Leveraging Generative AI for Enhancing Freelance Worker Efficiency: A Comparative Usability Study Using OfficeMind

Vahid Agbortoko

Summer-Fall 2024

Abstract

As gig economies continue to grow, there is an increasing need for innovative technologies that assist freelance workers in improving their productivity, communication, and self-management. This study explores the application of generative AI tools in the context of freelance work, focusing on enhancing communication between freelancers and their clients, improving work efficiency, and supporting self-marketing efforts. Through a structured usability study, we recruited 50 freelance workers via Upwork to participate in two distinct scenarios: a control group using conventional freelance tools, as well as popular AI tools like ChatGPT, and an intervention group utilizing a custom AI-based tool called OfficeMind. Participants were tasked with completing common freelance activities, including drafting biographies, emails, and personal work schedules. System Usability Scale (SUS) scores and task completion times were recorded for each group. Preliminary observations suggest that the AI-driven tool demonstrated potential in streamlining task completion, although participants highlighted areas for improvement in user experience. This study offers critical insights into how generative AI can be tailored to address the unique needs of gig workers, aiming to foster more effective collaboration between freelancers and technology.

1 Introduction

Freelancers constitute a growing segment of the global workforce, operating independently across various sectors. As technology evolves, artificial intelligence (AI) offers significant potential to enhance the daily workflows of these self-employed professionals. Generative AI, in particular, can streamline communication, improve task efficiency, and provide customized tools for freelancers to manage client interactions, personal schedules, and self-marketing efforts.

The objective of this study is to explore how generative AI can be adapted into assistive tools for freelance workers while comparing the usability and efficiency of a custom AI tool called OfficeMind with existing AI tools like ChatGPT. We specifically aim to investigate whether AI-driven platforms can offer tangible benefits in worker-client communication, task management, and overall productivity. Participants were recruited through Upwork, an online freelancing platform, and tasked with evaluating both conventional freelance tools and AI tools. Two user groups were established: a

control group, utilizing standard freelance tools and popular AI tools, and an intervention group, using OfficeMind to complete the same tasks.

By analyzing System Usability Scale (SUS) scores and task completion times between these two groups, we aim to gain a deeper understanding of the efficacy of AI-assisted tools in supporting the gig workforce. Participants engaged in a structured usability study, completing tasks such as drafting personal biographies, sending emails to clients, and organizing personal work schedules. These tasks represent common challenges freelancers face in their everyday work and were deliberately kept open-ended to allow participants flexibility, mimicking real-world scenarios.

Feedback was collected through interviews and surveys to assess user experience, identify areas for improvement, and explore the impact of generative AI on freelance work. By comparing the two groups, we aim to uncover insights on how AI may improve productivity and worker satisfaction, with implications for developing more advanced AI tools in the future. This study contributes to the growing body of research on human-centered AI and its role in enhancing the capabilities and efficiency of freelancers in the gig economy.

2 Related Work

The integration of artificial intelligence (AI) into gig work platforms has been a topic of increasing interest in recent years. Various tools have been developed to enhance the working conditions, professional growth, and earnings of gig workers, and this paper builds upon these advancements.

Improving Gig Workers' Labor Conditions: In a recent publication by the CIVIC AI Lab at Khoury College at Northeastern, the focus was placed on crowdworkers—the individuals behind many AI tasks, such as tagging, selecting, and identifying data for machine learning. The study highlights the development of tools designed to help gig workers quantify their labor conditions and unpaid labor time. Additionally, it discusses tools aimed at helping workers develop skills and increase wages, improving professional outcomes for those in the gig economy [1]. This paper draws upon these concepts by tailoring AI to aid in worker-client communications and increase overall worker efficiency.

AI-Driven Collective Action: Another study from our institution explores the use of AI-enhanced technologies to support collective action among gig workers on platforms such as Upwork and Amazon Mechanical Turk. The proposed systems, including "Gig-Sousveillance" and "GigAction," aim to empower gig workers by enabling them to monitor their workplace, strategize solutions, and implement collective action to improve labor conditions [2]. While this research focuses on collective action, our project narrows in on individual worker efficiency and communication through AI tools.

