Australian Government | GovAI

Whole of Australian Government Occupation Coding Service

API capability streamlining the coding of occupation data

Australian Bureau of Statistics

Australian Government | GovAI

Executive Summary

Problem Faced

The Occupation Standard Classification for Australia (OSCA) is used to measure and understand the labour market. Currently, however, coding of occupation data is not consistent across agencies.

Existing labour-intensive manual coding processes take time, are costly, and result in variability in coding practices between agencies. This fragments and lowers the coherence of data, affecting research and policy decisions.

Solution Overview

The ABS has developed a Whole of Australian Government (WoAG) Occupation Coding Service which uses machine learning (ML) to train models for automated coding of occupation data to the latest classifications.

Automated coding: The ML models are trained on ABS data to interpret occupations and tasks described in the vocabulary used by Australians (e.g. the service recognises "brickie" as "bricklayer").

Real-time coding: Supports immediate occupation coding for online surveys and agency website forms.

API access: Uses Application Programming Interfaces to allow for easy and secure data submission and retrieval.

Benefits and Impact

- **Efficiency gains**: By increasing the rates of automated occupation data coding, the manual effort needed is reduced.
- **Multimethod capability**: The hierarchical support vector machine (HSVM) model is more controllable and explainable than large language models (LLMs), but the solution supports both.
- **Improved decision-making**: Consistent data coding enhances the coherence of information available for policymaking.
- **Broader impacts**: Reliable, consistent, and efficient occupation coding, improving comparability across the data landscape which benefits Australian communities.

Target Audience and Stakeholders

The service primarily targets data producing agencies that will use it to efficiently code their occupation data for analysis and dissemination. The audience also includes policy makers who use insights from the resulting data to inform decisions and shape public policies.

ABS staff are responsible for managing the development, maintenance, and promotion of the service.

Consultation and engagement with over 50 commonwealth, state and territory agencies across Australia was integral to shaping the solution. This collaborative approach ensured the service met the diverse needs and expectations of its users.

Risks and Mitigation Overview

- **Data security**: Models function within ABS owned cloud infrastructure which is IRAP Certified at the PROTECTED level, supporting secure handling of data.
- **Model accuracy**: High-quality training data prevents overfitting and ensures model performance. Models will be regularly updated and retrained in a controlled process.
- **User data**: Static models, that do not learn from processed data, mitigate the risks of unintended data retention and privacy concerns.
- **Transparency**: A HSVM model has been used due to its better explainability and lower computational costs.

Use Case Status

Implemented

Use case timeline

- **2022**: Proof of concept
- **2023**: Project funded, and development initiated.
- **2024**: Testing phase with key agencies.
- **June 2025**: Public Beta release.
- **October 2025**: Live release of the service.

Additional Information

The digital infrastructure is built to handle data securely, ensuring compliance with relevant regulations.

The service is available via an API which can be integrated into agency platforms, systems and online forms. It can code single records, small batches of records, or large batches up to many millions of records at a time.

The service has been released and is accepting registrations. It is ready to support real-world data coding for agencies.

Lessons Learned

- Ensuring high-quality training data required significant efforts in cleaning and relabelling.
- HSVMs can be more explainable and have lower computational costs (training the HSVM model cost ~20x less than training an LLM per run).
- Building secure and segregated digital infrastructure to support training and inference is important.
- Static models and not retaining user inputs can mitigate some privacy risks to user data.
- Scalable deployment and broader adoption can be supported through API access.

Contact information

Responsible Entity Name

Australian Bureau of Statistics

Open for Collaboration?

Yes, the ABS is happy to collaborate.

Area of Entity

ABS Business Register and Statistical Standards Branch.

Use Case Contact

coding.capability@abs.gov.au

Use Case Website/s

Whole of Australian Government Occupation Coding Service.
About the Coding Service.

Use Case Owner

Aidan Kent, Coding Redevelopment Section
Craig Lindenmayer, ABS AI Accountable Official

Screenshot/s (example API service requests, responses, etc can be found here: WoAG Occupation Coding Service User Guide)

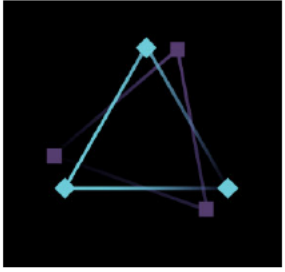
Detailed Overview

Version Control

Version	Date	Author	Description of Changes
1.0	3 Feb 2025	GovAI	Version 1 created
1.1	17 Mar 2025	GovAI	Modified based on feedback
1.2	21 Jul 2025	ABS	Draft – reviewed by ABS Strategic Comms
1.3	22 Jul 2025	ABS	Draft for ABS internal approval
1.4	25 July 2025	ABS	Approved draft

Index

Note: For details about category items in the detailed overview, see *APS AI Use Case Repository Guidance- Guidance for Use Case Owners and Editors*.



