Department for Business, Innovation, Science and Trade

Evaluation of the Redbox@DBT trial in the Department for Business and Trade

28 September 2026

Contents

Acknowledgements

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

Executive summary

Background

Redbox@DBT is an internally developed large language model (LLM) artificial intelligence (AI) tool. The Department for Business and Trade (DBT) conducted an evaluation of Redbox@DBT from December 2024 to May 2025. It aimed to understand the issues and benefits of deploying Redbox@DBT across the department, including identifying attitudes towards AI and LLMs more generally.

Throughout the trial period, 200 DBT staff were granted access to Redbox@DBT. The sample of DBT staff contained a mixture of volunteers (~36%), randomly selected participants (~48%), and participants selected by the Deputy Director in each directorate group area (~16%). Randomly selected participants were stratified by directorate group, grade and role to improve representation of the department in the trial sample.

This report covers the theory-based and evidence counterfactual evaluation of Redbox@DBT 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 Redbox@DBT to users and wider colleagues.

Methodology

The evaluation adopted a mixed-methods approach, comprising usage data analysis, before-and-after surveys, diary studies, and qualitative interviews. The department's performance analysts developed a Redbox@DBT that collected usage data.

The before and after surveys aimed to capture users' perception and experience at work before and after having access to Redbox@DBT. The before survey contributed to a baseline of expectations and ways of working of the trial participants to be compared to the findings from the after survey. The overall response rate for the participants who responded to both surveys was ~44%. Statistical tests comparing before and after survey responses were conducted to identify any significant changes.

The diary study collected self-reported metrics on usage, satisfaction, accuracy and time savings for one week within the evaluation period.

The diary study achieved a response rate of 28%. It was not possible to truly check for representatives of the department, but statistical analysis identified that responses could be considered representative of the trial participants. This is because no characteristics showed significant differences in both Chi tests against expected population size and Mann-Whitney U tests against satisfaction metrics.

However, it is important to note that other biases may exist in the respondent population that we were not able to account for, particularly given that 36% of participants self-selected to join the trial.

Finally, qualitative interviews aimed to provide further user experiences and perceptions of Redbox@DBT and gather insights on the emerging results of the

diary study. Ten individuals were interviewed, consisting of a mixture of grade, directorate groups and both volunteers and randomly selected participants.

Findings

Satisfaction

Evaluation results show generally positive satisfaction with Redbox@DBT across user groups, with 59% of respondents reporting being satisfied or very satisfied with Redbox@DBT throughout the trial period. Redbox@DBT's Net Promoter Score, a metric based on how likely respondents are to recommend a service, was 0, which is considered neither positive nor negative score.

Most use cases had high satisfaction levels reported. However, there was some variation in satisfaction rates; tasks such as translating documents and synthesising information tended to have higher satisfaction ratings (100%). Other tasks such as asking Redbox@DBT simple questions or the ones labelled as "other" reported poorer satisfaction ratings (33.3%).

Time savings

Modest time savings were observed across most use cases, with tasks related to document interaction, such as producing literature reviews (2.5 hours) or producing frameworks (1.8 hours) presenting the largest time savings. However, some tasks, like data analysis (-0.5 hours) and editing documents (-0.1 hours), incurred additional time to complete the task when participants used Redbox@DBT. Additional time to complete tasks was primarily caused by either Redbox@DBT being unable to produce adequate quality outputs or the task being a novel task, therefore only completed due to users having Redbox@DBT. Participants that were interviewed confirmed noticing time savings in their own roles due to Redbox@DBT. However, the time savings findings rely on self-reported data and there is a potential for reporting bias, which constitutes a limitation of this analysis.

Acceptable use and accuracy

The evaluation highlighted inconsistencies in quality assuring Redbox@DBT's outputs across trial participants and task types. The diary study highlighted that for some tasks such as producing literature reviews and data analysis, participants reported the checking process to take longer when they were quality assuring outputs produced by Redbox@DBT. Around 20% of the respondents reported observing some hallucinations presented by Redbox@DBT throughout the trial, with no direct impact observed on the satisfaction levels.

Attitudes towards AI, Issues and Barriers to adoption

Findings from the before and after surveys identified that trust in LLMs and experience in using AI tools increased among trial participants. The increase in participants' experience with AI after the pilot was statistically significant ($P = 0.006907$) while the increase in trust levels was not. Concerns identified included reliance on AI without sufficient human oversight and uncertainty around accuracy, alongside barriers to using AI disclaimers[1] due to unclear guidance and concerns

[1] A statement to inform others that artificial intelligence has been used to produce an output.

about perceived credibility. Additional training and clearer policies have since been introduced to mitigate these issues.

Limitations

The evaluation has several limitations. None of the diary study, surveys, or interviews included a control group, which limits attribution of outcomes and raises potential confounding. Additionally, the main quantitative data source – a diary study – had a low response rate (28%) and relied on self-reporting, increasing the risk of bias, especially in time-saving estimates. The surveys faced similar challenges with response rate and recall bias. Similarly, qualitative interviews involved a small, non-representative sample and served only exploratory purposes. Finally, the study did not assess long-term effects of using Redbox@DBT.

At the time of the evaluation, Redbox@DBT remained under active development. Several core features were still being iterated or tested, and the deployment was limited in scale and stability. Therefore, the findings reported here correspond to how the tool operated at that specific point in time. They should not be interpreted as reflective of the tool's present capabilities, performance, or user experience as further development, refinement, and feature expansion have since taken place.

Analysis of value for money and environmental impacts of the tool was not in scope for this evaluation but will be addressed in future DBT research.

Conclusion

In conclusion, the evaluation has demonstrated that Redbox@DBT has the potential to save time for users, particularly for document-related tasks, and generally positive levels of user satisfaction. The ongoing monitoring identified low usage of the tool with the evaluation highlighting uneven usage across use cases throughout the trial period. While some limitations and issues remain, the tool is broadly trusted by staff. 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 Redbox@DBT usage in the department.

Introduction

In December 2024, the Department for Business and Trade (DBT) initiated a trial period for its in-house large language model (LLM) tool, Redbox@DBT. The department initially allocated 200 accounts to colleagues across 4 Directorate Group (DG) areas. The trial ran until March 2025 and included DBT employees based in the UK as well as international locations. In early 2026, Redbox@DBT was rebranded to DBT Assist, however, given that this took place after the evaluation was conducted, we have referred to the AI tool as Redbox@DBT throughout. At the time of publication, the department that conducted the evaluation is known as the Department for Business, Innovation, Science and Trade (BIST). However, this report refers to the Department for Business and Trade (DBT), as that was the department's name during the evaluation period and when the report was redacted.

Redbox@DBT is a secure, department-specific LLM artificial intelligence (AI) tool designed to support natural language processing tasks such as summarising, drafting, and answering queries. Unlike external LLM tools, Redbox@DBT is tailored to DBT's internal environment and does not train on user input, making it safe to handle sensitive government information. Users interact with Redbox@DBT by entering prompts, receiving real-time responses that draw on work content they uploaded. These responses are designed to be contextually relevant to the user's role and task.

The Monitoring and Evaluation team in Digital, Data and Technology (DDaT) designed a mixed-method evaluation to assess Redbox@DBT's deployment and impact. The evaluation combined theory-based and evidence-without-a-counterfactual approaches, incorporating both process and impact components. The aims of the evaluation were structured around 3 key themes:

- Use: What do DBT staff use Redbox@DBT for, and how useful is it for these tasks? Are users following guidance to limit vulnerabilities within the department?
- Time savings: Are users experiencing time savings with Redbox@DBT? Which user groups or tasks benefit most? Has Redbox@DBT introduced additional workload?
- Satisfaction: How satisfied are users with Redbox@DBT? Are some groups more satisfied than others? What concerns or barriers to effective use have emerged?

This report summarises the evaluation's aims, methodology, findings and conclusions, providing insights into Redbox@DBT's role in supporting departmental work and informing future deployment decisions.

It is important to note that Redbox@DBT was at an early stage in the development process during this evaluation and the findings are reflective of this specific stage rather than its current capabilities.

Methodology

The evaluation methodology included both quantitative and qualitative data collection and analyses. This covered before and after surveys to identify attitudes and perceptions towards AI, and a diary study to investigate usage, accuracy, time savings and satisfaction. Qualitative interviews were used to support the findings that resulted from the quantitative methods. The data was collected at different stages during the evaluation period that started in December 2024 with the before survey and ended in May 2025 with the qualitative interviews. The timeline of the evaluation is detailed in table 1.

Table 1: evaluation timeline

Table 1 illustrates the timeline of the evaluation split by each research method used.

Method	Phase start date	Phase end date
Before survey	December 2024	February 2025
Diary study	February 2025	February 2025
Qualitative interviews	March 2025	April 2025
After survey	April 2025	May 2025

Consent was requested from all users who participated in each data collection exercise for the department to process their data.

Account allocation

Redbox@DBT was deployed initially to 200 users between December 2024 and March 2025, as part of a trial. The account allocation was split across 4 different directorate group (DG) areas that volunteered to take part of the trial.

Trial participants' selection was either random, based on Deputy Directors' decision for each DG area or on a voluntary basis. Some people were volunteered by their Deputy Directors because their role was believed to benefit from having access to an internal LLM. Around ~48% of the accounts were allocated to users that were randomly selected, ~16% were allocated to users that were selected by their Deputy Director and ~36% were distributed to volunteers. Randomly selected colleagues were stratified by directorate group, grade and role to improve representation of the department in the trial sample.

Usage data

Usage data derived from backend chat logs was partly obtained from the Redbox@DBT performance dashboard, which was developed by DBT performance analysts to monitor the usage and performance of the tool. The analysts generated reports to capture the number of active users and engagement. During the diary study, daily reports for these same metrics were collected.

Before and after survey

The purpose of the before and after survey was to capture users' perception and experience at work before and after having access to Redbox@DBT. The before survey contributed to a baseline covering trial participants' expectations, tasks performed in their role, and perceived experience and trust with LLM AI tools. The after survey was used to collect the same data to be compared with the before survey, to improve understanding of the changes in users' experiences after using Redbox@DBT.

Data collection

The Monitoring and Evaluation team designed both the before and after surveys in Qualtrics and distributed them via email using a unique link for every trial participant. The data collection period was initially 2 weeks for both, but it was extended to 4 weeks to account for annual leave or any similar circumstances.

The questions for both surveys were designed to assess users' attitudes around LLM AI tools, their perception around Redbox@DBT and the nature of the tasks performed. The before and after survey's implementation and questions used were agreed together with multiple analysts from different professions. During the before survey planning stage, the newly developed questions were cognitively tested with 3 participants. The feedback received was used to improve the wording and ensure that the questions were interpreted in the same way by different users.

The participants were asked to provide details about previous AI experience, level of trust in LLM AI tools, as well as information about themselves and their role. This included the role type, grade, profession, age, gender and health. Some questions differed from one survey to the other to allow for a better comparison and understanding of users' perception. During the before survey, participants were asked about what their expectations were around Redbox@DBT and what tasks they were performing before having access to Redbox@DBT. Then in the after survey, they were asked about their experience after using Redbox@DBT and what tasks they were using the tool for.

Response rate

The before survey was sent to 172 trial participants between December 2024 and February 2025, resulting in 131 responses of which 120 were completed; this corresponds to a response rate of approximately 70%. The after survey was administered to 290 users from April 2025 to May 2025 and received 155 responses, with 136 fully completed, indicating a response rate of 47%. The sample size for the after survey was expanded to include feedback from both the initial participants, and those who joined the trial at a later stage.

Both before and after surveys were sent to 137 participants, with 60 completing both. The response rate for completing both surveys was approximately 44%. The respondents were identified using email addresses to derive this response rate.

Analysis

Descriptive statistics were conducted on the aggregate results of each survey separately to assess the relevance of work tasks in participants' roles both before and after users could access Redbox@DBT. The surveys also captured respondents'

experiences and expectations regarding Redbox@DBT and AI more broadly. For participants who completed both surveys (N=60), responses were matched using email addresses to address 2 questions:

(i) among the tasks for which Redbox@DBT was utilised, which were considered most relevant to participants' roles
(ii) whether participants' experiences and expectations of the tool and AI more generally changed during the trial

Survey responses were analysed using 2 non-parametric tests to check for differences between both distributions and means in the before survey versus the after survey responses. To assess changes in distributions, a Mann-Whitney U test was conducted, and a permutation test was carried out to check the differences in means.

These tests were performed to establish if there are statistically significant differences in trust levels, experience using AI LLM tools, and perceptions of Redbox's impact between the before and after responses.

Diary study

The diary study aimed to collect data on Redbox@DBT usage by task, including accuracy, satisfaction and time saving per task. It also aimed to collect demographic information from participants and overall satisfaction with the tool.

Data collection

The diary study consisted of 3 Excel sheets, which all account holders were requested to complete over a designated one-week period in February 2025. The initial sheet was used to gather information on user characteristics for each participant. The second sheet required respondents to log every task they carried out using Redbox@DBT during that week, along with a set of questions relating to each task. The final sheet prompted participants to reflect on their experiences with Redbox@DBT throughout the entirety of the trial period so far.

Where possible, questions were adopted from established surveys, notably the Civil Service People Survey, which has undergone thorough validation. Any new questions were cognitively tested with 4 account holders to ensure clarity and consistency in interpretation. All account holders, regardless of whether they had used Redbox@DBT during the diary week or trial, were asked to participate, ensuring a comprehensive dataset from both users and non-users. Those on annual leave, sick leave, or with other commitments were given the option to complete the diary either the preceding or following week.

Participants were asked for details about themselves and their roles, including AI experience, role type, profession, age, gender, and health. They also provided information on the amount of time spent attending Redbox@DBT training. Participants recorded each instance of Redbox@DBT use during the diary week and answered follow-up questions regarding the nature of the task, time taken with Redbox@DBT compared to an estimated time without it, satisfaction, perceived accuracy, and usefulness. Table 2 provides an overview of the different use cases documented during the diary study.

Table 2: list of use cases for Redbox@DBT and prevalence

This table provides a summary of the use cases identified during the evaluation for Redbox@DBT. Throughout this report, these use cases are referenced. Less frequently performed use cases that were also bespoke and not comparable to other core use-cases included in the evaluation have been combined into an “other” category. For each use case, the table also displays the number of reported instances from the Redbox@DBT diary study.

Tasks list	Number of reported cases (N)
Summarising documents	30
Transcribing or summarising a meeting	10
Checking own writing for grammar, tone or spelling	12
Asking Redbox simple questions	19
Drafting documents	24
Other	11
Producing literature reviews	2
Producing frameworks, templates, guidance and other documentation	2
Writing an email	8
Reviewing documents against existing guidance, laws and policy	5
Editing documents	5
Searching for existing information or resources	9
Summarising written communications	7
Brainstorming ideas	11
Data analysis	1
Producing content for users	3

Tasks list	Number of reported cases (N)
Synthesising information	2
Translating documents	3
Planning stakeholder engagement	2

Analysis

The Monitoring and Evaluation team compiled diary submissions into master tables to generate population summary statistics. Response rates and experiences with Redbox@DBT were analysed by participant and use case, as appropriate. For instance, satisfaction was assessed both individually and per use case. A Net Promoter Score was also calculated to benchmark satisfaction against other departmental digital services. Details about the calculation are provided in annex C.

