Australian Government
Digital Transformation Agency

Evaluation of the whole-of-government trial of Microsoft 365 Copilot

Summary of evaluation findings

dta.gov.au

Digital Transformation Agency

The Digital Transformation Agency has tried to make the information in this product as accurate as possible. However, it does not guarantee that the information is totally accurate or complete. Therefore, you should not solely rely on this information when making a commercial decision.

Digital Transformation Agency is committed to providing web accessible content wherever possible.

If you are having difficulties with accessing this document, please email: ai@dta.gov.au.

Version: 1.1

Contents

Executive summary

Preface

The uptake of publicly available generative artificial intelligence (AI) tools, like ChatGPT, has grown. In the few years since its public introduction, generative artificial intelligence has become available and accessible to millions.

This meant the Australian Public Service (APS) had to respond quickly to allow its workforce to experiment with generative AI in a safe, responsible and integrated way. To make this experimentation possible, an appropriate generative AI tool needed to be selected.

This decision was dependent on:

- how swiftly and seamlessly the tool could be deployed for rapid APS experimentation purposes
- the ability for staff to experiment and learn using applications familiar to them.

One solution to enable the APS to experiment with safe and responsible generative AI was Microsoft 365 Copilot (formerly Copilot for Microsoft 365). On 16 November 2023, the Australian Government announced a 6-month whole-of-government trial of Copilot. Copilot is a supplementary product that integrates with the existing applications in the Microsoft 365 suite and it's nested within existing whole-of-government contracting arrangements with Microsoft. This made it a rapid and familiar solution to deploy.

Broadly, the trial and evaluation tested the extent the wider promise of generative AI capabilities would translate into real-world adoption by workers. The results will help the Australian Government consider future opportunities and challenges related to the adoption of generative AI.

This was the first trial of a generative AI tool in the Australian Government. The future brings exciting opportunities to understand what other tools are available to explore a broad landscape of use cases.

Overarching findings

There are clear benefits to the adoption of generative AI but also challenges with adoption and concerns that need to be monitored.

Copilot use was moderate and focused on a few use cases.

Use of Copilot was moderate. However most trial participants across classifications and job families were optimistic about Copilot and wished to keep using it.

- Only a third of trial participants across classifications and job families used Copilot daily.
- Copilot was predominantly used to summarise information and re-write content.
- Copilot in Microsoft Word and Teams were viewed favourable and used most frequently.
- Access barriers prevented Copilot use in Outlook.

Perceived improvements to efficiency and quality

Trial participants estimated time savings of up to an hour when summarising information, preparing a first draft of a document and searching for information.

The highest efficiency gains were perceived by APS levels 3-6, Executive Level (EL) 1 staff and ICT roles.

The majority of managers (64%) perceived uplifts in efficiency and quality in their teams.

40% of trial participants reported their ability to reallocate their time to higher-value activities such as staff engagement and strategic planning.

There is potential for Copilot to improve inclusivity and accessibility in the workplace and in government communication.

Adoption requires a concerted effort to address barriers

There are key integration, data security and information management considerations agencies must consider prior to Copilot adoption, including scalability and performance of the GPT integration and understanding of the context of the large language model.

Training in prompt engineering and use cases tailored to agency needs is required to build capability and confidence in Copilot.

Clear communication and policies are required to address uncertainty regarding the security of Copilot, accountabilities and expectation of use.

Adaptive planning is needed to reflect the rolling feature release cycle of Copilot alongside governance structures that reflect agencies' risk appetite, and clear roles and responsibilities across government to provide advice on generative AI use. Given its infancy, agencies would need to consider the costs of implementing Copilot in its current version. More broadly this should be a consideration for other generative AI tools.

Broader concerns on AI that require active monitoring

There are broader concerns on the potential impact of generative AI on APS jobs and skills, particularly on entry-level jobs and women.

Large language model (LLM) outputs may be biased towards western norms and may not appropriately use cultural data and information.

There are broader concerns regarding vendor lock-in and competition, as well as the use of generative AI on the APS' environmental footprint.

Recommendations

The overarching findings reveal several considerations for the APS in the context of future adoption of generative AI.

Detailed and adaptive implementation

1.1.1.1 Product selection

Agencies should consider which generative AI solution are most appropriate for their overall operating environment and specific use cases, particularly for AI Assistant Tools.

1.1.1.2 System configuration

Agencies must configure their information systems, permissions, and processes to safely accommodate generative AI products.

1.1.1.3 Specialised training

Agencies should offer specialised training reflecting agency-specific use cases and develop general generative AI capabilities, including prompt training.

1.1.1.4 Change management

Effective change management should support the integration of generative AI by identifying 'Generative AI Champions' to highlight the benefits and encourage adoption.