Responsible Organisation Category

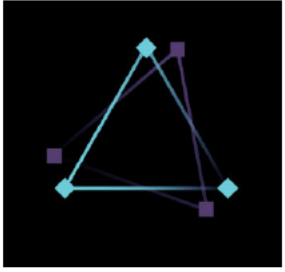
Select the Classification of the Functions of Government – Australia (COFOG-A) 3-digit category that best identifies the functional area associated with your AI use case.

☒ **01 - General Public Services**	Choose an item.013 - General personnel services
	Overall planning and statistical services (COFOG-A 0132)

Scope of the Use Case

Use the dropdown menus below to identify the scope of your use case.

Geographical focus Choose the region for implementation from the dropdown list	National
Primary type of government interaction Choose the type of government interaction from the dropdown list	Government-to-government (G2G)
Cross-features - Sector Indicate if the use case describes a solution that can be used across sectors or in cross-sector scenarios (Yes/No).	Yes
Cross-features – Jurisdiction Indicate if the use case describes a solution that can be used across State/Federal borders or in cross-border scenarios (Yes/No)	Yes

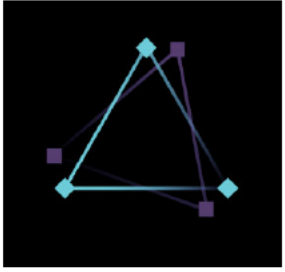


Ethical Considerations

Accuracy, Fairness, Accessibility, Bias and Discrimination	**Accuracy** The models are built to align and code to ABS classifications. Model performance is assessed against: • how the model codes the category labels within the classification, and • coherence of the coded output against existing (legacy) data coding tools and technology to ensure comparability with previous coding methods and therefore data sources. The machine learning models were developed using training data that had been cleaned and prepared for official statistical production. The models were assessed against the Australian Government *Pilot AI Assurance Framework* to assure their quality and fitness for purpose. **Fairness, bias and discrimination** The training data reflects Australian people's descriptions of their jobs and the tasks they perform. Where there were small cohorts or specific demographics undertaking a particular job, the related training records are given increased weight to elevate the confidence of the model to support fair coding outcomes. A ML model is only as good as the data it is trained on. The service is trained on data containing the broadest set of job titles and tasks described using everyday Australian words and phrases. Users of the service can then subsequently describe their job and tasks performed in their own vocabulary, and the model will code appropriately.

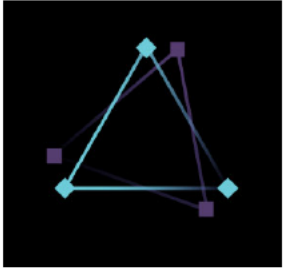
	Accessibility The model is available for integration across government via a secure API. This requires some level of technical skill / knowledge and infrastructure.
Privacy	A Privacy Threshold Assessment (PTA) and an independent IRAP (Infosec Registered Assessors Program) security assessment were carried out to assure the security of environment and secure handling of data. The service was purpose-built for the specific task of coding data and cannot be used for anything outside of this single purpose. The ML models underpinning the service are trained entirely within a dedicated, segregated, and highly secure ABS account. The models are static, so they cannot learn from user input data, and the service does not re-use or retain the input text data from user requests.
Rights of Users	Information about the way models are trained, the context of the text being sought to be coded, and the way that user data is handled, is available on the service website. The Service Terms of Use and Service Level Expectations outline what users can expect from the service. The Service Terms of Use need to be agreed to as part of user registration, including an acknowledgement of fairly using the service, which helps to ensure service stability and availability for all registered users. Instructions for connecting to the Service are in the online User Guide, and a range of security, assurance and support documentation is also provided on our web page, About the Coding Service. Users can contact the ABS for support via email coding.capability@abs.gov.au.

Value of the Use Case



Identify the public value that the solution provides or is expected to provide. Select from the multi-select options.