Response rate statistical analysis

Responses were considered valid when a returned diary included a “yes” answer to the initial consent question permitting data analysis for the evaluation. A total of 61 diaries were received, resulting in a response rate of 28%.

Trial participants submitted 61 diaries, while also consenting to their data being analysed. Of the 61 diaries which were submitted with consent for analysis, 36% were from the population randomly allocated an account for the trial, and 64% were from participants who had volunteered to join the trial.

Weighting was not applied following no significant Chi-Squared and Mann-Whitney U results. However, there could be unknown biases in the respondent population across the department.

Completion rate analysis

Regrettably, the completion rate was significantly reduced with certain questions consistently left unanswered across diary submissions. The diary had an uneven completion: 27% of participants achieved full completion, 45% nearly completed the diary (95%), and 72% reached three-quarters (75%). Consequently, this has impacted the number of statistical tests possible for this analysis. Any unanswered questions were grouped into ‘did not answer’ category in the results.

Anderson Darling tests established that variables of interest for the statistical analysis (satisfaction and Net Promoter Score) were non-normally distributed[OBJ]. Therefore, Fisher and Mann-Whitney U tests were conducted to establish whether statistically significant differences existed between various populations of respondents. These tests were selected due to data being non-normally distributed and the small sample of Redbox@DBT diaries received (n = 61) and resulting group sizes of the various populations.

Mann Whitney-U tests were conducted to establish if there was a statistically significant difference between the Net Promoter Score outcomes within various populations. Mann-Whitney U tests were also performed to establish if there was

statistically significant difference between the satisfaction outcomes within various populations. Further Mann-Whitney U tests were conducted to establish whether there were statistically significant differences in the number of training hours or tasks recorded by the voluntary population vs the randomly selected population. A Chi-Squared test was performed on the voluntary and randomly selected populations to establish whether sample weighting was necessary. Table 3 summarises the tests conducted.

Table 3: Mann-Whitney U tests completed on diary submissions

Population variable	**Numeric variables tested**
Previous experience with AI	Satisfaction and Net Promoter Score
Attended DBT provided training	Satisfaction and Net Promoter Score
Hours spent on 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

Based on these tests, and the outcome of the Chi-Squared test[2], we determined weighting did not need to be applied, meaning the respondent population could be considered representative of the Redbox@DBT trial participants. The Chi[2] tests (P = 1) did not show a significant difference between users who volunteered and the ones who were randomly added to the pilot. 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 trial participants population.

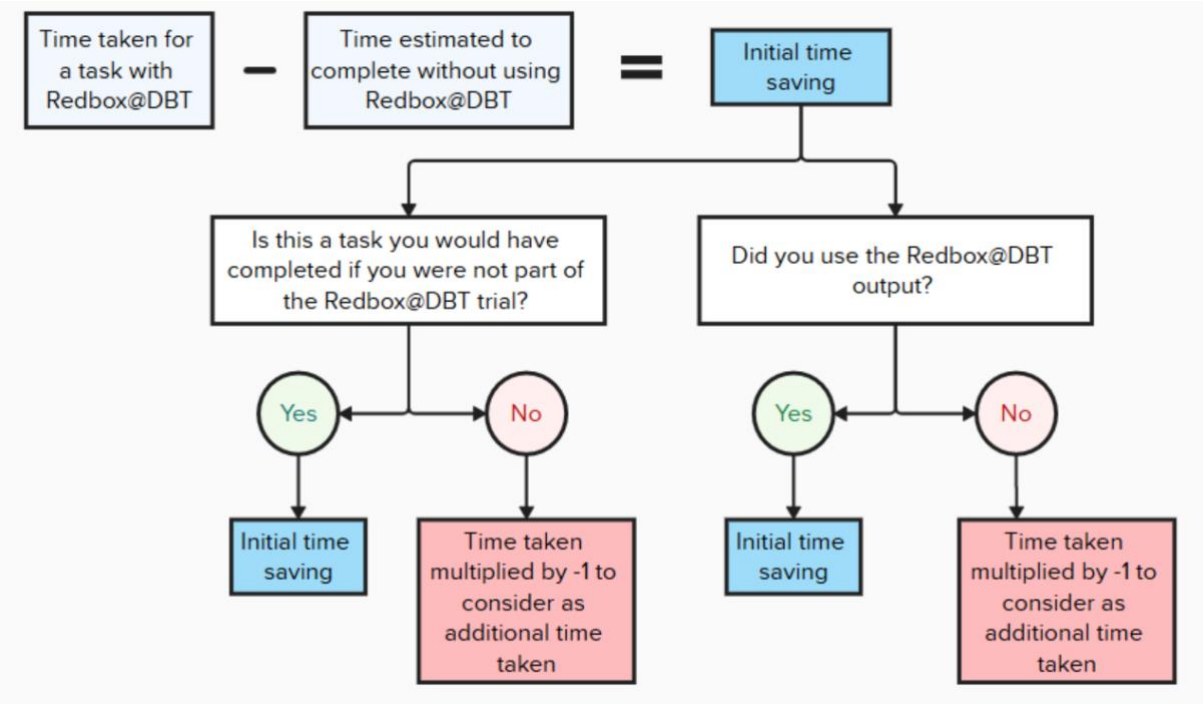
[2] P-value = 1, indicating no significant differences between the representation of random participants and volunteers in our sample, compared with the population of Redbox participants.

Time saving analysis

The Monitoring and Evaluation team analysed time-saving variables to determine the mean time saved per task type. The calculation was subsequently adjusted to account for unused Redbox@DBT outputs and novel tasks that respondents indicated they only performed because they had access to a Redbox@DBT licence. When respondents reported completing a task solely due to having Redbox@DBT, or not using the Redbox@DBT output, the recorded time was multiplied by –1 to indicate negative time savings, reflecting additional work generated by Redbox@DBT. The methodology for calculating and adjusting time savings is summarised in figure 1.

Figure 1: summary of time saving calculation and adjustment methodology

This figure provides an overview of the time saving calculation, including adjustments made for novel tasks created by access to Redbox@DBT and tasks where the output was not utilised. These adjustments were implemented to enhance the accuracy of reported time savings and represent both reductions and increases in time spent.



As shown in figure 1, the initial time saving was multiplied by –1 in either of the following circumstances: the respondent did not use the Redbox@DBT's output, or they used the tool for a novel task that they would not normally undertake as part of their routine work.

Quality assurance identified outliers among certain tasks as observed in distribution box plots. To determine whether these outliers represented valid time saving results, a detailed review of diary data for each instance was conducted. Based on this review, some cases were excluded from final calculations and results while others remained in the dataset. Justification for removal included qualitative evidence such

as aggregation of multiple tasks into one entry or misreporting units (for example recording time savings in minutes rather than hours).

A consistent method for handling time-saving outliers was established through agreement among several analysts and applied across Redbox@DBT and other AI evaluation analyses.

Qualitative analysis

Open-text diary responses were analysed by being sorted into themes and subthemes. The allocation was peer-reviewed and quality assured before finalising themes.

Qualitative interviews

The Monitoring and Evaluation team conducted qualitative interviews conducted to explore user experiences and perceptions of Redbox@DBT and gather insights on the emerging results of the diary study.

Interview design and sampling

Ten interviewees were sampled from a pool of diary respondents who had access to Redbox@DBT during the trial. Each interview represented a unique participant, with no repeat sessions. Interview followed a semi-structured format. Discussion guides were developed to support consistency across sessions, though interviewers were encouraged to adapt questions to their own style and to probe further where necessary, so the discussion guide was not cognitive tested.

Participants were sampled based on their role, satisfaction with Redbox@DBT, and reported health conditions or disabilities. These characteristics were selected for sampling as they aligned with evaluation objectives and the aims of the interviews, namely, to collect insights to complement the findings from the diary study. The sample included a mix of policy, digital, and operational roles across the department, ensuring a diverse range of perspectives.

All interviews were conducted as one-on-one online sessions, using Microsoft Teams. Participants were asked to reflect on their experiences using Redbox@DBT, including perceived benefits, challenges, and any changes in their work. Interviewers took notes during each session, and recordings were transcribed to support detailed analysis.

Analysis

A thematic analysis was conducted on the interview transcripts and notes. Responses were grouped into themes aligned with the evaluation objectives, such as usability, satisfaction and time saved. Key quotes were extracted to illustrate each theme. The analysis was quality assured and peer reviewed to ensure consistency and accuracy in interpretation.

Additional analysis

Although the comparison of Redbox@DBT with other AI tools is not within the scope of this report and does not form part of the Redbox@DBT evaluation, it is relevant to highlight that the Monitoring and Evaluation team has also employed the methodologies discussed above – specifically, performance data analysis, diary studies, and qualitative interviews – to compare Redbox@DBT's outcomes against other AI tools accessible to DBT staff during the same timeframe.

It should be noted that while not included in the initial phase of the evaluation, the Department for Business and Trade is currently developing a methodology to quantify the environmental impact of AI tools used by the department. Once this methodology is finalised, it will be used to estimate the environmental impact of Redbox@DBT.

Finally, the Monitoring and Evaluation team will use the results from this evaluation, alongside other data sources, to measure the value for money of Redbox@DBT.

Limitations of the evaluation methodology

It is important to acknowledge the limitations inherent in the evaluation methodology.

Across all approaches – diary study, surveys, and interviews – there was no control group. This absence limits the ability to attribute observed outcomes directly and solely to the intervention and increases the risk of confounding factors influencing the results.

The diary study, which generated most of the quantitative data, achieved a response rate of only 28%. There was not enough evidence to suggest the findings were representative of the department and the reliance on self-reported data introduces the possibility of unknown bias. Participants' recollections and subjective interpretations may have affected the accuracy of reported outcomes, particularly in relation to time-saving estimates, which are highly sensitive to individual perceptions.

Similarly, the before-and-after surveys also suffered from low response rates (approximately 44%) and were based entirely on self-reported data. Without a control group or baseline measures, it was not possible to control for recall bias or other distortions in participant responses.

The qualitative interviews, though valuable, were conducted with a relatively small sample and did not cover all role types within the department. It remains unclear whether any unobserved biases were present in the interview population, especially those not addressed through the stratified random sampling approach. These interviews were designed to explore themes and experiences, not to generate generalisable conclusions. They should be viewed as indicative rather than definitive.

In addition, the evaluation was not designed to assess long-term impacts of Redbox@DBT usage. No baseline data was collected prior to the trial, which prevents comparison of outcomes before and after Redbox@DBT accounts allocation.

It is also important to note that Redbox@DBT was at an early stage in the development process during this evaluation and the findings are reflective of this specific stage rather than its current one.

Finally, value for money and environmental impact assessments were not included in this evaluation, but DBT analysts are developing new methods to address them in future research.

Findings

Usage

This section covers findings from the Redbox@DBT performance dashboard during the evaluation period: from February to the end of May 2025. The data for charts A to E is derived from backend chat logs.

Chart A: number of weekly active users on Redbox@DBT

Chart A shows the number of active users per week throughout the evaluation period. Active users are defined as having initiated a new chat with Redbox@DBT. The number of active users per week peaked at 173 towards the end of the evaluation period (in May 2025) and averaged at 101 per week.

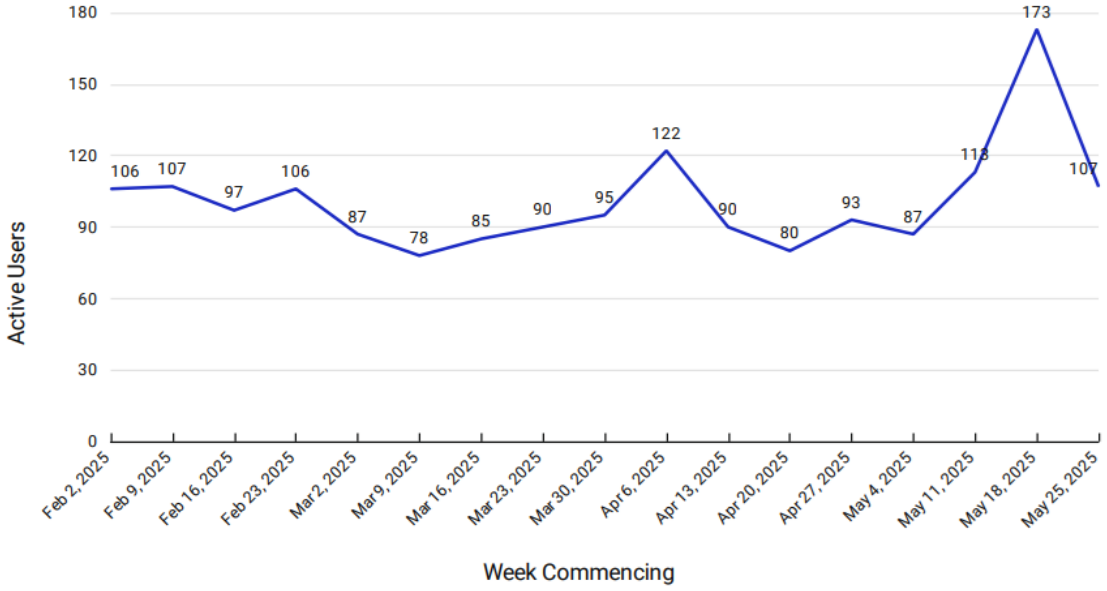


Chart B: proportion of respondents who used Redbox@DBT during the diary week

Chart B illustrates the number of diary study respondents who used Redbox@DBT. Almost 97% of respondents stated they used the tool at least once during the diary week.

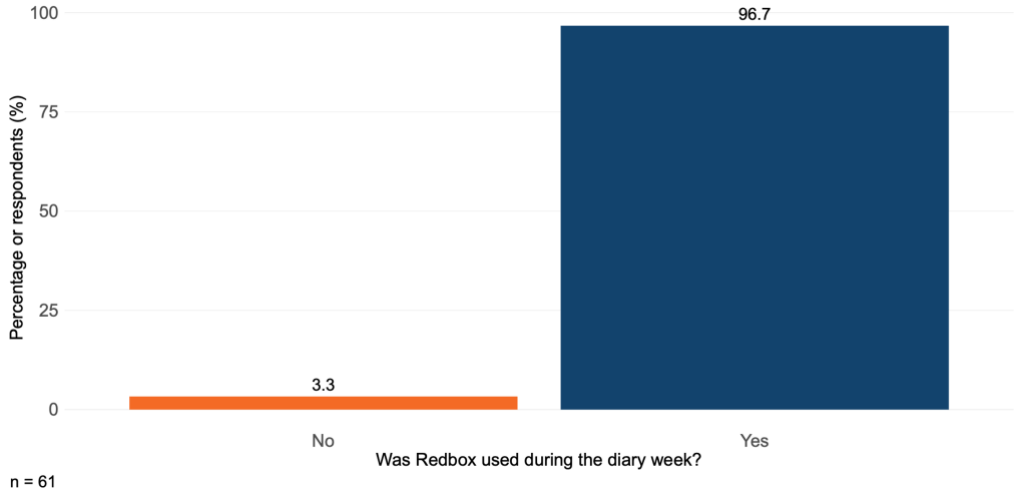


Chart C: number of new chats started by users on Redbox@DBT

Chart C covers the number of new chats started per week during the evaluation period. A new chat started is defined as sending a prompt in a new chat window to Redbox@DBT. An average of 346 chats with Redbox@DBT were initiated each week between February and May 2025, with the peak of 596 showing towards the end of the evaluation period, in the third week of May.