Career Development in Gig Work: The role of online technologies in career development has also been explored in earlier research by Smith [3], which examined how technologies shape freelance work and career trajectories. In particular, the paper highlighted the shift in responsibility for career development from employers to workers, with technologies like social media and freelance platforms playing an increasingly central role. This work is relevant to our research as we focus on how generative AI can assist gig workers in areas such as self-marketing and communication with clients.

Together, these studies provide a strong foundation for exploring the application of generative AI to enhance gig workers' client interactions and task efficiency. Our research

expands on these efforts by focusing on the practical integration of AI in real-world freelancing scenarios.

3 Methodology

3.1 Study Design

This study was designed as a comparison between a Control group using conventional freelance tools (and popular AI systems like ChatGPT) and an Intervention group using a tailored tool, OfficeMind, developed to assist freelancers. Both groups were given three tasks to complete: drafting a biography, composing an email, and creating a work schedule. There was no time limit for individual tasks; however, the entire session was capped at 60 minutes. All participants were informed that the session would be stopped after 60 minutes, regardless of whether all tasks were completed. Surveys were administered post-session and were not timed.

3.2 Participants

Participants were recruited from Upwork, a freelancing platform, and were compensated at an hourly rate. Eligible participants were fluent in English, at least 18 years old, and had prior freelancing experience. Recruitment targeted freelancers with diverse backgrounds to ensure a broad representation. A total of 50 participants completed the study, with 25 in each group (Control and Intervention). Recruitment was conducted through an Upwork job post, inviting freelancers to evaluate an AI tool aimed at improving task efficiency and communication with clients.

3.3 Tasks and Procedures

Participants were assigned three tasks:

- **Task 1**: Draft a personal biography suitable for uploading to a freelancing platform.
- **Task 2**: Compose an email that would be sent to a client on a freelancing platform.
- **Task 3**: Create a personal work schedule for completing freelancing tasks.

These tasks were deliberately designed to be open-ended, allowing participants to approach them independently. Both the control and intervention groups followed the same task order. The time taken to complete each task was recorded, though participants were not aware of the time tracking. Prior to starting the tasks, a brief trial run was conducted to familiarize participants with the tools, without time constraints.

3.4 Tool: OfficeMind

The Intervention group used OfficeMind, a customized AI-based tool designed to assist freelancers in completing tasks. The tool provided the following functionalities:

- **Biography Generator**: Suggested professional biography templates.
- **Email Composer**: Assisted in drafting formal emails with adjustable tone and style.

- **Scheduling Assistant**: Offered task scheduling templates and suggestions to streamline workload planning.

In contrast, the Control group was allowed to use their preferred tools, including popular AI systems such as ChatGPT and any other tool they typically used for freelancing tasks.

3.5 Data Collection and Analysis

Two primary types of data were collected: task completion times and System Usability Scale (SUS) scores. After completing the tasks, participants provided feedback through the SUS survey and additional qualitative questions regarding their experience.

- **Quantitative Analysis**: Descriptive statistics were calculated for task completion times and SUS scores. Hypothesis testing was conducted to determine if significant differences existed between the intervention and control groups in terms of task efficiency and user satisfaction.

- **Normality Tests**: Tests for normal distribution were conducted to guide the choice of parametric (e.g., t-test) or non-parametric tests (e.g., Mann-Whitney U test) for comparison.

- **Correlation Analysis**: A correlation analysis was planned to explore the relationship between task completion times and SUS scores. However, due to differences in dataset structure and the lack of a unique identifier for participants across datasets, this analysis was not performed.

4 Results and Discussion

4.1 SUS Scores Analysis

4.1.1 Descriptive Statistics

The mean SUS score for the Intervention group was 88.2, while the Control group had a slightly higher mean of 90.14. Both groups had a maximum SUS score of 100, but the Intervention group had a lower minimum score (55 compared to 67.5 for the Control group). The standard deviation was slightly higher for the Intervention group (12.32) than for the Control group (11.2), indicating more variability in the SUS scores within the Intervention group.

4.1.2 Normality Test Results

Normality tests were performed using the Shapiro-Wilk and Kolmogorov-Smirnov tests. Both groups failed the normality tests, as evidenced by p-values well below the significance threshold of 0.05:

- Shapiro-Wilk for Intervention group: p-value = 0.0043

- Shapiro-Wilk for Control group: p-value = 0.0069

These low p-values indicate a strong rejection of the null hypothesis of normality for both groups.

