

Evaluation of Task Specific Productivity Improvements Using a Generative Artificial Intelligence Personal Assistant Tool

Brian S. Freeman^{1, *}, Kendall Arriola², Dan Cottell², Emmett Lawlor³, Matt Erdman², Trevor Sutherland⁴, Brian Wells⁴

¹Advanced Technologies, Trane Technologies, Davidson NC, USA

²Data & Analytics, Trane Technologies, Davidson NC, USA

³Engineering, Trane Technologies, Galway, Ireland

⁴AI Foundry, Trane Technologies, Davidson NC, USA

Email address:

brian.freeman@tranetechnologies.com (Brian S. Freeman)

*Corresponding author

To cite this article:

Brian S. Freeman, Kendall Arriola, Dan Cottell, Emmett Lawlor et al. (2024). Evaluation of Task Specific Productivity Improvements Using a Generative Artificial Intelligence Personal Assistant Tool. *American Journal of Artificial Intelligence*, 8(2), 68-80.

<https://doi.org/10.11648/j.ajai.20240802.16>

Received: 29 October 2024; **Accepted:** 14 November 2024; **Published:** 18 December, 2024

Abstract: This study evaluates the productivity improvements achieved using a generative artificial intelligence personal assistant tool (PAT) developed by Trane Technologies. The PAT, based on OpenAI's GPT 3.5 model, was deployed on Microsoft Azure to ensure secure access and protection of intellectual property. To assess the tool's productivity effectiveness, an experiment was conducted comparing the completion times and content quality of four common office tasks: writing an email, summarizing an article, creating instructions for a simple task, and preparing a presentation outline. Sixty-three (63) participants were randomly divided into a test group using the PAT and a control group performing the tasks manually. Results indicated significant productivity enhancements, particularly for tasks involving summarization and instruction creation, with improvements ranging from 3.3% to 69%. The study further analyzed factors such as the age of users, response word counts, and quality of responses, revealing that the PAT users generated more verbose and higher-quality content. Writing email content improved by 3.3%, summarizing text improved by 69%, creating instructions improved by 45.9%, and preparing an outline improved by 24.8%. An 'LLM-as-a-judge' method employing GPT-4 was used to grade the quality of responses, which effectively distinguished between high and low-quality outputs. The findings underscore the potential of PATs in enhancing workplace productivity and highlight areas for further research and optimization.

Keywords: Generative AI, Productivity, Employee Assistant, Task Analysis, AI Integration

1. Introduction

Trane Technologies developed a personal assistant tool (PAT) in the autumn of 2023 using a version of OpenAI's ChatGPT 3.5-Turbo-16k on Microsoft Azure cloud platform. Over the course of several weeks, the platform was introduced throughout all business units and international offices of the company and made available to all of its 40,000+ staff.

To evaluate the improvement in productivity this tool

provides, an experiment was initiated that compared the completion time and content of four common office tasks between a test group of users who could use the PAT and a control group of users who completed the tasks manually. The tasks included:

1. Writing an email to a supervisor
2. Summarizing an article
3. Creating instructions for a simple task
4. Preparing a presentation outline

2. Background

While large language models (LLMs) have been around since 2017 [1], their capabilities were not widely adopted until November 2022 when OpenAI announced the release of ChatGPT 3 [2]. Using generative pre-trained (GPT) models as personal assistants for text-based tasks became widespread, until several leaks of intellectual property were reported in the media and information on how the LLMs were trained using scraped content from the internet forced many companies to block access to Gen AI websites and services.

Realizing the need to allow controlled and secure access to GenAI tools, Trane Technologies developed an in-house tool based on OpenAI's GPT 3.5-Turbo-16k-0613 model but deployed on a Microsoft Azure platform that established usage guardrails and secured company intellectual property. Other companies followed suit, such as Procter & Gamble with their *chatPG* [3].

In order to quantify productivity improvement using PAT's, estimates of time savings are applied to usage metrics across the board, without consideration of the individual tasks. Estimations were based on published reports suggesting up to 54% efficiencies for certain office tasks could be achieved using a PAT [4].

The Harvard Business School polled 758 business consultants with access to ChatGPT-4 with 18 different tasks [5]. They found that use of AI allowed workers to complete 12.2% more tasks 25.1% faster with 40% better quality, with less experienced workers improving even more than experienced workers. The tasks were specific to business consulting such as marketing, branding, and product management. The study introduced the idea of retainment - how much of the content generated by the AI was included in the final submitted answer. The degree of retainment was linked to how users applied the AI. The two categories of use were defined as *Centaur* and *Cyborg* behavior. Centaur behavior is inspired by the legendary creature that combines human and horse traits. It involves a close collaboration between humans and machines, with users alternating between AI and human tasks based on their strengths and capabilities. They assess which tasks are best suited for humans and which can be effectively managed by AI. Cyborg behavior was inspired by science fiction hybrids and combines machine components with human biology for seamless integration. Users go beyond a simple division of labor, intertwining their efforts with AI at the cutting edge of capabilities. Both behaviors impact how individuals utilize AI and PATs.

The National Bureau of Economic Research reported improvements of 14% by customer service agents with higher increases by less-experienced workers compared to more-experienced workers [6]. One explanation of this enhancement is the reduction of required working memory load to handle large amounts of data or work under stressful conditions [7].

An MIT report published in the journal *Science* sampled 444 college educated professionals with specific writing tasks such as preparing a press release, writing a short report, analyzing

a plan, and writing a delicate email [4]. The participants' objective was to improve grades on their assignments and incentivized with cash rewards based on the points received. In addition to improving grade quality by 27%, average task time was improved by 54%. Noy et al. found that higher quality results improved at the cost of time needed to complete the task, and deduced that participants were spending the time editing the gen AI responses or modifying prompts to enhance responses. A key observation was that human input added value to the responses, even if it increased the overall time it took to complete the task. An interesting component about their methodology was that after users were given training on the AI and completed the first task, they were allowed to choose whether they wanted to use the AI for additional tasks. Follow-up questions were also provided that helped explain why gen AI PATs might not be used for some tasks. Respondents explained that the technology did not work well with specialized tasks or tasks that required up to date information to complete.

3. Methods

A web-based survey was prepared that included an introduction page that collected metadata on the respondent. After this data was collected the survey began. The time a respondent was on each page was measured and stored.

Two versions of the survey were prepared. The control version blocked the ability of the respondent to paste content from the operating system's clipboard (No Paste), while the test version allowed respondents to paste from the clipboard. This allowed test group users to access the PAT and copy the results from their prompts and paste into the survey collection window. Control group users had to type their responses into the survey collection windows. Control group users could theoretically access the PAT and use it to generate a response, however no instance of this was identified.

The surveys were written in ReactJS and Tailwind, with the results stored in a Postgress database. Results were downloaded into a comma separated value (csv) file for analysis. Respondents were randomly given one of the two URLs from 15-30 April 2024. Initially, the respondents were limited to individuals attending an on-site event at the corporate headquarters in Davidson, NC. Survey participants used their own laptops in reserved rooms with a monitor to explain how to access the survey and respond to specific questions. Later, URLs were emailed to volunteers in the company who were not in attendance. The respondents could ask questions through email and chat services.

4. Survey Content

Tasks were selected that were considered common for all users and did not need specific skills or subject matter expertise. These included:

1. Task 1: Writing an email to a supervisor (Email)

2. Task 2: Summarizing an article (Summary)
 3. Task 3: Creating instructions for a simple task (Instructions)
 4. Task 4: Preparing a presentation outline (Outline)
- The actual tasks are provided in Appendix I.

