Department for
Business & Trade

The Evaluation of the M365 Copilot Pilot in the Department for Business and Trade

August 2025

Contents

Acknowledgements

This is a report of evaluation carried out by the Digital Data and Technology Monitoring and Evaluation team for the Department for Business and Trade.

Executive Summary

Background

An evaluation of Microsoft 365 (M365) Copilot was conducted by the Department for Business and Trade from October 2024 to March 2025. It aimed to understand the risks and benefits of deploying M365 Copilot across the department, including identifying user groups who may benefit more than others.

One thousand licences were available for a 3-month pilot from October to December 2024 and were allocated to UK-based staff. The sample contained a mixture of volunteers (~70%) and randomly selected participants (~30%) to ensure that findings were as representative as possible across a range of characteristics.

This report covers the theory-based and evidence-without-a-counterfactual evaluation of M365 Copilot conducted by the Department for Business and Trade. The evaluation contained both process and impact evaluation components to evaluate user experiences and understand the impact of M365 Copilot to users and wider colleagues.

Methodology

Usage data for each Microsoft application was collected from Microsoft's M365 Copilot dashboard and provided the percentage of active users of M365 Copilot licences each week during the pilot period in total, and within each application.

The primary data collection method was a diary study involving all pilot participants, aiming to collect detailed data on use cases, satisfaction, accuracy, and time savings. The diary study achieved a response rate of 32%. Statistical analysis identified that responses could be considered representative of the department, as no characteristics showed significant differences in both Chi^2 tests against expected population size and Mann-Whitney U tests against satisfaction metrics. Therefore, weighting was not applied to results. It is important to note that other biases may exist in the respondent population that could not be identified.

Qualitative interviews aimed to provide further insights into user experiences and attitudes of both pilot participants and control groups in the department. Interviews also aimed to identify the impacts of M365 Copilot on different role types and user groups. Nineteen individuals were interviewed, consisting of 13 pilot participants and 6 control group participants. Observed tasks were conducted to support diary study findings on quality, accuracy, and time savings of specific use cases, and to address concerns of self-reported bias within diary study data.

Findings

Satisfaction

Evaluation results show high satisfaction with M365 Copilot across user groups, with 72% of respondents reporting being satisfied or very satisfied with M365 Copilot throughout the pilot period. M365 Copilot's Net Promoter Score, a metric based on how likely respondents are to recommend a service, was 31, which is considered a good score. The majority of use cases had high satisfaction levels reported. However, there was some variation in satisfaction rates; written tasks, such as reviewing and summarising information and drafting, tended to have higher satisfaction ratings, and using M365

Copilot to produce PowerPoints and schedule meetings reported poorer satisfaction ratings.

Time savings

Small time savings were observed across most use cases, with written tasks presenting the largest time savings. However, some tasks, like scheduling and generating images, incurred additional time to complete the task when participants used M365 Copilot. Additional time to complete tasks was primarily caused by either M365 Copilot being unable to produce high quality outputs or the task being additional workload only completed due to users having M365 Copilot. The evaluation did not find evidence that time savings have led to improved productivity, and control group participants had not observed productivity improvements from colleagues taking part in the M365 Copilot pilot. However, many pilot participants reported noticing time savings in their own roles due to M365 Copilot.

Impacts on specific user groups

M365 Copilot was found to be particularly beneficial for neurodiverse colleagues, who were statistically significantly more satisfied than other users. Non-native English speakers also anecdotally reported significant benefits, including improvements to wellbeing, technical abilities, and future career ambitions. Training was identified as a significant factor in user satisfaction, with self-led training being more effective than formal departmental training sessions at increasing user satisfaction with M365 Copilot.

Acceptable use and accuracy

The evaluation identified a lack of knowledge from control group colleagues who were not part of the pilot, regarding capabilities, limitations, and security concerns. The evaluation also highlighted inconsistencies in quality assuring M365 Copilot outputs across pilot participants and task types. Respondents reported observing hallucinations presented by M365 Copilot throughout the pilot.

Attitudes

Findings identified that usage was sometimes limited by ethical concerns, particularly the environmental impacts of large language model development and maintenance. Environmental impact costing methodologies are currently being developed by the department to address this. Users were also anecdotally less likely to use M365 Copilot if their colleagues and line managers were hesitant about its use.

Conclusion

In conclusion, the evaluation demonstrated that M365 Copilot has the potential to save time and demonstrate high satisfaction among users, particularly those with specialised needs. However, the evaluation identified some limitations and the requirement for training and support to maximise M365 Copilot benefits and mitigate risks. The evaluation findings will feed into a series of internal recommendations and next steps. Further evaluation is needed to assess the value for money and environmental impact of M365 Copilot usage in the department.

Introduction

In October 2024, the Department for Business and Trade entered a trial period for the Microsoft 365 (M365) Copilot pilot. The department had 1,000 licences to allocate to UK-based colleagues until 30 December 2024.

M365 Copilot is Microsoft's corporate package for its large-language model (LLM) artificial intelligence (AI) tool. M365 Copilot consists of a browser chatbot and embedded M365 Copilot in the Microsoft suite applications such as Outlook, Teams, Word and Excel. Like other LLM AI tools, M365 Copilot is trained on large amounts of text data and is designed for natural language processing tasks such as generating and summarising text. Unlike other LLM AI tools, including Microsoft Copilot Chat, which is not embedded within Microsoft applications, M365 Copilot is not trained on the data users input to the tool. This means M365 Copilot is a more-secure alternative to other LLM AI tools and more suitable for handling business and government information.

To interact with M365 Copilot, users enter a prompt which is responded to in real-time. Responses can draw on internet-based information or work content that the user has permission to access. Responses are tailored to the individual and should be relevant to their work and the context of the Microsoft 365 application they are using.

The Monitoring and Evaluation team in Digital, Data and Technology designed an evaluation methodology to capture data throughout the pilot period with the aim of increasing the department's understanding of risks and benefits of M365 Copilot. The evaluation design was a combination of theory-based and evidence-without-a-counterfactual evaluation, including elements of both process and impact evaluation components. The aims of the evaluation were centred around three key themes:

- Use: what do colleagues use M365 Copilot for, and how useful is M365 Copilot for these tasks? Are users adhering to acceptable use policies and guidance to minimise risk to the department?
- Productivity: are users experiencing time savings from M365 Copilot? Are some user groups or tasks more likely to experience time savings? Has M365 Copilot led to additional tasks and workloads for users?
- Satisfaction: how satisfied are users with M365 Copilot? Are some user groups more likely to be satisfied than others? What concerns and barriers to high-quality outcomes are colleagues experiencing?

This report summarises the evaluation aims, methodology, findings and conclusions of the department's pilot.

Methodology

The evaluation methodology included both qualitative and quantitative data collection and analyses. This included a diary study to capture usage, accuracy, and time saving metrics, live usage data from Microsoft to understand uptake, and user interviews and observed tasks to support diary study findings. Volunteers to the pilot had to agree to take part in the evaluation to receive an M365 Copilot licence. Consent was requested from all users who participated in each data collection exercise for the department to process their data.

Licence allocation

The department had 1,000 licences to issue. Initially, UK-based colleagues were able to volunteer for a licence, until around ~70% of licences had been allocated. The remaining ~30% of licences were assigned to a random sample of UK-based colleagues stratified by directorate group and grade to improve representation of the department in the pilot sample. Licences were allocated this way in the month of October 2024, with a small number of exceptions.

Usage Data

Usage data reports from Microsoft's M365 Copilot dashboard were obtained, which was part of the M365 Copilot offer. Reports were generated to capture the number of active users, the percentage of active users per week, and the percentage of active users by application. During the diary study, daily reports for the same metrics were obtained.

Diary Study

Data collection

The diary study consisted of three Excel sheets. All licence holders, excluding analysts part of the M365 Copilot evaluation, were asked to complete during a one-week period in November 2024. The first sheet collected user characteristics data on each respondent. The second sheet asked respondents to record each task they used M365 Copilot for that week, plus a series of questions about each task. The third sheet asked respondents to reflect on M365 Copilot across the whole pilot period so far.

Questions were designed to address the key aims of the evaluation: M365 Copilot's accuracy, quality, ability to create time savings, and to understand how M365 Copilot impacts different user groups. The diary design and questions were reviewed by multiple analysts. Questions were taken from other established surveys where possible, with many characteristic questions coming from the Civil Service People Survey which has already undergone rigorous testing. Newly developed questions were cognitively tested with four licence holders to ensure wording was clear and interpreted in the same way by different users.

Licence holders were asked to complete the diary regardless of whether they had used M365 Copilot throughout both the diary study week and the entire pilot period. This was to ensure the diary captured data from both users and non-users. Participants who were on annual leave, sickness leave, or had other commitments were given the opportunity to complete a diary the previous or following week. This meant all diaries collected captured a three-week period.

Licence holders were asked to provide details about themselves and their role. This included information about previous AI experience, role type and profession, age, gender and health. Licence holders were also asked to record time they had spent on both departmental and other M365 Copilot training. Licence holders were asked to provide the name of a colleague in a similar role and grade who was not part of the M365 Copilot pilot. These names provided a sampling list for control group interviews and observed tasks.

More importantly, licence holders were asked to record each time they used M365 Copilot during the diary study week and respond to a range of questions about that task. Questions included the type of use case, the time taken with M365 Copilot and an estimated completion time without M365 Copilot, satisfaction, accuracy and usefulness. Table 1 summarises the use cases captured during the diary study.

Table 1: List of use cases for M365 Copilot and prevalence

This table summarises the use cases identified in the evaluation for M365 Copilot. Throughout this report, these use cases are referred to. Other types of use cases less commonly carried out were grouped into an "other" category. For each use case, the number of reported instances from the M365 Copilot diary study are also reported.

Use case list	Number of reported cases (N)
Asking M365 Copilot simple questions (e.g. alternative to an internet search)	124
Brainstorming ideas	82
Checking own writing for grammar, tone or spelling	89
Data analysis	48
Drafting a written report or briefing	114
Editing written materials (e.g. reports or briefings)	92
Generating images	22
Other	85
Producing or editing slide decks and presentations	81
Reviewing code	17
Scheduling	14
Searching for existing information or resources	57
Summarising research	45
Summarising written communication (e.g. emails)	153
Transcribing or summarising a meeting	234
Writing an email	167

Analysis
Diary submissions were collated into master tables and used to produce summary statistics on the population. Response rate analysis was conducted across a range of characteristics and experiences of M365 Copilot. Analysis was conducted based on both individual pilot participants, and individual use cases when appropriate; for example, satisfaction was measured on an individual basis and a use case basis. A Net Promotor Score calculation was completed to provide a satisfaction metric comparable to other digital services in the department.

Response rate statistical analysis
Responses were defined as a returned diary where the user answered “yes” to the first question asking for consent for their data to be analysed for the evaluation. Comparisons of the response rate statistics to the original licence holder population and DBT people data were made on all characteristics possible. To establish whether weighting should be applied, Chi^2 were conducted to determine whether there were statistically significant differences in population sizes versus expected population sizes. Mann-Whitney U tests also examined whether there were statistically significant differences in satisfaction between various populations.

Based on these tests, it was determined that weighting did not need to be applied to results, meaning the respondent population could be considered representative of the department. Whether users volunteered or were randomly added to the pilot did show a significant difference from Chi^2 tests (P = 0.000001067$^{),}$ as did directorate group (P = 0.03034). However, these two characteristics did not show significantly different satisfaction outcomes after Mann-Whitney U tests, and therefore weighting was not deemed necessary. No other characteristics tested showed a significant difference in the respondent population sizes versus expected population sizes from Chi^2 tests, meaning weighting was not appropriate. However, there may be unknown biases within the respondent population which we were not able to identify, meaning the response sample may not be representative of the department. Table 2 summarises the Mann-Whitney U tests conducted to identify significant differences between respondent populations.

Table 2: Mann-Whitney U tests completed on diary submissions
The table summarises the Mann-Whitney U tests conducted on diary study data. Presented is the population variable and what numeric variables were analysed to identify differences between populations.