To further assess the distribution, histograms and Q-Q plots were generated. The histograms show that the SUS scores are not normally distributed for either group, with visible skewness and deviations from the bell curve shape expected in a normal distribution.

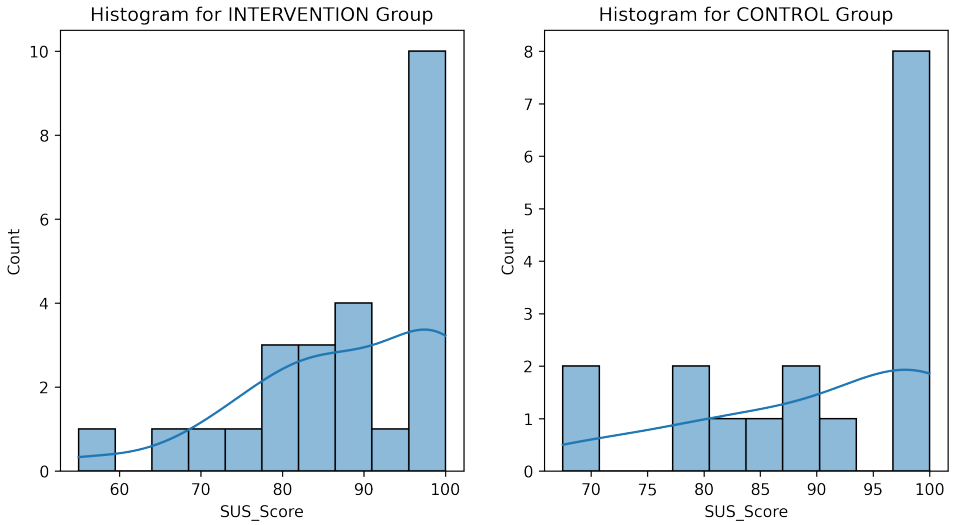

Figure 1: Histograms of SUS Scores for Intervention and Control Groups

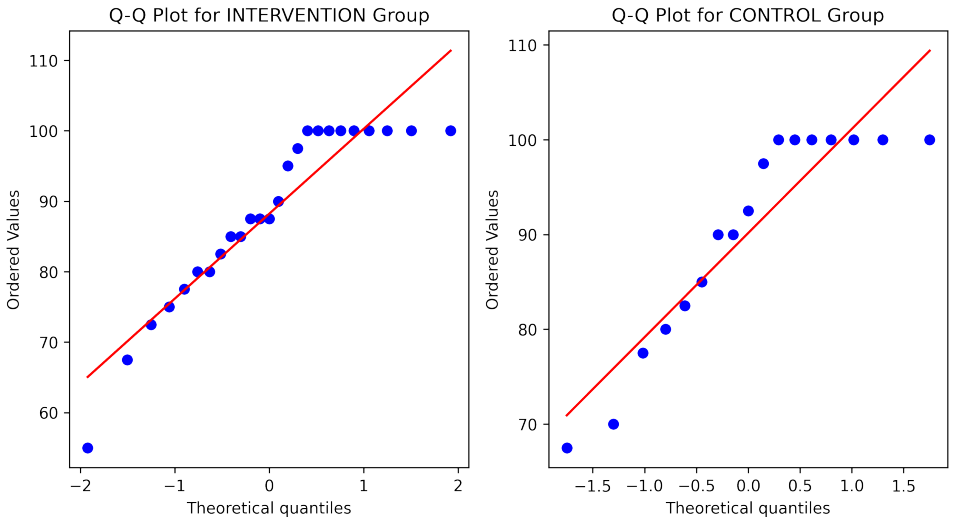

Figure 2: Q-Q Plots of SUS Scores for Intervention and Control Groups

The Q-Q plots illustrate that the data points deviate from the reference line, especially at the tails, confirming the non-normal distribution of the SUS scores in both groups.

4.1.3 Interpretation of Normality Tests

The results from both statistical tests and the visual assessments indicate that the SUS scores are not normally distributed for either group. Given these findings, it is appropriate

to use non-parametric tests for further analysis, as they do not assume normality of the data. Therefore, the Mann-Whitney U test was selected to compare the SUS scores between the two groups.