3. Beginner - Regular use (at least weekly) of a PAT, but only the web interface
4. Expert - Regular use of a PAT and gen AI model using API interfaces

Respondents readily identified themselves with a category without additional prompting.

5. Excluded Tasks

Tasks that included specific knowledge such as coding or third party tools such as translations were not considered due to access to limited sub population groups within the company. Current use of the company PAT suggests that over 55% of all prompts in the company are code related. Further research in this area as it related to company requirements should be considered. Current studies suggest coding tasks can be completed 56% faster using gen AI resources [8], while another study found that over 52% of PAT or Co-pilot assisted code contained errors [9].

6. Metadata

Basic metadata was collected on all respondents as shown in Table 1. All results were anonymous.

While most of the metadata parameters are evident, the *Experience with AI* parameter was broken down as follows:

1. None - No previous experience using the PAT
2. Some - Some experience using the PAT, but not weekly

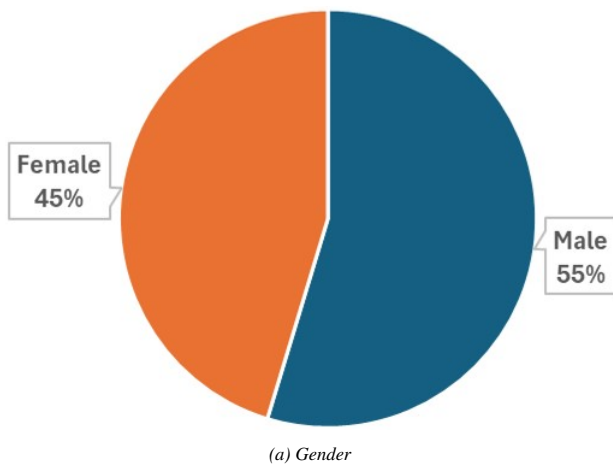


Table 1. Required metadata and selection options.

| Parameter | Options |
|--------------------|---------------------------|
| Age | |
| Gender | |
| Education | HS/Bachelor/Masters/PhD |
| Company Position | |
| Location | |
| Work Location | Remote/Hybrid/On-Site |
| Experience with AI | None/Some/Beginner/Expert |

7. Results

Sixty-three (63) respondents provided input to this study. Thirty-two (32) were part of the Test Group (could paste from clipboard) and thirty-one (31) were part of the Control Group and could not paste. Their breakdowns are shown below.

7.1. Metadata Results

There were slightly more male respondents than females with a median age of 42 years as shown in Figure 1.

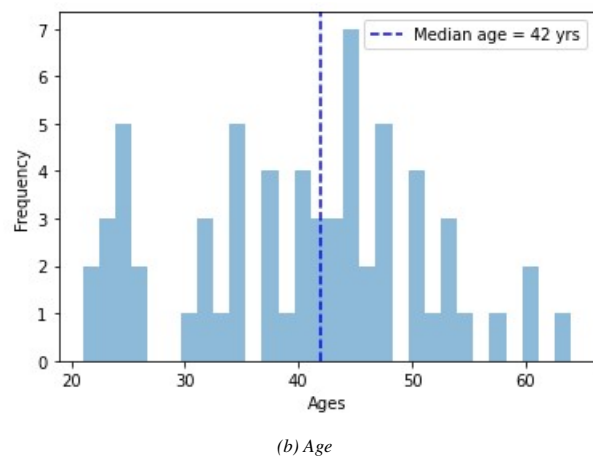


Figure 1. Respondent gender and age.

Almost all of the respondents had college degrees, with almost half of them having graduate degrees. This might be attributed to the specialized population participating at the internal company data conference where the survey was

given. This also could explain the large number of individuals identifying as Beginner level (having regular exposure to the PAT prior to taking the survey) as shown in Figure 2.

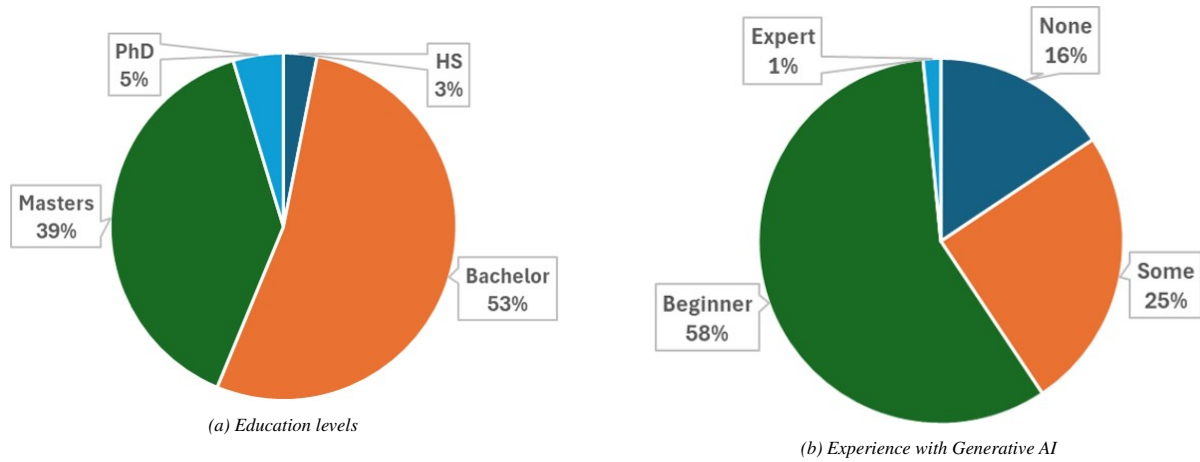


Figure 2. Respondent education and experience with gen AI.

Job positions were manually entered by the respondent resulting in 51 different position names. In order to classify each position into more manageable and comparable categories, the list of position titles was categorized by the PAT using the prompt:

User message = "Given this list of position titles, please group into no more than 5 different categories: << list of position names >>"

The resulting categories were:

1. Quality Management System (QMS)

2. Data and Analytics

3. IT and Software Development

4. Finance and Pricing

5. Project and Operations Management

Each category was assigned to the position titles given by the respondents. Figure 3 shows that the majority of participants were in Data & Analytics or Project Management positions. Seven of the respondents were interns or junior staff representing new employees.