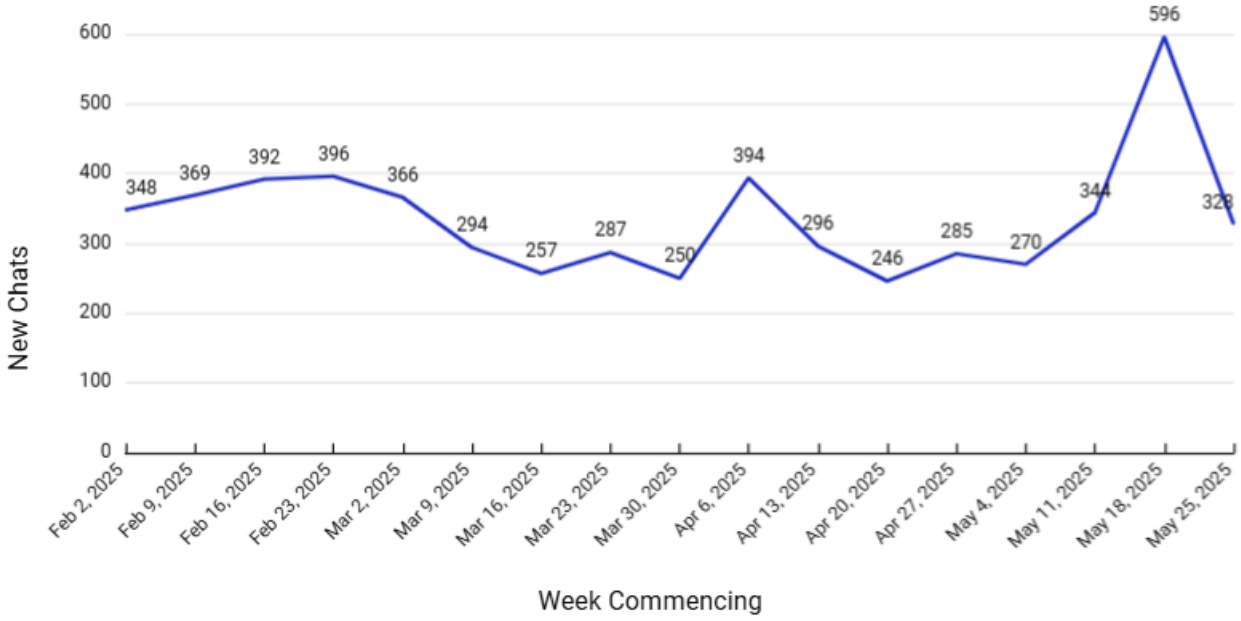


Chart D: number of prompts sent by users on Redbox@DBT

Chart D looks at the number of prompts per week sent by Redbox@DBT users during the pilot period. A prompt is defined as a piece of text written by the user and sent to Redbox@DBT which details a question, command or details of a task and, on average, 1,162 prompts were sent to Redbox@DBT each week between February and May 2025.

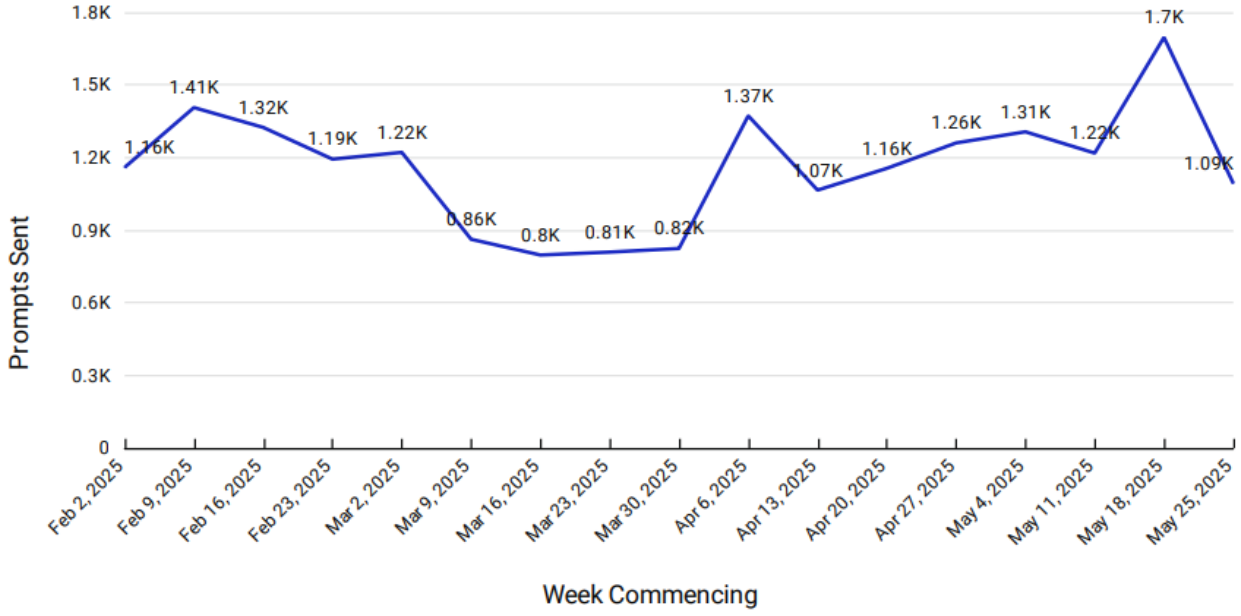


Chart E: number of documents uploaded by users on Redbox@DBT

Users can upload documents to Redbox@DBT for it to utilise in response to prompts. Chart E shows the number of documents uploaded during the evaluation period. An average of 92 documents were uploaded by users each week. The maximum recorded was 232 in February 2025.

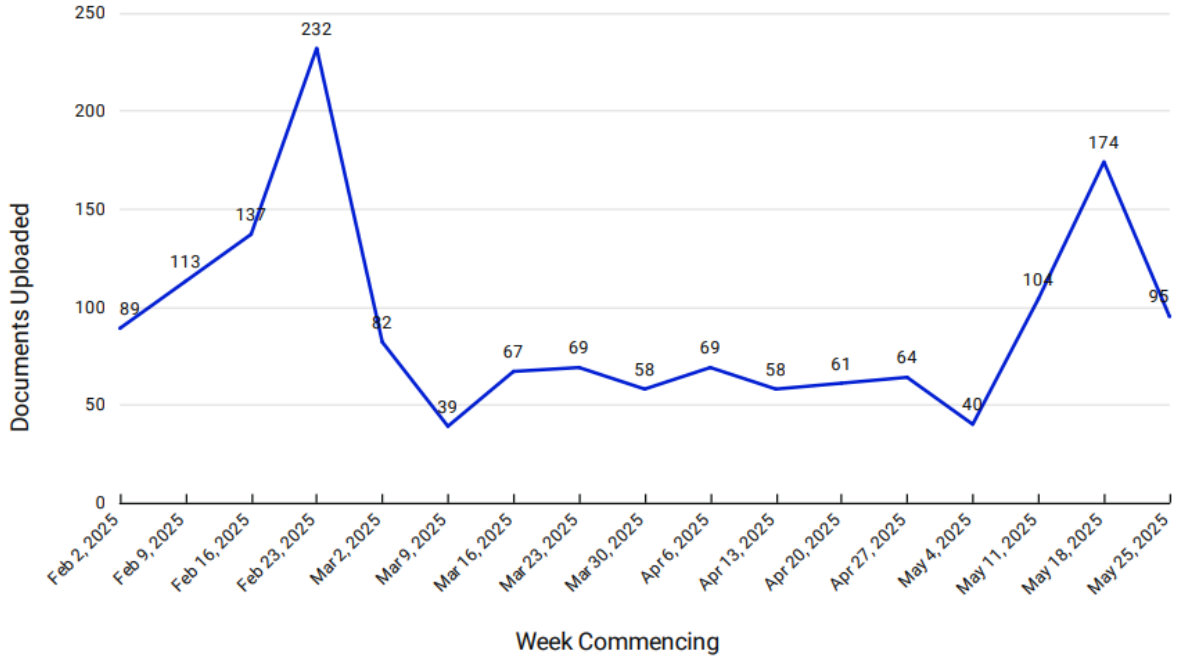
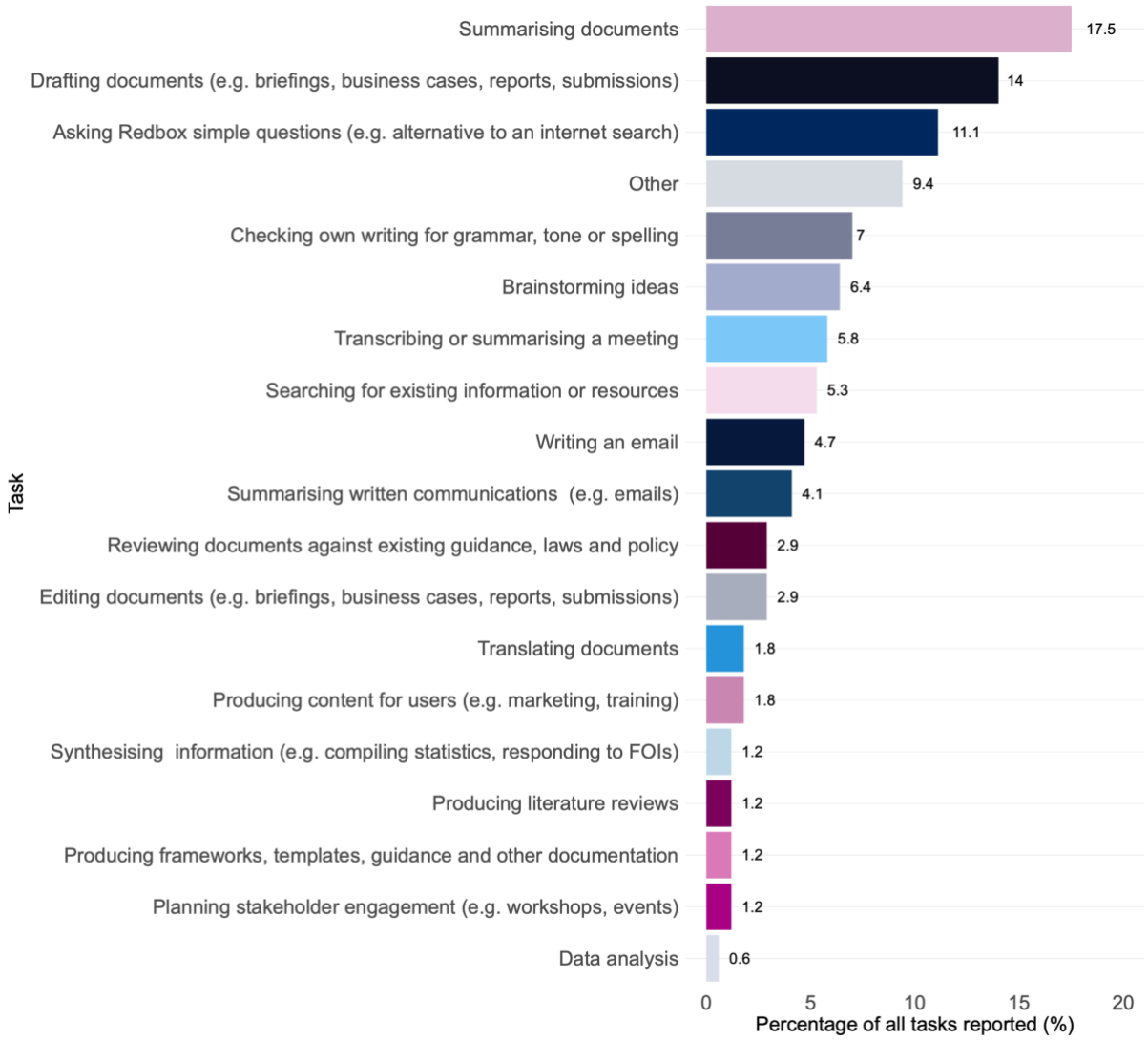


Chart F: percentage of use cases reported by diary respondents

Chart F shows the percentage of each use case out of all diary study submissions' cases. Individuals may appear multiple times if they submitted more than one use case or repeated a use case.



What task did you use Redbox for?
Please note that individuals may be counted multiple times in the chart if they reported multiple tasks, or reported a task more than once within the one-week period the diary was populated.
n = 171

Chart F summarised the findings from the diary study. The most frequently reported use cases included “summarising documents” with 17.5% of use cases, “drafting documents” with 14% and “asking Redbox@DBT simple questions” with 11.1%. Conversely, the least frequently noted use cases were “data analysis” with 0.6%, “planning stakeholder engagement”, and “producing frameworks, templates, guidance, and documentation”, both being reported in 1.2% of use cases.

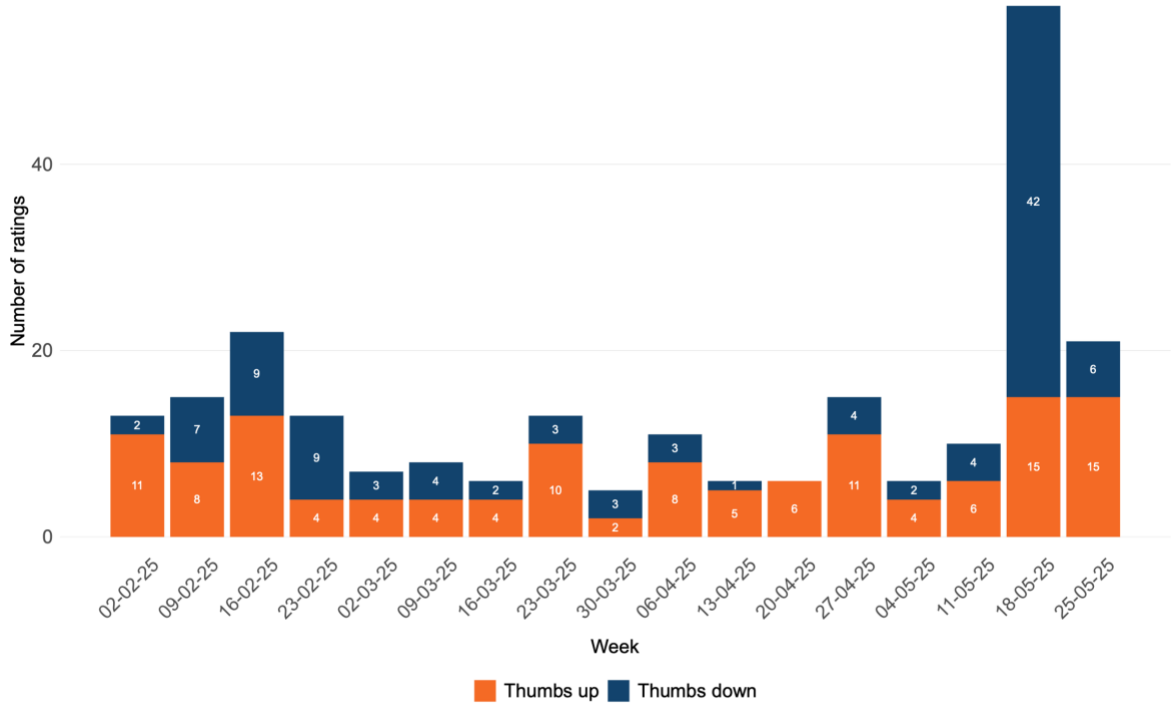
Satisfaction outcomes

This section presents both quantitative and qualitative findings on participant satisfaction with Redbox@DBT.

