**BILC POLICY RECOMMENDATION**

**on**

**the Safe and Ethical Use of Generative Artificial Intelligence for Language Assessment in the Military**

**Introduction**

In an era marked by rapid advancements in Artificial Intelligence (AI), particularly in generative AI, its applications in language assessment present both opportunities and challenges for language testing organisations in NATO member nations. AI tools are increasingly employed to assist in developing language test items and evaluating language proficiency across the skill modalities. These guidelines aim to establish a framework for the safe and ethical use of generative AI in language assessment for defence purposes. They are intended to complement, not replace, existing national or internal procedures and regulations on the use of AI tools. Given the swiftly evolving nature of AI technologies, this document is considered a living document, subject to regular updates to reflect new developments and best practices in AI governance and language assessment methodologies. By adhering to the principles outlined in this document, the integrity and reliability of language testing processes are upheld while ensuring the protection of sensitive data and the ethical treatment of individuals.

**Understanding Generative AI**

Generative AI (GenAI) refers to technologies capable of producing language-based outputs based on patterns learned from vast datasets. GenAI represents a major advancement beyond traditional, ‘non-generative’ AI systems by its capability to autonomously create new content, such as text, speech, images, or even music. Unlike simpler AI systems that are designed for specific tasks with predefined rules and responses, generative AI operates with a higher level of autonomy and creativity. It can analyse large datasets to understand underlying patterns and then generate outputs that mimic or expand upon those patterns in a way that is often indistinguishable from human-generated content. For example, in language assessment, generative AI tools like ChatGPT, Gemini, MidJourney and MS Copilot can generate coherent text responses based on prompts or questions, adapting to the context and style of language usage learned from extensive training data. This ability to generate new content rather than simply react to predefined inputs distinguishes generative AI, making it particularly powerful for tasks that require creativity, adaptation, and nuanced under­standing of human language and behaviour.

**Application of Generative AI in Language Assessment**

GenAI holds significant potential for enhancing various facets of the language assessment process. Its efficiency and creative capabilities can greatly elevate the production of test content by automating the creation of diverse and sophisticated language prompts, scenarios, and multimedia materials. This automation enables the development of more varied, engaging, and challenging assessments, broadening the range of test items and ensuring comprehensive coverage across different proficiency levels and linguistic domains.

During test administration, GenAI can significantly improve the testing experience through adaptive and interactive environments. AI-driven systems can adapt the selection of test items in real-time, responding to the test taker’s performance. Additionally, AI tools such as Spoken Dialogue Systems (SDS) dynamically engage with test takers in real-time conversations, adjusting their responses based on the test taker’s answers. This personalised approach allows for a more accurate and tailored assessment, better reflecting the individual’s true language proficiency. Moreover, AI tools can enhance the integrity of the testing process by supporting remote proctoring and monitoring test takers, helping to detect and prevent unauthorised use of aids and resources during testing sessions.

In scoring and rating, GenAI technologies offer substantial advantages by automating the evaluation process. These systems can analyse test taker responses across multiple linguistic dimensions, including grammar, vocabulary, coherence, pronunciation, and fluency. This automation not only accelerates the scoring and rating process but also minimises human bias, resulting in more objective and reliable evaluations. The consistency and objectivity provided by AI-driven scoring systems contribute to fairer assessments, with results that are less susceptible to subjective interpretation.

Furthermore, in the realm of test quality control, GenAI can play an important role in the continuous calibration of test items and scoring algorithms. This ongoing calibration ensures that assessments remain valid, reliable, and aligned with established proficiency standards. AI tools can also be used to monitor the consistency of test administration and scoring across different environments, identifying any variations that might affect the accuracy of results. Additionally, GenAI can assist in ensuring the equivalency of different test forms by helping to create and validate multiple versions of a test that are of comparable difficulty and content coverage. This contributes to the overall standardisation, fairness, and quality of language assessments.

**Potential Risks of Generative AI**

While the integration of GenAI systems in language assessment presents numerous opportunities, improper design or irresponsible use of these systems can introduce risks that require careful management and human oversight. These risks include perpetuating biases inherent in AI training data, which could result in unfair evaluations based on demographics or cultural backgrounds. The complexity of AI systems can obscure decision-making processes, raising concerns about transparency and accountability. Additionally, security vulnera­bili­ties in AI models could compromise sensitive data, posing privacy and data security risks for individuals undergoing language testing. Without clear guidelines and robust safeguards, the deployment of GenAI in sensitive military contexts may result in unintended consequences and ethical dilemmas.