Population variable	**Numeric variables tested**
Previous experience with AI	Satisfaction and Net Promoter Score
Hours spent on DBT training and self-led training	Satisfaction and Net Promoter Score
Gender	Satisfaction and Net Promoter Score
Age	Satisfaction and Net Promoter Score
Health conditions or disabilities	Satisfaction and Net Promoter Score
Civil Service grade	Satisfaction and Net Promoter Score

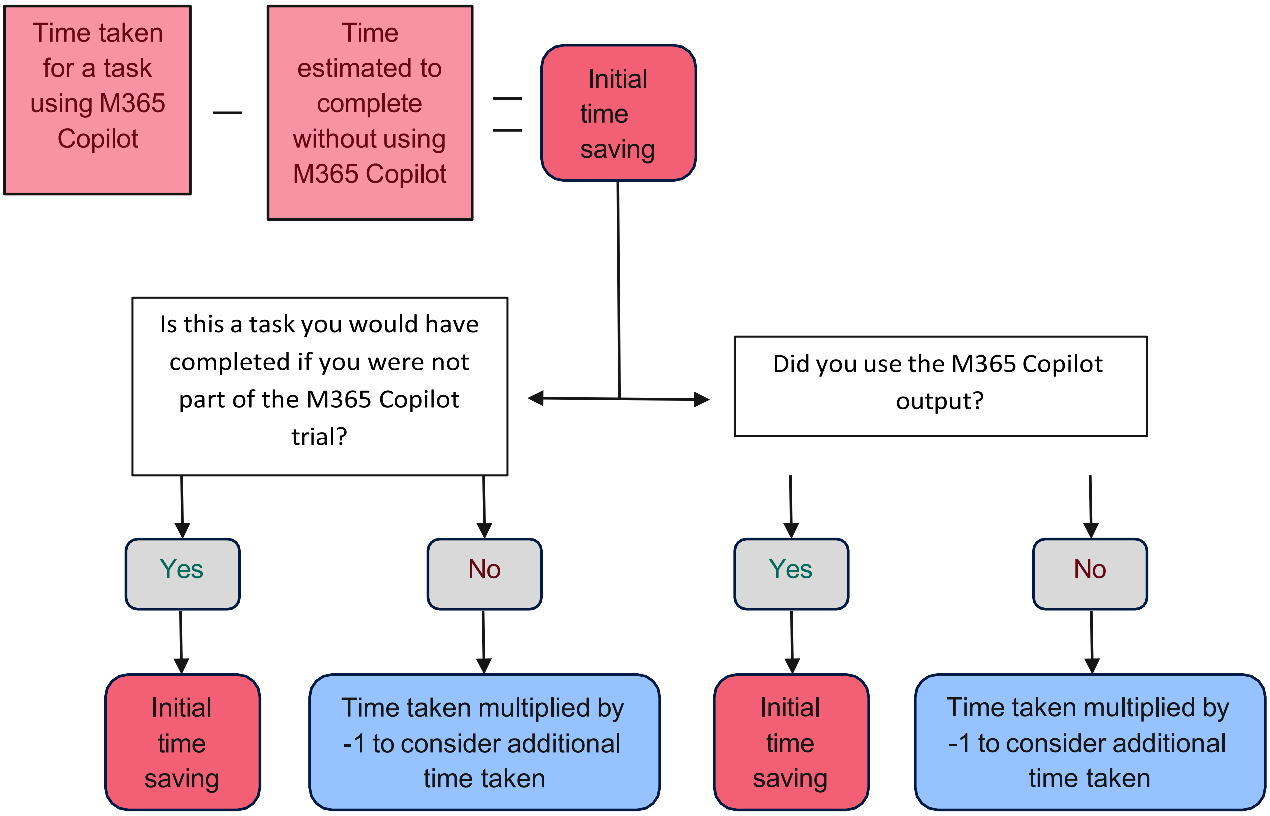
Directorate group	Satisfaction and Net Promoter Score
Whether users volunteered or were randomly assigned a licence	Satisfaction and Net Promoter Score
Whether users volunteered or were randomly assigned a licence	Hours of training
Whether users volunteered or were randomly assigned a licence	Number of tasks recorded in the diary

Time saving analysis

Time saving variables were analysed to calculate a mean time saving per task type. This calculation was then adjusted to discount unused M365 Copilot outputs, and novel tasks which respondents stated they only completed due to having an M365 Copilot licence. If respondents reported they completed a task only because they had M365 Copilot, or they did not use the M365 Copilot output, the time taken was multiplied by –1 to create a negative time saving, reflecting the additional work M365 Copilot had created. The calculation for timesaving is shown in figure 1.

Figure 1: Timesaving calculation and adjustment methodology summary

A summary of the timesaving calculation and how outputs were adjusted. Tasks were adjusted if they were novel, meaning they were additional work created due to having access to M365 Copilot. Tasks were also adjusted if the M365 Copilot output was not used. Adjustments were applied to improve the accuracy of time savings reported on and reflect both time savings and additional time taken.


Time taken for a task using M365 Copilot
—
Time estimated to complete without using M365 Copilot
=
Initial time saving
Is this a task you would have completed if you were not part of the M365 Copilot trial?
Did you use the M365 Copilot output?
Yes
No
Yes
No
Initial time saving
Time taken multiplied by -1 to consider additional time taken
Initial time saving
Time taken multiplied by -1 to consider additional time taken

Quality assurance identified certain tasks had outliers reported in the diaries, visible when producing distribution box plots per task. To assess if outliers were genuine cases, a detailed review of the diary data for each case was conducted. Following these checks, some cases were removed from the final calculations and results. Evidence used to justify removing cases included insights from qualitative fields. Some qualitative fields identified that the respondent aggregated multiple tasks into one when reporting time savings. Other questions provided evidence that the respondent misread the question and submitted time savings in minutes, not hours.

A standard approach to time saving outliers was agreed with several analysts and applied to M365 Copilot and other AI evaluation analysis.

Qualitative analysis

Qualitative analysis from open-text variables within the diary was analysed by sorting responses into themes and subthemes. Analysis was peer-reviewed and quality assured before themes were finalised.

Interviews and Observed Tasks

Interview design and sampling

Three types of interviews were conducted, each of which had a different design and sampling strategy. Interview discussion guides were produced but were not cognitively tested as each interviewer was encouraged to frame questions in their own personal style. Interviewers were also able to clarify, prompt further, and rephrase questions as required. Table 3 shows the number of participants in each interview type. Interviewees were sampled from the pool of diary respondents, and were selected for further research only once, meaning each interview conducted represents a unique research participant.

Table 3: Summary of interview quantities and sampling approaches

Table 3 summarises the qualitative interviews and observed tasks conducted. Presented is the number of sessions and sampling approach for each type of interview.

Interview type	Number of interviews conducted	Sampling approach
Role-based interviews with pilot participants	13	M365 Copilot diaries, using information provided by the respondents on role, Copilot use, and disabilities or health conditions.
Control group interviews with non-users	6	M365 Copilot diaries, using information provided by the respondents on a suitable colleague for a control group, and people data on grade
Observed tasks: test groups	6	M365 Copilot diaries, using information provided by the respondent on role, and people data on grade

Observed tasks: control group	5	M365 Copilot diaries, using information provided by the respondents on a suitable colleague for a control group, and people data on grade

(1) Role-based interviews with pilot participants

The first interview type was a role-based interview with pilot participants aiming to understand how M365 Copilot impacted different role types. Sixteen role types identified in the diaries were selected for interviews. Two participants from each role were sampled, one to be interviewed and one as a back-up should the first participant be unable to attend an interview. These interviews also aimed to understand the impacts M365 Copilot had on users who reported having health conditions or disabilities, and to understand why some pilot participants did not want to use M365 Copilot.

Participants for role-based interviews were sampled from a list of diary returnees, as role-based data was collected from diaries. Other variables considered in sampling were whether the participant used M365 Copilot throughout the pilot, and whether the participant had any health conditions or disabilities. This enabled the interviews to address all research aims.

(2) Control group interviews with non-users

The second interview type was a focus group with DBT colleagues who did not have access to M365 Copilot during or after the pilot, acting as a control group. The aim of these interviews was to understand attitudes to AI across the department. They also aimed to identify if colleagues working closely to those with access to M365 Copilot experienced any impacts, referred to in this report as “spillover effects”.

A focus group was deemed most appropriate for this interview as participants were likely to have limited knowledge or experience to discuss to justify an individual session. To mitigate the influence of the group on the individual, the focus grouped was conducted using a private Mural board to facilitate discussion. Interviewees added thoughts to the board, not visible to other interviewees. This also ensured all participants contributed to the discussion. The board was then used to prompt group discussions around each question.

The focus group was sampled from the control group names provided by pilot participants. Therefore, participants were colleagues in a similar role and grade to pilot participants but did not have M365 Copilot licences themselves. Sampling was conducted based on grade to ensure the sample captured a range of grades, allowing interviews to address how spillover effects are observed by both senior leadership and within teams. One participant who was not able to attend the focus group was asked to respond to key questions via email to ensure the full range of grades were covered by data collection. This respondent was still considered a participant of the control group interviews as their responses were included in the analysis.

(3) Observed Tasks

The third research session type was an observed task. Observed tasks were conducted with test groups (those with access to M365 Copilot) and control groups (those without access to M365 Copilot, obtained from the control group names provided by pilot participants). They aimed to assess time savings, quality, and accuracy of outputs when tasks are completed with M365 Copilot versus without.

Tasks were designed to assess the use cases reported as the most positive or negative time savings following the diary study: summarising information; drafting a written report; data analysis; producing presentations and slide decks. The tasks were short and able to be completed within a one-hour session both with and without M365 Copilot. Tasks were tested by another member of the evaluation team before the sessions were conducted with interviewees.

Test group participants were sampled from the diary respondents, and control group participants were sampled from the matched pairs provided by diary respondents. For the data analysis and presentation task, interviewees also had to have reported as either an analysis, finance, or delivery professional in the diary. This was in attempt to control for skill differences which may influence results. For the summarising and drafting task, users were selected from a range of role types which included Policy, Human Resources, Analysis, Communications, and International Trade. Sampling by role type ensured that interviewees were able to conduct the tasks without the use of M365 Copilot, and that inexperience was not contributing to the quality and accuracy of outputs.

Interviewees were required to complete the task while being observed and sharing their screen and were encouraged to think aloud.

Analysis

Sessions for the role-based and control group interviews were recorded and transcribed using M365 Copilot, and notes were also taken by interviewers or observers in each session. Thematic analysis was conducted on the notes and transcriptions from each interview. Responses were grouped into relevant themes linked to the evaluation aims, and relevant quotes were highlighted. Analysis was quality assured by other members of the team, ensuring the themes produced during the analysis represented the views shared by all interviewees and quotes were accurate. One control group interviewee was unable to attend the group session and instead sent feedback via an email, which was analysed and quality assured in the same way as interview notes and transcripts.

Outputs of the observed task sessions were blind assessed by a secondary analyst to score each output on quality and accuracy, and the time taken for each task was recorded by the interview facilitator. Any comments made by interviewees were analysed using the same methodology as the role-based and control group interview results. Mark schemes were drafted for each task type to ensure scoring on accuracy and quality were standardised. Accuracy scores were based on the output answering the questions correctly. Quality scores were based on the formatting, tone, and presentation of findings being appropriate based on the question and were specific to each task type.

Environmental Impact and Value for Money Analysis

The Department is currently reviewing options to measure the environmental cost of M365 Copilot. For example, a bottom-up approach can be used to estimate the environmental impact of the department's M365 Copilot usage. This would utilise usage data from Microsoft dashboards, proportions of use cases from diary study data, published data to convert types of tasks completed with LLMs to KwH (*Luccioni et al, 2024*), and the Department for Energy, Security and Net Zero's carbon calculator to convert KwH to cash figures. However, this and other methodologies remain in development, and as such findings and detailed methodologies are not included in this publication.

A value for money assessment of M365 Copilot is currently being conducted by the department and as such is not included in this publication.

Findings

Response rate analysis

This section covers the findings of the diary study response rate.

312 diaries were submitted by pilot participants, of which 300 consented to their data being analysed, representing a 32% response rate. Of the 300 diaries, 18% were from the population randomly allocated a licence for the pilot, and 82% were from licence holders volunteering to join the pilot.

The response rate was analysed by a range of characteristics. As discussed in the methodology section, weighting was not applied as statistical tests identified that the response rate could be considered representative of the department. However, unknown biases may be present in the respondent population and therefore the respondent population may not be representative of the department.

Usage data

This section covers the findings using diary study data, and Microsoft usage data provided to the department for the pilot period, covering 01 October to 30 December 2024.

Active users

Chart A shows the percentage of active users each day within the department, obtained from Microsoft dashboards. An active user is defined as an individual who used M365 Copilot at least once that week within any application. Figures are not adjusted for annual leave, sickness, or other absences. Therefore, weekends are visible by clear drops in the percentage of active users to around 1%.

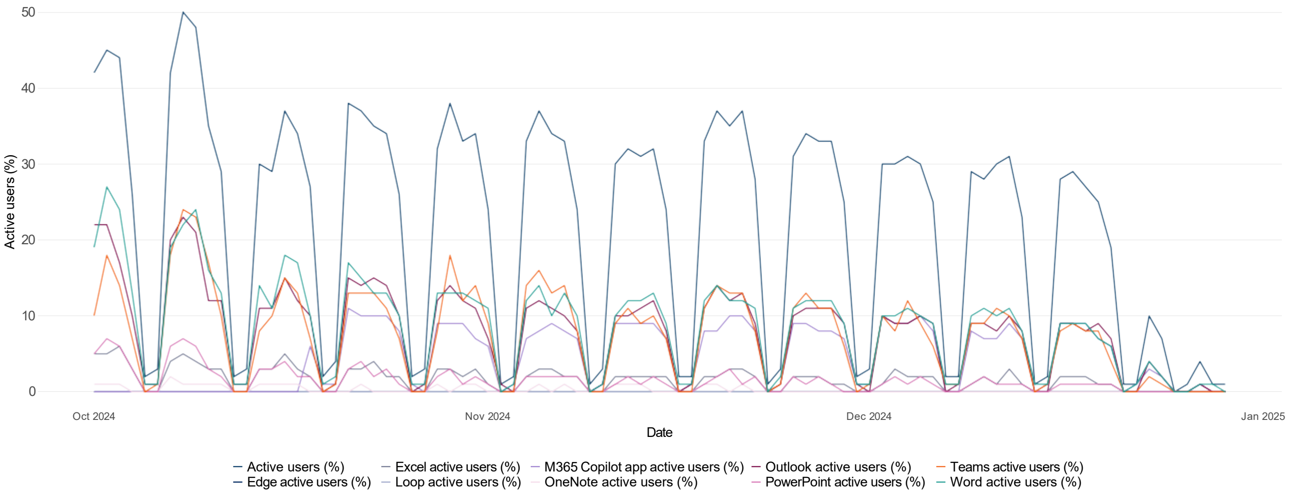
The mean percentage of daily active users was 21%, or 30% when weekends and bank holidays are excluded. The percentage of daily active users peaked on 8 October, where 50% of licence holders used M365 Copilot at least once.