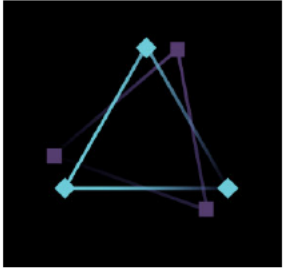
Improved public service This category refers to solutions that enhance the services provided to end users, whether they are citizens or businesses.	☐ Personalised services ☐ Public (citizen)-centred services ☒ Increased quality of public information and services ☒ More responsive, efficient and cost-effective public services ☒ New services or channels
Improved administrative efficiency This category refers to solutions that increase efficiency, effectiveness, and quality while reducing costs within administrative processes, systems, and services.	☒ Cost reduction ☒ Responsiveness of government operation ☐ Improved management of public resources ☒ Increased quality of processes and systems ☐ Better collaboration and better communication ☐ Reduced risk of corruption and abuse of the law by public servants ☐ Greater fairness, honesty and equality enabled
Open government capabilities This category refers to solutions that enhance the level of openness, transparency, engagement, and communication within public organisations.	☒ Increased transparency of public sector operations ☐ Increased public participation in government actions and policymaking ☐ Improved public control of and influence on government actions and policies



AI Process Type

Select the types of tasks within government operations that the AI solution is performing or expected to perform

Supporting Decision Making- Tasks that support formal or informal agency decision-making on benefits or rights.	☐ Taking decisions on benefits ☐ Managing copyright and intellectual property rights
Analysis, monitoring and regulatory research – Tasks that collect or analyse information that shapes agency policymaking.	☐ Information analysis processes ☐ Monitoring policy implementation ☐ Innovating public policy ☐ Prediction and planning
Enforcement - Tasks that identify or prioritise targets of agency enforcement action.	☐ Smart recognition processes ☐ Management of auditing and logging ☐ Predictive enforcement processes ☐ Supporting inspection processes ☐ Improving cybersecurity ☐ Registration and data notarisation processes ☐ Certification and validation processes
Internal management - Tasks that support agency management of resources, including **employee management**, procurement, and maintenance of technology systems.	☒ Internal primary processes ☒ Internal support processes ☐ Internal management processes ☐ Procurement management ☐ Financial management and support
Public services and engagement - Tasks that support the direct provision of services to the public or facilitate communication with the public for regulatory or other purposes.	☐ Engagement management ☒ Data-sharing management ☐ Governance and voting ☐ Payments and international transactions ☐ Supporting disintermediation ☐ Authentication of self-sovereign digital ID services ☒ Service (data) integration ☐ Service personalisation ☐ Tracking of goods and assets along the supply chain



AI Technologies Utilised

Select the types of AI technologies proposed / utilised to deliver the use case.

Reasoning or Knowledge Representation AI systems that store, structure, and process knowledge to make inferences, derive conclusions, or support decision-making.	☐ Knowledge Representation ☒ Automated Reasoning ☐ Commonsense Reasoning
Planning and Optimisation AI techniques that generate, refine, and optimise action sequences or resource allocation to achieve specific goals efficiently.	☐ Planning and Scheduling ☒ Searching ☒ Optimisation
Learning and Adaptation AI systems that identify patterns, extract insights, and improve performance over time based on data.	☒ Machine Learning ☒ Deep Learning (large language models) ☐ Generative AI
Communication and Natural Language Processing AI systems that process, interpret, and generate human language for interaction, comprehension, and automation.	☒ Natural Language Processing (NLP) ☐ Text Generation ☐ Text Mining ☐ Machine Translation
Perception through the Senses AI systems that process and interpret sensory data, such as visual, auditory, or tactile inputs, to understand and respond to their environment.	☐ Computer Vision ☐ Audio Processing
Integration and Interaction with the Environment AI systems that interact with physical or digital environments, including autonomous agents, robotics, and interconnected systems.	☐ Multi-agent Systems ☐ Robotics and Automation ☐ Connected and Automated Vehicles (CAVs)
AI as a Service AI capabilities delivered through cloud-based platforms, offering tools, models, and infrastructure for AI-powered applications.	☒ AI Services (e.g., cognitive computing, machine learning frameworks, bots) ☐ Infrastructure as a Service (IaaS) ☐ Platform as a Service (PaaS) ☒ Software as a Service (SaaS)

Additional Comments or Explanation:	If you have selected any of the subcategories above, feel free to provide more detailed comments or a description of how these elements apply to your specific use case.