**Safety Measures**

GenAI applications create texts and images based on data collected from the internet, which means the accuracy or correctness of the content produced cannot be guaranteed and must always be independently verified. Users of GenAI applications should be aware that information sourced from or uploaded to the internet may be shared with external parties. To address these concerns, specific safety rules for the use of generative AI in military settings should be established:

* **Confidential information**: do not share confidential defence information in queries to AI applications. This includes operational methods, doctrines, personal data, and information about weapon systems, troop readiness and deployments.
* **Email usage**: never use official defence email addresses to create accounts for AI applications.
* **Document handling**: do not input defence documents (such as memos, reports, briefings, manuals, and instructions) into AI applications for generating test items or prompts unless these documents are publicly available.

Language testing organisations typically handle substantial amounts of educational data, including personal information about test takers, examiners, management, and suppliers, as well as confidential information about testing materials, procedures, and assessment grades. To mitigate risks associated with GenAI deployment, defence language testing centres must implement several key measures concerning *privacy* and *data security*.

The protection of sensitive information processed by AI systems requires robust encryption, access controls, and secure storage practices. It is essential to avoid inputting or uploading sensitive or personal materials, such as test taker output from writing tests, into AI tools without obtaining prior consent from the respective test taker. Regular risk assessments should be conducted to identify and address potential vulnerabilities in AI applications. Furthermore, organisations must ensure compliance with relevant legal and regulatory frameworks governing AI use in defence contexts. This involves establishing appropriate policies and procedures for the protection and secure handling of personal data, in accordance with national or defence regulations.

**Ethical Considerations**

Ethical principles, centred around the concepts of fairness, accountability and transparency (FAT), must guide the deployment of GenAI in language assessment within the military. Important considerations to take into account are:

* **Fairness**: it must be ensured that AI tools do not discriminate against individuals based on race, gender, religion, or other protected characteristics. To enhance fairness, it is essential to regularly audit and evaluate AI models for potential algorithmic biases. This involves analysing AI outputs across different demographic groups to ensure equitable treatment. Implementing a diverse dataset during the training phase of the AI tool can help minimise biases. Additionally, incorporating human oversight in reviewing and adjusting AI-generated content ensures that the assessments remain inclusive and culturally sensitive.
* **Accountability**: mechanisms should be established to attribute responsibility for AI-generated outputs and decisions, with clear protocols for monitoring and documenting AI decision-making processes. This includes maintaining logs of AI interactions and decisions, which can be reviewed in case of disputes or errors. In addition, clear protocols should be established and widely communicated for handling cases of fraud or plagiarism. If AI-generated content is found to be plagiarised or used fraudulently, there should be defined steps for investigation, reporting, and remediation. This may involve disciplinary actions, revision of the AI tools or content used, and communication with stakeholders about the issue and its resolution. Assigning specific personnel to oversee AI operations and incorporating feedback loops for continuous improvement ensures that accountability is maintained. Establishing a clear chain of responsibility helps in attributing actions and decisions to either the AI system or human supervisors, thereby promoting transparency and trust.
* **Transparency**: it is crucial to provide clear explanations of how AI assessments are conducted and how decisions are made. Any text, speech, or image generated using AI should be explicitly identified in the documentation, with a proper reference to the AI model used (e.g., ChatGPT-4 by OpenAI) to acknowledge AI’s contribution and avoid plagiarism. Ensuring that AI-generated content is original, does not infringe on intellectual property rights, and complies with copyright laws and ethical standards can be supported through regular audits to detect and prevent plagiarism. Using clear symbols or icons to mark AI-generated content will inform stakeholders about its origin and nature, thereby enhancing transparency and fostering awareness about GenAI’s capabilities and limitations. This practice not only demonstrates ethical accountability but also promotes consistent communication about AI’s role in language assessments. Establishing standardised symbols aligns with international best practices, enhances consistency across BILC member nations, and ensures uniformity in conveying AI’s involvement in language assessment.