Redbox@DBT users can optionally click a thumb-up or thumb-down icon for every reply given by the tool to indicate whether they are satisfied with the tool outputs

Between February and May 2025, 94 trial participants rated Redbox@DBT responses. Chart G shows the breakdown of responses by week, while chart H summaries the overall rating during the pilot period.

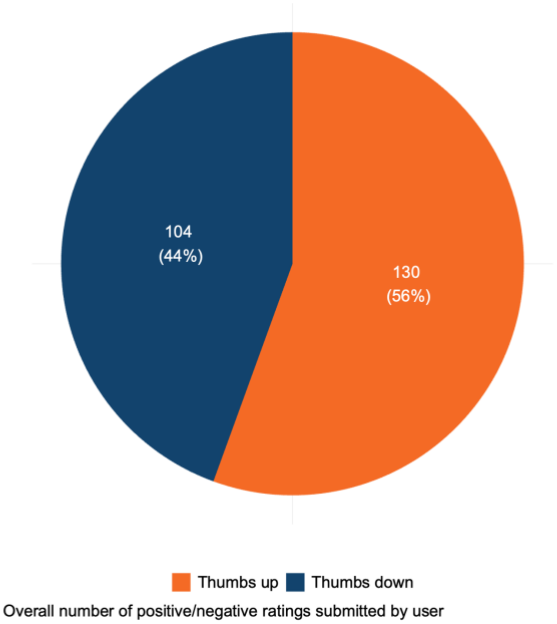
Chart G: number of positive or negative ratings submitted by users by week



Number of positive/negative ratings submitted by users by week

It should be noted that 30 negative ratings were submitted by a single user during week commencing 18 May. This was the week when the negative ratings exceeded the positive ones, with 42 negative ratings and only 15 positive ones. The following week, Redbox@DBT received 15 positive ratings and 6 negative ones.

Chart H: overall number of positive or negative ratings submitted by users

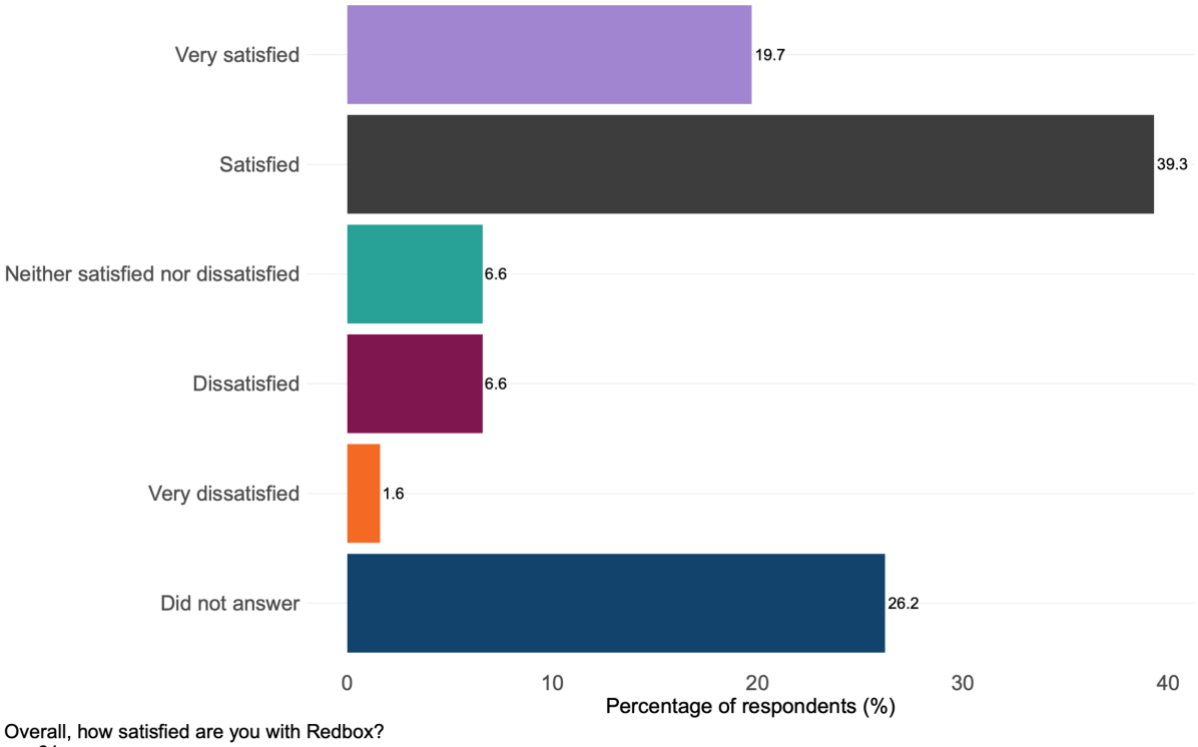


Overall number of positive/negative ratings submitted by user

Throughout the trial period, there were 56% of positive ratings and 44% of negative ones.

Chart I: overall satisfaction with Redbox@DBT

Chart I illustrates the percentage of responses from the diary study who were satisfied or dissatisfied with Redbox@DBT overall during the trial period.

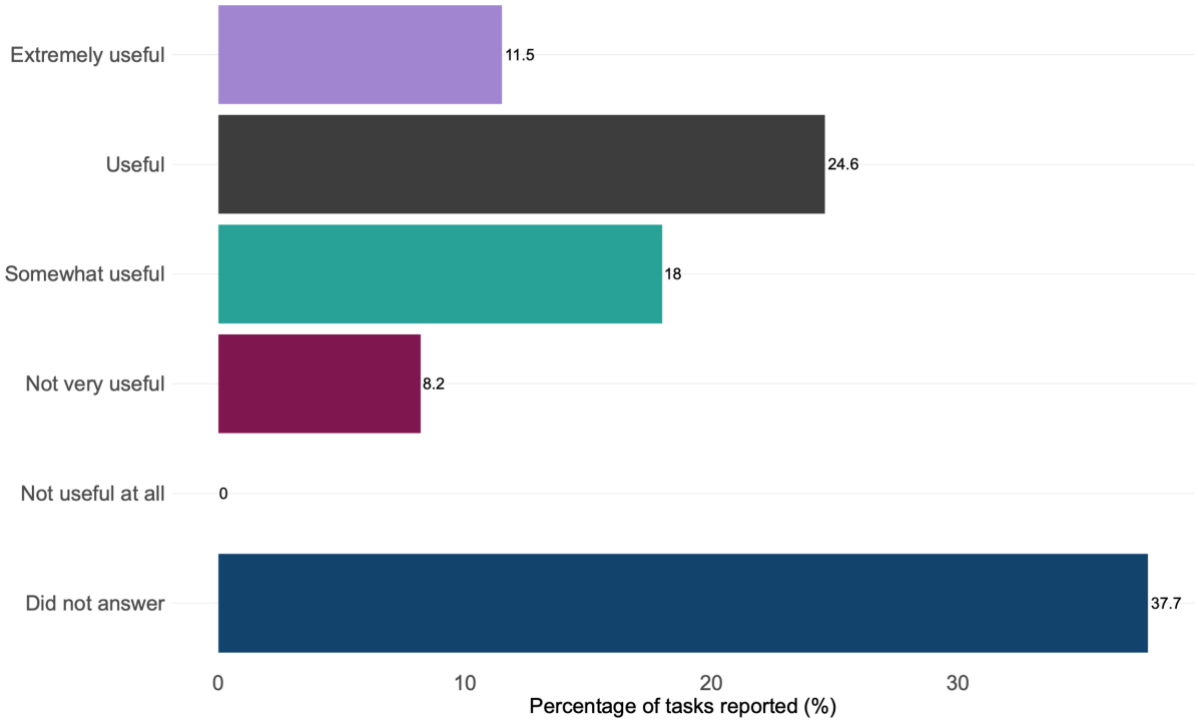


Overall, how satisfied are you with Redbox?
n = 61

Overall, respondents were positive regarding Redbox@DBT, with 59% indicating they were satisfied or very satisfied, while around 8% reported being dissatisfied. Fewer than 2% of participants reported being very dissatisfied. Notably, 26% of diary returners did not provide an answer to this question. These results are illustrated in chart I.

Chart J: overall usefulness of Redbox@DBT

Chart J shows the share of diary study respondents who found Redbox@DBT useful or not during the trial period.



Overall, how useful is Redbox in your day-to-day work activities?
n = 171

Chart J indicates that just over 8% of respondents did not find Redbox@DBT useful, while about 54% rated the tool as somewhat useful, useful or extremely useful. Notably, 38% did not answer this question.

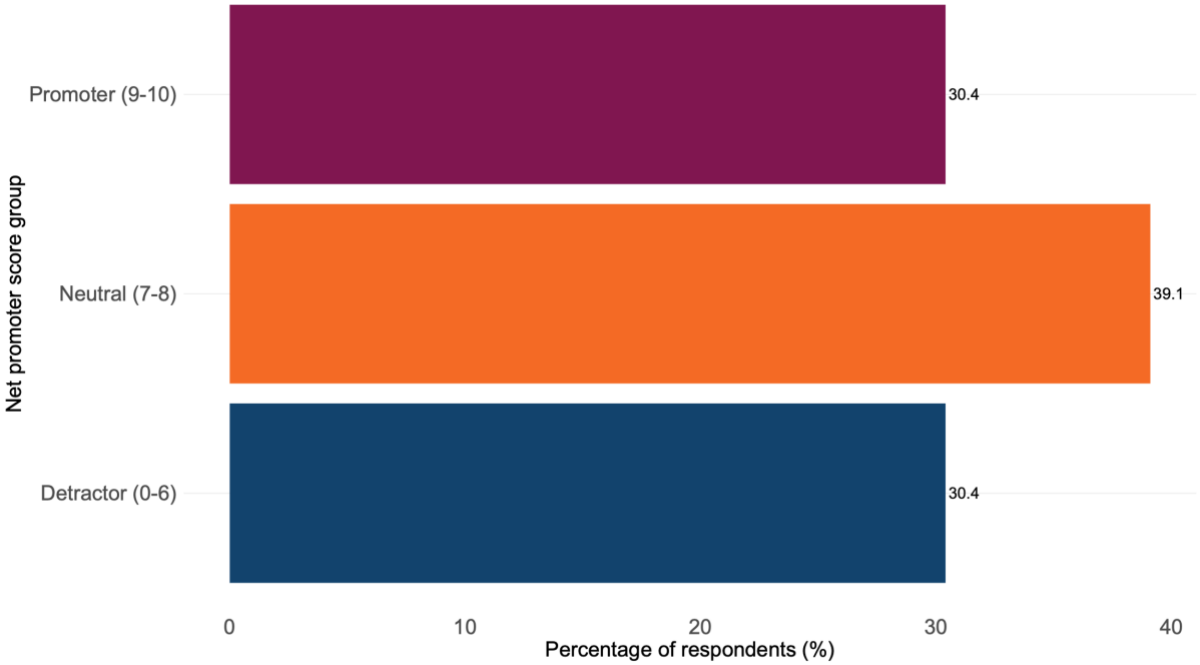
Qualitative interviews confirmed that overall satisfaction with Redbox@DBT was positive among users, with some highlighting that Redbox@DBT has become an additional resource or "problem-solving ability" in their daily work. Ease of use was also identified as a recurring theme.

"I always feel like I have a tech body inside that I could refer to if I get stuck"

"Yes, it seems like a really really straight forward tool (…) I don't think it can be any simpler".

Chart K: net promoter score

Chart K 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 question asks users "On a scale of 0 to 10, how likely is it that you would recommend the service to a colleague?".

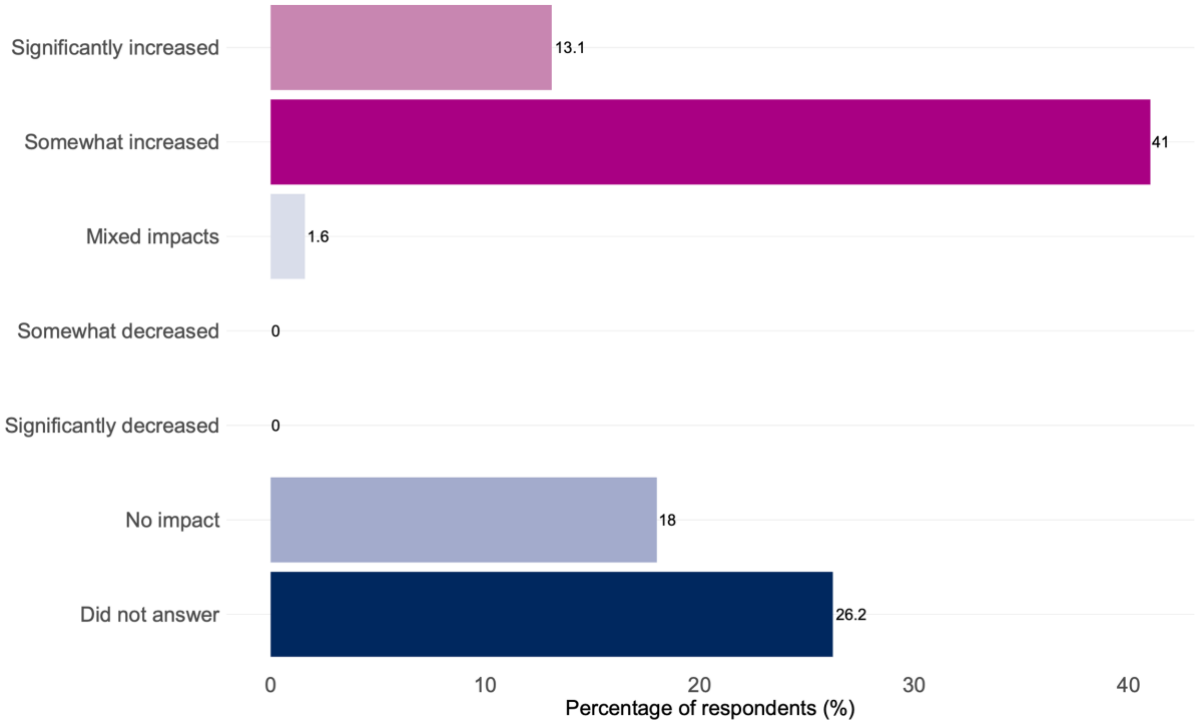


On a scale of 0 to 10, how likely is it that you would recommend Redbox to a colleague?
The Net Promoter Score is calculated by subtracting the percentage of detractors from the percentage of promoters.
Redbox's Net Promoter Score result is 0. Exlcuding respondents from the dual evaluation, Redbox's Net Promoter Score is also 0.
n =46

Diary respondents were presented with the Net Promoter Score (NPS) question, a standardised metric utilised for benchmarking across services. The NPS is determined by assessing the likelihood of users recommending the service, with the percentage of detractors subtracted from the percentage of promoters to yield the final score. For Redbox@DBT, the calculated Net Promoter Score was 0. Detailed results are illustrated in chart J: 30.4% of respondents rated Redbox@DBT as a 9 or 10 (promoters), 39.1% provided a score of 7 or 8 (neutral), and 30.4% rated Redbox@DBT between 0 and 6 (detractors).