1.1.1.5 Clear guidance

The APS must provide clear guidance on using generative AI, including when consent and disclaimers are needed, such as in meeting recordings, and a clear articulation of accountabilities.

Encourage greater adoption

1.1.1.6 Workflow analysis

Agencies should conduct detailed analyses of workflows across various job families and classifications to identify further use cases that could improve generative AI adoption.

1.1.1.7 Use case sharing

Agencies should share use cases in appropriate whole-of-government forums to facilitate the adoption of generative AI across the APS.

Proactive risk management

1.1.1.8 Impact monitoring

The APS should proactively monitor the impacts of generative AI, including its effects on the workforce, to manage current and emerging risks effectively.

Evaluation Objectives

The evaluation assessed the use, benefits, risks and unintended outcomes of Copilot in the APS during the trial.

The Digital Transformation Agency (DTA) designed 4 evaluation objectives, in consultation with:

- the AI in Government Taskforce
- the Australian Centre for Evaluation (ACE)
- advisors from across the APS designed four evaluation objectives.

Employee-related outcomes

Evaluate APS staff sentiment about the use of Copilot, including:

- staff satisfaction
- innovation opportunities
- confidence in the use of Copilot
- ease of integration into workflow.

Productivity

Determine if Copilot, as an example of generative AI, benefits APS productivity in terms of:

- efficiency
- output quality
- process improvements
- agency ability to deliver on priorities.

Adoption of AI

Determine whether and to what extent Copilot, as an example of generative AI:

- can be implemented in a safe and responsible way across government
- poses benefits and challenges in the short and longer term
- faces barriers to innovation that may require changes to how the APS delivers on its work.

Unintended consequences

Identify and understand unintended benefits, consequences, or challenges of implementing Copilot as an example of generative AI and the implications on adoption of generative AI in the APS.

Evaluation findings

Employee related outcomes

- 77% were optimistic about Microsoft 365 Copilot at the end of the trial.
- 1 in 3 used Copilot daily.
- Over 70% of used Microsoft Teams and Word during the trial, mainly for summarising and re-writing content
- 75% of participants who received 3 or more forms of training were confident in their ability to use Copilot, 28 percentage points higher than those who received one form of training.

Most trial participants were positive about Copilot and wish to continue using it

- 86% of trial participants wished to continue to use Copilot.
- Senior Executive Service (SES) staff (93%) and Corporate (81%) roles had the highest positive sentiment towards Copilot.

Despite the positive sentiment, use of Copilot was moderate

Moderate usage was consistent across classifications and job families but specific use cases varied. For example, a higher proportion of SES and Executive Level (EL) 2 staff used meeting summarisation features, compared to other APS classifications.

Microsoft Teams and Word were used most frequently and met participants' needs. Poor Excel functionality and access issues in Outlook hampered use.

Content summarisation and re-writing were the most used Copilot functions.

Other generative AI tools may be more effective at meeting users' needs in reviewing or writing code, generating images or searching research databases.

Tailored training and propagation of high-value use cases could drive adoption.

Training significantly enhanced confidence in Copilot use and was most effective when it was tailored to an agency's context.

Identifying specific use cases for Copilot could lead to greater use of Copilot.

Productivity

- **69%** of survey respondents agreed that Copilot improved the speed at which they could complete tasks.
- **61%** agreed that Copilot improved the quality of their work.
- **40%** of survey respondents reported reallocating their time for:
  - mentoring / culture building
  - strategic planning
  - engaging with stakeholders
  - product enhancement.

Most trial participants believed Copilot improved the speed and quality of their work

Improvements in efficiency and quality were perceived to occur in a few tasks with perceived time savings of around an hour a day for these tasks. These tasks include:

- summarisation
- preparing a first draft of a document
- information searches.

Copilot had a negligible impact on certain activities such as communication.

APS 3-6 and EL1 classifications and ICT-related roles experienced the highest time savings of around an hour a day on summarisation, preparing a first draft of a document and information searches.

Around 65% of managers observed an uplift in productivity across their team.

Around 40% of trial participants were able to reallocate their time to higher value activities.

Copilot's inaccuracy reduced the scale of productivity benefits.

Quality gains were more subdued relative to efficiency gains.

Up to 7% of trial participants reported Copilot added time to activities.

Copilot's potential unpredictability and lack of contextual knowledge required time spent on output verification and editing which negated some of the efficiency savings.

Whole-of-government adoption of generative AI

61% of managers in the pulse survey could not confidently identify Copilot outputs.

There is a need for agencies to engage in adaptive planning while ensuring governance structures and processes appropriately reflect their risk appetites.

Adoption of generative AI requires a concerted effort to address key barriers.

Technical

There were integration challenges with non-Microsoft 365 applications, particularly JAWS and Janusseal[1], however it should be noted that such integrations were out of scope for the trial.