**Implementing safety and ethical standards for the use of GenAI**

Successful integration of GenAI in language assessment requires adherence to best practices that ensure ethical standards and effective deployment within defence language testing organisations. These practices emphasise the importance of informed decision-making and ethical conduct in leveraging GenAI to enhance language assessment processes within military settings. By promoting transparency and establishing clear guidelines for AI use, defence language testing organisations can maintain the integrity of language testing and uphold principles of fairness and validity in proficiency assessment.

Effective strategies include:

* **Establishing an AI Ethics Committee** tasked with overseeing the ethical development, deployment, and use of GenAI in language assessment. The committee’s responsibilities include setting guidelines and policies to ensure fairness, accountability, and transparency in AI-driven assessments. It also reviews AI models for algorithmic biases and potential harmful effects, monitors their performance, addresses ethical concerns raised by stakeholders and language testers, and handles cases of plagiarism to ensure the integrity of assessments.
* **Implementing a risk-based categorisation of AI Models** to classify AI tools used in language assessment according to the level of risk they pose to fairness, privacy, and assessment integrity. This categorisation helps better manage the potential risks associated with the use of AI tools and can be broken down into:
	+ *High-Risk Tools:* AI systems involved in scoring and evaluation, where biases could significantly affect test outcomes, should undergo rigorous validation and continuous monitoring to ensure fairness and accuracy.
	+ *Moderate-Risk Tools:* AI tools used for generating test items and prompts, which might inadvertently introduce bias or cultural insensitivity, should be subject to careful oversight and thorough human review.
	+ *Low-Risk Tools:* AI applications that assist in administrative tasks, such as scheduling or test delivery, may require less stringent regulation but should still be monitored for any unintended and harmful consequences.
* **Completing an evaluation form for AI tool usage** to ensure that each AI tool undergoes thorough vetting before being employed. This form allows for a systematic assessment of the tool’s suitability based on key criteria such as practicality, security, reliability, validity, and fairness. It serves as a practical guide for informed decision-making and helps maintain high standards in testing practices. See *Appendix 2* for a sample evaluation form.
* **Providing training and raising awareness** to educate language testers and stakeholders, including test takers, about the capabilities and limitations of GenAI tools used in language assessment. Offer clear guidance on what constitutes acceptable and ethical use of AI-generated content, emphasising adherence to testing protocols and the avoidance of external AI assistance during assessments.
* **Fostering collaboration and consultation** with AI developers and subject matter experts to enhance AI capabilities in providing accurate and reliable assessments of language proficiency. Engaging stakeholders in the development and refinement of AI systems tailored to specific language assessment needs ensures that the AI tools employed are both effective and ethically sound.

**Documentation, Reporting and Case Studies**

Continuous monitoring and evaluation are critical components of the effective deployment of generative AI in language assessment within defence language testing organisations. These processes ensure that AI-driven assessments maintain accuracy, fairness, and reliability over time.

Regularly evaluate the effectiveness of AI in accurately measuring language proficiency across domains such as speaking, listening, reading, and writing. This involves comparing AI-generated assessments against established benchmarks and proficiency standards to validate their reliability. Ongoing *performance monitoring* helps identify both strengths and areas for improvement within AI systems, enabling adjustments that enhance accuracy and consistency in language assessment outcomes.

To support this, develop *evaluation mechanisms* for gathering feedback from language testers, test takers, stakeholders, and users of AI-driven assessment tools. Collecting input on user experiences, satisfaction, and the perceived effectiveness of AI-generated assessments is essential for continuous improvement and refinement of AI models. These feedback loops are critical for identifying potential biases, addressing usability issues, and implementing enhancements that align with NATO or national standards, as well as user expectations.

These continuous monitoring and feedback processes not only ensure the ongoing effectiveness of generative AI in language assessment but also foster a collaborative approach to improving assessment practices within military contexts. By using feedback to refine AI models and assessment methodologies, defence language testing organisations can enhance the reliability and relevance of language proficiency evaluations, thereby supporting mission readiness and operational effectiveness.

Finally, documenting real-world case studies provides concrete examples of successful AI deployment in language assessment within military organisations. These case studies illustrate best practices, lessons learned, and the impact of AI on enhancing language testing capabilities. By compiling and sharing these case studies, military language testing organisations can build a valuable repository of knowledge that supports the safe, ethical, and effective use of AI in language assessments across defence contexts.