Chart L: Redbox@DBT and job satisfaction

Chart L illustrates the percentages of diary responses in response to a question on whether access to Redbox@DBT impacted their job satisfaction.

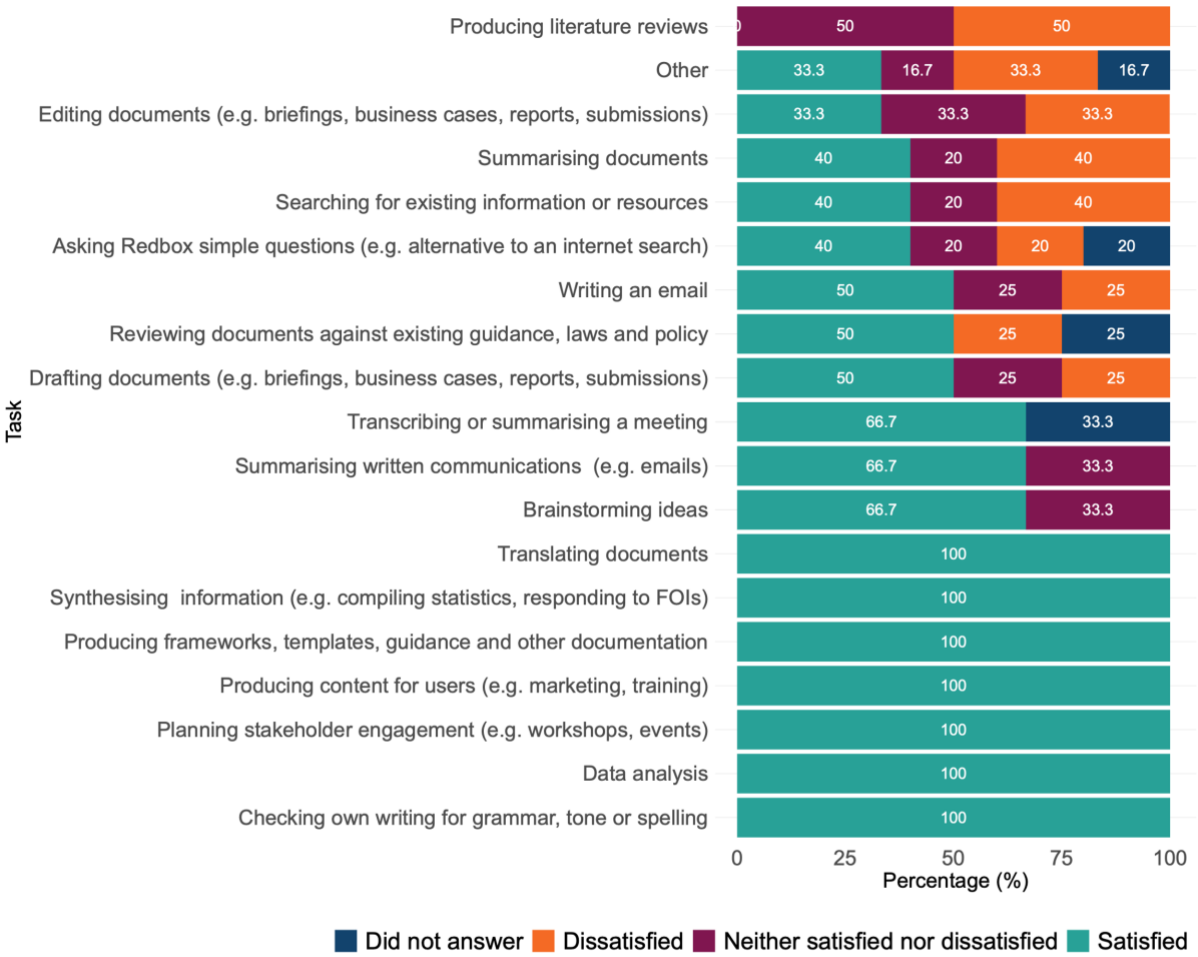


Has having a Redbox license impacted your overall satisfaction in work?
n = 61

Respondents were also asked whether Redbox@DBT had affected their job satisfaction. The results are presented in chart L. Approximately 54% of respondents indicated positive impacts, less than 2% reported mixed impacts (both positive and negative), and 26% did not answer this question. A key limitation of these findings is the absence of a comparison group, making it difficult to attribute changes solely to Redbox@DBT. Other factors may have influenced job satisfaction that could not be identified, and there is potential for unconscious bias among respondents. For instance, individuals with strong views on AI adoption within the department may have reported more pronounced effects on their job satisfaction.

Chart M: satisfaction by use case

Chart M displays the percentage distribution of satisfaction response options for each use case. For each reported use case, responses were categorised as 'satisfied', 'neither satisfied nor dissatisfied', 'dissatisfied', or 'did not answer'.



For this task, how satisfied were you with the quality of Redbox's output?
'Very Satisfied' and 'Satisfied' have been grouped into 'Satisfied'.
'Very dissatisfied' and 'Dissatisfied' have been grouped into 'Dissatified'.
n = 171
N per task: Checking own writing for grammar, tone or spelling = 2, data analysis = 1, planning stakeholder engagement = 1, producing content for users = 2, producing frameworks, templates, guidance and other documentation = 2, synthesising information = 2, translating documents = 2, brainstorming ideas = 3, summarising written communications = 3, transcribing or summarising a meeting = 3, drafting documents = 4, reviewing documents against existing guidance, laws and policy = 4, writing an email = 4, asking Redbox simple questions = 5, searching for existing information or resources = 5, summarising documents = 5, editing documents = 3, other = 6, producing literature reviews = 2

As illustrated in chart M, there are tasks that achieved 100% satisfaction rate, but only a small number of use cases were reported. Tasks like “translating documents”, “synthesising information”, “producing frameworks, templates, guidance and other documentation”, “producing content for users”, and “checking own writing for grammar tone or spelling” had 2 use cases reported each. Other tasks like “planning stakeholder engagement” and “data analysis” had only 1 use case reported. The rest of the tasks received more mixed feedback, with “producing literature reviews”, “summarising documents” and “searching for existing information or resources” having the highest dissatisfaction rate (50%, 40% and 40% respectively).

Qualitative interviews also confirmed synthesising information, drafting and translating documents to be useful Redbox@DBT tasks.

Impacts to specific user groups

The Monitoring and Evaluation team conducted statistical analysis to determine if characteristics of respondents were linked to satisfaction outcomes. Analysis assessed if there were differences in satisfaction with Redbox@DBT or Net

Promoter Score outcomes for various populations. Table 4 summarises the results where significant differences were identified.

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

Presenting only characteristics with statistically significant results. Statistical significance is indicated by asterixis (* = to 90%, ** = to 95%, *** = to 99%).

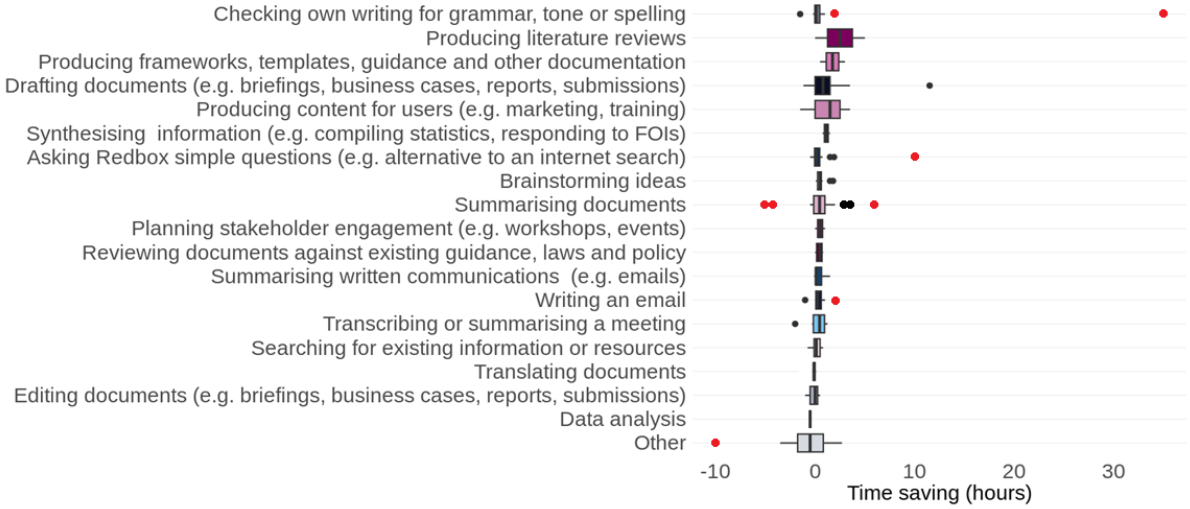
Characteristic	Group details	Satisfaction P value	Net Promoter Score P value
Grade	Assistant officer and executive office vs others	0.7472	0.2922
Grade	Higher executive officers and Fast Streamers vs others	0.7955	0.8815
Grade	Senior executive officers vs others	0.4638	0.619
Grade	Grade 7s vs others	0.9554	0.9028
Grade	Grade 6 and Senior Civil Servants vs others	0.03608 ** (less satisfied)	0.09098 * (less satisfied)

Time savings

As part of the diary, respondents were asked to record the time taken for each task with Redbox@DBT, and an estimate of the time this task would take without Redbox@DBT. Respondents were also asked if they used Redbox@DBT's output, and whether this task is something they would normally do in their role, or whether the task was "novel", meaning a task they only completed due to having access to Redbox@DBT.

Raw data was analysed to identify outliers as described in the methodology. Box plots of the raw data are shown in chart N. Analysis of outliers was undertaken and those likely to be errors were removed before calculations. Removed outliers are shown in red in the chart.

Chart N: box plots of time savings, before adjustments



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 = 171
N per task: Checking own writing for grammar, tone or spelling = 2, data analysis = 1, planning stakeholder engagement = 1, producing content for users = 2, producing frameworks, templates, guidance and other documentation = 2, synthesising information = 2, translating documents = 2, brainstorming ideas = 3, summarising written communications = 3, transcribing or summarising a meeting = 3, drafting documents = 4, reviewing documents against existing guidance, laws and policy = 4, writing an email = 4, asking Redbox simple questions = 5, searching for existing information or resources = 5, summarising documents = 5, editing documents = 3, other = 6, producing literature reviews = 2

Following outlier removal, calculations were conducted to produce the meantime saving per task. Time savings were adjusted for unused outputs, and both unused outputs and novel tasks (reducing the timesaving to 0 in both circumstances). 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 Redbox@DBT from the estimated time to complete the task without Redbox@DBT. Columns 3 and 4 show the mean time savings when the calculation is adjusted for unused Redbox@DBT 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	Mean hours saved adjusted for unused outputs and novel tasks
Producing literature reviews	2.5	2.5	2.5
Producing frameworks, templates, guidance and other documentation	1.8	1.8	1.8
Drafting documents	1.3	1.2	1.2
Producing content for users	1.2	1.2	1.2
Synthesising information	1.1	1.1	1.1
Summarising documents	1.3	1.1	0.8
Brainstorming ideas	0.6	0.6	0.6
Planning stakeholder engagement	0.5	0.5	0.5
Reviewing documents against existing guidance, laws and policy	0.4	0.4	0.4
Summarising written communications	0.5	0.5	0.4
Transcribing or summarising a meeting	0.8	0.6	0.3
Asking Redbox simple questions	0.5	0.3	0.2
Writing an email	0.3	0.2	0.2
Searching for existing information or resources	1.3	0.2	0.1
Translating documents	0.1	0.0	0.0

Task	Unadjusted mean hours saved	Mean hours saved adjusted for unused outputs	Mean hours saved adjusted for unused outputs and novel tasks
Checking own writing for grammar, tone or spelling	0.1	0.1	0.0
Editing documents	0.2	0.0	-0.1
Other	1.6	0.9	-0.3
Data analysis	2.5	2.5	-0.5

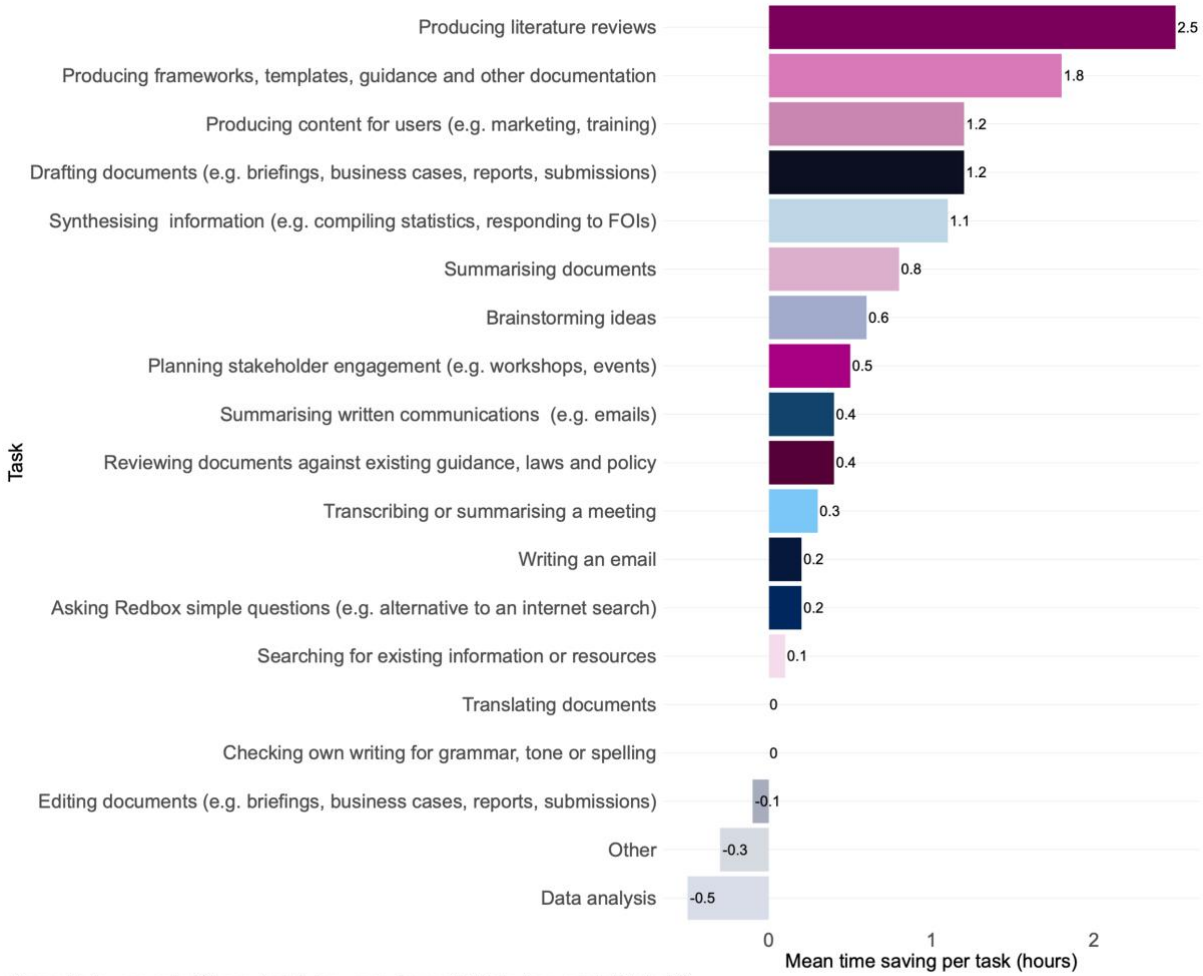
The use cases with the most pronounced time savings are “producing literature reviews”, with a mean time saving of 2.5 hours per task, and “producing frameworks” with a mean time saving of 1.8 hours per task. Both “drafting documents” and “producing content for users” had a mean time savings of 1.2 hours per task.