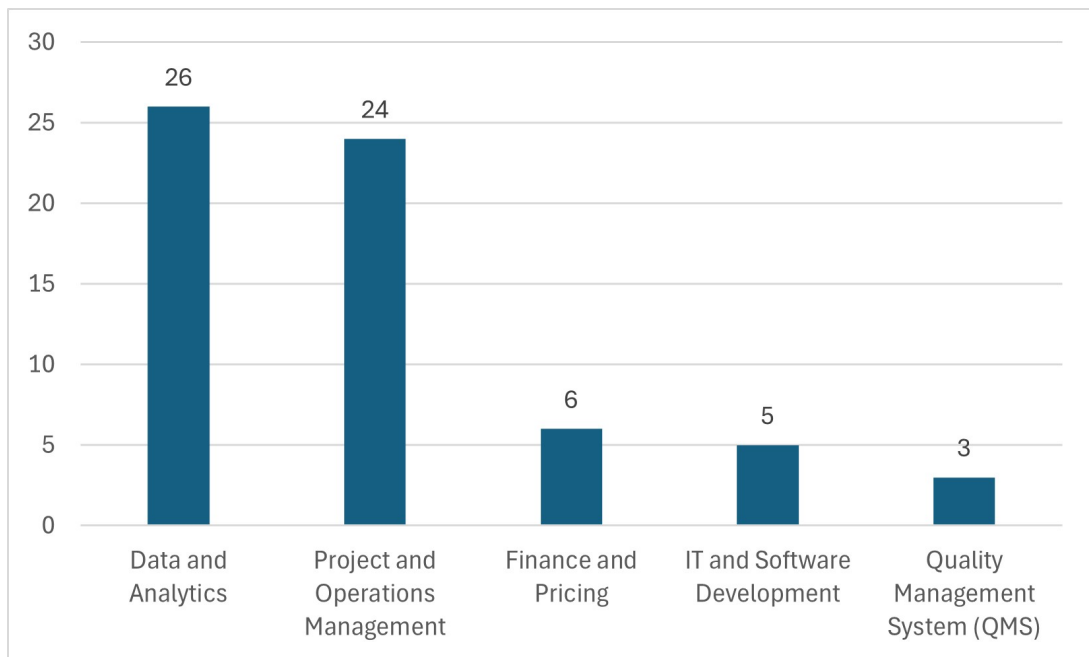


Figure 3. Respondent roles.

7.2. Task Results

The aggregated results are shown in Figure 4 for each task. The figures include all datasets without segregation by subgroups.

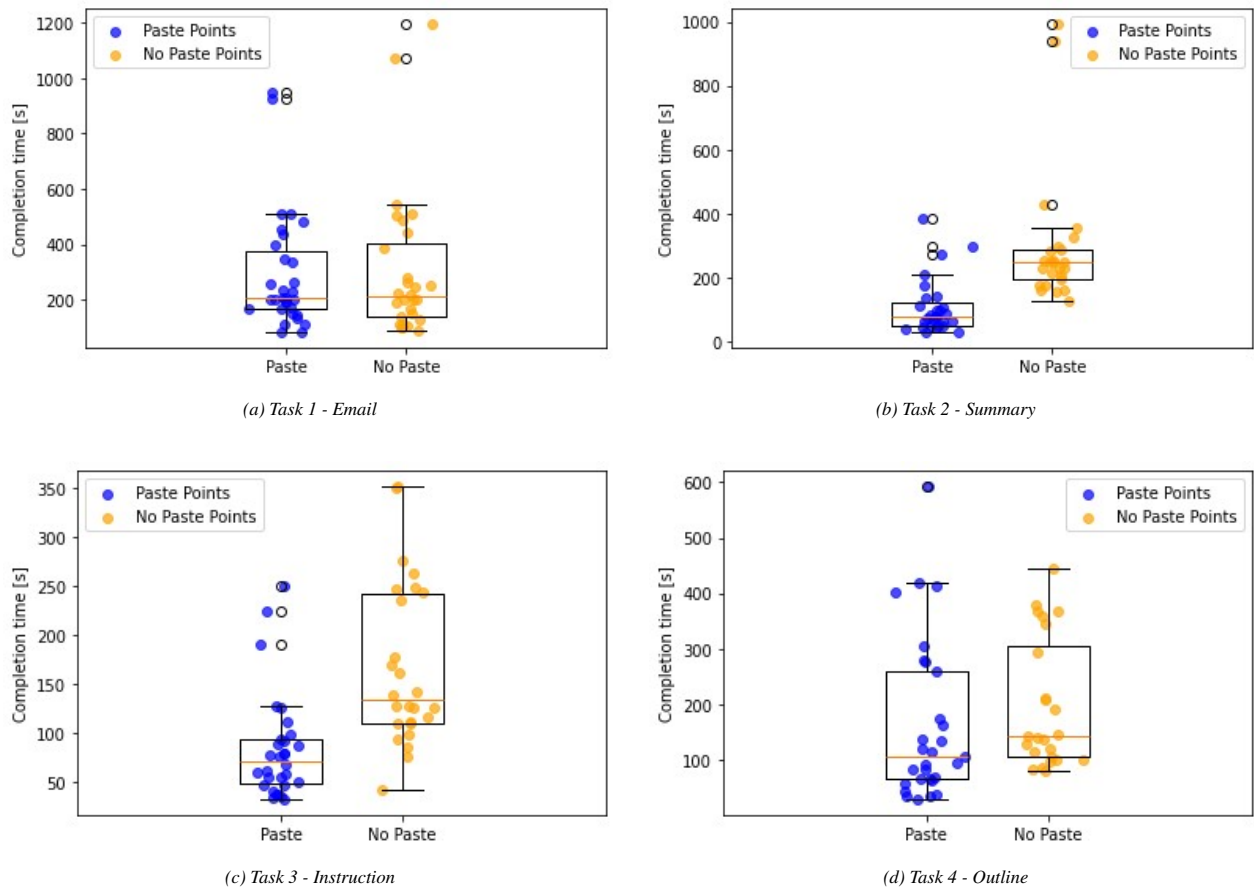


Figure 4. Aggregated results for tasks.

A review of the distributions of the aggregated data in each task suggested that all subsets were heavily skewed and not normally distributed. The median of each subset was used for descriptive purposes and shown in Table 2.

Table 2. Median results for each task.

| Task | No Paste [s] | Paste [s] | Improvement |
|-------------|--------------|-----------|-------------|
| Email | 211 | 204 | -3.3% |
| Summary | 248 | 77 | -69.0% |
| Instruction | 133 | 72 | -45.9% |
| Outline | 141 | 106 | -24.8% |

Reviewing the Improvement results in Table 2 shows a wide variation of productivity enhancements, from 3.3% for email

generation to 69% for summarization of text.

7.2.1. Statistical Evaluation of Aggregated Results

To determine statistical significance of the aggregated results, a non-parametric test was used to account for the non-normal distribution of the data. In this case the Mann-Whitney U test (also known as the Wilcoxon rank sum test) was used [10]. The null hypothesis (H_0) and alternative hypothesis (H_1) were defined as:

1. H_0 : There is no difference between the Control Group and the Test Group.
2. H_1 : There is a significant difference between the Control Group and the Test Group.

with a level of significance of $p = 0.05$ (5%). The results of the test are shown in Table 3.

Table 3. Mann-Whitney U Test for significance ($p = 0.05$).

| Task | U Statistic | p-value | Result |
|-------------|-------------|----------|---|
| Email | 426.5 | 0.915 | Fail to reject (Not Significant - Strong) |
| Summary | 626.5 | 8.73E-07 | Reject (Significant - Strong) |
| Instruction | 658.5 | 1.06E-05 | Reject (Significant - Strong) |
| Outline | 453 | 0.062 | Fail to Reject (Not Significant - Weak) |

Table 3 shows that the Email and Outline tasks do not pass and the null hypothesis should not be rejected (no significant differences between the groups). However, the p-value for the Outline task is very close to the critical value of 0.05 and could be assumed to be significant (reject null). Both the Summary

and Instruction tasks have strong differences from the null hypothesis and can safely be considered significantly different. The interquartile range shown in the box-plots of Figure 4 are shown in Table 4 for 25% and 75% percentiles.