**Conclusion**

These guidelines underscore BILC’s commitment to the safe and responsible use of AI in language testing processes within defence organisations. By upholding ethical standards, enhancing transparency, and implementing robust safety measures, the full potential of generative AI can be harnessed to improve language testing efficacy, support mission readiness, and maintain fairness for all stakeholders.

Content partially generated with AI

**Sources**

Selected sources and literature consulted for the guidelines

ChatGPT (2024). *Strategies for implementing ethical AI*. OpenAI; <https://chat.openai.com/chat>

EALTA Special Interest Group: Artificial Intelligence for Language Assessment; <https://sites.google.com/view/ealtasigai/home>

European Commission - *Ethics guidelines for trustworthy AI*; <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>

European Council - *Artificial intelligence (AI) act*; [https://www.consilium.europa.eu/en/ policies/artificial-intelligence/](https://www.consilium.europa.eu/en/%20policies/artificial-intelligence/)

European Union - *Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators*; <https://op.europa.eu/en/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1/language-en>

Item Writing for Language Testing - Generative AI and item writing; [https://itemwriting.co/ generative-ai-for-item-writing/](https://itemwriting.co/%20generative-ai-for-item-writing/)

O’Sullivan, B., et al. (2021). *Validating a Language Test’s Auto-Scoring Model.* British Association for Applied Linguistics – Testing, Evaluation and Assessment Special Interest Group; <https://www.baalteasig.co.uk/>

Perkins, M., Furze, L., et al. (2024). The AI Assessment Scale (AIAS): A Framework For Ethical Integration Of Generative AI In Educational Assessment. *Journal of University Teaching and Learning Practice, 21(06);* <https://doi.org/10.53761/q3azde36>

The Institute for Ethical AI in Education;<https://www.buckingham.ac.uk/research/research-in-applied-computing/the-institute-for-ethical-ai-in-education/>

UNESCO – *Artificial Intelligence*; <https://www.unesco.org/en/artificial-intelligence/>