Technical Elements

Platform implementation	**Hosting platforms** The ABS used AWS SageMaker end points as an integration standard. This enabled the service to standardise different model topics and ML classifications across that platform. It also enabled lineage and traceability to be captured as part of the process steps for governance, and when developing and validating models. This flowed into testing and assurance of accuracy and reporting for clearance. The ABS implemented AWS standard components and pipelines, which facilitated the creation of custom steps for: • storing and cleaning training data via Confident Learning (pipeline and UI algorithm implementation for data exploration and relabelling), and • supporting model training and deployment. The ABS also implemented AWS Triton inference to reduce costs by having multiple models on a single end point. **Security considerations** Development of the service adopted secure ISM security guidelines right from the start. Solutions were chosen that allowed for better security controls, such as AWS serverless components. Training and inference accounts

	were segregated with controlled access. The external service is only available to registered users who authenticate as part of calling the API. The service itself does not store or re-use external input data. **Integrations with existing platforms** The service is open to external users via an API for automated calling and returning of data, but the service itself does not integrate with external systems at a system level. This design minimises overheads to support adoption and uptake of the service in a financially and operationally sustainable manner for the ABS. **Identity service** User registration and authentication processes control access to the service, while enabling automation and consumption. **Cost model** The service is currently offered openly to registered users under fair use terms and within configurable operational thresholds and allowable limits based on user's advised throughput.
Model / Algorithm used	The service offers HSVM trained models for ANZSCO 2022 and OSCA 2024. The models were trained with millions of de-identified responses to occupation of employment questions in previous Australian Censuses. The project explored the use of different Large Language Models (i.e. DistilBERT) as well as the legacy HSVM being re-implemented as a SageMaker native model, however these were considered less fit for external application.

Data Sources Select the types of data sources used and provide relevant details.	☒ Internal ☐ Public	☐ Third-party ☒ Synthetic
	Details: Internal and synthetic data used to train models. Individual agency data is submitted to (ingested by) the service for coding.	
Risk Assessment and Mitigation Details	**Integration** The project was designed to provide a service that met a wide range of government use cases (identified through user research). The service is currently available via an API only. **Data migration** Not directly applicable. **Prompt ingestion techniques** The service has character limits on the size of ingestible input text records, to control excessive compute draw down and support accurate coding outcomes. A range of parameters are used to refine models, such as n-grams. For example, the models consider individual words, as well as bigrams, or combinations of two consecutive words, to help capture contextual meaning and word order. **Model tuning** Starting with high quality training data, models are iteratively trained following a range of tests that measure classification coverage, conceptual coding accuracy, coding rate, an optimal and a minimum model confidence threshold. Confident Learning and supplementary data are used between model iterations to achieve acceptable coding outcomes against a pre-defined business case (quantified targets).	

	Models are also measured against traditional ML measures such as f-score, precision, recall, accuracy and match rate. **Data protection** The service does not re-use or retain the input text data from user requests, as user data is deleted after processing. The data in transit is encrypted and passed only through ABS owned AWS infrastructure.	
Security and Compliance Frameworks Select the security and compliance frameworks and measures implemented. Provide details or additional artifacts if relevant.	☒ Authority to Operate (ATO) ☒ System Security Plan (SSP) ☒ Security Risk Management Plan (SRMP)	☒ Information Security Registered Assessors Program (IRAP) ☒ Penetration Testing
	Details: Full security assurance processes and documentation suite completed.	
Assurance and Government Frameworks	The service complies with the *Responsible use of AI in government* policy and was assessed against the *Pilot AI Assurance Framework*. The service was self-assessed as low risk under that framework. The service is noted within the ABS's AI Transparency Statement (using AI to enhance operations, and improve productivity by supporting and automating business functions, including machine learning to apply standard classifications to data).	
Record maintenance	Models are developed using an internally created Model Development Workflow, outlining the iterative training, approval and deployment process. An internally created templated / proforma Model Approval Report ensures consistent and comprehensive testing, including against an ABS Model Suitability Framework (including outlining the specific business case) for traceable testing, approval and deployment.	

	Iterative model development is recorded in detailed end-to-end training records for model lineage (outcomes of testing, parameters used, files used, etc.). A Model Register is maintained to ensure tracking of deployed models to the inference accounts (including internal versus external inference accounts) and versions.
Disengagement	Users who do not comply with the Terms of Use, or do not establish connections with the service that follow the User Guide, may have their access revised or revoked. Users will be notified per instance, before and after any action is taken. At the service level, the ABS has an operational procedure in place to revoke all access to the service if required.
Performance Metrics and Results	The coding service performance measure is the number of government agencies registered and using the service (active users). As no user input text is kept or reused, the coding performance of the ML models cannot be monitored or measured by the ABS. To inform this, the ABS engages with select users for feedback on the model's fitness for purpose. A range of API and service level operation performance measures, thresholds, and costs are also monitored.