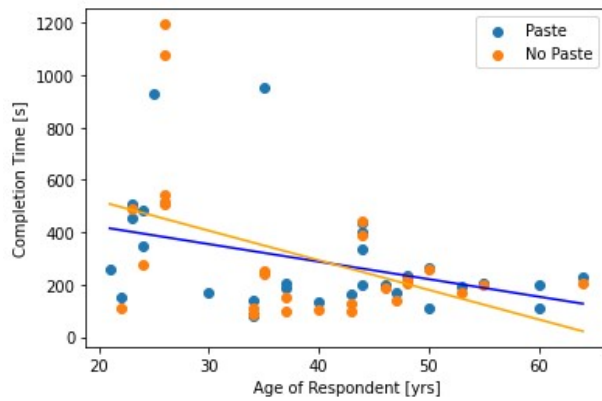
Table 4. Lower (25%) and upper (75%) quartiles in Figure 4.

| Task | No Paste [s] | | Paste [s] | |
|-------------|----------------|----------------|----------------|----------------|
| | 25% percentile | 75% percentile | 25% percentile | 75% percentile |
| Email | 87 | 542 | 83 | 511 |
| Summary | 128 | 355 | 29 | 210 |
| Instruction | 41 | 352 | 32 | 128 |
| Outline | 80 | 445 | 29 | 419 |

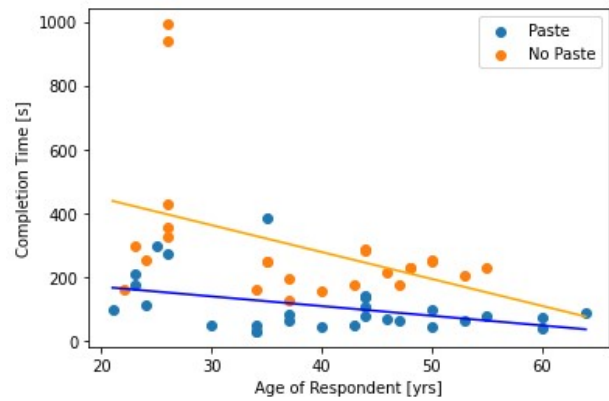
7.2.2. Evaluation of Age vs Completion Time

Determining how different users utilized the PAT was an important objective of the study. Physical age was used to indirectly represent user experience with the general assumption that the participant had similar experience

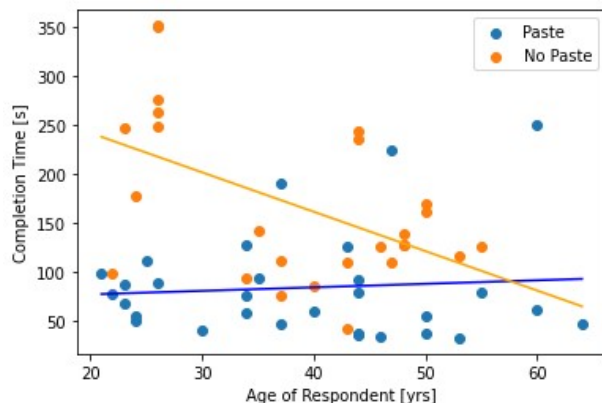
throughout their individual careers [11]. Figure 5 shows scatter plots of completion time to complete tasks by age of the participants with simple overlaid linear regression trend lines. The slopes and intercepts of the trend lines are given in Table 5.



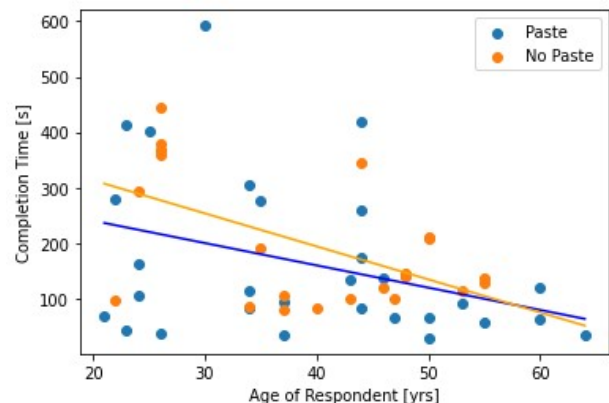
(a) Task 1 - Email



(b) Task 2 - Summary



(c) Task 3 - Instruction



(d) Task 4 - Outline

Figure 5. Evaluation of respondent's age vs completion time for tasks.

Spearman's Rank Correlation or Spearman's ρ [12] was used to show correlation between respondent age and completion time. This statistic is a non-parametric measure of correlation and commonly used when analyzing the

relationship between two variables that may not follow a normal distribution and is shown along with the trend line coefficients in Table 5.

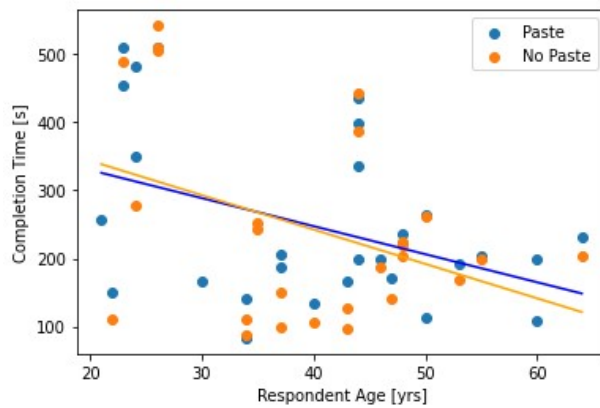
Table 5. Coefficients and correlations for trend lines in Figure 5.

| Task | Paste | | | No Paste | | |
|--------------|-------|--------|-------------------|----------|---------|-------------------|
| | Slope | Y-int | Spearman's ρ | Slope | Y-int | Spearman's ρ |
| Email | 0.26 | 212.26 | -0.086 | 3.24 | -41.66 | -0.182 |
| Summary | 0.06 | 101.31 | 0.113 | 8.42 | -100.41 | 0.181 |
| Instructions | -0.09 | 106.00 | 0.092 | 1.39 | 82.63 | -0.407 |
| Outline | 0.28 | 101.83 | 0.141 | 2.00 | 103.32 | -0.038 |

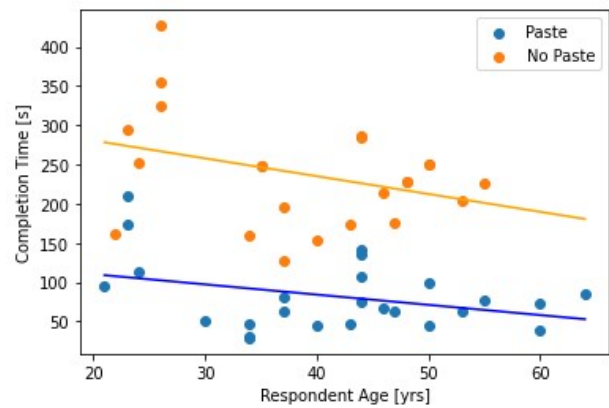
Correlation is not strong, with the highest being a negative correlation for Task 3 (Instructions) for the No Paste scenario.