Not presented in the chart, the mean percentage of active users per week was 64%, with the week of 13 to 19 October having the highest percentage of 68% of users being active. The predominant week of the diary study, 17 to 23 November during which 85.8% of diaries were returned, had an active user rate of 67%.

This data implies around 64% of licence holders used M365 Copilot at least once a week, while around 30% used it at least once per working day. Data on individual licence holders is not available from the M365 Copilot monitoring database. During the pilot period, data from the M365 Copilot monitoring dashboard showed that an average of 72 M365 Copilot actions were taken per user. Based on there being 63 working days during the pilot, this is an average of 1.14 M365 Copilot actions taken per user per day. This does not account for annual leave or sickness, nor those who are not full-time employees.

Chart A: active user time-series

Chart A presents the percentage of active M365 Copilot users per day across the pilot, from 01 October to 29 December 2024. The total active licence holder percentage is presented, as well as the percentage of active users within each Microsoft application.



Daily data of the percentage of active licence holders, from 01 October 2024 to 29 December 2024
The number of active licence holders varies across the time frame
Loop and Edge active user percentages were consistently below 1%, meaning these series are not visible in the chart

Use by application

Chart A also shows the percentage of active users within each Microsoft application during the pilot period, obtained from Microsoft dashboards. Microsoft Word, Microsoft Teams and Microsoft Outlook were the applications M365 Copilot was used in the most, with around 10% to 15% of licence holders using them daily. The M365 Copilot app also had higher use than other applications, at around 6% to 10% of licence holders using it daily. However, the M365 Copilot app had fewer users than the three most popular applications.

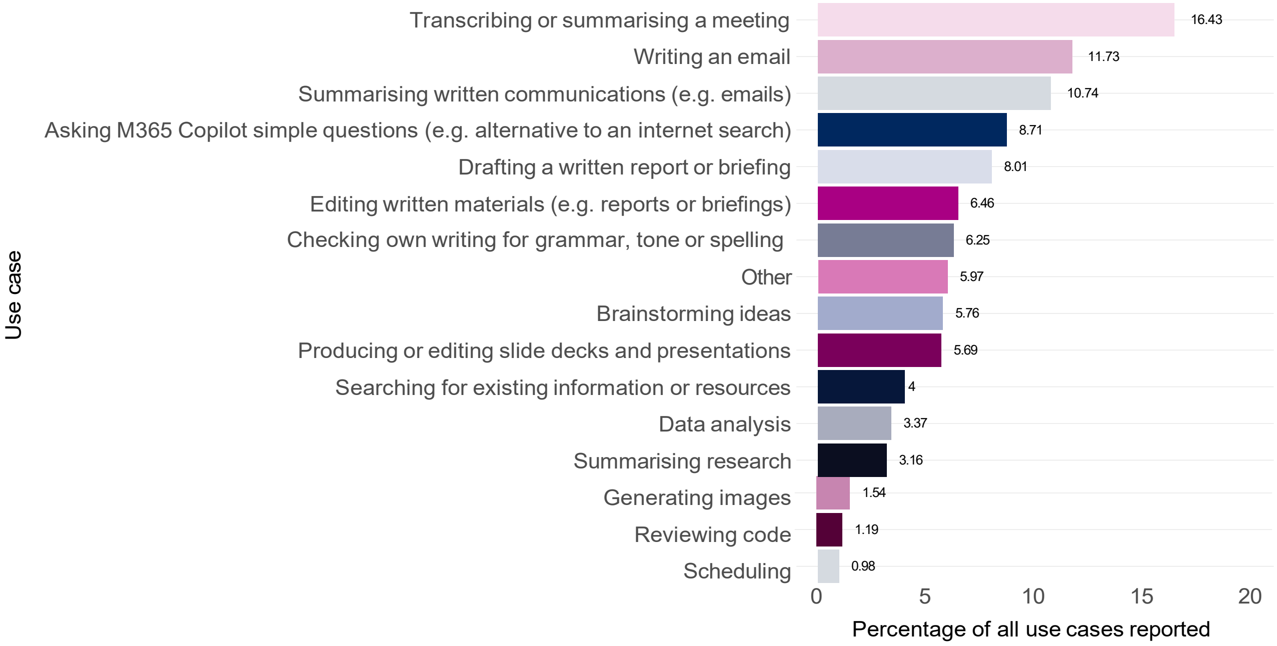
Microsoft Loop and Microsoft OneNote had very low usage rates; consistently across the pilot period, less than 1% used M365 Copilot within Loop per day, and less than 3% used M365 Copilot within OneNote per day. PowerPoint and Excel were slightly more popular; both experienced peak activity of 7% of licence holders using M365 Copilot in a single day within those applications.

Comparisons cannot be made to general use of these Microsoft applications as we did not have access to that data. Therefore, trends observed in chart A could reflect general usage trends of each application and may not be a reflection on M365 Copilot’s usefulness within each application.

Use case findings from the diary study are presented in chart B. The most reported use cases from the diary study were “transcribing or summarising a meeting”, “writing an email”, and “summarising written communications”. In contrast, the least reported use cases were “scheduling”, “reviewing code”, and “generating images”.

Chart B: Percentage of use cases reported by diary respondents

Chart B presents the percentage of each use case taken from the total number of use cases submitted in the diary study. Individuals may be counted multiple times in the chart if they reported multiple use cases or reported a use case more than once.

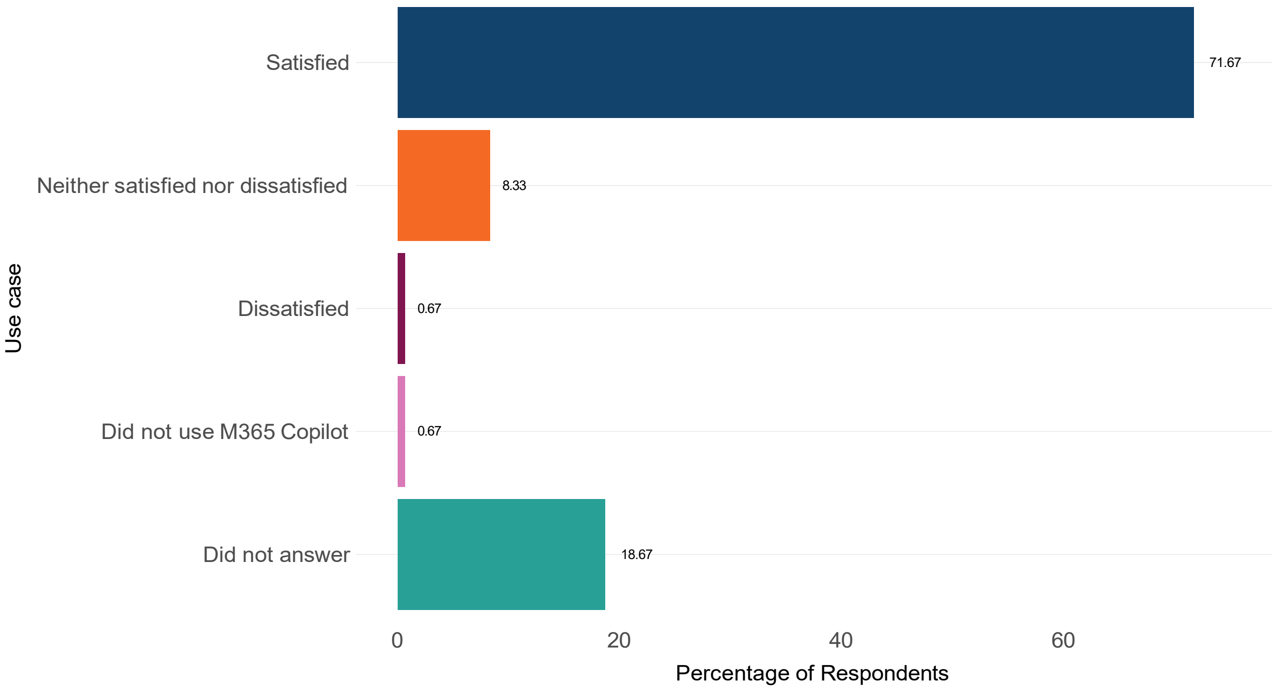


Respondents were asked to report each use case they used M365 Copilot for within a one week period.
Use case categories were predefined following research with users and cognitive testing.
Respondents selected the category from a drop-down list.
Please note that individuals may be counted multiple times in the chart if they reported multiple use cases, or reported a use case more than once.
N = 1,424

Satisfaction outcomes and statistical analysis

This section covers findings quantitative and qualitative findings relating to satisfaction with M365 Copilot from pilot participants.

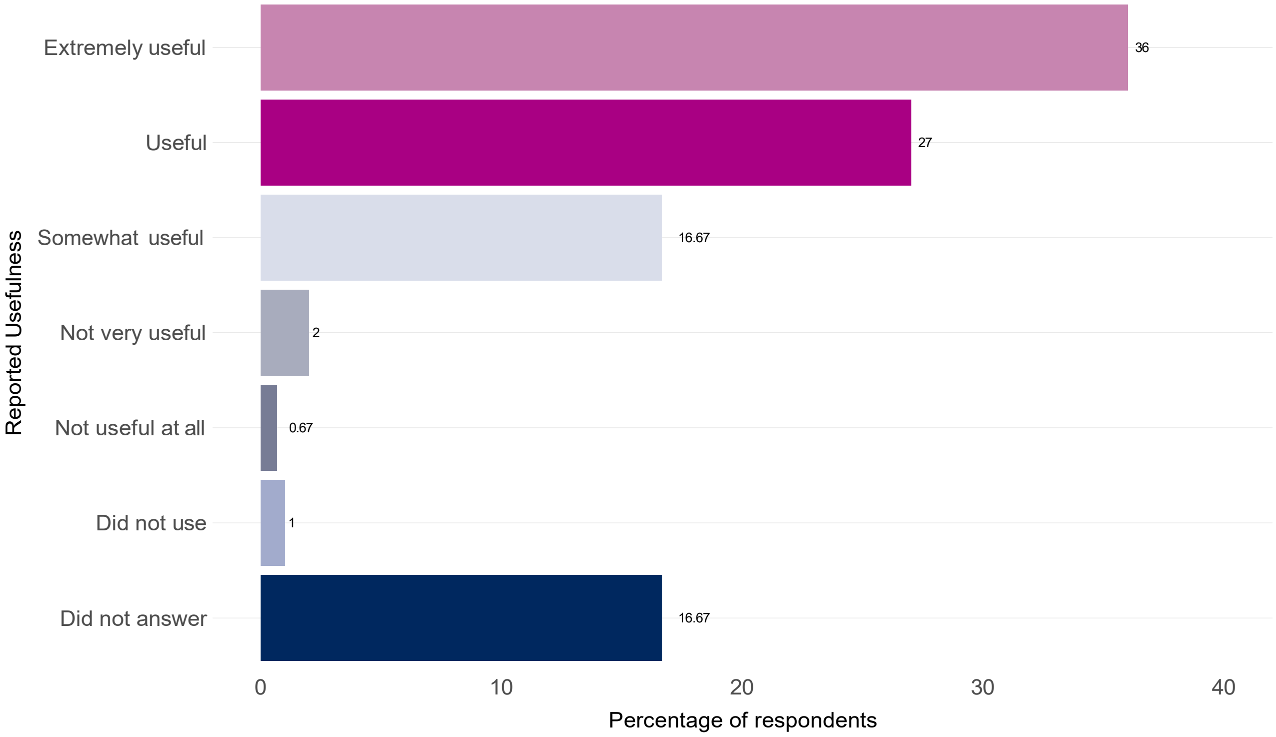
Chart C: Overall satisfaction with M365 Copilot.
Chart C presents the percentage of responses from the diary study who were satisfied or dissatisfied with M365 Copilot overall during the pilot period.



Respondents were asked `Overall, how satisfied are you with M365 Copilot?`
N = 300
Satisfied bar represents respondents who reported being satisfied or very satisfied.
Dissatisfied bar represents respondents who reported being dissatisfied or very dissatisfied.

Overall, diary submissions were positive about M365 Copilot; satisfaction was high with 72% of respondents reporting being satisfied or very satisfied overall. Less than 1% of respondents reported being dissatisfied or very dissatisfied. These findings are presented in chart C.

Chart D: Overall usefulness of M365 Copilot.
Chart D presents the percentage of responses from the diary study who found M365 Copilot useful or not useful overall during the pilot period.