4.1.4 Mann-Whitney U Test Results

The Mann-Whitney U test was conducted to determine if there was a significant difference in SUS scores between the Intervention and Control groups. The results are as follows:

- U-statistic: 194.5
- p-value: 0.644

Since the p-value (0.644) is much higher than the common significance level of 0.05, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference between the SUS scores of the two groups.

4.1.5 Conclusion

The analysis suggests that both the Intervention and Control groups rated their user experience similarly in terms of usability. Despite slight differences in mean SUS scores and variability, these differences are not statistically significant.

4.2 Task Completion Times Analysis

4.2.1 Descriptive Statistics

Task completion times were recorded for three tasks: drafting a biography (Bio-Secs), composing an email (Email-Secs), and scheduling freelance work (Sched-Secs). Descriptive statistics for each task were calculated for both the Intervention and Control groups.

- **Bio-Secs**: The Control group took more time on average to complete the biography task (mean = 411.12 seconds) compared to the Intervention group (mean = 316.88 seconds).
- **Email-Secs**: Similarly, the Control group showed a higher mean completion time (356.40 seconds) compared to the Intervention group (273.76 seconds).
- **Sched-Secs**: The scheduling task followed the same trend, with the Control group taking more time (mean = 393.68 seconds) compared to the Intervention group (280.24 seconds).

Task	Control Group (sec)	Intervention Group (sec)
Biography	411.12	316.88
Email	356.40	273.76
Schedule	393.68	280.24

Table 1: Descriptive Statistics for Task Completion Times

4.2.2 4.2.2 Visual Representation of Task Times

To visualize the differences in task completion times between the groups, box plots and histograms were created for each task.

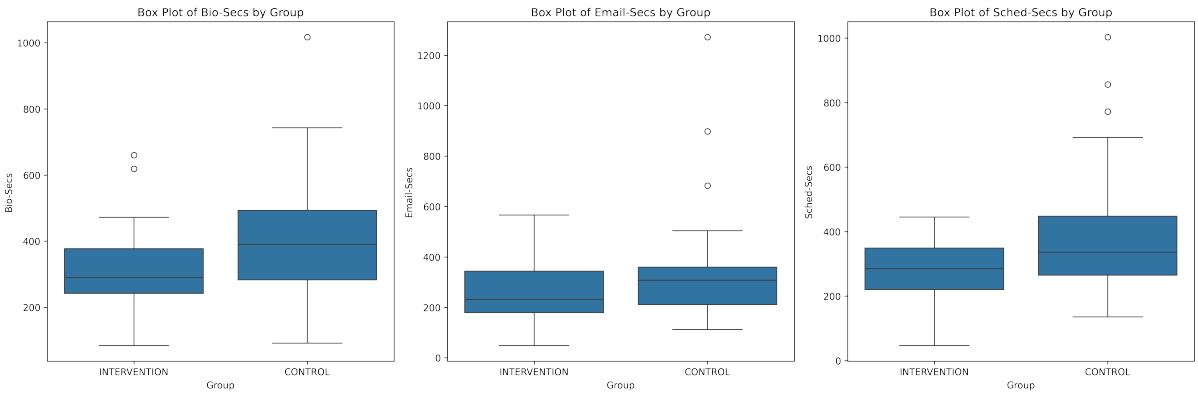

Figure 3: Box Plots Comparing Task Completion Times for Intervention and Control Groups