Overall trends suggest decreasing completion times the older and more experienced the user was. This was especially evident for the Control Group (No Paste) with all tasks

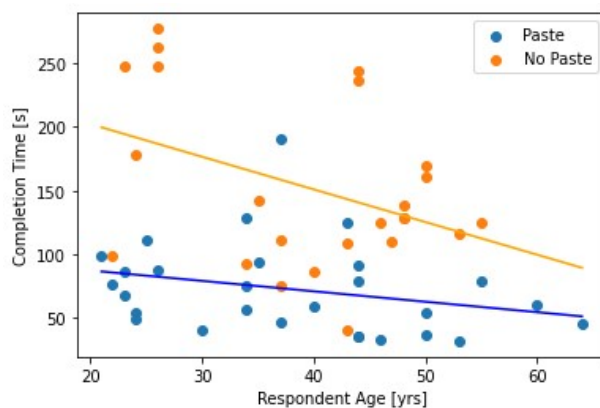
experiencing negative slopes on linear regression trend lines except for Task 3. In this case, the Test Group (Paste) showed a slight positive trend based on respondent age. For all cases, the No Paste scenario had steeper rates than the Paste scenarios.



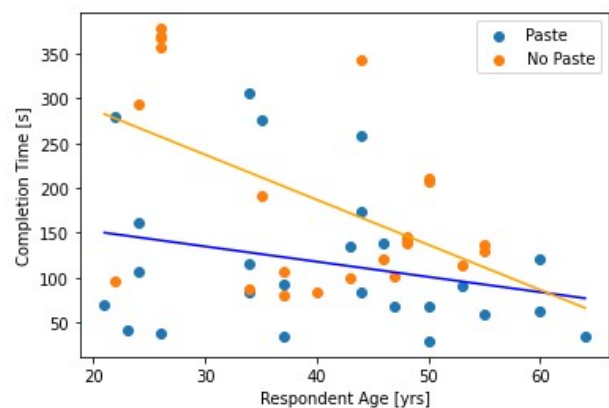
(a) Task 1 - Email



(b) Task 2 - Summary



(c) Task 3 - Instruction



(d) Task 4 - Outline

Figure 6. Evaluation of respondent's age vs completion time for tasks (filtered).

Concerns that some of the extreme points were influencing the linear regression lines led to evaluation of the main clusters for comparison. The Modified Z-Score test was used to

identify outliers. Unlike the traditional Z-Score method, which assumes data follows a normal distribution, the Modified Z-Score method leverages the median and the Median Absolute

Deviation (MAD), making it more resilient to the influence of outliers and suitable for non- normally distributed samples. The method calculates the modified z-scores for each data point by subtracting the median and then dividing by the MAD, scaled by a constant factor of 0.6745. This scaling factor ensures that the modified z-scores approximate standard z-scores for normally distributed data. Data points with an absolute modified z-score exceeding a chosen threshold of 3.5 for this case [13]. The filtered datasets are shown in Figure 6 with updated coefficients and Spearman's correlations in Table

6.

Filtering out extreme points shows more distinct clustering between the Paste and No Paste scenarios, especially with Task 2. Negative trends are present for all tasks and scenarios, with the No Paste slopes slightly steeper for all tasks compared to Paste scenarios. One explanation is that older respondents are also more experienced and can accomplish these routine tasks manually faster than less experienced respondents. Filtering also improves correlation ranking in general, with strong negative correlation present in Task 3 for the Paste scenario.

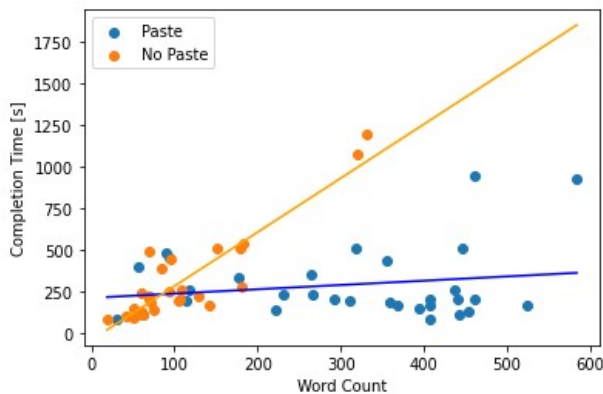
Table 6. Coefficients and correlations for filtered trend lines in Figure 6.

| Task | Paste | | | No Paste | | |
|--------------|-------|--------|-------------------|----------|-------|-------------------|
| | Slope | Y-int | Spearman's ρ | Slope | Y-int | Spearman's ρ |
| Email | -4.12 | 412.2 | -0.213 | -5.05 | 444.5 | -0.156 |
| Summary | -1.3 | 136.42 | -0.140 | -2.27 | 326 | -0.211 |
| Instructions | -0.82 | 103.7 | -0.910 | -2.56 | 253.4 | -0.156 |
| Outline | -1.7 | 185.8 | -0.298 | -5.04 | 388.4 | -0.175 |

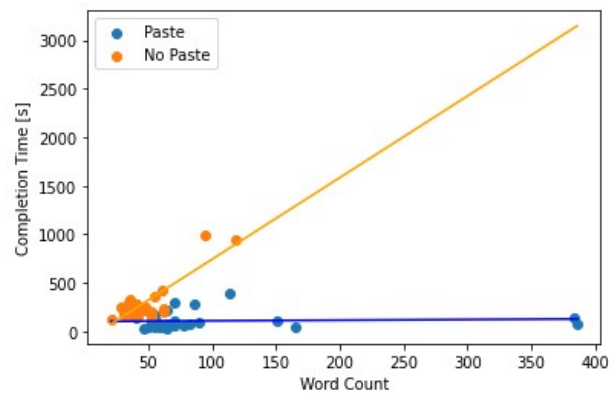
7.2.3. Evaluation of Response Word Counts

Another factor that was considered was how strong were the individual responses for both groups. In order to consider this,

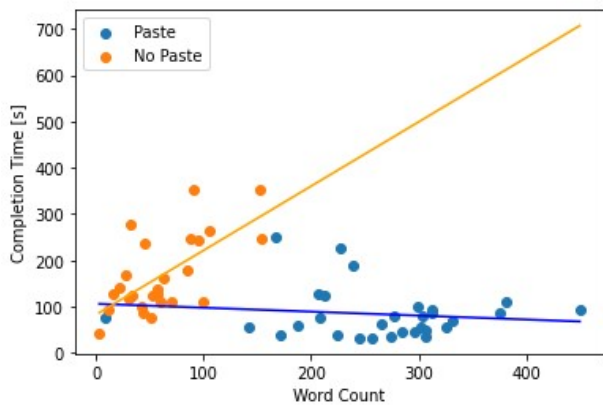
word counts for each task were evaluated. Word counts for each task and scenario are shown in Figure 7 with coefficients of the trend lines in Table 7.



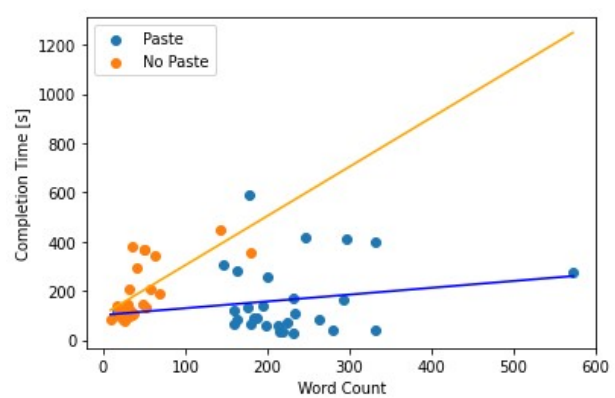
(a) Task 1 - Email



(b) Task 2 - Summary



(c) Task 3 - Instruction



(d) Task 4 - Outline

Figure 7. Evaluation of respondent's word count vs completion time for tasks.