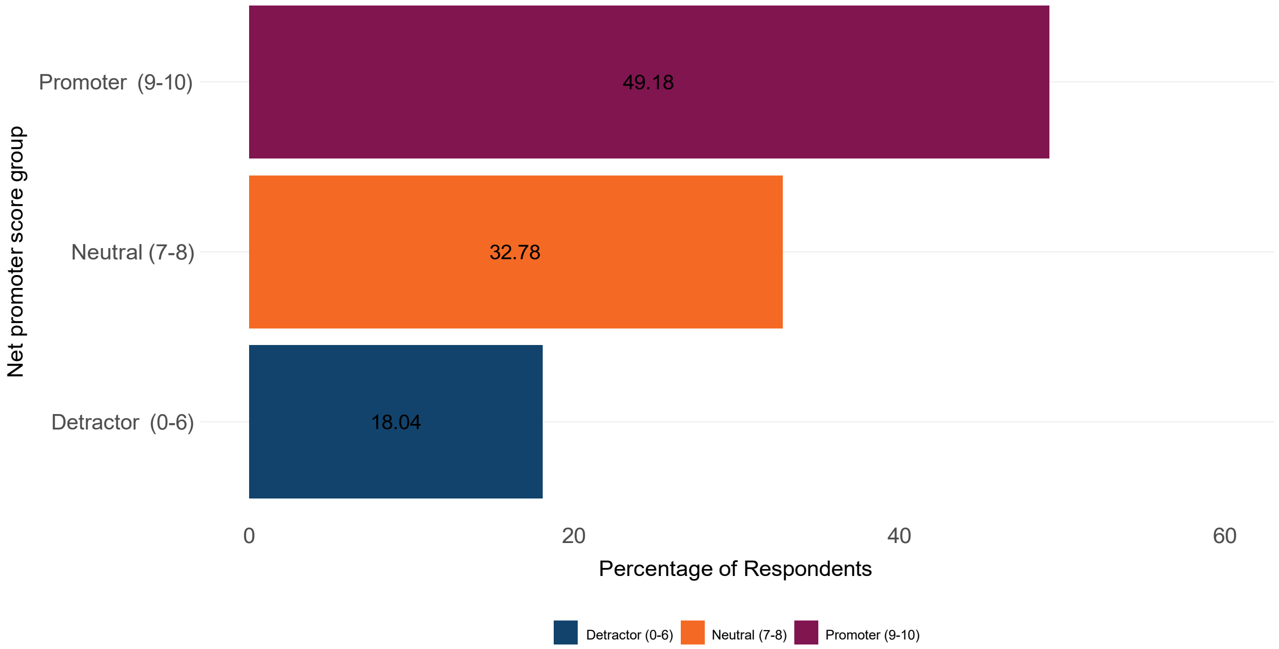


Respondents were asked 'Overall, how useful is M365 Copilot in your day-to-day work activities?'
N = 300

Eighty percent of respondents also reported M365 Copilot was somewhat useful, useful, or extremely useful, and only 3% reported M365 Copilot was not very useful or not useful at all. For both satisfaction and usefulness, 19% of respondents did not answer the question. These findings are presented in chart D.

Chart E: Net Promoter Score results.

Chart E presents the percentage of responses who are promoters, detractors or neutral in responding to the Net Promoter score question from the diary study. The Net Promoter Score questions asks users “On a scale of 0 to 10, how likely is it that you would recommend the service to a colleague?”.

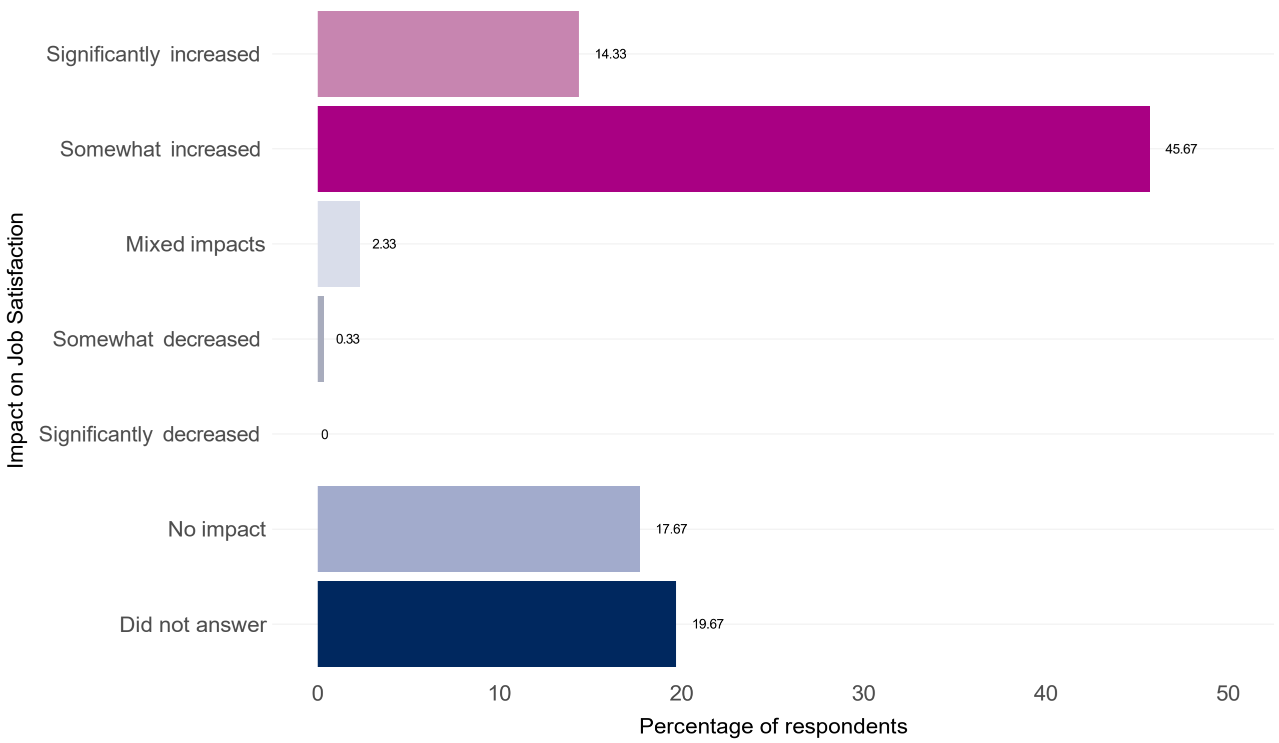


Respondents were asked `On a scale of 0 to 10, how likely is it that you would recommend M365 Copilot to a colleague?`
The Net Promoter Score is calculated by subtracting the percentage of detractors from the percentage of promoters.
M365 Copilot's Net Promoter Score result was 31.
Users who did not respond to the question are excluded from the Net Promoter Score calculation.
N = 244

Diary respondents were asked the Net Promoter Score question, which is a comparable metric across services. It is calculated from the results of asking how likely users are to recommend the service. Detractors are subtracted from promoters to produce the Net Promoter Score for a service. M365 Copilot’s Net Promoter Score was calculated as 31, which is considered a good score. The results of this question are presented in chart E. Forty-nine percent of users scored M365 Copilot a 9 or 10 (promoters), 32% scored it 7 or 8 (neutral), and 18% scored M365 Copilot between 0 and 6 (detractors).

Chart F: Impact of M365 Copilot on job satisfaction

Chart F presents the percentage of responses from the diary study for each category of impacts to job satisfaction caused by having M365 Copilot.



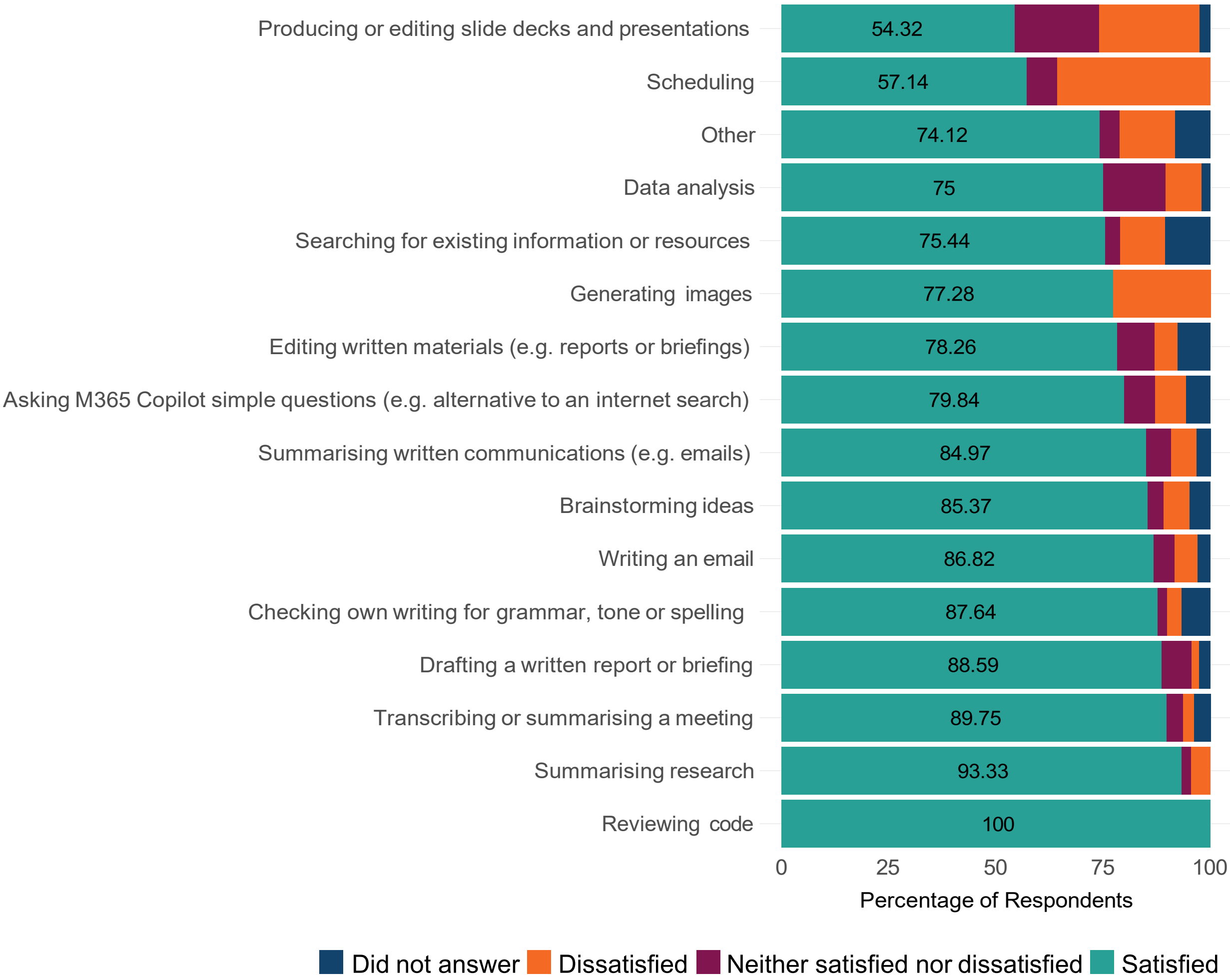
Respondents were asked 'Has having an M365 Copilot licence impacted your overall satisfaction in work?'
N = 300

Respondents were also asked whether M365 Copilot had impacted their job satisfaction. The results of this are presented in chart F. The chart shows that 60% of respondents reported positive impacts to job satisfaction. However, 18% of respondents reported no impacts, and 20% of respondents chose not to answer this question. Two percent of respondents reported mixed impacts, meaning they experienced a combination of positive and negative impacts to their job satisfaction.

An important limitation of the job satisfaction findings is the lack of a counterfactual group. It is possible that factors other than M365 Copilot have impacted job satisfaction results which we cannot identify. It is also possible that unconscious bias from respondents has impacted job satisfaction results. For example, users who are overly keen or overly hesitant for AI to be adopted by the department may have reported more extreme impacts to job satisfaction.

Chart G: Satisfaction by use case

Chart G presents the percentage of each satisfaction response option for each use case. For each use case reported, responses were grouped into ‘satisfied’, ‘neither satisfied nor dissatisfied’, ‘dissatisfied’, and ‘did not answer’ categories.



Respondents were asked `For this task, how satisfied were you with M365 Copilot?`.
Individuals may be counted multiple times across the chart. Percentages of satisfied users are presented for each use case.
Satisfied and very satisfied users have been grouped to 'satisfied'.
Dissatisfied and very dissatisfied users have been grouped to 'very dissatisfied'
Total N = 1,424
N per task: Transcribing or summarising a meeting = 234, writing an email = 167, summarising written communications = 153, asking M365 Copilot simple questions = 124, drafting a written report or briefing = 114, editing written materials = 92, checking own writing for grammar, tone or spelling = 89, other = 85, brainstorming ideas = 82, producing or editing slide decks and presentations = 81, searching for existing information or resources = 57, data analysis = 48, summarising research = 45, generating images = 22, reviewing code = 17, scheduling = 14

Chart G shows each use case broken down by satisfaction ratings. Most use cases reported high proportions of users being either satisfied or very satisfied with M365 Copilot. On average, 80% of use cases recorded were reported as satisfied or very satisfied with M365 Copilot by the respondent, shown by the green bars. However, “producing or editing slide decks and presentations” and “scheduling” had poorer satisfaction ratings, with only 54% and 57% of recorded cases respectively reporting being satisfied or very satisfied. “Scheduling” also had the highest percentage of dissatisfied recorded cases at 36%. This is significantly higher than “producing or editing slide decks and presentations” in which 23% of recorded cases reported dissatisfaction with M365 Copilot. All other use cases had at least 74% of recorded cases reporting being satisfied or very satisfied.

