BRITISH COUNCIL

**BAAL TEASIG 2024**

# AI and its potential for innovation in language assessment

13th September 2024

# A history of innovations

Language learning and assessment solutions have long benefitted from technological advances, especially from Natural Language Processing.

Taggers and parsers

Language models

Essay scoring

Error correction

Machine translation

Speech recognition

# A history of innovations

New tools embracing the latest technology have quickly emerged throughout the years. However, changes to traditional assessment instruments have been slow.

Let's have a look...

# A history of innovations



ELIZA, the first chatbot.

**1965**

*Project Essay Grade (PEG)*, one of the first essay scorers.

**1968**

AI is born.

Alan Turing proposes the "Turing Test".

**1950**

**1960**

**1958**

Perceptron created by Rosemblatt.

**1970**

Rule-based NLP systems.

"AI winter"

Hidden Markov Models and resurgence of neural networks.

**1988**

Statistical methods in NLP.

**1980**

**1990**

**1982**

*Writer's Workbench*, first grammar checker.

NLP takes off.

Statistical Machine Translation

e-Rater, by ETS.

**1999**

**2005**

**2000**

**1997**

Intelligent Essay Assessor.

**2004**

*Criterion Online Writing Evaluation Service* by ETS.

Duolingo.

**2011**

**2010**

**2009**

Grammarly.

SpeechRater, by ETS.

**2015**

Write&Improve.

CAMBRIDGE

**2016**

**2013**

Word2Vec (embeddings).

OpenAI
ChatGPT

**2022**
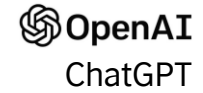
BERT

**2018**

**2017**

Transformers

Generative AI revolution.

**2020**

The *Duolingo English Test* gains popularity.

www.britishcouncil.org

# The assessment landscape is changing

A number of events have challenged traditional assessment methods.

★ A move towards digital assessments after the COVID-19 pandemic.

★ A need for more modern, engaging, realistic and interactive assessment tasks to keep up with the modern world:

  ★ From separate skills to multimodal tasks.

  ★ Communication-focused vs linguistically-focused.

★ Challenges posed by disruptive technology, e.g. ChatGPT.

★ A need to re-think assessment and its purpose.

★ A need for more cost and time-efficient assessments.

# AI in standardised testing

| Organisation | Test | Content | Adaptivity | Scoring |
|---|---|---|---|---|
| ETS | TOEFL iBT Home Edition | | | S+W |
| Pearson | PTE & Versant | | | S+W |
| Duolingo | DET | ✓ | ✓ | S+W |
| Cambridge | Linguaskill | | ✓ | S+W |
| British Council | Primary English Test | | ✓ | S+W |

# How can we use AI to innovate?

# New tasks and constructs

**Content creation**

Generative AI can facilitate content creation (text, images, audio and video). This can help item writers with test design and provide students with practice material on demand.

**New tasks**

AI allows us to generate more realistic tasks that meet the demands of the modern world (e.g. collaborative tasks, workplace situations, etc.) or were too complex to produce and implement (group discussions, simulated scenarios, culturally diverse speakers, etc).

**New constructs**

The ability to create new task types also allows us to assess new constructs that were harder to evaluate in traditional settings, e.g. collaborative problem solving, pragmatic competence, mediation, etc.

# Personalisation

**Tailored content**

AI can tailor content to a student's individual needs, interests, cultural context, proficiency level, pace of learning, preferred task types, progress, strengths and weaknesses, etc.

**Personalised feedback**

AI-powered tools can assess grammar, style, coherence, and other aspects of writing, offering personalised feedback to improve.

**Customised platforms**

AI tutors can provide personalised learning experiences, focusing on specific needs and interests and in a way that is appropriate to the user.

**Examples**    Duolingo (lessons based on the learner's progress, strengths, and weaknesses)
Memrise (personalised vocabulary and tutor)
ChatGPT (can generate, simplify and summarise content + provide tailored feedback)

# Adaptivity

### Adapting to special needs

AI systems can be tailored to accommodate students with special needs, such as learning disabilities, neurodivergence, and speech disorders. E.g. text-to-speech (for students with dyslexia), simplified language or visual aids, slower learning pace (for neurodivergent students), embedded speech therapy tools.