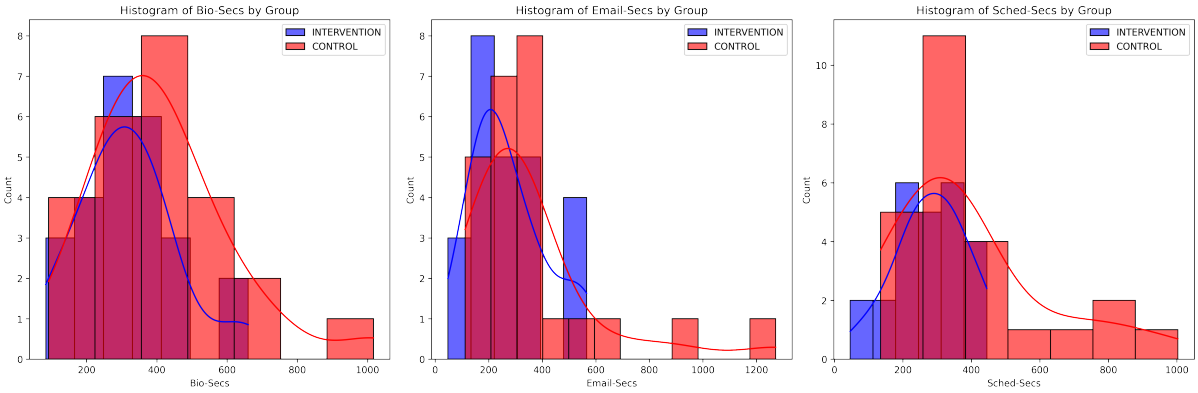

Figure 4: Histograms of Task Completion Times for Bio-Secs, Email-Secs, and Sched-Secs for both groups

- **Bio-Secs**:
  - **Box Plot**: The Control group exhibited a wider range of times with more outliers, indicating greater variability in task completion times compared to the Intervention group.
  - **Histogram**: The Control group's distribution is more spread out, while the Intervention group's times are more clustered around the mean.
- **Email-Secs**:
  - **Box Plot**: Both groups showed similar medians, but the Control group had higher variance and more extreme values (outliers).

  - **Histogram**: The Control group's distribution extends further into higher time values, suggesting some participants took significantly longer to compose emails.

- **Sched-Secs**:
  - **Box Plot**: The Control group had a higher median time and more outliers, indicating a broader spread in scheduling times. The Intervention group's times were more consistent.
  - **Histogram**: The Control group's task times are more dispersed, showing greater variability in how long participants took to complete the scheduling task.

These visualizations highlight that, although the Control group generally exhibited higher task completion times and greater variability, the differences were not statistically significant based on the tests performed.

4.2.3 Normality Test Results

Normality tests were conducted on the task completion times to determine the appropriate statistical tests for group comparisons. The Shapiro-Wilk test was used for assessing normality.

- **Bio-Secs**:
  - Intervention group p-value: 0.3102 ($p > 0.05$)
  - Control group p-value: 0.0583 ($p > 0.05$)
  - Interpretation: Both groups passed the normality test, indicating that the data is approximately normally distributed.

- **Email-Secs**:
  - Intervention group p-value: 0.0704 ($p > 0.05$)
  - Control group p-value: 0.0000 ($p \leq 0.05$)
  - Interpretation: The Control group failed the normality test, indicating non-normal distribution.

- **Sched-Secs**:
  - Intervention group p-value: 0.7278 ($p > 0.05$)
  - Control group p-value: 0.0045 ($p \leq 0.05$)
  - Interpretation: The Control group failed the normality test, indicating non-normal distribution.

4.2.4 Statistical Test Results

Based on the normality test results, appropriate statistical tests were selected for each task:

- **Bio-Secs**:
  - Test Used: Independent t-test (assuming equal variances, as supported by the Fligner-Killeen test)
  - t-statistic: -1.9480
  - p-value: 0.0579
  - Conclusion: The p-value is slightly higher than 0.05; therefore, we fail to reject the null hypothesis. There is no statistically significant difference in the time taken to complete the biography task between the two groups.
- **Email-Secs**:
  - Test Used: Mann-Whitney U test (due to non-normal distribution)
  - U-statistic: 250.0
  - p-value: 0.2290
  - Conclusion: The p-value is greater than 0.05; we fail to reject the null hypothesis. There is no significant difference in email composition times between the groups.
- **Sched-Secs**:
  - Test Used: Mann-Whitney U test (due to non-normal distribution)
  - U-statistic: 226.5
  - p-value: 0.0971
  - Conclusion: The p-value is above 0.05, though closer to the threshold. We fail to reject the null hypothesis. There is no statistically significant difference in scheduling task completion times between the groups.

4.2.5 Variance Analysis: Fligner-Killeen Test

To assess differences in variance between the groups, the Fligner-Killeen test was conducted for each task:

- **Bio-Secs**:
  - Fligner-Killeen statistic: 0.8356
  - p-value: 0.3607
  - Interpretation: No significant difference in variance between groups.
- **Email-Secs**:
  - Fligner-Killeen statistic: 0.0864
  - p-value: 0.7688

  - Interpretation: No significant difference in variance between groups.