In contrast, “data analysis” resulted in a mean time savings of -0.5, followed by the use cases not captured in the list and labelled as “other” that resulted in mean time savings of -0.3 hours per task. “Editing documents” that resulted in a mean time savings of -0.1 hours per task. This means that these tasks took longer when Redbox@DBT was used. Assessing the data from all 3 columns, it is worth noting that “data analysis” resulted in a mean time savings of 2.5 both before the adjustments were made and after the adjustment for unused output was applied. The mean time savings only dropped to -0.5 after the adjustment for novel tasks was applied, meaning that many tasks undertaken were novel ones. These tasks were most likely completed by participants that were testing out Redbox@DBT during the initial trial period.

It is recommended to report on the right end hand-side column in table 5 which presents the adjusted time savings. These results are more representative of time saved by individuals as they account for additional work created by having access to Redbox@DBT. The results of the adjusted time savings are shown in chart O.

Chart O: mean time saving per task

Chart O 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 Redbox?`. Respondents were then asked if they used Redbox's output, and if the task is something they normally do in their role, or if they only completed the task due to having Redbox.
Time savings are adjusted to remove unused outputs and tasks not normally completed.
Outliers have been manually inspected and inaccurate observations were removed for each task.

It is important to highlight that the self-reported time saving data may contain reporting biases, for example by users who are overly enthusiastic or critical of Redbox@DBT, LLMs, or adopting AI tools. Potential biases are further discussed in the limitations section of this report. Time savings were therefore investigated further during the qualitative interviews in an attempt to understand any biases present in the data.

Interviews

Qualitative interviews confirmed that Redbox@DBT was perceived to help save time for tasks such as “producing literature reviews”, “summarising documents”, and “synthesising information”. This benefit seemed more pronounced for participants due to tasks they complete in their role. Some interviewees reported that using Redbox@DBT for these tasks increased the engagement and collaboration with colleagues from both their team and other teams.

“We were actually using it (Redbox@DBT) together for the first time, so that was really quite interesting to do it like that. And the rest of the team encourage us to use it (Redbox@DBT) and they're more than happy when we suggest to them, you know, is this task suitable for Redbox?”

Acceptable use

This section covers findings on participants' knowledge and application of departmental acceptable use policies and guidance relating to the use of Redbox@DBT. It also covers the extent to which users quality assured Redbox@DBT's outputs.

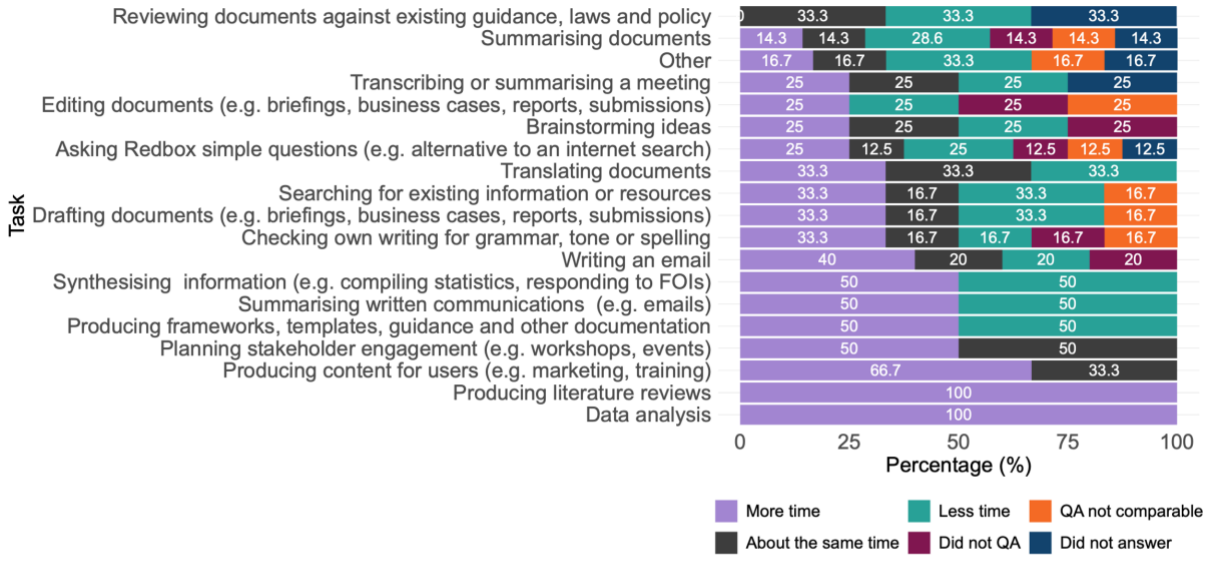
Diary study

Respondents were asked a series of questions to assess how diligent users were in reviewing outputs, and how much users adhere to acceptable use policies.

Chart P: quality assurance time comparison with and without Redbox@DBT by task

Chart P shows a comparison of quality assurance time taken by specific use cases, comparing using Redbox@DBT to completing the task without Redbox@DBT. Participants were asked about quality assurance time to assess if quality assurance was faster or slower when Redbox@DBT was used for a task. It also aimed to identify if users were being diligent in quality assuring Redbox@DBT outputs.

Responses are grouped into "more time", "less time", "about the same time", "quality assurance is not comparable", "the user did not quality assure" and "did not answer". The percentage of use cases split by each of these categories is presented.



Compared to completing this task without Redbox, how long did you spend on quality assurance?
'Much less time' and 'Slightly less time' have been grouped into 'Less time'
'Much more time' and 'Slightly more time' have been grouped into 'More time'.
n = 168
N per task: Checking own writing for grammar, tone or spelling = 2, data analysis = 1, planning stakeholder engagement = 1, producing content for users = 2, producing frameworks, templates, guidance and other documentation = 2, synthesising information = 2, translating documents = 2, brainstorming ideas = 3, summarising written communications = 3, transcribing or summarising a meeting = 3, drafting documents = 4, reviewing documents against existing guidance, laws and policy = 4, writing an email = 4, asking Redbox simple questions = 5, searching for existing information or resources = 5, summarising documents = 5, editing documents = 3, other = 6, producing literature reviews = 2

In 100% of the reported cases, both "Producing literature reviews" and "Data analysis" took more time to quality assure. This was followed by "synthesising information", "summarising written documentation", and "producing frameworks" which had an even split of 50% of tasks taking more time and 50% taking less time to quality assure. Assessing the data further, it is worth noting that for "editing

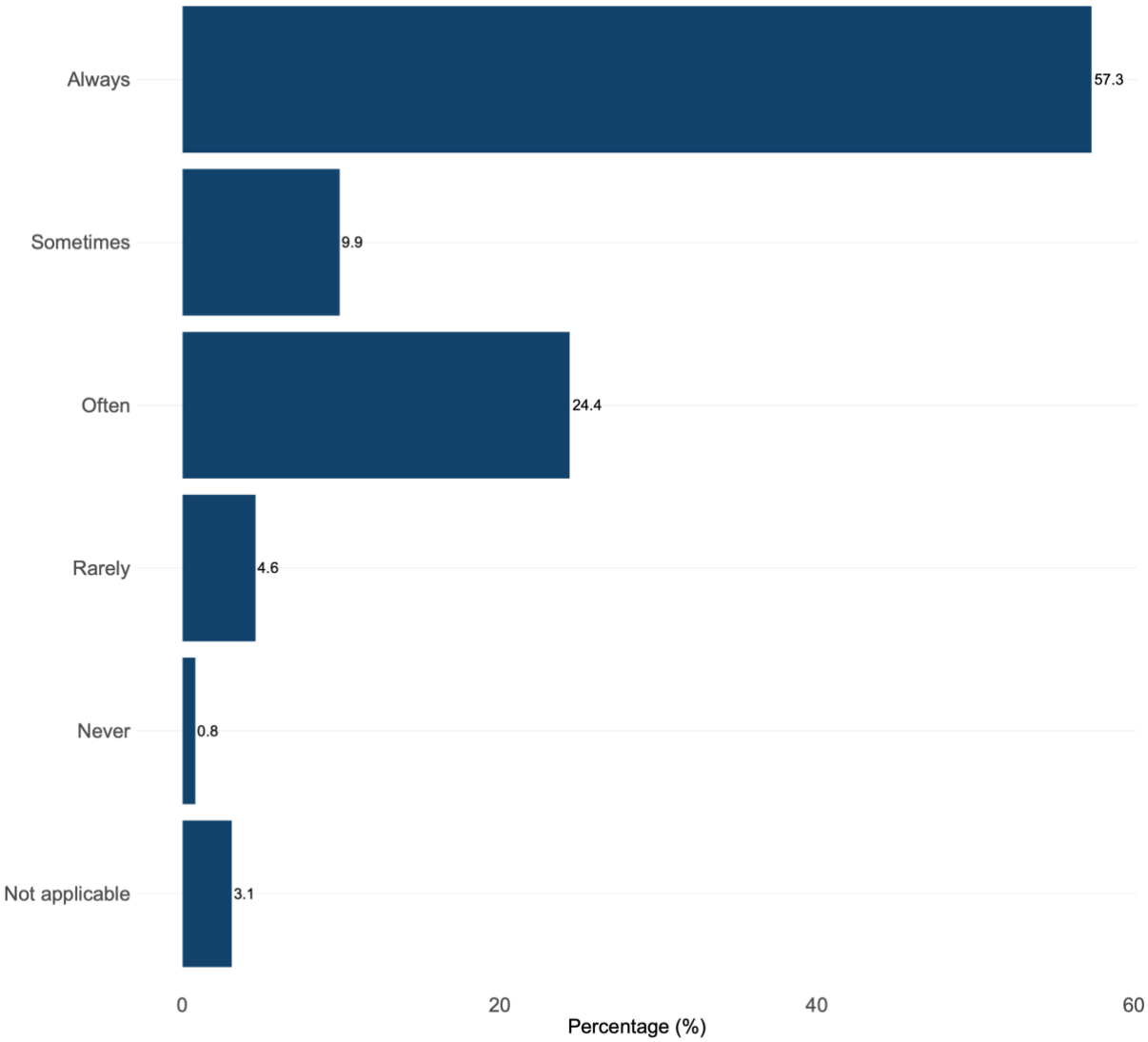
documents” it was reported in 25% of the cases that quality assurance was not comparable with outputs not produced by Redbox@DBT. This task was also not quality assured by 25% of users. The same percentage of not being quality assured is reported for “brainstorming ideas” (25%).

After survey

Respondents were asked a question related to acceptable use in the after survey too. This was to assess how often do users perform quality assurance checks on Redbox@DBT outputs.

Chart Q: quality assurance check performed

Chart Q presents the percentage of responses by frequency of quality assurance checks undertaken on Redbox@DBT outputs, collected from the after survey.



How often do you carry out quality checks on Redbox's outputs?
n = 131*
* 5 non-responses have been filtered out

Around 67% of the after survey respondents reported to always or sometimes perform quality assurance checks, and 24.4% reported to often perform these checks on outputs produced by Redbox@DBT. Only 0.8% of respondents reported

to never undertake quality assurance on outputs, and around 3.1% reported this question to not be applicable for their specific use case. These results are presented in chart P.

Interviews

The qualitative interviews assessed the topic of acceptable usage through asking questions about the frequency and nature of the quality checks performed.

Interviewees generally reported to undertake quality assurance on Redbox@DBT's output regardless of the time taken to complete these checks. The total time saved per task, including the time to quality assure, was still perceived as a time saving.

"It helps so much with time in general that for me to just slightly go over it, spend 10 or 15 minutes going through the responses and just doing a few changes is still massively helpful."

"You can't 100% rely on its output, still needs to be human intervention to ensure that it's accurate."

However, some users reported that they did not check every output depending on the task and context. Some were relying more on their ability to immediately spot errors without feeling the need to go over the content of the output. For example, personal notes were flagged as not being checked as often, especially if not shared with others.

Accuracy

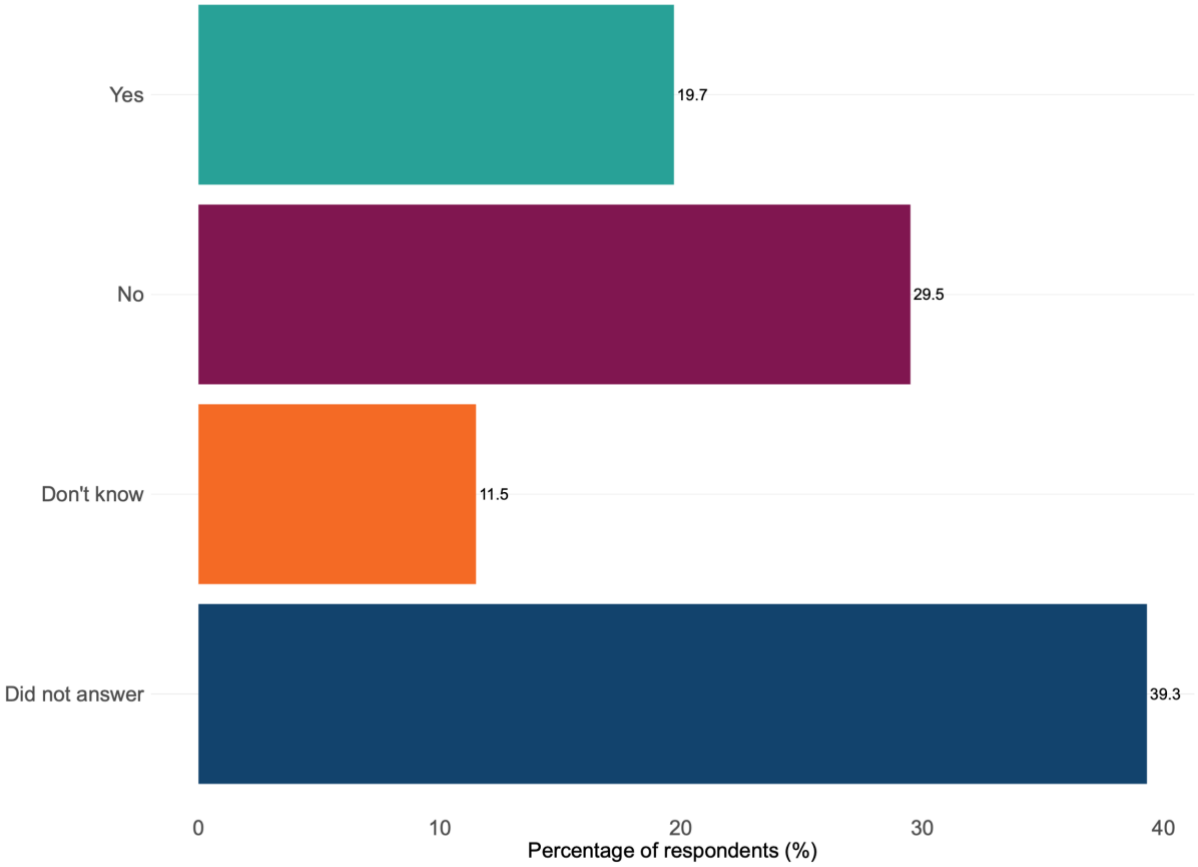
This section covers findings relating to the accuracy of Redbox@DBT and hallucinations.