### Adaptive testing

Use AI to dynamically adjust the difficulty of test questions based on the student's performance. Adaptive testing can reduce the number of questions needed to accurately assess a student's abilities, making the testing process more efficient and less stressful.

### Content adaptation

Modifying educational materials to suit the cultural, linguistic, and contextual needs of different student populations. This ensures that assessments are relevant and engaging for students from diverse backgrounds. Adaptation can make content more relatable and easier to understand, and therefore more effective.

# Multimodality

Multimodal assessments make use of various forms of input, such as voice, video, and text, to evaluate a student's communication skills in a more holistic manner.

Multimodality can be achieved by leveraging the generative capabilities of AI together with ASR and text-to-speech technology.

# Linguistic diversity

### Low resource languages

AI can support the assessment of low-resource languages by leveraging transfer learning and multilingual models. Most LLMs can understand and produce content in many languages as well as translating and switching between them.

### Code-switching

AI can detect and analyse code-switching behaviour, providing insights into a student's language proficiency and usage patterns.

### English as a Lingua Franca

AI can help create scenarios that reflect the cultural diversity of English speakers worldwide.

### Non-Native and Regional Accents

Many text-to-speech systems can generate speech with different regional and international accents. This facilitates the creation of tasks that expose learners to a wide variety of accents and help them improve listening skills in more realistic settings.

# Natural interfaces

## Conversational agents

AI-powered chatbots can engage in natural conversations using text-only interfaces or speech. This provides a more familiar and pleasant experience than traditional testing scenarios.

## Personal devices

Mobile phones can offer a familiar interface for interacting with an AI system whereas other devices like VR headsets can provide immersive experiences that can make assessment more realistic, such as navigating a virtual city or participating in a simulated conversation.

## Emotion and sentiment analysis

AI can analyse the emotional tone and sentiment in spoken or written language, providing insights into the user's confidence and engagement levels. This can help create more supportive and motivating learning environments.

# Interaction

AI can enhance interaction by creating engaging, real-life scenarios through interactive tasks and chatbots. These technologies make it easy to design role-playing tasks where students can interact with one or more AI agents.
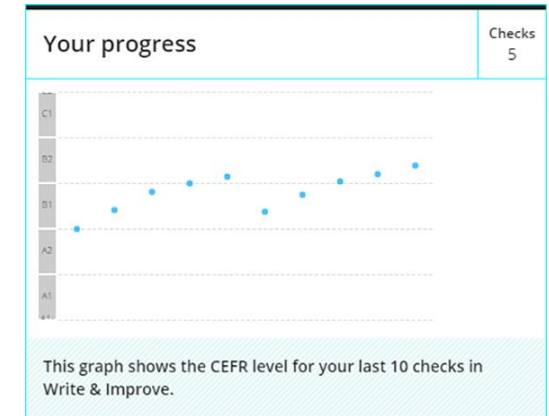
Such scenarios can provide an ideal environment to assess interactional competence and interpersonal communication abilities.

# Continuous assessment

AI can facilitate a dynamic and ongoing evaluation of a student's language skills, providing a more holistic and accurate picture of their progress. This approach could replace traditional examinations, reducing the need for tests and alleviating student anxiety and negative washback effects.

*Open Learner Models* (OLMs) provide students with a transparent view of their learning progress, strengths, and areas for improvement. By continuously updating based on the student's performance, OLMs encourage self-regulation and motivate learners to take an active role in their education.



**97%**

You're acing French lessons today!

You got an average accuracy of **97%** and have practiced **33** words so far

SHARE MY PROGRESS

CONTINUE



Your progress

Checks
5

This graph shows the CEFR level for your last 10 checks in Write & Improve.

# Considerations

# Feasibility

## Technical

**Expertise – AI literacy – Skills**
Do we have the knowledge and skills to implement these solutions ourselves? Do we need a team? Do we need to train a model or can we use generative AI tools? Do we have model training or prompt engineering skills?