- **Sched-Secs**:
  - Fligner-Killeen statistic: 3.1986
  - p-value: 0.0737
  - Interpretation: Variance is closer to being significantly different but still above the 0.05 threshold.

Conclusion: The results suggest that the assumption of equal variances holds for all tasks, supporting the choice of statistical tests used.

4.3 SUS Scores and Task Times: Correlation Analysis

As mentioned in the Methodology section, we initially planned to conduct a correlation analysis between SUS scores and task completion times. However, due to differences in dataset structures (e.g., no unique participant identifiers across the two datasets), this analysis could not be performed. In future iterations of the study, improving data collection processes by introducing unique participant IDs will allow for a more thorough analysis of potential relationships between user satisfaction and task efficiency.

5 Conclusion

This study set out to evaluate the effectiveness of OfficeMind, a custom AI-based tool designed to enhance the productivity and efficiency of freelance workers. By comparing task completion times and user satisfaction scores between freelancers using OfficeMind and those using conventional tools—including popular AI platforms like ChatGPT—we aimed to understand the potential benefits of integrating generative AI into freelance workflows.

The results of our structured usability study indicated that while the Intervention group (using OfficeMind) exhibited slightly lower mean task completion times across all tasks, these differences were not statistically significant when compared to the Control group. Similarly, the System Usability Scale (SUS) scores showed no significant difference between the two groups, suggesting that user satisfaction with OfficeMind was comparable to that of existing tools.

These findings suggest that, in its current form, OfficeMind does not provide a measurable improvement in task efficiency or usability over the tools freelancers are already using. The lack of significant differences could be attributed to several factors:

- **Familiarity with Existing Tools**: Freelancers in the control group may have leveraged their familiarity with their preferred tools, offsetting any potential advantages offered by OfficeMind.
- **Learning Curve**: Participants in the intervention group might have experienced a learning curve with OfficeMind, reducing its immediate effectiveness.
- **Task Complexity**: The tasks selected may not have been complex enough to highlight the potential benefits of OfficeMind's features.

Despite the quantitative results, qualitative feedback from participants highlighted specific features of OfficeMind that were appreciated, such as the tailored assistance in drafting biographies and composing emails. Some participants suggested areas for improvement, including enhancements to the user interface and additional customization options. This feedback underscores the potential of OfficeMind to better meet the unique needs of gig workers with further development. **Implications for Future Research and Development**:

- **Enhanced Features**: Incorporating user feedback to refine OfficeMind's functionalities could enhance its effectiveness and user satisfaction.
- **Larger Sample Size**: Future studies with a larger participant pool may provide more definitive insights into the tool's impact.
- **Longitudinal Studies**: Assessing the tool's effectiveness over a longer period could reveal benefits not apparent in a single-session study.
- **Integration with Existing Workflows**: Exploring how OfficeMind can seamlessly integrate with freelancers' existing tools and platforms may increase its adoption and utility.

In conclusion, while OfficeMind did not demonstrate a significant advantage over conventional tools in this study, the insights gained are valuable for guiding the future development of AI tools tailored to freelancers. The potential for generative AI to enhance freelance work remains significant, especially as tools become more sophisticated and better aligned with user needs. Continued collaboration between freelancers and technology developers is essential to create solutions that genuinely enhance productivity, communication, and self-management in the gig economy.

References

[1] S. Savage and M. Garcia-Murillo, "Tools for crowdworkers coding data for AI," in *Political Science and Public Policy*, Chapter 5, pp. 76–94, 2024. DOI: https://doi.org/10.4337/9781800889972.00012

[2] S. Savage, "Unveiling AI-driven collective action for a worker-centric future," in *Proc. 17th ACM Int. Conf. Web Search and Data Mining*, WSDM '24, pp. 6–7, 2024. DOI: https://doi.org/10.1145/3616855.363763

[3] J. Hui, E. M. Gerber, L. Dombrowski, M. L. Gray, A. Marcus, and N. Salehi, "Computer-supported career development in the future of work," Authors Info & Claims, 2018.