Diary study

When asked to recall hallucinations, 29.5% of respondents reported they did not identify a hallucination at any point throughout the trial. However, 19.7% of respondents did identify hallucinations, and 11.5% of respondents were unsure if they encountered hallucinations. Thirty-nine percent of respondents also chose not to answer the question. It is unclear if this is because users were unable to identify hallucinations, unable to recall if they had observed hallucinations or simply did not want to answer this question. This data is presented in chart R.

Chart R: hallucinations

Chart R illustrates the percentage of diary respondents for each possible answer to the question asking if users identified hallucinations during the Redbox@DBT trial.



In your time using Redbox, have you encountered any false or misleading information presented as fact?
In the contect of Large Language Models (LLMs), a hallucination refers tot he generation of information that is incorrect, nonsensical, or entirely fabricated, despite being presented in a credible manner.
n = 61

Interviews

Hallucinations were also assessed during the qualitative interviews, with most users being aware that hallucinations were possible when using Redbox@DBT. Interviewees also reported that hallucinations could be difficult to spot. Some users flagged human expertise as an important aspect in identifying hallucinations.

"Because the way it [Redbox@DBT] laid it [the output] out was very convincing that it was correct."

Generally, hallucinations were not reported as a significant concern by the interviewees with minimal impact on satisfaction with Redbox@DBT. However, the timing of hallucinations was highlighted as influential to future use of Redbox@DBT. For example, if a user identified a hallucination during their first use of Redbox@DBT, they were anecdotally less likely to attempt to use the tool again.

Attitudes to Redbox@DBT and artificial intelligence generally

This section covers attitudes from trial participants to the use of Redbox@DBT, or AI more generally. It also covers how attitudes changed as the trial progressed and the influence of others' attitudes on Redbox@DBT use.

Before and after survey

In the before and after survey, respondents were asked questions to assess the attitudes towards Redbox@DBT. This includes trust level and perceived experience with LLMs, as well as perceived impact on their role.

An important limitation of the before and after findings is that the respondents were not prevented from using other LLM tools in their personal life, and the department was also conducting trials of other LLMs at the same time which users of Redbox@DBT could access. This means that any changes presented in the following charts cannot be attributed solely to Redbox@DBT usage.

Statistical analysis was conducted to determine whether differences between the before and after responses are statistically significant. The analysis indicated that differences in trust level were not statistically significant before versus after the pilot. On the other side, the experience using AI tools, and the perceived impact using Redbox@DBT would have on their role showed statistically significant differences. For the shift from expectations to perceived impact, the Mann-Whitney U test showed a difference in distribution, and the permutation test showed a difference in the mean. Table 6 shows the results of the variables tested.

Table 6: results of Mann-Whitney U tests and permutation tests on before and after variables

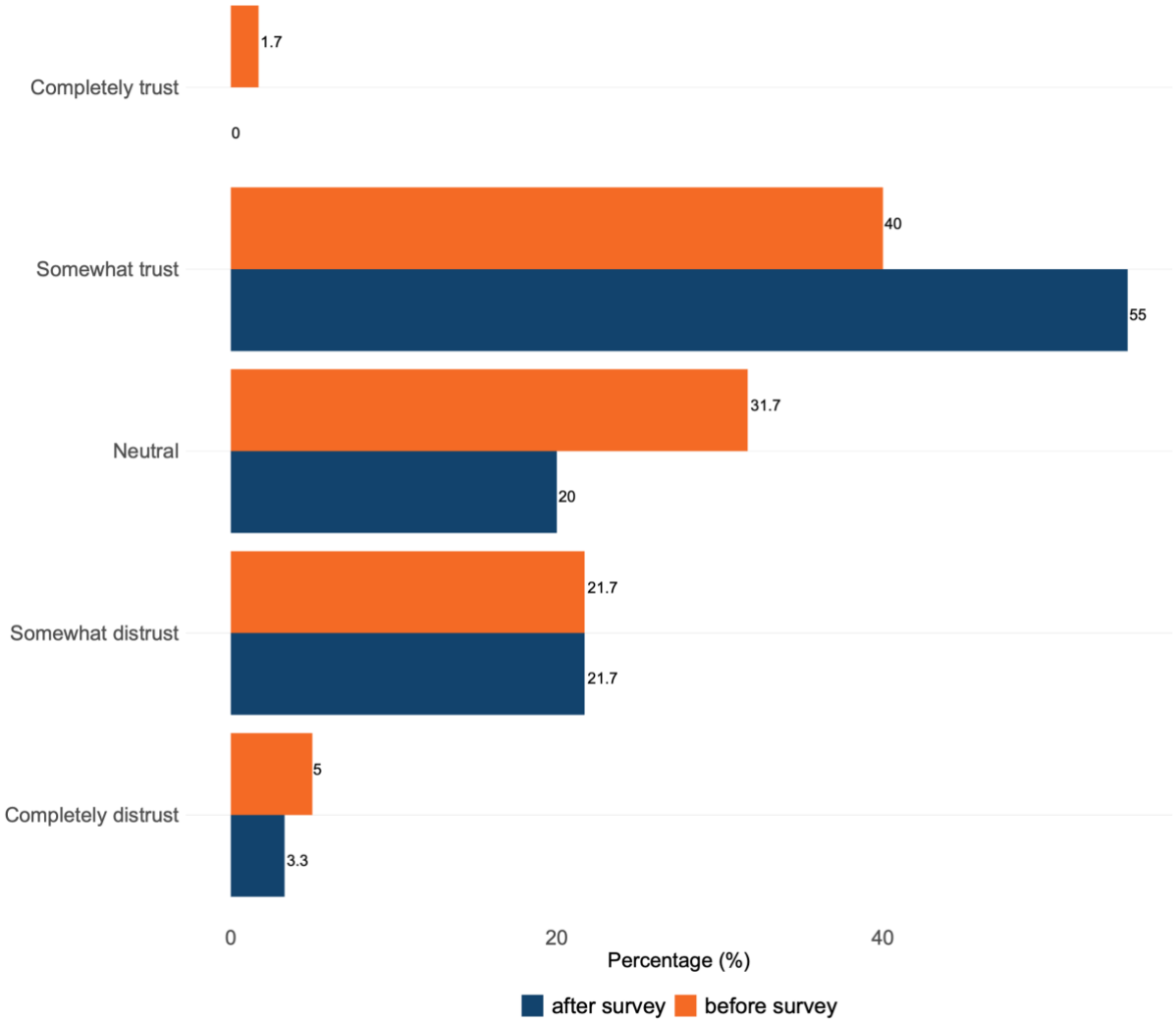
Table 6 presents the results of the Mann-Whitney U and the permutation tests. The direction of the difference is specified in brackets for the statistically significant results. Statistical significance is indicated by asterixis (* = to 90%, ** = to 95%, *** = to 99%).

Characteristic	**Mann-Whitney U test** **P value**	**Permutation test** **P value**
Experience in using AI LLM tools	0.00823*** (experience increased after trial)	0.0059*** (experience increased after trial)
Trust level	0.1986	0.1362
Expectations before and experience after	0.03782** (expectations of Redbox@DBT were higher before trial)	0.0206** (expectations of Redbox@DBT were higher before trial)

The results of the statistical analysis provide insight into which variables demonstrated meaningful change and which remained stable. The following section summarises these findings.

Chart S: level of trust with large language models

Chart S presents the percentages of before and after survey's respondents by levels of trust from "completely distrust" to "completely trust".



Which of the following statements best describes your attitude towards Large Language Model (LLM) AI tools in general?

This chart only contains people that completed both surveys.

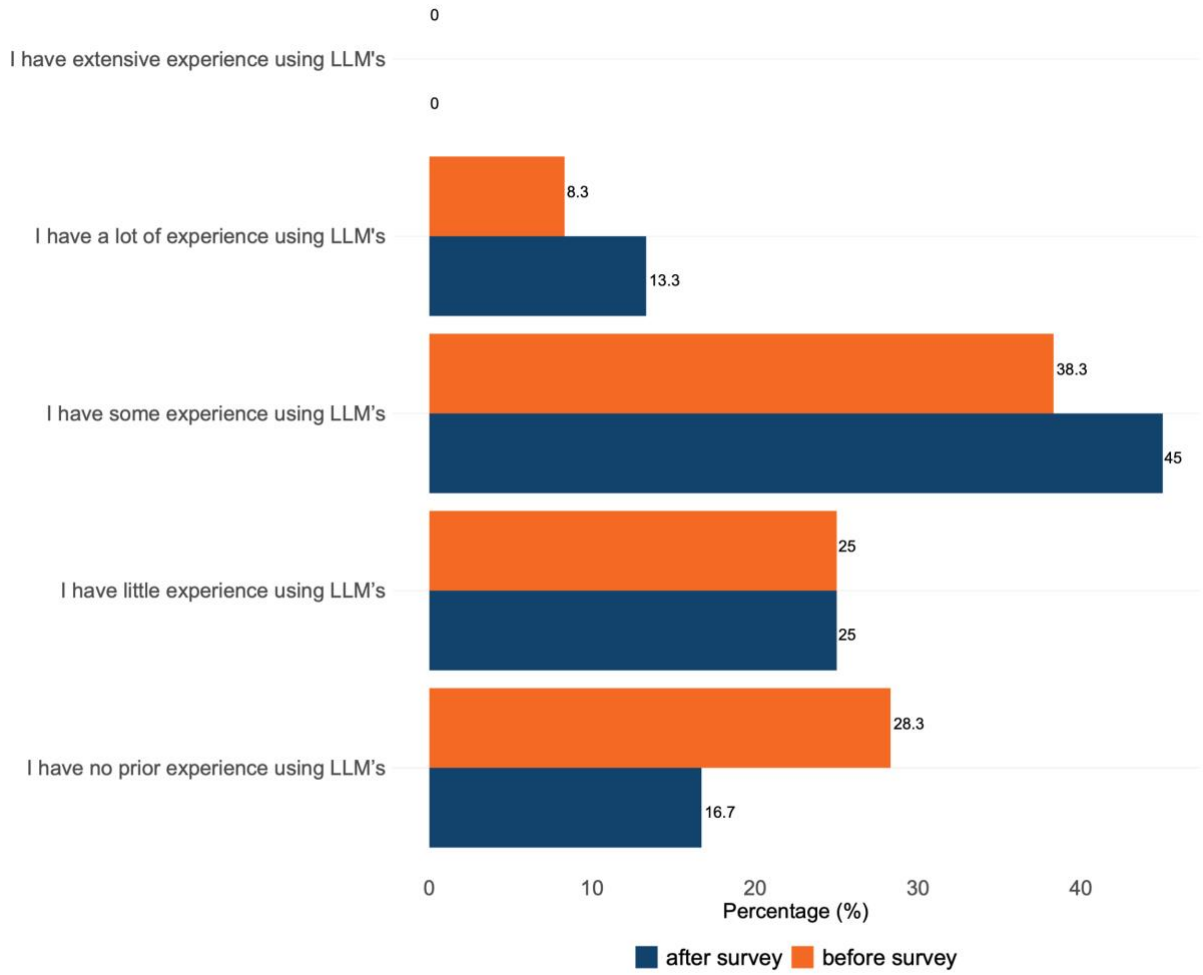
n = 60

Chart S shows a comparison between the trust level before the respondents had access to Redbox@DBT and after they used it. There was a decrease in the extreme options, with 1.7% of respondents reported to completely trust LLMs before being part of the trial and 0% of respondents completely trusting LLMs after having access to Redbox@DBT. For the completely distrust option, chart S shows a decrease between the before and after survey respondents, with 1.7% less respondents reporting to completely distrust LLMs after being part of the trial.

An increase in trust levels with LLMs was seen for the “somewhat trust” option, where 15.0% more respondents reported to somewhat trust LLMs after having access to Redbox@DBT. The increase in trust was not deemed statistically significant following the pilot.

Chart T: experience with large language models

Chart T presents the percentages of before and after respondents by each level of self-reported experience with LLMs from “I have no prior experience using LLMs” to “I have extensive experience using LLMs”.



Which statement best describes your experience with Large Language Model (LLM) AI tools prior to receiving access to Redbox?

E.g. Chat GPT, Microsoft Copilot. Please include experience from your personal life.

This chart only contains people that completed both surveys.

n = 60

Chart T shows that the respondents' confidence and perceived experience in using AI and LLMs has increased over time. Less respondents reported no prior experience, 16.7% after having access to Redbox@DBT compared to 28.3% before. Additionally, 6.7% more respondents reported some experience using LLMs, and 5.0% more reported a lot of experience using LLMs after taking part in the trial. The differences in the level of experience using an AI LLM tool identified between the before and after respondents were found to be statistically significant.

Both the trust levels and the experience with LLMs were confirmed during the qualitative interviews. Interviewees reported an increased level of confidence after using Redbox@DBT more often and after taking the time to get used to the tool.