**Infrastructure – Complexity**
What infrastructure is required? (networks, services, devices) How complex is the implementation?

## Financial

**Cost**
What is the cost of the solution? Is it within budget? Is it sustainable?

**Resources**
What resources are required by developers and end users? Are they available? Are they accessible?

# Validity

The validity of AI models can be undermined by a number of issues.

| Technology | Key issues |
|---|---|
| Spoken dialogue systems | ASR accuracy (WER), NLU accuracy, latency, appropriateness |
| Automatic scoring | Accuracy, features, training data, transparency, biases, impact |
| Automated feedback | Accuracy (precision, recall), effects |
| Remote proctoring | Underlying technology, security, privacy, vulnerabilities |
| Large Language Models | Accuracy, hallucinations, bias, reproducibility |
| Generative AI | Appropriateness, bias, originality |

Adapted from **Advancing Language Assessment with AI and ML–Leaning into AI is Inevitable, but Can Theory Keep Up?**
Xi (2023)

# Test security

Interacting with a machine (either by speaking, writing or scoring) can open the door to potential "cheating" or "gaming".

Consider forms of online proctoring or supervision (e.g. via the front camera on mobiles). The non-deterministic nature of generative AI tools also contributes to the security of test items, as they differ across test takers and cannot be memorised.

# Ethics

**Ethics by Design:** Think about and address potential ethics concerns during the whole system development lifecycle.

| | | |
|---|---|---|
| Respect for Human Agency | Privacy and Data Governance | Fairness |
| Individual, Social and Environmental Well-being | Transparency | Accountability and Oversight |

An AI system that violates these principles/values is considered **unethical**.

# Innovation in practice
# A case study at the British Council

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is…

Innovative

Multimodal

Interactive

Natural

Adaptive

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is...

**Innovative**

Multimodal

Interactive

Natural

Adaptive

★ New task and constructs

**Challenges:** No initial data to train the required assessment models.

**Solution:** Use LLMs for assessment.

**Outcome:** LLMs are vague and not as consistent or reliable as human raters.

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is…

Innovative

**Multimodal**

Interactive

Natural

Adaptive

★ New content

**Challenges:** Generate text, speech, images and videos.

**Solution:** Use generative AI tools.

**Outcome:** Generating high-quality content requires good prompt engineering skills and many iterations.

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is…

Innovative

**Multimodal**

Interactive

Natural

Adaptive

★ Spoken dialogue

**Challenges:** Accurate ASR for non-native speakers.

**Solution:** Use a state-of-the-art ASR service.

**Outcome:** Transcriptions are not accurately punctuated and require post-processing.

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is...

Innovative

Multimodal

**Interactive**

Natural

Adaptive

★ Interacting with multiple bots simultaneously

**Challenges:** Generate multiple bots that can successfully interact with the user and one another.

**Solution:** Use an LLM, defining multiple agents and personas.

**Outcome:** Looks promising but requires extensive testing.

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is…

Innovative

Multimodal

Interactive

Natural

Adaptive

★ On-topic non-scripted natural conversations

**Challenges:** Ensure conversations stay on topic and limited to the input material.

**Solution:** Careful prompt engineering and guardrails.

**Outcome:** Conversations now stay on track.

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is…

**Innovative**

**Multimodal**

**Interactive**

**Natural**

**Adaptive**

★ Natural turn-taking

**Challenges:** The difficulty detecting end of turns and text-to-speech latency makes conversation sound unnatural.

**Solution:** Tweak the end-of-turn pause threshold.

**Outcome:** Moderate improvement to turn-taking. Interruptions are still not possible (too complex).

# A dialogue system for goal-oriented tasks

Build a platform to practise speaking skills that is…

Innovative

Multimodal

Interactive

Natural

**Adaptive**

★ Support or challenge the student based on their performance

**Challenges:** Embed an adaptive mechanism into dialogue generation.

**Solution:** Dynamic prompting?

**Outcome:** Unknown yet.

# Thank you!

Mariano.Felice@britishcouncil.org