**Appendix 1:**

**Checklist for the safe and ethical use of GenAI in Language Assessment**

*By following this checklist, language testers, administrators, management, and stakeholders can ensure the safe, ethical, and effective use of generative AI in language assessment within defence organisations. It aims to ensure adherence to ethical standards, mitigate risks, and promote effective deployment of AI technologies.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. | *Policy and Guidelines Development* | Yes | No | N/A |
| a. | Have clear policies been established specifically for the safe use of GenAI in language assessment? |[ ] [ ] [ ]
| b. | Do policies include data privacy considerations in AI deployment? |[ ] [ ]  [ ]  |
| c. | Have guidelines been developed to address the ethical use of GenAI in language assessment? |[ ] [ ] [ ]
| d. | Do policies or guidelines include the principle of fairness in AI deployment? |[ ] [ ]  [ ]  |
| e. | Do policies or guidelines include the principle of transparency in AI deployment? |[ ] [ ] [ ]
| f. | Do policies or guidelines include the principle of accountability in AI deployment? |[ ] [ ]  [ ]  |
| g. | Are policies or guidelines regularly updated to reflect advancements in AI governance and language assessment methodologies? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 2. | ***Understanding Generative AI*** | **Yes** | **No** | **N/A** |
| a. | Have language test developers been trained to understand the capabilities and limitations of GenAI in producing language-based outputs? |[ ] [ ]  [ ]  |
| b. | Are examiners knowledgeable about the capabilities and limitations of GenAI in language assessment? |[ ] [ ]  [ ]  |
| c. | Have stakeholders been informed about the potential and limits of GenAI in language assessment? |[ ] [ ]  [ ]  |
| d. | Is training provided on the ethical considerations and potential risks associated with AI deployment in language testing? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 3. | ***Risk Management and Security*** | **Yes** | **No** | **N/A** |
| a. | Are risk assessments conducted to identify potential biases, security vulnerabilities, and privacy concerns associated with AI models? |[ ] [ ] [ ]
| b. | Are robust data security measures, including encryption and access controls, implemented to protect sensitive information processed by AI systems? |[ ] [ ]  [ ]  |
| c. | Is the use of sensitive, confidential, or classified defence information (such as operational methods, doctrines, and weapon systems) avoided in queries to AI applications? |[ ] [ ]  [ ]  |
| d. | Are official defence email addresses avoided when creating accounts for AI applications? |[ ] [ ] [ ]
| e. | Are defence documents like memos, reports, briefings excluded from use in AI-generated test items unless they are publicly available? |[ ] [ ]  [ ]  |
| f. | Are organisational, national, or NATO safety regulations regarding the use of GenAI consistently followed? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 4. | ***Test Development and Administration*** | **Yes** | **No** | **N/A** |
| a. | Is AI-generated content supervised to prevent biases and cultural insensitivities in test items? |[ ] [ ] [ ]
| b. | Are intellectual property rights respected in the use of AI-generated materials? |[ ] [ ]  [ ]  |
| c. | Is there a process to ensure that AI-generated content does not infringe on existing copyrights or other protected intellectual property? |[ ] [ ]  [ ]  |
| d. | Has a decision been made regarding whether external AI tools are allowed during test sessions? |[ ] [ ]  [ ]  |
| e. | Are test takers informed if the use of external AI tools is prohibited during assessments |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 5. | ***Rating and Scoring Procedures*** | **Yes** | **No** | **N/A** |
| a. | Are validation protocols in place to ensure the accuracy and reliability of AI-generated scores and ratings? |[ ] [ ]  [ ]  |
| b. | Are regular calibration exercises conducted to align AI assessments with established scoring rubrics and proficiency standards? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 6. | ***Ethical Considerations*** *(Fairness, Accountability and Transparency)* | **Yes** | **No** | **N/A** |
| a. | Is there open communication with stakeholders about AI’s role, limitations, and impact on language assessment processes? |[ ] [ ]  [ ]  |
| b. | Is fairness promoted by monitoring AI-generated assessments for biases? |[ ] [ ]  [ ]  |
| c. | Are efforts made to mitigate biases found in AI-generated assessments? |[ ] [ ]  [ ]  |
| d. | Is accountability ensured through thorough documentation of decisions based on AI outputs? |[ ] [ ]  [ ]  |
| e. | Are standardised symbols or icons used to indicate AI involvement in test item production? |[ ] [ ]  [ ]  |
| f. | Are standardised symbols or icons used to indicate AI involvement in scoring and rating assessments? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 7. | ***Monitoring and Evaluation*** | **Yes** | **No** | **N/A** |
| a. | Is the performance of AI systems continuously monitored for accuracy in assessing language proficiency? |[ ] [ ]  [ ]  |
| b. | Is the performance of AI systems continuously monitored for effectiveness across different skill modalities? |[ ] [ ]  [ ]  |
| c. | Are end users informed about the level of precision, responsiveness and accuracy of the AI system? |[ ] [ ]  [ ]  |
| d. | Is feedback regularly solicited from language testers, administrators, test takers, and other stakeholders to continuously improve AI models and assessment methodologies? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 8. | ***Collaboration and Training*** | **Yes** | **No** | **N/A** |
| a. | Is ongoing collaboration with AI developers and SME’s in place to enhance AI capabilities for language assessment? |[ ] [ ]  [ ]  |
| b. | Is regular training provided for language testers and stakeholders on the safe and responsible use of AI in language assessment? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 9. | ***Compliance and Adherence*** | **Yes** | **No** | **N/A** |
| a. | Is compliance with national or international regulations governing AI use in defence contexts ensured? |[ ] [ ]  [ ]  |
| b. | Are compliance procedures periodically reviewed and updated to adapt to evolving AI technologies and regulatory requirements? |[ ] [ ]  [ ]  |
|  |  |  |  |  |
| 10. | ***Documentation, Reporting and Case Studies*** | **Yes** | **No** | **N/A** |
| a. | Are detailed records maintained of AI deployment in language assessments for transparency? |[ ] [ ]  [ ]  |
| b. | Are detailed records maintained of outcomes from AI-generated assessments for auditability? |[ ] [ ]  [ ]  |
| c. | Are comprehensive reports prepared on AI’s impact on language testing efficacy and ethical compliance for management and regulatory review? |[ ] [ ]  [ ]  |
| d. | Are case studies of successful GenAI implementations in language assessment documented? |[ ] [ ]  [ ]  |
| e. | Are case studies shared to demonstrate best practices in AI integration? |[ ] [ ]  [ ]  |
| f. | Are examples included to highlight lessons learned from AI integration in language testing? |[ ] [ ]  [ ]  |
| g. | Are improvements in language testing capabilities through AI integration documented? |[ ] [ ]  [ ]  |

