doanh Duong, Joanna Ejdys, Rūta Kazlauskaitė, Pawel Korzynski, Grzegorz Mazurek, Joanna Paliszkiewicz, and Ewa Ziemba. 2023. The dark side of generative artificial intelligence: A critical analysis of controversies and risks of ChatGPT. *Entrepreneurial Business and Economics Review* 11, 2 (2023), 7–30.

[60] Qiaosi Wang, Koustuv Saha, Eric Gregori, David Joyner, and Ashok Goel. 2021. Towards Mutual Theory of Mind in Human-AI Interaction: How Language Reflects What Students Perceive About a Virtual Teaching Assistant. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 384, 14 pages. https://doi.org/10.1145/3411764.3445645

[61] Justin D. Weisz, Michael Muller, Stephanie Houde, John Richards, Steven I. Ross, Fernando Martinez, Mayank Agarwal, and Kartik Talamadupula. 2021. Perfection Not Required? Human-AI Partnerships in Code Translation. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 402–412. https://doi.org/10.1145/3397481.3450656

[62] Justin D. Weisz, Michael Muller, Steven I. Ross, Fernando Martinez, Stephanie Houde, Mayank Agarwal, Kartik Talamadupula, and John T. Richards. 2022. Better Together? An Evaluation of AI-Supported Code Translation. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 369–391. https://doi.org/10.1145/3490099.3511157

[63] Christopher D Wickens, Benjamin A Clegg, Alex Z Vieane, and Angelia L Sebok. 2015. Complacency and automation bias in the use of imperfect automation. *Human factors* 57, 5 (2015), 728–739.

[64] Ziang Xiao, Q. Vera Liao, Michelle Zhou, Tyrone Grandison, and Yunyao Li. 2023. Powering an AI Chatbot with Expert Sourcing to Support Credible Health Information Access. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 2–18. https://doi.org/10.1145/3581641.3584031

[65] Yaqi Xie, Indu P Bodala, Desmond C. Ong, David Hsu, and Harold Soh. 2020. Robot Capability and Intention in Trust-Based Decisions across Tasks. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI '19)*. IEEE Press, Daegu, Republic of Korea, 39–47.

[66] Su-Fang Yeh, Meng-Hsin Wu, Tze-Yu Chen, Yen-Chun Lin, XiJing Chang, You-Hsuan Chiang, and Yung-Ju Chang. 2022. How to Guide Task-Oriented Chatbot Users, and When: A Mixed-Methods Study of Combinations of Chatbot Guidance Types and Timings. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 488, 16 pages. https://doi.org/10.1145/3491102.3501941

[67] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300509

[68] Zelun Tony Zhang, Cara Storath, Yuanting Liu, and Andreas Butz. 2023. Resilience Through Appropriation: Pilots' View on Complex Decision Support. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 397–409. https://doi.org/10.1145/3581641.3584056

[69] Albert Ziegler, Eirini Kalliamvakou, X. Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity Assessment of Neural Code Completion. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming* (San Diego, CA, USA) *(MAPS 2022)*. Association for Computing Machinery, New York, NY, USA, 21–29. https://doi.org/10.1145/3520312.3534864

A SURVEY QUESTIONS

A.1 Pre-Task Survey

In addition to the Trust in Automated Systems survey (*Appendix D*), we asked the following to gauge background and self-described expertise. Many questions were adopted from a company-internal longitudinal survey on engineering productivity and satisfaction.

Programming languages

(1) In the last three months, which languages have you used the most?
(2) Describe your level of familiarity in the following languages. (C++, Java, Python, Go, ...)

Knowledge resources

(1) In the past three months, how well have the following knowledge resources supported you in your development tasks? (Documentation, chat-based LLMs, forums, search tools, discussion spaces, ...)

Engineering satisfaction

(1) Overall, how satisfied are you with your experience as a developer at [company]?
(2) In the past three months, how productive have you felt at work at [company]?
(3) How often are you able to reach a high level of focus or achieve "flow" during development tasks?
(4) How satisfied are you with the quality of code that you produce?
(5) How satisfied are you with your engineering velocity?

LLM usage

(1) How often do you use LLM-assisted tools in your development tasks at work?
(2) If so, how well have LLM-assisted tools supported you in your development tasks in the past three months?
(3) Briefly describe any other ways LLM-assisted tools have supported you in your development tasks.
(4) Briefly describe any ways LLM-assisted tools have supported you in other tasks.

