CAMBRIDGE
UNIVERSITY PRESS & ASSESSMENT

Job Description

Role title	AI Evaluation Scientist
Role reports to	Principal Validation Manager
Direct reports	None
Indirect reports	None
Department	Applied AI Capability Team, Assessment Operations
Location	Cambridge, Hybrid Working
Effective date	10th February 2025

Purpose of the role

To carry out the development, provision and interpretation of statistical analyses and methodologies critical to ensuring the validity, fairness, accuracy and reliability of AI-driven assessment technologies; and to conduct research on innovative evaluation methods to enable advancements in the validation of current and next generation AI-driven assessment technologies.

Key accountabilities

- Conduct small-scale evaluation projects and contribute to large-scale evaluation programmes that support the development and deployment of AI-powered educational technologies through evidencing and enhancing their fairness, accuracy, transparency, and explainability
- Identify and recommend the application of new evaluation methods for better evidencing the validity of white and black box AI models, exploring and implementing these methods where agreed and appropriate
- Contribute to the design of quality assurance processes that enable the organisation to evidence the validity of deployed AI models in-line with regulatory requirements and industry best practice, carrying out components of these processes where appropriate and interpreting their outputs for senior decision-makers
- Ensure that research outputs are appropriately documented and communicated and that any associated documentation is regularly reviewed to keep it accurate and up-to-date
- Contribute to the dissemination of research on AI-enabled language assessment and learning, with a willingness to represent the organisation as a credible expert on AI evaluation methods
- Contribute to the culture of a team that includes experts from a range of disciplinary backgrounds

Key relationships (internal and external)

Internal

- Applied AI colleagues

- Assessment and Propositions product and process owners
- Research Managers
- Data & Analytics Analysts and Managers
- Assessment Managers
- Operations Managers

External

- . None

Role impact (people, policy, financial)

Budgets	**Income** *(that role is responsible for delivering or makes direct contribution to delivery of)* None
	Direct costs *(budgets set and controlled by post holder)* None
	Indirect costs *(post holder authorises spending on budget, delegated by budget holder)* None
People *(The number and grades of colleagues you manage)*	None

What is the major impact of this role on the business?

AI evaluation scientists are central to the provision and interpretation of statistical and psychometric analyses which are integral to the development, deployment and maintenance of the technologies that underpin our entire portfolio of AI-enabled products.

These analyses and approaches are essential as they help ensure that our AI-powered technologies robustly incorporate and demonstrate our validity and assessment principles throughout the product life cycle. They are also key for ensuring that such technologies meet any specifically AI-related regulatory or recognition requirements through enabling the provision and documentation of appropriate and timely evidence.

The expertise that goes into implementing these analyses, coupled with the interpretation and decisions made based on their outcomes, are central to the global recognition of the quality of our world leading products. It ensures that Cambridge Assessment English is able to maximise the benefits of its AI-powered capabilities, underpinning efficiency and growth efforts through the development and deployment of current and next generation AI technologies that positively impact our customers whilst also defending the business against tech-based disruptors.

Decisions and Recommendations

- Planning, scoping and prioritising agreed evaluation tasks in line with business objectives
- Making recommendations for further action (e.g. suitability of model deployment; model retraining; model improvements) within agreed parameters
- Making recommendations for alternative prioritization and evaluation approaches
- Identifying improvements to processes and techniques for further evidencing the quality of AI-powered assessment technologies

Essential knowledge

Essential

- A degree in a science/maths/statistical field with a strong analytic focus (or equivalent work experience)
- Theoretical and practical knowledge of quantitative research methods and inferential statistics
- Theoretical and practical knowledge of a range of data science and analytic techniques for building and interpreting multivariate statistical models

Desired

- A degree with a strong focus on methods and methodologies for evaluating interpretable and non-interpretable AI models

Essential skill(s) & experience

Essential skills

- Ability to conduct multi-variate statistical analyses using either Python or R; Python is desirable
- Ability to provide credible and effective recommendations to cross-departmental colleagues on the basis of these analyses
- Ability to identify and advise on appropriate and feasible evaluation approaches, as well as inferences that can be made from them
- Ability to learn and apply new statistical approaches appropriate for addressing particular challenges that arise in the context of evidencing the quality of cutting-edge AI models
- Effective research/statistical report writing and presentation skills, including the ability to present statistical analyses to both technical and non-technical audiences
- Ability to prioritise and time-manage tasks and larger scale projects
- Ability to work flexibly under pressure and with a good attention to detail
- Ability to collaborate effectively with colleagues from a variety of backgrounds on complex issues, and to seek out innovative solutions
- Understanding of, and commitment to, a non-profit making educational ethos which seeks to expand opportunity for learners and promote the best of international language assessment and learning practice

Essential experience

- Practical experience of evaluating AI models using a variety of metrics and statistical methodologies, whether in an academic or industry context, including methods relating to the domain of explainable AI
- Experience of building and interpreting multi-variate statistical models

Desired experience

- Experience of evaluating AI-powered technologies within the specific context of assessment and learning, whether in an industry or academic context
- Experience of prompt engineering as a means of evaluating AI models
- Experience of collaborating with ML engineers in the development and evaluation of AI models

- Experience of assessment and educational principles and methodologies
- Experience in contributing to published papers in relevant fields
- Experience of establishing and embedding replicable processes that simplify and standardise evaluations within an industry context

Key competencies (values or organisational values)

- **Collaboration:**
  - Proactiveness in communication and action
  - Ability to praise and encourage but also constructively criticize the work of others
  - Ability to resolve differences tactfully
  - Open, honest and constructive approach to cross-team relationships
- **Empowerment:**
  - Diligent, creative, tenacious, results-driven approach to problem solving
  - Clarity in goal setting and prioritization
- **Innovation:**
  - Proactive self-starter with ability to find solutions to problems
  - Learn continuously to improve and make things better for our customers
  - Open to taking on/incorporating goals outside own particular specialist area
- **Responsibility:**
  - Willingness to raise and take ownership of quality issues
  - Willingness to insist on agreed minimum standards and work with others to achieve them
  - Ability to anticipate and raise concerns over delivery timescales

Key competencies (functional)

- Data Literacy
- Statistical Analysis (Descriptive & Inferential)
- Research Methods & Design
- Programming (e.g. Python, R)
- Problem Solving
- Communication