**Appendix 2:**

**Sample AI Tool Evaluation Form**

*Please complete all sections of this form for each AI tool used in your language testing system to ensure that it meets essential criteria for safe and effective use. Provide detailed responses and supporting evidence where applicable.*

|  |  |
| --- | --- |
| 1. | *AI Tool Information* |
|  | AI Tool Name: |  |
|  | Version: |  |
|  | Developer/Provider: |  |
|  | Date of Evaluation: |  |
|  |  |
| 2. | ***Practicality*** | **Yes** | **No** | **Partially** |
| a. | Is the tool practical for the specific language assessment context? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| b. | Does the tool integrate smoothly into existing assessment processes and workflows? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| c. | Are there any operational or technical challenges associated with using the tool? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
|  |  |  |  |  |
| 3. | ***Security*** | **Yes** | **No** | **Partially** |
| a. | Does the tool have robust measures to handle sensitive data securely? |[ ] [ ] [ ]
|  | *Comments:* |  |  |  |
| b. | Are there documented security protocols to prevent data breaches and unauthorized access? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| c. | Is the tool compliant with relevant data protection regulations and standards? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
|  |  |  |  |  |
| 4. | ***Reliability*** | **Yes** | **No** | **Partially** |
| a. | Can the tool consistently produce accurate results across different scenarios? |[ ] [ ] [ ]
|  | *Comments:* |  |  |  |
| b. | Has the tool been tested across diverse test-taker groups to ensure unbiased performance? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| c. | Are there mechanisms in place for regular performance monitoring and maintenance? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
|  |  |  |  |  |
| 5. | ***Validity*** | **Yes** | **No** | **Partially** |
| a. | Does the tool align with established proficiency standards and assessment objectives? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| b. | Does the tool support the development of valid and reliable assessment items? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| c. | Are there validation studies or evidence supporting the tool’s effectiveness in meeting assessment goals? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
|  |  |  |  |  |
| 6. | 1. ***Fairness***
 | **Yes** | **No** | **Partially** |
| a. | Does the tool maintain fairness across diverse test-taker populations? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| b. | Are there mechanisms to identify and mitigate potential biases in the tool’s outputs? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
| c. | Does the tool ensure equitable treatment of all test-takers? |[ ] [ ]  [ ]  |
|  | *Comments:* |  |  |  |
|  |  |

|  |  |
| --- | --- |
| 7. | *Additional Comments* |
| a. | Summary of strengths and weaknesses of the tool |
|  | *Strengths:**Weaknesses:* |
| b. | Recommendations for improvement or further testing: |
|  |  |
|  |  |
| 8. | ***Evaluator Information and Acknowledgement*** |
|  | Evaluator Name: |  |
|  | Position: |  |
|  | Date: |  |
|  | Signature: |  |