A.2 Post-Task Survey

In addition to the Trust in Automated Systems survey (*Appendix D*) and comparative satisfaction questions (*Table 2*), we asked the following to elicit free-form commentary.

(1) What was your approach for using non-LLM resources to answer questions?
(2) What was your approach for using Bard to answer questions?
(3) What was your approach for using non-LLM resources to verify your previous responses?
(4) What was your approach for using Bard to verify your previous responses?
(5) Which non-LLM resources did you use? How did they help?
(6) In the space below, please feel free to share any thoughts you have on the study.

B DESCRIPTIVE STATISTICS

The following statistics about our sample are taken from both company-internal data and participant-reported pre-task survey responses.

C EXPERTISE PERCENTILE RANKS

Expertise [30] is a multi-dimensional construct. *Table 7* shows summary statistics of expertise measures taken from company-internal data. To simplify the analysis, we constructed a joint objective measured expertise metric, weighting the following measures in decreasing order:

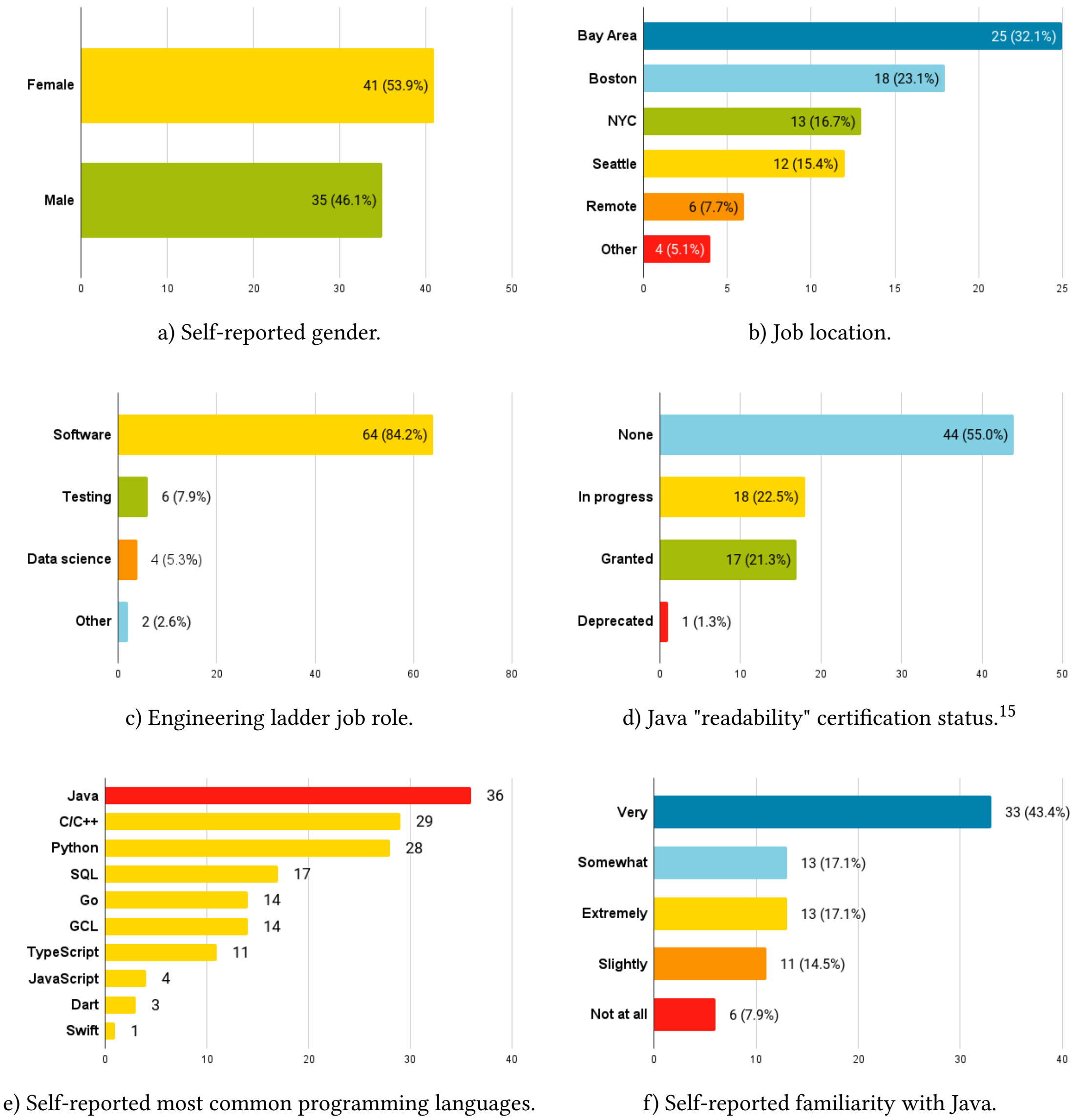
- Java readability certification and status (*Fig. 6d*), tiebreak by certifications in other languages
- Normalized number of Java changelists, tiebreak by changelists in other languages
- Most recent submitted Java, tiebreak by most recently submitted code in other languages