Copilot may magnify poor data security and information management practices.

Capability

Prompt engineering, identifying relevant use cases and understanding the information requirements of Copilot across Microsoft Office products were significant capability barriers.

[1] JAWS is a software product designed to improve the accessibility of written documents. Jannusseal is a data classification tool used to easily distinguish between sensitive and non-sensitive information.

Legal

Uncertainty regarding the need to disclose Copilot use, accountability for outputs and lack of clarity regarding the remit of Freedom of Information were barriers to Copilot use – particularly in regard to transcriptions.

Cultural

Negative stigmas and ethical concerns associated with generative AI adversely impacted its adoption.

Governance

Adaptive planning is needed to reflect the rolling release cycle nature of generative AI tools, alongside relevant governance structures aligned to agencies' risk appetites.

Unintended outcomes

There are both benefits and concerns that will need to be actively monitored.

Benefits

Generative AI could improve inclusivity and accessibility in the workplace particularly for people who are neurodiverse, with disability or from a culturally and linguistically diverse background.

The adoption of Copilot and generative AI more broadly in the APS could help the APS attract and retain employees.

Concerns

There are concerns regarding the potential impact of generative AI on APS jobs and skills needs in the future. This is particularly true for administrative roles, which then have a disproportionate flow on impact to marginalised groups, entry-level positions and women who tend to have greater representation in these roles as pathways into the APS.

Copilot outputs may be biased towards western norms and may not appropriately use cultural data and information such as misusing First Nations images and misspelling First Nations words.

The use of generative AI might lead to a loss of skill in summarisation and writing. Conversely a lack of adoption of generative AI may result in a false assumption that people who use it may be more productive than those that do not.

Participants expressed concerns relating to vendor lock-in, however the realised benefits were limited to specific features and use cases.

Participants were also concerned with the APS' increased impact on the environment resulting from generative AI use.

> *There's a concern of vendor lock-in as the APS becomes more dependent on this tool.*
>
> **Focus group participant**

> *It's difficult to account for a bias that you are yet to identify.*
>
> **Focus group participant**

> *Copilot could cause myself and colleagues to lack deep knowledge of topics.*
>
> **Pre-use survey respondent**

Approach and methodology

A mixed-methods approach was adopted for the evaluation.

Over 2,000 trial participants from more than 50 agencies contributed to the evaluation. The final report was written based on document/data review, consultations and surveys.

Document/data review

The evaluation synthesised existing evidence, including:

- government research papers on Microsoft 365 Copilot and generative AI
- the trial issue register
- 6 agency-led internal evaluations.

Consultations

It also involved thematic analysis through:

- 24 outreach interviews conducted by the DTA
- 17 focus groups facilitated by Nous Group
- 8 interviews facilitated by Nous Group.

Surveys

Analysis was conducted on data collected from:

- 1,556 respondents in pre-use survey
- 1,159 respondents in pulse survey
- 831 respondents in post-use survey.

Appendix

Methodological limitations

Evaluation fatigue may have reduced the participation in engagement activities.

Several agencies conducted their own internal evaluations over the course of the trial and did not participate in Digital Transformation Agency's overall evaluation.

Mitigations: where possible, the evaluation has drawn on agency-specific evaluation to complement findings.

The non-randomised sample of trial participants may not reflect the views of the entire APS.

Participants self-nominated to be involved in the trial, contributing to a degree of selection bias. The representation of APS job families and classifications in the trial differs from the proportions in the overall APS.

Mitigations: the over and underrepresentation of certain groups has been noted. Statistical significance and standard error were calculated, where applicable, to ensure robustness of results.

There was an inconsistent roll out of Copilot across agencies.

Agencies began the trial at different stages, meaning there was not an equal opportunity to build capability or identify use cases. Agencies also used different versions of Microsoft 365 Copilot due to frequent product releases.

Mitigations: there is a distinction between what may be a functionality limitation of Copilot and when a feature has been disabled by an agency.

Measuring the impact of Copilot relied on trial participants' self-assessment of productivity benefits.

Trial participants were asked to estimate the scale of Copilot's benefits, which may naturally under or overestimate its impact.

Mitigations: where possible, the evaluation has compared productivity findings against other evaluations and external research to verify its validity.

Statistical significance of outcomes

The trial of Microsoft 365 Copilot involved the distribution of nearly 5,765 Copilot licenses across 56 participating agencies. As part of engagement activities – consultations and surveys – the evaluation gathered the experience and sentiment from over 2,000 trial participants representing more than 45 agencies. Insights were further strengthened by the findings from internal evaluations completed by certain agencies. The sample size was sufficient to ensure 95% confidence intervals of reported proportions (at the overall level) were within a margin of error of 5%.