Some respondents recorded using M365 Copilot for tasks other than those already identified as use cases. Throughout this report, these are grouped into the “other” use case. Information on these use cases was collected by qualitative fields in the diary study.

Impacts to specific user groups

This section explores the impacts, both positive and negative, of M365 Copilot to specific user groups, using primarily satisfaction data. User groups include age groups, genders, professions and role types, grades, and those with disabilities. It also explores the impact training has had on user satisfaction.

Statistical analysis was conducted to determine which characteristics of respondents were linked to satisfaction outcomes. Age, gender, profession, role type, grade, and previous AI experience variables showed no statistical differences in satisfaction outcomes between groups. There were also no statistical differences in satisfaction outcomes between users who volunteered for the pilot versus those who were randomly sampled to join the pilot. Table 4 shows the results of test variables which did lead to statistically significant outcomes; training and health conditions or disabilities.

Table 4: Results of Mann-Whitney U tests on satisfaction outcomes for different populations

Table 4 presents the results of Mann-Whitney U tests. The table presents only characteristics with statistically significant results. Statistical significance is indicated by asterixis (* = to 90%, ** = to 95%, *** = to 99%)

Characteristic	Group details	Satisfaction P value	Net Promoter Score P value
Self-led training	At least 1 hour versus no training	0.2431	0.00005252***
Self-led training	At least 2 hours versus less than 2 hours	0.00132***	0.0002132***
Self-led training	At least 3 hours versus less than 3 hours	0.007694***	0.06036*
Self-led training	At least 4 hours versus less than 4 hours	0.002201***	0.01002**
Self-led training	At least 5 hours versus less than 5 hours	0.01041**	0.04333**
Departmental training sessions attendance	At least 1 hour versus no training	0.6714	0.7258
Departmental training sessions attendance	At least 2 hours versus less than 2 hours	0.5192	0.7795
Departmental training	At least 3 hours versus less than 3 hours	0.1367	0.9132

sessions attendance			
Departmental training sessions attendance	At least 4 hours versus less than 4 hours	0.1852	0.9629
Departmental training sessions attendance	At least 5 hours versus less than 5 hours	0.07844**	0.3290
Health conditions or disabilities	Reporting any condition versus all other respondents	0.2591	0.0692*
Health conditions or disabilities	Neurodiverse respondents versus all other respondents	0.09978*	0.03484**
Health conditions or disabilities	Partial hearing or deaf respondents versus all other respondents	0.5764	0.1215
Health conditions or disabilities	Respondents with a learning difficulty or disability versus all other respondents	0.7977	0.1539
Health conditions or disabilities	Respondents with a mental health condition versus all other respondents	0.7167	0.7694
Health conditions or disabilities	Respondents with a stamina condition versus all other respondents	0.6377	0.8886
Health conditions or disabilities	Respondents with vision impairment or blindness versus all other respondents	0.3367	0.9819

Health conditions and disability findings: Diary study and interviews
Mann-Whitney U tests identified that respondents identifying as neurodiverse were statistically more likely to be satisfied (to a 90% confidence level), and to recommend M365 Copilot to others (to a 95% confidence level). Respondents who reported having any health condition or disability were statistically more likely to recommend M365 Copilot to others (to a 90% confidence level).

Interviews confirmed observations in the diary studies that M365 Copilot has significant impacts on neurodiverse users. A user with attention deficit hyperactivity disorder (ADHD) said M365 Copilot has "*levelled the playing field*" for neurodiverse colleagues. Another user with dyslexia commented that the tool empowered them to perform a wider range of tasks with confidence, stating:

> *"It's not revolutionary, it's not going to change how I do my work, but it does make certain things easier... it's made me more confident being able to do reporting... it has actually empowered me".*

A third user made comparisons to other accessibility software currently available to them, saying:

> "*I'd definitely use [M365] Copilot over [REDACTED]... it does a hell of a lot more... the advantage of [M365] Copilot is that it's embedded in your applications*".

Interviews also identified benefits for users with visual or hearing disabilities. One user with a hearing disability reported that M365 Copilot allowed them to relax more in meetings, such as take a sip of water, without worrying they will miss key details. They said:

> "*It can be very tiring because every meeting I need to be 110% focused on what is going on… [with M365 Copilot] I can very quickly recall and be able to share my inputs… rather than sit quietly thinking I missed the point and it's best I don't say anything before I made a fool of myself*".

Non-native English speakers' findings: interviews
Users for whom English is not their first language reported positive impacts not previously captured in the diary study. One user described how M365 Copilot allowed them to communicate in a more effective and professional manner while enhancing the quality of their written outputs. These benefits led to increased career aspirations for this individual.

Training findings: Diary Study and interviews
Analysis identified that conducting at least 2 hours of independent training and upskilling meant respondents were more likely to be satisfied and to recommend M365 Copilot to a 99% confidence level.

The impact of training was also discussed during qualitative interviews. Sentiments regarding formal training sessions were generally positive. However, departmental and Microsoft training sessions were often viewed as introductory or high-level, and less relevant to pilot participant's roles. Individual, self-led M365 Copilot training and experimentation were seen as more useful and allowed users to identify where they personally could benefit from M365 Copilot.

Those who completed at least 5 hours of departmental training were statistically more satisfied with M365 Copilot, but no other departmental training group showed significantly different satisfaction outcomes.

Other characteristics: diary data
Populations with different experience levels with AI prior to the pilot also showed no significantly different satisfaction outcomes, and neither did populations based on age group, sex, or grade. Those who were randomly assigned a licence showed no difference in satisfaction outcomes to those who volunteered to take part in the pilot. One directorate group showed statistically significant differences in satisfaction (satisfaction outcomes

were higher, to a 95% confidence level) but no difference in the Net Promoter Score question satisfaction outcome. No other directorate group showed differences in satisfaction outcomes.

Satisfaction and impact to specific roles: interviews
Statistical tests were not conducted on satisfaction outcomes for each role type population as the sample sizes for some role types were small. Therefore, interviews aimed to identify role types which M365 Copilot was particularly suitable or unsuitable for. Interviewees felt that M365 Copilot was best suited for roles with a heavy administrative burden or roles that require handling large volumes of information and documentation. Benefits were identified for colleagues in Human Resources, Communications, and Commercial roles. However, roles requiring high levels of contextual awareness, nuance, and attention-to-detail, such as Legal and Policy roles, found M365 Copilot less suitable. Individuals in these roles experienced minor benefits by using M365 Copilot for simple tasks but reported that M365 Copilot was not particularly useful for their role.

> *“In a work context, especially dealing with legal text, it’s important that every word is right”*
>
> *“You can’t take all those things and put them together with AI. Only a person can do that”*

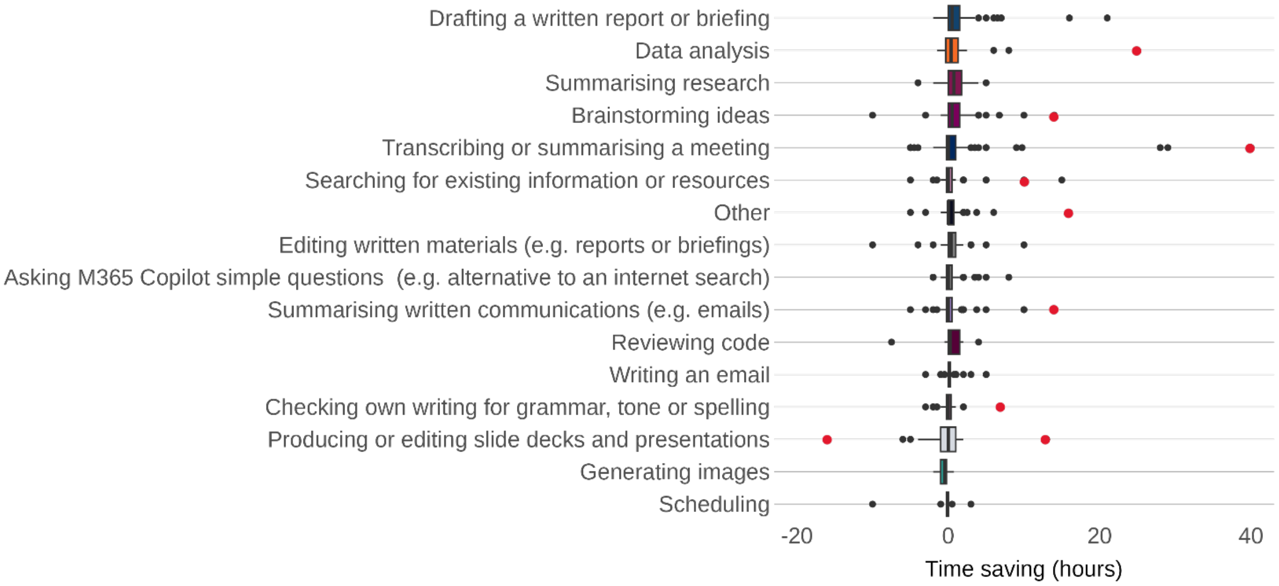
For these roles, M365 Copilot appears more likely to impact individuals based on their personal needs and uses. For example, neurodiverse colleagues in Legal and Policy roles may still significantly benefit from M365 Copilot. However, the number of tasks individuals in roles such as Legal and Policy can use M365 Copilot for are likely to be more limited than other roles, such as Human Resources and Communications.

Time saving findings

This section covers findings relating to time savings and additional time taken to complete a task, due to M365 Copilot.

As part of the diary study, respondents were asked to record the time taken for each task with M365 Copilot, and an estimate of the time this task would take without M365 Copilot. Respondents were asked if they used the M365 Copilot output. Respondents were also asked whether the task was something they would normally do in their role, or whether the task was "novel", meaning they only completed the task due to having access to M365 Copilot. These questions were used to adjust time savings by accounting for additional workload created by M365 Copilot, to make time saving outcomes more accurate.

Chart H: Box plots of time savings, before adjustments
Chart H presents a boxplot of the raw time savings data, calculated by removing the time taken using M365 Copilot from the estimated time to complete the task without M365 Copilot. Nine removed outliers are highlighted in red and split across use cases. Some results which were not outliers were deemed inaccurate when reviewed and were removed from the time estimates, but these are not highlighted in this chart.



Boxplots were produced to identify outliers in time saving calculations. Presented are the time saving results for each individual use case, adjusted for time spent on unused M365 Copilot outputs, and novel tasks the respondent did using M365 Copilot only due to having M365 Copilot, but is not part of their normal day-to-day tasks. Outliers were assessed. Those deemed inaccurate were removed, and are highlighted in red on the chart.
Total N = 1,424
N per task: Transcribing or summarising a meeting = 234, writing an email = 167, summarising written communications = 153, asking M365 Copilot simple questions = 124, drafting a written report or briefing = 114, editing written materials = 92, checking own writing for grammar, tone or spelling = 89, other = 85, brainstorming ideas = 82, producing or editing slide decks and presentations = 81, searching for existing information or resources = 57, data analysis = 48, summarising research = 45, generating images = 22, reviewing code = 17, scheduling = 14

Raw data was analysed to identify outliers as described in the methodology section. Box plots of the raw data are shown in chart H. Analysis of outliers was undertaken and those likely to be errors were removed before calculations. Box plots for all use cases are centred around 0 hours saved, with most outliers being in the positive end of the axis representing reported time savings. A smaller number of outliers are in the negative end of the axis representing tasks took longer with M365 Copilot.

Following outlier removal, calculations were conducted to produce the mean time saving per task. Time savings were adjusted for unused outputs, and both unused outputs and novel tasks using the methodology described previously in figure 1. The results are presented in table 5.

Table 5: Time savings per task type presented unadjusted, adjusted for unused outputs, and adjusted for both unused outputs and novelty
Unadjusted mean hours saved were produced by calculating the mean of subtracting the time taken to complete a task using M365 Copilot from the estimated time to complete the task without M365 Copilot. Columns 3 and 4 show the mean time savings when the calculation is adjusted for unused M365 Copilot outputs, and additional tasks not normally completed by the user, as described in figure 1.