Correlations in Table 7 were calculated using the same method as described in Table 5 with Spearman's ρ . Overall correlations are better than the previous section, especially for No Paste scenarios.

Table 7. Coefficients and correlations of trend lines in Figure 7.

| Task | Paste | | | No Paste | | |
|--------------|-------|--------|-------------------|----------|---------|-------------------|
| | Slope | Y-int | Spearman's ρ | Slope | Y-int | Spearman's ρ |
| Email | 0.26 | 212.26 | 0.029 | 3.24 | -41.66 | 0.813 |
| Summary | 0.06 | 101.31 | 0.307 | 8.42 | -100.41 | 0.352 |
| Instructions | -0.09 | 106.00 | 0.017 | 1.39 | 82.63 | 0.518 |
| Outline | 0.28 | 101.83 | -0.012 | 2.00 | 103.64 | 0.711 |

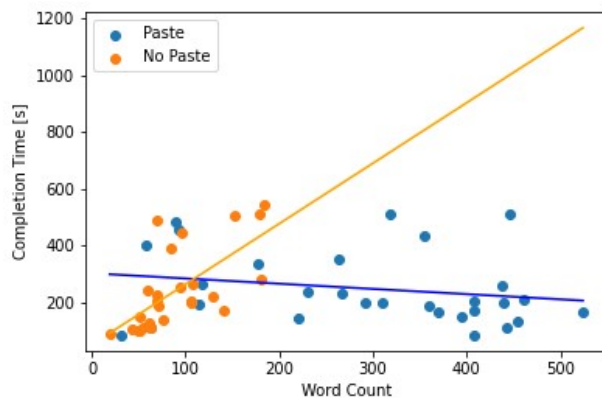
All tasks show similar patterns. The Paste scenario have more verbose answers with little fluctuation of completion time, while the No Paste scenario follow a positive correlation between word count and completion time with samples clustered around a central tendency. The median word counts for each task and scenario are shown in Table 8. A five-fold increase in text should be expected for most PATs, consistent with verbose reports in other studies [4].

As with concern of extreme values in the previous section comparing age to completion time, similar filtering was accomplished to evaluate the influence of outliers to the main

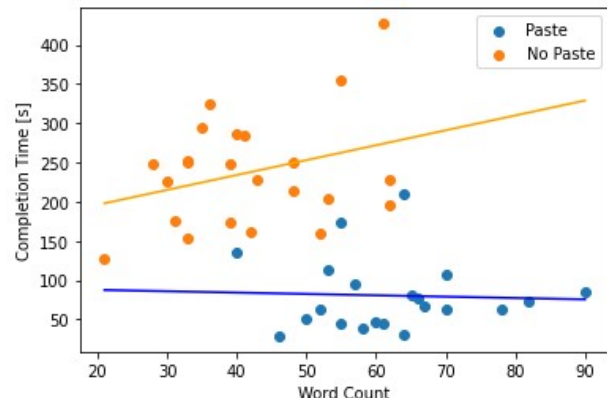
clusters. A similar approach, using the Modified Z-Score Test, was used. Results are shown in Figure 8 with coefficients of the trend lines in Table 9.

Table 8. Median word counts for each task and scenario.

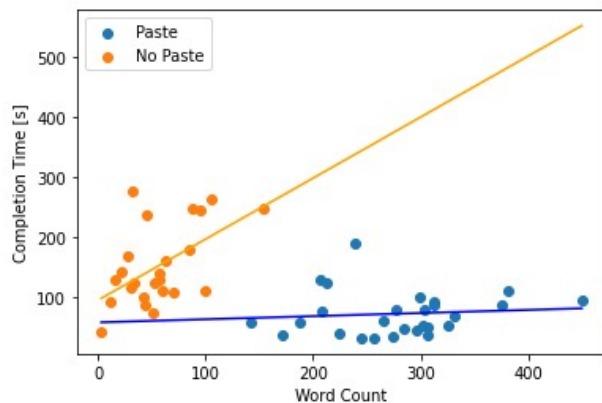
| Task | No Paste | Paste | Increase |
|-------------|----------|-------|----------|
| Email | 81 | 355 | 438% |
| Summary | 41 | 66 | 161% |
| Instruction | 55 | 276 | 502% |
| Outline | 36 | 213 | 592% |



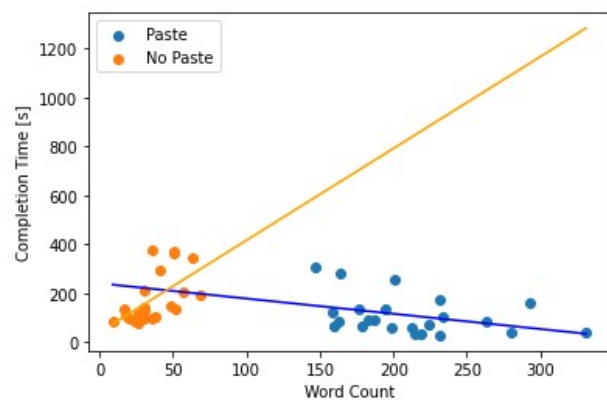
(a) Task 1 - Email



(b) Task 2 - Summary



(c) Task 3 - Instruction



(d) Task 4 - Outline

Figure 8. Evaluation of respondent's word count vs completion time for tasks (filtered).

Figure 8 shows more distinct clustering between the Paste/No Paste scenarios for each task, but reinforces the general patterns from Figure 7 in that the No Paste scenario

tasks have a high word count rates but within a very short range (not exceeding 100 words for most tasks) while Paste scenarios generate more than 5x the words.

Table 9. Coefficients and correlations of filtered trend lines in Figure 7.

| Task | Paste | | | No Paste | | |
|--------------|-------|-------|-------------------|----------|-------|-------------------|
| | Slope | Y-int | Spearman's ρ | Slope | Y-int | Spearman's ρ |
| Email | -0.18 | 301.4 | -0.175 | 2.13 | 50.1 | 0.766 |
| Summary | -0.17 | 91.2 | 0.120 | 1.9 | 158.2 | 0.160 |
| Instructions | 0.05 | 58.40 | 0.193 | 1.02 | 95.1 | 0.415 |
| Outline | -0.62 | 241 | -0.364 | 3.74 | 42.1 | 0.654 |