We similarly create a *self-reported expertise* measure weighting self-reported Java experience (*Fig. 9e, 9f*), engineering productivity (Appendix A.1), and LLM expertise in decreasing order. These measures are then transformed into percentile ranks, which yields more robust estimates by reducing the influence of outliers [13]. The position of each individual's expertise score in the distribution of scores is relative to all others in the primary analysis sample. Results are largely robust across most of the individual objective measures as well as the aggregated objective measure. Self-reported measures of expertise have low predictive power on the outcomes of interest.

D TRUST IN AUTOMATED SYSTEMS SURVEY

We administered Jian's Trust in Automated Systems survey before and after the task. Each statement is assessed on a 5-point, bipolar Likert scale with the possible options: *Strongly disagree, Somewhat disagree, Neutral, Somewhat Agree, Strongly Agree.* The chart below shows frequencies of each option. To calculate the means and regression coefficients for the comparative satisfaction survey in *Table 2* with the same Likert responses, we map these options to numeric values [-2, -1, 0, 1, 2].

[15] Those with *In progress* or *Granted* readability status have already taken this exam. Those with *Deprecated* readability withdrew from the process of obtaining readability due to failure to take the exam.



a) Self-reported gender.

b) Job location.

c) Engineering ladder job role.

d) Java "readability" certification status.[15]

e) Self-reported most common programming languages.

f) Self-reported familiarity with Java.

Figure 6: Categorical statistics (n=76).

	μ	σ	Min	Max
Java expertise				
# lines of code (Java)	41,718	151,850	0	1,197,572
# submitted changelists (Java)	157		0	2452
Most recent submitted Java	October 2022	-	June 2016	June 2023
Coding expertise				
# lines of code (all languages)	268,858	680,043	130	4,054,273
# submitted changelists (all)	777	1306	9	7132
Least recent submitted code	October 2019	-	March 2007	May 2023
Most recent submitted code	May 2023	-	September 2021	June 2023
# of readability certifications	1	0.87	0	4

Table 7: Summary statistics on objective expertise scores.

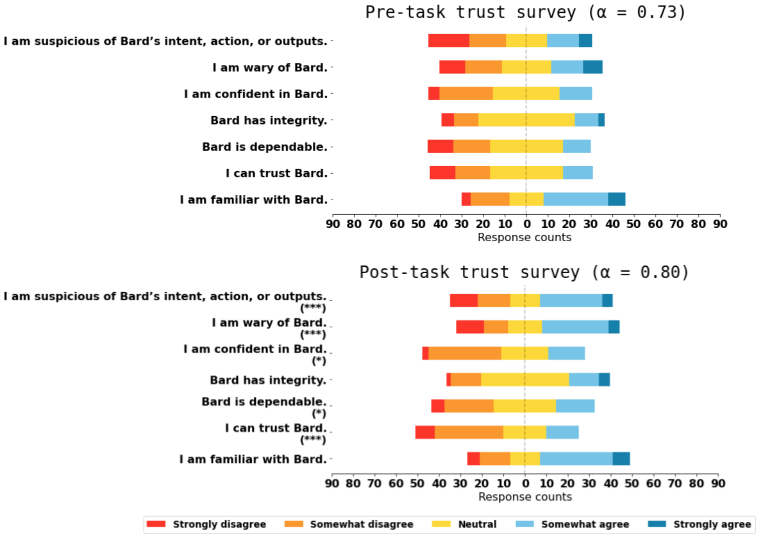


Figure 7: Pre- and post- task survey responses. α is Cronbach's alpha, a reliability measurement [55]. Asterisks shows that the pre- and post- task distributions are significantly different, calculated using a χ^2 independence test.