Task	Unadjusted mean hours saved	Mean hours saved adjusted for unused outputs	Means hours saved adjusted for unused outputs and novel tasks
Drafting written documents	1.5	1.4	1.3
Summarising research	1.6	1.4	0.8
Transcribing or summarising meetings	1.2	1.1	0.7
Searching for existing information	1.1	1	0.7
Data analysis	2	1.7	0.6
Brainstorming ideas	2.1	1.6	0.5
Editing written documents	1.7	1.6	0.4
Asking simple questions	0.6	0.5	0.3
Reviewing code	1.6	1.6	0.3
Other	0.8	0.6	0.3
Writing an email	0.2	0.2	0.2
Summarising emails	0.6	0.5	0.2
Presentations	0.8	0.3	0
Generating images	1.5	0.3	-0.5
Scheduling	0	-0.5	-0.6

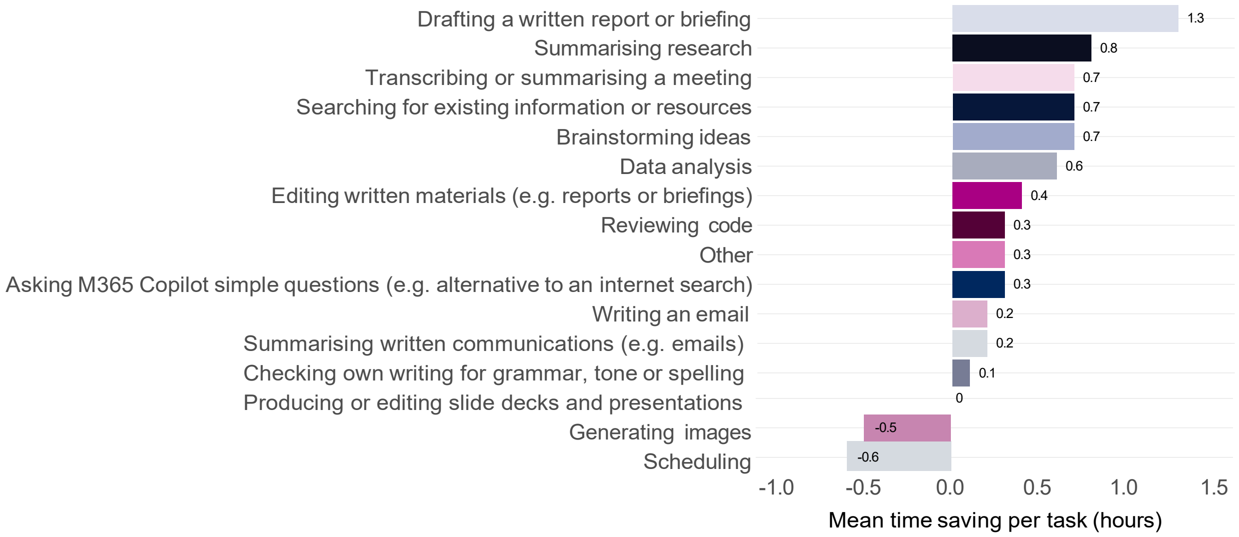
The use cases with the most pronounced time savings are “drafting written documents”, with a mean time saving of 1.3 hours per task, and “summarising research” with a mean time saving of 0.8 hours per task. Both “searching for existing information” and “transcribing or summarising meetings” had mean time savings of 0.7 hours per task.

In contrast, “scheduling” resulted in mean time savings of -0.6 hours per task, and “generating images” resulted in mean time savings of -0.5 hours per task, meaning they both took longer when M365 Copilot were used. Assessing the data from all three columns, it is worth noting that “scheduling” took the same time with and without M365 Copilot before adjustments were made, but some M365 Copilot “scheduling” outputs were unused or novel. Similarly, “generating images” was faster with M365 Copilot, but many outputs were unused or novel.

It is recommended to report on the final column in table 5 with the adjusted time savings as this accounts for additional work created by having access to M365 Copilot. The results of the final column are shown in chart I.

Chart I: Mean time saving per task

Chart I presents the mean time saving per task in hours following adjustments for novelty and unused outputs, and outlier points removed, as described earlier in the paper.



Respondents were asked `Approximately how many hours did it take to complete this task?`, and `Approximately how many hours would you expect this task to have taken without M365 Copilot?`.
Respondents were then asked if they used the M365 Copilot output, and if the task is something they normally do in their role, or if they only completed the task due to having M365 Copilot.
Time savings are adjusted to remove unused outputs and tasks not normally completed.
Outliers were investigated and results deemed inaccurate were excluded from the time saving calculations.
Total N = 1,411
N per task: Transcribing or summarising a meeting = 233, writing an email = 165, summarising written communications = 152, asking M365 Copilot simple questions = 124, drafting a written report or briefing = 114, editing written materials = 92, checking own writing for grammar, tone or spelling = 88, other = 83, brainstorming ideas = 81, producing or editing slide decks and presentations = 78, searching for existing information or resources = 56, data analysis = 47, summarising research = 45, generating images = 22, reviewing code = 17, scheduling = 14

It is important to highlight that this self-reported data may contain reporting biases, for example by users who are overly enthusiastic or critical of M365 Copilot, large-language models, or adopting AI tools. Time savings were therefore investigated further during the observed task research sessions.

Results of the observed tasks are presented in table 6 and somewhat mirror observations from the diary study. Observed task sessions showed that M365 Copilot users produced summaries of reports and wrote emails faster and to a higher quality and accuracy than non-users. Time savings observed for writing emails were extremely small, similar to what was observed in chart I. However, M365 Copilot users completed Excel data analysis more slowly and to a worse quality and accuracy than non-users, conflicting time savings reported in the diary study for data analysis. M365 Copilot users were able to produce

PowerPoint slides over 7 minutes faster on average, but to a worse quality and accuracy than non-users. Again, this conflicts time savings reported in the diary which showed neutral PowerPoint time savings.

Table 6: results of observed task sessions

Table 6 presents the mean accuracy and quality scores from the observed task sessions, and the mean time taken. The results are split by task type, and M365 Copilot users versus non-M365 Copilot users. Accuracy and quality scores were recorded on a 1-5 scale, with 5 being a perfect score and 1 being a poor score.

Task	Average accuracy score M365 Copilot user	Average accuracy score Non-user	Average quality score M365 Copilot user	Average quality score Non-user	Average time taken (MM:SS) M365 Copilot user	Average time taken (MM:SS) Non-user
Data analysis in Excel	1.5	2.7	1.5	2.7	25:01	20:33
PowerPoint	1.5	5	1	2	10:47	18:30
Summarising reports	4	2.5	4	3	12:37	41:34
Writing emails	4.3	4	4.7	3.5	07:30	07:43

The observed task findings are limited by the small sample sizes, and it is possible external factors additional to the use or non-use of M365 Copilot may have impacted the results presented. For example, it may be that sampled participants had different skill levels which has impacted results. Therefore, they should only be considered supplementary to existing findings on accuracy, quality and time savings.

Qualitative findings also provided evidence of time savings. Many diary respondents reported observing time savings in their own work, and many interviewees stated that M365 Copilot allowed routine administrative tasks to be carried out with greater efficiency. Some interviewees discussed that time savings allowing them to redirect time towards tasks seen as more strategic or of higher value, while others reported using these time savings to attend training sessions or take a lunchtime walk:

> *"[M365 Copilot] has great potential for allowing people to focus their energy and time on the intellectually demanding and engaging aspects of their work, by freeing up time spent on things like note taking and simple communication"*
>
> *"[M365 Copilot] has made my job easier for administrative tasks, allowing me to dedicate more time to policy making, strategic thinking, and tasks that add the most value to my role and the team."*

We did not find robust evidence to suggest that time savings are leading to improved productivity. However, this was not a key aim of the evaluation and therefore limited data was collected to identify if time savings have led to productivity gains.

Acceptable Use

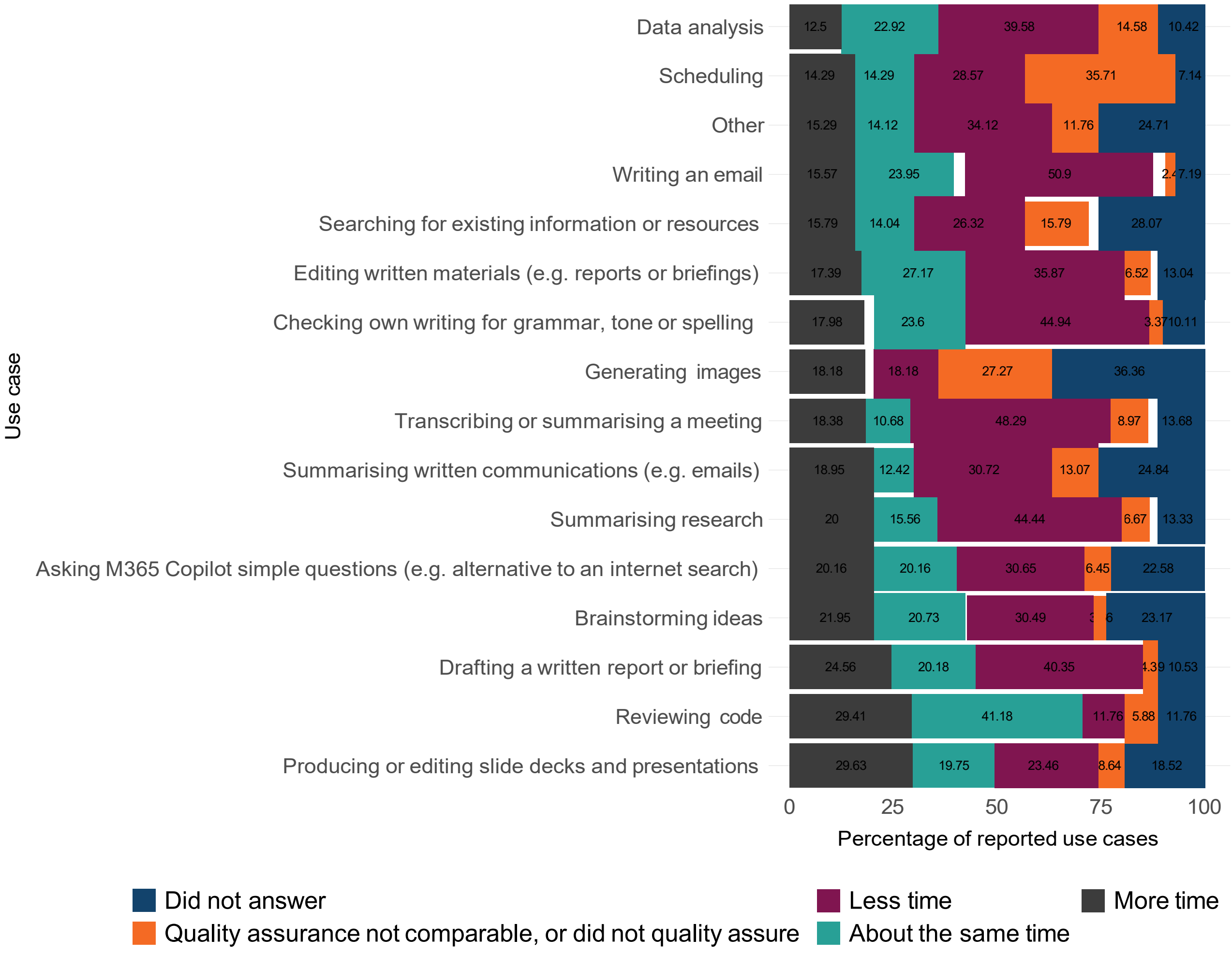
This section covers findings on participant and control group knowledge and application of acceptable use policies and departmental guidance relating to the use of M365 Copilot. It also covers the extent to which users quality assured M365 Copilot outputs.

Diary study

Respondents were asked a series of questions to assess the accuracy of M365 Copilot and its outputs. The evaluation also aimed to assess how diligent users were in reviewing outputs, and how much users adhere to acceptable use policies.

Chart J: Quality assurance time comparison with and without M365 Copilot by use case

Chart J presents the percentages of quality assurance time answers by each use case. Responses are grouped into more time, less time, about the same time, quality assurance is not comparable or the user did not quality assure, and did not answer.



Respondents were asked `Compared to completing this task with M365 Copilot, how much time did you spend on quality assurance?`
Respondents were asked this question for each use case reported, so individuals will may counted more than once in this chart.
Responses for 'Much more time' and 'Slightly more time'were grouped into 'More time'.
Responses for 'Much less time' and 'Slightly less time' were grouped into 'Less time'.
Total N = 1,424
N per task: Transcribing or summarising a meeting = 234, writing an email = 167, summarising written communications = 153, asking M365 Copilot simple questions = 124, drafting a written report or briefing = 114, editing written materials = 92, checking own writing for grammar, tone or spelling = 89, other = 85, brainstorming ideas = 82, producing or editing slide decks and presentations = 81, searching for existing information or resources = 57, data analysis = 48, summarising research = 45, generating images = 22, reviewing code = 17, scheduling = 14

Chart J shows a comparison of quality assurance time taken by specific use cases, comparing using M365 Copilot to completing the task without M365 Copilot. Presented is the percentage of use cases reported split by the quality assurance comparison. The diary question relating to quality assurance time aimed to assess if quality assurance was faster or slower when M365 Copilot was used for a task. It also aimed to identify if users were being diligent in quality assuring M365 Copilot outputs.

Some use cases were less comparable to without using M365 Copilot, such as "generating images". Sixty-four percent of these cases recorded did not have the question answered, did not review the output, or stated the output is not comparable to completing the task without using M365 Copilot.

In 51% of reported cases, "writing an email" took less time to quality assure. This was followed by "transcribing or summarising a meeting" and "checking own writing for grammar, tone or spelling" where 48% and 45% of cases respectively reported quality assurance took less time with M365 Copilot. In contrast, 30% of reported cases of "producing and editing slide decks or presentations" and 29% of reported cases of "reviewing code" reported quality assurance took longer when M365 Copilot had been used.

"Scheduling" was the use case with the highest percentage of users reporting that did not quality assure the output or quality assurance was not comparable (36%). After "generating images", "Data analysis" was the next highest use case where quality assurance was not comparable, or the user did not quality assure the output, with 15% of use cases reporting this.

Interviews

During qualitative interviews, users reported confidence in the department's ability to mitigate security risks of M365 Copilot. However, there was significant variance of knowledge of acceptable use policies across users, with some users reporting not reading security guidance or feeling it was not relevant to their role.