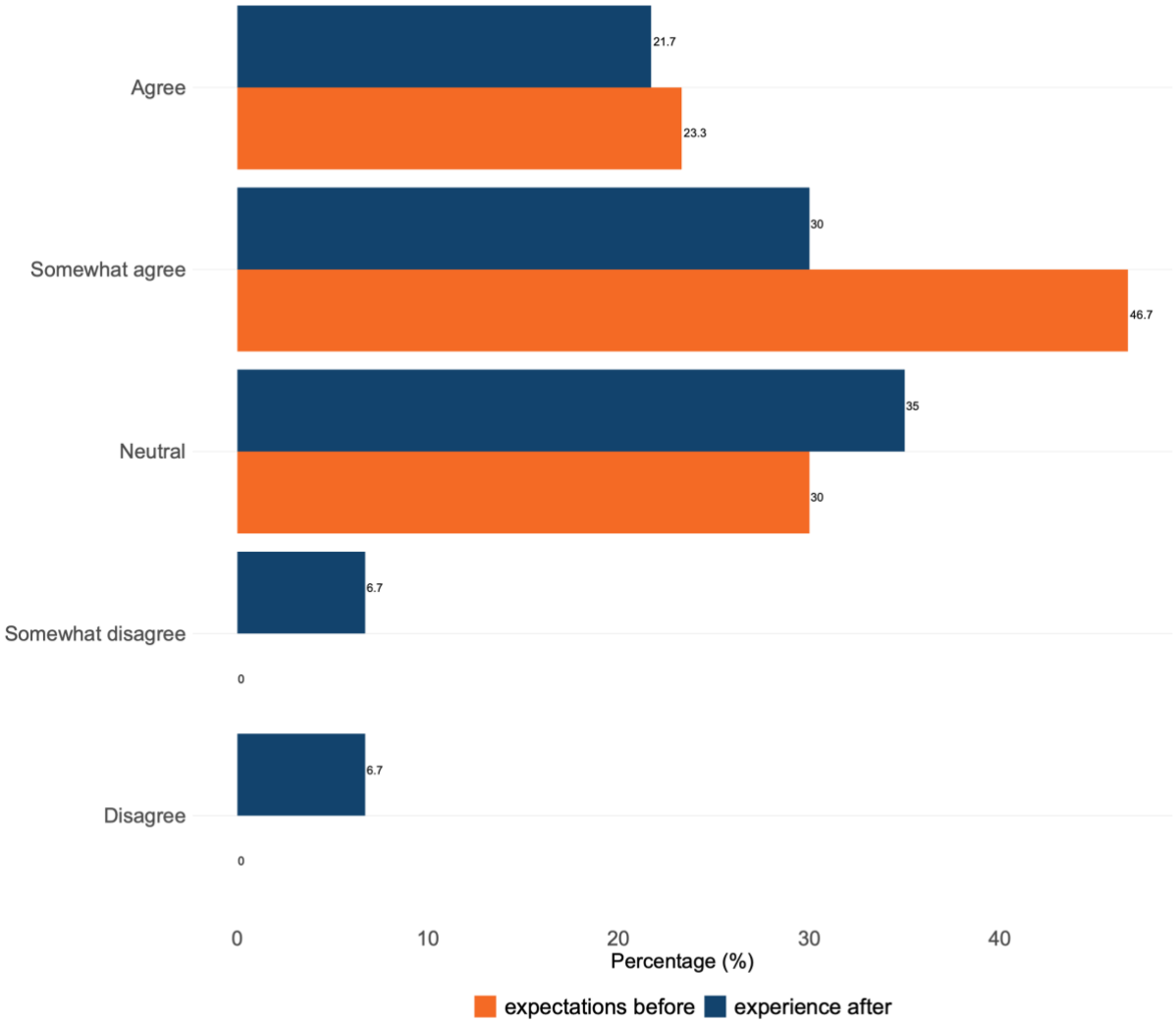
"The more I use it the more I trust it."

Interviewees generally suggested that their trust in Redbox@DBT's output to be comparable to the level of trust they would have in a colleague's output.

To assess changes in perceptions, the Monitoring and Evaluation team compared responses to 2 related statements across the before and after surveys. Prior to implementation, participants indicated their level of agreement with the statement: “I expect Redbox will enhance my ability to perform my role.” After implementation, they responded to: “Redbox has enhanced my ability to perform my role.” This approach allowed us to measure the shift from expectation to actual experience and identify whether perceived benefits aligned with initial expectations.

Chart U: expectations versus experience

Chart U presents the percentages of before and after survey respondents for 2 different questions. Respondents were asked to report how much or how little they agree with the following statements: “I expect Redbox to enhance my ability to perform my role” (in the before survey) and: “Redbox has enhanced my ability to perform my role” (in the after survey).



I expect Redbox will enhance my ability to perform my role vs Redbox has enhanced my ability to perform my role

This chart only contains people that completed both surveys.

n = 60

As shown in chart U, expectations were high initially, with most respondents anticipating positive impact on their ability to perform their role. After having access

to Redbox@DBT, less respondents reported to agree that the tool has enhanced their ability to perform their role. Respondents who agreed decreased slightly from 23.3% in the before survey to 21.7% in the after survey. The “Somewhat agree” option decreased from 46.7% in the before survey to 30.0% in the after survey. Many respondents reported more neutral or stronger disagreement options after having access to Redbox@DBT. Statistical testing showed differences in both the mean values, and the distribution of responses between the before and after surveys, for this specific question.

Nevertheless, during the qualitative interviews, users reported that Redbox@DBT had a minimal yet positive impact on their role and working life. Users also mentioned that Redbox@DBT had potential for further improvement.

Concerns and barriers of using Redbox@DBT and AI disclaimers

This section covers any issues and barriers associated with the use of Redbox@DBT raised by evaluation participants.

Interviews

One of the concerns highlighted during the qualitative interviews was AI tools being used without enough human input or expertise. This was also mentioned when discussing hallucinations; some participants identified lack of topic related knowledge and accuracy during the quality assurance process as a concern. It’s worth noting that since these interviews took place, more AI related training has been made available to staff to address this issue.

Hallucinations or errors were identified during the interviews as affecting adoption of Redbox@DBT, especially if experienced at the beginning of the trial. There was no other aspect identified that influenced the adoption of Redbox@DBT, however other aspects that weren’t identified in the evaluation may impact adoption rates of AI.

Security issues were mentioned by interviewees, some of them believed that security concerns were being dealt with before reaching the end user and are not something individual users need to worry about. Others reported to avoid using Redbox@DBT in certain meetings to protect the trust built between colleagues.

“I’m just the end user, so I’m trying not worry about the things outside of my control.”

“I wouldn’t use (Redbox@DBT) in a very confidential meeting where there was a trust relationship.”

The evaluation interviews also assessed the user’s attitudes to using AI disclaimers and helped identify some of the barriers around including these. Respondents were split between raising multiple concerns about using an AI disclaimer and reporting to be happy to use it if needed.

One of the reasons behind participants’ decision not to use the disclaimer was the fact that they amended the output, therefore it was not solely produced by the AI tool. Also, the end use of their work influenced this decision; users reported to not feel the need to use the AI disclaimer if the output was not to be shared publicly.

“I didn't have to [use the AI disclaimer]. I didn't produce anything for the general public. I didn't feel the need to mention that I use AI to help draft an email because I add to it, I change it, I use it as a starting point, not just take it as it is.”

A barrier reported by interviewees to using an AI disclaimer was other people's perception of it. This is perceived to be negative and some participants reported that it would affect their credibility and the way that their work is perceived by others, especially by seniors. Additionally, others were not using the disclaimer as they felt it was not clear enough when they should use it.

"You know you'd be looked down on as you're using the shortcut, but very soon everybody will be using this as default."

"Not wanting to look like you don't know what you're talking about, or that you want help with something, or that you're nervous about talking to him."

Conclusion

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

Diary study data and qualitative interviews identified generally positive levels of satisfaction from users, with more than half of respondents reporting Redbox@DBT had a positive impact on job satisfaction.

The largest time savings were reported for document-related tasks such as producing literature reviews and frameworks and drafting documents. Time savings are evident, albeit in small quantities. Some tasks, such as data analysis and editing documents, resulted in negative time savings, suggesting Redbox@DBT may introduce additional workload for certain activities.

Qualitative interviews reinforced both the satisfaction and time savings findings from the diary study, highlighting Redbox@DBT's ease of use and its role as a valuable resource for problem solving and daily tasks.

Most users reported performing quality assurance checks on Redbox@DBT's outputs. While some tasks required more quality assurance time, there was still an overall time saving perceived by respondents. Hallucinations were also identified by some of the respondents but not seen as major concern.

Confidence in using AI tools increased among participants as the trial progressed. The areas for further consideration identified during the evaluation included reliance on AI without human input and potential security concerns. Some users also reported to be hesitant to use AI disclaimers due to perceived negative perceptions from colleagues and managers.

Findings from the evaluation will feed into internal recommendations and next steps regarding the use and further development of Redbox@DBT.

Annex

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, Redbox@DBT is referred to as Redbox, which was the term being used to describe the tool in the department at the time. Questions 11 to 35 were asked to respondents for each task they reported using Redbox@DBT 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 Redbox diary study?	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 Redbox log-in?	Drop down choice
8	What is your sex?	Drop down choice
9	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
10	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

Question number	Question	Answer format
11	What task did you use Redbox for?	Drop down choice
12	If ‘other’, please describe the task you used Redbox for	Free text field
13	Why did you choose to use Redbox to complete this task?	Drop down choice
14	If ‘other’, please specify	Free text field
15	How important would you consider this task for your overall role?	Drop down choice
16	Is this a task you would have carried out this week if you didn’t have a Redbox account?	Drop down choice
17	How frequently do you perform this task in your current role?	Drop down choice
18	How many prompts did it take for Redbox to produce your desired output for this task?	Drop down choice
19	Approximately how many hours did it take to complete this task?	Data-validated field requiring a number and allowing decimals
20	Approximately how many hours would you expect this take to have taken without using Redbox?	Data-validated field requiring a number and allowing decimals
21	Did you use Redbox’s output for this task?	Drop down choice
22	For this task, how useful was Redbox’s output?	Drop down choice
23	For this task, how accurate was Redbox's output?	Drop down choice

Question number	Question	Answer format
24	If you encountered any false or inaccurate information for this task please provide more detail here	Free text field
25	For this task, how complete was Redbox's output?	Drop down choice
26	If Redbox's output did not address all aspects of your question, please provide more detail here	Free text field
27	If you provided Redbox with specific instructions for this task, did the output follow these?	Drop down choice
28	Please use this box to provide more detail on the instructions you gave Redbox	Free text field
29	For this task, how did you find the level of detail in Redbox's output?	Drop down choice
30	For this task, how clear and easy to understand was Redbox's output?	Drop down choice
31	For this task, how satisfied were you with the quality of Redbox's output?	Drop down choice
32	Did you quality assure (QA) Redbox's output?	Drop down choice
33	Compared to completing this task without Redbox, how long did you spend on quality assurance?	Drop down choice
34	What kind of manual changes did you make to Redbox's output, if any?	Free text field
35	Do you have any additional comments and observations?	Free text field
36	Have you used Redbox since receiving a license?	Drop down choice
37	If you did not use Redbox since receiving an account, please explain why here	Free text field

Question number	Question	Answer format
38	Have you attended live training or watched any recorded training sessions organised by the AI Enablement Team about Redbox?	Drop down choice
39	If you attended any training sessions on Redbox, how useful were they?	Drop down choice
40	Outside of the AI Enablement Team training sessions, approximately how many hours have you spent training, learning to use, or upskilling on Redbox?	Drop down choice
41	Overall, how useful is Redbox in your day-to-day work activities?	Drop down choice
42	Have you encountered any technical issues in using or accessing Redbox?	Free text field
43	In your time using Redbox, have you encountered any false or misleading information presented as fact?	Drop down choice
44	Are there any tasks which Redbox was unable to perform, but which you would like to use AI tools for in the future?	Free text field
45	Has having a Redbox license impacted your overall satisfaction in work?	Drop down choice
46	On a scale of 0 to 10, how likely is it that you would you recommend Redbox to a colleague?	Drop down choice
47	Overall, how satisfied are you with Redbox?	Drop down choice
48	Has having a Redbox account affected how much you socialise and interact with colleagues in work?	Drop down choice
49	Do you have any further comments, feedback or reflections on your experience of using Redbox which you would like to share?	Free text field

Annex B: before and after question list

Before survey

Question number	Question	Answer format
1	Which statement best describes your experience with Large Language Model (LLM) AI tools prior to receiving your Redbox seat?	Multiple choices
2	Which of the following statements best describes your attitude towards Large Language Model (LLM) AI tools in general?	Multiple choices
3	How little or how much do you agree with the following statement: “I expect Redbox will enhance my ability to perform my role”:	Multiple choices
4	Which of the following tasks do you perform in your current role?	Multiple choices
5	How important would you consider this task to your overall role?	Multiple choices
6	How frequently do you perform this task in your current role?	Multiple choices
7	How often do you carry out quality checks for this task?	Multiple choices
8	Thinking about the last time you undertook this task, approximately how long did it take you to complete?	Free text field
9	Thinking about the last time you undertook this task, approximately how long did you spend on quality checks?	Free text field
10	Which DG area do you belong to?	Multiple choices
11	Which directorate do you belong to?	Multiple choices
12	How long have you been in your current role?	Multiple choices
13	What is your grade in the UK Civil Service?	Multiple choices
14	Which of the following professional roles best reflects the type of work you do in your main job?	Multiple choices
15	Which of the following categories best reflects the type of work you do in your main job?	Multiple choices
16	What is your age?	Multiple choices
17	How would you describe yourself?	Multiple choices
18	Do you have any physical or mental health conditions or illnesses lasting, or expected to last 12 months or more, that	Multiple choices

Question number	Question	Answer format
	affect your ability to carry out your day-to-day work activities?	

After survey

For questions 9 to 11, respondents were given a list of tasks, and they were requested to answer the questions for each of them.

Question number	Question	Answer format
1	Did you fill in the before survey prior to receiving access to Redbox?	Multiple choices
2	Which statement best describes your experience with Large Language Model (LLM) AI tools after receiving access to Redbox?	Multiple choices
3	Which of the following statements best describes your attitude towards Large Language Model (LLM) AI tools in general?”:	Multiple choices
4	How little or how much do you agree with the following statement: "Redbox has enhanced my ability to perform my role"?	Multiple choices
5	Overall, how satisfied are you with Redbox?	Multiple choices
6	Please use the box below to explain your answer to the previous question.	Free text field
7	On a scale of 0 to 10, how likely is it that you would recommend Redbox to a colleague?	Multiple choices
8	How often do you carry out quality checks on Redbox’s output?	Multiple choices
9	Did you use Redbox for this task?	Multiple choices
10	Why did you decide not to use Redbox for this task?	Free text field
11	Is there anything else you would like to tell us about your experience with Redbox?	Free text field

Annex C: discussion guide questions for interviews

Background

1. Could you tell me a bit about what your day-to-day role looks like?
2. In general, how confident would you say you are in learning to use new technologies?

Engagement and uptake of the tool

3. Had you heard about Redbox or about the development of an in-house LLM before receiving your account?
4. After receiving your account, how eager were you to begin using Redbox?
5. When approaching a task, what informed your decision to use or not use Redbox?
6. How would you describe your experience of using Redbox's features?
7. When you did attempt to use Redbox, did you encounter any barriers or difficulties?
8. Do you feel the training available adequately prepared you to use Redbox?

User and role specific benefits or challenges

9. What are the main challenges you face in your daily work?
10. Was Redbox useful for any parts of your role?
11. Are there any parts of your role that Redbox wasn't useful for?

Acceptable use and diligence

12. Coming on to talk about some risks now, what do you think could go wrong when using AI?
13. To what extent would you say you trust Redbox?
14. How would you describe the quality of the Redbox's outputs?
15. After using Redbox for a task, what was your general approach to reviewing outputs?
16. In your time using Redbox, did you notice any inaccurate or misleading information?
17. How confident did you feel in using Redbox with materials at various levels of security classification?
18. As you were using the tool, did you feel adequately prepared to navigate potential risks around inaccuracies or information management?

Satisfaction and forward look

19. Overall, how would you say that Redbox has affected your day-to-day working life?
20. As the trial progressed, did your view of Redbox change?
21. Coming back to the expectations from the beginning, in what ways did Redbox manage to meet these expectations?
22. Would you wish to continue using Redbox in your work?

23. Going forwards, how would you like to see the department approach the use of Redbox or AI tools more broadly?

Annex 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 2 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 2 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.

Legal disclaimer
Whereas every effort has been made to ensure that the information in this document is accurate, the Department for Business, Innovation, Science 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.

© Crown Copyright 2026
You may re-use this publication (not including logos) free of charge in any format or medium, under the terms of the Open Government Licence.
To view this licence, visit: nationalarchives.gov.uk/doc/open-government-licence/version/3.

Where we have identified any third-party copyright information in the material that you wish to use, you will need to obtain permission from the copyright holder(s) concerned.

Published by
Department for Business, Innovation, Science and Trade
28 September 2026