7.2.4. Evaluation of Response Quality

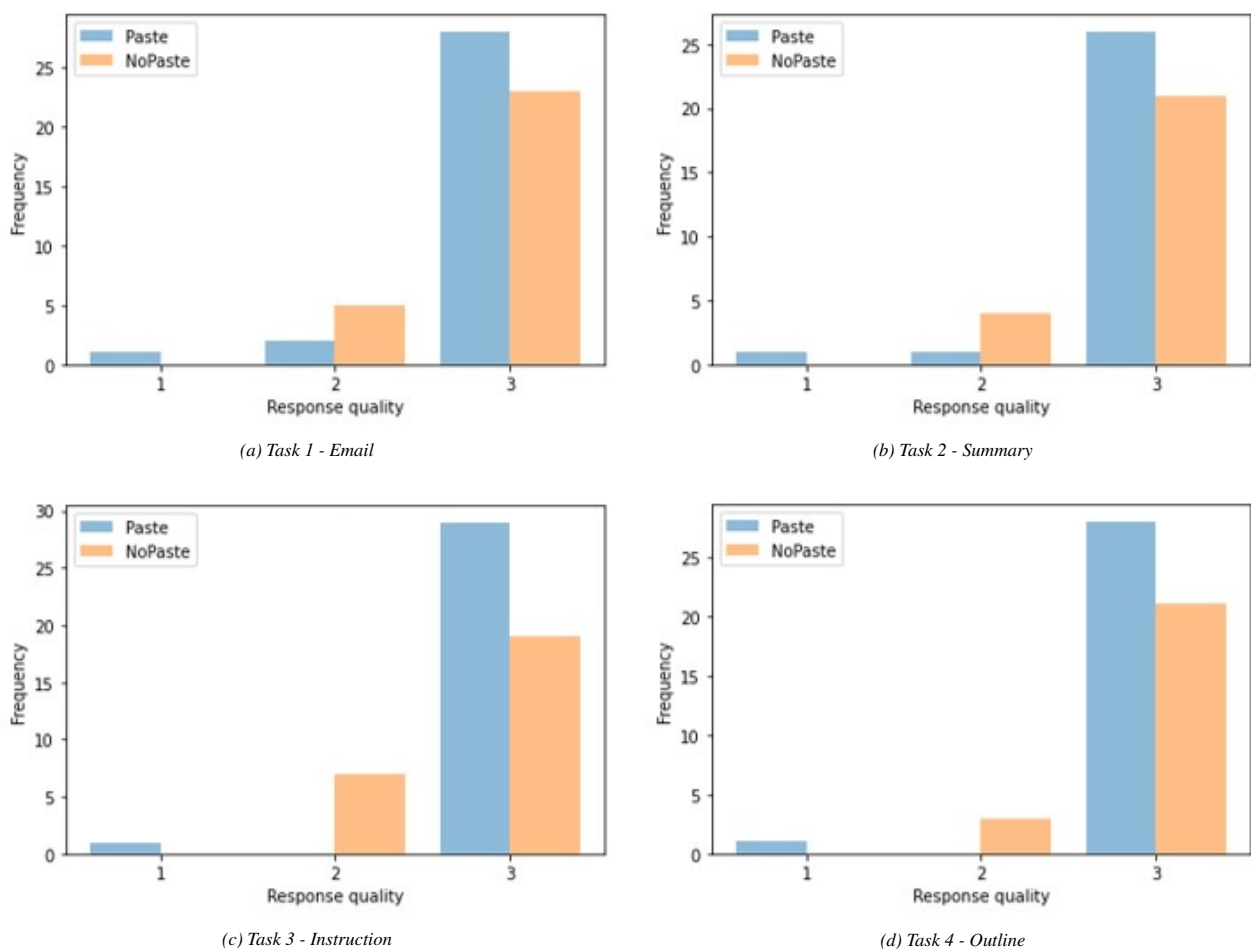


Figure 9. Evaluation of respondent's writing quality for tasks.

Multiple methods were provided by literature to evaluate response quality including Google's Universal Sentence Encoder [14] and LIWC [15]. However, a zero-shot 'LLM-as-a-judge' method was used employing OpenAI's instance of GPT-4 on Microsoft Azure to act as a grader for the content. The GPT temperature was set to 0.0 to reduce hallucination effects and provide result reproducibility [15]. Each response, both No Paste and Paste scenario, was given to the GPT in the

form of a prompt along with the task and a grading method. The prompt looked like this:

System message = "You are an English language expert who will grade a response to a question. Give a grade of 1 for a response that does not answer the question. Give a grade of 2 for a response that answers the question with

spelling and grammar errors. Give a grade of 3 for a response that answers the question with good English. Only provide a response that is 1, 2, or 3."

User message = "The question is: task. The user response is response"

The result was a numerical grade from 1 to 3 for each task response [16, 17]. The review was run three times and an integer average of the three runs was used as the grade. Using a GPT-4 model to evaluate GPT-3.5 results from the PAT was assumed. Evaluating GPT-3.5 results from the PAT using a GPT-4 model was considered appropriate due to the significant differences in training sets and methodologies between the two models, ensuring sufficient separation for objective evaluation.

The GPT-4 reviewer correctly identified incomplete responses which were removed from the analysis pool. The remaining response samples were relatively close to each other, with predominately more lower-quality results from the No Paste respondents than the Paste group.

8. Discussion

Evaluation of the respondents' results looked at overall completion time to complete a task and then evaluating the task

$$t_{paste} = t_{ingest} + 2t_{navigate} + \sum_{i=1}^n (t_{prompt_i} + t_{edit_i} + t_{latency_i} + t_{response_i}) + t_{copy} + t_{paste} \quad (2)$$

where $2t_{navigate}$ is the time taken to navigate from the survey URL to the PAT URL, and then return to the survey URL after the PAT response has been copied. It can be assumed that the same time is required for both actions.

For other terms, t_{prompt} is the time used to write the prompt request for the PAT, $t_{latency}$ is the time the PAT uses to ingest the prompt, generate a response, and display the response, $t_{response}$ is the time to review the review the response, t_{copy} is the time needed to copy the response from the PAT window,

times against age, word count and quality of the responses. In general, the results reinforced observations from previous studies as well as highlighting some issues. Overall, using a PAT reduces the overall time to accomplish certain tasks while improving the quality of the output. Using PATs also generate responses that are more verbose.

Most interesting was the time required to write an email was almost the same between the Control and Test Groups. Although other studies acknowledged that users often spend time editing outputs or re-prompting to obtain better results, they did not evaluate how the final responses were achieved in terms of time spent and information retention.

Decomposing the different times in the total completion time for the Control Group that could not use the PAT (No Paste scenario) shows the following terms for $t_{nopaste}$:

$$t_{nopaste} = t_{ingest} + t_{transcribe} + t_{edit} \quad (1)$$

where t_{ingest} is the total time needed to read the task question and understand the requirements, $t_{transcribe}$ is the time needed to write the initial response in the content window, and t_{edit} is the time used to edit the response. In many cases $t_{edit} = 0$ when the the initial typed response is accepted. All terms are independent random variables.

For the Test group using the PAT, completion time, t_{paste} terms becomes more complicated.

and t_{paste} is the time needed to paste in the survey content window. The act of creating a prompt, editing it, waiting for the response and reviewing the response may be iterative in order to get a satisfactory result. To account for these iterations, n represents the number of iterations of prompting used. As with 1, all the variables are independent and random.

Assuming $2t_{navigate}$, t_{copy} , and t_{paste} are negligible, Eq 2 can be reduced to

$$t_{paste} = t_{ingest} + \sum_{i=1}^n (t_{prompt_i} + t_{edit_i} + t_{latency_i} + t_{response_i}) \quad (3)$$

Comparing the terms associated with Eq 1 and 3 it is clear that not using a PAT is less complicated in regards to user steps, especially for simple tasks. More complicated tasks or tasks that may require more working memory load, such as summarizing a large body of text, will benefit from using a PAT, even if the text uses specialty language.

9. Conclusions and Recommendations

A survey was prepared and presented to 63 employees of Trane Technologies to evaluate the productivity improvements

of a generative AI base personal assistant tool (PAT). Half of the participants received a link to a survey where they could paste the results of the survey tasks from the Trane Technology PAT - Employee Virtual Assistant into the survey, while the other half could not paste content and had to complete the tasks manually. The resulting analysis showed that improvements varied extensively by task:

1. Write an email: 3.3% improvement using a PAT
2. Summarize text: 69% improvement using a PAT
3. Create instructions: 45.9% improvement using a PAT
4. Prepare an outline: 24.8% improvement using a PAT

Further analysis showed that age of the user influenced overall time required to complete tasks, especially when summarizing a document without using a PAT. PATs were also shown to generate more verbose responses. This may ultimately result in longer emails, reports, and other text-based products, requiring more time to ingest by a user, or creating the need to use a PAT to complete more work tasks.