Of those who had attempted to familiarise themselves with guidance, there was confusion between M365 Copilot and Copilot Chat, both of which users had access to. Many users who were unsure on policies reported being overly cautious, which limited the extent to which they used the tool:

> *"I was probably more hesitant to use it than I should have been at certain points, worrying that the information might be sensitive"*

Control group interviews with non-users identified minimal knowledge of acceptable use policies outside of the pilot population, with many being unable to recall any departmental guidance or information.

In both user groups and control groups, respondents were unsure of where the data they input may lead to, leading to concerns about security and data privacy:

> *"A bit like your Alexa or Google Pod is always listening, is Copilot running in the background? […] Is it picking up anything I write, whether I'm using it or not?"*

Users did not mention using M365 Copilot for any use cases against departmental policies during interviews. Users and control group participants did not mention the need to use the departmental disclaimer on work M365 Copilot, or other AI tools, had been used to support. However, some M365 Copilot users in the observed task sessions did add the departmental disclaimer to their work.

Accuracy

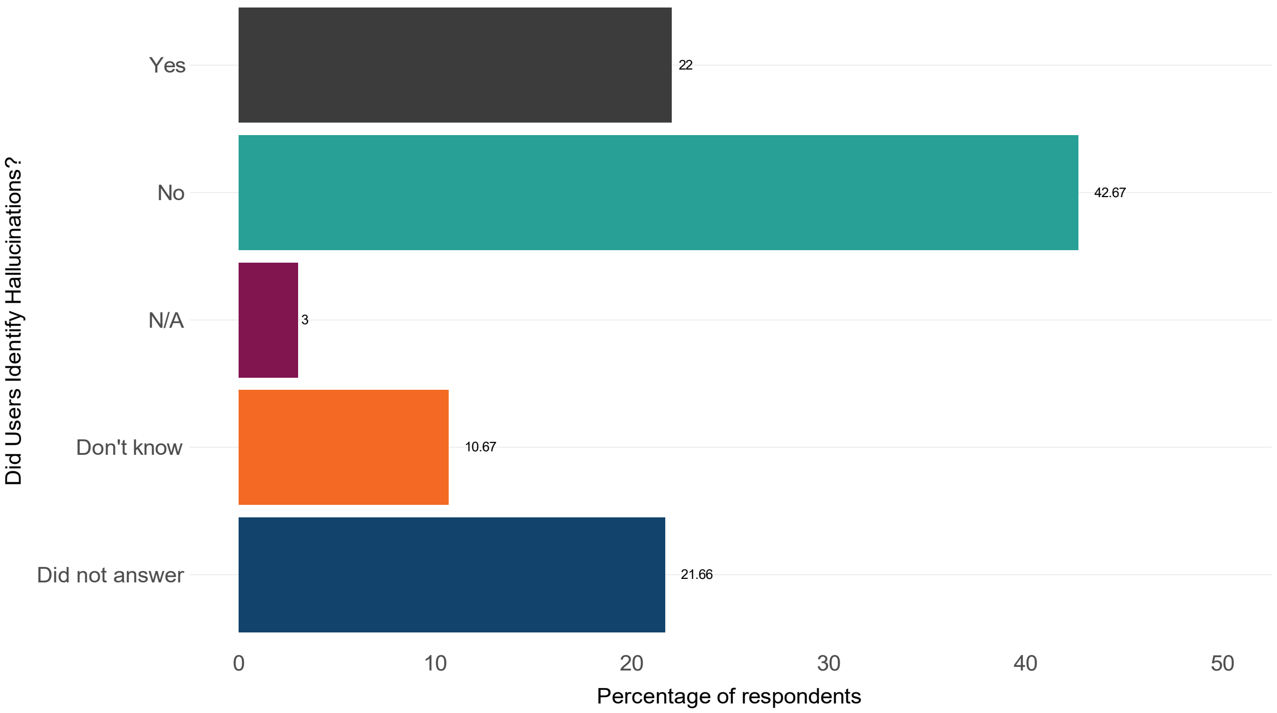
This section covers findings relating to the accuracy of M365 Copilot and hallucinations.

Diary study

When asked to recall hallucinations, 43% of respondents reported they did not identify a hallucination at any point throughout the pilot. However, 22% of respondents did identify hallucinations, and 11% of respondents were unsure if they encountered hallucinations. Twenty-two percent of respondents also chose not to answer the question. It is unclear if this is because users were unable to identify hallucinations, users were unable to recall if they had observed hallucinations, or simply did not want to answer this question. 3 percent of users selected “N/A” meaning they did not use M365 Copilot throughout the trial period. This data is presented in chart K.

Chart K: Proportion of Users who Identified Hallucinations from M365 Copilot

Chart K presents the percentage of respondents for each possible answer to the question asking if users identified hallucinations during the M365 Copilot pilot.



Respondents were asked `In your time using M365 Copilot, have you encountered any false or misleading information presented as fact?`
Respondents selected 'N/A' if they had not used M365 Copilot throughout the trial period.
N = 300

Respondents were also asked to describe any manual changes they made to M365 Copilot outputs. Many respondents discussed adjusting the grammar, spelling and tone of M365 Copilot outputs, such as correcting American English spellings and making the tone of writing more appropriate or personal to the user’s communication style. Several respondents also identified that M365 Copilot produced inaccurate information which they had to rectify. Some respondents reported that they used M365 Copilot as a starting point for a task and then built on the output to produce a final output, making many manual changes but still finding the M365 Copilot helpful to start a task.

Interviews

Interviewees were generally aware that M365 Copilot is capable of producing hallucinations, but users had experienced hallucinations to varying degrees. Those whose use of M365 Copilot was more limited tended to report they did not experience inaccuracies or hallucinations very frequently. For example, users who used M365 Copilot to rewrite a drafted email to improve the tone and grammar did not experience hallucinations. The impact of hallucinations and inaccuracies on user attitudes also varied, with some being more concerned than others about their ability to identify them. Users generally trusted M365 Copilot to complete basic tasks but were not fully trusting of outputs and regularly checked outputs before using them:

> *"Relying on AI carries inherent risks, such as the potential for inaccuracies or contextually inappropriate suggestions."*

> "*My advice here is always to double check the outcome from Copilot and not rely entirely on it*"

However, some users reported that they did not check every output depending on the task and context. For example, some acknowledged that they may not review outputs thoroughly if they were not sharing the output with others.

Cultural attitudes to M365 Copilot, and Artificial Intelligence generally

This section covers attitudes from pilot participants and control groups to the use of M365 Copilot, or AI more generally. It also covers how attitudes changed as the pilot progressed and the influence of others' attitudes on M365 Copilot use.

Current cultural attitudes

Qualitative interviews identified that users tended to increase their use of M365 Copilot as the pilot progressed, even if initially sceptical.

> "*As soon as I did start using it, I felt like I really should've been using it from the day dot*".

Users generally expressed disappointment regarding the ending of the pilot and having their M365 Copilot licences removed. Some users reported they found it difficult to readjust to working without the assistance of M365 Copilot as the pilot ended, with these difficulties appearing more pronounced among users with disabilities.

Users encountered a mixture of attitudes from colleagues towards their use of M365 Copilot. Some users experienced positive attitudes and reported that they were encouraged by colleagues to complete tasks with M365 Copilot, whereas others experienced negative attitudes and were discouraged by colleagues from using M365 Copilot. It was identified that the attitudes of an individual's team and line manager appeared to influence how much and the way in which the individual used M365 Copilot.

Control group interviewees were intrigued by AI tools such as M365 Copilot but were also cautious about their security and potential misuse. One interviewee asked:

> *"what you put in, where does it go?"*

Control group interviews revealed a lack of knowledge regarding the capabilities of M365 Copilot and other AI tools among those who did not hold licences. Interviewees were unsure of how M365 Copilot could support them in their individual role or what they could use M365 Copilot for. Control group interviewees reported that evidence of successful use cases from the pilot would inform their decision as to whether they would adopt M365 Copilot within their roles and wider teams.

Some pilot interviewees raised ethical concerns regarding the environmental impact of M365 Copilot use, as well as the environmental impact of any other AI tool. Some users simply raised concerns about this, whereas others identified environmental impacts as a barrier to their use and were unwilling to use M365 Copilot for this reason. Interviewees expressed interest in having access to information about the environmental impact of M365 Copilot to inform their decisions as to whether they use AI tools. As a result, they stressed the need for transparency about the environmental impacts of AI technologies from both the department and providers.

Future hopes and concerns

Licence holders generally wanted the department to continue using M365 Copilot and expected the benefits to grow more pronounced as the technology and departmental capabilities develop. However, very few users were aware of other AI tools being trialled within the department and most were unable to identify which AI tool may be the best for their role. Some users identified that they would require AI tools with different capabilities to experience the maximum benefits in the future, such as AI tools that are integrated with departmental systems to reduce administrative burden.

Users recognised potential long-term risks and challenges around the adoption of M365 Copilot. Qualitative interviews identified concerns around the increasing reliance on AI tools, with some users concerned increased use could limit individual development and departmental capabilities. For example, one user had observed AI tools in their previous private sector role be used as "*more of a crutch than a tool*".

Conclusion

The M365 Copilot evaluation conducted in the Department for Business and Trade utilised qualitative and quantitative data collection and analysis to assess M365 Copilot's accuracy, quality, satisfaction, and time savings.

Diary study data and qualitative interviews identified high levels of satisfaction from users, particularly for those who are neurodiverse, are not native English speakers, and those with hearing or vision disabilities. Those who invested their own time to upskill in M365 Copilot demonstrated higher levels of satisfaction than others.

M365 Copilot performed best in written tasks, such as drafting and summarising, across satisfaction and time saving metrics. Time savings are evident, albeit in small quantities, and do not necessarily lead to productivity improvements. Time saving metrics may include self-reported biases but observed task sessions conducted show similar trends to those reported by pilot participants, suggesting bias had small impacts on time saving findings.

Observed task sessions also quantified the quality of M365 Copilot outputs and supports findings from the diary study that M365 Copilot produces outcomes of differing quality based on the type of task. Diary study data and qualitative interviews identified that M365 Copilot outputs are mostly accurate, but hallucinations do occur, and quality assurance should be conducted to review each output.

Certain roles appear to benefit more from M365 Copilot than others; those with heavy administrative burdens and lack of complex data anecdotally reported M365 Copilot was extremely useful for their role. In contrast, roles with complex data and information reported more limited benefits.

Use and uptake of M365 Copilot was somewhat limited by ethical concerns of users, particularly regarding environmental impacts of LLM data centre development and maintenance. Attitudes of colleagues and managers also appear to impact how much an individual utilises M365 Copilot in their role.

Findings from the evaluation will feed into internal recommendations and next steps regarding the use of M365 Copilot.

Limitations of the evaluation

It is important to recognise limitations of the evaluation methodology.

The diary study, which produced the majority of quantitative data, had a response rate of just 32%. While statistical analysis suggests the responding population is representative of the department, it is possible that the responding population contains unknown bias. The diary study also relied on self-reporting from participants, meaning it is possible that respondents submitted inaccurate information by accident. Their answers may also be influenced by personal biases, and their recollection of use cases may lead to inaccurate information being submitted. This is particularly important when considering the time saving calculations, which heavily relies on accurate data being submitted by pilot participants.

Qualitative interviews were not able to cover every role type. They also had a small sample size compared to the size of the pilot, and it is unclear if there were biases in that population, based on characteristics not included in the stratified random sampling.

Similarly, observed tasks had a very small sample size and only covered four tasks. It is also possible that external factors, other than use or non-use of M365 Copilot, influenced the outcomes of the observed tasks. These findings should be treated with caution.

The evaluation was unable to assess the long-term impacts of M365 Copilot use, and the evaluation did not capture data from pilot participants before they began the M365 Copilot pilot. Therefore, no comparison can be made on licence holders before versus after the pilot. The control group was only utilised for qualitative interviews and observed tasks, meaning comparisons to non-users are limited.

The evaluation findings presented in this report are not comparable to findings from other Government departments who also participated in an M365 Copilot pilot period during 2024 and 2025. This is because methodologies and evaluation aims were prioritised differently across Government to allow each department to address benefits and limitations of M365 Copilot specific to their operational purpose.

Annex

A – Diary study question list

Annex A shows a table of questions respondents were asked to answer in the diary study. Not presented is additional clarifying information that was provided to respondents for some questions. In the diary study, M365 Copilot is referred to as Copilot Pro, which was the term being used to describe the tool in the department at the time. Questions 13 to 29 were asked to respondents for each task they reported using M365 Copilot for that week.