**Appendix 3:**

**Glossary of terms** used in relation to AI in language assessment

|  |  |
| --- | --- |
| accountability | The responsibility of organisations and individuals to ensure that AI-generated outputs and decisions are accurate, fair, and ethical. Establishing accountability involves documenting AI processes and decisions, assigning responsibility for AI operations, and maintaining transparent communication with stakeholders. |
| AI Ethics Committee | A panel or board responsible for establishing guidelines, policies, and ethical standards governing the development, deployment, and use of AI technologies in military language assessment. |
| AI-generated content | Text, prompts, or responses produced by generative AI tools like ChatGPT, MS Copilot, MidJourney or Google Gemini, used for or in language assessment tasks. |
| algorithm | A process or set of rules to be followed in calculations or other problem-solving operations, especially by a computer. |
| algorithmic bias | Systematic and repeatable errors in an AI model that create unfair outcomes, such as amplifying prejudices present in the training data or “privileging” one category over another in ways different from the intended function of the algorithm. In language assessment, algorithmic bias can lead to unfair evaluations, particularly if the AI model lacks diversity in its data.  |
| Artificial Intelligence (AI) | The simulation of human intelligence processes by machines, especially computer systems. AI encompasses learning, reasoning, problem-solving, perception, language understanding, and other cognitive capabilities. |
| automation | The use of technology to perform (usually repetitive) tasks without human intervention or continuous supervision. |
| bias | Systematic errors introduced into AI models or assessment processes that unfairly advantage or disadvantage certain individuals or groups based on characteristics such as race, gender, religion, or cultural or professional background. |
| calibration | The process of adjusting and fine-tuning AI models and scoring algorithms to ensure consistent and reliable performance across diverse language assessment tasks and contexts. |
| dataset | A collection of related sets of information composed of separate elements but can be manipulated as a unit by a computer. In AI, datasets are used to train models and make predictions. |
| ethical AI | AI systems designed and implemented in accordance with ethical principles, ensuring respect for human rights, privacy, and fairness in all aspects of operation, including language assessment. |
| ethical principles | Fundamental guidelines governing the ethical use of AI in language assessment, including principles of fairness, accountability, transparency, and respect for individual rights and cultural diversity. |
| fairness | The principle of ensuring that AI tools do not discriminate against individuals based on race, gender, religion, or other protected characteristics. In language assessment, fairness involves creating an equitable testing environment where all test takers have the same opportunity to demonstrate their language proficiency without bias or discrimination. |
| generative AI (GenAI) | A subset of artificial intelligence technology capable of autonomously generating outputs, such as text or speech, based on learned patterns from large datasets. Generative AI goes beyond traditional AI by creating new content rather than simply reacting to predefined inputs. |
| human oversight | Supervision and review conducted by human experts to ensure the ethical and accurate deployment of AI in language assessment tasks, including monitoring for bias and validating assessment outcomes. |
| inclusive | In the context of language assessment and generative AI, inclusive refers to practices and approaches that ensure all individuals, regardless of their background, language, culture, or personal characteristics, have equal access to and fair treatment in the assessment process. |
| integrity | The quality of being honest and having strong moral principles. In the context of data and systems, it refers to the assurance that information is accurate, consistent, and trustworthy, preventing unauthorised modifications. Integrity is crucial in defence language testing organisations, as it guarantees that all data and systems operate as intended without being compromised by internal or external threats.  |
| intellectual property (IP) | A category of property that includes intangible creations of the human intellect. This encompasses inventions, designs, symbols, names, images, and literary and artistic works. In the context of language assessment and AI-generated content, it is crucial for test developers to respect IP rights, ensuring that AI-generated material does not infringe on existing copyrights or other protected intellectual property. |
| language assessment | Evaluation of an individual’s proficiency in a specific language, including skills in speaking, listening, reading, and writing. |
| personal data | Any information related to an identified or identifiable individual, such as names, addresses, identification numbers, or any factors specific to their identity. The protection of personal data is crucial to privacy and compliance with data protection laws. |
| privacy and data security | Measures to protect personal information and ensure secure handling of data used in AI models and language assessments to prevent unauthorised access or misuse. |
| remote proctoring | A method of monitoring and supervising language assessments remotely using AI technology to ensure test integrity and prevent cheating. |
| scoring rubric | A standardised set of criteria used to evaluate and assign scores to language assessment tasks based on predefined proficiency levels and performance indicators. |
| spoken dialogue systems (SDS) | AI systems designed to engage in natural language conversations with users, used in language assessment for interactive tasks such as oral proficiency testing and dialogue-based assessments. |
| subject matter expert (SME) | An individual with specialised knowledge and expertise in a particular field or subject area relevant to military operations, such as cybersecurity, who provides insights and guidance to enhance the development and implementation of AI technologies in another particular field of expertise, such as language assessment. |
| transparency | The practice of providing clear and accessible information about how AI assessments are conducted, how decisions are made, and the role of AI in generating test items or scoring. Transparency fosters trust and understanding among stakeholders and test takers. |
| validation protocols | The specific procedures and methods used to verify the accuracy, reliability, and validity of AI-generated assessments through empirical studies and comparison with established benchmarks. |
| validity | The degree to which the interpretations and decisions made based on test results are meaningful, appropriate, and justified. In the context of AI-driven assessments, validity ensures that the test accurately measures the intended language skills and that the results can be confidently used to make reliable predictions and decisions about a test taker’s language proficiency. |