An '*LLM-as-a-judge*' method using a GPT-4 model was used to review content quality that assigned a grade to each response based on the task and completion response. The method was able to filter out bad responses and assign grades to each response. Users who could not paste from the PAT had slightly lower quality than PAT users.

The small sample size of the study compared to other studies (63 participants compared to 444 in the lowest of other studies) may have biased results by not capturing enough variance in task completion time, word count, and content quality. Nonetheless, the diversity of participants that were collected suggest that results may not vary significantly, and at the very least, validate the expectation that different administrative tasks have different associated productivity efficiencies.

9.1. Recommendations

This study looked at use cases at the nascent stage of PAT adoption. Evaluating productivity after a year of use where more people feel comfortable using the tools and have integrated them into work routines is recommended. Additionally, a similar study to see how well PATs impact coding and software development with light coders (data scientist, analysts, and engineers - users who use software and generate code but are not full time developers) should be conducted, especially since over 50% of the queries received by company's PAT are code related.

Abbreviations

| | |
|-----|---------------------------|
| AI | Artificial Intelligence |
| GPT | Generative Pre-Trained |
| LLM | Large Language Model |
| MAD | Median Absolute Deviation |
| PAT | Personal Assistant Tool |

ORCID

0000-0002-9760-063X (Brian S. Freeman)

Supplementary Information

Raw survey datasets are available in csv format by contacting the corresponding author.

Acknowledgments

This project was an initiative executed under the Trane Technology AI Foundry. We wish to thank Rebecca Wells and Stephanie Benson for their support and encouragement. Also thanks to Phillip Sime for technical support and access to cloud infrastructure.

Conflict of Interest

The authors declare no conflicts of interest.

Appendix

Appendix I: Survey Tasks

The following four tasks were presented in a web-based survey.

Appendix II: Task 1 - Email

Write an email to your supervisor requesting approval to go to the Data Analytics Summit. Include 3 reasons why he/she should approve it.

Appendix III: Task 2 - Summary

Summarize the following text:

1. Generative Artificial Intelligence (AI) has revolutionized machine learning by enabling computers to exhibit creativity and generate content. The inception of generative AI can be traced back to the 1950s and 1960s when computer scientists began experimenting with automated systems capable of producing original content. The 1980s saw significant progress through the development of algorithms and computational techniques, resulting in the creation of systems capable of generating content in various domains.
2. The 1990s marked a significant milestone with the introduction of recurrent neural networks (RNNs) by researchers like Sepp Hochreiter and Jurgen Schmidhuber. RNNs allowed machines to generate sentences, scripts, and even poetry, resulting in increasing excitement in the field. However, it was the early 2000s that witnessed a remarkable resurgence in generative AI. This resurgence was fueled by the rise of deep learning and the introduction of generative adversarial networks (GANs) by Ian Goodfellow and his colleagues in 2014.
3. Deep learning algorithms utilizing neural networks with multiple layers greatly enhanced the generative capabilities of computers. GANs, in particular, revolutionized the field by introducing a generator-discriminator architecture, dramatically improving the quality and realism of generated content. This breakthrough enabled the generation of high-resolution

images, music, and even human-like videos, opening up new possibilities and applications for generative AI.

4. Generative AI has found diverse applications across various industries. In creative fields such as art, design, and entertainment, it has pushed the boundaries of creativity and facilitated the generation of novel content. Industries like marketing and advertising have utilized generative AI to create personalized campaigns and generate content at scale, enhancing customer engagement.
5. Looking towards the future, researchers are continuously exploring methods to further enhance the diversity, coherence, and controllability of generated content. However, challenges, including ethical considerations related to the misuse of generative AI, need to be addressed. With ongoing innovations, the potential impact of generative AI on artistic expression, human-computer interaction, and society at large is vast, making it an exciting frontier in technology.

Word count = 323 words

Appendix IV: Task 3 - Instructions

Explain how to address a letter for postage.

Appendix V: Task 4 - Outline

Prepare an outline for a presentation that shows the major events from your last trip.

References

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017, 30, 5998-6008.
- [2] OpenAI. Introducing ChatGPT. [Internet]. Available from: <https://openai.com/index/chatgpt/>. [Accessed 22 May 2024].
- [3] Lindsey Wilkinson. How P&G rolled out its internal generative AI model. [Internet]. Available from: <https://www.ciodive.com/news/procter-gamble-PG-chatgpt-AI-openAI/697067/>. [Accessed 25 May 2024].
- [4] Shakked Noy, Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*. 2023, 381(6654), 187-192. <https://doi.org/10.1126/science.adh2586>
- [5] F. Dell'Acqua, E. McFowland III, E. Mollick, H. Lifshitz-Assaf, K. C. Kellogg, S. Rajendran, L. Kraye, F. Candelon, K. R. Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. Technical Report 24-013, Harvard Business School, September 2023.
- [6] Erik Brynjolfsson, Danielle Li, Lindsey Raymond. Generative AI at work. [Internet]. Available from: https://www.nber.org/system/files/working_papers/w31161/w31161.pdf
- [7] Jakob Nielsen. AI improves employee productivity by 66%. [Internet]. Available from: <https://www.nngroup.com/articles/ai-tools-productivity-gains/>. [Accessed 27 May 2024].
- [8] Sida Peng, Eirini Kalliamvakou, Peter Cihon, Mert Demirel. The impact of AI on developer productivity: Evidence from github copilot, 2023.
- [9] Samia Kabir, David Udo-Imeh, Bonan Kou, Tianyi Zhang. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions, 2024.
- [10] Patrick E. McKnight, Julius Najab. Mann-Whitney U Test. In *Methods in behavioral research*, Edition. John Wiley: New York, NY, USA; 2010, pp. 1-1. <https://doi.org/10.1002/9780470479216.corpsy0524>
- [11] M Morris, V Venkatesh. Age differences in technology adoption decisions: Implications for a changing work force. *Personnel Psychology*. 2006. <https://doi.org/10.1111/j.1744-6570.2000.tb00206.x>
- [12] C. Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*. 1904, 15(1), 72-101.
- [13] David C. Hoaglin. Volume 16: How to detect and handle outliers, 2013. [Internet]. Available from: <https://api.semanticscholar.org/CorpusID:208231456>
- [14] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. Universal sentence encoder. *CoRR*, abs/1803.11175, 2018. [Internet]. Available from: <http://arxiv.org/abs/1803.11175>
- [15] R. L. Boyd, A. Ashokkumar, S. Seraj, J. Pennebaker. The development and psychometric properties of LIWC-22. University of Texas at Austin, 2022. [Internet]. Available from: <https://www.liwc.app>
- [16] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, Ion Stoica. Judging LLM-as-a-judge with mt-bench and chatbot arena, 2023.
- [17] Quinn Leng, Kasey Uhlenhuth, Alkis Polyzotis. Best practices for LLM evaluation of RAG applications. [Internet]. Available from: <https://www.databricks.com/blog/LLM-auto-eval-best-practices-RAG>. [Accessed 7 June 2024].