There were 3 questions asked in the post-use survey that were originally included in either the pre-use or pulse survey. These questions were repeated to compare responses of trial participants before and after the survey and measure the change in sentiment. A **t-test** was used to determine whether changes were statistically significant at a 5% level of significance.

A **t-test** is a statistical method to test whether the difference between 2 groups, such as a 'before' and 'after' samples, are statistically significant.

The survey aligned with the APS Job Family Framework and APS job families and classifications were aggregated in survey analysis to reduce standard error and ensure statistical robustness. Post-use survey responses from Trades and Labour, and Monitoring and Audit job families were excluded from reporting as their sample size was less than 10, but their responses were still included in aggregate findings.

For APS classifications, APS 3-6 have been aggregated.

Table A: Aggregation of APS job families for survey analysis

Group	Job families
Corporate	Accounting and Finance Administration Communications and Marketing Human Resources Information and Knowledge Management Legal and Parliamentary
ICT and Digital Solutions	ICT and Digital Solutions
Policy and Program Management	Policy Portfolio, Program and Project Management Service Delivery
Technical	Compliance and Regulation Data and Research Engineering and Technical Intelligence Science and Health

Survey participation by APS classification and job family

Table B: Participation in surveys according to APS level classification

	Percentage of all APS employees	Percentage of pre-use survey respondents	Percentage of post-use survey respondents
SES	1.9	4.7	5.3
EL 2	9.0	20.0	20.2
EL 1	20.8	36.9	34.0
APS 6	23.4	23.4	22.3
APS 5	14.7	8.5	9.6
APS 3-4	26.0	6.0	7.4
APS 1-2	4.2	10.5	1.1

Table C: Participation in surveys according to job family

	Percentage of all APS employees	Percentage of pre-use survey respondents	Percentage of post-use survey respondents
Accounting and Finance	5.1	5.3	3.5
Administration	11.4	9.0	8.9
Communication and Marketing	2.5	4.9	5.8
Compliance and Regulation	10.3	6.6	6.5

	Percentage of all APS employees	Percentage of pre-use survey respondents	Percentage of post-use survey respondents
Data and Research	3.7	9.9	8.3
Engineering and Technical	1.8	1.3	1.5
Human Resources	3.9	5.3	5.0
ICT and Digital Solutions	5.0	19.6	22.3
Information and Knowledge Management	1.1	2.5	1.6
Intelligence	2.4	0.9	2.1
Legal and Parliamentary	2.6	4.1	3.5
Monitoring and Audit	1.5	1.1	1.0
Policy	7.9	13.7	14.4
Portfolio, Program and Project Management	8.3	8.6	7.5
Science and Health	4.2	1.6	2.1
Senior Executive	2.1	2.3	1.5
Service Delivery	25.5	2.7	4.0
Trades and Labour	0.7	0.9	-

Participating agencies

Table D: List of participating agencies by portfolio

Portfolio	Entity
Agriculture, Fisheries and Forestry	Department of Agriculture, Fisheries and Forestry Grains Research and Development Corporation Regional Investment Corporation Rural Industries Research and Development (trading as AgriFutures Australia)
Attorney-General's	Australian Criminal Intelligence Commission Australian Federal Police Australian Financial Security Authority Office of the Commonwealth Ombudsman
Climate Change, Energy, the Environment and Water	Australian Institute of Marine Science Australian Renewable Energy Agency Department of Climate Change, Energy, Environment and Water Bureau of Meteorology
Education	Australian Research Council Department of Education Tertiary Education Quality and Standards Agency
Employment and Workplace Relations	Comcare Department of Employment and Workplace Relations Fair Work Commission
Finance	Commonwealth Superannuation Corporation Department of Finance Digital Transformation Agency
Foreign and Trade Affairs	Australian Centre for International Agricultural Research Australian Trade and Investment Commission Department of Foreign Affairs and Trade Tourism Australia

Portfolio	Entity
Health and Aged Care	Australian Digital Health Agency Australian Institute of Health and Welfare Department of Health and Aged Care
Home Affairs	Department of Home Affairs (Immigration and Border Protection)
Industry, Science and Resources	Australian Building Codes Board Australian Nuclear Science and Technology Organisation Commonwealth Scientific and Industrial Research Organisation Department of Industry, Science and Resources Geoscience Australia IP Australia
Infrastructure, Transport, Regional Development, Communication and the Arts	Australian Transport Safety Bureau
Parliamentary Departments (not a portfolio)	Department of Parliamentary Services
Social Services	Australian Institute of Family Studies National Disability Insurance Agency
Treasury	Australian Prudential Regulation Authority Australian Securities and Investments Commission Australian Charities and Not-for-profits Commission Australian Taxation Office Department of the Treasury Productivity Commission