Question number	Question	Answer format
1	Please enter your DBT email	Free text field
2	Do you consent to take part in the Copilot Pro evaluation and have your responses recorded?	Drop down choice of "yes" or "no"
3	Which of the following professional roles best reflects the type of work you do in your main job?	Drop down choice
4	Which of the following categories best reflects the type of work you do in your main job?	Drop down choice
5	If 'other', please specify	Free text field
6	How long have you been in your current role?	Drop down choice
7	Which statement best describes your experience with similar AI tools prior to receiving your Copilot Pro licence?	Drop down choice
8	What is your sex?	Drop down choice
9	What is your age?	Drop down choice
10	Do you have any physical or mental health conditions or illnesses lasting or expected to last 12 months or more, that affect your ability to carry out your day-to-day work activities?	Drop down choice
11	Please use this space to expand on how your physical or mental health conditions or illnesses impact your ability to carry out your day-to-day work activities	Free text field
12	Please add the email address of a person in a similar role at the same grade	Free text field

13	What task did you use Copilot Pro for?	Drop down choice
14	If ‘other’, please describe the task you used Copilot Pro for	Free text field
15	Why did you choose to use Copilot Pro to complete this task?	Drop down choice
16	If ‘other’, please specify	Free text field
17	How important would you consider this task for your overall role?	Drop down choice
18	Is this a task you would have carried out this week if you didn’t have a Copilot Pro licence?	Drop down choice
19	How frequently do you perform this task in your current role?	Drop down choice
20	How many prompts did it take for Copilot Pro to produce your desired output for this task?	Drop down choice
21	Approximately how many hours did it take to complete this task?	Data-validated field requiring a number and allowing decimals
22	Approximately how many hours would you expect this take to have taken without using Copilot Pro?	Data-validated field requiring a number and allowing decimals
23	Did you use Copilot Pro’s output for this task?	Drop down choice
24	For this task, how useful was Copilot Pro’s output?	Drop down choice
25	For this task, how satisfied were you with the quality of Copilot Pro’s output?	Drop down choice
26	Did you quality assure (QA) Copilot Pro’s output?	Drop down choice
27	Compared to completing this task without Copilot Pro, how long did you spend on quality assurance?	Drop down choice
28	What kind of manual changes did you make to Copilot Pro’s output for this task, if any?	Free text field

29	Do you have any additional comments and observations specific to this task?	Free text field
30	Have you used Copilot Pro since receiving a licence?	Drop down choice
31	If you did not use Copilot Pro since receiving a licence, please explain why here	Free text field
32	Approximately how many hours have you spent attending live or watching recorded training sessions organised by DBT about Copilot Pro?	Drop down choice
33	If you attended any DBT training sessions on Copilot Pro, how useful were they?	Drop down choice
34	Outside of the DBT training hours, approximately how many hours have you spent training, learning to use, or upskilling on Copilot Pro?	Drop down choice
35	Overall, how useful is Copilot Pro in your day-to-day activities?	Drop down choice
36	Have you encountered any technical issues in using or accessing Copilot Pro?	Free text field
37	In your time using Copilot Pro, have you encountered any false or misleading information presented as fact?	Drop down choice
38	Are there any tasks which Copilot Pro was unable to perform, but which you would like to use AI tools for in the future?	Free text field
39	Has having a Copilot Pro licence impacted your overall satisfaction in work?	Drop down choice
40	On a scale of 0 to 10, how likely is it that you would recommend Copilot Pro to a colleague?	Drop down choice
41	Overall, how satisfied are you with Copilot Pro?	Drop down choice
42	Has having a Copilot Pro licence affected how much you socialise and interact with colleagues in work?	Drop down choice
43	Do you have any further comments, feedback or reflections on your experience of using Copilot Pro which you would like to share?	Free text field

Annex B – Qualitative interview research methodology

Summary

Annex B covers the qualitative interview research methodologies for data collection design, data collection, and analysis. Interviews and focus groups were conducted to complement data from the diary study. Thirteen individual semi-structured interviews were conducted with a stratified, random sample of diary respondents. One focus group was conducted with colleagues who were not part of the M365 Copilot pilot and had not used the tool. Fieldwork notes were analysed to identify key, recurring themes.

Research aims

Interviews and focus groups were conducted with a range of participants, designed to complement the diary study data by expanding on those themes and providing insights on areas not previously addressed, such as ethical concerns and cultural attitudes regarding AI. Interview questions and sampling were centred around 4 key aims:

(i) Engagement with licences, and how the department might address any barriers going forwards

(ii) User group-specific benefits and challenges

(iii) Awareness of risks and best-practise when using LLM

(iv) Satisfaction with M365 Copilot and attitudes to future deployment

Sampling

Interviews were conducted with colleagues who were part of the M365 Copilot pilot, and a separate focus group was conducted with participants who were not part of the pilot. Stratified random sampling was used to select participants, using diary study responses as a sampling base. Oversampling was conducted to produce a reserve list of participants to contact should colleagues decline to take part in interviews.

A total of 13 individual interviews were conducted with M365 Copilot pilot participants, with the sample including:

(i) One participant for 13 of the 16 role types identified in the diaries. Due to non-response and small population bases, no users were interviewed from ‘Science and Engineering’, ‘Internal Audit’ or ‘Legal’ roles

(ii) Two participants who reported a neurodiverse condition

(iii) Two participants who held an M365 Copilot licence as part of the original pilot but reported that they did not use M365 Copilot throughout the pilot period

Sampling firstly considered neurodiversity and non-users to ensure that these characteristics were included. Stratified random sampling was then completed for each remaining role type.

Focus group participants were selected via stratified random sampling from the control group sampling list, obtained from diary respondents providing a name of a colleague in a similar role and grade. Sampling was stratified on grade to ensure the control group captured wider impacts to both members and leaders of teams. Oversampling was carried out to account for non-attendance. Ultimately, 5 participants attended the focus group. A further participant was unable to attend the focus group but shared their thoughts separately via email.

Data collection

Interviews were conducted by 6 researchers, running between 13 and 25 February. Interviews were semi-structured; a discussion guide containing questions addressing research aims was produced to ensure consistency in the themes covered, however, each interviewer was encouraged to adapt the discussion guide to suit their preferences and communication style. Researchers were also encouraged to adapt their use of their discussion guides throughout interviews based on the interview participant needs and answers.

The focus group followed a similar semi-structured approach, focussing on key overarching questions to address research aims, which were explored in more detail through follow-up questions based on the participants' answers. Participants of the control group were also encouraged to add thoughts to a private Mural board throughout the session (meaning participants could not view others' thoughts added) to minimise the influence of the group on individuals, and ensure that views were collected from all participants, including those less confident speaking in a group discussion. Thoughts recorded on the Mural board were also used to facilitate the session, providing further prompts and follow-up questions.

Each interview was recorded and transcribed using M365 Copilot. Detailed notes were taken throughout each interview session by either the interviewer or an observer, covering each of the key questions and including verbatim quotes where appropriate.

Analysis

Interview notes were used as the basis for thematic analysis. Analysis was initially conducted with a deductive approach, drawing on a list of themes developed previously when analysing qualitative diary responses as well as an existing theory of change. However, additional themes emerged inductively through close reading of the interview notes, incorporating topics which were not discussed in diaries.

Notes were coded manually in Microsoft Word, with excerpts being labelled by theme. An excerpt could be coded with multiple themes where appropriate. When all interviews had been coded in this way, excerpts under each theme were collated, analysed and summarised. A second researcher quality assured thematic analysis by reviewing how notes were coded, whether the themes created were representative of responses and whether all views shared by participants were captured.

Annex C – Observed tasks methodology

Summary

Annex C covers the research methodologies for observed tasks, including data collection design, data collection, and analysis. Observed tasks, where participants completed a fictional work task with or without M365 Copilot, were conducted to assess the validity of self-reported diary data. A total of 11 sessions took place, divided between data analysis tasks and written tasks, and control and test groups. Outputs of observed tasks were blind-assessed to quantify quality and accuracy of outputs and the time taken to complete the task, to compare M365 Copilot users to non-users.

Research aims

Observed task sessions aimed to assess the validity of timesaving calculations calculated through the self-reported diary study data, using observational data to check for biases or inaccuracies. Secondary aims included comparing the quality and accuracy of outputs produced with and without M365 Copilot, and to observe how users interact with M365 Copilot when carrying out tasks.

Sampling

Sixteen participants were contacted to take part in observed task research sessions, with the aim of conducting 12 sessions. Of those, 11 participants were recruited for observed exercises, which were divided between a data analysis task in Excel, a presentation task in PowerPoint, a summarising task using PDFs and Word, and a written task in Outlook. The data analysis and presentation tasks were given to 3 participants with access to M365 Copilot, who formed the test group, and 3 participants without access, who formed the control group. The summarising and written tasks were given to 3 test group participants and to 2 control group participants.

Test group participants were identified through a stratified random sample of M365 Copilot diary respondents, and stratified on role type for the data analysis task to ensure the participant was in an Analysis, Finance, or Delivery role in an attempt to ensure participants had enough Excel experience to not influence the results of the observed tasks.

Control group participants were sampled from the control group sampling list, obtained from diary respondents providing a name of a colleague in a similar role and grade who had not had access to M365 Copilot. To minimise unfair comparisons between treatment and control groups, sampling excluded users who reported having a physical or mental health condition affecting their ability to carry out day-to-day work activities. Additionally, treatment and control groups were balanced in distribution of grades where possible.

Task design

Observed tasks were divided between two types of session:

(i) **Data analysis and presentation task**: performing descriptive analysis on a fictional Microsoft Excel dataset and presenting the results in a Microsoft PowerPoint slide aimed at seniors.

(ii) **Summarising and writing task**: producing a written summary covering the key similarities and differences between 2 reports, and composing a short email aimed at seniors summarising the results.

These exercises were chosen to capture tasks which showed pronounced time saving results in the diaries and could feasibly be completed in a single observed session. Drafting, summarising, and analysis showed large time savings, while presentations showed poor time savings. Transcribing and summarising meetings showed large time savings, but could not feasibly be assessed using this methodology, hence it was excluded as a task.

Each task was designed by one researcher, who produced a set of instructions covering the required outputs as well as their intended purpose and target audience. Subsequently, the alternate researcher tested the task to ensure that instructions were clear, and the task could be completed within the allocated time, whether M365 Copilot was used, before sessions were held.

Data collection

Observed tasks were conducted by 2 researchers between 24 February and 5 March, with each researcher covering one type of observed task session.

Prior to the session, the participant was provided with written instructions on how to complete the task, alongside the required materials (a dataset for the data analysis task, and 2 written reports for the written task) and a privacy notice with information on data collection, storage, analysis and reporting.

During the session, participants were asked to read through the provided instructions before performing the specified tasks. Participants shared their screen and were asked to 'think aloud' as they worked but were otherwise instructed to go about the task as they normally would. The researcher remained in the call to observe the participant, provide clarification on tasks, and would occasionally ask questions to clarify what the participant was doing and why.

All sessions were recorded for time saving quantification and stored in SharePoint. Task outputs were anonymised and stored in SharePoint.

Analysis

Timings were assessed by watching back recordings and measuring time taken from start of task to final output.

Anonymised outputs from each session were assessed by the alternate researcher who was unaware which outputs were from test and control groups. Assessment criteria were created to score the quality and accuracy of each output on a five-point scale to allow for direct comparison between test and control groups and between different tasks.

Mean accuracy scores, quality scores, and completion times were calculated for each task, split by test and control groups to assess which tasks M365 Copilot may produce better quality outputs, more accurate outputs, or be able to produce outputs more quickly.

Due to the small sample sizes, these results are only presented to support other, more robust, quantified results from the M365 Copilot diary study. Due to the small sample size, it is also possible that factors other than the use or non-use of M365 Copilot may have impacted the results presented.

D – Quantitative analysis methodologies

This annex covers the theory of quantitative analytical methodologies used in the M365 Copilot evaluation's diary study analysis, and why these tests were appropriate.

Net promoter score

The Net Promoter Score is a satisfaction metric based on the results of asking users of a service "On a scale of 0 to 10, how likely is it that you would recommend the service to a colleague?". Scores of 0 to 6 are grouped as "detractors", scores of 7 to 8 are grouped as "neutral", and scores of 9 to 10 are grouped as "promoters". The percentage each group makes up of total responses is calculated, excluding respondents who did not choose to answer the question. The Net Promoter Score is then calculated by subtracting the percentage of detractors from the percentage of promoters. The neutral group are excluded from the calculation.

Mann-Whitney U tests

Mann-Whitney U is a nonparametric statistical test used to determine whether there is a significant difference between the distributions of two independent samples. Unlike other tests, such as T tests, it does not assume that the data follows a normal distribution, making it particularly useful for small sample sizes. This is why Mann-Whiteny U tests were chosen for statistical comparison between pilot populations.

The results of a Mann-Whitney U test are a P value, which is interpreted in the same way as other statistical P values; if the P value is greater than or equal to 0.1, then there is no statistically significant difference between the populations being compared. If the P value is less than 0.1 then there is a statistically significant difference between the two populations to a 90% confidence level. Mann-Whitney U tests can also be known as Wilcoxon rank-sum tests.

Chi^2 tests

A Chi^2 test, also known as an χ^2 test, is a statistical hypothesis test used to determine whether there is a significant difference between the expected and observed frequencies in one or more categories of data.

Chi^2 tests were used in the M365 Copilot diary study evaluation to determine if the observed frequencies of diary responses were significantly different from the pilot population frequencies on a range of characteristics such as grade, directorate, gender, age, and whether respondents were volunteers or randomly added to the pilot population.

Department for Business and Trade

The Department for Business and Trade is an economic growth department. We ensure fair, competitive markets at home, secure access to new markets abroad and support businesses to invest, export and grow. Our priorities are the industrial strategy, make work pay, trade and the plan for small business.

Legal disclaimer

Whereas every effort has been made to ensure that the information in this document is accurate, the Department for Business and Trade does not accept liability for any errors, omissions or misleading statements, and no warranty is given or responsibility accepted as to the standing of any individual, firm, company or other organisation mentioned.

Please be aware that artificial intelligence has been used to support the generation of this content. Content has not been written by AI.

Published by
Department for Business and Trade

August 2025